Fixed-Smoothing Uniform Inference for Quantile Regression

Kaicheng Chen[1] Antonio F. Galvao[2] Seunghwa Rho[3] Timothy J. Vogelsang[2]
Jungmo Yoon[3]

[1]Shanghai University of Finance and Economics

[2]Michigan State University

[3]Hanyang University

IESR, Jinan University, Guangzhou
June 5, 2026

Motivation and challenge

- **Motivation**:
  - Quantile regression (QR): effect heterogeneity over the distribution of an outcome;
  - Time series applications: tail outcomes, downside risk, local persistence, asymmetric dynamics;
  - Pointwise inference is invalid when the question is about a curve or a region.
- **Goal:** fixed-smoothing HAR inference for the quantile process $\{\beta(\tau) : \tau \in \mathcal{T}\}$, accounting for serial and cross-quantile dependence.
- **Key challenge:** under temporal dependence, the test statistics are NOT asymptotically pivotal.

Main contributions

- **Approach:** two HAR estimators (kernel and orthonormal-series); fixed-smoothing (fixed-b, fixed-K) approximations; implementations:

  1. **Direct method:** simulate the limiting Gaussian process with estimated covariance kernel.

  2. **Stack method:** for a fixed finite grid of quantiles, nuisance cancels out asymptotically–use standard pivotal fixed-smoothing limits.

- **Applications:** (i) Wald/t tests for hypotheses across multiple or a continuum of quantiles, (ii) uniform confidence bands, (iii) tests for shape restrictions (monotonicity, homogeneity).

Roadmap

1. Motivation
2. Model and Inference Problem
3. Fixed-b Limit Theory
4. Inference Procedures
5. Simulation Evidence
6. Empirical Application
7. Conclusion and Takeaways

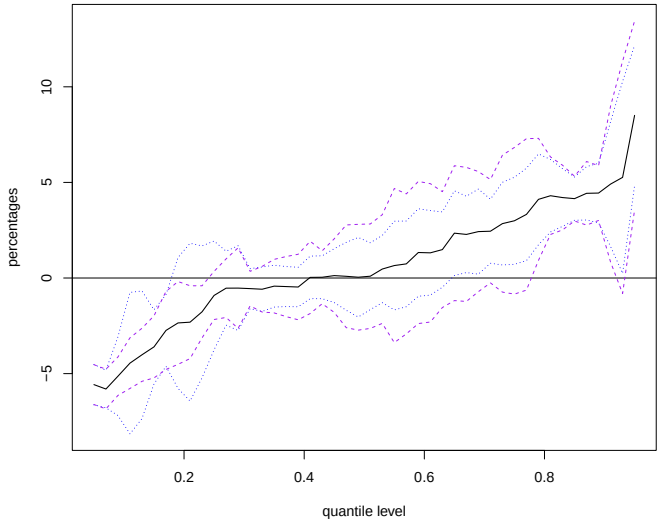
Motivation

Quantile regression answers distributional questions

- Mean regression asks how x_t shifts the conditional mean.
- Quantile regression asks how x_t shifts different parts of the outcome distribution.
- Questions commonly involve different parts of the outcome distribution.
- We will see an example of stock return prediction across the return distribution–highly heterogeneous.

Heterogeneouts stock return predictability

Predicting monthly S&P 500 excess returns using lagged market-wide volatility:



Why pointwise inference can mislead

- **Pointwise question**: Is $\beta(\tau_0) = 0$ at one pre-specified τ_0?
- **Multiple testing question**: Is $\beta(\tau_1) = \cdots = \beta(\tau_m) = 0$ for a finite set $(\tau_1, \ldots, \tau_m)$?
- **Uniform question**: Is $\beta(\tau) = 0$ for all $\tau \in \mathcal{T}$?
- **Shape question**: Is $\beta(\tau)$ constant or monotone over $\mathcal{T}$?
- **Confidence band**: critical values that account for cross-quantile dependence?

Where the paper fits

- Existing time-series QR methods are pointwise in τ: Gregory, Lahiri, and Nordman (2018), Galvao & Yoon (2024); Hoga & Schulz (2025); Cai and Long (2026).
- Existing uniform QR inference is mostly designed for iid or non-HAR settings:
  - Quantile process: Koenker & Xiao (2002); Qu & Yoon (2015); Belloni, Chernozhukov, Chetverikov, and Fernández-Val (2019).
  - Quantile effects stability: Qu (2008); Su & Xiao (2008); Galvao, Kato, Montes-Rojas, and Olmo (2014);
  - **One exception:** Chernozhukov & Fernández-Val (2005).
- Fixed-smoothing HAR inference improves finite-sample approximation of variance estimators and tests: Kiefer & Vogelsang (2002, 2005); Sun (2011, 2013); Chen & Vogelsang (2024); Hwang & Valdés (2025).

Model and Inference Problem

Econometric model

For $t = 1, \ldots, T$ and $\tau \in \mathcal{T}$, where $\mathcal{T}$ is a subset of $(0, 1)$,

$$Q_{y_t}(\tau \mid x_t) = x_t' \beta_0(\tau).$$

The QR estimator is

$$\widehat{\beta}(\tau) = \arg\min_{\beta} \sum_{t=1}^{T} \rho_\tau(y_t - x_t'\beta), \qquad \rho_\tau(u) = u\{\tau - 1(u \leq 0)\}.$$

It is well known that, for given τ and under regularity conditions, as $T \to \infty$,

$$\sqrt{T}(\widehat{\beta}(\tau) - \beta_0(\tau)) \xrightarrow{d} N(0, \Sigma(\tau)),$$

where $\Sigma(\tau) = D(\tau)^{-1}\Lambda(\tau)D(\tau)^{-1}$, $D(\tau) = \mathrm{E}\left[f(0|x_t)x_t x_t'\right]$, and $\Lambda(\tau)$ is the long-run covariance.

The QR score process

However, when we treat the quantile effect as a process indexed by $\tau \in \mathcal{T}$, we need to consider cross-quantile dependence.

Define the score

$$z_t(\tau) = \{\tau - 1(e_t(\tau) \leq 0)\}x_t.$$

The long-run covariance kernel is

$$\Lambda(\tau_1, \tau_2) = \sum_{j=-\infty}^{\infty} \text{Cov}\{z_0(\tau_1), z_j(\tau_2)\}.$$

And $\Lambda(\tau) = \Lambda(\tau, \tau)$.

Remark: $\Lambda(\tau, \tau)$ handles serial dependence at one quantile, and $\Lambda(\tau_1, \tau_2)$ also handles dependence across quantiles.

Non-separable Gaussian process

If $e_t(\tau)$ is conditionally **i.i.d**, or $\tau - 1\{e_t(\tau) \leq 0\}$ is **MDS**, $\text{Cov}\{z_s(\tau_1), z_t(\tau_2)\} = 0$ for $s \neq t$.

In general, $\text{Cov}\{z_s(\tau_1), z_t(\tau_2)\} = -\tau_1\tau_2 E[x_s x_t'] + E[1(e_s(\tau_1) \leq 0)1(e_t(\tau_2) \leq 0)x_s x_t']$

For $s = t$, the last term reduces to $E\left[E[1(e_t(\tau_1) \leq 0)1(e_t(\tau_2) \leq 0)|x_t]x_t x_t'\right]$. Furthermore,

$$\begin{aligned} &E[1(e_t(\tau_1) \leq 0)1(e_t(\tau_2) \leq 0)|x_t] = E[1\{y_t \leq x_t\beta(\tau_1); y_t \leq x_t\beta(\tau_2)\}|x_t] \\ &= \Pr[y_t \leq Q_y(\tau_1|x_t); y_t \leq Q_y(\tau_2|x_t)|x_t] = \Pr[F_y(y_t|x_t) < \tau_1; F_y(y_t|x_t) < \tau_2|x_t] = \tau_1 \wedge \tau_2 \end{aligned}$$

where the second last equality follows if $F_y(.|x)$ is continuous and strictly increasing.

Non-separable Gaussian process

It can be shown (by empirical process methods; e.g., Arcones & Yu, 1994) that $\frac{1}{\sqrt{T}} \sum_{t=1}^{[rT]} z_t(\tau)$ has a limiting Gaussian process $G(r, \tau)$ with covariance function

$$\text{Cov}(G(r_1, \tau_1), G(r_2, \tau_2)) = (r_1 \wedge r_2)\Lambda(\tau_2, \tau_2).$$

We have shown, **for i.i.d or MDS errors**, $\Lambda(\tau_1, \tau_2) = (\tau_1 \wedge \tau_2 - \tau_1\tau_2) \sum_{j=-\infty}^{\infty} E[x_0 x_j']$

Note that, for a standard Kiefer process $K(r, \tau)$, $\text{Cov}[K(r_1, \tau_1)K(r_1, \tau_2)] = (\tau_1 \wedge \tau_2 - \tau_1\tau_2)$

Therefore, we can write $G(r, \tau) = \Gamma K(r, \tau)$, where $\Gamma\Gamma' = \sum_{j=-\infty}^{\infty} E[x_0 x_j']$.

But, in general, $G(r, \tau)$ **cannot** be written as a separable structure.

HAR covariance estimation for a given quantile level

A Powell's kernel estimator for the outer part is given by, with $k(.)$ being a density kernel,

$$\widehat{D}(\tau) = \frac{1}{hT} \sum_{t=1}^{T} k\left(\frac{\widehat{e}_t(\tau)}{h}\right) x_t x_t' \xrightarrow{p} D(\tau) \quad \text{(following Kato, 2012)}$$

A kernel HAR estimator takes the form, with $\mathcal{K}(.)$ being a lag-window kernel,

$$\widehat{\Lambda}(\tau) = \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \mathcal{K}\left(\frac{t-s}{M}\right) \widehat{z}_t(\tau) \widehat{z}_s(\tau)'.$$

- Small-b asymptotics: $M/T \to 0$.
- Fixed-b asymptotics: $M = [bT]$ with $b \in (0, 1]$ fixed.
- Fixed-b reference distributions account for the smoothing choice.

Testing at one quantile

For a restriction $r(\beta(\tau)) = 0$, define

$$\mathcal{W}(\tau) = Tr(\widehat{\beta}(\tau))' \left\{ R\left(\widehat{\beta}(\tau)\right) \widehat{\Sigma}(\tau) R\left(\widehat{\beta}(\tau)\right)' \right\}^{-1} r(\widehat{\beta}(\tau)),$$

where $R(\beta) = \partial_\beta r(\beta)$ and $\widehat{\Sigma}(\tau) = \widehat{D}(\tau)\widehat{\Sigma}(\tau)\widehat{D}(\tau)$. Define

$$t(\tau) = \frac{\sqrt{T}\, r(\widehat{\beta}(\tau))}{\left\{ R\left(\widehat{\beta}(\tau)\right) \widehat{\Sigma}(\tau) R\left(\widehat{\beta}(\tau)\right)' \right\}^{1/2}}.$$

Pointwise tests use these statistics one quantile at a time. The paper instead studies hypotheses such as $r\left(\beta_j(\tau)\right) = 0\ \forall \tau \in \mathcal{T}$, $\beta_j(\tau_1) = \cdots = \beta_j(\tau_m)$, and $\beta_j(\tau_1) \leq \beta_j(\tau_2)$.

Fixed-b Limit Theory

Partial-sum empirical process

For $r \in [0,1]$, define

$$S([rT],\tau) = \frac{1}{\sqrt{T}} \sum_{t=1}^{[rT]} z_t(\tau).$$

It is shown that $S([rT],\tau)$ is sequentially Donsker:

$$S([rT],\tau) \Rightarrow G(r,\tau) \quad \text{in } \ell^\infty([0,1]\times\mathcal{T}),$$

with covariance

$$\mathrm{Cov}\{G(r_1,\tau_1), G(r_2,\tau_2)\} = (r_1 \wedge r_2)\Lambda(\tau_1,\tau_2).$$

Why cross-quantile covariance matters

For a single quantile, one can write

$$G(r,\tau) = \Gamma(\tau) W_p(r), \qquad \Gamma(\tau)\Gamma(\tau)' = \Lambda(\tau,\tau).$$

Across quantiles, this representation generally fails because

$$\Lambda(\tau_1,\tau_2) \neq \Gamma(\tau_1)\Gamma(\tau_2)'$$

Therefore, pointwise fixed-b intuition does **not** automatically extend to the quantile process.

Theorem 1: estimated score process

Let

$$\widehat{S}([rT],\tau) = \frac{1}{\sqrt{T}} \sum_{t=1}^{[rT]} \widehat{z}_t(\tau).$$

Theorem 1: under common regularity conditions for QR,

$$\widehat{S}([rT],\tau) = S([rT],\tau) - rS(T,\tau) + o_p(1)$$

uniformly in (r,τ). Hence

$$\widehat{S}([rT],\tau) \Rightarrow \widetilde{G}(r,\tau) = G(r,\tau) - rG(1,\tau).$$

Remark: This result is new regardless of fixed-smoothing. It implies that tests are a functional of the partial sum process would not be asymptotically pivotal for either small-b or fixed-b approximation, echoing Chernozhukov & Fernández-Val (2005).

Theorem 2: fixed-b limit of the HAR estimator

Under the same set of assumptions, Let $M = bT$ for fixed values $b \in (0, 1]$. As $T \to \infty$,

$$\widehat{\Lambda}(\tau) \Rightarrow P\left(\widetilde{G}(r,\tau), b\right) \quad \text{in} \quad l^{\infty}(\mathcal{T}).$$

For a C^2 kernel $\mathcal{K}(.)$,

$$P(\widetilde{G}(r,\tau), b) = -\frac{1}{b^2}\int_0^1 \int_0^1 \mathcal{K}''\left(\frac{r-s}{b}\right) \widetilde{G}(r,\tau)\widetilde{G}(s,\tau)' dr ds.$$

For the Bartlett kernel $\mathcal{K}(.)$,

$$\begin{aligned} P =& \frac{2}{b}\int_0^1 \widetilde{G}(r,\tau)\widetilde{G}(r,\tau)' dr \\ & - \frac{1}{b}\int_0^{1-b} \widetilde{G}(r,\tau)\widetilde{G}(r+b,\tau)' dr - \frac{1}{b}\int_0^{1-b} \widetilde{G}(r+b,\tau)\widetilde{G}(r,\tau)' dr. \end{aligned}$$

Fixed-b limit of the t statistic

For a scalar restriction,

$$t(\tau) \Rightarrow \frac{R_0 D(\tau)^{-1} G(1,\tau)}{\{R_0 D(\tau)^{-1} P(\widetilde{G}, b) D(\tau)^{-1} R_0'\}^{1/2}} =: t^{\infty}(\tau), \text{ in } l^{\infty}(\mathcal{T}).$$

The limit depends on $\Lambda(\tau_1, \tau_2)$ through the full process $G(r, \tau)$. Thus, the fixed-b statistic is generally non-pivotal.

Inference Procedures

Two solutions

- **Direct method:** simulate the Gaussian process directly, with estimated nuisance covariance kernel.
  - Discretizing (the covariance matrix across quantiles) after asymptotic approximation.

- **Stack method:** Stack moments for a finite set of quantiles and exploit the nuisance cancellation.
  - Discretizing before asymptotic approximation.

Direct method: intuition

By (matrix-valued) Mercer's theorem, the covariance kernel can be represented as

$$\Lambda(\tau_1, \tau_2) = \sum_{l=1}^{\infty} \lambda_l e_l(\tau_1) e_l(\tau_2)$$

where $\lambda_l \geq 0$ and $e_l(\tau)$ are the eigenvalues and eigenfunctions.
Under this representation, $G(r, \tau)$ admits the Karhunen–Loéve expansion:

$$G(r, \tau) = \sum_{l=1}^{\infty} \sqrt{\lambda_l} W_l(r) e_l(\tau).$$

We can verify that

$$\mathrm{Cov}\left(G(r_1, \tau_1), G(r_2, \tau_2)\right) = (r_1 \wedge r_2) \sum_{l=1}^{\infty} \lambda_l e_l(\tau_1) e_l(\tau_2)' = (r_1 \wedge r_2) \Lambda(\tau_1, \tau_2).$$

Direct method: implementation

Let $\widehat{\lambda}_l$ and $\widehat{e}_l(\tau)$ be some feasible approximations to the leading eigenvalues and eigenfunctions, then the Gaussian limit process can be approximated by

$$G(r,\tau) \approx \sum_{l=1}^{L} \sqrt{\widehat{\lambda}_l} W_l(r) \widehat{e}_l(\tau).$$

Note that $\Lambda(\tau_1, \tau_2)$ is an infinite dimensional object; and the eigenfunction $e_l(\tau)$ is a vector-valued function defined on a continuum of quantile levels, too.

Direct method: implementation

We propose to discretize the covariance kernel by a finite-dimensional covariance matrix.

For a grid of quantiles $\bar{\tau}_m = (\tau_1, \ldots, \tau_m)$, estimate the block covariance matrix

$$\Lambda(\bar{\tau}_m, \bar{\tau}_m) = [\Lambda(\tau_i, \tau_j)]_{i,j=1}^m .$$

Let $\widehat{\Lambda}(\bar{\tau}_m, \bar{\tau}_m)$ denote the HAC variance matrix estimator with the (i,j)-th block given by:

$$\widehat{\Lambda}(\tau_i, \tau_j) = \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \mathcal{K}\left(\frac{t-s}{M}\right) \widehat{z}_t(\tau_i) \widehat{z}_s(\tau_j)'.$$

Let $\widehat{\Lambda}(\bar{\tau}_m, \bar{\tau}_m) = \widehat{\mathcal{E}}_n \widehat{\Lambda}_n \widehat{\mathcal{E}}_n'$ be the eigenvalue decomposition, where the columns of $\widehat{\mathcal{E}}_n$ contains eigenvectors $(\widehat{e}_l(\tau_1)', \ldots, \widehat{e}_l(\tau_m)')'$; and $\widehat{\Lambda}_n$ is the diagonal matrix of eigenvalues.

Direct method: sup-t/sup-Wald test

Consider $H_0 : r(\beta(\tau)) = 0, \forall \tau \in \mathcal{T}$.

1. Choose a fine grid $\bar{\tau}_m = \{\tau_1, \ldots, \tau_m\}$.
2. Estimate $\widehat{\Lambda}(\tau_i, \tau_j)$ for all grid pairs.
3. Simulate paths from $\widehat{G}(r, \tau)$.
4. Compute simulated $\max_{\tau \in \bar{\tau}_m} |t^{\infty}(\tau)|$ or $\max_{\tau \in \bar{\tau}_m} \mathcal{W}^{\infty}(\tau)$.
5. Compare $\max_{\tau \in \bar{\tau}_m} \mathcal{W}(\tau)$ or $\max_{\tau \in \bar{\tau}_m} |t(\tau)|$ with the corresponding simulated critical values.

Stack method: intuition

For fixed $\bar{\tau}_m = \{\tau_1, \ldots, \tau_m\}$, consider $H_0 : r\left(\beta\left(\bar{\tau}_m\right)\right) = 0$.

Stack the scores:

$$s_t(\bar{\tau}_m) = \begin{pmatrix} x_t\{\tau_1 - 1(y_t \leq x_t'\beta_0(\tau_1))\} \\ \vdots \\ x_t\{\tau_m - 1(y_t \leq x_t'\beta_0(\tau_m))\} \end{pmatrix}.$$

This converts cross-quantile dependence into a finite-dimensional moment problem.

Stack method: intuition

Denote the $mp \times mp$ block diagonal matrix, $D(\bar{\tau}_m)$, as

$$D(\bar{\tau}_m) := \begin{bmatrix} D(\tau_1) & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & D(\tau_m) \end{bmatrix}.$$

By standard fixed-b arguments with $M = bT$, we can obtain a set of results for the given set of quantiles $\bar{\tau}_m$:

$$\begin{aligned} \sqrt{T}\left(\widehat{\beta}(\bar{\tau}_m) - \beta(\bar{\tau}_m)\right) &\Rightarrow D(\bar{\tau}_m)^{-1}\Gamma(\bar{\tau}_m)W_{mp}(1), \\ \widehat{S}_{[rT]}(\bar{\tau}_m) &= \Gamma(\bar{\tau}_m)\widetilde{W}_{mp}(r), \\ \widehat{\Lambda}(\bar{\tau}_m, \bar{\tau}_m) &\Rightarrow \Gamma(\bar{\tau}_m)P(\widetilde{W}_{mp}(r), b)\Gamma(\bar{\tau}_m)'. \end{aligned}$$

Stack method: nuisance cancellation

For a set of quantiles $\bar{\tau}_m$, consider a null hypothesis $r\left(\beta(\bar{\tau}_m)\right) = 0$. Define the t statistic:

$$
\begin{aligned}
t(\bar{\tau}_m) =& \frac{\sqrt{T}\ r(\widehat{\beta}(\bar{\tau}_m))}{\sqrt{R(\widehat{\beta}(\bar{\tau}_m))\ \widehat{D}^{-1}(\bar{\tau}_m)\widehat{\Lambda}(\bar{\tau}_m,\bar{\tau}_m)\widehat{D}^{-1}(\bar{\tau}_m)\ R(\widehat{\beta}(\bar{\tau}_m))'}}. \\
\Rightarrow& \frac{\bar{R}_0 D(\bar{\tau}_m)^{-1}\Gamma(\bar{\tau}_m)W_{mp}(1)}{\sqrt{\bar{R}_0 D(\bar{\tau}_m)^{-1}\Gamma(\bar{\tau}_m)P(\widetilde{W}_{mp}(r),b)\Gamma(\bar{\tau}_m)'D(\bar{\tau}_m)^{-1}\bar{R}_0'}}
\end{aligned}
$$

Because a finite linear combination of Gaussian processes is also Gaussian, $\bar{R}_0 D(\bar{\tau}_m)^{-1}\Gamma(\bar{\tau}_m)W_{mp}(1)$ can be written as $\Upsilon^{1/2}W_q(r)$ where $\Upsilon^{1/2}(\Upsilon^{1/2})' = \Upsilon$. Therefore,

$$
t(\bar{\tau}_m) \Rightarrow \frac{W_q(1)}{\sqrt{P(\widetilde{W}_q(r),b)}},
$$

Direct method v.s. Stack method

- Direct method is designed for questions that involve a **continuum** of quantiles; stack method deals with a **fix set** of quantiles.
- Direct method is also **valid** for a fix set of quantiles, while stack method is **invalid** for a continuum of quantiles.
- When both are feasible, stack method is **nuisance-free** and works better when T is small; when T is large enough, they are expected to behave very **similar**.

Uniform confidence band/Sup-t band

Consider a simultaneous band for one coefficient $\beta_j(\tau)$ across multiple quantiles.

For a given $\alpha \in (0,1)$, define the rectangle

$$I_\alpha(\bar{\tau}_m) = \prod_{i=1}^{m} [\widehat{\beta}_j(\tau_i) - \widehat{\sigma}_j(\tau_i) c_\alpha, \widehat{\beta}_j(\tau_i) + \widehat{\sigma}_j(\tau_i) c_\alpha],$$

where $\widehat{\sigma}_j(\tau) := \left[\widehat{D}(\tau)^{-1} \widehat{\Lambda}(\tau) \widehat{D}(\tau)^{-1} / T\right]_{jj}^{1/2}$.

Question: choose c_α such that asymptotically

$$\Pr\left(\beta_j(\bar{\tau}_m) \in I_\alpha(\bar{\tau}_m)\right) \geq 1 - \alpha.$$

Uniform confidence band/Sup-t band

Using the main results, we show

$$\Pr\left(\beta_j(\bar{\tau}_m) \in I_\alpha(\bar{\tau}_m)\right) \to \Pr\left(\max_{1\leq i\leq m}\left|t_j^\infty(\tau_i)\right| \leq c_\alpha\right).$$

Therefore, choose c_α as the $1-\alpha$ quantile of the simulated distribution of $\max_i |t_j^\infty(\tau_i)|$.

The confidence-band analogue of the Sup-t test.

See Montiel Olea & Mikkel (2019) for additional methods to construct the confidence band.

Shape-restriction tests

The same framework handles hypotheses about the whole curve:

$$\begin{aligned} &\text{Homogeneity:} && \beta_j(\tau_1) = \beta_j(\tau_2) \quad \forall \tau_1, \tau_2, \\ &\text{Monotonicity:} && \beta_j(\tau_1) \leq \beta_j(\tau_2) \quad \forall \tau_1 < \tau_2. \end{aligned}$$

Shape restriction test: Monotonicity

$H_0 : \beta_j(\tau_2) - \beta_j(\tau_1) \geq 0$ for all $\tau_2 > \tau_1 \in \mathcal{T}$.
$H_1 : \beta_j(\tau_2) - \beta_j(\tau_1) < 0$ for some $\tau_2 > \tau_1 \in \mathcal{T}$.

The test statistic: $t^M(\mathcal{T}) = \min_{i>l} \sqrt{T} \left\{ \widehat{\beta}_j(\tau_i) - \widehat{\beta}_j(\tau_l) \right\} / \widehat{\sigma}_j(\tau_i, \tau_l)$.

Under the least favorable scenario of the null, i.e. $\beta_j(\tau)$ is some constant, we show

$$t^M(\mathcal{T}) \to \min_{i>l} \frac{\left\{ e_i' D(\bar{\tau}_m)^{-1} G(1, \bar{\tau}_m) - e_l' D(\bar{\tau}_m)^{-1} G(1, \bar{\tau}_m) \right\}}{\sqrt{e_{i,l}' D(\tau)^{-1} P\left(\widetilde{G}(r, \tau), b \right) D(\tau)^{-1} e_{i,l}}}.$$

where $e_1 = (1, 0, \ldots, 0)'$ and $e_{2,1} = (-1, 1, 0 \ldots, 0)'$

What else in the paper

- Orthonormal-series (OS) HAR estimator and fixed-K asymptotic approximation;
- Test-based data-driven bandwidth rules for both kernel-HAR and OS-HAR methods;
- The stack method for the Sup-t band and its equivalence to the direct approach.
- Local power analysis.

Practical tuning choices

- Kernel-HAR: choose bandwidth $M = [bT]$.
- OS-HAR: choose number of series terms K.
- Direct method: choose grid size m, truncation rank L, Gaussian increments B, simulation replications n.

Simulation Evidence

Simulation design

Location-shift design:

$$
\begin{aligned}
y_t &= \beta_0 + \beta_1 x_t + e_t, \\
e_t &= \rho e_{t-1} + v_t, \\
x_t &= \rho x_{t-1} + \varepsilon_t.
\end{aligned}
$$

- True slope is constant: $\beta_1(\tau) = 1$.
- Serial dependence: $\rho \in \{0, 0.5, 0.8\}$.
- Sample sizes: $T \in \{200, 500, 800\}$.

Simulation question 1: size control by fixed-smoothing approximation

Main comparison: Fixed-smoothing HAR methods versus small-b HAC inference.

- Single-quantile tests reduce to conventional fixed-b methods.
- The limiting distribution reduces to that for a given quantile, which coincides with the conventional fixed-b result.
- For tests that involve a finite number of quantile levels, a fine grid in simulating the Gaussian process does not matter.
- We report results for $\tau \in \{0.50, 0.75, 0.90\}$.
- Overall, fixed-b approach delivers sizable gains, across different parts of the outcome distribution.

Simulation question 2: joint tests across two quantiles

Main comparison: Comparing Wald tests of $H_0 : \beta(\tau_1) = \beta(\tau_2)$ using the direct and stack methods.

- Proposed fixed-smoothing Wald tests have reasonable finite-sample performance.
- We report results for the nulls $\beta(0.5) = \beta(0.75)$, $\beta(0.5) = \beta(0.90)$, $\beta(0.10) = \beta(0.90)$.
- Direct and stack kernel methods perform similarly in the two-quantile design.

Simulation question 3: Sup-t/Wald, Sup-t band and shape tests

Main comparison: Comparing the performance of tests the direct method across different DGPs and the tuning parameter m (quantile grid).

- Kernel-HAR appears to be the more stable across bandwidths in finite samples.
- Results are relatively insensitive to the quantile grid for a sufficiently large sample.
- The Sup-t band test shows better size control than a joint Wald test (e.g., $H_0 : \beta(\tau_1) = \ldots = \beta(\tau_m) = 0$).
- Homogeneity and monotonicity tests follow similar patterns.

Simulation: takeaway

- Comparing the direct and stack methods:
  - When the number of quantile levels is large relative to the sample size: stack method;
  - When the sample size is sufficiently large: direct method;
  - For hypotheses defined over a continuum of quantile levels: direct method is the only choice.
- Comparing the Sup-t and Wald test:
  - Sup-t test as multiple testing: better size control than Wald joint test.
- Comparing the Bartlett-HAR and OS-HAR estimators:
  - Bartlett-HAR is more stable across bandwidths for both size and power;
  - OS-HAR may have better size control when bandwidth is well-tuned.

Empirical Application

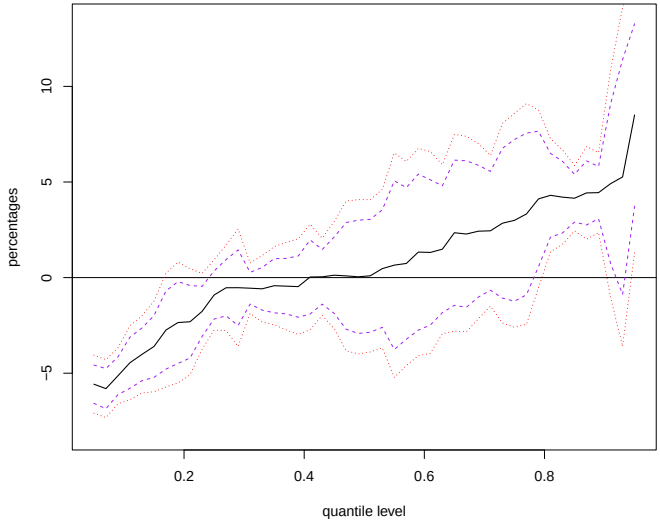
Application: predictive quantile regression

Monthly S&P 500 excess returns are modeled as

$$y_t = \alpha(\tau) + \beta_1(\tau)x_{t-1} + e_t(\tau).$$

- Updated version of the monthly data from Welch & Goyal (2008): January 1926 to December 2024.
- Predicted variable: stock return − short-term U.S. Treasury-bill rate.
- Predictors (avoiding strong persistence): stock variance, inflation, long-term returns, net equity expansion.
- Main focus: whether predictability differs across quantiles.

Stock variance: uniform band



Stock variance: three tests using the direct fixed-b method

For $\tau_1 = 0.5$ and $\tau_2 = 0.9$:

Hypothesis	Statistic	95% critical value
$\beta(\tau_1) = \beta(\tau_2) = 0$	2.959	2.863
$\beta(\tau_1) = \beta(\tau_2)$	5.236	2.492
$\beta(\tau_1) \leq \beta(\tau_2)$	5.236	-1.845

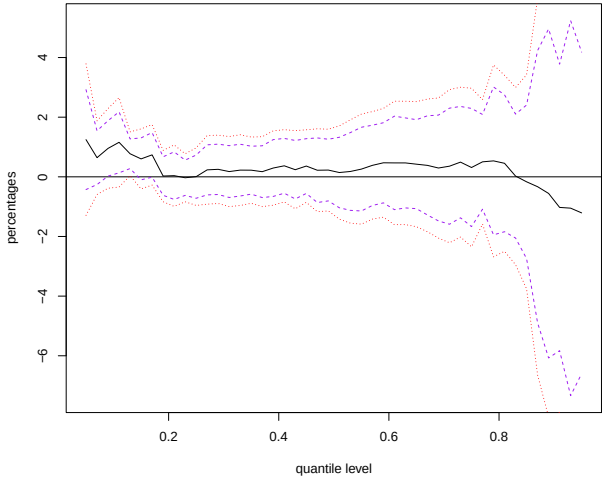
The effect is nonzero at some quantile and larger in the right tail than at the center.

Stock variance: global tests over $\mathcal{T} = [0.05, 0.95]$

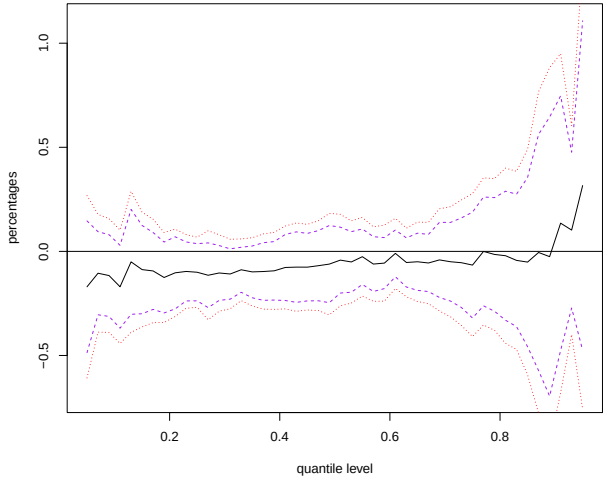
Global hypothesis	Statistic	95% critical value
No effect	9.059	3.719
Homogeneity	6.476	4.902
Monotonicity	-1.038	-4.245

Evidence supports a nonzero and heterogeneous effect, with no rejection of monotone increase.

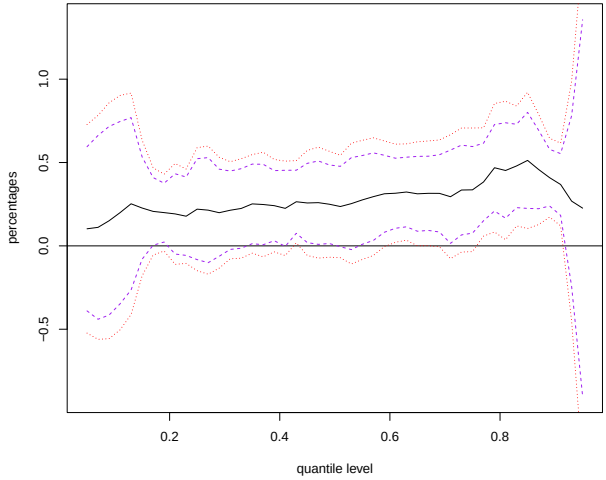
Other predictors: inflation rate



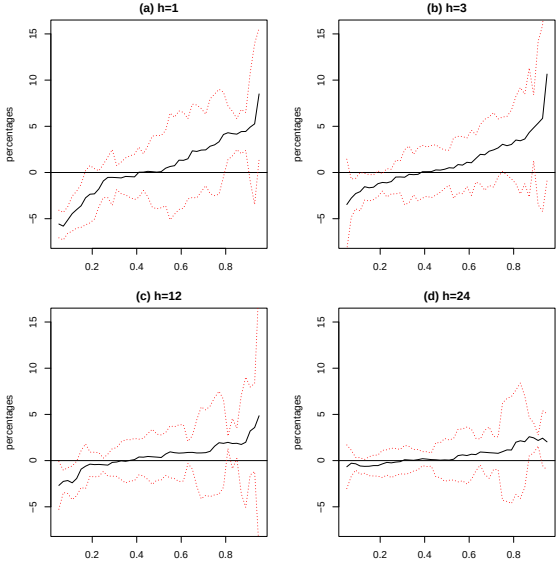
Other predictors: long-term return



Other predictors: net equity expansion



Forecast horizons



Conclusion and Takeaways

Empirical Recommendations

Empirical goal	Recommended method
Uniform confidence band/Multiple testing	Direct method
Continuum Sup-t test	Direct/Stack method
Shape restriction	Direct method
Joint test for a finite set of quantiles, small T	Stack method
Joint test for a finite set of quantiles, large T	Direct/Stack method

Conclusion

- Quantile methods is commonly used in times series settings.
- However, with temporal dependence, existing inference methods are invalid for questions that involves multiple/a continuum of quantiles.
- We derive fixed-smoothing asymptotic approximation of the quantile estimator, kernel/OS variance estimators, and test statistics, all uniform in (r, τ).
- We propose fixed-b and fixed-K inference methods based on the results.
- Two practical implementations are proposed: direct and stack-moment methods.
- We propose several tests using our methods, and illustrate their performances in simulations and a study of stock return prediction.

Backup Slides

Regularity conditions for Theorem 1

- The Bahadur representation;
- finite 4th moment of x;
- strict stationarity and beta-mixing with exponential rate of (y, x');
- smooth, (upper and lower) bounded conditional density function;
- bounded joint conditional density function;
- bounded derivatives of the conditional density function;
- uniform WLLN for the Gram matrix;
- uniform WLLN for the 3rd-order cross-moment of x.

Back

Simulation result 1: size control by fixed-smoothing approximation

Table: Inference for the single quantile

Setting		fixed-b			fixed-K			small-b	
		kernel-HAR			OS-HAR			kernel-HAC	
T	ρ	M	10%	5%	K	10%	5%	10%	5%
$\tau = 0.5$									
200	0.0	1	0.093	0.055	200	0.108	0.046	0.103	0.051
200	0.5	7	0.123	0.061	49	0.132	0.079	0.131	0.073
200	0.8	25	0.139	0.080	20	0.182	0.107	0.191	0.137
500	0.0	1	0.114	0.060	500	0.083	0.043	0.103	0.059
500	0.5	15	0.089	0.043	84	0.110	0.062	0.103	0.053
500	0.8	54	0.115	0.064	33	0.123	0.069	0.140	0.083
800	0.0	1	0.092	0.044	800	0.105	0.063	0.090	0.046
800	0.5	15	0.103	0.054	134	0.101	0.050	0.118	0.073
800	0.8	54	0.118	0.058	51	0.144	0.089	0.144	0.080

Back

Simulation result 1: size control by fixed-smoothing approximation

Table: Inference for the single quantile

Setting		fixed-b			fixed-K			small-b	
		kernel-HAR			OS-HAR			kernel-HAC	
T	ρ	M	10%	5%	K	10%	5%	10%	5%
$\tau = 0.75$									
200	0.0	1	0.101	0.046	200	0.106	0.052	0.093	0.055
200	0.5	6	0.124	0.064	51	0.143	0.090	0.139	0.072
200	0.8	24	0.161	0.094	20	0.151	0.087	0.209	0.147
500	0.0	1	0.095	0.062	500	0.105	0.055	0.086	0.044
500	0.5	14	0.098	0.050	89	0.125	0.080	0.120	0.063
500	0.8	50	0.125	0.074	35	0.138	0.085	0.167	0.106
800	0.0	1	0.097	0.050	800	0.101	0.054	0.101	0.044
800	0.5	14	0.095	0.048	141	0.114	0.059	0.109	0.061
800	0.8	51	0.121	0.064	54	0.139	0.078	0.146	0.086

Back

Simulation result 1: size control by fixed-smoothing approximation

Table: Inference for the single quantile

Setting		fixed-b			fixed-K			small-b	
		kernel-HAR			OS-HAR			kernel-HAC	
T	ρ	M	10%	5%	K	10%	5%	10%	5%
$\tau = 0.9$									
200	0.0	1	0.124	0.078	200	0.108	0.046	0.115	0.065
200	0.5	5	0.139	0.091	49	0.132	0.079	0.136	0.087
200	0.8	21	0.198	0.136	20	0.182	0.107	0.237	0.168
500	0.0	1	0.114	0.050	500	0.083	0.043	0.115	0.066
500	0.5	11	0.124	0.076	84	0.110	0.062	0.109	0.063
500	0.8	44	0.158	0.090	33	0.123	0.069	0.190	0.116
800	0.0	1	0.122	0.066	800	0.105	0.063	0.104	0.058
800	0.5	11	0.115	0.057	134	0.101	0.050	0.125	0.069
800	0.8	44	0.133	0.074	51	0.144	0.089	0.182	0.113

Back

Simulation result 2: Inference for two quantile effects

Table: Inference for the slope coefficient at two quantile levels

Setting			Direct			Stacking		Direct		
			kernel-HAR			kernel-HAR		OS-HAR		
(τ_1, τ_2)	T	ρ	M	10%	5%	10%	5%	K	10%	5%
(0.5, 0.75)	200	0.0	1	0.096	0.051	0.099	0.048	200	0.109	0.058
	200	0.5	7	0.119	0.076	0.109	0.059	48	0.139	0.079
	200	0.8	28	0.141	0.083	0.156	0.093	17	0.160	0.093
	500	0.0	1	0.095	0.042	0.092	0.060	500	0.111	0.068
	500	0.5	14	0.126	0.065	0.113	0.051	85	0.135	0.075
	500	0.8	57	0.107	0.057	0.118	0.063	30	0.147	0.083
	800	0.0	1	0.119	0.061	0.087	0.048	800	0.108	0.058
	800	0.5	14	0.104	0.047	0.113	0.063	136	0.118	0.059
	800	0.8	57	0.114	0.054	0.119	0.062	47	0.143	0.087

Back

Simulation result 2: Inference for two quantile effects

Table: Inference for the slope coefficient at two quantile levels

Setting			Direct			Stacking		Direct		
			kernel-HAR			kernel-HAR		OS-HAR		
(τ_1, τ_2)	T	ρ	M	10%	5%	10%	5%	K	10%	5%
(0.5, 0.9)	200	0.0	1	0.134	0.084	0.108	0.055	200	0.117	0.075
	200	0.5	7	0.132	0.080	0.133	0.075	46	0.136	0.070
	200	0.8	29	0.158	0.097	0.193	0.128	17	0.180	0.123
	500	0.0	1	0.126	0.070	0.125	0.066	500	0.137	0.073
	500	0.5	15	0.115	0.060	0.112	0.053	83	0.153	0.088
	500	0.8	62	0.131	0.074	0.130	0.078	30	0.158	0.096
	800	0.0	1	0.120	0.062	0.108	0.056	800	0.112	0.061
	800	0.5	15	0.113	0.053	0.130	0.067	133	0.129	0.070
	800	0.8	61	0.128	0.074	0.135	0.065	47	0.146	0.098

Back

Simulation result 2: Inference for two quantile effects

Table: Inference for the slope coefficient at two quantile levels

Setting			Direct			Stacking		Direct		
			kernel-HAR			kernel-HAR		OS-HAR		
(τ_1, τ_2)	T	ρ	M	10%	5%	10%	5%	K	10%	5%
$(0.1, 0.9)$	200	0.0	1	0.138	0.078	0.131	0.081	200	0.134	0.092
	200	0.5	7	0.162	0.098	0.152	0.097	49	0.177	0.109
	200	0.8	27	0.218	0.134	0.213	0.139	20	0.266	0.188
	500	0.0	1	0.135	0.074	0.123	0.060	500	0.135	0.084
	500	0.5	14	0.135	0.082	0.131	0.075	87	0.167	0.095
	500	0.8	57	0.177	0.113	0.180	0.116	34	0.196	0.128
	800	0.0	1	0.121	0.065	0.133	0.085	800	0.134	0.081
	800	0.5	14	0.120	0.058	0.134	0.082	140	0.117	0.060
	800	0.8	58	0.140	0.072	0.163	0.089	53	0.194	0.126

Back