IDEAL Inference on Conditional Quantiles via Interpolated Duals of Exact Analytic L-statistics

David M. Kaplan[*]
Department of Economics, UC San Diego

December 26, 2012; first version May 30, 2012

Abstract

We examine inference on conditional quantiles from the nonparametric perspective of local smoothing. This paper develops a framework for translating the powerful, high-order accurate IDEAL results (Goldman and Kaplan, 2012) from their original unconditional context into a conditional context, via a uniform kernel. Under mild smoothness assumptions, our new conditional IDEAL method's two-sided pointwise coverage probability error is $O(n^{-2/(2+d)})$, where d is the dimension of the conditioning vector and n is the total sample size. For $d \leq 2$, this is better than the conventional inference based on asymptotic normality or a standard bootstrap. It is also better for other d depending on smoothness assumptions. For example, conditional IDEAL is more accurate for $d = 3$ unless 11 or more derivatives of the unknown function exist and a corresponding local polynomial of degree 11 is used (which has 364 terms since interactions are required). Even as $d \to \infty$, conditional IDEAL is more accurate unless the number of derivatives is at least four, and the number of terms in the corresponding local polynomial goes to infinity as $d \to \infty$. The tradeoff between the effective (local) sample size and bias determines the optimal bandwidth rate, and we propose a feasible plug-in bandwidth. Simulations show that IDEAL is more accurate than popular current methods, significantly reducing size distortion in some cases while substantially increasing power (while still controlling size) in others. Computationally, our new method runs much more quickly than existing methods for medium and large datasets (roughly $n \geq 1000$). We also examine health outcomes in Indonesia for an empirical example.

1 Introduction

Ideally, we would like to know the full joint distribution of every variable we care about. Practically, with a finite amount of data, we can learn a lot from estimating quantiles of

[*]Email: dkaplan@ucsd.edu. Thanks to Andres Santos, Yixiao Sun, and Patrik Guggenberger for helpful comments. Thanks to Brendan Beare, Graham Elliott, James Hamilton, and other UCSD seminar participants for additional questions and suggestions.

conditional distributions. To gain knowledge from the data, rather than simply compute numbers, we need statistical inference on these conditional quantiles, which is this paper's concern.

In economics, conditional quantiles have appeared across diverse topics because they are such fundamental statistical objects. Conditional quantile studies of wages have looked at experience (Hogg, 1975), union membership (Chamberlain, 1994), and inequality in the U.S. wage structure (Angrist et al., 2006; Buchinsky, 1994), while Kordas (2006) examines married women's propensity to work. Examples from health economics include models of infant birthweight (Abrevaya, 2001) and demand for alcohol (Manning et al., 1995). Among others, Chesher (2003) employs conditional quantiles for identification in nonseparable models. Guerre and Sabbah (2012) give an example of estimating conditional quantiles of private values from bid data in a first-price sealed bids auction. Other empirical examples involving conditional quantiles include school quality and student outcomes (Eide and Showalter, 1998), spatial land valuations (Koenker and Mizera, 2004), welfare analysis (Belluzzo, 2004), and Engel curves (Nayyar, 2009). For similar reasons as in economics, conditional quantiles enrich empirical work in other areas, such as modeling temperature (Hyndman et al., 1996), limiting factors in ecology (Cade et al., 1999), and terrestrial mammal running speeds (Koenker, 2005), among other examples. Additionally, almost any study of conditional means could be extended to conditional quantiles to expose additional heterogeneity.

If our continuous dependent variable of interest is Y (e.g., high school GPA), and we want to know quantiles of its distribution conditional on vector X having value x_0 (e.g., a particular value of family income and other socioeconomic and demographic characteristics), we would like an infinite number of observations with $X = x_0$. With our finite sample, if X contains even one continuous component, we have zero probability of even one observation with $X = x_0$. One approach is to parameterize the conditional p-quantile function as $Q_{Y|X}(p) = X'\beta$, linear in X. This strong linearity assumption leads to an estimator where observations with very different X can influence the conditional quantile at $X = x_0$. If the true function is not linear in X, this misspecification can lead to a poor estimator for any given x_0. A second approach is to use a more flexible parameterization, which could be set in a sieve-type nonparametric framework. A third approach is nonparametric local kernel smoothing, using only observations with X close in value to the target x_0. Here, we develop inference via kernel smoothing.

Our strategy is to apply an accurate method for unconditional quantile inference to the observations with X close to the target x_0. Instead of relying on an asymptotic normal approximation, this method directly approximates the exact finite sample distribution using fractional order statistic theory. As usual for kernel smoothing, taking X not quite equal to

x_0 causes some bias, which increases as we include X values farther from x_0 to include more observations in our effective sample. Counter to this, the method's accuracy (for the biased value) improves as the effective sample size grows. This tradeoff determines the optimal bandwidth that minimizes overall coverage probability error. After more precisely deriving the objects involved, and estimating the unknown ones, we propose a bandwidth for use in practice. Specifically, we apply some of the IDEAL (interpolated dual of exact analytic L-statistic; see §2) results from Goldman and Kaplan (2012) on the Hutson (1999) method based on fractional order statistics. Joint confidence sets are constructed as Cartesian products of (Bonferroni-adjusted) confidence intervals over many x_0.

This strategy has advantages in theory and in practice. Theoretically, the coverage probability error is of a smaller order of magnitude than that for inference based on asymptotic normality or bootstrap in the most common cases. This reflects the same advantage of the fractional order statistic method for unconditional quantile inference. The only theoretical limitation is our implicit use of a uniform kernel, which is only a second-order kernel. This prevents reducing the coverage probability error by assuming higher degrees of smoothness than we do here, though to maintain robustness we would not want to assume more smoothness of unknown functions anyway. As it is, our method has a better rate than asymptotic normality even with infinite smoothness assumed if the conditioning vector X contains one or two continuous components, and similarly for any number of continuous conditioning variables in X if no more than four derivatives of the unknown function are (correctly) assumed to exist. It is also an advantage that our coverage error either leads to over-coverage or can be set to zero (at the dominant order of magnitude) by our choice of bandwidth since we derive the value (not just rate) of the optimal bandwidth when X contains only one continuous conditioning variable.

Practically, direct estimation of a plug-in version of the optimal bandwidth is straightforward, with details in §4.2 and §5.1 and code available in R for X with a single continuous component. An extension to additional continuous conditioning variables will require additional calculations and estimation, but the approach will be identical. In either case, there is no asymptotic variance or other nuisance parameters to estimate. Also, coverage probability is monotonically decreasing in the size of the bandwidth, which is nice for transparency and may be helpful for future refinements. Another practical advantage is computation times orders of magnitude smaller than those for existing methods on medium and large datasets (roughly $n \geq 1000$); see §5 for details.

Past research has focused instead on inference through asymptotic normal approximation or bootstrap. First-order accuracy has been shown for the asymptotic normality approaches in Bhattacharya and Gangopadhyay (1990, Thm. N2) and Hall et al. (1999, Thm. 1, eqns.

7 and 8) and for the bootstrap in Gangopadhyay and Sen (1990). Higher-order accuracy can be shown for inference using the asymptotic normality in Chaudhuri (1991); improved recently by Portnoy (2012), this result is compared to ours in detail following Theorem 3.

The remainder of this paper is organized as follows. Prior unconditional IDEAL work and intuition is given in Section 2. Section 3 details our setup and gives results for the bias. Section 4 describes the optimal bandwidth and its feasible plug-in counterpart. Section 5 contains a simulation study, while Section 6 contains empirical applications. Notationally, $\doteq$ should be read as "is equal to, up to smaller-order terms"; $\asymp$ as "has exact (asymptotic) rate/order of" (same as "big theta" Bachmann–Landau notation, $\Theta(\cdot)$); and $A_n = O(B_n)$ as usual, $\exists k < \infty$ s.t. $|A_n| \leq B_n k$ for sufficiently large n. Acronyms used are those for cumulative distribution function (CDF), confidence interval (CI), coverage probability (CP), coverage probability error (CPE), interpolated duals of exact analytic L-statistics (IDEAL), and probability density function (PDF). Proofs are reserved for the appendix.

2 Fractional order statistic theory

Fractional order statistic theory is key to the method developed in this paper. This section contains an overview of the relevant theory. For a more comprehensive and general development of this theory, see Goldman and Kaplan (2012), who also provide details of additional IDEAL (Interpolated Duals of Exact Analytic L-statistics) methods of quantile inference built upon the theory.

The approach is to construct confidence intervals (CIs) from order statistics observed in the sample. Consider constructing a two-sided CI for the p-quantile. By definition, the p-quantile is between the u_ℓ-quantile and the u_h-quantile when $u_\ell < p < u_h$. One way to construct a CI is to use the empirical u_ℓ- and u_h-quantiles as endpoints. If $(n+1)u_\ell$ and $(n+1)u_h$ are integers, then both endpoints are order statistics, and we can calculate the exact, finite-sample coverage probability because the joint distribution of order statistics is known. The only obstacle is that there almost surely (i.e., with probability one) does not exist a pair of order statistics that yields the exact coverage we desire. Thus, either randomization or interpolation between order statistics is needed; we pursue the latter approach.

In the unconditional case, there is an iid sample $\{Y_i\}_{i=1}^n$ of draws of an absolutely continuous, scalar random variable Y with unknown cumulative distribution function (CDF) denoted $F_Y(\cdot)$. By definition, the quantile function is the inverse CDF, so we also write $Q_Y(\cdot) \equiv F_Y^{-1}(\cdot)$. The kth order statistic $Y_{n:k}$ denotes the kth smallest value out of the n observations $\{Y_i\}_{i=1}^n$. Since $Y < Q_Y(p)$ is equivalent to $F_Y(Y) < p$, we can work with the uniformly distributed $U_i \equiv F_Y(Y_i)$ and the corresponding order statistics. For any $u \in (0,1)$

such that $(n+1)u$ is an integer, it is well known (and derived via combinatorics) that the uniform order statistics follow a beta distribution,

$$U_{n:(n+1)u} \sim \beta\big((n+1)u, (n+1)(1-u)\big).$$

These $U_{n:(n+1)u}$ are estimators on $[0,1]$ of true quantiles $Q_U(u) = u$, where $Q_U(\cdot)$ is the quantile function of the uniformly distributed U_i. As such, they may also be written as $\hat{Q}^I_U(u)$, where the 'I' superscript is for 'ideal' since the distribution is known exactly. This can be generalized beyond integer $(n+1)u$ to any $u \in (0,1)$, and the order statistics generalize to corresponding 'ideal' fractional order 'statistics'[1]

$$\hat{Q}^I_U(u) = U_{n:(n+1)u} \sim \beta\big((n+1)u, (n+1)(1-u)\big), \tag{1}$$

the same distribution as before.

For a two-sided equal-tailed CI for the median, we can solve for $u = u_h$ for the upper (high) endpoint with $P\Big(\hat{Q}^I_U(u) < 1/2\Big) = \alpha/2$ since we know the exact distribution of $\hat{Q}^I_U(u)$ for all u. We may solve for $u = u_\ell$ for the lower endpoint similarly. Since $P\Big(\hat{Q}^I_U(u) < 1/2\Big) = P\Big(Q_Y\Big(\hat{Q}^I_U(u)\Big) < Q_Y(1/2)\Big)$, an exact $(1-\alpha)$ CI for the median $Q_Y(1/2)$ is defined by the unobserved, *fractional* order statistic endpoints $\hat{Q}^I_Y(u_h)$ and $\hat{Q}^I_Y(u_\ell)$, where

$$\hat{Q}^I_Y(u) \equiv Q_Y\Big(\hat{Q}^I_U(u)\Big).$$

For quantiles besides the median, the endpoint indices u_ℓ and u_h are implicit functions of p and α, strictly monotonic in both p and α, determined by

$$\alpha/2 = P\Big(\hat{Q}^I_U(u_h) < p\Big), \quad \alpha/2 = P\Big(\hat{Q}^I_U(u_\ell) > p\Big), \tag{2}$$

with $\hat{Q}^I_U(u) \sim \beta\big((n+1)u, (n+1)(1-u)\big)$ as in (1). For one-sided CIs, only one of the two equalities in (2) is used, and with α instead of $\alpha/2$. Figure 1 is an example.

The unobserved, 'ideal' fractional order statistic endpoints of the exact CI can be approximated by linear (superscript 'L') interpolation between consecutive observed order statistics. For example, if $n = 8$ and $u = 1/2$, to approximate the 4.5th order statistic, we average the 4th and 5th: $\hat{Q}^L_Y(1/2) = (1/2)Y_{8:5} + (1/2)Y_{8:4}$. For any $u \in (0,1)$,

$$\hat{Q}^L_Y(u) \equiv (1-\epsilon)Y_{n:\lfloor(n+1)u\rfloor} + \epsilon Y_{n:\lfloor(n+1)u\rfloor+1}, \tag{3}$$

[1] Technically, when $(n+1)u$ is not an integer, $U_{n:(n+1)u}$ is not a statistic since it is not a function of the observed sample. Instead, $U_{n:(n+1)u}$ is a theoretical construct when it is not equal to an observed order statistic. Nonetheless, we follow the convention in the literature and call $U_{n:(n+1)u}$ a fractional order statistic for all u.

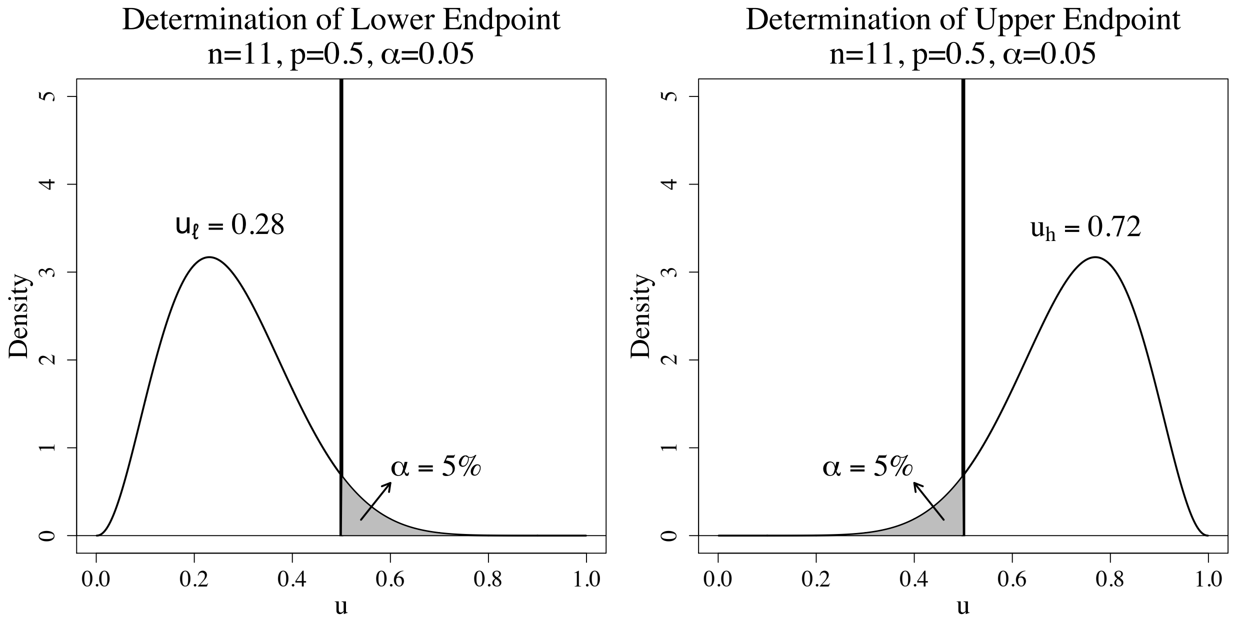

Figure 1: Example of selection of u_ℓ and u_h using (2), for one-sided CI endpoints. For two-sided, $\alpha/2$ would be in place of α. Note that for $p = 0.5$, $u_h = 1 - u_\ell$.

where $\epsilon \equiv (n+1)u - \lfloor (n+1)u \rfloor$ is the interpolation weight, with $\lfloor \cdot \rfloor$ denoting the floor function. The construction of a CI with endpoints $\hat{Q}^L_Y(u_h)$ and $\hat{Q}^L_Y(u_\ell)$ is first given in Hutson (1999). Theoretical justification in terms of coverage probability is first given in Goldman and Kaplan (2012), where the linear interpolation is shown to induce only $O(n^{-1})$ coverage probability error (CPE), as reproduced in the following lemma.

Lemma 1. *For quantile of interest $p \in (0,1)$ and iid sample $\{Y_i\}_{i=1}^n$, consider the two-sided CI constructed as $\Big(\hat{Q}^L_Y(u_\ell), \hat{Q}^L_Y(u_h)\Big)$ with $\hat{Q}^L_Y(u)$ as defined in* (3)*, where u_ℓ and u_h are defined by* (2)*. Alternatively, consider a lower or upper one-sided CI, with u_h or u_ℓ determined by the relevant equality in* (2) *but with α in place of $\alpha/2$.*

Assume that $F'_Y(Q_Y(p)) > 0$ and that the probability density function $F'_Y(\cdot)$ is twice continuously differentiable in a neighborhood of $Q_Y(p)$. Then the coverage probability of the lower one-sided CI is

$$P\Big(\hat{Q}^L_Y(u_h) > Q_Y(p)\Big) = 1 - \alpha + \frac{\epsilon_h(1-\epsilon_h) z_{1-\alpha} \exp\{-z^2_{1-\alpha}/2\}}{p(1-p)\sqrt{2\pi}} n^{-1} + O\big(n^{\delta - 3/2}\big)$$

for arbitrarily small $\delta > 0$, $\epsilon_h \equiv (n+1)u_h - \lfloor (n+1)u_h \rfloor$ similar to before, and $z_{1-\alpha}$ the $(1-\alpha)$-quantile of a standard normal distribution. The coverage probability of the upper one-sided CI is the same but with ϵ_ℓ instead of ϵ_h. Similarly, the coverage probability of the *two-sided CI is*

$$P\Big(\hat{Q}_Y^L(u_\ell) < Q_Y(p) < \hat{Q}_Y^L(u_h)\Big)$$
$$= 1 - \alpha + \frac{z_{1-\alpha}\exp\{-z_{1-\alpha}^2/2\}}{p(1-p)\sqrt{2\pi}}[\epsilon_h(1-\epsilon_h) + \epsilon_\ell(1-\epsilon_\ell)]n^{-1} + O\big(n^{\delta-3/2}\big).$$

Beyond the asymptotic rate of CPE, this method also has the advantage of approximating the exact finite-sample distribution directly rather than relying on asymptotic normality. To determine the optimal bandwidth in §4, Lemma 1 will be used along with results for CPE from the bias in §3.

The asymptotic (first-order) power of the IDEAL hypothesis test is equal to that of a test based on the asymptotic normality of the p-quantile estimator, against alternatives of order $n^{-1/2}$. The upper endpoint of a one-sided $(1-\alpha)$ CI is $\hat{Q}_Y^L(u_h)$ for IDEAL and can be written as $\hat{Q}_Y^L(p) + n^{-1/2}z_{1-\alpha}\sqrt{p(1-p)}/\hat{f}_Y\Big(\hat{Q}_Y^L(p)\Big)$ for normality. The quantile spacing estimator

$$\frac{1}{h}\Big(\hat{Q}_Y^L(p+h) - \hat{Q}_Y^L(p)\Big),$$

similar to Siddiqui (1960) and Bloch and Gastwirth (1968), consistently estimates the $1/f_Y$ object as long as $h \to 0$ and $nh \to \infty$. Lemma 3 in Goldman and Kaplan (2012) gives $u_h - p = n^{-1/2}z_{1-\alpha}\sqrt{p(1-p)} + O(n^{-1})$, so bandwidth $h = u_h - p$ satisfies the rate conditions for consistency. Then, to show the equivalence of the IDEAL endpoint with the normality endpoint,

$$\begin{aligned}
\hat{Q}_Y^L(u_h) &= \hat{Q}_Y^L(p) + \Big[\hat{Q}_Y^L(p+u_h-p) - \hat{Q}_Y^L(p)\Big] \\
&= \hat{Q}_Y^L(p) + \frac{u_h-p}{u_h-p}\Big[\hat{Q}_Y^L(p+u_h-p) - \hat{Q}_Y^L(p)\Big] \\
&= \hat{Q}_Y^L(p) + n^{-1/2}z_{1-\alpha}\sqrt{p(1-p)}\frac{1}{u_h-p}\Big[\hat{Q}_Y^L(p+u_h-p) - \hat{Q}_Y^L(p)\Big] + O\big(n^{-1}\big) \\
&= \hat{Q}_Y^L(p) + n^{-1/2}z_{1-\alpha}\sqrt{p(1-p)}/\hat{f}_Y\Big(\hat{Q}_Y^L(p)\Big) + O(n^{-1}).
\end{aligned}$$

Asymptotically, against alternatives of order $n^{-1/2}$, the $O(n^{-1})$ difference in the final line above is negligible, so the endpoints and thus power are the same. This shows that IDEAL attains better coverage probability without sacrificing optimal asymptotic power. It is reasonable to expect that this is also true of IDEAL methods in other settings.

Additional unconditional IDEAL methods are developed in Goldman and Kaplan (2012). Beyond inference on a single quantile, joint inference on multiple quantiles is also possible, with $O(n^{-1})$ CPE. IDEAL inference on linear combinations of quantiles, such as the interquartile range, has $O(n^{-2/3})$ CPE. In a two-sample setup (e.g., treatment and control groups), IDEAL inference on quantile treatment effects has $O(n^{-2/3})$ CPE. For the IDEAL quantile treatment effect inference, in addition to an empirical application revealing interesting heterogeneity in the "gift exchange" examined in Gneezy and List (2006), many simulation results are given in Goldman and Kaplan (2012). Specific to the median, and switching to the language of hypothesis testing, permutation-type tests (including the Mann–Whitney–Wilcoxon test) are often used because of their good power against pure location shifts, but they can be severely size distorted when exchangeability is violated. In contrast, IDEAL controls size in these cases, and it also has better power against certain alternatives that are not pure location shifts. IDEAL's robust size control extends to all quantiles, and power is consistently better than that of other methods for quantile treatment effects, including various bootstrap methods and the approach in Kaplan (2011). This strong performance of IDEAL in unconditional settings motivates this paper's development of a framework to extend these methods to a nonparametric conditional setting, where even fewer alternative methods exist.

3 Setup

Let $Q_{Y|X}(u; x)$ be the conditional quantile function of some scalar outcome Y given conditioning vector $X \in \mathcal{X} \subset \mathbb{R}^d$, evaluated at $X = x$ and quantile $u \in (0, 1)$. A sample of iid data $\{Y_i, X_i\}_{i=1}^n$ is drawn. If the conditional cumulative distribution function (CDF) $F_{Y|X}(\cdot)$ is strictly increasing and continuous at u, then $F_{Y|X}(Q_{Y|X}(u; x); x) = u$. For some value $X = x_0$ and some quantile $p \in (0, 1)$, interest is in inference on $Q_{Y|X}(p; x_0)$. Without loss of generality, let $x_0 = 0$, which can always be achieved in practice by translation of the X_i.

If X is discrete, we can take all the observations with $X_i = 0$ and then compute a confidence interval (CI) for the p-quantile of the corresponding Y_i values. The one-sample quantile inference method of Hutson (1999) gives CIs with coverage probability error (CPE) of order $O(n^{-1})$, as proved in Goldman and Kaplan (2012). The term of order n^{-1} is explicitly derived by Goldman and Kaplan (2012), too, which makes possible analytic calibration to reduce CPE to $O(n^{-3/2+\delta})$ for arbitrarily small $\delta > 0$. In the conditional setting, however, it is more practical not to calibrate, since the n^{-1} term always leads to over-coverage (as previously shown) and will be used to calculate an optimal bandwidth in §4. If there are N_n observations with $X_i = 0$, then the CPE is $O(N_n^{-1})$ (uncalibrated), where N_n is the effective sample size. For joint CIs, since the pointwise CIs are all independent (using different X), we can generate $(1-\alpha)$ joint coverage by pointwise $(1-\alpha)^{1/m}$ CIs, where m is the number of values in the support of X. Since X is discrete, these are also uniform confidence "bands." The key to the one-sample results going through (with N_n) is that we have N_n iid draws of

Y_i from the same $Q_{Y|X}(\cdot;0)$ conditional quantile function, which is the one of interest.

As always, if X is continuous (or N_n too small), we need to add observations with X_i near zero. Specifically, we include any observations with $X_i \in C_h$, where C_h is the interval $[-h,h]$ for some bandwidth h when $d=1$. For $d>1$, C_h is a hypersphere or hypercube with radius (or half side length) h, centered at the origin. Important objects are defined here for reference, as well as our concept of smoothness that, like our use of C_h, follows that of Chaudhuri (1991, pp. 762–3). Recall that $x_0 = 0$ is assumed without loss of generality.

Definition 1 (effective sample)**.** Let h denote the bandwidth and $p \in (0,1)$ the quantile of interest. The effective sample consists of Y_i values from observations with X_i inside some window $C_h \subset \mathbb{R}^d$ centered at the origin, and the effective sample size is N_n:

$$C_h \equiv \{x : x \in \mathbb{R}^d, \|x\| \le h\}, \tag{4}$$

$$N_n \equiv \#\big(\{Y_i : X_i \in C_h, 1 \le i \le n\}\big), \tag{5}$$

where $\|\cdot\|$ denotes any norm on $\mathbb{R}^d$. With $X = \big(X_{(1)},\ldots,X_{(d)}\big) \in \mathbb{R}^d$, C_h becomes a d-dimensional hypersphere if $\|X\| = \|X\|_2 \equiv \sqrt{X_{(1)}^2 + \cdots + X_{(d)}^2}$, the Euclidean norm ($L_2$ norm), and C_h becomes a d-dimensional hypercube if $\|X\| = \|X\|_\infty \equiv \max\{|X_{(1)}|,\ldots,|X_{(d)}|\}$, the max-norm ($L_\infty$ norm). Additionally, the p-quantile of Y when X is restricted to C_h is denoted $Q_{Y|C_h}(p)$, which satisfies

$$p = P\big(Y < Q_{Y|C_h}(p) \mid X \in C_h\big).$$

Definition 2 (local smoothness, differentiation)**.** For d-dimensional vector $v = (v_{(1)},\ldots,v_{(d)})$ of nonnegative integers, let D^v denote the differential operator $\partial^{\|v\|_1}/[\partial x_{(1)}^{v_{(1)}} \cdots \partial x_{(d)}^{v_{(d)}}]$, where $\|v\|_1 = v_{(1)} + \cdots + v_{(d)}$ is the L_1 norm. A function $h(x)$ is said to have local smoothness of degree $s = k+\gamma$, where k is a nonnegative integer and $\gamma \in (0,1]$, if $h(x)$ is continuously differentiable through order k in a neighborhood of the origin and has uniformly Hölder continuous kth derivatives at the origin, with exponent γ. More precisely, this means that there exists a positive, real constant c such that in some neighborhood of the origin

(i) $D^v h(x)$ exists and is continuous in x for all $\|v\|_1 \le k$, and

(ii) $|D^v h(x) - D^v h(0)| \le c\|x\|^\gamma$ for all $\|v\|_1 = k$, with $\|\cdot\|$ the norm on $x \in \mathbb{R}^d$ in (4).

At one extreme, if the conditional quantile function $Q_{Y|X}(\cdot;x)$ is the same for all $x \in C_h$, we will have the same results as for the discrete case taking $X_i = 0$ above. At the opposite extreme, if the conditional quantile function varies completely arbitrarily over $x \in C_h$, nothing can be learned from the data. If we make local smoothness assumptions on the conditional quantile function in between these two extremes, the method will be informative but subject to additional CPE due to bias.

Assuming a positive, continuous marginal density for X at the origin, N_n will asymptotically be proportional to the volume of C_h, which is proportional to h^d. There is a tradeoff: larger h lowers CPE via N_n but raises CPE via bias. This determines the optimal rate at which $h \to 0$ as $n \to \infty$. Using the precise rate results from Goldman and Kaplan (2012, §3.1) and new results established here, we determine the optimal value of h.

A Bonferroni approach gives joint CIs over $m < \infty$ different values of x_0. If the various x_0 yield non-intersecting C_h, then the m pointwise CIs will be independent since data are iid. In that case, pointwise $(1-\alpha)^{1/m}$ CIs can be used instead of $(1-\alpha/m)$, which is no longer conservative. Asymptotically, for a fixed number of x_0, this is always the case, and it may be a good approximation even if not exactly true in finite samples. However, the difference is small—for two-sided CIs with $\alpha = 0.05$, 0.975 vs. 0.9747 for $m = 2$, 0.999 vs. 0.99897 for $m = 50$, etc.—so the Bonferroni approach is always used here for convenience. As discussed in §5, an alternative Hotelling (1939) tube-based calibration of α yields similar results.

The unconditional IDEAL method is key to constructing our pointwise CI for $Q_{Y|X}(p; 0)$ in the conditional case, where we maintain $x_0 = 0$ as the point of interest.

Definition 3 (conditional IDEAL method)**.** Given continuous iid data $\{X_i, Y_i\}_{i=1}^{n}$, bandwidth $h > 0$, quantile $p \in (0, 1)$, and desired coverage probability $(1-\alpha)$, first C_h and N_n are calculated as in Definition 1. Using the values of Y_i from the effective sample of N_n observations with $X_i \in C_h$, the CI is then constructed as in Lemma 1. If additional discrete conditioning variables exist, this method may be run separately for each combination of discrete conditioning values, e.g. once for males and once for females. This procedure may be repeated for any number of x_0, too.

Remark. Code for the unconditional IDEAL method is publicly available in both MATLAB and R on the author's website. The only additional difficulty in the conditional setting is determining the optimal bandwidth (see §4.2 and §5.1). For $d = 1$, fully automated code for conditional IDEAL is available in R, also at the author's website.

Remark. Censoring and missing data can be accounted for in many cases. For instance, if Y_i is missing for some observations, there are two extreme cases to consider: replacing all the missing values with $Y_{\min}$ (the lower bound of the support of Y, or $-\infty$ if unbounded), or replacing all the missing values with $Y_{\max}$ (the upper bound of the support of Y, or ∞ if unbounded). A conservative CI is then the convex hull of the IDEAL CIs in the two extreme cases. If there are not too many missing values, this will still produce an informative CI.

Extensions of the IDEAL method to missing data (with or without assumptions like missing at random) and different types of censoring should be relatively straightforward and valuable.

In addition to the foregoing definitions, the following assumptions are maintained throughout. We continue using $x_0 = 0$ as the point of interest. Assumptions A1–A4(i) are needed for the bias calculation, while A4(ii)–A6 are needed to apply the unconditional IDEAL quantile inference method (Goldman and Kaplan, 2012).

Assumption A1. (X_i, Y_i) is iid across $i = 1, 2, \ldots, n$, where Y_i is a continuous scalar and X_i a continuous vector with support $\mathcal{X} \subset \mathbb{R}^d$.

Assumption A2. The marginal density of X, denoted $f_X(\cdot)$, satisfies $f_X(0) > 0$ and has local smoothness $s_X = k_X + \gamma_X > 0$ with constant c_X.

Assumption A3. For all u in a neighborhood of p, $Q_{Y|X}(u; \cdot)$ (as a function of the second argument) has local smoothness $s_Q = k_Q + \gamma_Q > 0$ with constant c_Q.

Assumption A4. As $n \to \infty$, (i) $h \to 0$, (ii) $nh^d/[\log(n)]^2 \to \infty$.

Assumption A5. The conditional density of Y is positive at the quantile and X values of interest: $f_{Y|X}\big(Q_{Y|X}(p; 0); 0\big) > 0$.

Assumption A6. For all y in a neighborhood of $Q_{Y|X}(p; 0)$ and x in a neighborhood of the origin, $f_{Y|X}(y; x)$ is twice continuously differentiable ($f_{Y|X} \in C^2$) in its first (Y) argument, i.e. has local smoothness $s_Y = k_Y + \gamma_Y > 2$ with constant c_Y.

Remark (smoothness). There is no minimum requirement of s_Q and s_X, though as $s_Q \to 0$ the inference becomes meaningless, as shown explicitly in §4. Since we are implicitly using a uniform kernel, which is a second-order kernel, there is no benefit to having smoothness greater than $(s_Q, s_X, s_Y) = (2, 1, 1) + \epsilon$ for some arbitrarily small $\epsilon > 0$, as stated in Lemma 2. Our s_Q corresponds to variable p in Chaudhuri (1991), who also notes that Bhattacharya and Gangopadhyay (1990) use $s_Q = 2$ and $d = 1$.

Remark (bandwidth). From A4(i), asymptotically C_h will be totally contained within the neighborhoods mentioned in A2, A3, and Definition 2. In order to get $N_n \overset{a.s.}{\to} \infty$, A4(ii) is a primitive condition. This in turn allows us to examine only a local neighborhood around quantile of interest p (e.g., as in A3), since asymptotically the CI endpoints will converge to the true value at a $\sqrt{N_n}$ rate. Reassuringly, the optimal bandwidth rate turns out to be inside the assumed bounds.

The bias may now be determined. Since our conditional quantile CI uses the subsample of Y_i with $X_i \in C_h$, rather than a subsample of Y_i with $X_i = 0$, our CI is constructed for the biased conditional quantile $Q_{Y|C_h}(p)$ (from Definition 1) rather than for $Q_{Y|X}(p;0)$. The bias is the difference between these two population conditional quantiles. For any h and p, the bias can be derived using A3.

Lemma 2. *Define* $b \equiv \min\{s_Q, s_X+1, 2\}$ *and* $B_h \equiv Q_{Y|C_h}(p) - Q_{Y|X}(p;0)$*. If Assumptions A2, A3, A4(i), and A6 hold, then the bias is of order*

$$|B_h| = O(h^b). \tag{6}$$

With $k_Q \geq 2$ *and* $k_X \geq 1$*, and defining*

$$Q^{(0,1)}_{Y|X}(p;0) \equiv \frac{\partial}{\partial x} Q_{Y|X}(p;x)\bigg|_{x=0}, \quad Q^{(0,2)}_{Y|X}(p;0) \equiv \frac{\partial^2}{\partial x^2} Q_{Y|X}(p;x)\bigg|_{x=0}, \quad \xi_p \equiv Q_{Y|X}(p;0),$$
$$f^{(0,1)}_{Y|X}(y;0) \equiv \frac{\partial}{\partial x} f_{Y|X}(y;x)\bigg|_{x=0}, \quad f^{(1,0)}_{Y|X}(\xi_p;0) \equiv \frac{\partial}{\partial y} f_{Y|X}(y;0)\bigg|_{y=\xi_p},$$

and similarly for $F^{(0,1)}_{Y|X}(\xi_p;0)$ *and* $F^{(0,2)}_{Y|X}(\xi_p;0)$*, the bias is*

$$\begin{aligned} B_h = \frac{h^2}{6}\Big\{ & 2Q^{(0,1)}_{Y|X}(p;0) f'_X(0)/f_X(0) + Q^{(0,2)}_{Y|X}(p;0) \\ & + 2 f^{(0,1)}_{Y|X}(\xi_p;0) Q^{(0,1)}_{Y|X}(p;0)/f_{Y|X}(\xi_p;0) \\ & + f^{(1,0)}_{Y|X}(\xi_p;0)\Big[Q^{(0,1)}_{Y|X}(p;0)\Big]^2 / f_{Y|X}(\xi_p;0)\Big\} + o(h^2) \\ = -h^2 & \frac{f_X(0)F^{(0,2)}_{Y|X}(\xi_p;0) + 2f'_X(0)F^{(0,1)}_{Y|X}(\xi_p;0)}{6 f_X(0) f_{Y|X}(\xi_p;0)} + o(h^2). \end{aligned}$$

Remark. The latter formulation for the bias is the same as that in Bhattacharya and Gangopadhyay (1990), who derive it using different arguments.

We discuss some intuition of the proof here. We start with an identity for $Q_{Y|C_h}(p)$ and subtract off the corresponding identity for $Q_{Y|X}(p;0)$. Using the smoothness assumptions, a Taylor expansion (of some order) of the remainder may be taken. The bias appears in the lowest-order term; some cancellation and rearrangement leads to the final expression. Further manipulations to replace $Q_{Y|X}$ with $F_{Y|X}$ lead to the exact same expression as in Bhattacharya and Gangopadhyay (1990).

Remark. Even when $k_Q \geq 2$ and $k_X \geq 1$, the bias will never shrink smaller than $O(h^2)$ since we are effectively using a second-order (uniform) kernel. It is unclear if the IDEAL fractional order statistic results can be used with a higher-order kernel. Alternatively, higher-order kernels could likely be used with a method as in Chernozhukov et al. (2009), who in the parametric quantile regression model use Bernoulli random variables and MCMC simulation.

4 Optimal bandwidth and CPE

4.1 Optimal rate of bandwidth and CPE

The optimal bandwidth minimizes the effect of the two dominant high-order terms on coverage probability error (CPE). In terms of coverage probability (CP) and nominal coverage $1-\alpha$, we follow convention and define CPE $\equiv$ CP $-(1-\alpha)$, so that CPE is positive when there is over-coverage and negative when there is under-coverage. This means the equivalent hypothesis test is size distorted when CPE is negative.

From Goldman and Kaplan (2012, §3.1), we know the IDEAL CPE in the unconditional one-sample case for Hutson's (1999) confidence interval (CI). In that case, we are interested in the p-quantile $F^{-1}(p)$ of scalar random variable Y. The Hutson (1999) method's CPE with respect to sample size n is of order n^{-1}. To apply this result, we need a more precise handle on the effective sample size N_n, which is random. From Chaudhuri (1991, proof of Thm. 3.1, p. 769), and under A1, we can choose $c_1, c_2, c_3, c_4 > 0$ such that

$$P(A_n) \geq 1 - c_3 \exp(-c_4 n h^d)$$

for all n, where $A_n \equiv \left\{c_1 n h^d \leq N_n \leq c_2 n h^d\right\}$ and C_h is the hypercube from Definition 1. (Adjusting the c_i appropriately, Chaudhuri's (1991) argument goes through for a hypersphere C_h also.) If the rate of h leads to $\sum_n [1 - P(A_n)] < \infty$, then the Borel–Cantelli Lemma gives $P(\liminf A_n) = 1$. This holds for the optimal bandwidth rates derived here, which all satisfy A4.

One-sided inference

In the lower one-sided case, we write $\hat{Q}^L_{Y|C_h}(u_h)$ as the Hutson (1999) upper endpoint, with notation similar to §2. This is a linearly interpolated fractional order statistic approximation calculated from the N_n values of Y_i with $X_i \in C_h$. The random variable $F_{Y|C_h}\left(\hat{Q}^I_{Y|C_h}(u_h)\right)$ follows a (collapsing) beta distribution, which has a continuously differentiable PDF in $(0,1)$ that converges to a (collapsing) normal PDF at a $\sqrt{N_n}$ rate (Goldman and Kaplan, 2012). Other than $O(N_n^{-1})$ CPE from interpolating between order statistics, the CI using $\hat{Q}^L_{Y|C_h}(u_h)$ is exact for $Q_{Y|C_h}(p)$. The fact that instead $Q_{Y|X}(p;0)$ is of interest introduces additional CPE from the bias. The CP of the lower one-sided CI is

$$\begin{aligned}
&P\Big(Q_{Y|X}(p;0) < \hat{Q}^L_{Y|C_h}(u_h)\Big) \\
&\quad = P\Big(Q_{Y|C_h}(p) < \hat{Q}^L_{Y|C_h}(u_h)\Big)
\end{aligned}$$

$$
\begin{aligned}
&+ \Big[P\Big(Q_{Y|X}(p;0) < \hat{Q}^L_{Y|C_h}(u_h)\Big) - P\Big(Q_{Y|C_h}(p) < \hat{Q}^L_{Y|C_h}(u_h)\Big)\Big] \\
&= 1 - \alpha + \mathrm{CPE}_{\mathrm{GK}} + \mathrm{CPE}_{\mathrm{Bias}}, \qquad (7)
\end{aligned}
$$

where $\mathrm{CPE}_{\mathrm{GK}}$ is due to the Goldman and Kaplan (2012) CPE from Lemma 1 and $\mathrm{CPE}_{\mathrm{Bias}}$ comes from the bias discussed in §3.

As before, define $B_h \equiv Q_{Y|C_h}(p) - Q_{Y|X}(p;0)$. From Lemma 2, we have $B_h = O(h^b)$ with $b \equiv \min\{s_Q, s_X + 1, 2\}$. Let $F_{\hat{Q}^{I,u_h}_{Y|C_h}}(\cdot)$ and $f_{\hat{Q}^{I,u_h}_{Y|C_h}}(\cdot)$ be the CDF and PDF, respectively, of $\hat{Q}^I_{Y|C_h}(u_h)$. If B_h is sufficiently small—i.e. if $h^b = o(N_n^{-1/2})$, which is true below since $h^b = N_n^{-3/2}$—then we can approximate

$$
\begin{aligned}
\mathrm{CPE}_{\mathrm{Bias}} &= P\Big(Q_{Y|X}(p;0) < \hat{Q}^L_{Y|C_h}(u_h)\Big) - P\Big(Q_{Y|C_h}(p) < \hat{Q}^L_{Y|C_h}(u_h)\Big) \\
&= P\Big(\hat{Q}^L_{Y|C_h}(u_h) < Q_{Y|C_h}(p)\Big) - P\Big(\hat{Q}^L_{Y|C_h}(u_h) < Q_{Y|X}(p;0)\Big) \\
&= P\Big(\hat{Q}^I_{Y|C_h}(u_h) < Q_{Y|C_h}(p)\Big) - P\Big(\hat{Q}^I_{Y|C_h}(u_h) < Q_{Y|X}(p;0)\Big) + O(B_h N_n^{-1/2}) \\
&= F_{\hat{Q}^{I,u_h}_{Y|C_h}}\big(Q_{Y|C_h}(p)\big) - F_{\hat{Q}^{I,u_h}_{Y|C_h}}\big(Q_{Y|X}(p;0)\big) + O(B_h N_n^{-1/2}) \\
&= B_h f_{\hat{Q}^{I,u_h}_{Y|C_h}}\big(Q_{Y|C_h}(p)\big) + O(B_h N_n^{-1/2} + B_h^2 N_n), \qquad (8)
\end{aligned}
$$

where the order of the approximation error from switching to $\hat{Q}^I_{Y|C_h}(u_h)$ from $\hat{Q}^L_{Y|C_h}(u_h)$ comes from the first theorem in Goldman and Kaplan (2012), and the other remainder is the subsequent B_h^2 term in the Taylor expansion that would be multiplied by an $O(N_n)$ PDF derivative as in (10). From the aforementioned PDF convergence of $F_{Y|C_h}(\hat{Q}^I_{Y|C_h}(u_h))$ to a normal, it can be shown that $f_{\hat{Q}^{I,u_h}_{Y|C_h}}\big(Q_{Y|X}(p;0)\big) \asymp N_n^{1/2}$. Since $B_h = O(h^b)$ from Lemma 2, the dominant term of $\mathrm{CPE}_{\mathrm{Bias}}$ is $O(N_n^{1/2}h^b)$. The expression in (8) holds for $B_h > 0$ (leading to over-coverage) or $B_h < 0$ (under-coverage).

The two dominant terms of $\mathrm{CPE}_{\mathrm{GK}}$ and $\mathrm{CPE}_{\mathrm{Bias}}$ are thus respectively $O(N_n^{-1})$ and $O(N_n^{1/2}h^b)$. These are both sharp except in the special case of $u_h(N_n + 1)$ being an integer or of $f_X(0)F^{(0,2)}_{Y|X}(\xi_p;0) + 2f'_X(0)F^{(0,1)}_{Y|X}(\xi_p;0) = 0$ when $k_Q \geq 2$ and $k_X \geq 1$ (or similar conditions otherwise); the following assumes we are not in a special case. The term $\mathrm{CPE}_{\mathrm{GK}}$ is always positive. If $\mathrm{CPE}_{\mathrm{Bias}}$ is negative, the optimal h will make them cancel, and it will set the orders equal:

$$
N_n^{-1} \asymp N_n^{1/2}h^b \implies (nh^d)^{3/2} \asymp h^{-b} \implies h \asymp n^{-3/(2b+3d)}.
$$

Overall CPE will then be $o(n^{-2b/(2b+3d)})$, which can likely be sharpened further. Even if h does not make the dominant CPE terms cancel, as long as it is the above asymptotic rate, the overall CPE will be $O(n^{-2b/(2b+3d)})$.

If instead the two terms are both positive, minimizing the sum will lead to a first-order condition like

$$0 = \frac{\partial}{\partial h}\left[N_n^{-1} + N_n^{1/2}h^b\right] = (-d)n^{-1}h^{-d-1} + (b+(d/2))n^{1/2}h^{b+(d/2)-1},$$

giving the same rate $h \asymp n^{-3/(2b+3d)}$. In that case, overall CPE will be positive (over-coverage) and of order

$$N_n^{-1} \asymp (nh^d)^{-1} \asymp n^{-1+3d/(2b+3d)} = n^{-2b/(2b+3d)}.$$

If the calibrated unconditional method from Goldman and Kaplan (2012) is used, $\text{CPE}_{\text{GK}} = O\Big(N_n^{-3/2+\rho}\Big)$ for arbitrarily small $\rho > 0$. Ignoring the ρ for simplicity, the optimal bandwidth rate is then $h \asymp n^{-2/(b+2d)}$, leading to overall CPE of $O\big(n^{-3b/(2b+4d)}\big)$.

In the upper one-sided case, with $\hat{Q}^L_{Y|C_h}(u_\ell)$ the lower endpoint, CPE_{GK} is of the same order and sign, while

$$\begin{aligned}
\text{CPE}_{\text{Bias}} &= P\Big(Q_{Y|X}(p;0) > \hat{Q}^L_{Y|C_h}(u_\ell)\Big) - P\Big(Q_{Y|C_h}(p) > \hat{Q}^L_{Y|C_h}(u_\ell)\Big) \\
&= F_{\hat{Q}^{I,u_\ell}_{Y|C_h}}\big(Q_{Y|X}(p;0)\big) - F_{\hat{Q}^{I,u_\ell}_{Y|C_h}}\big(Q_{Y|C_h}(p)\big) + O(B_h N_n^{-1/2}) \\
&= -B_h f_{\hat{Q}^{I,u_\ell}_{Y|C_h}}\big(Q_{Y|C_h}(p)\big) + O(B_h N_n^{-1/2} + B_h^2 N_n). \qquad (9)
\end{aligned}$$

Opposite before, $B_h > 0$ now contributes under-coverage and $B_h < 0$ over-coverage, but for now it suffices to note that the order of CPE_{Bias} is the same as before.

Two-sided inference

With two-sided inference, the lower and upper endpoints have opposite bias effects, but in general they will not cancel completely. Below, since by construction $\hat{Q}^L_{Y|C_h}(u_\ell) < \hat{Q}^L_{Y|C_h}(u_h)$, it is certain that $\hat{Q}^L_{Y|C_h}(u_\ell) < c$ if $\hat{Q}^L_{Y|C_h}(u_h) < c$ and that $\hat{Q}^L_{Y|C_h}(u_h) > c$ if $\hat{Q}^L_{Y|C_h}(u_\ell) > c$, for any c. With two-sided CI $(\hat{Q}^L_{Y|C_h}(u_\ell), \hat{Q}^L_{Y|C_h}(u_h))$, CP is

$$\begin{aligned}
&P\Big(\hat{Q}^L_{Y|C_h}(u_\ell) < Q_{Y|X}(p;0) < \hat{Q}^L_{Y|C_h}(u_h)\Big) \\
&\quad = 1 - P\Big(\hat{Q}^L_{Y|C_h}(u_\ell) > Q_{Y|X}(p;0)\Big) - P\Big(\hat{Q}^L_{Y|C_h}(u_h) < Q_{Y|X}(p;0)\Big) \\
&\quad = 1 - P\Big(\hat{Q}^L_{Y|C_h}(u_\ell) > Q_{Y|C_h}(p)\Big) \\
&\qquad + \Big[P\Big(\hat{Q}^L_{Y|C_h}(u_\ell) > Q_{Y|C_h}(p)\Big) - P\Big(\hat{Q}^L_{Y|C_h}(u_\ell) > Q_{Y|X}(p;0)\Big)\Big] \\
&\qquad - P\Big(\hat{Q}^L_{Y|C_h}(u_h) < Q_{Y|C_h}(p)\Big) \\
&\qquad + \Big[P\Big(\hat{Q}^L_{Y|C_h}(u_h) < Q_{Y|C_h}(p)\Big) - P\Big(\hat{Q}^L_{Y|C_h}(u_h) < Q_{Y|X}(p;0)\Big)\Big]
\end{aligned}$$

$$
\begin{aligned}
&= 1-\alpha+\mathrm{CPE}_{\mathrm{GK}}+\Bigl[1-F_{\hat{Q}^{I,u_\ell}_{Y|C_h}}\bigl(Q_{Y|C_h}(p)\bigr)\Bigr]-\Bigl[1-F_{\hat{Q}^{I,u_\ell}_{Y|C_h}}\bigl(Q_{Y|X}(p;0)\bigr)\Bigr]\\
&\quad+F_{\hat{Q}^{I,u_h}_{Y|C_h}}\bigl(Q_{Y|C_h}(p)\bigr)-F_{\hat{Q}^{I,u_h}_{Y|C_h}}\bigl(Q_{Y|X}(p;0)\bigr)+O(B_hN_n^{-1/2})\\
&= 1-\alpha+\mathrm{CPE}_{\mathrm{GK}}+B_h\Bigl[f_{\hat{Q}^{I,u_h}_{Y|C_h}}\bigl(Q_{Y|C_h}(p)\bigr)-f_{\hat{Q}^{I,u_\ell}_{Y|C_h}}\bigl(Q_{Y|C_h}(p)\bigr)\Bigr]\\
&\quad+(1/2)B_h^2\Bigl[f'_{\hat{Q}^{I,u_\ell}_{Y|C_h}}\bigl(Q_{Y|C_h}(p)\bigr)-f'_{\hat{Q}^{I,u_h}_{Y|C_h}}\bigl(Q_{Y|C_h}(p)\bigr)\Bigr]\\
&\quad+O\Bigl\{B_h^3f''_{\hat{Q}}\bigl(Q_{Y|C_h}(p)\bigr)+B_hN_n^{-1/2}\Bigr\}. \qquad (10)
\end{aligned}
$$

For the special case of the median, the B_h term zeroes out. This happens because the beta distribution PDFs of $F_{Y|C_h}\Bigl(\hat{Q}^I_{Y|C_h}(u)\Bigr)$ and $F_{Y|C_h}\Bigl(\hat{Q}^I_{Y|C_h}(1-u)\Bigr)$ are reflections of each other around $p=1/2$—i.e., $f_\beta(x;u)=f_\beta(1-x;1-u), \forall x\in(0,1)$—so the upper and lower u are symmetric around $p=1/2$; see Figure 1 for an example. Consequently, the bias effect on CPE is the B_h^2 term instead. This makes the overall CPE smaller. The optimal rate of h will equate $(nh^d)^{-1}\asymp h^{2b}(nh^d)$, so $h^*\asymp n^{-1/(b+d)}$ and CPE is $O(n^{-b/(b+d)})$. With the calibrated method (again suppressing the $\rho>0$), the rates would instead be $h^*\asymp n^{-5/(4b+3d)}$ and CPE $=O\bigl(n^{-6b/(4b+5d)}\bigr)$.

Even with $p\neq 1/2$, the same rates hold for two-sided inference. As seen in (11), the PDF difference multiplying B_h is only $O(1)$, smaller than the $O(N_n^{1/2})$ PDF value multiplying B_h in the one-sided expression (8). This makes the B_h and B_h^2 terms the same order, as discussed further in §4.2.

Theorem 3. *Let Assumptions A1–A6 hold, and define* $b\equiv\min\{s_Q,s_X+1,2\}$. *For a one-sided CI, the bandwidth* h^* *minimizing CPE for the Hutson (1999) method applied to observations falling inside* C_h *has rate* $h^*\asymp n^{-3/(2b+3d)}$. *This corresponds to overall CPE of* $O(n^{-2b/(2b+3d)})$.

For two-sided inference, the optimal bandwidth rate is $h^*\asymp n^{-1/(b+d)}$*, and the optimal CPE is* $O(n^{-b/(b+d)})$.

With the precise bandwidth value provided in §4.2, the two-sided CPE reduces to $o(n^{-b/(b+d)})$. *With* $k_Q\geq 2$ *and* $k_X\geq 1$*, the two-sided CPE becomes* $o\bigl(n^{-2/(2+d)}\bigr)$.

Using the calibrated method proposed in Goldman and Kaplan (2012), the two-sided CPE-optimal bandwidth rate is $h^*\asymp n^{-(5-2\rho)/(4b+5d-2d\rho)}$ *for arbitrarily small* $\rho>0$*, yielding CPE of* $O\bigl(n^{-(6b-4b\rho)/(4b+5d-2d\rho)}\bigr)$. *The optimal one-sided calibrated bandwidth rate is* $h^*\asymp n^{-(2-\rho)/(b+2d-d\rho)}$*, yielding CPE of* $O\bigl(n^{-(3b-2b\rho)/(2b+4d-2d\rho)}\bigr)$.

Discussion: smoothness and bandwidth

The following discussion is for the more common two-sided inference, with one-sided equivalents noted in parentheses. Let $\kappa \equiv 1/(b+d)$ (one-sided: $\kappa \equiv 3/(2b+3d)$), so that $h = n^{-\kappa}$ above, with CPE of order $n^{d\kappa-1}$. Since $0 < b \leq 2$, the optimal κ will depend on A2 and A3 but fall within the range $1/(2+d) \leq \kappa < 1/d$ (one-sided: $3/(4+3d) \leq \kappa < 1/d$). This corresponds to CPE order in the range $[n^{-2/(2+d)}, n^0)$ (one-sided: $[n^{-4/(4+3d)}, n^0)$). The high end of this range matches intuition: as the smoothness diminishes to zero ($s_Q \to 0$), we are unable to say anything informative. However, the smoothness levels $(s_Q, s_X, s_Y) = (2, 1, 2) + \epsilon$ for any small $\epsilon > 0$ are quite mild, so it is most helpful in practice to look at the other end of the range.

As d increases, $h \to 0$ more slowly (smaller κ). More smoothness also makes $h \to 0$ more slowly: we can afford a relatively larger window C_h if there is less variability near our point of interest.

With $d = 1$ and $b = 2$, we get $h^* \asymp n^{-1/3}$ and a CPE of order $n^{-2/3}$ (one-sided: $h^* \asymp n^{-3/7}$, CPE $n^{-4/7}$). With $d = 2$, the CPE order increases to $n^{-1/2}$ (one-sided: $n^{-2/5}$). We expect the method to work (relatively) well with even higher-dimensional X, though CPE continues to increase as d increases.

Regarding robustness to bandwidth, our CI will be asymptotically (first-order) correct as long as $N_n^{-1} \to 0$ (so $\text{CPE}_{\text{GK}} \to 0$) and $N_n h^{2b} \to 0$ (so $\text{CPE}_{\text{Bias}} \to 0$). These are equivalent to $1/(2b+d) < \kappa < 1/d$. With the common example $d = 1$ and $b = 2$, this says that any $h \asymp n^{-\kappa}$ with $1/5 < \kappa < 1$ will give CPE $= o(1)$. The range for κ is the same as when using the local polynomial asymptotic normality approach.

Discussion: comparison with asymptotic normality or bootstrap

Theorem 3 suggests that for most common combinations of smoothness s_Q and dimensionality d, our method is preferred to inference based on asymptotic normality or bootstrap with a local polynomial estimator. Since these are the only known alternatives for inference in the literature, the following discussion is detailed. The main limitation of our method is the uniform kernel required to leverage the fractional order statistic theory. Even though asymptotic normality has a larger error in terms of N_n, it could be smaller in terms of n if N_n is allowed to be much bigger. This could potentially happen if a high enough degree of smoothness s_Q is assumed and there are enough observations that it is appropriate to fit a correspondingly high-degree local polynomial. As we will see, though, our method has smaller CPE in the most important cases even with $s_Q = \infty$.

Results from Chaudhuri (1991) can be used to obtain the optimal CPE for inference based on asymptotic normality of the local polynomial estimator. Note that $\hat{Q}_{Y|X}(p;0) = \hat{\beta}_{(0)}$, the intercept term estimator from the local polynomial quantile regression, so inference on $\beta_{(0)}$ is equivalent to inference on $Q_{Y|X}(p;0)$. The goal of Chaudhuri (1991) is to show that the local polynomial estimator therein achieves the optimal rate of convergence given by Stone (1980, 1982) for nonparametric mean regression. A decomposition of the estimator is given as

$$\hat{\beta} - \beta_0 = V_n + B_n + R_n,$$

where R_n is a Bahadur-type remainder from Theorem 3.3, B_n is the bias from equation (4.1), and V_n is the term from Proposition 4.2 that when scaled by $\sqrt{N_n}$ converges to a Gaussian limit. To get the best rate, it is necessary to balance the squared bias with the variance. The given MSE-optimal bandwidth is $h \propto n^{-1/(2s_Q+d)}$ (p. 763, in our notation), balancing the square of the $B_n = O(h^{s_Q}) = O(n^{-s_Q/(2s_Q+d)})$ bias (Prop. 4.1; $r_n(x)$ on p. 765) with the N_n^{-1} variance (Prop. 4.2), where $N_n \overset{a.s.}{\asymp} n^{2s_Q/(2s_Q+d)}$ (Prop. 4.2 proof). However, with the bias the same order of magnitude as the standard deviation, CPE is $O(1)$. If CPE is the target instead of MSE, a smaller bandwidth is necessary.

To find the CPE-optimal bandwidth for Chaudhuri (1991), we balance the CPE from the bias with additional CPE from the Bahadur remainder from Theorem 3.3(ii), ignoring asymptotic variance estimation error. The CPE due to the bias is of order $N_n^{1/2}h^{s_Q}$. Chaudhuri (1991, Thm. 3.3) gives a Bahadur-type expansion of the local polynomial quantile regression estimator that has remainder $R_n\sqrt{N_n} = O(N_n^{-1/4})$ (up to log terms) as in Bahadur (1966), but recently Portnoy (2012) has shown that the CPE is nearly $O(N_n^{-1/2})$ in such cases. Solving $N_n^{1/2}h^{s_Q} = N_n^{-1/2} = (nh^d)^{-1/2}$, this yields $h^* \asymp n^{-1/(s_Q+d)}$. The optimal CPE is then (nearly) $O(\sqrt{N_n}B_n) = O(N_n^{-1/2}) = O(n^{-s_Q/(2s_Q+2d)})$.

Specifically, and very much theoretically, in Chaudhuri (1991), if $s_Q \to \infty$, then $N_n \to n$ and (nearly) CPE $\to O(n^{-1/2})$. In other words, with infinite smoothness, almost the entire sample n is used because the fitted infinite-degree polynomial can perfectly approximate the true function. Practically, when we have a finite sample, we can't even fit an nth-degree polynomial, let alone infinite-degree. Additionally, the argument in Chaudhuri (1991) is that even if the smoothness only holds in some tiny neighborhood V, asymptotically there will be an infinite number of observations within V (and C_h will be contained in V). But when the sample is finite, smoothness must hold over C_h, which may not be so small (and is always of larger-order volume than our C_h). Barring relevant a priori information, speculating more smoothness over a larger region may be prohibitively unpalatable in light of the more robust method we provide.

If $s_Q = 2$ (one Lipschitz-continuous derivative), then optimal CPE from asymptotic normality is nearly $O(n^{-2/(4+2d)})$. This goes to n^0 as $d \to \infty$ (same as above). With $d = 1$, this is $n^{-1/3}$, significantly larger than our $n^{-2/3}$ (one-sided: $n^{-4/7}$); with $d = 2$, $n^{-1/4}$ is larger than our $n^{-1/2}$ (one-sided: $n^{-2/5}$); and it remains larger for all d (even for one-sided inference). With $s_Q = 1$ (no derivatives, but Lipschitz continuity), then optimal CPE from asymptotic normality is nearly $O(n^{-1/(2+2d)})$, compared to our CPE of $O(n^{-1/(1+d)})$ (one-sided: $O(n^{-1/(2+2d)})$). With $d = 1$, this is $n^{-1/4}$, again much larger than our $n^{-1/2}$ (one-sided: $n^{-2/5}$); and again it remains larger for all d (including one-sided).

From another perspective: what amount of smoothness (and degree of local polynomial fit) is needed for asymptotic normality to match the CPE of our method? For any d, the bound on CPE for asymptotic normality is nearly $n^{-1/2}$ (with $s_Q \to \infty$). For the most common cases of $d \in \{1, 2\}$ (one-sided: $d = 1$), asymptotic normality will be worse even with infinite smoothness. With $d = 3$ (one-sided: $d = 2$), asymptotic normality needs $s_Q \geq 12$ (one-sided: $s_Q \geq 12$) to have as good CPE. If n is quite large, maybe that high of a local polynomial degree ($k_Q \geq 11$) will be appropriate, but often it is not. Note that interaction terms are required, so an 11th-degree polynomial has 364 terms. As $d \to \infty$, the required smoothness approaches $s_Q = 4$ (one-sided: 8/3) from above, though again the number of terms in the local polynomial grows with d as well as k_Q and may be prohibitive in finite samples. Even with very high-dimensional X, asymptotic normality will only be better under a stronger smoothness assumption and fourth-degree (i.e., includes x^4 term) local polynomial fit.

In finite samples, the asymptotic normality approach may have additional error from estimation of the asymptotic variance, which includes the "dispersion matrix" as well as the probability density of the error term at zero as discussed in Chaudhuri (1991, pp. 764–766). There is no direct error from nuisance parameter estimation in our method, only through the plug-in bandwidth.

For the bootstrap, the basic percentile method offers no higher-order refinement over first-order asymptotic normality. Consequently, our method has better CPE than the bootstrap in all the cases discussed above. Gangopadhyay and Sen (1990) examine the bootstrap percentile method for a uniform kernel (or nearest-neighbor) estimator with $d = 1$ (and $s_Q = 2$), but they only show first-order consistency. Based on their (2.14), the optimal error seems to be $O(n^{-1/3})$ when $h \asymp n^{-1/3}$ to balance the bias and remainder terms, improved by Portnoy (2012) from $O(n^{-2/11})$ CPE when $h \asymp n^{-3/11}$. Although we are unaware of any improvements on Gangopadhyay and Sen (1990) to date, it is still valuable to pursue different inference strategies, and developing strong resampling or subsampling competitors here is no exception.

4.2 Plug-in bandwidth

The terms CPE_{GK} and CPE_{Bias} can be calculated more precisely than simply the rates. Goldman and Kaplan (2012) give an exact expression for the $O(N_n^{-1})$ CPE_{GK} term. This can be used to reduce the CPE to almost $O(N_n^{-3/2})$ via analytic calibration, but we use it to determine the optimal bandwidth. In theory, it is better to use the analytic calibration and an ad hoc bandwidth of the proper rate. In practice, it is helpful to know that the $O(N_n^{-1})$ CPE term only leads to over-coverage, which implies that smaller h is always more conservative, and the detrimental effect of an ad hoc bandwidth can be significant in smaller samples.

As before, $p \in (0, 1)$ is the quantile of interest. Let $\phi(\cdot)$ be the standard normal PDF, $z_{1-\alpha}$ be the $(1-\alpha)$-quantile of the standard normal distribution, $u_h > p$ (or $u_\ell < p$) be the quantile determining the high (low) endpoint of a lower (upper) one-sided CI, $\epsilon_h \equiv (N_n+1)u_h - \lfloor (N_n+1)u_h \rfloor$, and $\epsilon_\ell \equiv (N_n+1)u_\ell - \lfloor (N_n+1)u_\ell \rfloor$. Let I_H denote a $100(1-\alpha)\%$ CI constructed using Hutson's (1999) method on a univariate data sample of size N_n. The coverage probability of a lower one-sided I_H (with the upper one-sided result substituting ϵ_ℓ for ϵ_h) is

$$P\{F^{-1}(p) \in I_H\} = 1 - \alpha + N_n^{-1} z_{1-\alpha} \frac{\epsilon_h(1-\epsilon_h)}{p(1-p)} \phi(z_{1-\alpha}) + o(N_n^{-1}),$$

or for a two-sided CI,

$$P\{F^{-1}(p) \in I_H\} = 1 - \alpha + N_n^{-1} z_{1-\alpha/2} \frac{\epsilon_h(1-\epsilon_h) + \epsilon_\ell(1-\epsilon_\ell)}{p(1-p)} \phi(z_{1-\alpha/2}) + o(N_n^{-1}).$$

In either case there is $O(N_n^{-1})$ over-coverage.

While we know the sign of CPE_{GK}, we don't always know the sign of the bias. Specifically, when the rate-limiting term of B_h is determined by Hölder continuity (of $Q_{Y|X}(\cdot;\cdot)$ or $f_X(\cdot)$), we lose the sign at that step. However, when $k_Q \geq 2$ and $k_X \geq 1$, the Hölder continuity terms all end up in the remainder while the rate-limiting terms are signed derivatives.

Since $(k_Q, k_X, k_Y) = (2, 1, 2)$ is only a mild smoothness assumption, which also gives the smallest attainable order of bias, we maintain it for our plug-in bandwidth. We also take $d = 1$ for simplicity.

From Lemma 2, we have two explicit expressions for B_h,

$$\begin{aligned} B_h = \frac{h^2}{6}\Big\{ & \Big[2Q_{Y|X}^{(0,1)}(p;0) f_X'(0)/f_X(0) + Q_{Y|X}^{(0,2)}(p;0)\Big] + 2f_{Y|X}^{(0,1)}(\xi_p;0) Q_{Y|X}^{(0,1)}(p;0)/f_{Y|X}(\xi_p;0) \\ & + f_{Y|X}^{(1,0)}(\xi_p;0)\Big[Q_{Y|X}^{(0,1)}(p;0)\Big]^2 / f_{Y|X}(\xi_p;0)\Big\} + o(h^2) \end{aligned}$$

$$= -h^2 \frac{f_X(0) F_{Y|X}^{(0,2)}(\xi_p; 0) + 2 f_X'(0) F_{Y|X}^{(0,1)}(\xi_p; 0)}{6 f_X(0) f_{Y|X}(\xi_p; 0)} + o(h^2).$$

For $f_X(0)$, $f_X'(0)$, and $f_{Y|X}(\xi_p; 0)$, we could either estimate them or use a parametric plug-in assumption (e.g., assume the distribution is Gaussian and estimate its parameters). For the first formulation above, $Q_{Y|X}^{(0,1)}(p; 0)$, $Q_{Y|X}^{(0,2)}(p; 0)$, $f_{Y|X}^{(1,0)}(\xi_p; 0)$, and $f_{Y|X}^{(0,1)}(\xi_p; 0)$ also must be estimated. Alternatively, for the second formulation, $F_{Y|X}^{(0,1)}(\xi_p; 0)$ and $F_{Y|X}^{(0,2)}(\xi_p; 0)$ must be estimated. Recall that $X = 0$ really means $X = x_0$, our target point of interest.

Additionally, CPE_{Bias} depends on $f_{\hat{Q}_u}\big(Q_{Y|C_h}(p)\big)$. From earlier, $F_{Y|C_h}(\hat{Q}_u) \sim \beta[(N_n + 1)u, (N_n + 1)(1 - u)]$. Let $f_\beta(\cdot; u)$ denote the corresponding beta distribution's PDF and $F_\beta(\cdot; u)$ its CDF. As shown in the appendix, some identities and calculus lead to

$$f_{\hat{Q}_u}\big(Q_{Y|C_h}(p)\big) = f_\beta(p; u) f_{Y|C_h}\Big(F_{Y|C_h}^{-1}(p)\Big).$$

The term $f_{Y|C_h}\Big(F_{Y|C_h}^{-1}(p)\Big)$ is equal to $f_{Y|X}(\xi_p; 0)$ up to smaller-order terms. The term $f_\beta(p; u)$ is well-approximated by a normal PDF. As detailed in the appendix, for either one-sided $(1 - \alpha)$ CI endpoint quantile $u = u_h$ or $u = u_\ell$ chosen by the Hutson (1999) method,

$$f_\beta(p; u) = N_n^{1/2} [u(1 - u)]^{-1/2} \phi(z_{1-\alpha}) \big[1 + O(N_n^{-1/2})\big].$$

For convenient reference, the plug-in bandwidth expressions are collected here. Some intuition follows; details of calculation may be found in the appendix. We continue to assume $d = 1$, $k_Q \geq 2$, and $k_X \geq 1$, with quantile of interest $p \in (0, 1)$ and point of interest $X = x_0$. Standard normal quantiles are denoted, for example, $z_{1-\alpha}$ for the $(1 - \alpha)$-quantile such that $\Phi(z_{1-\alpha}) = 1 - \alpha$. We let $\hat{B}_h$ denote the estimator of bias term B_h; $\hat{f}_X$ the estimator of $f_X(x_0)$; $\hat{f}_X'$ the estimator of $f_X'(x_0)$; $\hat{F}_{Y|X}^{(0,1)}$ the estimator of $F_{Y|X}^{(0,1)}(\xi_p; x_0)$; and $\hat{F}_{Y|X}^{(0,2)}$ the estimator of $F_{Y|X}^{(0,2)}(\xi_p; x_0)$, where $\xi_p \equiv Q_{Y|X}(p; x_0)$. To avoid a recursive definition of the plug-in bandwidth, we use $\epsilon_h = \epsilon_\ell = 0.2$ as a rule of thumb (which cancels nicely with other constants). The largest possible bandwidth would use 0.5 instead, which would give extremely similar bandwidths since, for example, $[(0.2)(0.8)]^{1/6} = 0.74$ while $[(0.5)(0.5)]^{1/6} = 0.79$ in the two-sided median case.

We recommend the following when $d = 1$.

- For one-sided inference, let

$$\hat{h}_{+-} = n^{-3/7} \left(\frac{z_{1-\alpha}}{3\Big[p(1 - p)\hat{f}_X\Big]^{1/2} \Big\{ \hat{f}_X \hat{F}_{Y|X}^{(0,2)} + 2 \hat{f}_X' \hat{F}_{Y|X}^{(0,1)} \Big\}} \right)^{2/7},$$

$$\hat{h}_{++} = -0.770 \hat{h}_{+-}.$$

- For lower one-sided inference, $\hat{h}_{+-}$ should be used if $\hat{B}_h < 0$, and $\hat{h}_{++}$ otherwise.
- For upper one-sided inference, $\hat{h}_{++}$ should be used if $\hat{B}_h < 0$, and $\hat{h}_{+-}$ otherwise.
- For two-sided inference on the median,

$$\hat{h} = n^{-1/3}\left(\frac{3\hat{f}_{Y|X}}{\left\{\hat{f}_X \hat{F}^{(0,2)}_{Y|X} + 2\hat{f}'_X \hat{F}^{(0,1)}_{Y|X}\right\}^2}\right)^{1/6}.$$

- For two-sided inference with $p \neq 1/2$ (and equivalent to above with $p = 1/2$),

$$\hat{h} = n^{-1/3}\left(\frac{(2p-1)(\hat{B}_h/|\hat{B}_h|) + \sqrt{(2p-1)^2 + (4/3)/\hat{f}_{Y|X}}}{(2/3)\left|\hat{f}_X \hat{F}^{(0,2)}_{Y|X} + 2\hat{f}'_X \hat{F}^{(0,1)}_{Y|X}\right|/\hat{f}_{Y|X}}\right)^{1/3}.$$

  Alternatively, the median-specific bandwidth may be used if u_h and u_ℓ are chosen such that $f_\beta(p; u_h) = f_\beta(p; u_\ell)$, which requires relaxing the equal-tailed restriction of (2). The benefit is a simpler expression for the bandwidth; the costs are an additional (though fast) numerical search and loss of the equal-tailed property.

Regarding the signs of the two CPE terms in (7), we know that $\text{CPE}_{\text{GK}} > 0$ (over-coverage). We can estimate the sign of CPE_{Bias} as the sign of B_h for lower one-sided inference and the opposite of the sign of B_h for upper one-sided inference. For two-sided inference, the sign of CPE_{Bias} depends on the bandwidth, and there always exists a bandwidth such that $\text{CPE}_{\text{Bias}} < 0$ and cancels CPE_{GK}. In the one-sided case, if $\text{CPE}_{\text{Bias}} < 0$, then the optimal bandwidth causes the two CPE terms to cancel out; if $\text{CPE}_{\text{Bias}} > 0$, the optimal bandwidth minimizes their sum. The only difference is an extra coefficient of $[2d/(2b+d)]^{1/(b+3d/2)} = [2d/(d+4)]^{2/(4+3d)}$ from the first-order condition in the latter case, where the initial exponents of h come down when taking a derivative. If $\text{CPE}_{\text{Bias}} < 0$, then overall CPE is $o(n^{-4/7})$; if $\text{CPE}_{\text{Bias}} > 0$, then overall CPE is $O(n^{-4/7})$ and positive (over-coverage).

In the rest of this subsection, we provide some additional details on the two-sided case. For two-sided inference with $p = 1/2$, the B_h term becomes zero, as is clear in (12) below. Then $\text{CPE}_{\text{Bias}} < 0$ since $B_h^2 > 0$, $f'_{\hat{Q}^{I,u_\ell}_{Y|C_h}}(p) < 0$, and $f'_{\hat{Q}^{I,u_h}_{Y|C_h}}(p) > 0$. The optimal h causes this to cancel with $\text{CPE}_{\text{GK}} > 0$. By the convergence of our beta to a normal distribution, the product rule for derivatives, and the invariance of $f_{Y|C_h}\Big(F^{-1}_{Y|C_h}(p)\Big)$ to u,

$$\begin{aligned} &f'_{\hat{Q}^{I,u_\ell}_{Y|C_h}}\left(Q_{Y|C_h}(p)\right) - f'_{\hat{Q}^{I,u_h}_{Y|C_h}}\left(Q_{Y|C_h}(p)\right) \\ &\qquad \doteq -z_{1-\alpha/2} N_n \phi(z_{1-\alpha/2}) 2[p(1-p)]^{-1} f_{Y|C_h}\Big(F^{-1}_{Y|C_h}(p)\Big), \end{aligned}$$

and

$$
\begin{aligned}
&N_n^{-1} z_{1-\alpha/2} \frac{\epsilon_h(1-\epsilon_h)+\epsilon_\ell(1-\epsilon_\ell)}{p(1-p)} \phi(z_{1-\alpha/2}) \\
&\quad = -(1/2) h^4 \left(\frac{f_X(0) F_{Y|X}^{(0,2)}(\xi_p;0) + 2 f_X'(0) F_{Y|X}^{(0,1)}(\xi_p;0)}{6 f_X(0) f_{Y|X}(\xi_p;0)} \right)^2 \\
&\qquad \times \left\{ -z_{1-\alpha/2} N_n \phi(z_{1-\alpha/2}) 2[p(1-p)]^{-1} f_{Y|X}(\xi_p;0) \right\}
\end{aligned}
$$

leads to the plug-in bandwidth.

For two-sided inference with $p \neq 1/2$, the following does not cancel but is of smaller order than the $O(N_n^{1/2})$ in the one-sided case:

$$
\begin{aligned}
f_\beta(p;u_h) - f_\beta(p;u_\ell) &= N_n^{1/2} \phi(z_{1-\alpha/2}) \big([u_h(1-u_h)]^{-1/2} - [u_\ell(1-u_\ell)]^{-1/2} \big) \\
&= N_n^{1/2} \phi(z_{1-\alpha/2}) \left(\frac{2p-1}{2[p(1-p)]^{3/2}} [(u_h - p) - (u_\ell - p)] + O(N_n^{-1}) \right) \\
&= z_{1-\alpha/2} \phi(z_{1-\alpha/2}) \frac{2p-1}{p(1-p)} + O(N_n^{-1/2}) = O(1). \qquad (11)
\end{aligned}
$$

Thus our two CPE terms are of orders $N_n^{-1} \asymp n^{-1}h^{-1}$ and h^2. This implies $h^* \asymp n^{-1/3}$ and that CPE is $O(n^{-2/3})$. But then the B_h^2 term is of order $h^4 N_n = h^5 n = n^{-2/3}$, so it must also be included. (The B_h^3 term is of order $h^6 N_n^{3/2}$, which is smaller.) Though the second term from the product rule derivative in the B_h^2 term is not zero this time, it is smaller-order and thus omitted.

The sign of $\mathrm{CPE}_{\mathrm{Bias}}$ is determined by the sign of $[B_h(2p-1) - B_h^2 N_n]$, as seen in (12). Since $B_h^2 N_n > 0$ always, if $B_h(2p-1) < 0$, then $\mathrm{CPE}_{\mathrm{Bias}} < 0$ irrespective of h (only the magnitude of B_h depends on h, not the sign). If $B_h(2p-1) > 0$, as shown in the appendix, it is still always possible to choose h such that $\mathrm{CPE}_{\mathrm{Bias}} < 0$ and cancels with $\mathrm{CPE}_{\mathrm{GK}}$. Since such a solution is always possible, we pick h to equate

$$
\begin{aligned}
&-N_n^{-1} z_{1-\alpha/2} \frac{\epsilon_h(1-\epsilon_h)+\epsilon_\ell(1-\epsilon_\ell)}{p(1-p)} \phi(z_{1-\alpha/2}) \\
&\quad \doteq B_h \left[f_{\hat{Q}_{Y|C_h}^{I,u_h}}\big(Q_{Y|C_h}(p)\big) - f_{\hat{Q}_{Y|C_h}^{I,u_\ell}}\big(Q_{Y|C_h}(p)\big) \right] \\
&\qquad + (1/2) B_h^2 \left[f'_{\hat{Q}_{Y|C_h}^{I,u_\ell}}\big(Q_{Y|C_h}(p)\big) - f'_{\hat{Q}_{Y|C_h}^{I,u_h}}\big(Q_{Y|C_h}(p)\big) \right] \\
&\quad \doteq f_{Y|X}(\xi_p;0) z_{1-\alpha/2} \phi(z_{1-\alpha/2}) [p(1-p)]^{-1} \big[B_h(2p-1) - B_h^2 N_n \big]. \qquad (12)
\end{aligned}
$$

Note that $2p-1 > 0$ is equivalent to $p > 1/2$, and that $p = 1/2$ zeroes the B_h term. Continuing to solve for h leads to the plug-in bandwidth given.

Once $\hat{h}$ is determined, for any of these cases, C_h can be constructed, and then the one-sample unconditional quantile inference method can be applied to $\{Y_i : X_i \in C_h\}$.

5 Simulation study

Code for the inference function in R will be available on the author's website, and simulation code in R is available upon request.

5.1 Computation of plug-in bandwidth

For our plug-in bandwidth, we need to estimate five objects. Consistent estimators exist for all five. For clarity, we now explicitly write x_0 as the point of interest, instead of taking $x_0 = 0$; we also take $d = 1$, $b = 2$, and focus on two-sided inference, both pointwise and joint.

For $f_{Y|X}(\xi_p; x_0)$, a kernel estimator such as `npcdens` from package `np` (Hayfield and Racine, 2008) can be used given an estimate $\hat{\xi}_p$, which can be computed using any nonparametric conditional quantile estimator. To estimate $\hat{\xi}_p$, we use `rq` from package `quantreg` (Koenker, 2012) with a cubic B-spline basis generated by `bs` from package `splines` (R Core Team, 2012).

For $f_X(x_0)$, any kernel density estimator will suffice, such as `kde` from package `ks` (Duong, 2012). From the same package, `kdde` can be used to estimate the density derivative $f'_X(x_0)$. Both functions work for up to six-dimensional X data.

Alternatively, we could use the Gaussian plug-in assumption for $f_X(x_0)$ and $f'_X(x_0)$ directly in $\hat{h}$, instead of estimating them. For the density derivative, another popular option is the local polynomial approach as in Fan and Gijbels (1996, pp. 50–52).

For $F^{(0,1)}_{Y|X}(\xi_p; x_0)$ and $F^{(0,2)}_{Y|X}(\xi_p; x_0)$, they enter a Taylor expansion of $G(\cdot) \equiv F_{Y|X}(\xi_p; \cdot)$

$$G(x) = G(x_0) + G'(x_0)(x - x_0) + (1/2)G''(x_0)(x - x_0)^2 + (1/6)G'''(\tilde{x})(x - x_0)^3,$$

where $\tilde{x}$ is between x and x_0.[2] Also, $F_{Y|X}(\xi_p; X = x) = E(1\{Y_i \le \xi_p\} \mid X = x)$, a mean regression of $1\{Y_i \le \xi_p\}$ on X. Correspondingly, we use `lm` from package `stats` (R Core Team, 2012) to fit a cubic B-spline generated by `bs` from package `splines` (R Core Team, 2012) and compute the first two derivatives at a grid of points using `mybs` (Weisberg, 2012). The degrees of freedom are chosen to match the effective degrees of freedom automatically selected by penalized spline function `qsreg` from package `fields` (Furrer et al., 2012).

[2] A quadratic would also give a consistent estimator since $\tilde{x} \to x_0$ as $n \to \infty$, but may be worse in finite samples, depending also on local smoothness.

In finite samples, sampling error can make the plug-in bandwidth differ from the infeasible version, and in turn the approximation error from the Taylor expansion (around x_0) used to calculate the bias can make even the infeasible bandwidth bigger or smaller than optimal. Since an excessively large bandwidth leads to under-coverage, we implement a "reality check" adjustment to our plug-in bandwidths. For points $x_1 < x_2$, the C_h window for x_2 cannot extend to the left of the window for x_1, and likewise the x_1 window cannot extend to the right of the x_2 window. This essentially lets information about local quantile function derivatives be shared among the points of interest. For example, imagine that the quantile function is estimated to be very smooth (small derivatives) at x_2, leading to a large bandwidth, but estimated to be highly variable (big derivatives) at x_1, leading to a small bandwidth. That additional information from x_1 should cause the x_2 window to shrink until its left edge matches that of the x_1 window (and arguably farther), rather than letting x_2 continue to blithely imagine that the function extends smoothly as far as the eye can see. We did not experiment with different (possibly data dependent) magnitudes or types of adjustment, opting simply for the intuitive version above, but more sophisticated ones may exist.

5.2 Results

For comparison, we show two approaches available in the popular `quantreg` package in R (Koenker, 2012). First, the function `rqss` ("regression quantile smoothing spline") is used as on page 10 of its vignette, with the Schwarz (1978) information criterion (SIC) for model selection. Both pointwise and uniform confidence bands are generated with `plot.rqss`. Second, the function `rq` is used with a cubic B-spline generated by `bs`, again with SIC model selection. The function `predict.rq` with `type="percentile"` then generates pointwise confidence intervals using a bootstrap. Similar but consistently worse results were obtained from using the same approach but with `type="direct"` to use the analytic method, so only the bootstrap version is presented. Joint CIs may be generated using a Bonferroni approach.

A third approach was tried, using `npqreg` from package `np` (Hayfield and Racine, 2008) and `boot` and `boot.ci` from package `boot` (R Core Team, 2012), with both percentile and adjusted percentile bootstraps (`"perc"` and `"bca"` types). In preliminary simulations, with `bwmethod="normal-reference"` (rule of thumb), the computation time was still a factor of ten larger (with 999 bootstrap replications), and over-smoothing caused severe under-coverage. With `bwmethod="cv.ml"` (likelihood-based cross-validation), the computation time was prohibitive, running for over three hours on the first simulation replication alone (which was stopped before completion). Consequently, this approach is omitted from the results.

A bias-corrected version of our IDEAL method is also omitted. Compared to the uncorrected IDEAL, it performed extremely similarly, usually the same or slightly worse. Median (over all simulation replications) confidence bands appeared marginally more centered around the true conditional quantile function, but the additional variance incurred from bias estimation likely caused the observed (slight) under-coverage.

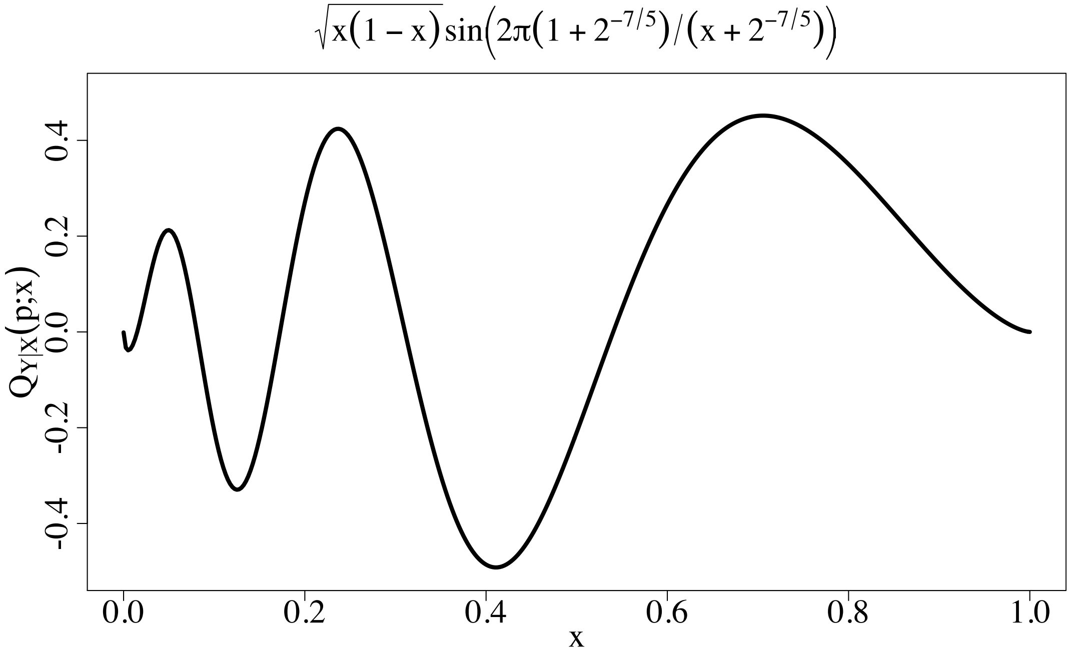

Figure 2: True conditional median function for simulations.

Our simulations repeat the simulation setup of the `rqss` vignette in Koenker (2012), which in turn was taken in part from Ruppert et al. (2003, §17.5.1). Parameters were set to $n = 400$, $p = 1/2$, $d = 1$, and $\alpha = 0.05$. Scalar $X_i \overset{iid}{\sim} \text{Unif}(0, 1)$,

$$Y_i = \sqrt{X_i(1 - X_i)}\sin\left(\frac{2\pi(1 + 2^{-7/5})}{X_i + 2^{-7/5}}\right) + \sigma(X_i)U_i,$$

where the U_i are iid Gaussian, t_3, Cauchy, or centered χ^2_3, and $\sigma(X) = 0.2$ or $\sigma(X) = 0.2(1 + X)$. The conditional median function is shown in Figure 2.

With all eight DGPs (four error distributions, homoskedastic or heteroskedastic), our IDEAL method had the most consistent accuracy over all points on the conditional median function, as shown in the first two columns of Figure 3. Coverage probability is near nominal for all x_0, all distributions, and in the presence or absence of heteroskedasticity.

In contrast, the other two methods are subject to under-coverage for a variety of reasons, as well as some over-coverage. Near $X = 0$, the true conditional median function varies

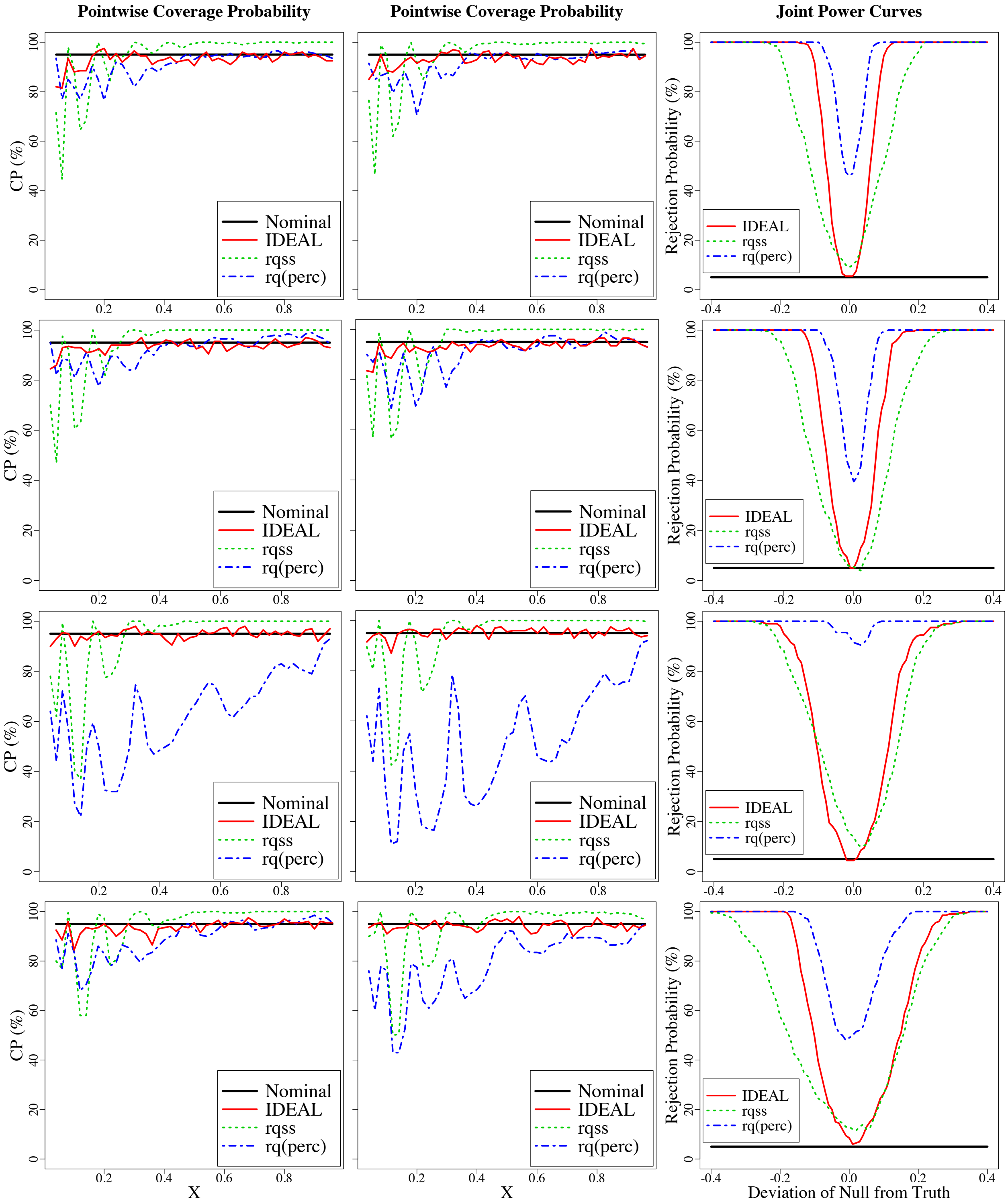

Figure 3: Pointwise coverage probabilities by X (first two columns) and joint power curves (third column), for conditional median 95% confidence intervals, $n = 400$, $X_i \overset{iid}{\sim} \text{Unif}(0,1)$, $Y_i = \sqrt{X_i(1-X_i)}\sin\bigl[2\pi(1+2^{-7/5})/(X_i+2^{-7/5})\bigr] + \sigma(X_i)U_i$. Distributions of U_i are, top row to bottom row: standard normal, t_3, Cauchy, and centered χ^2_3. Columns 1 & 3: homoskedastic, $\sigma(x) = 0.2$. Column 2: heteroskedastic, $\sigma(x) = (0.2)(1+x)$.

rapidly, and then it smooths out as X increases toward one. As seen in the first two columns of Figure 3, the rqss confidence intervals are too narrow when the function varies more, and too wide when the function varies less. This pattern holds for all eight DGPs. The under-coverage (as low as 40–60% CP depending on the error distribution) is quite significant for x_0 closer to zero, and over-coverage (as high as 100% CP for all distributions) occurs for most $x_0 \geq 0.5$. The rq approach avoids the over-coverage, but it can suffer even more severe under-coverage, as shown. This is particularly true with Cauchy errors (third row in figure), where there is under-coverage for all x_0 and CP dips below 30% with homoskedasticity and even below 10% with heteroskedasticity.

The easiest way to construct IDEAL joint CIs is by the Bonferroni approach. For example, when $\alpha = 0.05$ to give a 95% confidence level, if there are 47 points of interest x_0, pointwise CIs are constructed with $\alpha/47$ instead of α. Alternatively, instead of the Bonferroni $\alpha/47$, an adjusted value for α can be backed out from the uniform confidence bands provided by rqss, which uses a Hotelling tube approach. This gives extremely similar results in our simulations, so it has been omitted for simplicity.

Joint power curves are given in the third column of Figure 3. The x-axis of the graphs indicates the deviation of the null hypothesis from the true curve; for example, -0.1 refers to a test against a curve lying 0.1 below the true curve (at all X), and zero means the null hypothesis is true. The IDEAL joint CP is again very close to nominal (since the test's size is close to α) under all four error distributions. Heteroskedastic versions were similar and thus omitted. The rqss method has CP relatively close to nominal (within 10%); it appears that the too-wide part nearly balances the too-narrow part in these examples. However, IDEAL has significantly better power. (Note that this is true without size-adjusting power, even though rqss is size-distorted and IDEAL is not.) The power advantage for the asymmetric χ^2_3 distribution is much stronger for negative deviations, seemingly due to IDEAL's superior ability to adapt to skewed distributions. The rq method has significant under-coverage in all cases.

Figures 4 and 5 show pointwise power, by X, against points differing from the true conditional median by ± 0.1. This is calculated as the percentage of simulation replications where a method's CI excluded the point 0.1 below the true conditional median, averaged with the value for the point 0.1 above. To make the comparison fair, points with CP below 90% are omitted.

The rq method seems to have the best power (narrowest CIs) when it does succeed in controlling CP, though the difference with IDEAL is small for x_0 in the upper quartile of X. With very smooth quantile functions, like a constant function (in the extreme), and with error distributions close to normal, the rq method would be preferred since it controls

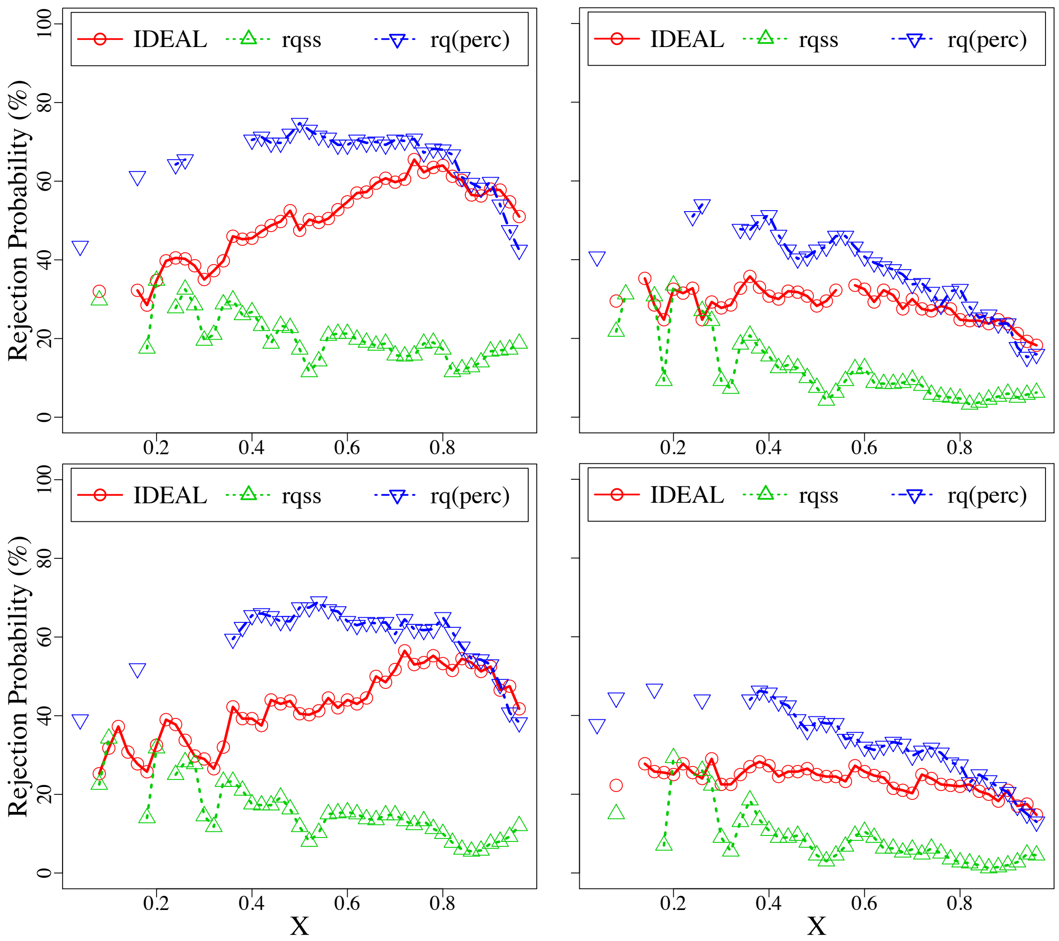

Figure 4: Pointwise power by X, against deviations of magnitude 0.1 (half negative, half positive), for conditional median 95% confidence intervals, $n = 400$, $X_i \overset{iid}{\sim} \text{Unif}(0,1)$, $Y_i = \sqrt{X_i(1-X_i)}\sin\left[2\pi(1+2^{-7/5})/(X_i+2^{-7/5})\right] + \sigma(X_i)U_i$. Points with CP below 90% not plotted. Top row: $U_i \overset{iid}{\sim} N(0,1)$. Bottom row: $U_i \overset{iid}{\sim} t_3$. Left column: homoskedastic, $\sigma(x) = 0.2$. Right column: heteroskedastic, $\sigma(x) = (0.2)(1+x)$.

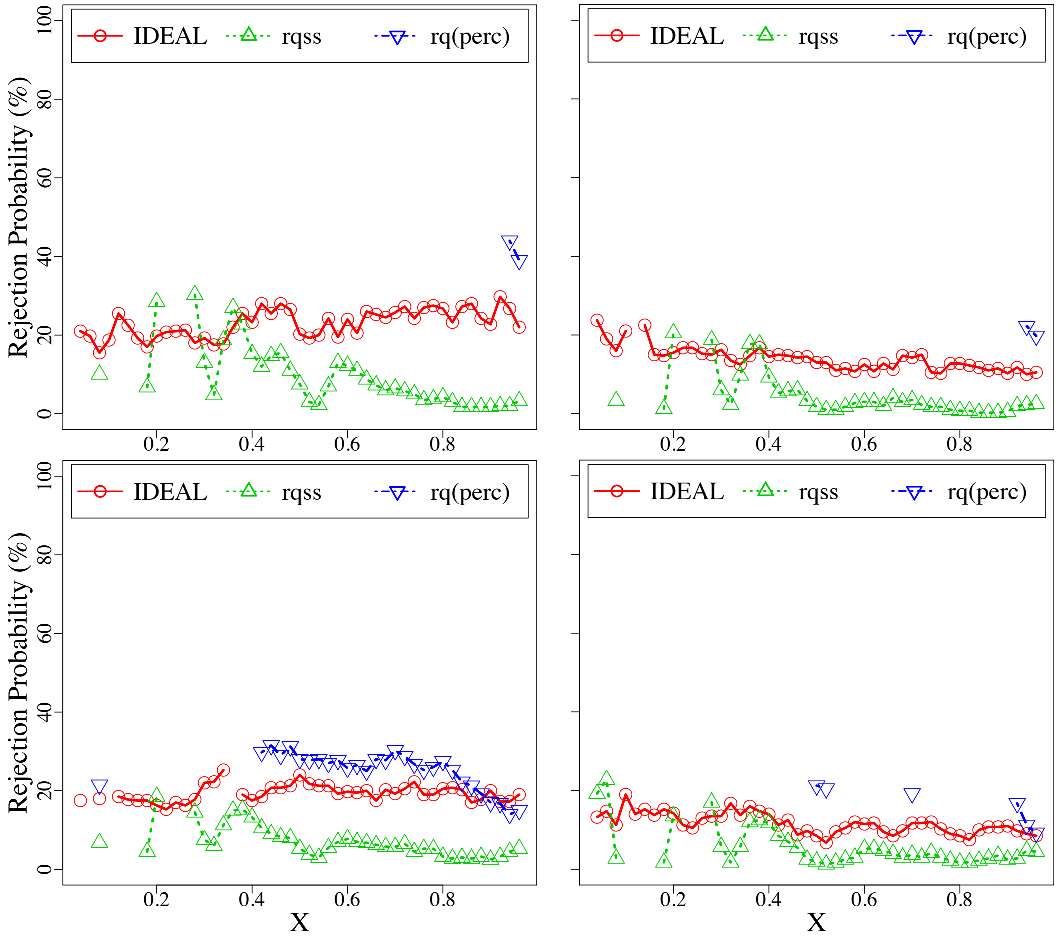

Figure 5: Pointwise power by X; same as Figure 4 but with Cauchy errors for top row and centered χ^2_3 for bottom row.

CP and has better power. Recall that rq is a parametric method, so this is essentially just saying that when the parametric model is (nearly) properly specified, inference will generally be more precise. However, barring a model selection method capable of reliably identifying situations where rq would be better (given that the quantile function and error distribution are both unknown), IDEAL provides a robust inference option with good power.

The IDEAL method is generally more powerful than rqss, as already hinted at in the pointwise CP graphs. Among the eight DGPs shown in Figures 4 and 5, IDEAL is more powerful by at least 10–20 percentage points for most x_0 and most DGPs, and by over 40 percentage points in some cases.

Using the same setup but with errors following an exponential distribution (with or without heteroskedasticity), the results look similar to a mix of the standard normal and χ^2_3 results. One difference is that the joint CP for rqss is almost down to 80%, while the IDEAL joint CP is near 90%. Using the standard normal errors but instead looking at the conditional upper quartile, results were again similar other than worse CP for the joint intervals: rqss was again around 80%, as was the Bonferroni IDEAL, while the rqss/Hotelling-tube-aided IDEAL was near 90%. In this case, bias correction actually improved joint coverage by a few percentage points. Moving up to the even more difficult $\Phi(1) \approx 0.84$-quantile of a standard normal, the joint IDEAL coverage is the same, but rqss worsens to below 70% and suffers power loss against the -0.1 and -0.2 alternatives. The pointwise CP is actually best for rq in this case, though the joint CP for rq is below 60%.

Overall, the simulation results show IDEAL to be significantly more accurate than rq and rqss. IDEAL always has CP near the nominal level, for both pointwise and joint CIs, even when both rq and rqss show severe under-coverage. IDEAL hypothesis tests also have better power than rqss in most cases, against both pointwise and joint alternatives.

5.3 Computation time

The simplicity of the IDEAL method leads to significant computational advantages over existing methods. To demonstrate this, we ran simulations with the same DGP as before, with homoskedastic Gaussian errors, varying the sample size over a wide range. In addition to the IDEAL method, rqss (and rqss.predict) is used to generate pointwise CIs. We also calculate a lower bound time for a bootstrap method with 99 bootstrap replications by multiplying the rqss estimation time by 99. (Calculating the optimal smoothing parameter takes most of the time, rather than the spline fit itself.) This provides a strict lower bound because more replications (possibly 10 or 100 times more) may be needed to ensure bootstrap accuracy, and additional computation is required beyond just the estimator.

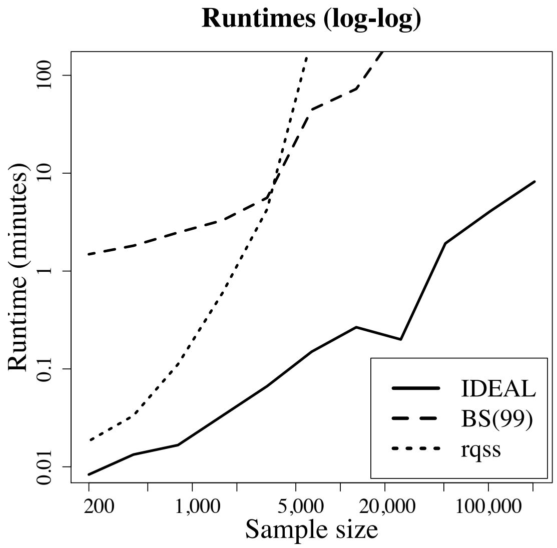

Figure 6: Computation time for different methods, as a function of sample size. DGP is the same as previous simulations, with homoskedastic Gaussian errors. Bootstrap times are for 99 bootstrap replications, estimated by multiplying the estimation time for `rqss` by 99; this gives a lower bound on the total bootstrap time.

Figure 6 shows computation times for these three methods as functions of sample size. For $n = 200$, 99 bootstrap replications take more than a minute to run, while IDEAL runs in less than a second. For $n = 3200$, both bootstrap and `rqss` take a few minutes to run, while IDEAL takes only a few seconds. For $n = 6400$, the `rqss` computation was stopped after two hours without having finished; 99 bootstrap replications would take at least 20 minutes, and 999 replications would take over three hours. Even with only 99 bootstrap replications, $n = 25{,}600$ still takes over five hours. In contrast, with the much larger $n = 204{,}800$ sample, IDEAL runs in under ten minutes. In addition to being orders of magnitude smaller at these sample sizes, the IDEAL runtime appears to scale roughly proportional to the sample size, whereas the other methods' runtimes appear to increase more rapidly. If these relationships continue to hold at larger sample sizes, the IDEAL computational advantage will grow proportionally bigger as the sample size increases.

6 Empirical application

This section contains an empirical application of the IDEAL confidence intervals for conditional quantiles. It may be replicated using the two R files and data files available at the author's website.

Measurements of hemoglobin concentration in the bloodstream are commonly used to screen for anemia, which in turn may indicate iron deficiency, a nutritional deficiency affecting well over 1 in 4 people worldwide (Khusun et al., 1999). Examining how quantiles of the hemoglobin distribution vary with different conditioning variables is of more interest than how the mean varies because it is the lower quantiles that indicate different degrees of anemia. Wave 4 (2007) of the Indonesian Family Life Survey (IFLS) contains measurements of hemoglobin concentration, among numerous other variables. The World Health Organization (WHO) has suggested threshold values[3] for "mild" anemia (acknowledging that even this level is a serious health concern), "moderate" anemia, and "severe" anemia, which seem valid at least for the (nonrandom) subsample of Indonesians examined in Khusun et al. (1999). These thresholds vary by age, sex, and pregnancy, so the raw IFLS values were scaled such that the threshold between non-anemia and mild anemia is 13 grams per deciliter (g/dl) for all individuals. Additionally, household per capita annual expenditure was computed from the raw IFLS data by summing the values in various categories of expenditure (scaled to annual for each) and dividing by household size; the natural log was then taken for more appropriate scaling. The highest level of education taken by the head of household was linked to the hemoglobin and expenditure data when available. The full Stata code for constructing the dataset starting from the raw IFLS data is available on the author's website.[4]

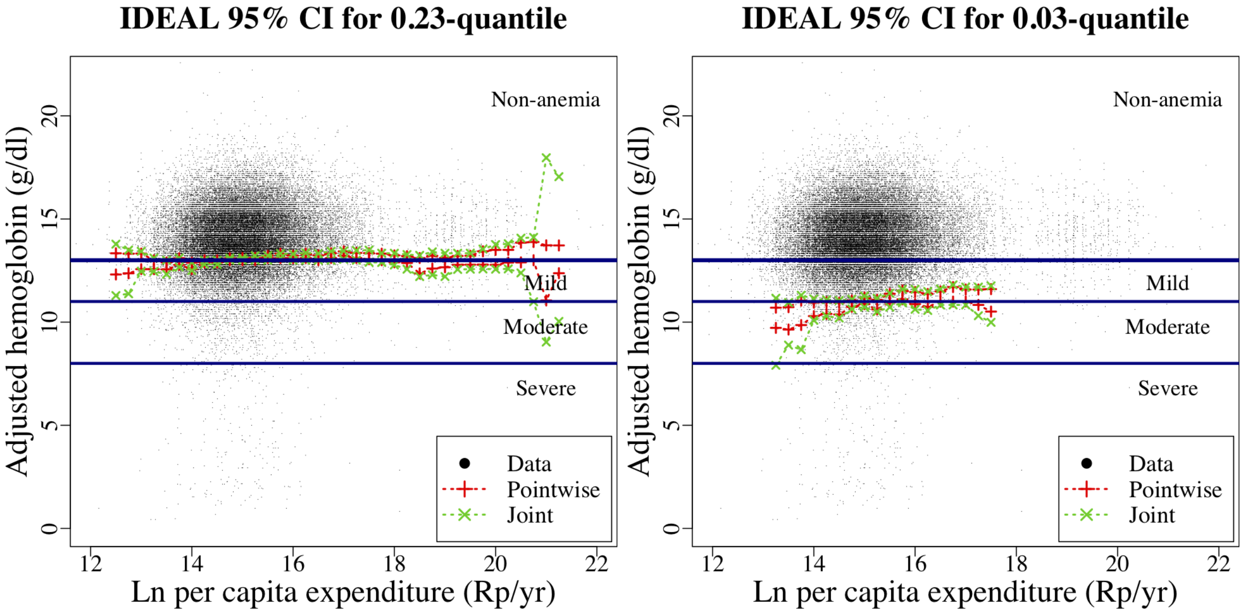

Figure 7: IDEAL 95% confidence intervals for 0.23-quantile (left) and 0.03-quantile (right) of adjusted hemoglobin concentration conditional on log per capital household expenditure.

[3] http://www.who.int/vmnis/indicators/haemoglobin.pdf

[4] The relationship between hemoglobin and expenditure (and education) is also examined in Li et al. (2013), which inspired this example.

Figure 7 shows the relationship between household per capita annual expenditure and adjusted hemoglobin concentration. For the lowest levels of expenditure, the IDEAL pointwise confidence intervals for the conditional 0.23-quantile (left panel) of hemoglobin fall mostly into the range for mild anemia, while the intervals for higher expenditure levels are mostly in the range for no anemia. The joint intervals are also tight in this case (and increasing in expenditure), rejecting any null hypothesis with a constant hemoglobin value for all levels of expenditure. The IDEAL confidence intervals are also helpful for seeing that the smaller cloud of data in the 18–22 log expenditure range is likely a decimal place error, because the intervals with expenditure 18 reset to the levels at the far left of the plot, and then slowly rise again. Note that $\ln(100) = 4.6$, so erroneously adding two zeroes to the expenditure would turn 13.4 into 18, 16 into 20.6, etc. In the right panel, we see similar results for the conditional 0.03-quantile and the mild/moderate threshold.

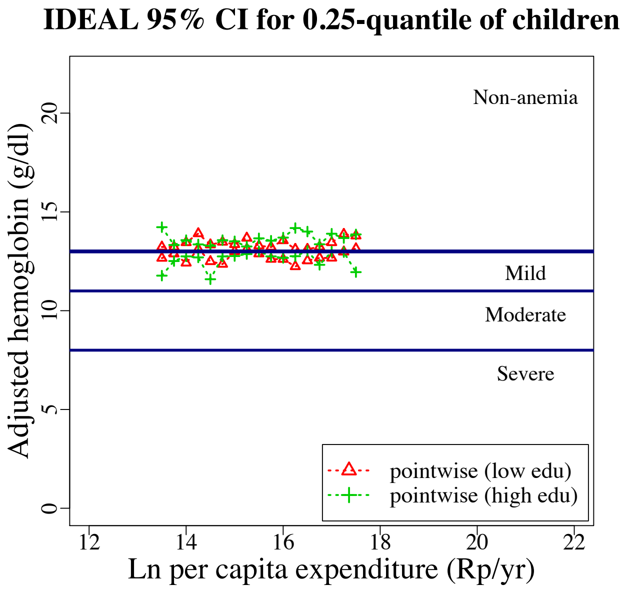

Figure 8: IDEAL 95% confidence intervals for 0.25-quantile of adjusted hemoglobin concentration conditional on log per capital household expenditure, by education level of head of household, for children 15 years old and younger. “High” education is defined as having attended at least some “senior high” education; “low” is no more than junior high.

Figure 8 restricts attention to children (15 years old or younger) and additionally conditions on a discrete (binary) variable for the educational attainment of the head of household. Having attended at least some “senior high” school is labeled “high,” while having attended only elementary or junior high is labeled “low.” The confidence intervals are somewhat wider than before due to restricting the sample to children in households whose head’s educational attainment was recorded, but they are still relatively tight. Still, there does not appear to be any systematic difference between the high and low groups over any range of expenditure.

Depending on the survey sampling scheme, the iid assumption may not quite hold in this case. It would be of great interest to extend the IDEAL method and its underlying fractional order statistic theory to account for sampling weights and/or various structures of dependence among observations.

7 Conclusion

We have provided a new method for inference on conditional quantiles, embedding the IDEAL theory of Goldman and Kaplan (2012) into a local smoothing context. Under mild smoothness assumptions, the two-sided coverage probability error is $O(n^{-2/(2+d)})$ for a d-dimensional conditioning vector. This is always better than conventional inference from asymptotic normality or bootstrap for $d \leq 2$, as well as for all $d \geq 3$ unless at least four (or more, depending on d) derivatives of the unknown function are correctly assumed and the corresponding local polynomial with hundreds or thousands of terms is fit. We also provide joint (over many values of X) confidence intervals. Our feasible plug-in bandwidth translates the superior theoretical properties into practice; simulations show improvements in size control and power over popular existing techniques, and R code for our new method is publicly available.

This paper develops a framework for successfully pushing unconditional IDEAL results through to a conditional context. This framework should also accommodate a conditional version of Kaplan (2011) for cases where Hutson (1999) cannot be applied (smaller samples and/or quantiles closer to zero or one). We plan to extend the unconditional two-sample IDEAL inference of Goldman and Kaplan (2012) to a conditional setting.

References

Abrevaya, J. (2001). The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics*, 26(1):247–257.

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica*, 74(2):539–563.

Bahadur, R. R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37(3):577–580.

Belluzzo, Jr., W. (2004). Semiparametric approaches to welfare evaluations in binary response models. *Journal of Business & Economic Statistics*, 22(3):322–330.

Bhattacharya, P. K. and Gangopadhyay, A. K. (1990). Kernel and nearest-neighbor estimation of a conditional quantile. *Annals of Statistics*, 18(3):1400–1415.

Bloch, D. A. and Gastwirth, J. L. (1968). On a simple estimate of the reciprocal of the density function. *The Annals of Mathematical Statistics*, 39(3):1083–1085.

Buchinsky, M. (1994). Changes in the U.S. wage structure 1963-1987: Application of quantile regression. *Econometrica*, 62(2):405–458.

Cade, B., Terrell, J., and Schroeder, R. (1999). Estimating effects of limiting factors with regression quantiles. *Ecology*, 80(1):311–323.

Chamberlain, G. (1994). Quantile regression, censoring, and the structure of wages. In *Advances in Econometrics: Sixth World Congress*, volume 2, pages 171–209.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *Annals of Statistics*, 19(2):760–777.

Chernozhukov, V., Hansen, C., and Jansson, M. (2009). Finite sample inference for quantile regression models. *Journal of Econometrics*, 152:93–103.

Chesher, A. (2003). Identification in nonseparable models. *Econometrica*, 71(5):1405–1441.

Duong, T. (2012). *ks: Kernel smoothing.* R package version 1.8.8.

Eide, E. and Showalter, M. H. (1998). The effect of school quality on student performance: A quantile regression approach. *Economics Letters*, 58(3):345–350.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, volume 66 of *Monographs on statistics and applied probability.* Chapman & Hall, London.

Furrer, R., Nychka, D., and Sain, S. (2012). *fields: Tools for spatial data.* R package version 6.6.3.

Gangopadhyay, A. K. and Sen, P. K. (1990). Bootstrap confidence intervals for conditional quantile functions. *Sankyā: the Indian Journal of Statistics, Series A*, 52(3):346–363.

Gneezy, U. and List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, 74(5):1365–1384.

Goldman, M. and Kaplan, D. M. (2012). IDEAL quantile inference via interpolated duals of exact analytic L-statistics. Working paper.

Guerre, E. and Sabbah, C. (2012). Uniform bias study and Bahadur representation for local polynomial estimators of the conditional quantile function. *Econometric Theory*, 28(1):87–129.

Hall, P., Wolff, R. C. L., and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94(445):154–163.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5).

Hogg, R. (1975). Estimates of percentile regression lines using salary data. *Journal of the American Statistical Association*, 70(349):56–59.

Hotelling, H. (1939). Tubes and spheres in n-space and a class of statistical problems. *American Journal of Mathematics*, 61:440–460.

Hutson, A. D. (1999). Calculating nonparametric confidence intervals for quantiles using fractional order statistics. *Journal of Applied Statistics*, 26(3):343–353.

Hyndman, R., Bashtannyk, D., and Grunwald, G. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5(4):315–336.

Kaplan, D. M. (2011). Improved population and two-sample quantile inference via fixed-smoothing asymptotics and Edgeworth expansion. Working paper.

Khusun, H., Yip, R., Schultink, W., and Dillon, D. (1999). World Health Organization hemoglobin cut-off points for the detection of anemia are valid for an Indonesian population. *The Journal of Nutrition*, 129(9):1669–1674.

Koenker, R. (2005). *Quantile regression*, volume 38 of *Econometric Society Monographs*. Cambridge university press.

Koenker, R. (2012). *quantreg: Quantile Regression*. R package version 4.81.

Koenker, R. and Mizera, I. (2004). Penalized triograms: total variation regularization for bivariate smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):145–163.

Kordas, G. (2006). Smoothed binary regression quantiles. *Journal of Applied Econometrics*, 21(3):387–407.

Li, Q., Lin, J., and Racine, J. S. (2013). Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. *Journal of Business & Economic Statistics*. Forthcoming.

Manning, W., Blumberg, L., and Moulton, L. (1995). The demand for alcohol: the differential response to price. *Journal of Health Economics*, 14(2):123–148.

Nayyar, G. (2009). Demand for services in India: A mirror image of Engel's law for food? Working Paper 451, University of Oxford. Department of Economics Working Paper Series.

Portnoy, S. (2012). Nearly root-N approximation for regression quantile processes. *Annals of Statistics*. Forthcoming.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Siddiqui, M. M. (1960). Distribution of quantiles in samples from a bivariate population. *J. Research of the NBS, B. Math. and Math. Physics*, 64B(3):145–150.

Stone, C. J. (1980). Optimal rates of convergence for non-parametric regression. *Annals of Statistics*, 8:1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for non-parametric regression. *Annals of Statistics*, 10:1040–1053.

Weisberg, S. (2012). mybs. `http://users.stat.umn.edu/~sandy/courses/8053/Data/mybs.R`.

A Lemma 2 proof

The case $d = 1$ is covered here. Generalizing to $d > 1$ should yield the same order and type of terms in the bias, just with vectors for first derivatives and matrices for second derivatives. The h^d (instead of h) from the change of variables does not determine the bias (see below).

Case $b = 2$

By definition, $Q_{Y|C_h}(p)$ satisfies

$$p = \int_{C_h} \left\{ \int_{-\infty}^{Q_{Y|C_h}(p)} f_{Y|X}(y;x)dy \right\} f_{X|C_h}(x)dx, \tag{13}$$

where $f_{Y|X}(y;x) \equiv f_{Y,X}(y,x)/f_X(x)$ is the conditional PDF of Y given X evaluated at $Y = y$ and $X = x$. Similarly, $f_{X|C_h}(x) = f_X(x)/P(X \in C_h) = O(h^{-1})$ is the conditional PDF of X within C_h. Also by definition, $Q_{Y|X}(p;x)$ satisfies

$$p = \int_{-\infty}^{Q_{Y|X}(p;x)} f_{Y|X}(y;x)dy$$

for all x.

Decomposing the $\{\cdot\}$ term, (13) becomes

$$\begin{aligned} p &= \int_{C_h} \left\{ \int_{-\infty}^{Q_{Y|X}(p;x)} f_{Y|X}(y;x)dy + \int_{Q_{Y|X}(p;x)}^{Q_{Y|C_h}(p)} f_{Y|X}(y;x)dy \right\} f_{X|C_h}(x)dx \\ &= p + \int_{C_h} \left\{ \int_{Q_{Y|X}(p;x)}^{Q_{Y|C_h}(p)} f_{Y|X}(y;x)dy \right\} f_{X|C_h}(x)dx, \end{aligned}$$

implying

$$0 = \int_{C_h} \left\{ \int_{Q_{Y|X}(p;x)}^{Q_{Y|C_h}(p)} f_{Y|X}(y;x)dy \right\} f_{X|C_h}(x)dx.$$

With change of variables $x = wh$, and converting $f_{X|C_h}(\cdot)$ to $f_X(\cdot)$,

$$0 = \frac{h}{P(X \in C_h)} \int_{-1}^{1} \left\{ \int_{Q_{Y|X}(p;wh)}^{Q_{Y|C_h}(p)} f_{Y|X}(y;wh)dy \right\} f_X(wh)dw.$$

With enough smoothness, a Taylor expansion around $w = 0$ can be taken of the integrand. An explicit expression for the bias comes out of the 0th order term; the 1st order terms will zero out; and the 2nd order terms will additionally determine the bias, through $O(h^2)$ terms. In the following calculations, we use the derivative rule

$$\frac{\partial}{\partial x} \int_{a(x)}^{b} f(y,x)dy = -f(a(x),x)a'(x) + \int_{a(x)}^{b} \frac{\partial}{\partial x} f(y,x)dy.$$

Schematically, we are taking a second-order Taylor expansion of

$$\left\{ \int_{a(x)}^{b} f(y;x)dy \right\} g(x)$$

around $x = 0$, which would be

$$\begin{aligned} &\int_{a(0)}^{b} f(y;0)dy\, g(0) \\ &+ x\left\{ \int_{a(0)}^{b} f(y;0)dy\, g'(0) + \left[\int_{a(0)}^{b} \frac{\partial}{\partial x} f(y;x)\bigg|_{x=0} dy - f(a(0);0)a'(0) \right] g(0) \right\} \\ &+ (x^2/2)\left\{ \int_{a(0)}^{b} f(y;0)dy\, g''(0) + \left[\int_{a(0)}^{b} \frac{\partial}{\partial x} f(y;x)\bigg|_{x=0} dy - f(a(0);0)a'(0) \right] g'(0) \right. \end{aligned}$$

$$
\begin{aligned}
&+\left[\int_{a(0)}^{b} \frac{\partial}{\partial x} f(y;x)\bigg|_{x=0} dy - f(a(0),0)a'(0)\right] g'(0) \\
&+\left[\int_{a(0)}^{b} \frac{\partial^2}{\partial x^2} f(y;x)\bigg|_{x=0} dy - \frac{\partial}{\partial x} f(a(0);x)\bigg|_{x=0} a'(0)\right. \\
&\qquad \left. - f(a(0);0)a''(0) - \left(\frac{\partial}{\partial y} f(y;0)\bigg|_{y=a(0)} a'(0) + \frac{\partial}{\partial x} f(a(0);x)\bigg|_{x=0}\right) a'(0)\right] g(0) \Bigg\}
\end{aligned}
$$

$$
+\, o(x^2).
$$

For compactness, we define

$$
Q^{(0,1)}_{Y|X}(p;0) \equiv \frac{\partial}{\partial x} Q_{Y|X}(p;x)\bigg|_{x=0}, \quad Q^{(0,2)}_{Y|X}(p;0) \equiv \frac{\partial^2}{\partial x^2} Q_{Y|X}(p;x)\bigg|_{x=0}, \quad \xi_p \equiv Q_{Y|X}(p;0),
$$

$$
f^{(0,1)}_{Y|X}(y;0) \equiv \frac{\partial}{\partial x} f_{Y|X}(y;x)\bigg|_{x=0}, \qquad f^{(1,0)}_{Y|X}(\xi_p;0) \equiv \frac{\partial}{\partial y} f_{Y|X}(y;0)\bigg|_{y=\xi_p}.
$$

The Taylor expansion of

$$
\left\{\int_{Q_{Y|X}(p;wh)}^{Q_{Y|C_h}(p)} f_{Y|X}(y;wh)dy\right\} f_X(wh)
$$

around $w=0$ yields

$$
\begin{aligned}
&\left\{\int_{\xi_p}^{Q_{Y|C_h}(p)} f_{Y|X}(y;0)dy\right\} f_X(0) \\
&+ wh\left\{\int_{\xi_p}^{Q_{Y|C_h}(p)} f_{Y|X}(y;0)dy\, f'_X(0) + \int_{\xi_p}^{Q_{Y|C_h}(p)} f^{(0,1)}_{Y|X}(y;0)dy\, f_X(0)\right. \\
&\qquad \left. - f_{Y|X}(\xi_p;0)Q^{(0,1)}_{Y|X}(p;0)f_X(0)\right\} \\
&+ \frac{(wh)^2}{2}\left\{\int_{\xi_p}^{Q_{Y|C_h}(p)} f_{Y|X}(y;0)dy\, f''_X(0)\right. \\
&\qquad + 2\left[\int_{\xi_p}^{Q_{Y|C_h}(p)} f^{(0,1)}_{Y|X}(y;0)dy - f_{Y|X}(\xi_p;0)Q^{(0,1)}_{Y|X}(p;0)\right] f'_X(0) \\
&\qquad + \left[\int_{\xi_p}^{Q_{Y|C_h}(p)} f^{(0,2)}_{Y|X}(y;0)dy - f_{Y|X}(\xi_p;0)Q^{(0,2)}_{Y|X}(p;0)\right] f_X(0) \\
&\qquad \left. - 2f^{(0,1)}_{Y|X}(\xi_p;0)Q^{(0,1)}_{Y|X}(p;0)f_X(0) - f^{(1,0)}_{Y|X}(\xi_p;0)\left[Q^{(0,1)}_{Y|X}(p;0)\right]^2 f_X(0)\right\}
\end{aligned}
$$

$$
\begin{aligned}
&+ o(h^2) \\
&= A + hwB - (1/2)h^2w^2C + o(h^2),
\end{aligned}
$$

with A, B, and C implicitly defined.

To extract the bias from A, we also need to expand

$$
\begin{aligned}
\int_{\xi_p}^{Q_{Y|C_h}(p)} f_{Y|X}(y;0)dy &= \int_{\xi_p}^{Q_{Y|C_h}(p)} \Big[f_{Y|X}(\xi_p;0) + f_{Y|X}^{(1,0)}(\tilde{y};0)(y-\xi_p)\Big]dy \\
&= f_{Y|X}(\xi_p;0)\big[Q_{Y|C_h}(p) - \xi_p\big] + O\Big(\big[Q_{Y|C_h}(p) - \xi_p\big]^2\Big),
\end{aligned}
$$

where $\tilde{y}$ is determined by the mean value theorem and so located between the lower and upper limits of integration. Anticipating that the bias is $\big[Q_{Y|C_h}(p) - \xi_p\big] = O(h^2)$, the term A is then

$$
A = f_{Y|X}(\xi_p;0)\big[Q_{Y|C_h}(p) - \xi_p\big]f_X(0) + O\big(h^4\big).
$$

Since there is no w in A, $\int_{-1}^{1} Adw = 2A$.

The B term zeroes out since the only w in it is the w in hwB. The integral over $[-1,1]$ is thus $\int_{-1}^{1}(hwB)dw = hB\int_{-1}^{1} wdw = (hB)(0) = 0$.

For the C term, the only w is the w^2 in $(1/2)h^2w^2C$, so

$$
\int_{-1}^{1}(1/2)h^2w^2Cdw = (1/2)h^2C\int_{-1}^{1} w^2dw = (1/2)h^2C(2/3) = h^2C/3.
$$

Anticipating that the bias is $O(h^2)$, and assuming $f_{Y|X}^{(0,2)}(y;0)$ is bounded in a neighborhood of $y = \xi_p$, the three definite integrals in C are $O(h^2)$. Thus C simplifies to

$$
\begin{aligned}
C = {}& 2f_{Y|X}(\xi_p;0)Q_{Y|X}^{(0,1)}(p;0)f_X'(0) + f_{Y|X}(\xi_p;0)Q_{Y|X}^{(0,2)}(p;0)f_X(0) \\
&+ 2f_{Y|X}^{(0,1)}(\xi_p;0)Q_{Y|X}^{(0,1)}(p;0)f_X(0) + f_{Y|X}^{(1,0)}(\xi_p;0)\Big[Q_{Y|X}^{(0,1)}(p;0)\Big]^2 f_X(0) \\
&+ O(h^2).
\end{aligned}
$$

Combining the results for A, B, and C,

$$2f_{Y|X}(\xi_p;0)\bigl[Q_{Y|C_h}(p)-\xi_p\bigr]f_X(0)+O\bigl(h^4\bigr)=h^2C/3+O(h^4)+o(h^2),$$

so the bias is

$$\begin{aligned}
&Q_{Y|C_h}(p)-\xi_p\\
&=h^2\frac{C}{6f_{Y|X}(\xi_p;0)f_X(0)}+o(h^2)\\
&=\frac{h^2}{6}\biggl\{\Bigl[2Q^{(0,1)}_{Y|X}(p;0)f'_X(0)/f_X(0)+Q^{(0,2)}_{Y|X}(p;0)\Bigr]+2f^{(0,1)}_{Y|X}(\xi_p;0)Q^{(0,1)}_{Y|X}(p;0)/f_{Y|X}(\xi_p;0)\\
&\qquad+f^{(1,0)}_{Y|X}(\xi_p;0)\Bigl[Q^{(0,1)}_{Y|X}(p;0)\Bigr]^2/f_{Y|X}(\xi_p;0)\biggr\}+o(h^2).
\end{aligned}$$

This is identical to the bias in Bhattacharya and Gangopadhyay (1990, Thm. K1), just in different notation. By definition, for all x,

$$F_{Y|X}\bigl(Q_{Y|X}(p;x);x\bigr)=p.$$

Differentiating once with respect to x yields

$$\begin{aligned}
0&=Q^{(0,1)}_{Y|X}(p;x)f_{Y|X}\bigl(Q_{Y|X}(p;x);x\bigr)+F^{(0,1)}_{Y|X}\bigl(Q_{Y|X}(p;x);x\bigr),\\
Q^{(0,1)}_{Y|X}(p;x)&=-\frac{F^{(0,1)}_{Y|X}\bigl(Q_{Y|X}(p;x);x\bigr)}{f_{Y|X}\bigl(Q_{Y|X}(p;x);x\bigr)}.
\end{aligned}$$

Differentiating again with respect to x gives

$$\begin{aligned}
0&=Q^{(0,2)}_{Y|X}(p;x)f_{Y|X}\bigl(Q_{Y|X}(p;x);x\bigr)+F^{(0,2)}_{Y|X}\bigl(Q_{Y|X}(p;x);x\bigr)\\
&\quad+Q^{(0,1)}_{Y|X}(p;x)f^{(0,1)}_{Y|X}\bigl(Q_{Y|X}(p;x);x\bigr)+\Bigl[Q^{(0,1)}_{Y|X}(p;x)\Bigr]^2f^{(1,0)}_{Y|X}\bigl(Q_{Y|X}(p;x);x\bigr)\\
&\quad+f^{(0,1)}_{Y|X}\bigl(Q_{Y|X}(p;x);x\bigr)Q^{(0,1)}_{Y|X}(p;x),\\
Q^{(0,2)}_{Y|X}(p;x)&=-\frac{1}{f_{Y|X}\bigl(Q_{Y|X}(p;x);x\bigr)}\\
&\quad\times\biggl\{F^{(0,2)}_{Y|X}\bigl(Q_{Y|X}(p;x);x\bigr)+2Q^{(0,1)}_{Y|X}(p;x)f^{(0,1)}_{Y|X}\bigl(Q_{Y|X}(p;x);x\bigr)\\
&\qquad+\Bigl[Q^{(0,1)}_{Y|X}(p;x)\Bigr]^2f^{(1,0)}_{Y|X}\bigl(Q_{Y|X}(p;x);x\bigr)\biggr\}.
\end{aligned}$$

Plugging these substitutions into the original bias expression gives

$$\begin{aligned}
\text{Bias} = \frac{h^2}{6}\Bigg\{ &-\frac{2F^{(0,1)}_{Y|X}(\xi_p;0)f'_X(0)}{f_X(0)f_{Y|X}(\xi_p;0)} \\
&- \frac{1}{f_{Y|X}(\xi_p;0)}\Bigg\{F^{(2,0)}_{Y|X}(\xi_p;0) - 2F^{(0,1)}_{Y|X}(\xi_p;0)f^{(0,1)}_{Y|X}(\xi_p;0)/f_{Y|X}(\xi_p;0) \\
&\qquad\qquad + \Big[F^{(0,1)}_{Y|X}(\xi_p;0)\Big]^2 f^{(1,0)}_{Y|X}(\xi_p;0)/f_{Y|X}(\xi_p;0)\Bigg\} \\
&- \frac{2f^{(0,1)}_{Y|X}(\xi_p;0)F^{(0,1)}_{Y|X}(\xi_p;0)}{\Big[f_{Y|X}(\xi_p;0)\Big]^2} + \frac{f^{(1,0)}_{Y|X}(\xi_p;0)\Big[F^{(0,1)}_{Y|X}(\xi_p;0)\Big]^2}{\Big[f_{Y|X}(\xi_p;0)\Big]^3}\Bigg\} \\
= \frac{h^2}{6[f_{Y|X}(\xi_p;0)]^3}& \\
\times \Bigg\{ &-2F^{(0,1)}_{Y|X}(\xi_p;0)[f_{Y|X}(\xi_p;0)]^2 f'_X(0)/f_X(0) - F^{(0,2)}_{Y|X}(\xi_p;0)[f_{Y|X}(\xi_p;0)]^2 \\
&+ 2F^{(0,1)}_{Y|X}(\xi_p;0)f^{(0,1)}_{Y|X}(\xi_p;0)f_{Y|X}(\xi_p;0) - [F^{(0,1)}_{Y|X}(\xi_p;0)]^2 f^{(1,0)}_{Y|X}(\xi_p;0) \\
&- 2F^{(0,1)}_{Y|X}(\xi_p;0)f^{(0,1)}_{Y|X}(\xi_p;0)f_{Y|X}(\xi_p;0) + [F^{(0,1)}_{Y|X}(\xi_p;0)]^2 f^{(1,0)}_{Y|X}(\xi_p;0)\Bigg\} \\
= -h^2\frac{f_X(0)F^{(0,2)}_{Y|X}(\xi_p;0) + 2f'_X(0)F^{(0,1)}_{Y|X}(\xi_p;0)}{6f_X(0)f_{Y|X}(\xi_p;0)}&.
\end{aligned}$$

This is equivalent to the bias in Bhattacharya and Gangopadhyay (1990, Thm. K1) since their bandwidth is $h/2$, so $(h/2)^2/6 = h^2/24$, and with $x_0 = 0$ and

$$g(\xi) \equiv f_{Y|X}(\xi_p;0), \quad G_x(\xi \mid x_0) \equiv F^{(0,1)}_{Y|X}(\xi_p;0), \quad G_{xx}(\xi \mid x_0) \equiv F^{(0,2)}_{Y|X}(\xi_p;0).$$

Case $b < 2$

In the foregoing calculations, we (implicitly) assumed $s_X > 1$, $s_Q > 2$, and $s_Y > 1$. (While the brief appearance of $f''_X(0)$ means we technically assumed $s_X > 2$, this can easily be weakened to $s_X > 1$ since it appears in a smaller-order term.)

By examining the original Taylor expansion, we can see how the order of the bias will diminish as we relax these smoothness assumptions. We consider the above expansion $A = h^2w^2C/2 - hwB + o(h^2)$.

The A term is the bias plus a remainder depending on s_Y. Even if $s_Y \le 1$, the remainder is

$$\begin{aligned}&\int_{\xi_p}^{Q_{Y|C_h}(p)} \big[f_{Y|X}(y;0) - f_{Y|X}(\xi_p;0)\big](y-\xi_p)dy \\ &= O\big(\mathrm{Bias}^{2+s_Y}\big),\end{aligned}$$

which is always smaller than $O(\mathrm{Bias})$ if the bias goes to zero asymptotically. Thus, s_Y here does not have an effect on the order of the bias, and Ah is the bias times $O(1)$ terms not dependent on smoothness.

The C term includes many derivatives, but some of the terms are already smaller-order. Assumption A6 already requires $f_{Y|X}^{(0,2)}\big(Q_{Y|X}(p;0)\big)$ to be continuous in a neighborhood of p, so those terms may be ignored. Specifically, $f_X''(0)$ only appears in a term of order (h^2Bias), which is always smaller than the bias, so s_X has no binding effect here.

The three key terms in C involve $Q_{Y|X}^{(0,1)}(p;0)f_X'(0)$, $Q_{Y|X}^{(0,2)}(p;0)$, and $f_{Y|X}^{(1,0)}(\xi_p;0)\Big[Q_{Y|X}^{(0,1)}(p;0)\Big]^2$. If $s_X < 1$, the first term will be replaced by a larger-order term; if $s_Q < 2$, the same will happen for the second term; and if $s_Y < 1$, same for the third term, but we already need $s_Y > 2$ to apply Goldman and Kaplan (2012). Essentially, with less smoothness, we are forced to replace (for example) $g(x) = g(0) + g'(0)x + (1/2)g''(\tilde{x})x^2$ with $g(x) = g(0) + g'(\tilde{x})x$, where $\tilde{x}$ is determined by the mean value theorem. Before, $g''(\tilde{x}) = g''(0) + [g''(\tilde{x}) - g''(0)]$, and the absolute value of the $[\cdot]$ term is bounded as $|\tilde{x}|^\gamma$ by Hölder exponent γ. After relaxing the smoothness assumptions some, we get a similar expression but with $[g'(\tilde{x}) - g'(0)]$: the number of derivatives k is smaller by one, and there is some new γ. With less smoothness, the B terms do not integrate to zero since there are $\tilde{w}$ floating around, instead of just $\mathrm{Const} \times \int_{-1}^{1} w dw = 0$. The result is that the bias is of order h^b for $b = \min\{s_Q, 1+s_X, 1+s_Y\}$ for $s_Q \in [1,2)$ and $s_X, s_Y \in (0,1)$. As we saw originally, the biggest b can be is two, and relaxing s_Q further continues to decrease the order in the same pattern; so in all, the bias is

$$Q_{Y|C_h}(p) - \xi_p = O(h^b), \quad b = \min\{2, s_Q, 1+s_X, 1+s_Y\}.$$

B Plug-in bandwidth calculations

The plug-in bandwidth is for $d = 1$ and $b = 2$, in which case the bias is

$$
\begin{aligned}
Q_{Y|C_h}(p) - \xi_p = \frac{h^2}{6}\Bigg\{ & \Big[2Q^{(0,1)}_{Y|X}(p;0)f'_X(0)/f_X(0) + Q^{(0,2)}_{Y|X}(p;0)\Big] \\
& + 2f^{(0,1)}_{Y|X}(\xi_p;0)Q^{(0,1)}_{Y|X}(p;0)/f_{Y|X}(\xi_p;0) \\
& + f^{(1,0)}_{Y|X}(\xi_p;0)\Big[Q^{(0,1)}_{Y|X}(p;0)\Big]^2/f_{Y|X}(\xi_p;0)\Bigg\} + o(h^2) \\
= -h^2 & \frac{f_X(0)F^{(0,2)}_{Y|X}(\xi_p;0) + 2f'_X(0)F^{(0,1)}_{Y|X}(\xi_p;0)}{6f_X(0)f_{Y|X}(\xi_p;0)} + o(h^2).
\end{aligned}
$$

To avoid iteration (and nicely cancel some constants), we plug in $\epsilon_h = \epsilon_\ell = 0.2$ as a rule of thumb. The maximum bandwidth would be obtained using $\epsilon = 0.5$, which would give extremely similar bandwidths since, for example, $[(0.2)(0.8)]^{1/6} = 0.74$ while $[(0.5)(0.5)]^{1/6} = 0.79$ in the two-sided median case. The stabilizing effect and computational gains of using a fixed ϵ outweigh the small benefit of iteration. We also consider a Gaussian plug-in assumption for $f_{Y|X}(\xi_p;0)$ and $f^{(1,0)}_{Y|X}(\xi_p;0)$, using an estimated variance of Y.

First, CPE_{Bias} depends on $f_{\hat{Q}_u}\big(Q_{Y|C_h}(p)\big)$. From earlier, $F_{Y|C_h}(\hat{Q}_u)$ has a beta distribution. Specifically, for $u \in (0,1)$,

$$F_{Y|C_h}(\hat{Q}_u) \sim \beta[(N_n+1)u, (N_n+1)(1-u)],$$

writing $f_\beta(\cdot)$ for the corresponding beta distribution's PDF and $F_\beta(\cdot)$ for the CDF. For the lower one-sided CI, upper endpoint quantile $u = u_h$ is chosen by the Hutson (1999) method such that $F_\beta(p) = \alpha$. By applying the chain rule and the fact that $F_{Y|C_h}(Q_{Y|C_h}(p)) = p$,

$$
\begin{aligned}
f_{\hat{Q}_u}(x) &= \frac{\partial}{\partial x}F_{\hat{Q}_u}(x) = \frac{\partial}{\partial x}F_\beta\big(F_{Y|C_h}(x)\big) = f_\beta\big(F_{Y|C_h}(x)\big)f_{Y|C_h}(x), \\
f_{\hat{Q}_u}\big(Q_{Y|C_h}(p)\big) &= f_\beta\big(F_{Y|C_h}\big(Q_{Y|C_h}(p)\big)\big)f_{Y|C_h}\big(Q_{Y|C_h}(p)\big) = f_\beta(p;u)f_{Y|C_h}\Big(F^{-1}_{Y|C_h}(p)\Big), \\
f_\beta(p;u) &\equiv \frac{\Gamma(N_n+1)}{\Gamma((N_n+1)u)\Gamma((N_n+1)(1-u))}p^{(N_n+1)u-1}(1-p)^{(N_n+1)(1-u)-1}.
\end{aligned}
$$

To avoid recursive dependence on h, we can approximate $f_{Y|C_h}\big(Q_{Y|C_h}(p)\big) \doteq f_{Y|X}(\xi_p;0)$ up to smaller-order terms.

We can also approximate

$$P_C = \int_{C_h} f_X(x)\,\mathrm{d}x = \mathrm{Vol}(C_h) f_X(0)[1+o(1)] = 2hf_X(0)[1+o(1)],$$

$$N_n \doteq nP_C = n\mathrm{Vol}(C_h) f_X(0)[1+o(1)] = 2nhf_X(0)[1+o(1)].$$

We know that $\mathrm{CPE}_{\mathrm{GK}} > 0$ (over-coverage), and we can estimate the sign of $\mathrm{CPE}_{\mathrm{Bias}}$. If they are opposite, the optimal bandwidth causes them to cancel out; if they are the same sign, the optimal bandwidth minimizes their sum. The only difference is an extra coefficient of $[2d/(2b+d)]^{1/(b+3d/2)} = [2d/(d+4)]^{2/(4+3d)}$ from the first-order condition in the latter case, where the initial exponents of h come down when taking a derivative.

Plug-in bandwidth: one-sided

For one-sided inference when $\mathrm{CPE}_{\mathrm{GK}}$ and $\mathrm{CPE}_{\mathrm{Bias}}$ are of opposite sign, with $\epsilon = \epsilon_h$ or $\epsilon = \epsilon_\ell$, the optimal h equates (up to smaller-order terms)

$$N_n^{-1} z_{1-\alpha} \frac{\epsilon(1-\epsilon)}{p(1-p)} \phi(z_{1-\alpha}) = h^2 \frac{f_X(0) F_{Y|X}^{(0,2)}(\xi_p;0) + 2f_X'(0) F_{Y|X}^{(0,1)}(\xi_p;0)}{6 f_X(0) f_{Y|X}(\xi_p;0)} f_\beta(p;u) f_{Y|X}(\xi_p;0).$$

After plugging in estimates of the unknown objects, $N_n \doteq 2nhf_X(0)$, and value for u (e.g., based on a pilot bandwidth), the above equation could be solved numerically for h with any standard statistical software.

Alternatively, we could approximate the beta PDF with a normal PDF (Goldman and Kaplan, 2012). Writing β for a random variable with the distribution $\beta[(N_n+1)u, (N_n+1)(1-u)]$ from above,

$$\sqrt{N_n}(\beta - u)/\sqrt{u(1-u)} \xrightarrow{d} N(0,1), \quad \text{so that (informally) } \beta \overset{a}{\sim} N(u, u(1-u)/N_n),$$

$$f_\beta(p;u) = \frac{\sqrt{N_n}}{\sqrt{u(1-u)}} \Big[\phi\Big([p-u]/\sqrt{u(1-u)/N_n} \Big) + O(N_n^{-1/2}) \Big],$$

where $\phi(\cdot)$ is the standard normal PDF. Using additional results from Goldman and Kaplan

(2012) that

$$\begin{aligned}
u_h &= p + z_{1-\alpha}\sqrt{u_h(1-u_h)/N_n} + O(N_n^{-1}) \text{ and} \\
u_\ell &= p - z_{1-\alpha}\sqrt{u_\ell(1-u_\ell)/N_n} + O(N_n^{-1}), \text{ then for } u = u_h \text{ or } u = u_\ell, \\
f_\beta(p;u) &= N_n^{1/2}[u(1-u)]^{-1/2}\left[\phi\left(\frac{\pm z_{1-\alpha}\sqrt{u(1-u)/N_n}}{\sqrt{u(1-u)/N_n}}\right) + O(N_n^{-1/2})\right] \\
&\doteq N_n^{1/2}[u(1-u)]^{-1/2}\phi(z_{1-\alpha})
\end{aligned}$$

since $\phi(z_{1-\alpha}) = \phi(-z_{1-\alpha})$.

Plugging in $N_n \doteq 2nhf_X(0)$ and $u = p + O(N_n^{-1/2})$, the optimal h is now an explicit function of known values and objects that can be estimated directly from the data. Denoting $\hat{h}_{++}$ as the plug-in bandwidth when both $\text{CPE}_{\text{GK}} > 0$ and $\text{CPE}_{\text{Bias}} > 0$ (both over-coverage), and $\hat{h}_{+-}$ when instead $\text{CPE}_{\text{Bias}} < 0$, we first solve for $\hat{h}_{+-}$. Up to smaller-order terms,

$$\begin{aligned}
[2nhf_X(0)]^{-1}z_{1-\alpha}\frac{\epsilon(1-\epsilon)}{p(1-p)}\phi(z_{1-\alpha}) &= h^2\frac{f_X(0)F_{Y|X}^{(0,2)}(\xi_p;0) + 2f_X'(0)F_{Y|X}^{(0,1)}(\xi_p;0)}{6f_X(0)f_{Y|X}(\xi_p;0)} \\
&\quad\times [2nhf_X(0)]^{1/2}[p(1-p)]^{-1/2}\phi(z_{1-\alpha})f_{Y|X}(\xi_p;0),
\end{aligned}$$

$$\begin{aligned}
h^{7/2} &= \frac{n^{-3/2}2^{-3/2}z_{1-\alpha}\epsilon(1-\epsilon)/\sqrt{p(1-p)}}{\sqrt{f_X(0)}\left\{f_X(0)F_{Y|X}^{(0,2)}(\xi_p;0) + 2f_X'(0)F_{Y|X}^{(0,1)}(\xi_p;0)\right\}/6}, \\
\hat{h}_{+-} &= n^{-3/7}\left(\frac{z_{1-\alpha}}{3\sqrt{p(1-p)f_X(0)}\left\{f_X(0)F_{Y|X}^{(0,2)}(\xi_p;0) + 2f_X'(0)F_{Y|X}^{(0,1)}(\xi_p;0)\right\}}\right)^{2/7}, \\
\hat{h}_{++} &= \hat{h}_{+-}[-2d/(2b+d)]^{2/(2b+3d)} \approx -0.770\hat{h}_{+-}.
\end{aligned}$$

We have plugged in the rule-of-thumb $\epsilon = 0.2$ to avoid iteration (more precisely, $0.5 - (1/2)\sqrt{1-4\sqrt{2}/9} \approx 0.20$, to get the constants to cancel). These $\hat{h}$ hold for both lower and upper one-sided inference.

With the approximation $u = p + O(N_n^{-1/2})$, the bias CPE for the upper endpoint is the negative of the bias CPE for the lower endpoint, up to smaller-order terms. Then, for two-sided inference, the dominant bias terms from the upper and lower CI endpoints cancel, and the CPE is of a smaller order, as in Theorem 3.

Plug-in bandwidth: two-sided, $p = 1/2$

For two-sided inference with $p = 1/2$, the B_h term becomes zero, as is clear in (12). Then $\mathrm{CPE}_{\mathrm{Bias}} < 0$ since $B_h^2 > 0$, $f'_{\hat{Q}^{I,u_\ell}_{Y|C_h}}(p) < 0$, and $f'_{\hat{Q}^{I,u_h}_{Y|C_h}}(p) > 0$. The optimal h causes this to cancel with $\mathrm{CPE}_{\mathrm{GK}} > 0$. By the convergence of our beta to a normal distribution, the product rule for derivatives, and the invariance of $f_{Y|C_h}\Big(F^{-1}_{Y|C_h}(p)\Big)$ to u,

$$
\begin{aligned}
&\frac{\partial}{\partial p} f_\beta(p; u_\ell) \doteq -z_{1-\alpha/2} N_n [u_\ell(1-u_\ell)]^{-1} \phi(z_{1-\alpha/2}), \\
&\frac{\partial}{\partial p} f_\beta(p; u_h) \doteq z_{1-\alpha/2} N_n [u_h(1-u_h)]^{-1} \phi(z_{1-\alpha/2}), \\
f'_{\hat{Q}^{I,u_\ell}_{Y|C_h}}&\big(Q_{Y|C_h}(p)\big) - f'_{\hat{Q}^{I,u_h}_{Y|C_h}}\big(Q_{Y|C_h}(p)\big) \\
&\doteq -z_{1-\alpha/2} N_n \phi(z_{1-\alpha/2}) \big([u_\ell(1-u_\ell)]^{-1} + [u_h(1-u_h)]^{-1}\big) f_{Y|C_h}\Big(F^{-1}_{Y|C_h}(p)\Big) \\
&\doteq -z_{1-\alpha/2} N_n \phi(z_{1-\alpha/2}) 2[p(1-p)]^{-1} f_{Y|X}(\xi_p; 0),
\end{aligned}
$$

so plugging into (10) yields

$$
\begin{aligned}
&N_n^{-1} z_{1-\alpha/2} \frac{\epsilon_h(1-\epsilon_h) + \epsilon_\ell(1-\epsilon_\ell)}{p(1-p)} \phi(z_{1-\alpha/2}) \\
&\quad = -(1/2) h^4 \left(\frac{f_X(0) F^{(0,2)}_{Y|X}(\xi_p; 0) + 2 f'_X(0) F^{(0,1)}_{Y|X}(\xi_p; 0)}{6 f_X(0) f_{Y|X}(\xi_p; 0)} \right)^2 \\
&\qquad \times \big\{ -z_{1-\alpha/2} N_n \phi(z_{1-\alpha/2}) 2[p(1-p)]^{-1} f_{Y|X}(\xi_p; 0) \big\}, \\
&[2nh f_X(0)]^{-2} 2\epsilon(1-\epsilon) \\
&\quad = \frac{h^4}{36 f_X(0)^2 f_{Y|X}(\xi_p; 0)} \Big(f_X(0) F^{(0,2)}_{Y|X}(\xi_p; 0) + 2 f'_X(0) F^{(0,1)}_{Y|X}(\xi_p; 0) \Big)^2, \\
&\hat{h} = n^{-1/3} \left(\frac{3 f_{Y|X}(\xi_p; 0)}{\Big\{ f_X(0) F^{(0,2)}_{Y|X}(\xi_p; 0) + 2 f'_X(0) F^{(0,1)}_{Y|X}(\xi_p; 0) \Big\}^2} \right)^{1/6},
\end{aligned}
$$

again using $\epsilon = 0.2$ as the rule of thumb (and rounding 2.88 up to 3).

Plug-in bandwidth: two-sided, $p \neq 1/2$

For two-sided inference with $d = 1$ and $p \neq 1/2$, the following does not cancel but is of smaller order than the $O(N_n^{1/2})$ in the one-sided case:

$$\begin{aligned}
f_\beta(p; u_h) - f_\beta(p; u_\ell) &= N_n^{1/2}\phi(z_{1-\alpha/2})\big([u_h(1-u_h)]^{-1/2} - [u_\ell(1-u_\ell)]^{-1/2}\big) \\
&= N_n^{1/2}\phi(z_{1-\alpha/2})\left(\frac{2p-1}{2[p(1-p)]^{3/2}}[(u_h - p) - (u_\ell - p)] + O(N_n^{-1})\right) \\
&= N_n^{1/2}\phi(z_{1-\alpha/2})\frac{2p-1}{2[p(1-p)]^{3/2}}N_n^{-1/2}z_{1-\alpha/2} \\
&\quad \times \left(\sqrt{u_h(1-u_h)} + \sqrt{u_\ell(1-u_\ell)}\right) + O(N_n^{-1/2}) \\
&= z_{1-\alpha/2}\phi(z_{1-\alpha/2})\frac{2p-1}{p(1-p)} + O(N_n^{-1/2}) = O(1).
\end{aligned}$$

Thus our two CPE terms are of orders $N_n^{-1} \asymp n^{-1}h^{-1}$ and h^2. This implies $h^* \asymp n^{-1/3}$ and that CPE is $O(n^{-2/3})$. But then the B_h^2 term is of order $h^4 N_n = h^5 n = n^{-2/3}$, so it must also be included. (The B_h^3 term is of order $h^6 N_n^{3/2}$, which is smaller.) Though the second term from the product rule derivative in the B_h^2 term is not zero this time, it is smaller-order and thus omitted below.

The sign of CPE_{Bias} is determined by the sign of $[B_h(2p-1) - B_h^2 N_n]$, as seen in (12). Since $B_h^2 N_n > 0$ always, if $B_h(2p-1) < 0$, then $\text{CPE}_{\text{Bias}} < 0$ irrespective of h (only the magnitude of B_h depends on h, not the sign). If $B_h(2p-1) > 0$, then there may exist some h that yields $\text{CPE}_{\text{Bias}} < 0$, but maybe not. (There will always exist some $h > 1$ that does this, but asymptotically $h \to 0$.) Specifically, for coefficients a and b, $ah^2 - bh^5 < 0$ for $h > \sqrt[3]{a/b}$. Here, $a = (2p-1)(B_h/h^2)$ and $b = 2nf_X(0)(B_h/h^2)^2$, so $\text{CPE}_{\text{Bias}} < 0$ for

$$\begin{aligned}
B_h/h^2 &= -\frac{f_X(0)F_{Y|X}^{(0,2)}(\xi_p; 0) + 2f_X'(0)F_{Y|X}^{(0,1)}(\xi_p; 0)}{6f_X(0)f_{Y|X}(\xi_p; 0)}, \\
h &> \left(\frac{2p-1}{2nf_X(0)B_h/h^2}\right)^{1/3} = n^{-1/3}\left(\frac{3(2p-1)f_{Y|X}(\xi_p; 0)}{-\left\{f_X(0)F_{Y|X}^{(0,2)}(\xi_p; 0) + 2f_X'(0)F_{Y|X}^{(0,1)}(\xi_p; 0)\right\}}\right)^{1/3}.
\end{aligned} \tag{14}$$

When h equals the RHS, $\mathrm{CPE}_{\mathrm{Bias}} = 0$, so the optimal h should be somewhat larger. In practice, using (14) as an equality for $\hat{h}$ probably works well, but we pursue the more exact solution below.

Since such a solution is always possible, we pick h to equate

$$\begin{aligned}
&-N_n^{-1} z_{1-\alpha/2} \frac{\epsilon_h(1-\epsilon_h) + \epsilon_\ell(1-\epsilon_\ell)}{p(1-p)} \phi(z_{1-\alpha/2}) \\
&\quad\doteq B_h \left[f_{\hat{Q}^{I,u_h}_{Y|C_h}}\big(Q_{Y|C_h}(p)\big) - f_{\hat{Q}^{I,u_\ell}_{Y|C_h}}\big(Q_{Y|C_h}(p)\big) \right] \\
&\qquad + (1/2) B_h^2 \left[f'_{\hat{Q}^{I,u_\ell}_{Y|C_h}}\big(Q_{Y|C_h}(p)\big) - f'_{\hat{Q}^{I,u_\ell}_{Y|C_h}}\big(Q_{Y|C_h}(p)\big) \right] \\
&\quad\doteq B_h f_{Y|C_h}\Big(F^{-1}_{Y|C_h}(p)\Big) [f_\beta(p;u_h) - f_\beta(p;u_\ell)] \\
&\qquad - (1/2) B_h^2 z_{1-\alpha/2} N_n \phi(z_{1-\alpha/2}) 2 [p(1-p)]^{-1} f_{Y|C_h}\Big(F^{-1}_{Y|C_h}(p)\Big) \\
&\quad\doteq f_{Y|X}(\xi_p;0) z_{1-\alpha/2} \phi(z_{1-\alpha/2}) [p(1-p)]^{-1} \big[B_h(2p-1) - B_h^2 N_n \big].
\end{aligned}$$

Note that $2p - 1 > 0$ is equivalent to $p > 1/2$, which implies $|u_h - 0.5| > |u_\ell - 0.5|$. In that case, the B_h term is the same sign as B_h itself. If $2p - 1 < 0$, or equivalently $p < 1/2$ or $|u_h - 0.5| < |u_\ell - 0.5|$, the term is the opposite sign of B_h. If $p = 1/2$, the term is zero, as covered in the special case.

Continuing to solve for h,

$$\begin{aligned}
&-[2nhf_X(0)]^{-1} 2\epsilon(1-\epsilon) \\
&\quad = f_{Y|X}(\xi_p;0) \big\{ h^2 (B_h/h^2)(2p-1) - h^4 (B_h^2/h^4)[2nhf_X(0)] \big\}, \\
&n^{-1} \big\{ [f_X(0)]^{-1} \epsilon(1-\epsilon) \big\} \\
&\quad = h^6 n \big\{ 2 f_{Y|X}(\xi_p;0) f_X(0) B_h^2/h^4 \big\} - h^3 \big\{ f_{Y|X}(\xi_p;0)(2p-1)(B_h/h^2) \big\}, \\
&\quad 0 = (h^3)^2 n\{a\} - (h^3)\{b\} - \{c\}/n, \\
&\quad h^3 = \frac{b \pm \sqrt{b^2 + 4ac}}{2an}, \\
&\quad \hat{h} = n^{-1/3} \left(\frac{b + \sqrt{b^2 + 4ac}}{2a} \right)^{1/3}
\end{aligned}$$

$$= n^{-1/3}\left(\frac{(2p-1)(B_h/|B_h|) + \sqrt{(2p-1)^2 + 4\big\{2f_X(0)/f_{Y|X}(\xi_p;0)\big\}\{[f_X(0)]^{-1}\epsilon(1-\epsilon)\}}}{2\{2f_X(0)|B_h|/h^2\}}\right)^{1/3}$$

$$= n^{-1/3}\left(\frac{(2p-1)(B_h/|B_h|) + \sqrt{(2p-1)^2 + (4/3)/f_{Y|X}(\xi_p;0)}}{(2/3)\Big|f_X(0)F^{(0,2)}_{Y|X}(\xi_p;0) + 2f'_X(0)F^{(0,1)}_{Y|X}(\xi_p;0)\Big|/f_{Y|X}(\xi_p;0)}\right)^{1/3},$$

again with rule-of-thumb $\epsilon = 0.2$ and approximating $(0.2)(0.8) \approx 1/6$ to match the median-specific bandwidth when $p = 1/2$ is used here. The other root of the equation yields $h < 0$ since $a > 0$ and $c > 0$, so it is ignored. Even if $b < 0$, we will get $\hat{h} > 0$, for similar reasons. Notice that when $B_h(2p-1) > 0$, if we had plugged in $\epsilon = 0$ to get the smallest possible $\hat{h}$, the resulting $\hat{h}$ would be equal to the lower bound from (14).

C Estimation of unknown objects

For both the one-sided and two-sided $\hat{h}$, estimates are needed for $f_X(0)$, $f'_X(0)$, $F^{(0,1)}_{Y|X}(\xi_p;0)$, and $F^{(0,2)}_{Y|X}(\xi_p;0)$. For the first two, standard kernel density (derivative) estimators suffice, since they are consistent. For the latter two, an estimate $\hat{\xi}_p$ is first needed, to know where to evaluate the derivatives. This may be obtained using any consistent nonparametric conditional quantile estimator; we use one of the many quantile spline fitting options available in R, specifically `rq` from package `quantreg` (Koenker, 2012) using `bs` (R Core Team, 2012) to generate a cubic B-spline basis. For a fixed $Y = \xi_p$, the conditional CDF is a conditional expectation, $F_{Y|X}(y;x) = E(1\{Y \le y\} \mid X = x)$, so it can be estimated by a (mean) regression of $1\{Y_i \le \hat{\xi}_p\}$ on X_i. We again use a cubic spline fit, from which the first two derivatives can be calculated.

For the two-sided $\hat{h}$, an estimate for $f_{Y|X}(\xi_p;0)$ is also needed. This is most complicated object to estimate, but kernel estimators are available, such as `npcdens` from package `np` (Hayfield and Racine, 2008).

There are of course numerous ways to estimate probability densities (and their derivatives) and derivatives of a regression function. New methods or further experimentation and theoretical analysis may reveal some choices, or combinations of choices, that lead to even better performance.