Identifying Unconditional Quantile Impulse Responses with an Application to Growth-at-Risk*

Robert Wojciechowski†

October 31, 2025

Latest Version

Abstract

I introduce Generalized Quantile Local Projections (GQLP), a novel methodology for identifying structural quantile impulse responses. Unlike existing methods that estimate conditional quantile effects, GQLP identifies causal effects on unconditional quantiles while still exploiting controls for identification. This distinction is crucial in macroeconomics. For instance, when studying output growth, low unconditional quantiles correspond to actual recessions rather than merely periods of lower-than-expected growth relative to control variables. I develop a general simulation algorithm to recover true structural quantile responses in analytically intractable models. I conduct Monte Carlo experiments demonstrating that GQLP successfully recovers structural quantile impulse responses, whereas conventional conditional quantile methods can yield misleading conclusions in the presence of control variables even when the true structural shock is observed. In a growth-at-risk application, I show that financial risk shocks have strongly asymmetric effects. Using timing restrictions for identification, I find that a one-standard-deviation credit shock reduces industrial production growth by 2 percentage points in the lower tail versus 0.5 percentage points at the median. In other words, GQLP reveals left-tail-to-median response ratios of four-to-one, double those found using conventional Quantile Local Projections, indicating that standard methods underestimate the effect of financial shock on downturns. These findings suggest that stabilizing financial conditions can help prevent painful recessions without sacrificing growth during expansions.

JEL Classification: C32, E44, C14, E32.

---

*Acknowledgments: I am indebted to Christian Brownlees and Andrea Caggese for their support and guidance. I would also like to thank Geert Mesters, Barbara Rossi, Vladislav Morozov, Valeria Gargiulo and the participants of the 48th Symposium of the Spanish Economic Association, the Euro Area Business Cycle Network: Advances in Local Projections and Empirical Methods for Central Banking conference, the 2024 European Winter Meeting of the Econometric Society, CREi Macroeconomics Lunch, and the 2025 American Economic Association Annual Meeting for comments and discussion. All remaining errors are my own.

†Universitat Pompeu Fabra, robert.wojciechowski@upf.edu

1 Introduction

Modern macroeconomics increasingly recognizes that shocks may affect the entire distribution of economic outcomes, beyond just the mean, and in particular the tails of the distribution. Explicit attention to tail risks has also become commonplace in policy. For instance, at the September 2025 FOMC meeting, the Committee motivated a federal funds rate cut by noting that "downside risks to employment have risen."

To understand the drivers of tail risks, researchers increasingly use quantile regression to measure how quantiles of the outcome distribution respond to shocks. However, a methodological complication arises: the inclusion of control variables for causal identification transforms the analysis from unconditional to conditional quantiles. This distinction is central across many areas of macroeconomics. When studying fluctuations in output growth or inflation, researchers are often interested in the drivers of extreme outcomes—such as recessions or destabilizing inflationary episodes. Unconditional tail quantiles directly correspond to these extreme outcome periods. In contrast, conditional tail quantiles correspond to periods that are extreme relative to what the control variables predict and thus do not always map onto actual crisis periods. For instance, a conditionally low quantile of output growth may correspond to a period of underperforming growth given favorable economic conditions, yet still occur during an expansion.

Quantile regression estimates coefficients using observations at specific quantiles of the conditional outcome distribution. Adding controls changes that conditional distribution, altering which observations are "local" to a given quantile and thereby shifting the estimated coefficients. In linear regression, adding controls orthogonal to the treatment leaves average treatment effects unchanged. However, in quantile regression even the addition of statistically independent controls can change the estimated quantile treatment effects. This creates a trade-off: unconditional quantiles without controls have clear economic interpretation but potentially biased estimates; including controls helps achieve causal identification but at the cost of shifting the analysis to conditional quantiles and thereby losing economic

interpretability.

This paper develops a new methodology that resolves this tension between causal identification and economically meaningful quantile interpretation. I introduce Generalized Quantile Local Projections (GQLP), which build on the generalized quantile regression of Powell (2020), the local projections approach of Jordá (2005), and the potential outcomes framework for time series of Rambachan and Shephard (2021). GQLP explicitly distinguish treatment variables (whose effects we seek to measure) from control variables (used solely for identification), allowing estimation of causal effects on unconditional quantiles while still exploiting controls for identification. Beyond disentangling the effect of including controls on identification versus interpretation, GQLP has three additional advantages. First, the framework allows to define quantile impulse responses (QIRs) as responses to one-off shocks, ensuring direct comparability with conventional mean impulse responses. Second, it imposes no linearity assumptions on the structural quantile function, accommodating complex nonlinear responses that depend on both baseline and counterfactual treatment values. Third, while this paper emphasizes identification through timing restrictions and control variables, the framework naturally extends to instrumental variable designs.

Growth-at-risk (GaR) provides a natural application for GQLP. GaR is defined as the fifth percentile ($\tau = 0.05$ quantile) of future growth, representing a pessimistic scenario that materializes five percent of the time. A large empirical literature documents that financial conditions are key drivers of GaR (Adrian et al. 2019). This literature is primarily focused on forecasting (Plagborg-Møller et al. 2020; Brownlees and Souza 2021; Chuliá et al. 2024), with policymakers using declining GaR forecasts as early warnings of rising distress in the economy (Prasad et al. 2019). Theoretical macrofinance models attribute the heterogeneous effect of financial conditions on upside versus downside risk to occasionally binding financial constraints, balance sheet interactions, and amplification mechanisms that can endogenously cause occasional financial crises (He and Krishnamurthy 2019; Gertler et al. 2019; Brunnermeier and Sannikov 2014).

I identify the causal effects of financial risk shocks on the unconditional quantiles of industrial production growth via timing restrictions, controlling for macroeconomic, financial, and monetary policy variables. The results indicate large and persistent output losses in low-growth environments, while effects at the median and upper quantiles are considerably smaller. For example, a one-standard-deviation credit risk shock reduces growth by up to 2 percentage points in the lower tail, compared to approximately 0.5 percentage points at the median. Relative to conventional Quantile Local Projections (QLP), GQLP estimates uncover substantially stronger asymmetries between the lower tail and the median, with response ratios of roughly four to one, compared to two to one under QLP. The difference arises because QLP identifies effects on conditional quantiles, whereas GQLP recovers effects on unconditional quantiles, preserving the interpretation of quantiles as corresponding to phases of the business cycle. These results are in line with the fact that conditional quantile methods may underestimate adverse impacts of financial shocks during economic stress periods. I rationalize this finding in the Monte Carlo section using a nonlinear equilibrium model of Gertler et al. (2019), where conditioning on bank sector health masks the asymmetry of responses of quantiles of output growth to a capital quality shock.

To establish the theoretical foundation for quantile analysis, I prove a quantile invariance theorem. The theorem states that any dependent process with a structural Wold representation and Gaussian innovations features quantile impulse responses that are identical across all conditional and unconditional quantiles and equal to the mean impulse response. This means that interesting quantile dynamics require departures from linearity, Gaussianity, or stationarity. To evaluate the GQLP framework in non-linear settings, I develop a general algorithm to recover the true QIRs from simulating the model in cases where they are otherwise analytically intractable. The algorithm generates potential outcomes by experimentally fixing the treatment variable to values from a grid. Doing so over many histories of all other model variables recovers the distribution of the potential outcomes. This enables computation of the true value of the QIRs or other statistics of interest. This method follows directly from

the definition of potential outcomes and is readily applicable to most structural models. It remains computationally feasible for unconditional quantile responses since they require perturbing only one variable at a time. Conditional quantile responses are substantially more expensive to recover as each additional variable adds another dimension to the simulation grid, causing an exponential increase in computation time.

I apply my simulation algorithm to both an endogenous volatility structural vector autoregression (SVAR) and a nonlinear dynamic stochastic general equilibrium (DSGE) model of Gertler et al. (2019). I show that in these models shocks have asymmetric effects across the distribution, with responses differing substantially between upper and lower quantiles. Moreover, the structural quantile functions exhibit nonlinear functional forms at some horizons for some quantiles. I then compare the performance of QLP with GQLP in a Monte Carlo study. Using the SVAR I show that GQLPs with timing restrictions produce estimates equivalent to a regression on the unobservable structural shock only—a property shared by standard SVARs and local projections but not by QLPs. This ensures GQLP-estimated QIRs identify the causal effect of a structural shock to the treatment on the quantile of the outcome. Using the DSGE model I demonstrate the importance of distinguishing between conditional and unconditional quantile responses in structural models. I show that GQLP and QLP without controls both capture unconditional effects of capital quality shocks on output across all states of banking sector health. However, QLP with controls estimates conditional responses within given states of financial vulnerability, suggesting symmetric effects across quantiles and masking the amplified downside responses during financial crises. Since capital quality shocks are independent, controls are unnecessary for identification, yet researchers routinely include them with plausibly exogenous shocks. While this practice is innocuous when studying average effects, it fundamentally alters the economic interpretation of quantile effects.

This work contributes to several strands of literature. I extend the generalized quantile regression of Powell (2020) from the cross-section quantile treatment effects literature

(Chernozhukov and Hansen 2013; Angrist et al. 2006) to impulse response analysis. In doing so, I build on the potential outcomes for time series framework (Rambachan and Shephard 2021; Angrist and Kuersteiner 2011; Angrist et al. 2018). The methodology advances the local projections literature (Jordá 2005; Jorda and Taylor 2025) by enabling identification of unconditional quantile impulse responses. By establishing a novel causal link between financial risk shocks and the unconditional tail risks of output growth, I contribute to an emerging empirical literature that studies the structural drivers of GaR (Adrian et al. 2020; Adrian et al. 2022; Loria et al. 2025). My empirical findings can be rationalized by the macrofinance literature (Kiyotaki and Moore 1997; Brunnermeier and Sannikov 2014; He and Krishnamurthy 2019; Gertler et al. 2019) and are related to work on the cyclicality of volatility and the skewness of macroeconomic variables (see Bloom (2014) for a general review, and Cascaldi-Garcia et al. (2023) for a review of measures of uncertainty, volatility and risk).

Various alternative definitions and estimators of quantile impulse responses exist in the literature, but none addresses the fundamental challenge of identifying causal effects on unconditional quantiles while using control variables for identification. The QLP framework in Linnemann and Winkler (2016), Adrian et al. (2019), Adrian et al. (2022), Jordà et al. (2022), and Bochmann et al. (2023) is the closest to my approach. This framework recovers the QIR from local projection coefficients estimated using the quantile regression of Koenker and Bassett 1978, with estimation done separately for each quantile and horizon. In the absence of control variables, there is no difference between GQLP and QLP. Chavleishvili and Manganelli 2024 achieve identification by imposing timing restrictions on a recursive quantile vector autoregressive model. The reported QIRs assume a realization of a median sample path for the shock variable over the response horizon. Montes-Rojas 2019 and Lee et al. 2021 use a SVAR model to identify a structural shock since their multivariate quantile models are reduced-form. The QIR proposed by Montes-Rojas 2019 describes the cumulative impact of a series of shocks, not a one-off shock, because persistent realizations of lower (or

upper) quantiles are assumed in its construction. Han et al. 2022 and Jung and Lee 2022 study QIRs in models where the quantile itself is autoregressive, as in the CAViaR model of Engle and Manganelli 2004. In the applied literature, Mumtaz and Surico 2015 estimate structural QIRs of output growth in response to monetary policy shocks using the quantile autoregressive-distributed lag model of Galvao et al. 2013. Loria et al. 2025 estimate QIRs within a conventional local projections framework, but using as dependent variables the fitted quantiles of year-ahead output growth obtained from a quantile regression on current macro-financial conditions.

The remainder of the paper is structured as follows. Section 2 introduces the econometric framework. Section 3 motivates the frameworks using a simulation study. Section 4 contains the empirical analysis. Concluding remarks follow in Section 5.

2 Econometric framework

2.1 Potential Outcomes

My generalized quantile local projections (GQLP) framework builds upon the literature on local projections (Jordá 2005), quantile treatment effects in the presence of covariates (Powell 2020), and potential outcomes for time series (Rambachan and Shephard 2021).

Let $Y_{i,t}$ denote a scalar outcome variable and $Y_{j,t}$ the scalar treatment variable, both of which are part of a multivariate process $Y_t \in \mathbb{R}^k$. Let the vector X_t denote a finite history subset of the sigma-algebra generated by Y_t, i.e. $X_t \subset \mathcal{F}_t^Y$ where $\mathcal{F}_t^Y = \sigma\big(\{Y_s : s \leq t\}\big)$. Let the vector $W_t \in \mathbb{R}^k$ denote the assignments (or shocks in the terminology of macroeconomics). Throughout, capital letters represent random variables, while lowercase letters denote the fixed values these variables may take.

Following Rambachan and Shephard (2021) I define a potential outcomes process as a multivariate time series process that satisfied assumptions 1 and 2.

Assumption 1 (Non-anticipating Potential Outcomes)**.** *For each $t \geq 1$ and all sequences*

$\{w_s\}_{s\geq 1}$ *and* $\{w'_s\}_{s\geq 1}$ *in* $\mathcal{W}$,

$$Y_t(w_{1:t}, \{w_s\}_{s\geq t+1}) = Y_t(w_{1:t}, \{w'_s\}_{s\geq t+1}) \text{ almost surely.}$$

I thus denote the time-t potential outcome as $Y_t(w_{1:t})$*, where* $Y_t(w_{1:t}) \in \mathcal{Y} \subseteq \mathbb{R}^k$.

Assumption 2 (Sequentially Probabilistic Assignment Process)**.** *The assignment process satisfies* $0 < \Pr(W_t = w \mid \mathcal{F}_{t-1}^Y) < 1$ *almost surely for all* $w \in \mathcal{W}$*. The probability structure is determined by the filtered probability space generated by* $\{W_t, \{Y_t(w_{1:t}) : w_{1:t} \in \mathcal{W}^t\}\}_{t\geq 1}$.

I focus on the dynamic effects on outcome variable i to a time t change in treatment variable j, thus I use the shortcut notation:

$$Y_{i,t+h}(w_j) := Y_{i,t+h}(W_{1:t-1}, W_{1:j-1,t}, w_j, W_{j+1:k,t}, W_{t+1:t+h}).$$

$Y_{i,t+h}(w_j)$ represents the potential outcomes that $Y_{i,t+h}$ would take had assignment $W_{j,t}$ been experimentally (exogenously) fixed to $W_{j,t} = w_j$.[1] The potential outcome is a random variable which depends on assignments up to time $t + h$. The set of potential outcomes includes the observed outcome, which in this shorter notation is denoted $Y_{i,t+h} \equiv Y_{i,t+h}(W_{j,t})$. Similarly, a potential outcome can be defined in terms of experimentally fixing $Y_{j,t} = y_j$ as $Y_{i,t+h}(y_j)$, in which case the observed outcome is $Y_{i,t+h} \equiv Y_{i,t+h}(Y_{j,t})$.

Before moving on to identification of quantile impulse responses, it is instructive to consider what assumptions are needed to identify the structural mean impulse response when the assignments W_t are unobservable. This is a standard setting in macroeconomics, where researchers often observe only a vector of endogenous variables Y_t but not the underlying assignments W_t. The typical model used in such settings is the structural vector autoregression. SVAR(1) assumes that the potential outcome process satisfies $A_0 Y_t(w_{1:t}) = w_t + A_1 Y_{t-1}(w_{1:t-1})$.

[1] Throughout I focus on the potential outcomes for a fixed value value of a variable in time t. As such the notation does not make explicit the fact that both the timing and the horizon of the treatment matter, for instance $Y_{i,t+h}(w_j)$ and $Y_{i,(t+1)+(h-1)}(w_j)$ may differ even though both occur in period $t + h$, because $Y_{i,t+h}(w_j)$ does not constrain the assignment in period $t + 1$ to equal y_j.

The following assumptions are closely related but not identical to those used for identification in the SVAR literature, and although they can be applied in the SVAR setting they are more general and can be used to study identification in other (also nonlinear) settings.

Assumption 3 (Independent Assignments)**.** *Assignments are independent across time and across units. That is, for all $t \neq s$, $W_t \perp\!\!\!\perp W_s$, and for all $i \neq j$, $W_{i,t} \perp\!\!\!\perp W_{j,t}$.*

Assumption 4 (Deterministic Potential Outcomes)**.** *Potential outcomes $Y_t(w_{1:t})$ are deterministic functions of the assignment sequence for all $t \geq 1$ and $w_{1:t} \in \mathcal{W}^t$.*

Assumption 5 (Partial Invertibility)**.** *The outcome $Y_{j,t}$ is a function only of the structural shock $W_{j,t}$ and the time-t observables X_t, i.e., $Y_{j,t} = g_j(W_{j,t}, X_t)$. Furthermore, the inverse $W_{j,t} = g_j^{-1}(Y_{j,t}, X_t)$ exists.*

Assumption 3 ensures that the assignments can be interpreted as structural shocks. It is a stronger version of the uncorrelatedness of structural shocks assumptions commonly used in the SVAR literature (if structural shocks are Gaussian the two assumptions are equivalent). Assumption 4 ensures that the potential outcomes (including the observed outcome) are a function of past and present structural shocks only, i.e. there are no "external" sources of randomness driving the potential outcomes. Assumption 5 ensures that although the structural shock of interest $W_{j,t}$ is not observed, it can be recovered from observable data. This is a strong assumption, but it is weaker than the full invertibility assumption used in the SVAR literature – which requires that all the structural shocks W_t can be recovered. Note that assumptions 3 and 5 imply that $Y_{i,t+h}(y_j) \mid Y_{j,t}, X_t \sim Y_{i,t+h}(y_j) \mid X_t$. In other words, since the treatment variable is assumed to be only a function of observables and an independent assignment, by conditioning on the appropriate variables the treatment is conditionally independent of the potential outcomes.

Theorem 1 (Impulse Response Identification with Unobservable Assignments)**.** *Under*

Assumptions 1 through 5, the causal marginal filtered treatment effect is identified as:

$$\frac{\partial}{\partial y_j}\mathbb{E}[Y_{i,t+h} \mid Y_{j,t} = y_j, X_t] = \mathbb{E}\left[\frac{\partial Y_{i,t+h}(w_j)}{\partial w_j} \cdot \frac{\partial g_j^{-1}(y_j, X_t)}{\partial y_j} \mid X_t\right].$$

Theorem 1 is similar to the theorem 10 in Rambachan and Shephard (2021), except that the addition of the partial invertibility assumption 5 makes the interpretation of the identified impulse response more straightforward as it measures the causal effect of an intervention to a single scalar assignment rather than the effect of "simultaneously shifting all assignments from time $t = 1$ to t". The proof of theorem 1 is in the appendix section A1.1. This theorem says that the structural impulse response can be identified without having to observe $W_{j,t}$. Moreover, it takes a particularly simple form made up of two terms, where the second term $\frac{\partial g_j^{-1}(y_j, X_t)}{\partial y_j}$ does not depend on the horizon of the response nor the dependent variable and is a constant in linear settings. In particular, in a SVAR it is easy to show that $\frac{\partial g_j^{-1}(y_j, X_t)}{\partial y_j} = 1$ and in Local Projections settings it can be normalized to 1 as outlined in Plagborg-Møller and Wolf (2021).

2.2 Quantile Impulse Response definition

In what follows, I will assume that for a fixed treatment y_j and for each horizon h, $Y_{i,t+h}(y_j)$ has a structural quantile function (SQF) denoted $q_h(\tau \mid y_j)$. Notably, covariates X_t do not enter into this SQF, which distinguishes it from the conditional covariates SQF denoted $q_h(\tau \mid y_j, x)$.

Assumption 6 (Structural Quantile Function)**.** *For each $Y_{i,t+h}(y_j)$, there exists a structural quantile function $q_h(\tau \mid y_j)$ that is non-decreasing and left-continuous in $\tau \in [0, 1]$.*

Assumption 6 implies that the structural quantile function exists and does not change with time, ruling out models that feature structural breaks or other violations of stationarity. Each

potential outcome can be related to its structural quantile function as follows:

$$Y_{i,t+h}(y_j) = q_h(U_{i,t+h}(y_j) \mid y_j), U_{i,t+h}(y_j) \sim \mathsf{Uniform}[0,1].$$

$U_{i,t+h}(y_j)$ is responsible for heterogeneity of outcomes among time periods with the same treatment y_j. I refer to it as a rank variable as it determines the placement in the h-periods-ahead outcome distribution for a given treatment y_j. $U_{i,t+h}(y_j)$ contains information up to time $t+h$.

The goal of this paper is to identify the structural quantile impulse response, defined as:

$$\mathsf{QIR}_\tau(h) = \frac{\partial q_h(\tau \mid y_j)}{\partial y_j}. \tag{1}$$

If the SQF is linear i.e. $q_h(\tau \mid y_j) = \alpha_h(\tau) + \beta_h(\tau) y_j$, then the $\mathsf{QIR}_\tau(h) = \beta_h(\tau)$ does not depend on y_j. I discuss whether linearity of the SQF can be justified later. Importantly, I contrast this definition with the structural conditional quantile impulse response defined as:

$$\mathsf{cQIR}_\tau(h) = \frac{\partial q_h(\tau \mid y_j, x)}{\partial y_j}. \tag{2}$$

The QIR and the cQIR may differ even if the treatment and control variables are independent ($Y_{j,t} \perp\!\!\!\perp X_t'$). Furthermore, the same observation $Y_{i,t+h}$ might fall below $q_h(\tau \mid y_j, x)$ but above $q_h(\tau \mid y_j)$ or vice versa.

The structural mean impulse response can be defined as:

$$\mathsf{IR}(h) = \frac{\partial \mathbb{E}[Y_{i,t+h}(y_j)]}{\partial y_j}, \tag{3}$$

and the structural conditional mean impulse response can be defined as:

$$\mathsf{cIR}(h) = \frac{\partial \mathbb{E}[Y_{i,t+h}(y_j, x)]}{\partial y_j}, \tag{4}$$

The expectation and quantile operators have different mathematical properties, which is why quantile and mean impulse responses behave differently under conditioning. To illustrate this, consider a simple data generating process: $Y = X_1 \cdot W + X_2$, where X_1, X_2 and W are all independently and identically distributed uniform random variables with W representing an unobserved shock. The linearity of the expectations operator means that $\mathbb{E}[Y \mid x_1] = x_1 \cdot \mathbb{E}[W] + \mathbb{E}[X_2]$ and $\mathbb{E}[Y \mid x_1, x_2] = x_1 \cdot \mathbb{E}[W] + x_2$. After taking the partial derivative with respect to x_1, both expressions yield $\mathbb{E}[W]$ making the IR and cIR identical. This result generalizes to any model where the treatment variable and other covariates enter additively[2], the linearity of expectations ensures that the additive component drops out when computing partial derivatives, regardless of whether it is conditioned on or not. In contrast, $q_{A+B}(\tau) \neq q_A(\tau) + q_B(\tau)$ for random variables A and B unless A and B are comonotonic (Koenker 2005). After conditioning on both X_1 and X_2 they become constants, so $q_Y(\tau \mid x_1, x_2) = q_{x_1 W + x_2}(\tau) = x_1 \cdot q_W(\tau) + x_2$ by the affine property of quantiles. Taking the partial derivative with respect to x_1 then yields $\frac{\partial q_Y(\tau|x_1,x_2)}{\partial x_1} = q_W(\tau)$. However, after conditioning on $X_1 = x_1$ only, the conditional quantile $q_Y(\tau \mid x_1)$ cannot in be written in separable form. Thus, the derivative $\frac{\partial q_Y(\tau|x_1)}{\partial x_1} = \frac{\partial q_{x_1 \cdot W + X_2}(\tau)}{\partial x_1}$ depends on the distribution of the random variable $x_1 \cdot W + X_2$ and differs from the conditional case. In effect, the coefficients on X_1 from quantile regressions on X_1 only versus on X_1 and X_2 will differ, even though X_1 and X_2 are independent. Adding covariates that are uncorrelated with the treatment to conditional mean models never changes the coefficient on the treatment variable as per the Frisch-Waugh-Lovell theorem, but this theorem does not apply to quantile regression.

Comparing equations 1 and 3, the QIR and mean impulse response share the same structure but target different aspects of the outcome distribution. While the mean impulse response captures how treatment affects the expected value, the QIR describes how treatment affects specific quantiles. Both the QIR and IR should be interpreted as responses to shocks that cause a one-off time-t unit change in the treatment variable, consistent with the local

[2]In models where the covariates do not enter linearly, Generalized IR functions that are a function of the covariates are needed and a simple linear model is misspecified.

projections literature. This interpretation differs slightly from SVAR impulse responses, which measure responses to unit variance innovations, so comparisons with SVAR impulse responses require scaling by an appropriate constant of proportionality (Plagborg-Møller and Wolf 2021).

A word of caution is in order when dealing with cumulative quantile impulse responses. To calculate cumulative impact on growth in the level of the variable of interest (e.g. Industrial Production IP_t) using local projections, the outcome variable is usually transformed to $Y_{i,t+h} = \log(IP_{t+h}) - \log(IP_{t-1})$. This is also the transformation used in this paper. This transformation is innocuous in the case of the mean impulse response as linearity of the expectations operator ensures that the cumulative effect equals the sum of period-by-period effects. However, quantiles of sums generally do not equal sums of quantiles. For example, the effect on the median annual growth rate will not generally equal to the sum of the effects on the 12 consecutive median monthly growth rates. Therefore, when $Y_{i,t+h}$ represents cumulative growth, the quantile impulse response describes how treatment affects the τ quantile of the h-periods-ahead cumulative growth distribution, not the sum of consecutive period effects.

To motivate QIR analysis it is instructive to discuss in what class of models the QIR is not equal to the mean impulse response. In particular, I state a quantile invariance result that proves that in a rich class of models quantile impulse responses are the same for each conditional and unconditional quantile and equal to the mean impulse response. This result is useful for two reasons. First, it establishes a lower bound on model complexity necessary for a model to exhibit non-trivial quantile dynamics. This has implication for theorists interested in writing models which feature interesting quantile impulse responses consistent with empirical evidence. Second, it cautions against overly restrictive identifying assumptions for quantile impulse response identification, as that could lead to a paradox in which identification relies on assumptions that imply quantile invariance.

Suppose the process $Y_t \in \mathbb{R}^k$ admits a structural Wold representation of the form

$$Y_t = \sum_{j=0}^{\infty} \Psi_j W_{t-j},$$

where $W_t \in \mathbb{R}^k$ are orthogonal innovations satisfying $E[W_t] = 0$, $E[W_t W_t'] = \mathbf{1}$, and $E[W_t W_s'] = 0$ for all $t \neq s$. The marginal treatment effect of a shock in variable j at time t on variable i at horizon h is then

$$\frac{\partial E[Y_{i,t+h}(w_j)]}{\partial w_j} = [\Psi_h]_{i,j}.$$

Additionally assuming Gaussian innovations, i.e. $W_t \sim \mathcal{N}(0, \mathbf{1})$, implies that the components of W_t are independent over time and across variables.

Theorem 2 (Quantile Invariance Theorem)**.** *If $\{Y_t\}$ has a structural Wold representation with Gaussian innovations, then for all quantiles $\tau \in (0,1)$ and all horizons h,*

$$\mathsf{cIR}(h) = \mathsf{IR}(h) = \mathsf{QIR}_\tau(h) = \mathsf{cQIR}_\tau(h)$$

As any purely nondeterministic, zero-mean covariance stationary process has a Wold representation, and since invertibility means it is possible to orthogonalize innovations, the most restrictive assumption in Theorem 2 is the Gaussianity of innovations. Critically, if the underlying data-generating process is non-linear, it may not admit a Wold representation with Gaussian innovations even if the structural shocks driving the process are Gaussian. For instance, processes with stochastic volatility (like the example in section 3.1) have Wold representations but with non-Gaussian innovations. As such, Theorem 2 tells us that breaking quantile invariance requires either departures from Gaussian innovations, linearity or covariance stationarity. However, these departures are not sufficient conditions for quantile non-invariance. For instance, non-Gaussian but i.i.d. symmetric innovations in a linear

process may still exhibit identical responses across quantiles. The proof of Theorem 2 follows from the linearity of the Wold representation and the independence of Gaussian innovations. The full proof is in Appendix A1.2.

2.3 Quantile Impulse Response identification

If the observed treatment $Y_{j,t}$ is randomly assigned i.e. $U_{i,t+h}(y_j) \mid Y_{j,t} \sim U_{i,t+h}(y_j) \sim \mathsf{Uniform}[0,1]$, then a quantile local projection model $Y_{i,t+h} = q_h(U_{i,t+h} \mid Y_{j,t})$ estimated using a standard quantile regression restriction $P(Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid Y_{j,t}) = \tau$ identifies the QIR as defined in equation 1. In non-experimental settings typical in macroeconomics, an endogeneity problem arises because the realized treatment $Y_{j,t}$ is not randomly assigned. I address the endogeneity problem with an identification by controls strategy. In particular, I relax the assumption that $U_{i,t+h}(y_j) \mid Y_{j,t} \sim U_{i,t+h}(y_j)$ and replace it with $U_{i,t+h}(y_j) \mid Y_{j,t}, X_t \sim U_{i,t+h}(y_j) \mid X_t$.[3] In other words, I assume that the treatment is conditionally on (observable) controls randomly assigned. Consistent with assumption 5, I think of the observed treatment as a function of the observable controls and an unobserved structural shock $W_{j,t}$, i.e. $Y_{j,t} = g_j(X_t, W_{j,t})$. As such the object of causal analysis is the quantile impulse response to a structural shock to the treatment variable.

The Frisch-Wough-Lovell theorem does not apply to quantile regression making disentangling effect of controls on identification versus interpretation more difficult. In particular, the quantile local projections model with controls $Y_{i,t+h} = q_h(U^*_{i,t+h} \mid Y_{j,t}, X_t)$ estimated using a restriction $P(Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}, X_t) \mid Y_{j,t}, X_t) = \tau$ deals with the endogeneity issue, but estimates a different structural function $q_h(\tau \mid y_j, x)$ instead of $q_h(\tau \mid y_j)$. As such it estimates the cQIR defined in equation 2 instead of the QIR defined in equation 1. The addition of controls into the equation changes the interpretation of the model. As such, even in cases when the treatment is randomly assigned, inclusion of control variables could

[3]Note that this allows for the rank variable to have different distributions for different values of the controls X_t. I.e. the controls can help predict whether the outcome will be below/above its conditional (on treatment) quantile.

change the quantile regression coefficients on the treatment variable. Note also that the conditional on controls rank variable $U^*_{i,t+h}(y_j, x)$ is distinct from $U_{i,t+h}(y_j)$. In particular, $U_{i,t+h}(y_j) = \lambda(X_t, U^*_{i,t+h}(y_j, x))$ for some function λ that depends on the fixed treatment and the horizon, but not time.

Exploiting control variables for causal identification while still modeling the conditional on treatment only SQF $q_h(\tau \mid y_j)$ is possible thanks to the Powell 2020 generalized quantile regression estimator, which explicitly distinguishes between treatment and control variables. One of the contributions of this paper is to adapt this cross-sectional framework to the time-series setting. In what follows I only consider identification by controls, Powell 2020 also considers identification using instrumental variables making extension of GQLP to instrumental variable designs straightforward. To identify quantile impulse responses one more assumption is needed:

Assumption 7 (Rank Similarity)**.** *For all* y_j, y'_j*:* $\mathbb{P}[Y_{i,t+h}(y_j) \leq q_h(\tau \mid y_j) \mid Y_{j,t}, X_t] = \mathbb{P}[Y_{i,t+h}(y'_j) \leq q_h(\tau \mid y'_j) \mid Y_{j,t}, X_t]$.

The rank similarity assumption 7 posits that, conditional on current observables the rank of the potential outcome within its distribution does not systematically vary with different realizations of the treatment variable. In other words, if we know the current treatment and controls, whether a time period would have a high-rank or low-rank outcome does not depend on which treatment value we are considering. For example, if economic conditions suggest a period would experience an above-median outcome given one treatment value, those same conditions suggest it would also experience an above-median outcome for a different treatment value (though the median levels themselves would differ). It is a key assumption for identification, along with assumptions 3 and 5, which taken together establish conditional (on controls) independence of the treatment assignment.

Before stating the moment conditions used to recover the QIR, I reformulate the Theorem 1 from Powell 2020 except in the time series setting. The proof of the theorem is in the appendix section A1.3.

Theorem 3. *Suppose Assumptions 1 through 7 hold* $\forall h$*, then* $\forall h \in \{0, 1, 2, \ldots, H\}$ *and for each* $\tau \in (0, 1)$*:*

$\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid Y_{j,t}, X_t] = \mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid X_t]$,

$\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t})] = \tau$.

The first equation in theorem 3, states that after conditioning on controls X_t, the treatment $Y_{j,t}$ does not provide additional information about the probability that the outcome is below its quantile function. The second equation in theorem 3, ensures that the quantile function is correctly scaled. Together, these equations imply that the conditional probability $\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid X_t]$ is allowed to vary based on controls X_t, but in expectation it is equal to the quantile level τ. When there are no control variables in the model (i.e. $X_t = 0$), the two conditions in theorem 3 collapse into one standard quantile regression restriction $\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid Y_{j,t}] = \tau$. This restriction is used to estimate QIRs in the quantile local projections framework. As such, quantile local projections are a special case generalized quantile local projections, corresponding to a setting where all the model variables are treatment variables and there are no controls. Therefore, GQLP "nests" the QLP framework.

Theorem 3 provides the moment conditions needed for the estimation of the generalized quantile local projections. In particular, it implies two moment conditions for each horizon $h \in \{0, 1, 2, \ldots, H\}$ and quantile τ of interest:

$$\mathbb{E}\big[Y_{j,t}[\mathbf{1}\{Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t})\} - \mathbb{P}(Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid X_t)]\big] = 0,$$
$$\mathbb{E}\big[\mathbf{1}\{Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t})\} - \tau\big] = 0,$$

where $\mathbf{1}\{\}$ is the indicator function that equals 1 if the condition in braces is true and 0 otherwise. Estimation is done separately for each horizon h and each τ as in the quantile local projections. For a given h and τ and assuming a linear specification $q_h(\tau \mid y_j) = \alpha_h(\tau) + \beta_h(\tau)d$, estimation proceeds in three steps:

1. Postulate a candidate $\tilde{\beta}_h(\tau)$. For each candidate $\tilde{\beta}_h(\tau)$ there exists an intercept $\tilde{\alpha}_h(\tau)$ such that $\mathbb{P}(Y_{i,t+h} \leq \tilde{\alpha}_h(\tau) + \tilde{\beta}_h(\tau)Y_{j,t}) = \tau$. This means that we need to search over the slope coefficients only.

2. Given the pair $(\tilde{\alpha}_h(\tau), \tilde{\beta}_h(\tau))$, estimate a linear probability model (Logit or Probit could also be used) for the event that $Y_{i,t+h} \leq \tilde{\alpha}_h(\tau) + \tilde{\beta}_h(\tau)Y_{j,t}$ as a function of controls X_t. Save the predicted probabilities as $\hat{\tau}_{X_t}$.

3. $\hat{\beta}_h(\tau) = \text{argmin}_{\tilde{\beta}_h(\tau)}\, g'Ag$, where $g = \frac{1}{T}\sum_{t=1}^{T} Y_{j,t}[\mathbf{1}\{Y_{i,t+h} \leq \tilde{\alpha}_h(\tau) + Y_{j,t}\tilde{\beta}_h(\tau)\} - \hat{\tau}_{X_t}]$. $A = [\hat{E}(gg')]^{-1}$ is the optimal GMM weighting matrix constructed using starting values from standard quantile regression of $Y_{i,t+h}$ on $Y_{j,t}$.

Note that misspecification in the binary outcome model of step 2 does not pose issues for identification, as long as the misspecifcation errors are orthogonal to the treatment variable. For more details about the estimation algorithm I refer the reader to Powell 2020.

I calculate confidence intervals using moving block bootstrap, the description of the algorithm is in the appendix section A2. This procedure preserves the time-dependency by resampling blocks of M consecutive observations instead of resampling individual time points (Kilian and Lütkepohl 2017). After re-estimating the model B times using these pseudo-samples, the confidence intervals are based on the distribution of the estimated parameters across the B replications of the procedure. I test the coverage of the confidence intervals obtained using this method in the Monte Carlo study in section 3.1.

3 Monte Carlo

3.1 Endogenous Volatility SVAR

Consider a SVAR(1) augmented by an endogenous volatility term:

$$Y_{i,t} = a_{1,11}Y_{i,t-1} + a_{1,12}Y_{j,t-1} + \frac{1+\phi\sqrt{\exp(Y_{j,t-1})}}{1+\phi}W_{i,t}$$

$$Y_{j,t} + a_{0,22}Y_{i,t} = a_{1,21}Y_{i,t-1} + a_{1,22}Y_{j,t-1} + W_{j,t}$$

Where, $W_{i,t}, W_{j,t} \overset{iid}{\sim} N(0,1)$ are unobserved independent structural shocks. If parameter $\phi = 0$ the model collapses to a standard SVAR with $Y_{i,t}$ ordered first ($Y_{i,t}$ predetermined with respect to $Y_{j,t}$). When $\phi > 0$ the stochastic endogenous volatility term $\sqrt{\exp(Y_{j,t-1})}$ creates a relationship between $Y_{j,t-1}$ and the volatility of $Y_{i,t}$. This generates volatility dynamics that give rise to a skewed ergodic distribution of $Y_{i,t}$ and QIRs that vary across quantiles. The mean impulse responses in this model do not depend on the value of the volatility parameter ϕ, they are the same as in the linear SVAR (case when $\phi = 0$).

parameter	$a_{1,11}$	$a_{1,22}$	$a_{1,12}$	$a_{1,21}$	$a_{0,22}$	ϕ
value	0.5	-0.1	-0.25	-0.1	-0.2	4

Table 1: Model parameters used in the simulation.

Although this is not an economic model, to keep the discussion less abstract, think of $Y_{i,t}$ as output growth and $Y_{j,t}$ as the change in financial conditions (with positive values meaning tightening financial conditions). Thus, if $\phi > 0$ tightening financial conditions lead to an increase in the volatility of output growth. This endogenous volatility together with the negative relationship between the two variables generates an output growth distribution that is left skewed, consistent with empirical evidence.

To study the cumulative impulse responses of the level of output I define a transformed dependent variable $Y^c_{i,t+h} \equiv \sum_{j=0}^{h} Y_{i,t+j}$. The structural mean impulse responses can be identified using local projections with appropriate timing restrictions (Jordá 2005; Plagborg-

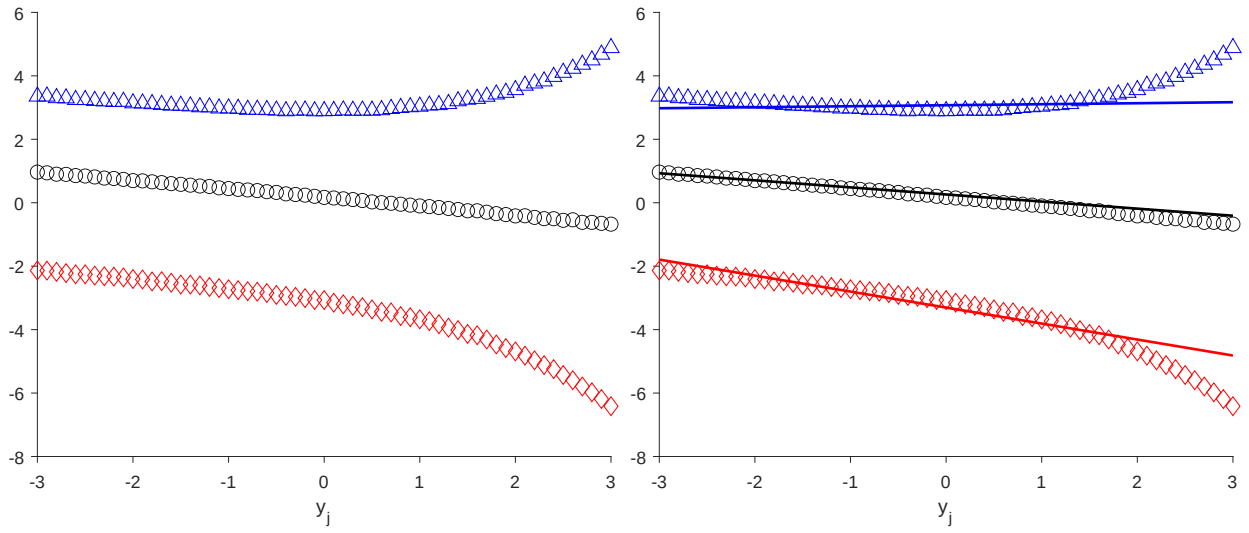
Møller and Wolf 2021). Estimating by least squares separately for each $h \in \{1, 2, \ldots, H\}$:

$$Y^c_{i,t+h} = \alpha_h + Y_{j,t}\beta_h + X_t^\top \theta_h + \varepsilon_{i,t+h}$$

with $X_t = \{Y_{i,t}, Y_{j,t-1}, Y_{i,t-1}\}$, β_h recovers the structural mean impulse response. I know which variables need to be included in X_t from looking at the equation for $Y_{j,t}$ in the data generating process and using the fact that $Y_{j,t} = g_j(X_t, W_{j,t})$. The inclusion of the controls vector X_t is necessary as $Y_{j,t}$ is endogenous. Failing to include the correct variables in X_t would result in biased estimates of the impulse response. If the structural shock $W_{j,t}$ were directly observable, replacing $Y_{j,t}$ with $W_{j,t}$ as the treatment variable would identify the structural impulse response without the need for controls X_t (although their inclusion may still be desirable to improve the precision of the estimates).

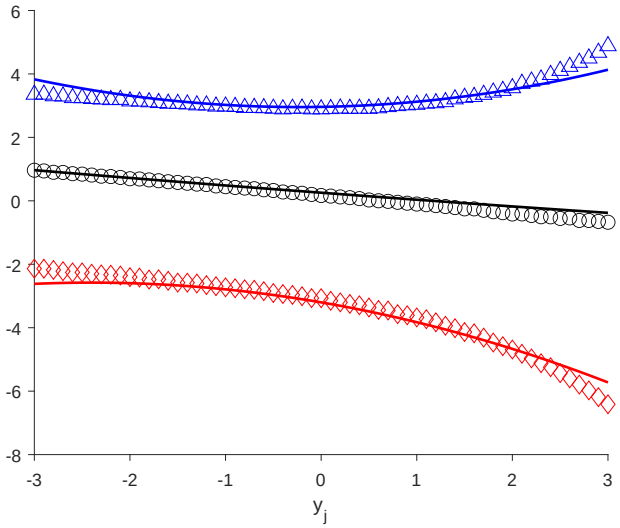
When interest lies in identifying the QIR as defined in equation 1, employing the Koenker and Bassett 1978 estimator in a local projections setting might not be enough. Firstly, a linear quantile regression may be misspecified if the functional form of the SQF $q_h(\tau \mid y_j)$ is not linear. In short time-series typical in macroeconomics, nonparametric estimation of the SQF may be unfeasible, especially for more extremes quantiles. For a given model for the underlying data generating process we can try to characterize the implied functional form of $q_h(\tau \mid y_j)$. Depending on the model, a closed-form solution for the SQF may be hard to find from the model's equations. For example, a linear SVAR model (case when $\phi = 0$) has linear SQFs for endogenous variables to structural shocks. Furthermore, SVAR quantile impulse responses equal to the mean impulse response for all quantiles. On the other hand, the stochastic volatility SVAR ($\phi > 0$) which features non-trivial QIRs – ones that vary with depending on the quantile τ – also features nonlinear SQFs for some quantiles and horizons. Even if the model implied SQF might be hard to characterize in closed-form, the shape of the SQF can be recovered from simulating the model.

The simulation algorithm follows from the definition of potential outcomes and the SQF. Potential outcomes can be generated based on either assignment counterfactuals $Y_{i,t+h}(w_j)$ or



True SQF

Fitted linear SQF using QR



Fitted quadratic SQF using QR

Figure 1: Simulation results for the first horizon SQF $q_1(\tau \mid y_j)$. The top-left panel plots the simulated quantiles of potential outcomes $Y^c_{i,t+1}(y_j)$ over a grid of values y_j for quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$ (obtained from $MC = 100{,}000$ simulation repetitions). The other two panels re-plot these simulated quantiles, with the overlayed solid lines showing the fitted SQF using a quantile regression of $Y^c_{i,t+1}$ on the structural shock $W_{j,t}$ for the same three quantiles. The fit in the top-right panel comes from a linear quantile regression while the bottom panel fit comes from a quadratic quantile regression. The regression coefficients used to plot the fitted SQFs are averaged estimates from a Monte Carlo simulation with $MC = 100$ replications and time-series of length $T = 500$ (after dropping 1,000 initial observations).

counterfactuals relating to a realization of another endogenous variable $Y_{i,t+h}(y_j)$. The first method involves letting the model run for a number of periods before fixing the structural

shock in one period to $W_{j,t} = w_j$ for a grid of values, then simulating forward for each w_j grid value and computing the empirical quantiles of $Y_{i,t+h}(w_j)$. This recovers the quantile of the potential outcomes $Y_{i,t+h}(w_j)$ as a function of the fixed assignment w_j, i.e. the SQF. The second method is the same except it requires forcing the endogenous variable to take a value $Y_{j,t} = y_j$, which — to remain consistent with the structural equations — requires choosing the shock value that solves the structural mapping $y_j = g_j(w_j, X_t)$ by setting $w_j = g_j^{-1}(y_j, X_t)$.

Intervening on $Y_{j,t}$ and intervening on $W_{j,t}$ are conceptually equivalent, in that an intervention on one can be expressed as an intervention on the other through the inverse mapping g_j^{-1}. When the function g_j is linear (as in the current example) the mapping becomes particularly simple. In particular, if $g_j(W_{j,t}, X_t) = \theta(X_t) + \kappa W_{j,t}$ with $\kappa \neq 0$. Solving for the shock gives $W_{j,t} = g_j^{-1}(Y_{j,t}, X_t) = \kappa^{-1}[Y_{j,t} - \theta(X_t)]$, so $Y_{j,t} = y_j$ is equivalent to setting $W_{j,t} = w_j = \kappa^{-1}[y_j - \theta(X_t)]$.. Equivalently, compared to an unperturbed draw $\{X_t, W_{j,t}\}$ that produced $Y_{j,t} = \theta(X_t) + \kappa W_{j,t}$, the required additive perturbation in the shock is $\delta = \kappa^{-1}[y_j - Y_{j,t}]$. In the current example that features an additive unit-slope ($\kappa = 1$) this further reduces to $\delta = y_j - Y_{j,t}$, so fixing $Y_{j,t} = y_j$ is exactly the same as perturbing $W_{j,t}$ by adding $\delta = y_j - Y_{j,t}$. For nonlinear but invertible g_j the same conceptual equivalence holds (interventions map into one another via g_j^{-1}), but the mapping need not be an additive shift and must generally be computed pointwise for each X_t.

The top-left panel of Figure 1 shows the first horizon SQFs for three quantiles $\tau \in \{0.1, 0.5, 0.9\}$, recovered from simulations based on experimentally fixing $Y_{j,t} = y_j$ for a grid of values for y_j. Visual examination suggests that the SQF is quadratic for quantiles $\tau \in \{0.1, 0.9\}$ and linear for the median $\tau = 0.5$. In a simulation setting, the structural shock $W_{j,t}$ is observable and statistically independent by construction. This suggest another strategy to recover the true SQF by estimating a quantile local projection model $Y_{i,t+h}^c = q_h(U_{i,t+h} \mid W_{j,t})$. Again, this requires either the knowledge of the functional form of the SQF, or the use of some nonparametric method to approximate it. Alternatively, we could (incorrectly) assume a linear specification, which although misspecified may nevertheless

be a good approximation to the truth[4]. Figure 1 shows that a linear quantile regression $Y^c_{i,t+1} = \alpha(U_{t+1}) + \beta(U_{t+1})W_{j,t}$ does well at approximating the true SQF around $W_{j,t} = 0$, but is outperformed by a quadratic specification. If the nonlinearity of the quantile function is of primary concern, higher order polynomial approximations could be used. Visual inspection of the SQF for the first horizon response in Figure 1 suggests a quadratic specification is sufficient. Fortunately, local projections are flexible making it is easy to add higher order terms into the estimation equation.

Outside of simulation settings, the structural shocks $W_{j,t}$ are usually unobserved, so researchers need to rely on the time-series of the endogenous model variables $\{Y_t, X_t\}$ to estimate the SQF. Figure 2 compares the performance of three quadratic models for the estimation of the true SQF at horizon $h = 1$. The first is a model without controls estimated using quantile regression given by:

$$Y^c_{i,t+1} = \alpha_{h=1}(U_{t+1}) + \beta_{1,h=1}(U_{t+1})Y_{j,t} + \beta_{2,h=1}(U_{t+1})Y^2_{j,t}.$$

The second model adds controls X_t into the estimation equation and uses the quantile regression to estimate the parameters.

$$Y^c_{i,t+1} = \alpha_{h=1}(U_{t+1}) + \beta_{1,h=1}(U_{t+1})Y_{j,t} + \beta_{2,h=1}(U_{t+1})Y^2_{j,t} + X_t^\top \theta_{h=1}(U_{t+1}).$$

The third model also estimates the quadratic equation, but uses the controls X_t for identification, while modeling the quadratic SQF that is not conditional on controls. This is possible as my framework uses the generalized quantile regression of Powell 2020 instead of the quantile regression of Koenker and Bassett 1978.

Comparing the performance of these three models in recovering the SQF shows that the standard quantile regression is unable to recover the true shape of the SQF. The quantile regression model without controls suffers from endogeneity bias, while the quantile regression

[4]Angrist et al. 2006 study the properties of Quantile Regression under misspecification and show that it minimizes a weighted mean-squared error loss function for specification error.

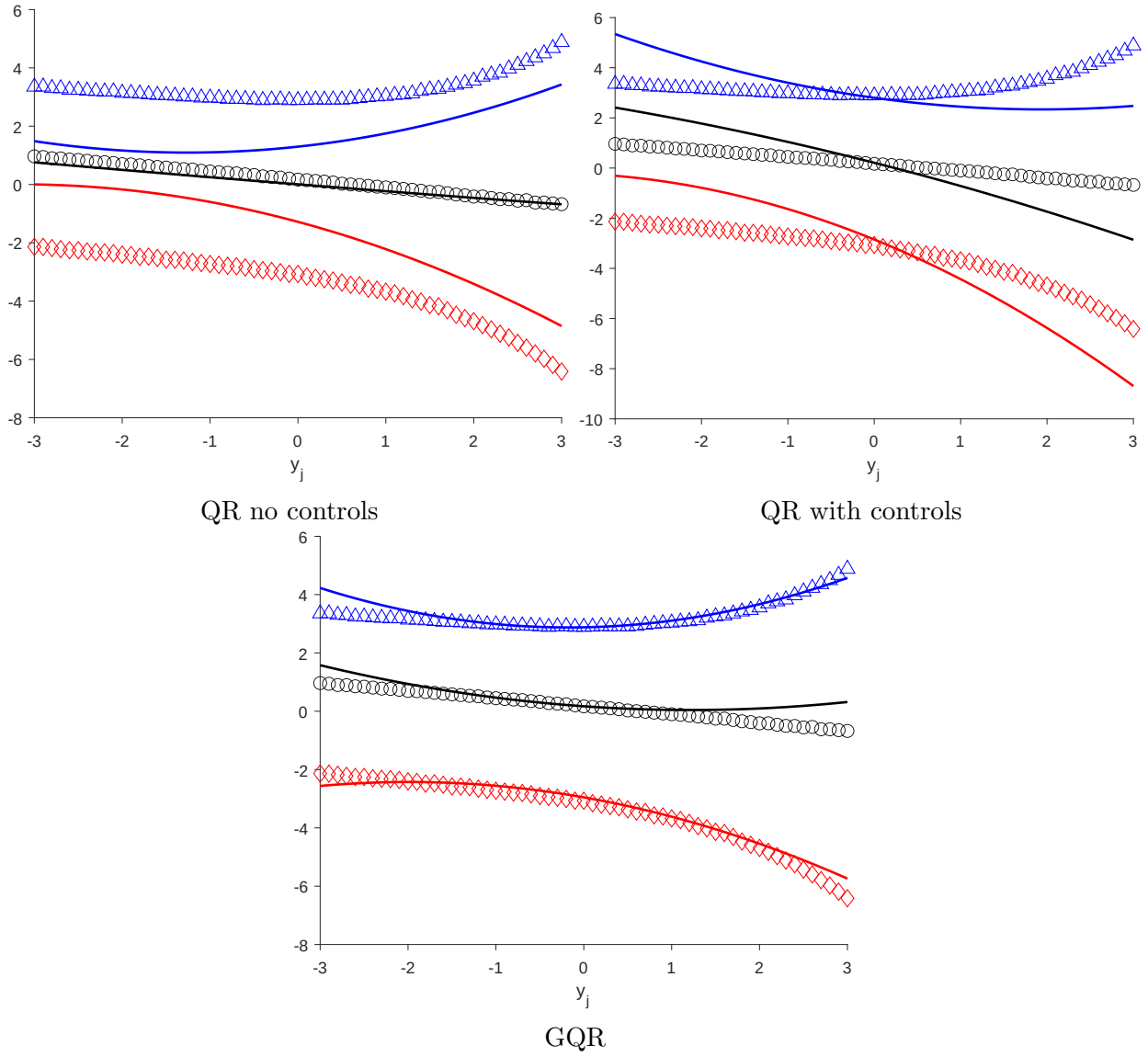
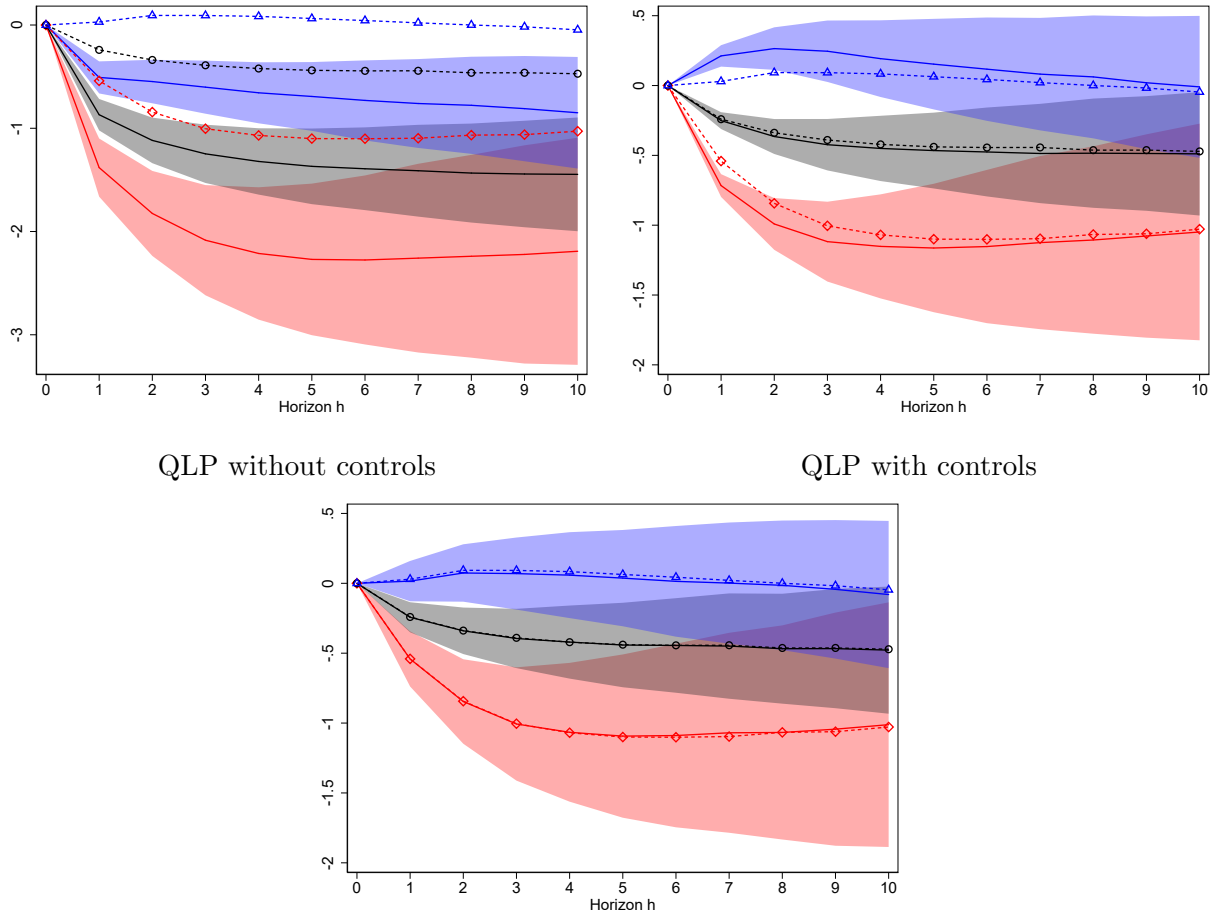


Figure 2: Simulation results for the first horizon SQF $q_1(\tau \mid y_j)$. The diamonds, circles and triangles are the same across the three panels and show the simulated quantiles of potential outcomes $Y^c_{i,t+1}(y_j)$ over a grid of values y_j for quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$(obtained from $MC = 100{,}000$ simulation repetitions). The three panels compare the performance of three estimators for the first horizon SQF. QR refers to the Koenker and Bassett 1978 estimator, GQR is the generalized quantile regression estimator introduced by Powell 2020. The regression coefficients used to plot the fitted SQFs are averaged estimates from a $MC = 1{,}000$ simulation replications and time-series of length $T = 500$ (after dropping 1,000 initial observations).

model with controls estimates a conditional SQF. On the other hand, the generalized quantile regression estimator targets the correct (unconditional on controls) SQF, while simultaneously

being able to address the endogeneity of the treatment in a controls-based identification strategy.



QLP without controls

QLP with controls

GQLP

Figure 3: Simulation results for the cumulative quantile impulse response. The diamonds ◇, circles ∘ and triangles △ are the same across the three panels and show the (linear approximation to the) true quantile impulse response as estimated by $Y^c_{i,t+h} = \alpha_h(U_{i,t+h}) + \beta_h(U_{i,t+h})W_{j,t}$, for quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. Solid lines show the results from the three estimators considered. QLP refers to the quantile local projection framework which uses the Koenker and Bassett 1978 estimator. GQLP is my local projections based framework which builds on the generalized quantile regression estimator introduced by Powell 2020. Results are averaged over $MC = 1{,}000$ simulation replications, with a time-series of length $T = 500$ (after dropping 1,000 initial observations). Y-axis plots $\hat{\beta}_h(\tau)$ and x-axis shows the horizon h. Shaded areas show the Monte Carlo standard error of the estimator equal to the estimate ± one standard deviation across the $MC = 1{,}000$ Monte Carlo iterations.

Since nonlinear SQFs imply that the QIRs will vary not only with the quantile but also with value of the treatment variable, they make plotting and analyzing the QIRs more

complicated. Thus for the sake of simplicity, a linear model may be deemed preferable even if it is misspecified. Ignoring the nonlinearity of the SQF for the moment, an approximation to the QIR can be recovered as the β_h from the quantile local projection:

$$Y^c_{i,t+h} = \alpha_h(U_{i,t+h}) + \beta_h(U_{i,t+h})W_{j,t}.$$

This simple strategy is possible in a simulation setting where the true structural shock $W_{j,t}$ is observable and by construction independent ($U_{i,t+h} \mid W_{j,t} \sim U_{i,t+h}$), making controls redundant for causal identification. A reasonable goal for an estimator of a structural QIR would be to recover the same QIR using only the time-series of the observed endogenous model variables $\{Y_{i,t}, Y_{j,t}\}$, similarly to how local projections identify the structural mean impulse response when the correct set of controls is included. Figure 3 shows that quantile local projections fail at achieving this goal.[5] In particular, a quantile local projection model without controls:

$$Y^c_{i,t+h} = \alpha_h(U_{i,t+h}) + \beta_h(U_{i,t+h})Y_{j,t},$$

suffers from endogeneity of $Y_{j,t}$ and as expected it fails to recover the structural QIR. Perhaps more surprisingly, a quantile local projection with the correct controls $X_t = \{Y_{i,t}, Y_{j,t-1}, Y_{i,t-1}\}$, given by:

$$Y^c_{i,t+h} = \alpha_h(U_{i,t+h}) + \beta_h(U_{i,t+h})Y_{j,t} + X_t^\top \theta_h(U_{i,t+h}),$$

solves the endogeneity of $Y_{j,t}$ problem, but in doing so models a conditional on controls SQF which has a different meaning than the conditional on treatment only SQF. In effect, it recovers the cQIR rather than the QIR, which in this case are not equal.

The GQLP estimator models the unconditional SQF, while still addressing the endogeneity of $Y_{j,t}$. As such, GQLP with the dependent variable $Y^c_{i,t+h}$, treatment variable $Y_{j,t}$ and controls $X_t = \{Y_{i,t}, Y_{j,t-1}, Y_{i,t-1}\}$ recovers the same QIRs as the (unfeasible in practice) QLP of $Y^c_{i,t+h}$

[5] In the appendix section A3, I provide a table that compares the mean bias and root mean squared error of the three estimators up to horizon 10, to complement Figure 3.

on the structural shock $W_{j,t}$. Shaded areas on Figure 3 plot the Monte Carlo standard errors of the estimators. It is clear from the plot that GQLP suffers from slightly higher estimation uncertainty than QLP at the short horizons, at which QLP is also biased. Meanwhile, at longer horizons where both estimators recover the unbiased effect, the standard errors are almost identical.

		Nominal Level		
Quantile	Horizon	68 %	90 %	95 %
0.1	1	71%	89%	94%
	5	68%	91%	96%
	10	70%	89%	94%
0.5	1	67%	92%	95%
	5	69%	90%	95%
	10	70%	91%	95%
0.9	1	68%	88%	93%
	5	70%	89%	96%
	10	72%	91%	96%
Average coverage		69.4%	90.0%	95.0%

Table 2: The table reports the coverage of moving block bootstrap confidence intervals for three quantiles, three horizons and three nominal confidence levels. The coverage was computed in a Monte Carlo simulation with $MC = 500$ repetitions, with sample size $T = 500$ and $B = 1000$ Bootstrap repetitions. The block size used in the bootstrap procedure was the same as in the empirical results section and equal to $m = 7$.

Table 2 reports the coverage of the moving block bootstrap confidence intervals (CIs) for three horizons $h \in \{1, 5, 10\}$ for the GQLP estimator. The algorithm used to compute the CIs, for which coverage is reported here, is the same as the one used in the empirical section and is described in the appendix section A2. Coverage was calculated by computing the CIs for the QIR at three selected horizons and three quantiles of interest $\tau \in \{0.1, 0.5, 0.9\}$ over 500 Monte Carlo simulation repetitions, and then recording the percentage of repetitions in which the true value of the estimator lay inside the CI. If the confidence intervals are correctly sized, the coverage should be close to the nominal level. The results in Table 2 suggest that the bootstrap confidence intervals are indeed correctly sized. The small deviations from the nominal levels are well within the Monte Carlo uncertainty (approximately ± 4 pp at 68%,

±2 pp at 90%, and ±2 pp at 95%). Moreover, the average coverage across quantiles and horizons is essentially equal to the nominal levels, further confirming that the intervals are appropriately calibrated.

3.2 Nonlinear DSGE model

"A Macroeconomic Model with Financial Panics" of Gertler et al. (2019) is a fully micro-founded nonlinear DSGE model that features bank panics and financial accelerator mechanisms. It is solved globally and so it can be used as a nonlinear data generating process by sampling random innovations. In the model households and bankers interact through capital accumulation, adjustment costs, and an agency problem. Households can invest and manage capital but incur convex adjustment and management costs; bankers finance their operations with their net worth and household deposits, but face diversion risk and potential bankruptcy if their net worth turns zero or negative. The only exogenous uncertainty in the model is a capital-quality shock, while endogenous bank runs are triggered by sunspot disturbances whenever liquidation prices fall sufficiently—reflecting households' relative inefficiency at managing capital—to push bank net worth below zero. Although the steady-state is free of bank runs, a sequence of adverse capital-quality shocks can erode net worth and open the door to panic equilibria. This is illustrated on Figure 4, which shows how the economy affected by a sequence of negative capital-quality shocks can move from the steady state to a bank-run equilibrium if a sunspot shock occurs.

Figure 4 was generated by Gertler et al. (2019) using a typical approach used in theoretical macroeconomic modeling. It is based on a single simulation repetition and reports the impulse responses to a so-called "MIT shock". The "MIT shock" approach starts the economy at steady state, then hits the economy with a one-off shock (or in this case sequence of three shocks) while setting all other current and future shocks to zero. Notably, this type of impulse response has no empirical counterpart. In empirical setting, researchers generally study the counterfactual changes in the outcome of interest, averaging out over all other shocks rather

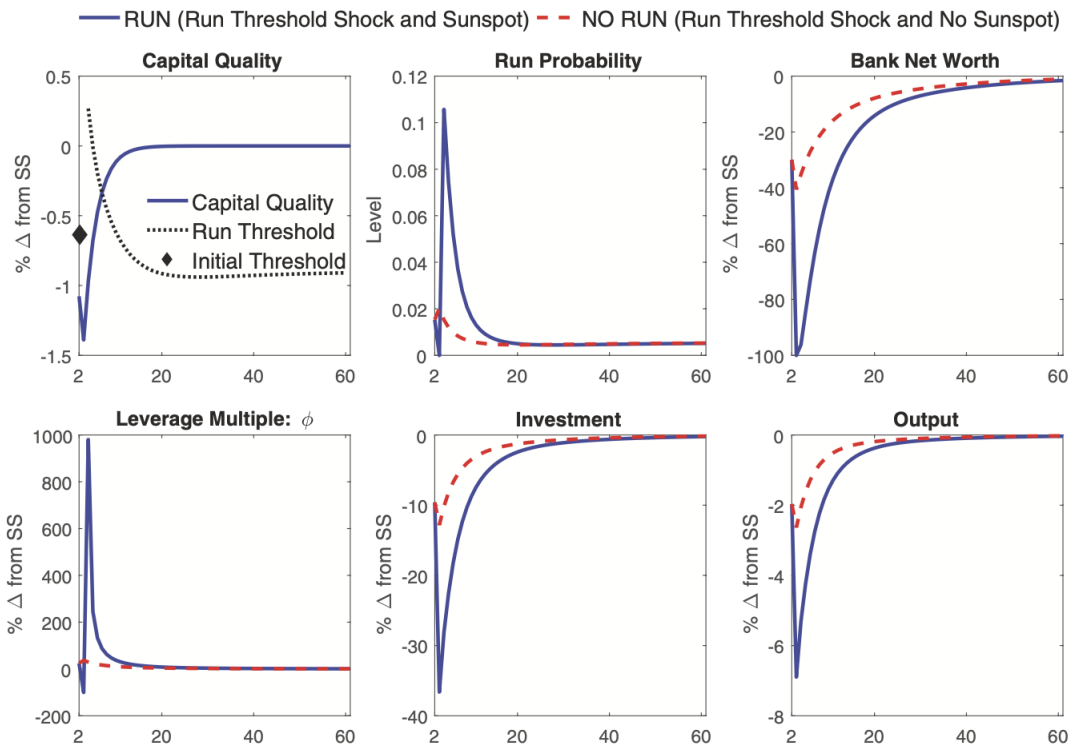


Figure 4: From Gertler et al. (2019). Response of the economy to a sequence of three small negative capital quality shocks combined with a sunspot that triggers a bank run. The plot starts in period 2, economy is in steady state in period 0 then it experiences the three shocks and no shocks thereafter.

than conditioning on them being zero (Kolesár and Plagborg-Møller 2025).

To recover the structural quantile function (SQF) in this nonlinear setting, I follow the same algorithm as for the SVAR model in the previous section. In each of the Monte Carlo runs, I draw random realizations of the capital-quality and sunspot shocks over the first $T+h$ periods, then I replace the realization of the capital-quality shock at date T with each of the grid values. This procedure generates potential outcomes of output for each fixed shock value from the grid. Doing this many times recovers the distribution of the potential outcomes. Then for a given quantile and horizon of interest, I simply compute the quantiles of the simulated time $T+h$ potential outcomes corresponding to each fixed grid point to recover the SQF. The results of applying this procedure for the first horizon $h=1$ and three quantiles $\tau \in \{0.1, 0.5, 0.9\}$ is presented in Figure 5. It is clear that the SQF for the extreme quantiles

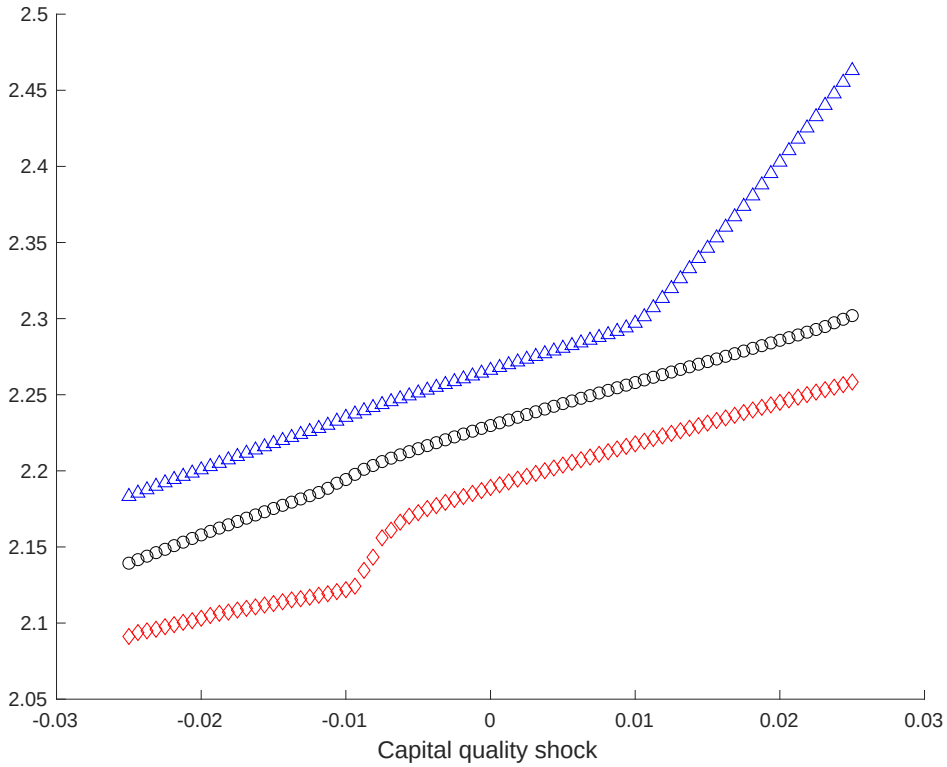


Figure 5: Structural quantile function of output to capital quality shocks at horizon 1, plotted for quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. X-axis shows the capital quality shocks (the variance of the capital quality shocks is $\sigma = 0.005$), output at horizon 1 is on the vertical axis. The results are by simulating the Gertler et al. (2019) model for each shock grid point over $MC = 1{,}000$ simulation repetitions.

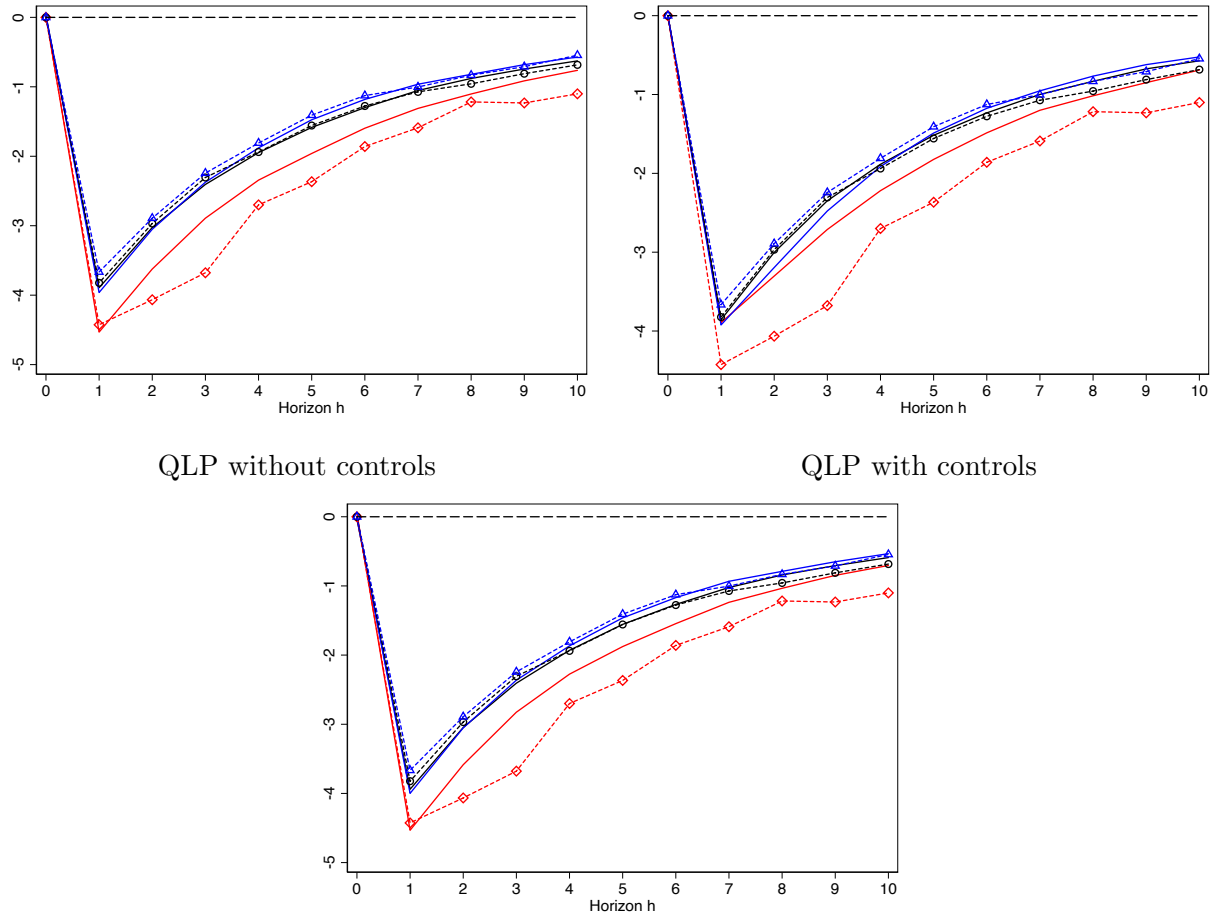
$\tau \in \{0.1, 0.9\}$ is nonlinear. Moreover, in the case of the $\tau = 0.1$ quantile which captures the downside risk to growth, the SQF does not look like it could be well approximated by a quadratic function. Using higher order polynomials or nonparametric methods for estimation of the SQF at extreme quantiles in small sample settings is unlikely to be yield satisfactory results, as the precision of the estimates is likely going to be too low to draw any substantive conclusions.

I compute the true QIR using a perturbation method. I sample random shocks for the first T periods, at T I create a counterfactual series where a small perturbation δ is added to the capital quality shock.[6] I then simulate both the original and the counterfactual series through $T + H$ by drawing more random shocks. Finally, I compute the difference in the

[6] I use $\delta = 0.2\sigma$ where σ is the standard deviation of the capital quality shock. Theoretically δ should be an infinitesimally small perturbation, but due to rounding errors excessively small values are impractical.

quantiles of $T+H$ output between the unperturbed and counterfactual series (and scale it by δ), i.e. $\frac{1}{\delta}[q_h(\tau \mid W_{j,t}) - q_h(\tau \mid W_{j,t} + \delta)]$. I assess the estimation performance of three models: QLP with and without controls and GQLP, comparing each against the true QIRs. In all specifications, I use the capital quality shock as an observed treatment variable. The controls are the model's state-variables, namely: bankers' net worth, capital quality and lagged capital stock. Since capital quality shocks are independent, the controls are actually not needed for causal identification. However, in applied work the tendency is to include controls even in settings where a possibly exogenous shock is observed. This generally poses little issue when focus lies in identification of average treatment effects, but can have dramatic consequences for identification of quantile treatment effects.

As shown on Figure 6, the GQLP estimator estimates the same QIRs as the QLP estimator without controls. These QIRs do not exactly coincide with the counterfactual QIR as the functional form of the model is misspecified. However, from the independence of the captial quality shocks and approximation properties of QR (Angrist et al. 2006), we know that QLP without controls identifies a linear approximation to the truth. On the other hand, QLP with controls estimates a completely different QIR for the $\tau = 0.1$ quantile. The results of the QLP model with controls could lead the researcher to conclude that there is no difference across the responses of quantiles of outcome to the capital quality shock. The difference arises because including controls means that the estimated QIRs are conditional on the state of the economy, rather than unconditional. In the context of the Gertler et al. (2019) model, this distinction is crucial. The unconditional lower quantiles capture episodes when the economy is fragile and bank runs can be triggered, so they exhibit strong nonlinearities in response to capital quality shocks. By contrast, once the state variables are conditioned on, the quantiles reflect outcomes within given states of vulnerability or resilience. In this conditional setting, the capital quality shock shifts the distribution of output more homogeneously, and the extreme quantiles no longer display the same amplified downside responses.



QLP without controls

QLP with controls

GQLP

Figure 6: True QIR (dashed) versus estimated QIR (solid) under the linear specification for quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. Results averaged over $MC = 1{,}000$ simulations of length $T = 500$ (after dropping 500 initial observations); horizons $h = 1, \ldots, 10$.

4 Empirical Results

I apply the GQLP methodology to reexamine how financial risk shocks affect the distribution of industrial production growth in the United States. The empirical exercise follows an established literature that documents how adverse financial shocks disproportionately increase downside risks to growth (Adrian et al. 2019; Chavleishvili and Manganelli 2024; Loria et al. 2025). My choice of financial risk variables – capturing credit and volatility risk – is motivated by both theoretical considerations and empirical precedents. Theoretically, tightening financial conditions can amplify economic downturns through multiple channels including credit market

frictions that constrain firm investment when external finance premiums rise (Gilchrist and Zakrajšek 2012), and uncertainty-driven real options effects that delay irreversible investments (Bloom 2009). Previous studies have used similar financial indicators to identify these channels, with Gilchrist and Zakrajšek (2012) focusing on the excess bond premium and Bloom (2009) focusing on stock market volatility. By contrasting results from GQLP and QLP using these indicators, I demonstrate that distinguishing between conditional and unconditional quantiles is both statistically and economically important for understanding the causal drivers of growth-at-risk.

My monthly dataset covers the US economy during the period between January 1984 and June 2025 (T=498). All the data used is publicly available, with majority of it contained in the FRED-MD database published by the St. Louis Fed (McCracken and Ng 2015). I use monthly data for a larger sample size, with Industrial Production as the dependent variable. In particular, the dependent variable $Y_{i,t+h}$ is defined as the h-months cumulative log growth rate $Y_{i,t+h} = 100 * [\log(IP_{t+h}) - \log(IP_{t-1})]$. I multiply the log growth rates by 100 to interpret the QIR in terms of percentage points. I normalize the treatment variable $Y_{j,t}$ to interpret the QIRs as responses to a one standard deviation change.

The first treatment variable $Y_{j,t}$ I consider measures movements in credit risk. I will refer to this variable as credit risk and I define it as the first-difference of the monthly Excess Bond Premium (EBP) of Gilchrist and Zakrajšek 2012, i.e. $Y_{j,t} = EBP_t - EBP_{t-1}$. The EBP is the residual of corporate bond credit spreads that cannot be explained by movements in expected default risk, as such it measures the investor sentiment or risk appetite in the corporate bond market.

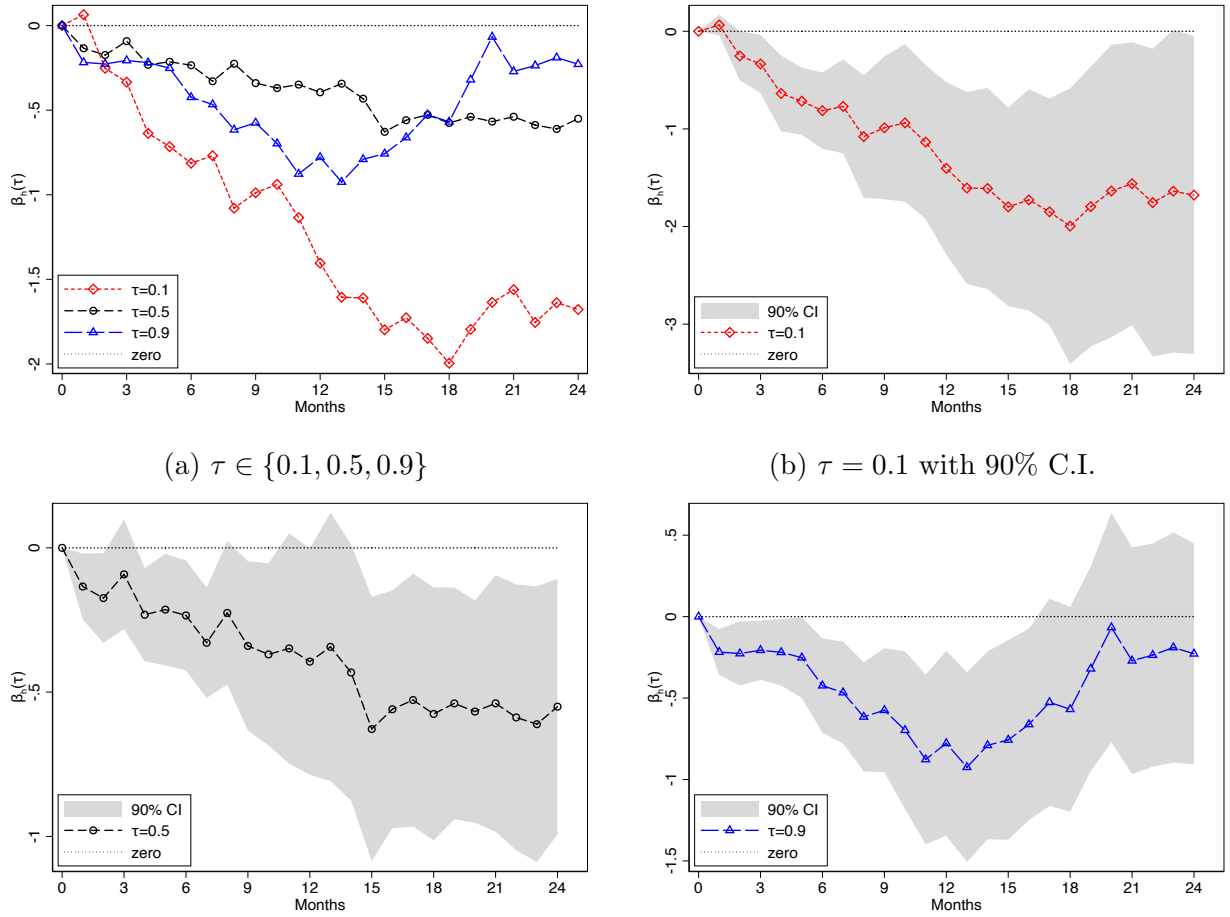
The second treatment variable $Y_{j,t}$ I consider measures the volatility risk premium in the equity markets, defined as the difference between realized and implied volatility of the S&P500 index. I will refer to it as volatility risk for short. I compute realized volatility by computing the standard deviation of daily returns (based on close prices) in each month. I use the VIX as a measure of implied volatility. I normalize both variables before taking the

difference. If option markets are efficient, implied volatility should be an efficient forecast of future volatility, it should subsume the information contained in all other variables in the market information set in explaining future volatility. Thus, $Y_{j,t} = realized_t - implied_t$ captures realized volatility that was unexpected by the financial markets (Christensen and Prabhala 1998).

I order the financial risk variable after macroeconomic variables but before financial markets and monetary policy variables. This assumes that financial conditions are affected contemporaneously by macroeconomic shocks but respond with a lag to shocks to monetary policy. Assuming that financial variables adjust quicker than real variables is justified by the speed at which financial markets respond to news and is a common assumption in the macroeconomic literature (Sims 1980; Christiano et al. 1996; Bloom 2009; Gilchrist and Zakrajšek 2012; Chavleishvili and Manganelli 2024). My variables are ordered as follows: {consumption growth, investment growth, industrial production growth, inflation, financial risk variable $Y_{j,t}$, S&P500 monthly return, change in the ten-year (nominal) Treasury yield, change in the effective (nominal) federal funds rate}. This ordering implies that controls vector X_t must include the contemporaneous values of the four variables ordered before the treatment variable $Y_{j,t}$. Additionally, to control for the broad state of the economy in the recent past, I include the first two lags of all eight variables contained in my ordering in X_t. In short, my timing restriction assumption allows for financial conditions to adjust within the period to consumption growth, investment growth, industrial production growth and inflation, but not to the stock market return, changes of the Treasury yields and changes to the Fed's funds rate. To test the robustness of the conclusions to the timing restriction chosen, I report results obtained using two alternative orderings in the appendix section A5. In particular, I report results with the treatment variable ordered first and last.

Throughout, I focus on three quantiles $\tau \in \{0.1, 0.5, 0.9\}$. The $\tau = 0.1$ quantile is of primary interest as it measures downside-risk. I also report results for a richer set of quantiles for four selected horizons, including the $\tau = 0.05$ quantile corresponding to the

usual definition of GaR. To simplify the analysis I assume a linear specification for the SQF $q_h(\tau \mid y_j) = \alpha_h(\tau) + \beta_h(\tau) y_j$, this ensures that the $\mathsf{QIR}_\tau(h) = \beta_h(\tau)$ does not depend on y_j. I choose a 90% confidence level for reporting the moving block bootstrap confidence intervals, which are computed using a block length of 7 and 1,000 bootstrap replications.



(a) $\tau \in \{0.1, 0.5, 0.9\}$

(b) $\tau = 0.1$ with 90% C.I.

(c) $\tau = 0.5$ with 90% C.I.

(d) $\tau = 0.9$ with 90% C.I.

Figure 7: Cumulative response of Industrial Production (in % pts.) from a shock that increases credit risk by one standard deviation, plotted for three quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. Y-axis is the estimated response $\hat{\beta}_h(\tau)$, x-axis is the horizon h in months. Dashed lines plot the quantile impulse response. Shaded area is the moving block bootstrap 90% Confidence Interval (with block length of 7, and 1,000 bootstrap replications). Note that the impact response (horizon $h = 0$) is by assumption zero, given my timing restrictions.

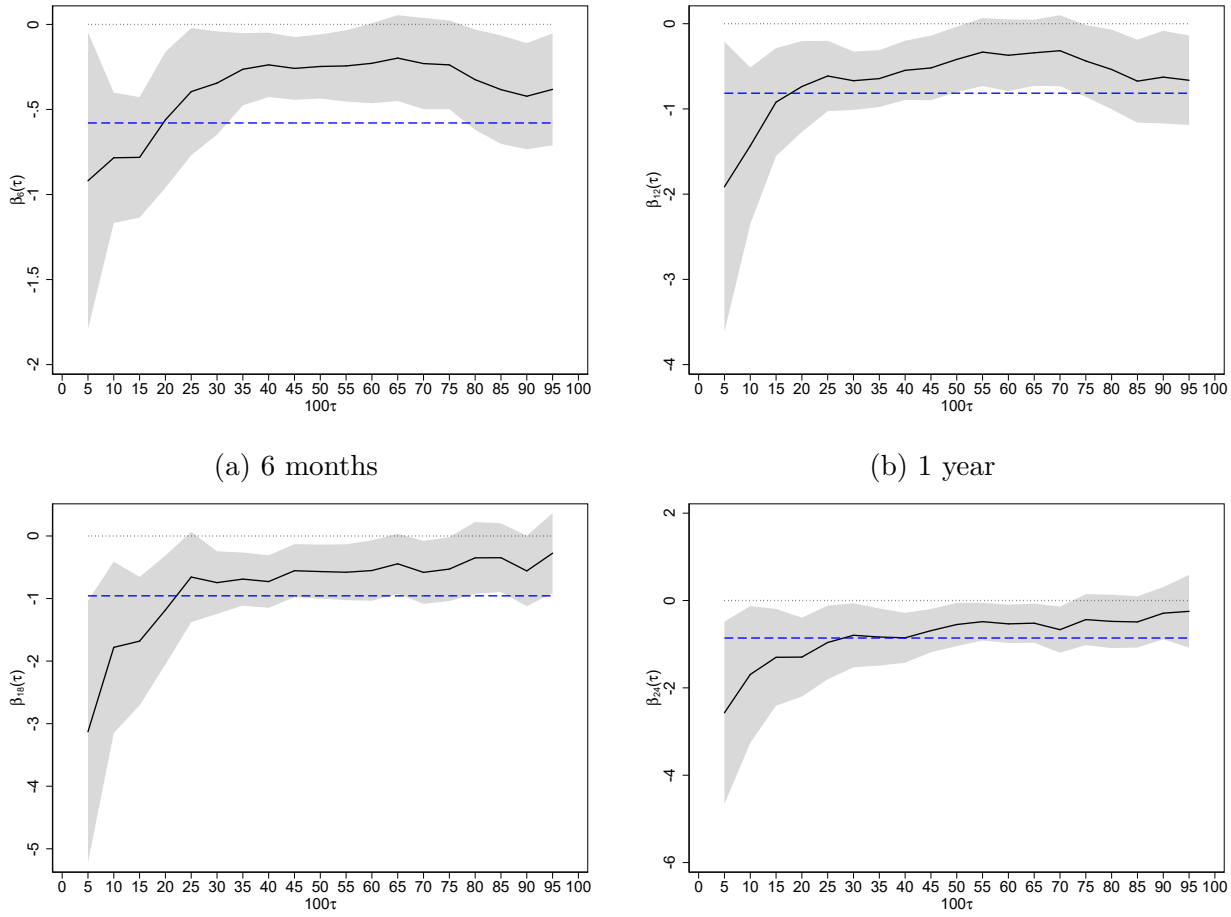
Figure 7 shows the recovered QIRs of industrial production to a shock which increases credit risk by one standard deviation. The upper-left panel in Figure 7 plots the QIRs for the three quantiles on the same axis. It is clear that the response at the $\tau = 0.1$ quantile is

much more pronounced than the response at the other quantiles considered. This is a feature of the data and not of the model, as nothing is restricting the responses of lower quantiles to be lower than those of the upper quantiles. For instance, a shock that lowers the variance of a distribution would give rise to positive QIRs for quantiles below the median and negative QIRs for quantiles above the median.

My findings suggest economically large and statistically significant (at 90% confidence level) growth losses of about 2 percentage points when a credit risk shock propagates in a low growth environment ($\tau = 0.1$). The median losses ($\tau = 0.5$) are considerably smaller at around 0.5 percentage points. The upside-risk response ($\tau = 0.9$) is similar to the median response, except that the effect is not statistically significant beyond the fifteen months horizon. The estimation uncertainty measured by the moving block bootstrap confidence intervals increases with the horizon, it is also higher for the $\tau = 0.1$ quantile than the median and the $\tau = 0.9$ quantile.

The four panels of Figure 8 report results of the same model estimated for a richer set of quantiles (from $\tau = 0.05$ to $\tau = 0.95$ in 0.05 increments) for four fixed horizons (6-months, 1-year, 2-years and 3-years). It shows that at all four of these horizons the slope of the structural quantile function is more negative for lower quantiles. The effect of financial shock on growth is statistically significant but not for all quantiles. Quantiles below the median are more affected by financial shocks and the effect is more likely to be statistically significant even though it is estimated less precisely than the effect around the median growth scenario. The confidence intervals for the GaR $\tau = 0.05$ quantile are considerably wider than for the $\tau = 0.1$ quantile, this is why the $\tau = 0.1$ quantile is often preferred as a measure of downside risk.

Figure 9 shows the results of estimating the same model but using volatility risk in place of credit risk as the treatment variable. Comparing Figure 9 to Figure 7 suggests that the relationship between volatility risk shocks and growth is similar to the relationship between credit risk shocks and growth. The timing and magnitude of the quantile impulse responses



(a) 6 months

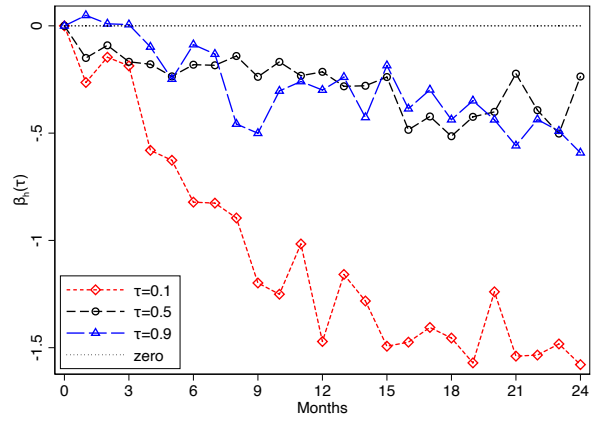
(b) 1 year

(c) 18 months

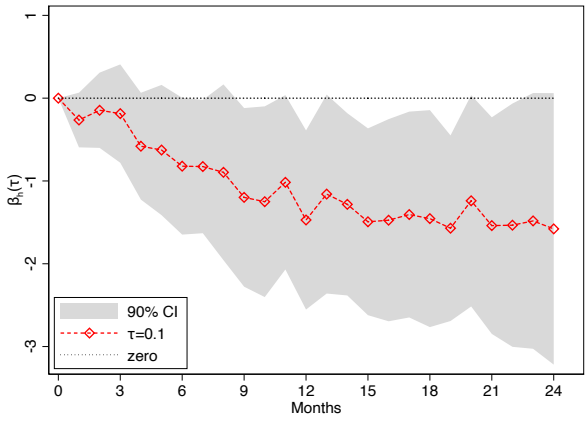
(d) 2 years

Figure 8: Responses of Industrial Production (in % pts.) to a shock that increases credit risk by one standard deviation, plotted for three horizons $h \in \{6, 12, 18, 24\}$ (panels from top left to bottom right). The responses were estimated for quantiles from $\tau = 0.05$ to $\tau = 0.95$ in 0.05 increments. Y-axis is the estimated response $\hat{\beta}_h(\tau)$, x-axis is the quantile τ (multiplied by 100 for legibility). Shaded area is the moving block bootstrap 90% Confidence Interval (with block length of 7, and 1,000 bootstrap replications). Blue dashed line reports the response of the mean estimated from conventional local projections.

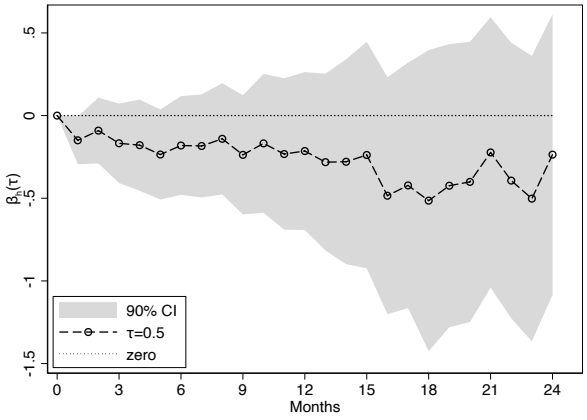
are almost identical following increases in volatility risk and credit risk. Both volatility and credit risk affect down-side more than upside-risk. The similarities are striking considering the fact that the sample correlation coefficient between these two variables is very low at 0.1. These findings suggest either the existence of a common non-linear propagation mechanism (as argued for by Loria et al. 2025) or the fact that it is the overall financial conditions – of which credit and volatility are both components – that have an asymmetric effect on the



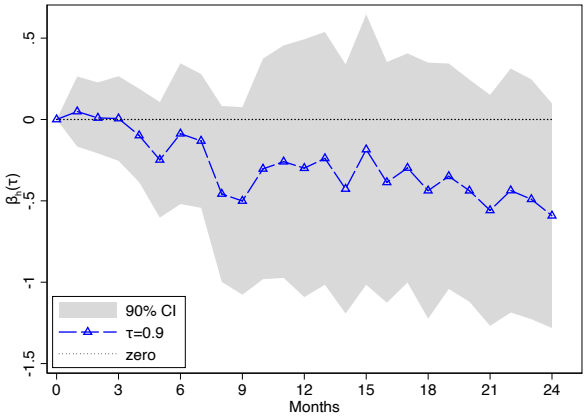
(a) $\tau \in \{0.1, 0.5, 0.9\}$



(b) $\tau = 0.1$ with 90% C.I.



(c) $\tau = 0.5$ with 90% C.I.



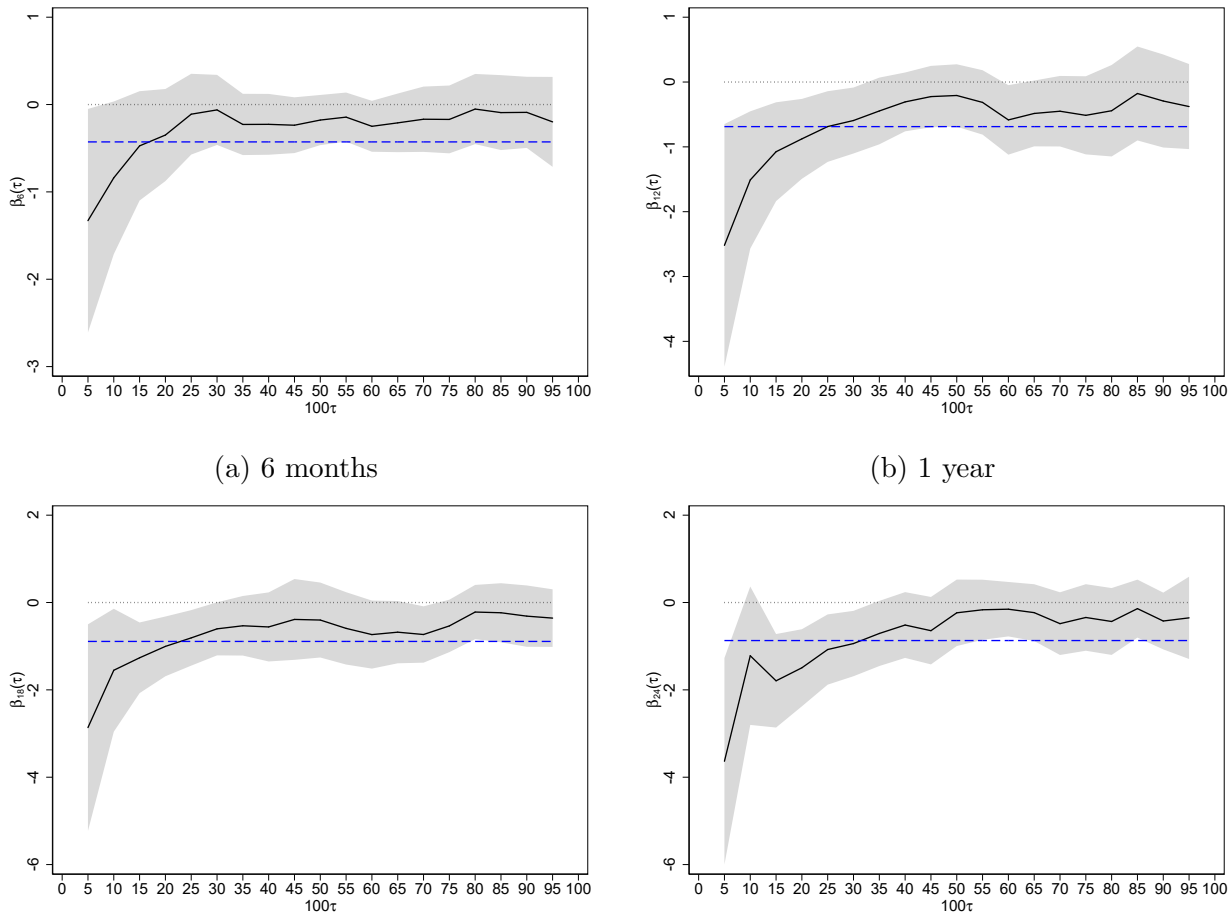
(d) $\tau = 0.9$ with 90% C.I.

Figure 9: Cumulative response of Industrial Production (in % pts.) from a shock that increases volatility risk by one standard deviation, plotted for three quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. Y-axis is $\hat{\beta}_h(\tau)$, x-axis is the horizon h in months. Dashed lines plot the quantile impulse response. Shaded area is the moving block bootstrap 90% Confidence Interval (with block length of 7, and 1,000 bootstrap replications). Note that the impact response (horizon $h = 0$) is by assumption zero, given my timing restrictions.

distribution of output growth.

As before, I report the results for more quantiles at four fixed horizons for the volatility risk in Figure 10. Figures 8 and 10 are nearly identical. Again, this implies that the relationship between financial shocks to the left-tail of growth does not depend on whether the shocks pertain to credit risk or volatility risk.

To highlight the practical implications of using QLP versus GQLP in the context of GaR, I compare the results obtained using both methodologies side by side on Figure 11. Figure



(a) 6 months

(b) 1 year

(c) 18 months

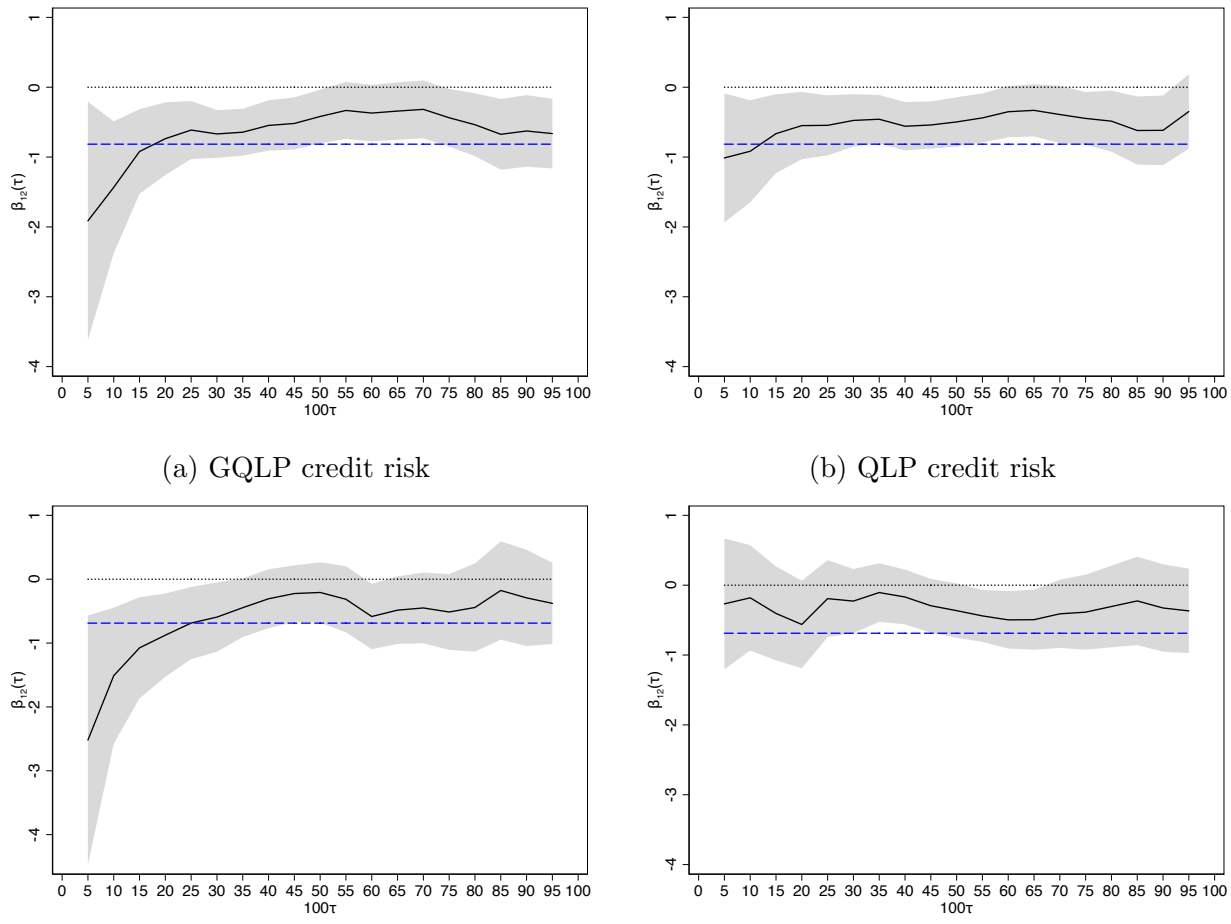
(d) 2 years

Figure 10: Responses of Industrial Production (in % pts.) to a shock that increases volatility risk by one standard deviation, plotted for three horizons $h \in \{6, 12, 18, 24\}$ (panels from top left to bottom right). The responses were estimated for quantiles from $\tau = 0.05$ to $\tau = 0.95$ in 0.05 increments. Y-axis is the estimated response $\hat{\beta}_h(\tau)$, x-axis is the quantile τ (multiplied by 100 for legibility). Shaded area is the moving block bootstrap 90% Confidence Interval (with block length of 7, and 1,000 bootstrap replications). Blue dashed line reports the response of the mean estimated from conventional local projections.

11 plots response of quantiles from $\tau = 0.05$ to $\tau = 0.95$ in 0.05 increments at the one-year horizon. The estimated shape of the quantile function ($\hat{\beta}_{12}(\tau)$) tells us how much asymmetry there is in the response of different parts of the one-year ahead output growth distribution to a financial shock. A flat line would suggests that all quantiles of the distribution are affected equally meaning that the effect of a financial shock is a local shift of the output growth distribution, implying that quantile analysis is redundant. The fact that the line

is upwards sloping means that financial shocks skew the output growth distribution to the left making large negative growth realizations substantially more likely. By comparing the estimates from GQLP versus QLP it is clear that GQLP estimates vary more across quantiles suggesting larger asymmetry between effects of financial shocks on downside versus median and upside growth scenarios. As the reported results from GQLP and QLP are based on the same timing restrictions and the same data, the difference in the estimates comes from the fact that GQLP captures the effect on unconditional quantiles while QLP captures the effect on conditional on controls quantiles. This means that volatility and credit risk shocks have a larger negative effect on unconditionally low quantiles of growth than on conditionally low growth quantiles. Therefore, relying on conditional quantile models can understate the importance of these shocks as causes of recessions.

Table 3 compares the estimates obtained from QLP versus GQLP for two horizons $h \in \{12, 24\}$ and three quantiles $\tau \in \{0.1, 0.5, 0.9\}$ using four different lag length specifications. I use this opportunity to point out another potential advantage of GQLP as a tool for quantile impulse response analysis. Since GQLP allows for inclusion of covariates for identification without affecting the interpretation of the coefficient on the treatment variable, it is less sensitive to potentially arbitrary modeling choices such as the choice of how many lags to include. In fact, for each quantile and horizon in table 3 the standard deviation of the estimated response across the 4 different lag length specifications is greater for the QLP than the GQLP estimator. At the 10th quantile, QLP estimates exhibit considerable sensitivity to lag specification, with credit risk effects at the 12-month horizon varying from -0.55 to -0.94 percentage points and from -0.99 to -1.54 at the 24-month horizon, a range of 0.39 and 0.55 percentage points respectively. For the same quantile and horizons the range of estimates obtained by GQLP is within 0.1 percentage point. For volatility risk both estimators are more sensitive to the choice of lag length, but GQLP estimates are still more stable. Notably, GQLP tends to identify more negative effects at the lower tail (with the exception of the 24-month horizon for volatility risk), suggesting that conditional quantile



(a) GQLP credit risk

(b) QLP credit risk

(c) GQLP volatility risk

(d) QLP volatility risk

Figure 11: Responses of Industrial Production (in % pts.) to a shock that increases credit risk (top panels) or volatility risk (bottom panels) by one standard deviation, plotted for the one year ahead horizon ($h = 12$). The responses were estimated for quantiles from $\tau = 0.05$ to $\tau = 0.95$ in 0.05 increments using GQLPs (left panels) and QLPs (right panels). Y-axis is the estimated response $\hat{\beta}_{12}(\tau)$, x-axis is the quantile τ (multiplied by 100 for legibility). Shaded area is the moving block bootstrap 90% Confidence Interval (with block length of 7, and 1,000 bootstrap replications). Blue dashed line reports the response of the mean estimated from conventional local projections. All models include the same variables, assume the same ordering, and have the same lag length specification.

analysis may underestimate the magnitude of adverse impacts during economic stress periods.

To facilitate comparison with previous studies, I focus on the effects at the twelve-month horizon for the median and the 10th quantile. My findings indicate that a one-standard deviation shock to credit or volatility risk lowers the 10th quantile of growth by approximately 1.5 percentage points, while the losses for median growth are around 0.5 percentage points.

Quantile	Horizon	Lags	Credit risk		Volatility risk	
			QLP	GQLP	QLP	GQLP
0.1	12	1	-0.94	-1.37	-0.22	-1.44
		2	-0.92	-1.43	-0.18	-1.51
		3	-0.64	-1.33	-0.03	-1.46
		4	-0.55	-1.33	-0.33	-1.44
	24	1	-1.32	-1.69	-2.04	-1.61
		2	-0.99	-1.69	-1.53	-1.22
		3	-1.54	-1.73	-1.64	-1.02
		4	-1.29	-1.73	-1.22	-1.02
0.5	12	1	-0.44	-0.57	-0.22	-0.17
		2	-0.50	-0.42	-0.36	-0.21
		3	-0.30	-0.34	-0.45	-0.26
		4	-0.33	-0.42	-0.36	-0.30
	24	1	-0.61	-0.66	-0.20	-0.24
		2	-0.35	-0.55	-0.76	-0.24
		3	-0.32	-0.50	-0.50	-0.06
		4	-0.20	-0.44	-0.63	-0.06
0.9	12	1	-0.84	-0.63	-0.28	-0.22
		2	-0.62	-0.63	-0.33	-0.29
		3	-0.77	-0.63	-0.19	-0.29
		4	-0.90	-0.83	-0.25	-0.26
	24	1	-0.62	-0.28	-0.33	-0.24
		2	-0.69	-0.29	-0.43	-0.42
		3	-1.03	-0.08	0.10	-0.52
		4	-0.50	-0.17	-0.54	-0.47

Table 3: Responses of Industrial Production (in % pts.) to a shock that increases credit risk or volatility risk by one standard deviation, computed for two horizons $h \in \{12, 24\}$ and quantiles $\tau \in \{0.1, 0.5, 0.9\}$. Lags refers to the number of lags included as covariates (2 lags is the baseline specification). QLP refers to the quantile local projection framework which uses the Koenker and Bassett 1978 estimator. GQLP is my local projections based framework which builds on the generalized quantile regression estimator introduced by Powell 2020.

To the best of my knowledge, this paper is the first to identify the effects of financial shocks on unconditional quantiles of growth, so there are no directly comparable findings. However, a rich literature studies this relationship using conditional quantile models. Adrian et al. 2019, in a quantile regression of one-year-ahead GDP growth on National Financial Conditions Index (NFCI) and current GDP, find losses from a one-standard-deviation increase in NFCI of 1.75 and 0.75 percentage points for the $\tau = 0.1$ and $\tau = 0.5$ quantiles, respectively. Importantly, the authors do not assign causal interpretation to these estimates as they do not control for

lags in the regression. Ruzicka 2021 uses QLP (with smoothing) to estimate the effects of NFCI on GDP growth using quarterly data from 1973 to 2015, finding losses of approximately 2.5 and 1.5 percentage points for the $\tau = 0.1$ and $\tau = 0.5$ quantiles, respectively. Loria et al. 2025 study the effects of the EBP on Industrial Production growth using their two-stage methodology. They rescale responses across quantiles such that the median falls by 25 basis points on impact, making direct magnitude comparisons difficult. However, we can compare the ratio of the 10th quantile response to the median response, which Loria et al. 2025 report as 3.7 (averaged over the first year). I find this ratio equals 2.8 for credit risk shocks and 4.1 for volatility risk shocks. Loria et al. 2025 examine the effects of various shocks on growth and find similar asymmetries, suggesting a common, non-linear propagation mechanism. This hypothesis is consistent with my findings that volatility risk and credit risk shocks have nearly identical quantile impulse responses despite being practically uncorrelated. For the euro area, Chavleishvili and Manganelli 2024 use a quantile vector autoregression to study the effects of shocks to the composite indicator of systemic stress (CISS) on euro area industrial production growth, reporting considerable asymmetry with downside risk losses exceeding median losses by a ratio of approximately 4. Chavleishvili et al. 2021 reach similar conclusions using Bayesian methods.

5 Conclusion

Conventional econometric methods that model the mean impulse responses of growth to financial shocks can underestimate the true importance of financial shocks as causes of recessions. This is widely appreciated by academics and policy-makers alike, which explains why a lot of research effort is put devoted to understanding the downside risks to growth.

I offer a new methodology to identify the causal drivers of growth-at-risk. My identification strategy is based on controls, yet it identifies treatment effects on unconditional quantiles. In my view, the distinction between conditional and unconditional quantiles of growth is important in the context of GaR. Conditionally low growth rates map to periods when the

economy under-performs expectations, for example in a favorable macroeconomic climate this would mean high-yet-disappointing growth. On the other hand, unconditionally low growth rates always map to downturns and recessions, and as such are of primary concern for policymakers and academics. My framework allows to study the latter while using familiar controls-based identification strategies based on timing restrictions.

Understanding the structural drivers of growth vulnerability can help discipline theoretical work and macroprudential policy efforts. My empirical findings show that financial shocks have very large effects on downside risks with little upside. This suggests that stabilizing them can help avoid painful recessions, without large growth losses during the expansions.

Several avenues for future research emerge from this work. First, the methodology could be extended to incorporate instrumental variables identification strategies. This extension is straightforward given that the generalized quantile regression estimator of Powell 2020 that underlies GQLP accommodates instrumental variable identification. Second, future research could extend smoothing techniques to the GQLP framework, similar to how Ruzicka (2021) applies smoothing to QLP, building on the smooth local projection approach of Barnichon and Brownlees (2019). Third, the framework offers promising applications beyond growth-at-risk analysis. One potential application of GQLP could be to study inflation-at-risk, examining how monetary policy shocks differentially affect inflation outcomes in high versus low inflation environments. Such analysis could provide valuable insights into the asymmetric transmission of monetary policy and inform optimal policy design across different inflationary regimes. Another application where GQLP could provide novel insights is studying how the size of the fiscal multiplier changes depending on whether the government spending shocks occur during expansions versus contractions. Such insights could enhance the effectiveness of fiscal policy in supporting growth by informing the optimal timing of government spending.

References

Adrian, Tobias, Nina Boyarchenko, and Domenico Giannone (2019). "Vulnerable Growth". In: *American Economic Review* 109.4, pp. 1263–89.

Adrian, Tobias, Fernando Duarte, Nellie Liang, and Pawel Zabczyk (2020). "NKV: A New Keynesian Model with Vulnerability". In: *AEA Papers and Proceedings* 110, pp. 470–76.

Adrian, Tobias, Federico Grinberg, Nellie Liang, Sheheryar Malik, and Jie Yu (2022). "The Term Structure of Growth-at-Risk". In: *American Economic Journal: Macroeconomics* 14.3, pp. 283–323.

Angrist, Joshua, Victor Chernozhukov, and Iván Fernández-Val (2006). "Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure". In: *Econometrica* 74.2, pp. 539–563.

Angrist, Joshua D., Òscar Jordà, and Guido M. Kuersteiner (2018). "Semiparametric Estimates of Monetary Policy Effects: String Theory Revisited". In: *Journal of Business & Economic Statistics* 36.3, pp. 371–387.

Angrist, Joshua D. and Guido M. Kuersteiner (2011). "Causal Effects of Monetary Shocks: Semiparametric Conditional Independence Tests with a Multinomial Propensity Score". In: *The Review of Economics and Statistics* 93.3, pp. 725–747. (Visited on 03/08/2023).

Barnichon, Regis and Christian Brownlees (2019). "Impulse response estimation by smooth local projections". In: *Review of Economics and Statistics* 101 (3).

Bloom, Nicholas (2009). "The Impact of Uncertainty Shocks". In: *Econometrica* 77.3, pp. 623–685.

— (2014). "Fluctuations in Uncertainty". In: *Journal of Economic Perspectives* 28.2, 153–76.

Bochmann, Paul, Daniel Dieckelmann, Stephan Fahr, and Josef Ruzicka (2023). "Financial stability considerations in the conduct of monetary policy". In: *ECB Working Paper* 2870.

Brownlees, Christian and André B.M. Souza (2021). "Backtesting global Growth-at-Risk". In: *Journal of Monetary Economics* 118, pp. 312–330.

Brunnermeier, Markus K. and Yuliy Sannikov (2014). "A Macroeconomic Model with a Financial Sector". In: *American Economic Review* 104.2, 379–421.

Cascaldi-Garcia, Danilo et al. (2023). "What Is Certain about Uncertainty?" In: *Journal of Economic Literature* 61.2, 624–54.

Chavleishvili, Sulkhan, Stephan Fahr, Manfred Kremer, Simone Manganelli, and Bernd Schwaab (2021). "A Risk Management Perspective on Macroprudential Policy". In: *ECB Working Paper Series* 2556.

Chavleishvili, Sulkhan and Simone Manganelli (2024). "Forecasting and stress testing with quantile vector autoregression". In: *Journal of Applied Econometrics* 39.1, pp. 66–85.

Chernozhukov, Victor and Christian Hansen (2013). "Quantile models with endogeneity". In: *Annual Review of Economics* 5.1, pp. 57–81.

Christensen, B.J. and N.R. Prabhala (1998). "The relation between implied and realized volatility." In: *Journal of Financial Economics* 50.2, pp. 125–150.

Christiano, Lawrence, Martin Eichenbaum, and Charles Evans (1996). "The Effects of Monetary Policy Shocks: Evidence from the Flow of Funds". In: *The Review of Economics and Statistics* 78.1, pp. 16–34.

Chuliá, Helena, Ignacio Garrón, and Jorge M. Uribe (2024). "Daily growth at risk: Financial or real drivers? The answer is not always the same". In: *International Journal of Forecasting* 40.2, pp. 762–776.

Engle, Robert F. and Simone Manganelli (2004). "CAViaR: Conditional autoregressive value at risk by regression quantiles". In: *Journal of Business and Economic Statistics* 22 (4).

Galvao, Antonio F., Gabriel Montes-Rojas, and Sung Y. Park (2013). "Quantile Autoregressive Distributed Lag Model with an Application to House Price Returns". In: *Oxford Bulletin of Economics and Statistics* 75 (2), pp. 307–321.

Gertler, Mark, Nobuhiro Kiyotaki, and Andrea Prestipino (2019). "A Macroeconomic Model with Financial Panics". In: *The Review of Economic Studies* 87.1, pp. 240–288.

Gilchrist, Simon and Egon Zakrajšek (2012). "Credit Spreads and Business Cycle Fluctuations". In: *American Economic Review* 102.4, pp. 1692–1720.

Han, Heejoon, Whayoung Jung, and Ji Hyung Lee (2022). "Estimation and Inference of Quantile Impulse Response Functions by Local Projections: With Applications to VaR Dynamics". In: *Journal of Financial Econometrics* 22.1, pp. 1–29.

He, Zhiguo and Arvind Krishnamurthy (2019). "A Macroeconomic Framework for Quantifying Systemic Risk". In: *American Economic Journal: Macroeconomics* 11.4, pp. 1–37.

Jordá, Ó. (2005). "Estimation and inference of impulse responses by local projections". In: *American Economic Review* 95 (1).

Jorda, Oscar and Alan M. Taylor (2025). "Local Projections". In: *Journal of Economic Literature* 63.1, 59–110.

Jordà, Òscar, Martin Kornejew, Moritz Schularick, and Alan M Taylor (2022). "Zombies at Large? Corporate Debt Overhang and the Macroeconomy". In: *The Review of Financial Studies* 35.10, pp. 4561–4586.

Jung, Whayoung and Ji Hyung Lee (2022). "Quantile Impulse Response Analysis with Applications in Macroeconomics and Finance". In: *Advances in Econometrics*.

Kilian, Lutz and Helmut Lütkepohl (2017). *Structural Vector Autoregressive Analysis*. Themes in Modern Econometrics. Cambridge: Cambridge University Press.

Kiyotaki, Nobuhiro and John Moore (1997). "Credit cycles". In: *Journal of Political Economy* 105 (2).

Koenker, Roger (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

Koenker, Roger and Gilbert Bassett (1978). "Regression Quantiles". In: *Econometrica* 46 (1).

Kolesár, Michal and Mikkel Plagborg-Møller (2025). "Dynamic Causal Effects in a Nonlinear World: the Good, the Bad, and the Ugly". In: *arXiv preprint*.

Lee, Dong Jin, Tae-Hwan Kim, and Paul Mizen (2021). "Impulse response analysis in conditional quantile models with an application to monetary policy". In: *Journal of Economic Dynamics and Control* 127, p. 104102.

Linnemann, Ludger and Roland Winkler (2016). "Estimating nonlinear effects of fiscal policy using quantile regression methods". In: *Oxford Economic Papers* 68 (4).

Loria, Francesca, Christian Matthes, and Donghai Zhang (2025). "Assessing Macroeconomic Tail Risk". In: *The Economic Journal* 135.665, pp. 264–284.

McCracken, Michael W. and Serena Ng (2015). "FRED-MD: A Monthly Database for Macroeconomic Research". In: *Journal of Business & Economic Statistics* 2015-012B. Revision of 2015-012.

Montes-Rojas, Gabriel (2019). "Multivariate Quantile Impulse Response Functions". In: *Journal of Time Series Analysis* 40 (5), pp. 739–752.

Mumtaz, Haroon and Paolo Surico (2015). "The transmission mechanism in good and bad times". In: *International Economic Review* 56.4, pp. 1237–1260.

Plagborg-Møller, Mikkel, Lucrezia Reichlin, Giovanni Ricco, and Thomas Hasenzagl (2020). "When is growth at risk?" en. In: *Brookings Pap. Econ. Act.* 2020.1, pp. 167–229.

Plagborg-Møller, Mikkel and Christian K. Wolf (2021). "Local Projections and VARs Estimate the Same Impulse Responses". In: *Econometrica* 89 (2).

Powell, David (2020). "Quantile Treatment Effects in the Presence of Covariates". In: *The Review of Economics and Statistics* 102.5, pp. 994–1005.

Prasad, Ananthakrishnan, Selim Elekdag, Phakawa Jeasakul, Romain Lafarguette, Adrian Alter, Alan Feng, and Changchun Wang (2019). "Growth at Risk: Concept and Application in IMF Country Surveillance". In: *IMF Working Paper* 19.36.

Rambachan, Ashesh and Neil Shephard (2021). "When do common time series estimands have nonparametric causal meaning?" In: *Manuscript, Harvard University*.

Ruzicka, Josef (2021). "Quantile local projections: Identification, smooth estimation, and inference". In: *Universidad Carlos III de Madrid*.

Sims, Christopher A (1980). "Macroeconomics and Reality". In: *Econometrica* 48.1, pp. 1–48.

Appendix

A1 Proofs

A1.1 Proof of theorem 1

Using conditional independence, I write:

$$\begin{aligned}\mathbb{E}[Y_{i,t+h}\mathbf{1}\{Y_{j,t}=y_j\}\mid X_t] &= \mathbb{E}[Y_{i,t+h}(W_{1:t+h})\mathbf{1}\{Y_{j,t}=y_j\}\mid X_t]\\ &= \mathbb{E}[Y_{i,t+h}(g_j^{-1}(y_j,X_t))\mathbf{1}\{Y_{j,t}=y_j\}\mid X_t].\end{aligned}$$

Expand via expectation and covariance:

$$\begin{aligned}&= \mathbb{E}[Y_{i,t+h}(g_j^{-1}(y_j,X_t))\mid X_t]\,\mathbb{E}[\mathbf{1}\{Y_{j,t}=y_j\}\mid X_t]\\ &\quad+\ \mathrm{Cov}\big(Y_{i,t+h}(g_j^{-1}(y_j,X_t)),\mathbf{1}\{Y_{j,t}=y_j\}\mid X_t\big).\end{aligned}$$

Under conditional random assignment, the covariance is zero

$$\begin{aligned}&\mathbb{E}[Y_{i,t+h}\mathbf{1}\{Y_{j,t}=y_j\}\mid X_t]\\ &= \mathbb{E}[Y_{i,t+h}(g_j^{-1}(y_j,X_t))\mid X_t]\,\mathbb{E}[\mathbf{1}\{Y_{j,t}=y_j\}\mid X_t].\end{aligned}$$

Use the identity $\mathbb{E}[A\mid B]=\frac{\mathbb{E}[A\mathbf{1}\{B\}]}{\mathbb{E}[\mathbf{1}\{B\}]}$ to obtain:

$$\mathbb{E}[Y_{i,t+h}\mid Y_{j,t}=y_j,X_t]=\mathbb{E}\big[Y_{i,t+h}(g_j^{-1}(y_j,X_t))\mid X_t\big].$$

Taking partial derivative w.r.t. y_j, exchanging derivative/expectation and applying the chain rule yields:

$$\begin{aligned}\frac{\partial}{\partial y_j}\,\mathbb{E}[Y_{i,t+h} \mid Y_{j,t} = y_j, X_t] &= \mathbb{E}\Big[\frac{\partial}{\partial y_j} Y_{i,t+h}(g_j^{-1}(y_j, X_t)) \mid X_t\Big] \\ &= \mathbb{E}\Big[\frac{\partial Y_{i,t+h}(w_j)}{\partial w_j}\,\frac{\partial g_j^{-1}(y_j, X_t)}{\partial y_j} \mid X_t\Big].\end{aligned}$$

A1.2 Proof of theorem 2

$Y_t \in \mathbb{R}^k$ admits a structural Wold representation

$$Y_t = \sum_{\ell=0}^{\infty} \Psi_\ell W_{t-\ell}, \qquad \sum_{\ell=0}^{\infty} \|\Psi_\ell\| < \infty,$$

where $\{W_t\}_{t\in\mathbb{Z}}$ are i.i.d. $\mathcal{N}(0, I_k)$. Fix $i, j \in \{1, \dots, k\}$ and a horizon $h \geq 0$. Write the i-th component at horizon h as

$$Y_{i,t+h} = [\Psi_h]_{i,j}\, W_{j,t} + R_{i,t+h},$$

where

$$R_{i,t+h} := \sum_{m\neq j} [\Psi_h]_{i,m} W_{m,t} + \sum_{\ell\neq h} \sum_{m=1}^{k} [\Psi_\ell]_{i,m} W_{m,t+h-\ell}$$

Because the innovations are i.i.d. Gaussian, $W_{j,t}$ is independent of $R_{i,t+h}$ (uncorrelated jointly Gaussian variables are independent). Applying the definition of potential outcomes yields:

$$Y_{i,t+h}(w_j) = [\Psi_h]_{i,j}\, w_j + R_{i,t+h}.$$

Mean impulse response (IR):

Taking expectations yields:

$$\mathbb{E}[Y_{i,t+h}(w_j)] = \mathbb{E}[R_{i,t+h} + [\Psi_h]_{i,j} w_j] = \mathbb{E}[R_{i,t+h}] + [\Psi_h]_{i,j}\, w_j.$$

Differentiating w.r.t. w_j yields:

$$\mathsf{IR}(h) = \frac{\partial\, \mathbb{E}[Y_{i,t+h}(w_j)]}{\partial w_j} = [\Psi_h]_{i,j}.$$

Conditional mean impulse response (cIR):

Let X_t be any vector of covariates measurable with respect to past and current information. Using the structural Wold representation,

$$Y_{i,t+h}(w_j, x) = [\Psi_h]_{i,j}\, w_j + R_{i,t+h}(x),$$

where $R_{i,t+h}(x)$ collects all terms independent of $W_{j,t}$.

$$\mathbb{E}[Y_{i,t+h}(w_j, x)] = [\Psi_h]_{i,j}\, w_j + \mathbb{E}[R_{i,t+h}(x)].$$

Differentiating with respect to w_j gives

$$\mathsf{cIR}(h) = \frac{\partial \mathbb{E}[Y_{i,t+h}(w_j, x)]}{\partial w_j} = [\Psi_h]_{i,j}.$$

Quantile impulse response (QIR):

$$\begin{aligned} q_h(\tau \mid w_j) &\equiv q_{Y_{i,t+h}(w_j)}(\tau \mid W_{t,j} = w_j) \\ &= q_{R_{i,t+h}+[\Psi_h]_{i,j}w_j}(\tau \mid W_{t,j} = w_j) \\ &= q_{R_{i,t+h}}(\tau \mid W_{t,j} = w_j) + [\Psi_h]_{i,j}w_j \\ &= q_{R_{i,t+h}}(\tau) + [\Psi_h]_{i,j}w_j \end{aligned}$$

The first equivalence is stated to remind the reader of a notational short-cut used throughout the paper. The first equality sign follows from the formula for the potential outcome. The second equality sign follows from the fact that $[\Psi_h]_{i,j}w_j$ is a constant and so it can go outside of the quantile function. The last equality sign follows from the independence

of $W_{j,t}$ w.r.t $R_{i,t+h}$.

Differentiating w.r.t. w_j yields:

$$\mathsf{QIR}_\tau(h) = \frac{\partial\, q_h(\tau \mid W_{j,t} = w_j)}{\partial w_j} = [\Psi_h]_{i,j}.$$

Therefore $\mathsf{QIR}_\tau(h) = [\Psi_h]_{i,j}$ for all $\tau \in (0,1)$.

Conditional quantile impulse response (cQIR):

Let X_t be any vector measurable w.r.t. past and current observables.

$$\begin{aligned} q_h(\tau \mid w_j, x) &\equiv q_{Y_{i,t+h}(w_j,x)}(\tau \mid W_{j,t} = w_j,\, X_t = x) \\ &= q_{R_{i,t+h}+[\Psi_h]_{i,j}\, w_j}(\tau \mid W_{j,t} = w_j,\, X_t = x) \\ &= q_{R_{i,t+h}}(\tau \mid W_{j,t} = w_j,\, X_t = x) + [\Psi_h]_{i,j}\, w_j \\ &= q_{R_{i,t+h}}(\tau \mid X_t = x) + [\Psi_h]_{i,j}\, w_j \end{aligned}$$

The first equivalence is stated to remind the reader of a notational short-cut used throughout the paper. The first equality substitutes the formula for the potential outcome. The second equality sign follows from the fact that $[\Psi_h]_{i,j} w_j$ is a constant and so it can go outside of the quantile function. The last equality sign follows from the independence of $W_{j,t}$ w.r.t $R_{i,t+h}$.

Differentiating w.r.t. w_j yields:

$$\mathsf{cQIR}_\tau(h) = \frac{\partial\, q_h(\tau \mid W_{j,t} = w_j, , X_t = x)}{\partial w_j} = [\Psi_h]_{i,j}.$$

Therefore $\mathsf{cQIR}_\tau(h) = [\Psi_h]_{i,j}$ for all $\tau \in (0,1)$.

This completes the proof that $\mathsf{cIR}(h) = \mathsf{IR}(h) = \mathsf{QIR}_\tau(h) = \mathsf{cQIR}_\tau(h)\ \forall h$ and $\forall \tau$.

A1.3 Proof of theorem 3

I reformulate the theorem 3 from Powell 2020 in my setting. First, I want to show: $\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid Y_{j,t}, X_t^\top] = \mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid X_t^\top]$. Evaluating the left hand side of the equality yields:

$$\begin{aligned}
\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid Y_{j,t}, X_t^\top] &= \mathbb{P}[Y_{i,t+h}(Y_{j,t}) \leq q_h(\tau \mid Y_{j,t}) \mid Y_{j,t}, X_t^\top] \\
&= \mathbb{P}[Y_{i,t+h}(y_j) \leq q_h(\tau \mid y_j) \mid Y_{j,t}, X_t^\top] \\
&= \mathbb{P}[Y_{i,t+h}(y_j) \leq q_h(\tau \mid y_j) \mid X_t^\top].
\end{aligned}$$

The first equality sign follows from the definition of a potential outcome. The second equality sign comes from the rank similarity assumption 7 which must hold for all d, d' and thus also for $d = Y_{j,t}$. The third equality sign follows from the conditional (on X_t) independence of $Y_{j,t}$ which follows from assumptions 3 and 5. Evaluating the right hand side of the equality yields:

$$\begin{aligned}
\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid X_t^\top] &= \mathbb{P}[Y_{i,t+h}(Y_{j,t}) \leq q_h(\tau \mid Y_{j,t}) \mid X_t^\top] \\
&= \int \mathbb{P}[Y_{i,t+h}(Y_{j,t}) \leq q_h(\tau \mid Y_{j,t}) \mid X_t^\top, Y_{j,t}] d\mathbb{P}(Y_{j,t} \mid X_t^\top) \\
&= \int \mathbb{P}[Y_{i,t+h}(y_j) \leq q_h(\tau \mid y_j) \mid Y_{j,t}, X_t^\top] d\mathbb{P}(Y_{j,t} \mid X_t^\top) \\
&= \mathbb{P}[Y_{i,t+h}(y_j) \leq q_h(\tau \mid y_j) \mid X_t^\top].
\end{aligned}$$

The first equality follows from the definition of a potential outcome. The third equality follows from the rank similarity assumption 7. The second and fourth equality follow directly from properties of marginal probability functions.

Now I want to show: $\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t})] = \tau$.

$$\begin{aligned}
\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t})] &= \int \mathbb{P}[Y_{i,t+h}(Y_{j,t}) \leq q_h(\tau \mid Y_{j,t}) \mid X_t^\top, Y_{j,t}] d\mathbb{P}(X_t^\top, Y_{j,t}) \\
&= \int \mathbb{P}[Y_{i,t+h}(y_j) \leq q_h(\tau \mid y_j) \mid X_t^\top, Y_{j,t}] d\mathbb{P}(X_t^\top, Y_{j,t}) \\
&= \mathbb{P}[Y_{i,t+h}(y_j) \leq q_h(\tau \mid y_j)] \\
&= \tau
\end{aligned}$$

The second equality follows from the rank similarity assumption 7. The fourth equality follows from assumption 6.

A2 Bootstrap algorithm

1. Bootstrap Setup: Initialize B bootstrap replications and create empty storage for parameter estimates of interest.

2. Block Resampling Loop: For each bootstrap replication $b = 1, \ldots, B$, randomly select starting positions and construct blocks of M consecutive observations to create a pseudo-sample of size T that preserves temporal dependence.

3. Model Estimation: Re-estimate the econometric model on each bootstrap pseudo-sample and extract the parameter estimates of interest.

4. Bootstrap Storage: Store the difference between each bootstrap estimate and the original sample estimate: $\hat{\beta}_h(\tau)_b - \hat{\beta}_h(\tau)$ for bootstrap replication b.

5. Confidence Interval Construction: Compute the standard deviation of the B bootstrap estimates and construct normal-based confidence intervals as:

$$\text{CI} = \hat{\beta}_h(\tau) \pm z_{\alpha/2} \times \text{SE}_{\text{bootstrap}}(\hat{\beta}_h(\tau))$$

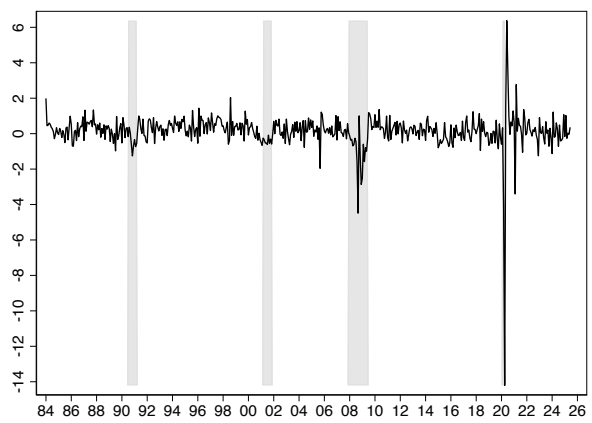
where $\text{SE}_{\text{bootstrap}}(\hat{\beta}_h(\tau)) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\beta}_h(\tau)_b - \hat{\beta}_h(\tau))^2}$.

A3 Monte Carlo results - endogenous volatility SVAR

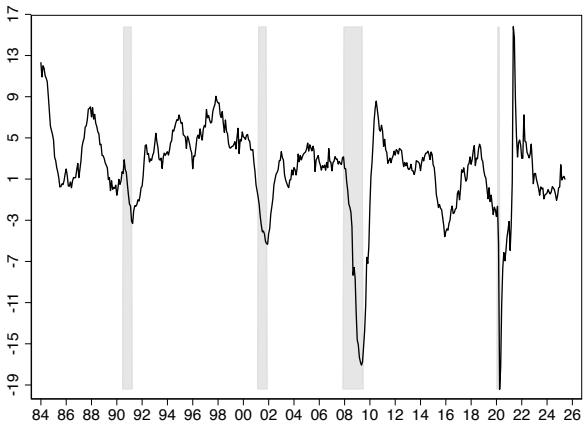
Quantile	Horizon	QLP no controls		QLP with controls		GQLP	
		Mean Bias	RMSE	Mean Bias	RMSE	Mean Bias	RMSE
0.1	1	-0.840	0.886	-0.176	0.194	0.000	0.197
	2	-0.981	1.064	-0.148	0.237	-0.002	0.302
	3	-1.079	1.204	-0.113	0.307	-0.002	0.405
	4	-1.143	1.310	-0.083	0.381	0.003	0.496
	5	-1.170	1.381	-0.064	0.464	0.007	0.584
	6	-1.174	1.431	-0.052	0.551	0.012	0.657
	7	-1.161	1.477	-0.029	0.620	0.027	0.716
	8	-1.173	1.530	-0.040	0.672	-0.001	0.766
	9	-1.160	1.569	-0.016	0.728	0.018	0.834
	10	-1.162	1.600	-0.020	0.776	0.017	0.875
0.5	1	-0.628	0.646	-0.010	0.061	-0.002	0.108
	2	-0.779	0.809	-0.026	0.128	-0.002	0.166
	3	-0.857	0.903	-0.034	0.187	-0.005	0.212
	4	-0.900	0.956	-0.029	0.235	0.001	0.261
	5	-0.930	0.999	-0.027	0.273	-0.003	0.302
	6	-0.948	1.029	-0.032	0.320	-0.001	0.338
	7	-0.968	1.065	-0.044	0.358	-0.006	0.377
	8	-0.971	1.083	-0.022	0.391	-0.005	0.393
	9	-0.980	1.108	-0.024	0.411	-0.005	0.426
	10	-0.974	1.119	-0.017	0.442	-0.007	0.455
0.9	1	-0.537	0.559	0.183	0.198	-0.014	0.145
	2	-0.640	0.674	0.172	0.230	-0.019	0.206
	3	-0.694	0.740	0.153	0.268	-0.023	0.259
	4	-0.740	0.797	0.109	0.295	-0.025	0.308
	5	-0.755	0.825	0.090	0.336	-0.026	0.345
	6	-0.774	0.865	0.074	0.377	-0.030	0.397
	7	-0.781	0.891	0.061	0.407	-0.020	0.434
	8	-0.777	0.908	0.062	0.444	-0.015	0.464
	9	-0.792	0.942	0.037	0.476	-0.026	0.496
	10	-0.803	0.967	0.037	0.509	-0.034	0.527

Table 4: Simulation results for the cumulative QIR of the illustrative example (complementing Figure 3 in the main text). The "true" QIR to which the estimators were compared with is in fact a linear approximation obtained from the quantile local projection model $Y_{t+h} = \alpha_h(U_{i,t+h}) + \beta_h(U_{i,t+h})W_{j,t}$. The true QIR was obtained by averaging the estimated $\beta_h(\tau)$ over the Monte Carlo replications. RMSE is the root mean squared error. QLP refers to the quantile local projection framework which uses the Koenker and Bassett 1978 estimator, GQLP uses the generalized quantile regression estimator introduced by Powell 2020. Results are from $MC = 1{,}000$ simulation replications, with a time-series of length $T = 500$ (after dropping 1,000 initial observations).

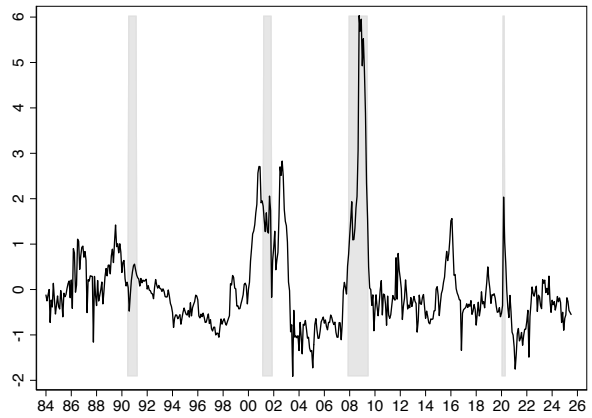
A4 Empirical results - additional figures and tables



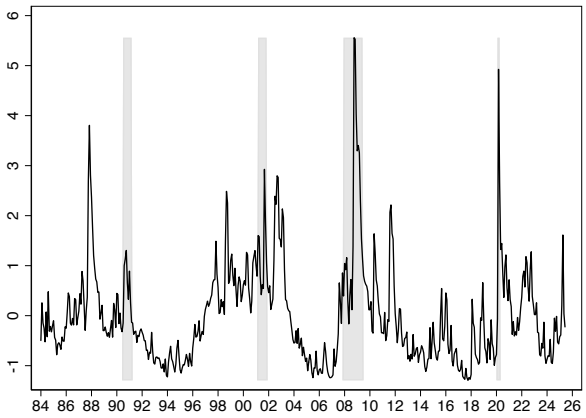
(a) $Y_{i,t+1}$ - Monthly IP growth



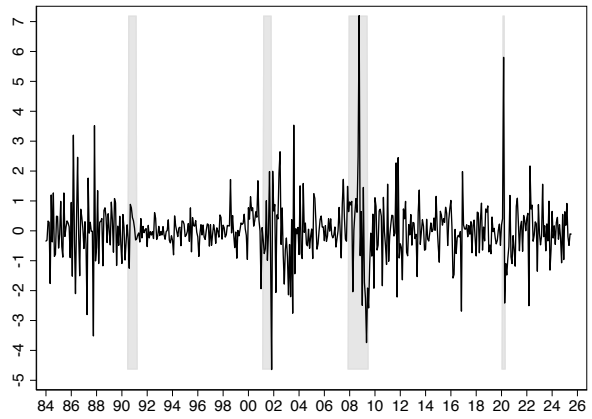
(b) $Y_{i,t+12}$ - Annual IP growth



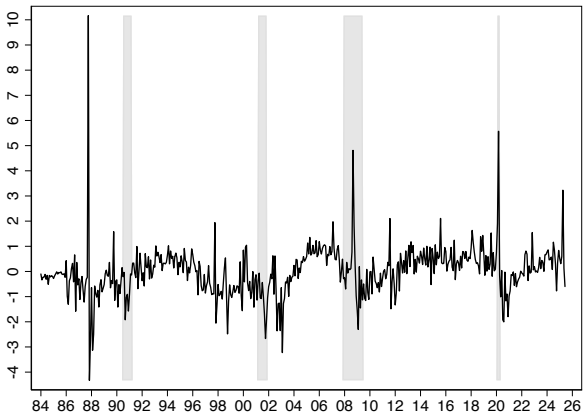
(c) Excess Bond Premium



(d) VIX Index



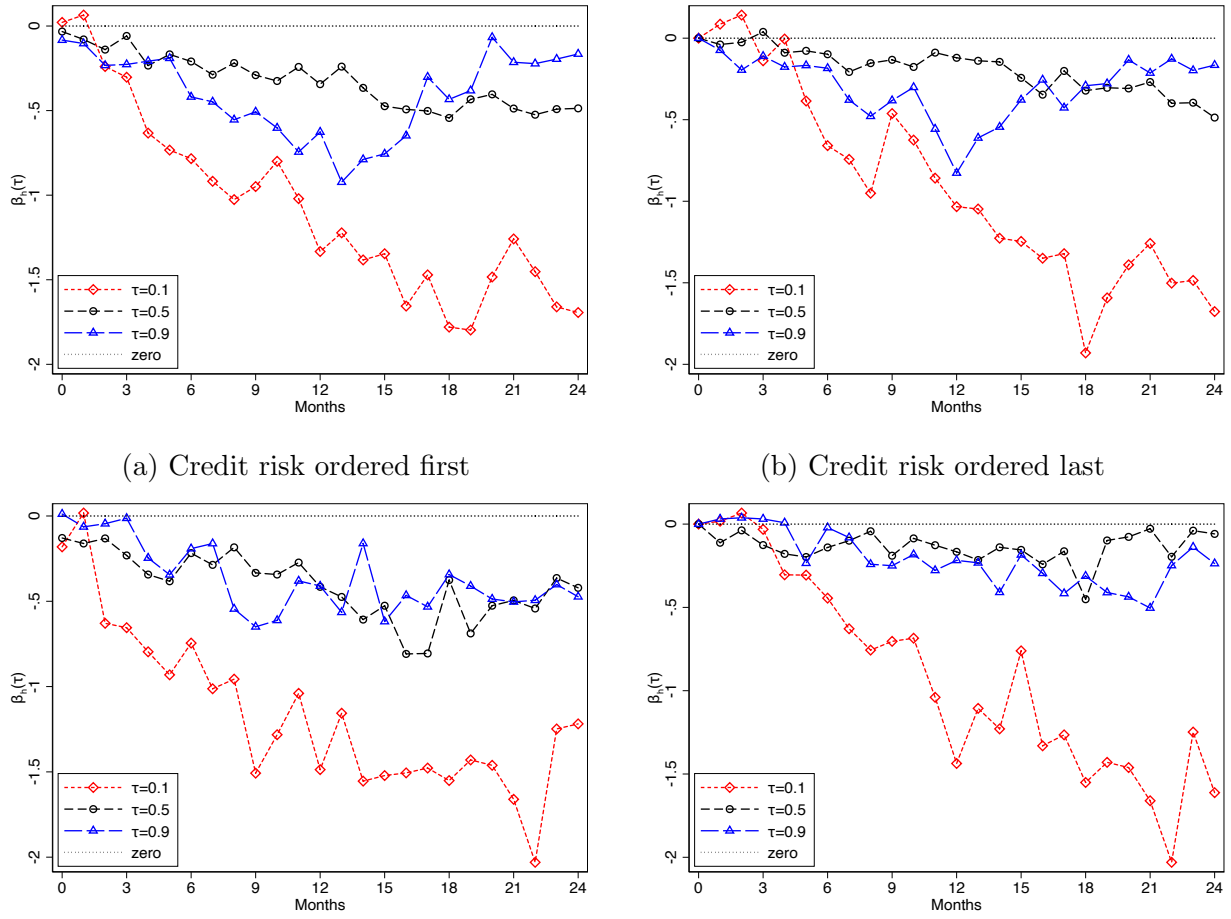
(e) $Y_{j,t}$ - Credit Risk



(f) $Y_{j,t}$ - Volatility Risk

Figure 12: Monthly time-series from January 1984 to June 2025. Grey bands indicate NBER recession dates. The series in the bottom panels have been normalized.

A5 Robustness



(a) Credit risk ordered first

(b) Credit risk ordered last

(c) Volatility risk ordered first

(d) Volatility risk ordered last

Figure 13: Cumulative response of Industrial Production (in % pts.) from a shock that increases credit (top panels) or volatility (bottom panels) risk by one standard deviation, plotted for three quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. Y-axis is the estimated response $\hat{\beta}_h(\tau)$, x-axis is the horizon h in months. Dashed lines plot the quantile impulse response. Note that the impact response (horizon $h = 0$) is by assumption zero when the shock is ordered after the dependent variable.