Journal of Econometrics 196 (2017) 331–346

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Fractional order statistic approximation for nonparametric conditional quantile inference

Matt Goldman [a], David M. Kaplan [b,*]

[a] *Microsoft Research, Redmond, WA 98052, United States*

[b] *Department of Economics, University of Missouri, Columbia, MO 65211, United States*

ARTICLE INFO

Article history:
Received 10 February 2015
Received in revised form
2 August 2016
Accepted 28 September 2016
Available online 31 October 2016

JEL classification:
C21

Keywords:
Dirichlet
High-order accuracy
Inference-optimal bandwidth
Kernel smoothing

ABSTRACT

Using and extending fractional order statistic theory, we characterize the $O(n^{-1})$ coverage probability error of the previously proposed (Hutson, 1999) confidence intervals for population quantiles using L-statistics as endpoints. We derive an analytic expression for the n^{-1} term, which may be used to calibrate the nominal coverage level to get $O\left(n^{-3/2}[\log(n)]^3\right)$ coverage error. Asymptotic power is shown to be optimal. Using kernel smoothing, we propose a related method for nonparametric inference on conditional quantiles. This new method compares favorably with asymptotic normality and bootstrap methods in theory and in simulations. Code is provided for both unconditional and conditional inference.

1. Introduction

Quantiles contain information about a distribution's shape. Complementing the mean, they capture heterogeneity, inequality, and other measures of economic interest. Nonparametric conditional quantile models further allow arbitrary heterogeneity across regressor values. This paper concerns nonparametric inference on quantiles and conditional quantiles. In particular, we characterize the high-order accuracy of both Hutson's (1999) L-statistic-based confidence intervals (CIs) and our new conditional quantile CIs.

Conditional quantiles appear across diverse topics because they are fundamental statistical objects. Such topics include wages (Hogg, 1975; Chamberlain, 1994; Buchinsky, 1994), infant birthweight (Abrevaya, 2001), demand for alcohol (Manning et al., 1995), and Engel curves (Alan et al., 2005; Deaton, 1997, pp. 81–82), which we examine in our empirical application.

We formally derive the coverage probability error (CPE) of the CIs from Hutson (1999), as well as asymptotic power of the corresponding hypothesis tests. Hutson (1999) had proposed CIs for quantiles using L-statistics (interpolating between order statistics) as endpoints and found they performed well, but formal proofs were lacking. Using the analytic n^{-1} term we derive in the CPE, we provide a new calibration to achieve $O\big(n^{-3/2}[\log(n)]^3\big)$ CPE, analogous to the Ho and Lee (2005a) analytic calibration of the CIs in Beran and Hall (1993).

The theoretical results we develop contribute to the fractional order statistic literature and provide the basis for inference on other objects of interest explored in Goldman and Kaplan (2016b) and Kaplan (2014). In particular, Theorem 2 tightly links the distributions of L-statistics from the observed and 'ideal' (unobserved) fractional order statistic processes. Additionally, Lemma 7 provides Dirichlet PDF and PDF derivative approximations.

High-order accuracy is important for small samples (e.g., for experiments) as well as nonparametric analysis with small *local* sample sizes. For example, if $n = 1024$ and there are five binary regressors, then the smallest local sample size cannot exceed $1024/2^5 = 32$.

For nonparametric conditional quantile inference, we apply the unconditional method to a local sample (similar to local constant kernel regression), smoothing over continuous covariates and also allowing discrete covariates. CPE is minimized by balancing the CPE of our unconditional method and the CPE from bias due to smoothing. We derive the optimal CPE and bandwidth rates, as well as a plug-in bandwidth when there is a single continuous covariate.

* Correspondence to: Department of Economics, University of Missouri, 909 University Ave, 118 Professional Bldg, Columbia, MO 65211-6040, United States.
E-mail address: kaplandm@Missouri.edu (D.M. Kaplan).
URL: http://faculty.missouri.edu/~kaplandm (D.M. Kaplan).

http://dx.doi.org/10.1016/j.jeconom.2016.09.015
0304-4076/

Our L-statistic method has theoretical and computational advantages over methods based on normality or an unsmoothed bootstrap. The theoretical bottleneck for our approach is the need to use a uniform kernel. Nonetheless, even if normality or bootstrap methods assume an infinitely differentiable conditional quantile function (and hypothetically fit an infinite-degree local polynomial), our CPE is still of smaller order with one or two continuous covariates. Our method also computes more quickly than existing methods (of reasonable accuracy), handling even more challenging tasks in 10–15 seconds instead of minutes.

Recent complementary work of Fan and Liu (2016) also concerns a "direct method" of nonparametric inference on conditional quantiles. They use a limiting Gaussian process to derive first-order accuracy in a general setting, whereas we use the finite-sample Dirichlet process to achieve high-order accuracy in an iid setting. Fan and Liu (2016) also provide uniform (over X) confidence bands. We suggest a confidence band from interpolating a growing number of joint CIs (as in Horowitz and Lee (2012)), although it will take additional work to rigorously justify. A different, ad hoc confidence band described in Section 6 generally outperformed others in our simulations.

If applied to a local constant estimator with a uniform kernel and the same bandwidth, the Fan and Liu (2016) approach is less accurate than ours due to the normal (instead of beta) reference distribution and integer (instead of interpolated) order statistics in their CI in Eq. (6). However, with other estimators like local polynomials or that in Donald et al. (2012), the Fan and Liu (2016) method is not necessarily less accurate. One limitation of our approach is that it cannot incorporate these other estimators, whereas Assumption GI(iii) in Fan and Liu (2016) includes any estimator that weakly converges (over a range of quantiles) to a Gaussian process with a particular structure. We compare further in our simulations. One open question is whether using our beta reference and interpolation can improve accuracy for the general Fan and Liu (2016) method beyond the local constant estimator with a uniform kernel; our Lemma 3 shows this at least retains first-order accuracy.

The order statistic approach to quantile inference uses the idea of the probability integral transform, which dates back to R.A. Fisher (1932), Karl Pearson (1933), and Neyman (1937). For continuous $X_i \overset{iid}{\sim} F(\cdot)$, $F(X_i) \overset{iid}{\sim} \text{Unif}(0, 1)$. Each order statistic from such an iid uniform sample has a known beta distribution for any sample size n. We show that the L-statistic linearly interpolating consecutive order statistics also follows an approximate beta distribution, with only $O(n^{-1})$ error in CDF. Although $O(n^{-1})$ is an asymptotic claim, the CPE of the CI using the L-statistic endpoint is bounded between the CPEs of the CIs using the two order statistics comprising the L-statistic, where one such CPE is too small and one is too big, for any sample size. This is an advantage over methods more sensitive to asymptotic approximation error.

Many other approaches to one-sample quantile inference have been explored. With Edgeworth expansions, Hall and Sheather (1988) and Kaplan (2015) obtain two-sided $O(n^{-2/3})$ CPE. With bootstrap, smoothing is necessary for high-order accuracy. This increases the computational burden and requires good bandwidth selection in practice.[1] See Ho and Lee (2005b, §1) for a review of bootstrap methods. Smoothed empirical likelihood (Chen and Hall, 1993) also achieves nice theoretical properties, but with the same caveats.

Other order statistic-based CIs dating back to Thompson (1936) are surveyed in David and Nagaraja (2003, §7.1). Most closely related to Hutson (1999) is Beran and Hall (1993). Like Hutson (1999), Beran and Hall (1993) linearly interpolate order statistics for CI endpoints, but with an interpolation weight based on the binomial distribution. Although their proofs use expansions of the Rényi (1953) representation instead of fractional order statistic theory, their n^{-1} CPE term is identical to that for Hutson (1999) other than the different weight. Prior work (e.g., Bickel, 1967; Shorack, 1972) has established asymptotic normality of L-statistics and convergence of the sample quantile process to a Gaussian limit process, but without such high-order accuracy.

The most apparent difference between the two-sided CIs of Beran and Hall (1993) and Hutson (1999) is that the former are symmetric in the order statistic index, whereas the latter are equal-tailed. This allows Hutson (1999) to be computed further into the tails. Additionally, our framework can be extended to CIs for interquantile ranges and two-sample quantile differences (Goldman and Kaplan, 2016b), which has not been done in the Rényi representation framework.

For nonparametric conditional quantile inference, in addition to the aforementioned Fan and Liu (2016) approach, Chaudhuri (1991) derives the pointwise asymptotic normal distribution of a local polynomial estimator. Qu and Yoon (2015) propose modified local linear estimators of the conditional quantile process that converge weakly to a Gaussian process, and they suggest using a type of bias correction that strictly enlarges a CI to deal with the first-order effect of asymptotic bias when using the MSE-optimal bandwidth rate.

Section 2 contains our theoretical results on fractional order statistic approximation, which are applied to unconditional quantile inference in Section 3. Section 4 concerns our new conditional quantile inference method. An empirical application and simulation results are in Sections 5 and 6, respectively. Proof sketches are collected in Appendix A, while the supplemental appendix contains full proofs. The supplemental appendix also contains details of the plug-in bandwidth calculations, as well as additional empirical and simulation results (see Appendix B).

Notationally, $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the standard normal PDF and CDF, $\doteq$ should be read as "is equal to, up to smaller-order terms", $\asymp$ as "has exact (asymptotic) rate/order of", and $A_n = O(B_n)$ as usual. Acronyms used are those for cumulative distribution function (CDF), confidence interval (CI), coverage probability (CP), coverage probability error (CPE), and probability density function (PDF).

2. Fractional order statistic theory

In this section, we introduce notation and present our core theoretical results linking unobserved 'ideal' fractional L-statistics with their observed counterparts.

Given an iid sample $\{X_i\}_{i=1}^n$ of draws from a continuous CDF denoted[2] $F(\cdot)$, interest is in $Q(p) \equiv F^{-1}(p)$ for some $p \in (0, 1)$, where $Q(\cdot)$ is the quantile function. For $u \in (0, 1)$, the sample L-statistic commonly associated with $Q(u)$ is

$$\hat{Q}_X^L(u) \equiv (1-\epsilon)X_{n:k} + \epsilon X_{n:k+1}, \quad k \equiv \lfloor u(n+1) \rfloor, \qquad (1)$$
$$\epsilon \equiv u(n+1) - k,$$

where $\lfloor \cdot \rfloor$ is the floor function, ϵ is the interpolation weight, and $X_{n:k}$ denotes the kth order statistic (i.e., kth smallest sample value). While $Q(u)$ is latent and nonrandom, $\hat{Q}_X^L(u)$ is a random variable,

[1] For example, while achieving the impressive two-sided CPE of $O(n^{-3/2})$, Polansky and Schucany (1997, p. 833) admit, "If this method is to be of any practical value, a better bandwidth estimation technique will certainly be required".

[2] F will often be used with a random variable subscript to denote the CDF of that particular random variable. If no subscript is present, then $F(\cdot)$ refers to the CDF of X. Similarly for the PDF $f(\cdot)$.