Second order improvements of sample quantiles using subsamples

Keith Knight[*]
Department of Statistics
University of Toronto

Gilbert W. Bassett, Jr.
Department of Finance
University of Illinois, Chicago

Abstract: Suppose that $X_1, \cdots, X_n$ are i.i.d. random variables with distribution function F. It is well known that if F is differentiable at the α-quantile $q(\alpha)$ with $F'(q(\alpha)) > 0$ then the sample quantile is asymptotically normal. In this note we compare this standard quantile estimator to one obtained by taking weighted averages of sample quantiles from non-overlapping subsamples or from balanced overlapping subsamples. It is straightforward to show that these "average-of-subsample-quantile" estimators are first-order equivalent to the standard estimator. Second order properties however differ in an interesting fashion. While the standard estimator might be intuitively expected best, it is possible to outperform that estimator in a certain sense. We also indicate connections to recently developed methods based on averaging of estimates from bootstrap samples (bagging and bragging) and from without replacement subsamples (subagging). Finally, we show how results generalize when the standard differentiability condition on F is relaxed.

1 Introduction

Suppose that $X_1, \cdots, X_n$ are independent, identically distributed (i.i.d.) random variables with distribution function F. The α quantile of F, $q(\alpha)$ can estimated non-parametrically by $\widehat{q}_n(\alpha)$ minimizing the objective function

$$h_n(t) = \sum_{i=1}^{n} \rho_\alpha(X_i - t) \tag{1}$$

[*]Research supported by a grant from the Natural Sciences and Engineering Research Council of Canada

where $\rho_\alpha(x) = x[\alpha - I(x < 0)]$. In the case where the minimizer of (1) is not unique, we can define $\widehat{q}_n(\alpha)$ to be the mid-point of the set of minimizers or, alternatively, one of the end-points of this set. In any event, if F is differentiable at $q(\alpha)$ with $F'(q(\alpha)) = f(q(\alpha)) > 0$ then

$$\sqrt{n}\,(\widehat{q}_n(\alpha) - q(\alpha)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{f^2(q(\alpha))}\right)$$

with the result holding for any sequence $\{\widehat{q}_n(\alpha)\}$ of minimizers of (1).

In this article, we will attempt to "improve" the asymptotic linearity of the estimator $\widehat{q}_n(\alpha)$ by combining quantile estimators from subsamples. More precisely when $f(q(\alpha)) > 0$, we have the Bahadur-Kiefer (Bahadur, 1966; Kiefer, 1967) representation

$$\sqrt{n}(\widehat{q}_n(\alpha) - q(\alpha)) = \frac{1}{f(q(\alpha))\sqrt{n}} \sum_{i=1}^{n} \{\alpha - I[X_i < q(\alpha)]\} + \widehat{R}_n(\alpha) \tag{2}$$

where the remainder term $\widehat{R}_n(\alpha) = O_p(n^{-1/4})$; this remainder term can be thought of as representing the deviation from linearity of $\widehat{q}_n(\alpha)$. It is straightforward to construct first order equivalent estimators $\widetilde{q}_n(\alpha)$ such that

$$\sqrt{n}(\widetilde{q}_n(\alpha) - q(\alpha)) = \frac{1}{f(q(\alpha))\sqrt{n}} \sum_{i=1}^{n} \{\alpha - I[X_i < q(\alpha)]\} + \widetilde{R}_n(\alpha) \tag{3}$$

where $\widetilde{R}_n(\alpha) \neq \widehat{R}_n(\alpha)$. A natural question to ask is whether or not it is possible to improve on the linearity of $\widehat{q}_n(\alpha)$, that is, to construct an estimator $\widetilde{q}_n(\alpha)$ satisfying (3) such that its remainder $\widetilde{R}_n(\alpha)$ is "smaller" than $\widehat{R}_n(\alpha)$ in (2).

The paper is organized as follows: In section 2, we discuss first order theory of quantile estimators that combine estimators from subsamples; in section 3, we derive some second order results for estimators satisfying (3); in section 4, we discuss second order theory for quantile estimators obtained via bagging and subagging; finally in section 5, we consider asymptotic theory in non-standard cases.

2 Combining information from subsamples; first order theory

There are a number of ways of constructing subsamples. We start by dividing $X_1, \cdots, X_n$ into k non-overlapping subsamples of length $n_1, \cdots, n_k$. Define $\widehat{q}_n^{(1)}(\alpha)$ to be the sample α quantile of $X_1, \cdots, X_{n_1}$, $\widehat{q}_n^{(2)}(\alpha)$ to be the sample α quantile of $X_{n_1+1}, \cdots, X_{n_1+n_2}$ and so on.

The quantile estimators $\widehat{q}_n^{(1)}(\alpha), \cdots, \widehat{q}_n^{(k)}(\alpha)$ can be combined in a number of ways. For example, we could define an estimator $\widetilde{q}_n(\alpha)$ to be a weighted average of the subsample estimators or the median (or some other order statistic); the latter type of estimator might be viewed as a generalization of Tukey's (1978) "ninther".

One possible consideration for combining subsample quantiles is computational. While a sample quantile requires only $O(n)$ computational effort (like the sample mean), it also requires $O(n)$ storage; combining subsample quantiles will typically involve the same computational effort while reducing the storage necessary. These issues are discussed, for example, in Rousseeuw and Bassett (1990), Hurley and Modarres (1995), and Dor and Zwick (1999).

The following result gives a convolution-type theorem for estimators that combine the information in $\widehat{q}_n^{(1)}(\alpha), \cdots, \widehat{q}_n^{(k)}(\alpha)$.

THEOREM 1. Assume $F'(q(\alpha)) = f(q(\alpha))$ and suppose that

$$\widetilde{q}_n(\alpha) = g_n\left(\widehat{q}_n^{(1)}(\alpha), \cdots, \widehat{q}_n^{(k)}(\alpha)\right)$$

where

(a) $n_i/n \to \lambda_i > 0$ as $n \to \infty$ for $i = 1, \cdots, k$,

(b) $g_n(\boldsymbol{x}_n) \to g_0(\boldsymbol{x}_0)$ for all sequences $\{\boldsymbol{x}_n\}$ converging to $\boldsymbol{x}_0$, and

(c) $\{g_n\}$ and g_0 are location and scale equivariant in the following sense: $g_n(a\,\boldsymbol{x} + b\mathbf{1}) = a\,g_n(\boldsymbol{x}) + b$ for all b and $a > 0$.

Then

$$\sqrt{n}(\widetilde{q}_n(\alpha) - q(\alpha)) \xrightarrow{d} \frac{1}{f(q(\alpha))}(W + V)$$

where $W \sim \mathcal{N}(0, \alpha(1-\alpha))$ and V is independent of W.

Proof. Let $W_1, \cdots, W_k$ be independent $\mathcal{N}(0, \alpha(1-\alpha))$ random variables. Then

$$\begin{pmatrix} \sqrt{n}(\widehat{q}_n^{(1)}(\alpha) - q(\alpha)) \\ \vdots \\ \sqrt{n}(\widehat{q}_n^{(k)}(\alpha) - q(\alpha)) \end{pmatrix} \xrightarrow{d} \frac{1}{f(q(\alpha))} \begin{pmatrix} \lambda_1^{-1/2} W_1 \\ \vdots \\ \lambda_k^{-1/2} W_k \end{pmatrix}$$

and so

$$\begin{aligned} \sqrt{n}(\widetilde{q}_n(\alpha) - q(\alpha)) &= g_n\left(\sqrt{n}(\widehat{q}_n^{(1)}(\alpha) - q(\alpha)), \cdots, \sqrt{n}(\widehat{q}_n^{(k)}(\alpha) - q(\alpha))\right) \\ &\xrightarrow{d} \frac{1}{f(q(\alpha))} g_0\left(W_1/\lambda_1^{1/2}, \cdots, W_k/\lambda_k^{1/2}\right). \end{aligned}$$

Finally, suppose that $W_1', \cdots, W_k'$ are independent random variables with

$$W_i' \sim \mathcal{N}(\theta, \lambda_i^{-1}\alpha(1-\alpha))$$

where θ is unknown and all other parameters are known. Then we have a one-parameter exponential family for which $S = \sum_{i=1}^{k} \lambda_i W'_i$ is a sufficient and complete statistic for θ while $g_0(W'_1, \cdots, W'_k) - S$ is ancillary. Thus by Basu's Theorem, S is independent of $g_0(W'_1, \cdots, W'_k) - S$ (for all θ). The conclusion of the theorem follows by noting that when $\theta = 0$, $g_0(W_1/\lambda_1^{1/2}, \cdots, W_k/\lambda_k^{1/2})$ has the same distribution as $g_0(W'_1, \cdots, W'_k)$ and S has the same distribution as W. □

It is easy to construct examples of estimators satisfying the conditions of Theorem 1 whose limiting distribution has $V \neq 0$ and $W + V$ is not normally distributed. For example, we could take $\widetilde{q}_n(\alpha)$ to be some fixed order statistic of $\widehat{q}_n^{(1)}(\alpha), \cdots, \widehat{q}_n^{(k)}(\alpha)$; further examples are given in Knight (2002).

Note that Theorem 1 can be extended to overlapping subsamples under certain conditions. For example, consider a situation where the subsamples each have size approximately λn with "adjacent" subsamples have an overlap of ρn observations so that every observation is included in $\ell > 1$ of the k subsamples. Then the subsample quantiles $\widehat{q}_n^{(1)}(\alpha), \cdots, \widehat{q}_n^{(k)}(\alpha)$ satisfy

$$\sqrt{n} \begin{pmatrix} \widehat{q}_n^{(1)}(\alpha) - q(\alpha) \\ \vdots \\ \widehat{q}_n^{(k)}(\alpha) - q(\alpha) \end{pmatrix} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$$

where Σ is a circulant matrix. In this case, Theorem 1 still holds for estimators of the form $\widetilde{q}_n(\alpha) = g_n(\widehat{q}_n^{(1)}(\alpha), \cdots, \widehat{q}_n^{(k)}(\alpha))$ for g_n satisfying the conditions of Theorem 1. To see this, note that the limiting distribution of $\sqrt{n}(\widetilde{q}_n(\alpha) - q(\alpha))$ can be represented as $g_0(\boldsymbol{W})$ where $\boldsymbol{W} \sim \mathcal{N}(\mathbf{0}, \Sigma)$; if $\boldsymbol{W}' \sim \mathcal{N}(\theta \mathbf{1}, \Sigma)$ then $\mathbf{1}^T \boldsymbol{W}'$ is sufficient and complete for θ when Σ is known (since Σ^{-1} is also a circulant matrix) and the remainder of the proof is the same.

Similarly, we can extend Theorem 1 to estimators based on "running quantiles". For some $\lambda \in (0, 1)$, we define $\widehat{q}_n(\alpha; u)$ to be the α quantile of the subsample $\{X_{\lfloor nu \rfloor + 1}, \cdots, X_{\lfloor n(u+\lambda) \rfloor}\}$ where $X_{n+k} = X_k$ for $k \geq 1$. Then the process

$$Q_n(u) = \sqrt{n}(\widehat{q}_n(\alpha; u) - q(\alpha))$$

converges weakly to a zero-mean Gaussian process $Q(u)$ with

$$E[Q(u)Q(v)] = \frac{\alpha(1-\alpha)}{\lambda f^2(q(\alpha))} C_\lambda(u, v)$$

where

$$C_\lambda(u, v) = \sum_{s=-1}^{1} \left(1 - \frac{1}{\lambda}|u - v + s|\right) I\left(|u - v + s| \leq \lambda\right). \tag{4}$$

Then for estimators $\widetilde{q}_n(\alpha) = g_n(\widehat{q}_n(\alpha;\cdot))$ where $g_n(f_n) \to g_0(f_0)$ whenever $f_n(t) \to f_0(t)$ uniformly over compact sets and $g_n(a\,f + b) = a\,g_n(f) + b$, we have

$$\sqrt{n}(\widetilde{q}_n(\alpha) - q(\alpha)) \xrightarrow{d} \int_0^1 Q(u)\,du + V \tag{5}$$

where the random variables in the limit are independent; note that

$$\int_0^1 Q(u)\,du \sim \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{f^2(q(\alpha))}\right)$$

and so we have another extension of Theorem 1. For any fixed $\lambda \in (0,1)$, the estimator

$$\widetilde{q}_n(\alpha) = \int_0^1 \widehat{q}_n(\alpha;u)\,du \tag{6}$$

satisfies (5) with $V = 0$.

In general, the estimators considered in Theorem 1 are in fact "argmin" estimators. More precisely, if $\widetilde{q}_n(\alpha) = g_n(\widehat{q}_n^{(1)}(\alpha), \cdots, \widehat{q}_n^{(k)}(\alpha))$ (where the subsample quantiles $\{\widehat{q}_n^{(j)}(\alpha)\}$ are from either disjoint or overlapping subsamples) then $\widetilde{q}_n(\alpha)$ minimizes the objective function

$$\widetilde{h}_n(t) = \inf\{h_n^{(1)}(t_1) + \cdots + h_n^{(k)}(t_k) : g_n(t_1, \cdots, t_k) = t\}$$

where

$$h_n^{(j)}(t_j) = \sum_{i \in I_j} \rho_\alpha(X_i - t_j)$$

for $j = 1, \cdots, k$ where I_j is the index set for subsample j.

EXAMPLE 1. We consider the case of four overlapping subsamples each containing 3/4 of the observations so that each observation is contained in exactly three subsamples. The covariance matrix of the asymptotic distribution of the subsample quantiles is

$$\Sigma = \frac{4\alpha(1-\alpha)}{3f^2(q(\alpha))}\begin{pmatrix} 1 & 2/3 & 2/3 & 2/3 \\ 2/3 & 1 & 2/3 & 2/3 \\ 2/3 & 2/3 & 1 & 2/3 \\ 2/3 & 2/3 & 2/3 & 1 \end{pmatrix}$$

We then define

$$\widetilde{q}_n(\alpha) = \frac{1}{4}\left[\widehat{q}_n^{(1)}(\alpha) + \cdots + \widehat{q}_n^{(4)}(\alpha)\right],$$

which has the same limiting distribution as the (full) sample quantile $\widehat{q}_n(\alpha)$. The "implied" objective function $\widetilde{h}(t)$ for $\widetilde{q}_n(\alpha)$ can be evaluated for each t by minimizing a quantile regression objective function. This objective function (for $\alpha = 1/2$) is shown in Figures 1

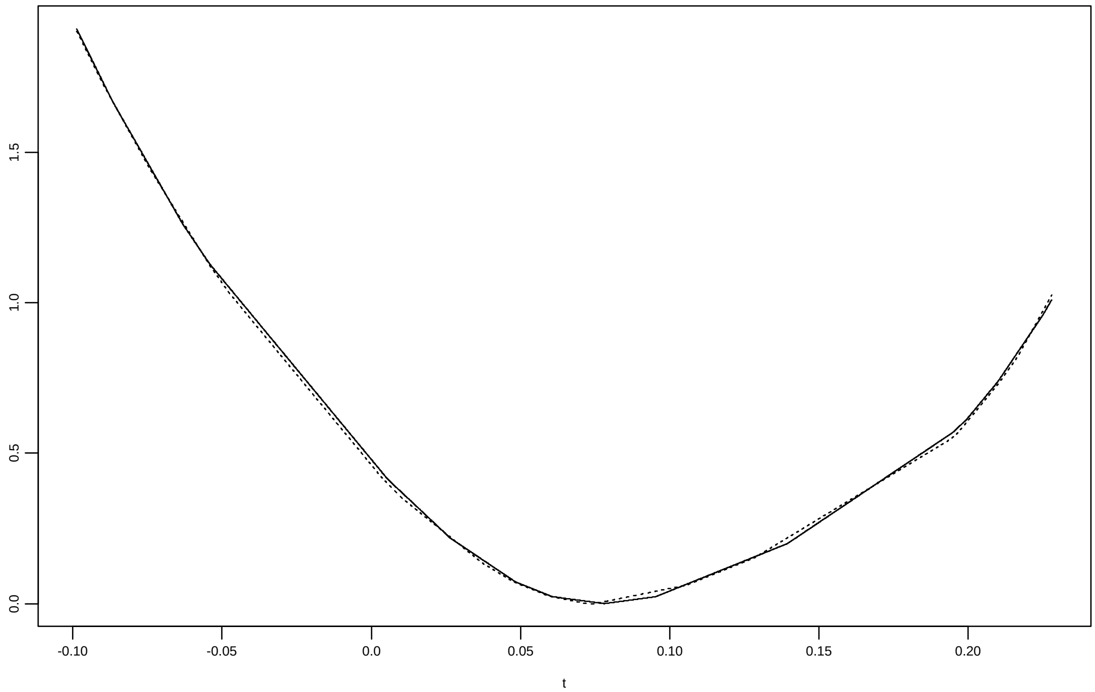

Figure 1: Comparison of $\widetilde{h}_n$ (solid line) to h_n (dashed line) for a sample of 100 $\mathcal{N}(0,1)$ random variables; both $\widetilde{h}_n$ and h_n have been scaled to facilitate the comparison.

and 2; in Figure 1, we take 100 observations from a $\mathcal{N}(0,1)$ distribution while in Figure 2, we have 100 observations from a density with $f(q(1/2)) = 0$ (for which the asymptotics described in Theorem 1 do not hold. (In both Figures 1 and 2, the range of values given on the x-axis is a non-parametric approximate 95% confidence interval for the median based on order statistics; see Efron (1982) for details.) In the latter case, the asymptotic distributions of the two estimators will be different and, perhaps not surprisingly, h_n and $\widetilde{h}_n$ are quite different; in contrast, in the "regular" case, the agreement between h_n and $\widetilde{h}_n$ can be very close as is the case in Figure 1. ◇

Under the conditions of Theorem 1 (assuming the sample quantiles are computed from disjoint subsamples), $\sqrt{n}(\widetilde{q}_n(\alpha) - q(\alpha))$ converges in distribution to the minimizer of

$$Z(u) = \inf\{Z^{(1)}(u_1) + \cdots + Z^{(k)}(u_k) : g_0(u_1, \cdots, u_k) = u\}$$

where

$$Z^{(j)}(u_j) = -\lambda_j^{1/2} W_j u_j + \frac{1}{2}\lambda_j f(q(\alpha)) u_j^2$$

where $W_1, \cdots, W_k$ are independent $\mathcal{N}(0, \alpha(1-\alpha))$ random variables.

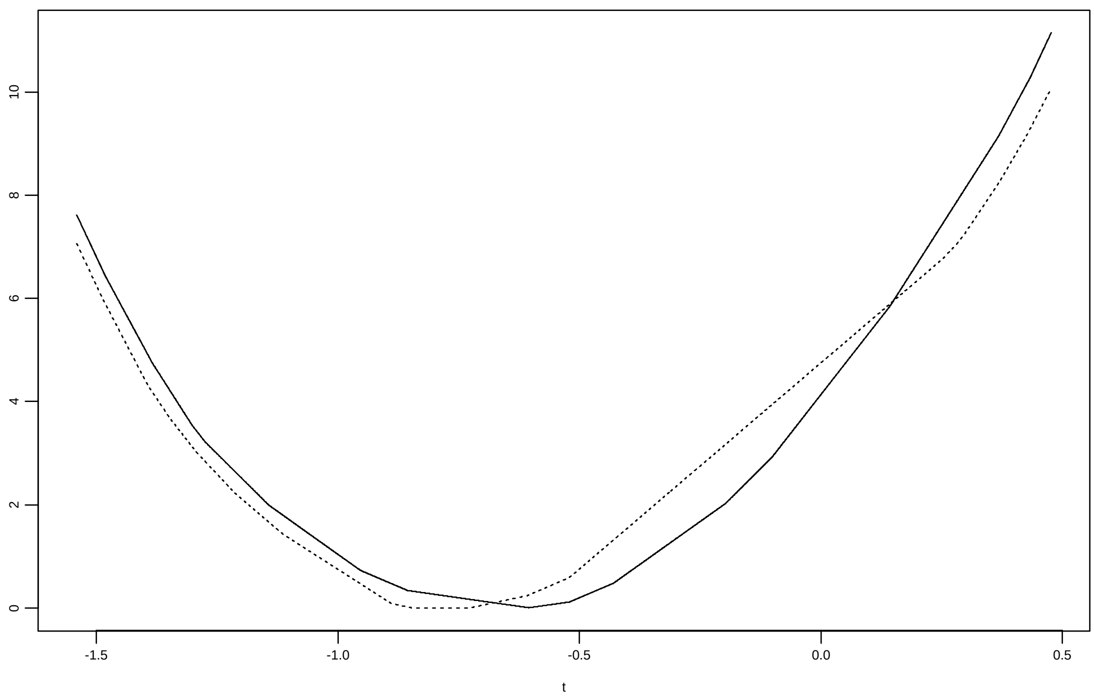

Figure 2: Comparison of $\widetilde{h}_n$ (solid line) to h_n (dashed line) for a sample of 100 random variables from a distribution with $f(q(1/2)) = 0$; both $\widetilde{h}_n$ and h_n have been scaled to facilitate the comparison.

3 Some second order theory

As indicated in section 2, there is a wide variety of estimators satisfying (3) with $\widetilde{R}_n(\alpha) = o_p(1)$. In this section, we will take a closer look at the second order behaviour of some of these estimators.

As a starting point, we will look at estimators of $q(\alpha)$ constructed by combining sample quantiles from disjoint subsamples. Under the conditions of Theorem 1, an optimal estimator of $q(\alpha)$ based on $\widehat{q}_n^{(1)}(\alpha), \cdots, \widehat{q}_n^{(k)}(\alpha)$ is

$$\bar{q}_n(\alpha) = \sum_{i=1}^{k} \frac{n_i}{n} \widehat{q}_n^{(i)}(\alpha) \approx \sum_{i=1}^{k} \lambda_i \widehat{q}_n^{(i)}(\alpha), \tag{7}$$

which satisfies $\sqrt{n}(\bar{q}_n(\alpha) - \widehat{q}_n(\alpha)) = o_p(1)$.

Intuitively, the sample quantile $\widehat{q}_n(\alpha)$ should be a better estimator of $q(\alpha)$. To investigate this, we look at the second order representations of both estimators:

$$\sqrt{n}(\widehat{q}_n(\alpha) - q(\alpha)) = \frac{1}{f(q(\alpha))\sqrt{n}} \sum_{i=1}^{n} \{\alpha - I[X_i < q(\alpha)]\} + \widehat{R}_n(\alpha) \tag{8}$$

$$\sqrt{n}(\bar{q}_n(\alpha) - q(\alpha)) = \frac{1}{f(q(\alpha))\sqrt{n}} \sum_{i=1}^{n} \{\alpha - I[X_i < q(\alpha)]\} + \bar{R}_n(\alpha). \tag{9}$$

The following theorem gives the joint asymptotic behaviour of $\widehat{R}_n(\alpha)$ and $\bar{R}_n(\alpha)$.

THEOREM 2. Define $\bar{q}_n(\alpha)$ as in (7) and suppose that
(a) $F(q(\alpha) + t) - \alpha = t\, f(q(\alpha)) + o(t^{3/2})$ as $t \to 0$, and
(b) $n_i/n = \lambda_i + o(n^{-1/4})$ for $i = 1, \cdots, k$.
Then for $\widehat{R}_n(\alpha)$ and $\bar{R}_n(\alpha)$ defined in (8) and (9), we have

$$n^{1/4} \begin{pmatrix} \widehat{R}_n(\alpha) \\ \bar{R}_n(\alpha) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \widehat{R}_0(\alpha) \\ \bar{R}_0(\alpha) \end{pmatrix}$$

where

$$\widehat{R}_0(\alpha) = \frac{1}{f(q(\alpha))} \sum_{i=1}^{k} \lambda_i^{1/4} B_i \left(\frac{\lambda_i^{1/2} W}{f(q(\alpha))} \right)$$

$$\bar{R}_0(\alpha) = \frac{1}{f(q(\alpha))} \sum_{i=1}^{k} \lambda_i^{1/4} B_i \left(\frac{W_i}{f(q(\alpha))} \right);$$

$B_1, \cdots, B_k$ are independent Gaussian processes with $E[(B_i(s) - B_i(t))^2] = f(q(\alpha))|s - t|$, $W_1, \cdots, W_k$ are independent $\mathcal{N}(0, \alpha(1-\alpha))$ random variables that are independent of the B_i's, and

$$W = \sum_{i=1}^{k} \lambda_i^{1/2} W_i.$$

Proof. For notational convenience, relabel the observations X_{ij} for $i = 1, \cdots, k$ and $j = 1, \cdots, n_i$. Define

$$Z_n^{(i)}(u) = \sum_{j=1}^{n_i} \left[\rho_\alpha\left(X_{ij} - q(\alpha) - u/\sqrt{n}\right) - \rho_\alpha\left(X_{ij} - q(\alpha)\right)\right]$$

for $i = 1, \cdots, k$ and

$$Z_n(u) = \sum_{i=1}^{k} Z_n^{(i)}(u).$$

Note that $\sqrt{n}(\widehat{q}_n^{(i)}(\alpha) - q(\alpha))$ minimizes $Z_n^{(i)}$ and $\sqrt{n}(\widehat{q}_n(\alpha) - q(\alpha))$ minimizes Z_n. Moreover, $Z_n^{(i)}$ and Z_n can be approximated by (respectively)

$$\begin{aligned}
\bar{Z}_n^{(i)}(u) &= -\frac{u}{\sqrt{n}} \sum_{j=1}^{n_i} \{\alpha - I[X_{ij} < q(\alpha)]\} + \frac{\lambda_i f(q(\alpha))}{2} u^2 \\
\bar{Z}_n(u) &= -\frac{u}{\sqrt{n}} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \{\alpha - I[X_{ij} < q(\alpha)]\} + \frac{f(q(\alpha))}{2} u^2 \\
&= \sum_{i=1}^{k} Z_n^{(i)}(u).
\end{aligned}$$

Then

$$\begin{aligned}
&n^{1/4}(Z_n^{(i)}(u) - \bar{Z}_n^{(i)}(u)) \\
&= n^{-1/4} \sum_{j=1}^{n_i} \int_0^u \left[\left\{I[X_{ij} \le q(\alpha) + t/\sqrt{n}] - I[X_{ij} \le q(\alpha)]\right\} - f(q(\alpha))\frac{t}{\sqrt{n}}\right] dt \\
&= \int_0^u Y_n^{(i)}(t)\, dt
\end{aligned}$$

and

$$n^{1/4}(Z_n(u) - \bar{Z}_n(u)) = \sum_{i=1}^{k} \int_0^u Y_n^{(i)}(t)\, dt$$

where $Y_n^{(i)} \xrightarrow{d} -\lambda_i^{1/4} B_i(\lambda_i^{1/2} \cdot)$ and $B_1, \cdots, B_k$ are independent Gaussian processes. Then following Knight (1998), we have

$$n^{1/4} \widehat{R}_n(\alpha) \xrightarrow{d} \frac{1}{f(q(\alpha))} \sum_{i=1}^{k} \lambda_i^{1/4} B_i\left(\frac{\lambda_i^{1/2} W}{f(q(\alpha))}\right)$$

and

$$\begin{aligned}
n^{1/4} \widehat{R}_n^{(i)}(\alpha) &= n^{1/4}\left[\sqrt{n}(\widehat{q}_n^{(i)}(\alpha) - q(\alpha)) - \frac{1}{\lambda_i f(q(\alpha))\sqrt{n}} \sum_{j=1}^{n_i} \{\alpha - I[X_{ij} < q(\alpha)]\}\right] \\
&\xrightarrow{d} \frac{1}{\lambda_i^{3/4} f(q(\alpha))} B_i\left(\frac{W_i}{f(q(\alpha))}\right)
\end{aligned}$$

where

$$\frac{1}{\sqrt{n_i}}\sum_{j=1}^{n_i}\{\alpha - I[X_{ij} < q(\alpha)]\} \xrightarrow{d} W_i$$

(as $n_i \to \infty$) and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\{\alpha - I[X_{ij} < q(\alpha)]\} \xrightarrow{d} W = \sum_{i=1}^{k}\lambda_i^{1/2}W_i.$$

Thus

$$\begin{aligned} n^{1/4}\bar{R}_n(\alpha) &= n^{1/4}\sum_{i=1}^{k}\lambda_i\widehat{R}_n^{(i)}(\alpha) + o_p(1) \\ &\xrightarrow{d} \frac{1}{f(q(\alpha))}\sum_{i=1}^{k}\lambda_i^{1/4}B_i\left(\frac{W_i}{f(q(\alpha))}\right), \end{aligned}$$

which completes the proof. □

Using properties of the two-sided Brownian motions, we have

$$\begin{aligned} \mathrm{Var}[\widehat{R}_0(\alpha)] &= \frac{[2\alpha(1-\alpha)]^{1/2}}{f^2(q(\alpha))\sqrt{\pi}} \\ \mathrm{Var}[\bar{R}_0(\alpha)] &= \frac{[2\alpha(1-\alpha)]^{1/2}}{f^2(q(\alpha))\sqrt{\pi}}\sum_{i=1}^{k}\lambda_i^{1/2} \\ \mathrm{Cov}[\widehat{R}_0(\alpha),\bar{R}_0(\alpha)] &= \frac{[2\alpha(1-\alpha)]^{1/2}}{2f^2(q(\alpha))\sqrt{\pi}}\sum_{i=1}^{k}\lambda_i^{1/2}\left[1+\lambda_i^{1/2}-(1-\lambda_i)^{1/2}\right]. \end{aligned}$$

Note that $\mathrm{Var}[\widehat{R}_0(\alpha)] < \mathrm{Var}[\bar{R}_0(\alpha)]$. (Duttweiler (1973) shows that if $F(x)$ is twice differentiable at $x = q(\alpha)$ then

$$\sqrt{n}E[\widehat{R}_n^2(\alpha)] = \mathrm{Var}(\widehat{R}_0(\alpha)) + o(n^{-1/4+\delta})$$

for any $\delta > 0$.) Moreover, both $\widehat{R}_0(\alpha)$ and $\bar{R}_0(\alpha)$ are uncorrelated with $W_1, \cdots, W_k$:

$$\begin{aligned} \mathrm{Cov}(\widehat{R}_0(\alpha), W_i) &= 0 \\ \mathrm{Cov}(\bar{R}_0(\alpha), W_i) &= 0 \end{aligned}$$

for $i = 1, \cdots, k$.

Since $\widehat{q}_n(\alpha)$ and $\bar{q}_n(\alpha)$ are equivalent to first order, any convex combination of the two will have the same first order representation although the variance of the second order term will vary. For some $t \in [0,1]$ define

$$\widetilde{q}_n(\alpha) = t\,\widehat{q}_n(\alpha) + (1-t)\,\bar{q}_n(\alpha)$$

and note that

$$\sqrt{n}(\tilde{q}_n(\alpha) - q(\alpha)) = \frac{1}{f(q(\alpha))\sqrt{n}} \sum_{i=1}^{n} \{\alpha - I[X_i < q(\alpha)]\} + t\,\widehat{R}_n(\alpha) + (1-t)\,\bar{R}_n(\alpha)$$

with

$$n^{1/4}\left[t\,\widehat{R}_n(\alpha) + (1-t)\,\bar{R}_n(\alpha)\right] \xrightarrow{d} t\,\widehat{R}_0(\alpha) + (1-t)\,\bar{R}_0(\alpha).$$

Since the remainder term above is asymptotically uncorrelated with the first order term, we can try to minimize the asymptotic variance of the remainder term; note that $\mathrm{Var}[t\,\widehat{R}_0(\alpha) + (1-t)\,\bar{R}_0(\alpha)]$ is minimized (for given $\lambda_1, \cdots, \lambda_k$) at

$$t(\lambda_1, \cdots, \lambda_k) = \frac{\mathrm{Var}[\bar{R}_0(\alpha)] - \mathrm{Cov}[\widehat{R}_0(\alpha), \bar{R}_0(\alpha)]}{\mathrm{Var}[\bar{R}_0(\alpha)] + \mathrm{Var}[\widehat{R}_0(\alpha)] - 2\mathrm{Cov}[\widehat{R}_0(\alpha), \bar{R}_0(\alpha)]}.$$

We can also minimize $\mathrm{Var}[t\,\widehat{R}_0(\alpha)+(1-t)\,\bar{R}_0(\alpha)]$ over all t, k, and non-negative λ_i's satisfying $\lambda_1 + \cdots + \lambda_k = 1$.

THEOREM 3. $\mathrm{Var}[t\,\widehat{R}_0(\alpha) + (1-t)\,\bar{R}_0(\alpha)]$ is minimized at $k = 2$, $\lambda_1 = \lambda_2 = 1/2$ and $t = 1/\sqrt{2}$.

Proof. For fixed k and t, $\mathrm{Var}[t\,\widehat{R}_0(\alpha) + (1-t)\,\bar{R}_0(\alpha)]$ is a symmetric function of $\lambda_1, \cdots, \lambda_k$ so we can assume that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$. Moreover, we can focus on local minima for which $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$. For fixed t and k, the local minima of $\mathrm{Var}[t\,\widehat{R}_0(\alpha) + (1-t)\,\bar{R}_0(\alpha)]$ occur at $\lambda_1 = \cdots = \lambda_k = 1/k$, in which case

$$\begin{aligned}&\mathrm{Var}[t\,\widehat{R}_0(\alpha) + (1-t)\,\bar{R}_0(\alpha)] \\ &= \frac{[2\alpha(1-\alpha)]^{1/2}}{f^2(q(\alpha))\sqrt{\pi}}\left[t^2 + (1-t)^2\sqrt{k} + t(1-t)\left(1 + \sqrt{k} - \sqrt{k-1}\right)\right].\end{aligned}$$

For given k, this is minimized over t at $t_k = (\sqrt{k} + \sqrt{k-1} - 1)/(2\sqrt{k-1})$, yielding

$$\mathrm{Var}[t_k\,\widehat{R}_0(\alpha) + (1-t_k)\,\bar{R}_0(\alpha)] = \frac{[2\alpha(1-\alpha)]^{1/2}}{f^2(q(\alpha))\sqrt{\pi}} \frac{\sqrt{k(k-1)} + \sqrt{k} + \sqrt{k-1} - k}{2\sqrt{k-1}}$$

which in turn is minimized over $k = 1, 2, \cdots$ at $k = 2$ with $t_2 = 1/\sqrt{2}$. □

At the optimal values $\lambda_1 = \lambda_2 = 1/2$ and $t_0 = 1/\sqrt{2}$, we have

$$\frac{\mathrm{Var}[t_0\widehat{R}_0(\alpha) + (1-t_0)\bar{R}_0(\alpha)]}{\mathrm{Var}[\widehat{R}_0(\alpha)]} = \sqrt{2} - \frac{1}{2} \approx 0.9142.$$

This ratio can be improved upon by considering average running quantile estimators defined

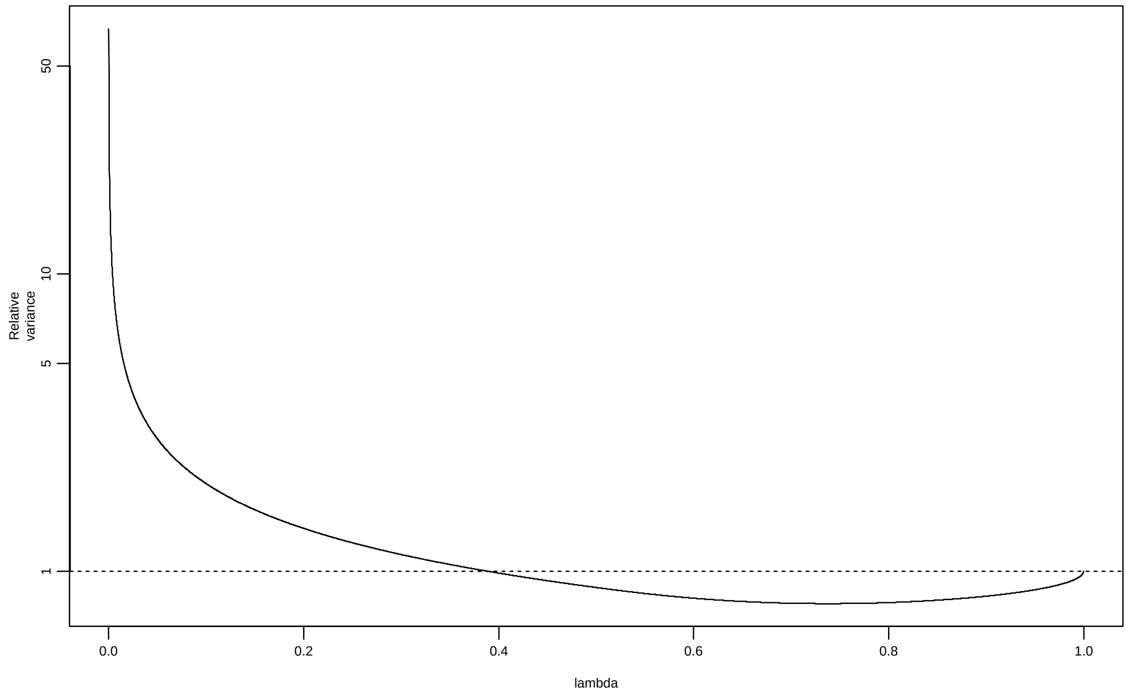

Figure 3: Plot of $\text{Var}[\widetilde{R}_0(\alpha;\lambda)]/\text{Var}[\widehat{R}_0(\alpha)]$ (log scale) versus λ.

at the end of section 2. For fixed $\lambda \in (0,1)$, define $\widetilde{q}_n(\alpha)$ as in (6); then

$$\sqrt{n}(\widetilde{q}_n(\alpha) - q(\alpha)) = \frac{1}{f(q(\alpha))\sqrt{n}} \sum_{i=1}^{n} \{\alpha - I[X_i \leq q(\alpha)]\} + \widetilde{R}_n(\alpha;\lambda)$$

with

$$n^{1/4}\widetilde{R}_n(\alpha;\lambda) \xrightarrow{d} \widetilde{R}_0(\alpha;\lambda) \frac{1}{\lambda^{3/4} f(q(\alpha))} \int_0^1 B_\lambda\left(\frac{W_\lambda(u)}{f(q(\alpha))}; u\right) du$$

where $B_\lambda(t;u)$ is a zero mean Gaussian process with

$$E[B_\lambda(t;u)B_\lambda(s;v)] = \frac{1}{2} f(q(\alpha))\,(|s| + |t| - |s-t|)\, C_\lambda(u,v)$$

and $W_\lambda(u)$ is a Gaussian process (which is independent of $B_\lambda(t;u)$) with $E[W_\lambda(u)W_\lambda(v)] = \alpha(1-\alpha)C_\lambda(u,v)$ where $C_\lambda(u,v)$ is defined in (4). Then

$$\begin{aligned}
&\text{Var}[\widetilde{R}_0(\alpha;\lambda)] \\
&= \frac{1}{\lambda^{3/2} f^2(q(\alpha))} \int_0^1 \int_0^1 E\left[B_\lambda\left(\frac{W_\lambda(u)}{f(q(\alpha))}; u\right) B_\lambda\left(\frac{W_\lambda(v)}{f(q(\alpha))}; v\right)\right] du\, dv \\
&= \frac{1}{2\lambda^{3/2} f^2(q(\alpha))} \int_0^1 \int_0^1 C_\lambda(u,v) \left\{E\left[|W_\lambda(u)| + |W_\lambda(v)| - |W_\lambda(u) - W_\lambda(v)|\right]\right\} du\, dv \\
&= \frac{[2\alpha(1-\alpha)]^{1/2}}{\sqrt{\pi}\lambda^{3/2} f^2(q(\alpha))} \int_0^1 \int_0^1 C_\lambda(u,v) \left\{1 - \frac{1}{2}\left[2 - 2C_\lambda(u,v)\right]^{1/2}\right\} du\, dv.
\end{aligned}$$

The double integral above is not difficult to evaluate as it is easy to see that the inner integral (with respect to u) does not depend on v although there does not appear to be a simple closed-form expression for it. Nonetheless, for given α and λ, it can be easily integrated numerically; a plot of $\mathrm{Var}[\widetilde{R}_0(\alpha;\lambda)]/\mathrm{Var}[\widehat{R}_0(\alpha)]$ versus λ is shown in Figure 3. $\mathrm{Var}[\widetilde{R}_0(\alpha;\lambda)]$ is minimized at $\lambda_0 \approx 0.7377$ with

$$\frac{\mathrm{Var}[\widetilde{R}_0(\alpha;\lambda_0)]}{\mathrm{Var}[\widehat{R}_0(\alpha)]} = 0.7773.$$

The ratio is less than 1 for values of λ in the interval $(0.3880, 1)$ and blows up as $\lambda \downarrow 0$.

The second order behaviour is somewhat different for smoother estimators; in particular, there is an interesting bias/variance tradeoff in the remainder term. For example, suppose that $\widehat{\theta}_n$ minimizes

$$\sum_{i=1}^{n} \rho(X_i;t)$$

where $\rho(x;t)$ is at least three times differentiable, convex function in t for each x. If $\psi(x;t)$, $\psi'(x;t)$, and $\psi''(x;t)$ are the first three derivatives (with respect to t) of $\rho(x;t)$ then (under appropriate regularity conditions), we have

$$\sqrt{n}(\widehat{\theta}_n - \theta) = -\frac{1}{E[\psi'(X_1;\theta)]\sqrt{n}} \sum_{i=1}^{n} \psi(X_i;\theta) + R_n.$$

For the remainder term R_n, we have

$$n^{1/2} R_n \xrightarrow{d} W\,V - \frac{1}{2} W^2 E[\psi''(X_1;\theta)] = R_0$$

where

$$\frac{1}{E[\psi'(X_1;\theta)]\sqrt{n}} \begin{pmatrix} \sum_{i=1}^{n} \psi(X_i;\theta) \\ \sum_{i=1}^{n} \{\psi'(X_i;\theta) - E[\psi'(X_i;\theta)]\} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} W \\ V \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, C).$$

On the other hand, if $\widehat{\theta}_n^{(1)}, \cdots, \widehat{\theta}_n^{(k)}$ are estimators based on subsamples of lengths $n_1, \cdots, n_k$ and

$$\bar{\theta}_n = \sum_{i=1}^{k} \frac{n_i}{n} \widehat{\theta}_n^{(i)}$$

then

$$\sqrt{n}(\bar{\theta}_n - \theta) = -\frac{1}{E[\psi'(X_1;\theta)]\sqrt{n}} \sum_{i=1}^{n} \psi(X_i;\theta) + R'_n$$

where

$$n^{1/2} R'_n \xrightarrow{d} \sum_{i=1}^{k} \lambda_i^{1/2} \left\{ W_i V_i - \frac{1}{2} W_i^2 E[\psi''(X_1;\theta)] \right\} = R'_0$$

and $(W_1, V_1), \cdots, (W_k, V_k)$ are i.i.d. pairs of random variables have the same distribution as (W, V) with

$$\begin{pmatrix} W \\ V \end{pmatrix} = \lambda_1^{1/2} \begin{pmatrix} W_1 \\ V_1 \end{pmatrix} + \cdots + \lambda_k^{1/2} \begin{pmatrix} W_k \\ V_k \end{pmatrix}.$$

Note that $\mathrm{Var}(R_0) = \mathrm{Var}(R_0')$ and

$$\mathrm{Cov}\,(R_0, R_0') = \mathrm{Var}(R_0) \sum_{i=1}^{k} \lambda_i^{3/2},$$

which implies that the variance of the limiting remainder term $t\,R_0 + (1-t)\,R_0'$ for the estimator $t\,\widehat{\theta}_n + (1-t)\,\bar{\theta}_n$ is always minimized when $t = 1/2$. Moreover, $\mathrm{Var}[(R_0 + R_0')/2]$ can be made close to $\mathrm{Var}(R_0)/2$ by taking k large and the λ_i's uniformly small (for example, $\lambda_i = 1/k$). On the other hand, if $E(R_0) \neq 0$ then $|E[t\,R_0 + (1-t)\,R_0']|$ is minimized at $t = 1$.

4 A step beyond: bagging, subagging, and bragging

In the previous section, we saw that we could achieve a second order improvement of a sample quantile by combining it with an estimator constructed by taking averages of subsample quantiles. A possible concern with these estimators is the fact that they are not invariant under permutations of the data and hence are not functions of the order statistics, which, if $X_1, \cdots, X_n$ are i.i.d. with unknown continuous distribution function F, is the sufficient statistic for F. In this section, we will consider estimators based on averaging (essentially) all possible subsample quantiles; these estimators will be invariant under permutations.

Suppose that $\widehat{q}_n^*(\alpha)$ minimizes

$$h_n^*(t) = \sum_{i=1}^{n} \Delta_{ni}^* \rho_\alpha(X_i - t) \tag{10}$$

where $\mathbf{\Delta}_n^* = (\Delta_{n1}^*, \cdots, \Delta_{nn}^*)$ is a random vector independent of the X_i's. We will construct an estimator of $q(\alpha)$ by averaging the $\widehat{q}_n^*(\alpha)$'s over the distribution of $\mathbf{\Delta}_n^*$:

$$\widetilde{q}_n(\alpha) = E^*[\widehat{q}_n^*(\alpha)].$$

If $\mathbf{\Delta}_n^*$ has a multinomial distribution with $\Delta_{n1}^* + \cdots + \Delta_{nn}^* = n$ then $\widetilde{q}_n(\alpha)$ is Breiman's (1996) "bagged" estimator. If the Δ_{ni}^*'s are exchangeable 0/1 random variables with $P^*(\Delta_{ni}^* = 1) = \lambda_n \to \lambda > 0$ and $\Delta_{n1}^* + \cdots + \Delta_{nn}^* = n\lambda_n$ then $\widetilde{q}_n(\alpha)$ is the average of all possible subsample quantiles from subsamples of length $n\lambda_n$; we will call these latter estimators "subagged estimators" after Bühlmann and Yu (2002). Asymptotics for bagged and subagged estimators have been considered by, among others, Bühlmann and Yu (2002), Buja and Stuetzle (2002), and Friedman and Hall (2000). Some properties of the bagged sample median are discussed in Grandvalet (2004). Bühlmann (2003) proposed replacing the averaging of estimators from bootstrap samples by a more robust estimator; he uses the median and calls the resulting procedure "bragging".

If $\widehat{q}_n^*(\alpha)$ minimizes (10) then we have the Bahadur-Kiefer representation

$$\begin{aligned}&\sqrt{n}(\widehat{q}_n^*(\alpha) - q(\alpha))\\ &= \frac{1}{f(q(\alpha))\sqrt{n}} \sum_{i=1}^{n} [\{\alpha - I[X_i \le q(\alpha)]\} + \{\Delta_{ni}^* - E(\Delta_{ni}^*)\}\{\alpha - I[X_i \le q(\alpha)]\}] + \widehat{R}_n^*(\alpha)\end{aligned}$$

where

$$P^* \left[n^{1/4} \widehat{R}_n^*(\alpha) \in A \right] \xrightarrow{d} P^* \left[\frac{1}{f(q(\alpha))} B(W + W^*) \in A \right]$$

where B is a two-sided Brownian motion, W and W^* are independent normal random variables (independent of B), and the randomness in the limiting probability is induced by B and W.

Second order representations for the bagged and subagged estimators can be obtained by averaging over the distribution of W^* conditional on B and W. Both the bagged and subagged estimators of $q(\alpha)$ satisfy

$$\sqrt{n}\,(\widetilde{q}_n(\alpha) - q(\alpha)) = \frac{1}{f(q(\alpha))\sqrt{n}} \sum_{i=1}^{n} \{\alpha - I[X_i \le q(\alpha)]\} + \widetilde{R}_n(\alpha).$$

For the bagged estimator, we have

$$n^{1/4} \widetilde{R}_n(\alpha) \xrightarrow{d} \int_{-\infty}^{\infty} B(W + u)\, \phi_\alpha\left(f(q(\alpha))\, u\right) du$$

where $\phi_\alpha(x)$ is the density function of a $\mathcal{N}(0, \alpha(1-\alpha))$ random variable, $W \sim \mathcal{N}(0, \alpha(1-\alpha)/f^2(q(\alpha)))$, and B is a two-sided Brownian motion with $E[(B(s) - B(t))^2] = f(q(\alpha))|s-t|$.

For the subagged estimators, the limiting distribution of $n^{1/4}\widetilde{R}_n(\alpha)$ depends on the subagging fraction λ; in particular, we have

$$n^{1/4} \widetilde{R}_n(\alpha) \xrightarrow{d} \frac{\lambda^{1/2}}{(1-\lambda)^{1/2}} \int_{-\infty}^{\infty} B(W+u)\phi_\alpha\left(\frac{\lambda^{1/2} f(q(\alpha))}{(1-\lambda)^{1/2}} u\right) du = \widetilde{R}_0(\alpha; \lambda)$$

where B and ϕ_α are defined as before. Note that the limit for the bagged estimator is the same as that of a subagged estimator with $\lambda = 1/2$ and that $\widetilde{R}_0(\alpha; 1) = \widehat{R}_0(\alpha)$, where $\widehat{R}_0(\alpha)$ is defined in the previous section. (More generally, if we take bootstrap samples of size $m \ne n$ then the resulting bagged estimators are equivalent to subagged estimators with $\lambda = m/(m+n)$.)

The variance of $\widetilde{R}_0(\alpha;\lambda)$ is

$$\mathrm{Var}[\widetilde{R}_0(\alpha;\lambda)] = \frac{[2\alpha(1-\alpha)]^{1/2}}{f^2(q(\alpha))\pi^{1/2}}\left[\frac{1}{\lambda^{1/2}} - \frac{(1-\lambda)^{1/2}}{(2\lambda)^{1/2}}\right]$$

and

$$\begin{aligned}\mathrm{Cov}[\widetilde{R}_0(\alpha;\lambda_1), \widetilde{R}_0(\alpha;\lambda_2)] &= \frac{[2\alpha(1-\alpha)]^{1/2}}{2f^2(q(\alpha))\pi^{1/2}}\left[\frac{1}{\lambda_1^{1/2}} + \frac{1}{\lambda_2^{1/2}} - \left(\frac{\lambda_1+\lambda_2-2\lambda_1\lambda_2}{\lambda_1\lambda_2}\right)^{1/2}\right] \\ &= K(\lambda_1,\lambda_2).\end{aligned}$$

The variance above is minimized at $\lambda = 1/2$ and so bagging or subagging with fraction 1/2 is optimal in this sense. In fact, we can go a step further. Defining $\widetilde{q}_n(\alpha;\lambda)$ to be the subagged estimator with fraction λ, if we define the estimator

$$\breve{q}_n(\alpha;\mu) = \int_{[0,1]} \widetilde{q}_n(\alpha;\lambda)\,\mu(d\lambda)$$

for a probability measure μ (or signed measure with $\int_{[0,1]}\mu(d\lambda) = 1$) then

$$\sqrt{n}(\breve{q}_n(\alpha;\mu) - q(\alpha)) = \frac{1}{f(q(\alpha))\sqrt{n}}\sum_{i=1}^{n}\{\alpha - I[X_i \le q(\alpha)]\} + \breve{R}_n(\alpha;\mu)$$

where

$$n^{1/4}\breve{R}_n(\alpha;\mu) \xrightarrow{d} \int_{[0,1]} \widetilde{R}_0(\alpha;\lambda)\,\mu(d\lambda) = \breve{R}_0(\alpha;\mu).$$

Defining

$$\begin{aligned}\varphi(\mu) &= \mathrm{Var}[\breve{R}_0(\alpha;\mu)] \\ &= \int_0^1\int_0^1 K(\lambda_1,\lambda_2)\,\mu(d\lambda_1)\,\mu(d\lambda_2),\end{aligned}$$

it follows that φ is minimized at the measure putting all its mass at 1/2; this follows since for any signed measure μ with $\int_{[0,1]}\mu(d\lambda) = 1$,

$$\begin{aligned}\mathrm{Cov}[\widetilde{R}_0(\alpha;1/2), \breve{R}_0(\alpha;\mu)] &= \int_{[0,1]} K(\lambda,1/2)\,\mu(d\lambda) \\ &= \int_{[0,1]} K(1/2,1/2)\,\mu(d\lambda) \\ &= \mathrm{Var}[\widetilde{R}_0(\alpha;1/2)]\end{aligned}$$

(since $K(\lambda,1/2) = K(1/2,\lambda) = K(1/2,1/2)$ for $0 \le \lambda \le 1$) and so

$$\mathrm{Var}[\widetilde{R}_0(\alpha;1/2)] \le \mathrm{Var}[\breve{R}_0(\alpha;\mu)]$$

by applying the Cauchy-Schwarz inequality. This suggests that bagging (or subagging with $\lambda = 1/2$) may be optimal from the point of view of reducing the nonlinearity of the sample quantile. Note that

$$\frac{\mathrm{Var}[\widetilde{R}_0(\alpha; 1/2)]}{\mathrm{Var}[\widehat{R}_0(\alpha)]} = \frac{1}{\sqrt{2}} \approx 0.7071,$$

which is smaller than the best ratio given in section 3.

Both the bagged and subagged estimators can be thought of as L-estimators of $q(\alpha)$ in the sense that they are weighted averages of order statistics. Suppose we define

$$\widetilde{q}_n(\alpha) = \int \widehat{q}_n(\alpha + t/\sqrt{n})\, \nu_n(dt) \tag{11}$$

where $\{\nu_n\}$ is a sequence of probability measures (or signed measures with $\int \nu_n(dt) = 1$ for all n) converging weakly to ν; that is, such that $\nu_n(A) \to \nu(A)$ for A such that $\nu(\partial A) = 0$. For bagging and subagging, the measure ν corresponds to a zero mean normal distribution whose variance depends on α and λ; for example, for subagging with fraction λ, this variance is $\alpha(1-\alpha)\lambda/(1-\lambda)$. (See also Grandvalet (2004).) First order equivalence of the estimators defined in (11) with the sample α quantile $\widehat{q}_n(\alpha)$ follows if $\int t\, \nu_n(dt) \to 0$; we then obtain the representation

$$\sqrt{n}(\widetilde{q}_n(\alpha) - q(\alpha)) = \frac{1}{f(q(\alpha))\sqrt{n}} \sum_{i=1}^{n} \{\alpha - I[X_i \le q(\alpha)]\} + \widetilde{R}_n(\alpha; \nu_n)$$

where

$$n^{1/4}\widetilde{R}_n(\alpha; \nu_n) \xrightarrow{d} \frac{1}{f(q(\alpha))} \int B(W + t/f(q(\alpha)))\, \nu(dt) = \widetilde{R}_0(\alpha; \nu)$$

where B and W are defined as above. In this case, the asymptotic variance $\mathrm{Var}[\widetilde{R}_0(\alpha; \nu)]$ is minimized over signed measures ν at the probability measure ν_0 corresponding to a normal distribution with mean 0 and variance $\alpha(1-\alpha)$; note that this optimal measure ν is the same as the limiting measure used for bagging. The proof of this follows by noting that

$$\mathrm{Cov}\left[\widetilde{R}_0(\alpha; \nu_0), \widetilde{R}_0(\alpha; \nu)\right] = \mathrm{Var}\left[\widetilde{R}_0(\alpha; \nu_0)\right]$$

for any probability measure ν; as before, this implies that

$$\mathrm{Var}\left[\widetilde{R}_0(\alpha; \nu_0)\right] \le \mathrm{Var}\left[\widetilde{R}_0(\alpha; \nu)\right]$$

with equality if, and only if, $\nu = \nu_0$.

An estimator $\widetilde{q}_n(\alpha)$ of the form (11) can itself be bootstrapped, for example, to estimate either its standard error or the standard error of the sample quantile $\widehat{q}_n(\alpha)$. Hall and Martin

(1988) show that if $\widehat{\sigma}_n^2(\alpha)$ is the variance of the bootstrap distribution of the sample quantile $\widehat{q}_n(\alpha)$ then

$$n^{1/4}\left(n\widehat{\sigma}_n^2(\alpha) - \frac{\alpha(1-\alpha)}{f^2(q(\alpha))}\right) \xrightarrow{d} -2\int_{-\infty}^{\infty} w\,B(w)\phi_\alpha(f(q(\alpha))\,w)\,dw$$
$$\sim \mathcal{N}\left(0, \frac{2[\alpha(1-\alpha)]^{3/2}}{\pi^{1/2}f^4(q(\alpha))}\right).$$

Using similar techniques for the estimators $\widetilde{q}_n(\alpha)$ with $\nu_n \xrightarrow{d} \nu$ and $\int t\,\nu_n(dt) \to 0$, the bootstrap variance estimator $\widetilde{\sigma}_n^2(\alpha)$ satisfies

$$n^{1/4}\left(n\widetilde{\sigma}_n^2(\alpha) - \frac{\alpha(1-\alpha)}{f^2(q(\alpha))}\right)$$
$$\xrightarrow{d} -2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w\,B(w + t/f(q(\alpha)))\phi_\alpha(f(q(\alpha))\,w)\,dw\,\nu(dt)$$
$$\sim \mathcal{N}\left(0, \gamma^2(\alpha;\nu)\right)$$

where

$$\gamma^2(\alpha;\nu)$$
$$= -\frac{2}{f^4(q(\alpha))}\int\int\int\int w\,w'|(w-w')+(t-t')|\phi_\alpha(w)\phi_\alpha(w')\,dw\,dw'\,\nu(dt)\nu(dt')$$

where, as before, $\phi_\alpha(x)$ is the density of a $\mathcal{N}(0, \alpha(1-\alpha))$ random variable.

Perhaps surprisingly, the variance $\gamma^2(\alpha;\nu)$ decreases as the dispersion of ν increases. More precisely, for a given measure ν, define ν_τ by $\nu_\tau(A) = \nu(\tau^{-1}A)$; then $\gamma^2(\alpha;\nu_\tau)$ is a decreasing function of τ. This suggests that the bootstrap standard error estimator improves by taking a weighted average of order statistics within a wider window; however, the bias of $\widetilde{q}_n(\alpha)$ as an estimator of $q(\alpha)$ will typically also increase with the dispersion of ν_n and this tradeoff must be taken into account in choosing ν_n. However, bootstrapping an estimator $\widetilde{q}_n(\alpha)$ of the form (11) in order to estimate the sampling distribution or standard error of $\widehat{q}_n(\alpha)$ is similar in spirit (although not equivalent) to the smoothed bootstrap (where bootstrap samples are drawn from a smoothed empirical distribution), which is known to have some attractive properties for sample quantiles.

Analysis of the bragged estimator (Bühlmann, 2003) is somewhat different. Here we will consider the general case where $\widetilde{q}_n(\alpha)$ minimizes

$$E^*\left[|\widehat{q}_n^*(\alpha) - t|^p\right]$$

where E^* denotes expectation with respect to the bootstrap distribution and $p \geq 1$; for $p = 2$, we obtained the bagged estimator while for $p = 1$, we obtain the median of the bootstrap distribution. (We can also replace $|\cdot|^p$ by some other symmetric loss function but for the asymptotics, it is the behaviour of the loss function near 0 that is important.) For each $p \geq 1$, we still obtain the representation

$$\sqrt{n}(\tilde{q}_n(\alpha) - q(\alpha)) = \frac{1}{f(q(\alpha))\sqrt{n}} \sum_{i=1}^{n} \{\alpha - I[X_i \leq q(\alpha)]\} + \widetilde{R}_n(\alpha)$$

where now

$$\begin{aligned} & n^{1/4}\widetilde{R}_n(\alpha) \\ \xrightarrow{d} \quad & \left(f(q(\alpha)) \int_{-\infty}^{\infty} |u|^{p-2} \phi_\alpha(f(q(\alpha))\, u)\, du \right)^{-1} \int_{-\infty}^{\infty} B(W+u)|u|^{p-2}\phi_\alpha(f(q(\alpha))\, u)\, du \\ = \quad & \frac{\pi^{1/2} f(q(\alpha))^{p-2}}{[2\alpha(1-\alpha)]^{(p-2)/2}\Gamma((p-1)/2)} \int_{-\infty}^{\infty} B(W+u)|u|^{p-2}\phi_\alpha(f(q(\alpha))\, u)\, du \end{aligned}$$

for $p > 1$ and

$$n^{1/4}\widehat{R}_n(\alpha) \xrightarrow{d} \frac{1}{f(q(\alpha))} B(W)$$

where the two-sided Brownian motion B and the normal random variable W are defined as above. Note that the limiting distribution of the remainder in the case of bragging with the median is identical to that for the sample quantile itself. For $p > 1$, bragging is asymptotically equivalent (to second order) to estimators of the form (11) where $\{\nu_n\}$ converges weakly to a probability measure ν whose density is

$$\varphi_\nu(t) = \frac{\pi^{1/2}}{[2\alpha(1-\alpha)]^{(p-2)/2}\Gamma((p-1)/2)} |t|^{p-2}\phi_\alpha(t).$$

It is worth noting that the effect bagging and subagging on smoother estimators is somewhat different. As in section 3, suppose that $\widehat{\theta}_n$ minimizes

$$\sum_{i=1}^{n} \rho(X_i; t)$$

where $\rho(x;t)$ is at least three times differentiable, convex function in t for each x. If $\psi(x;t)$, $\psi'(x;t)$, and $\psi''(x;t)$ are the first three derivatives (with respect to t) of $\rho(x;t)$ then (under appropriate regularity conditions), we have

$$\sqrt{n}(\widehat{\theta}_n - \theta) = -\frac{1}{E[\psi'(X_1;\theta)]\sqrt{n}} \sum_{i=1}^{n} \psi(X_i;\theta) + R_n.$$

For the remainder term R_n, we have

$$n^{1/2} R_n \xrightarrow{d} W\,V - \frac{1}{2} W^2 E[\psi''(X_1;\theta)] = R_0$$

where

$$\frac{1}{E[\psi'(X_1;\theta)]\sqrt{n}}\begin{pmatrix}\sum_{i=1}^{n}\psi(X_i;\theta)\\ \sum_{i=1}^{n}\{\psi'(X_i;\theta)-E[\psi'(X_i;\theta)]\}\end{pmatrix}\xrightarrow{d}\begin{pmatrix}W\\ V\end{pmatrix}\sim\mathcal{N}(\mathbf{0},C).$$

Defining $\widetilde{\theta}_n$ to minimize

$$\sum_{i=1}^{n}\Delta^*_{ni}\rho(X_i;t)$$

for the Δ^*_{ni}'s defined for bagging and subagging, we obtain

$$\sqrt{n}(\widetilde{\theta}_n-\theta)=-\frac{1}{E[\psi'(X_1;\theta)]\sqrt{n}}\sum_{i=1}^{n}\psi(X_i;\theta)+R'_n$$

where the limiting distribution of $\sqrt{n}R'_n$ depends on the type of subsampling. For bagging, we obtain

$$\sqrt{n}R'_n\xrightarrow{d}W\,V-\frac{1}{2}W^2E[\psi''(X_1;\theta)]+E(W\,V)-\frac{1}{2}E(W^2)E[\psi''(X_1;\theta)]$$

while for subagging with fraction λ, we get

$$\sqrt{n}R'_n\xrightarrow{d}W\,V-\frac{1}{2}W^2E[\psi''(X_1;\theta)]+\frac{1-\lambda}{\lambda}\left\{E(W\,V)-\frac{1}{2}E(W^2)E[\psi''(X_1;\theta)]\right\}.$$

From this analysis, it appears that bagging and subagging can have only a negative effect (by increasing the bias) of the estimator. A more complete analysis (c.f. Friedman and Hall, 2000; Chen and Hall, 2001) shows that bagging and subagging can improve the mean square error properties of non-linear estimators. Buja and Stuetzle (2002) examine the smoother effects of bagging and subagging on estimators $\widehat{\theta}_n$ having the form $\widehat{\theta}_n=\vartheta(X_1,\cdots,X_n)$ where $\vartheta(\cdot)$ is symmetric in its arguments; they show that the von Mises expansion of the appropriate functional is always finite.

5 Non-standard conditions

The differentiability condition on F assumed to this point can be generalized; the techniques used to determine limiting distributions remain more or less the same although the limiting distributions themselves will change. Define $\psi_\delta(t)$ to be a non-decreasing function satisfying (for some $\delta\geq 0$)

$$\begin{aligned}\psi_\delta(t)&\to\pm\infty\quad\text{as } t\to\pm\infty\\ \psi_\delta(a\,t)&=a^{1/\delta}\psi_\delta(t)\quad\text{for } a>0.\end{aligned}\tag{12}$$

When $\delta > 0$, ψ_δ has the form

$$\psi_\delta(t) = \begin{cases} c_+ t^{1/\delta} & \text{for } t > 0, \\ -c_-(-t)^{1/\delta} & \text{for } t < 0, \end{cases} \tag{13}$$

where $0 < c_+, c_- \leq \infty$ and at most one of c_+ and c_- is infinite. ψ_0 has the form

$$\psi_0(t) = \begin{cases} 0 & \text{for } -a_- < t < a_+, \\ +\infty & \text{for } t > a_+, \\ -\infty & \text{for } t < -a_-, \end{cases} \tag{14}$$

where $0 \leq a_+, a_- < \infty$ and at most one of a_+ and a_- is 0.

Given $\psi_\delta(t)$, suppose that for some sequence of constants $\{a_n\}$, we have

$$\lim_{n\to\infty} \sqrt{n}(F(q(\alpha) + t/a_n) - \alpha) \to \psi_\delta(t). \tag{15}$$

where ψ_δ is defined as in (13) or (14); the constants a_n are of the form $a_n = n^{\delta/2}L(n)$ where $L(n)$ is a slowly varying function. If F has a density f then $\delta < 1$ implies that $f(q(\alpha)) = 0$ while if $\delta > 1$ then $f(x) \uparrow \infty$ as $x \to q(\alpha)$.

The form of $\psi_\delta(t)$ and $\{a_n\}$ determine, respectively, the limiting distribution and the convergence rate of the sample quantile $\widehat{q}_n(\alpha)$ minimizing (1); in particular, we have

$$a_n(\widehat{q}_n(\alpha) - q(\alpha)) \xrightarrow{d} \psi_\delta^{\leftrightarrow}(W)$$

where $W \sim \mathcal{N}(0, \alpha(1-\alpha))$ and

$$\psi_\delta^{\leftrightarrow}(x) = \begin{cases} \inf\{t \leq 0 : \psi_\delta(t) \geq x\} & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ \sup\{t \geq 0 : \psi_\delta(t) \leq x\} & \text{if } x > 0. \end{cases} \tag{16}$$

When $\psi_\delta(t)$ is continuous (and strictly increasing) $\psi_\delta^{\leftrightarrow} = \psi_\delta^{-1}$. The scaling condition (12) for ψ_δ implies that each $\psi_\delta^{\leftrightarrow}$ satisfies a scaling condition: For each $a > 0$,

$$\psi_\delta^{\leftrightarrow}(a\,x) = a^\delta\,\psi_\delta^{\leftrightarrow}(x) \tag{17}$$

for some $\delta \geq 0$.

THEOREM 4. Suppose that $F(x)$ satisfies (15) at $x = q(\alpha)$ and

$$\widetilde{q}_n(\alpha) = g_n\left(\widehat{q}_n^{(1)}(\alpha), \cdots, \widehat{q}_n^{(k)}(\alpha)\right)$$

where

(a) $n_i/n \to \lambda_i > 0$ as $n \to \infty$ for $i = 1, \cdots, k$,

(b) $g_n(\boldsymbol{x}_n) \to g_0(\boldsymbol{x}_0)$ for all sequences $\{\boldsymbol{x}_n\}$ converging to $\boldsymbol{x}_0$, and
(c) $\{g_n\}$ and g_0 are equivariant under non-decreasing transformations:

$$g_n(\phi(x_1), \cdots, \phi(x_k)) = \phi(g_n(x_1, \cdots, x_k))$$

for all non-decreasing functions ϕ.
Then

$$a_n(\widetilde{q}_n(\alpha) - q(\alpha)) \xrightarrow{d} \psi_\delta^{\leftrightarrow}(W + V)$$

where $\psi_\delta^{\leftrightarrow}$ is defined in (16), $W \sim \mathcal{N}(0, \alpha(1-\alpha))$ and V is independent of W.

The proof of Theorem 4 follows along the same lines as that of Theorem 1 noting that

$$a_n(\widehat{q}_n^{(i)}(\alpha) - q(\alpha)) \xrightarrow{d} \psi_\delta^{\leftrightarrow}\left(\lambda_i^{-1/2} W_i\right)$$

for independent $\mathcal{N}(0, \alpha(1-\alpha))$ random variables $W_1, \cdots, W_k$. Theorem 4 can also be extended to cases where $\widehat{q}_n^{(1)}(\alpha), \cdots, \widehat{q}_n^{(k)}(\alpha)$ are obtained from balanced overlapping subsamples as in section 2.

The weighted estimator $\bar{q}_n(\alpha)$ defined in (7) does not satisfy the equivariance condition in Theorem 3. For this estimator (under the assumptions of Theorem 3), we have

$$\begin{aligned} a_n(\bar{q}_n(\alpha) - q(\alpha)) \quad &\xrightarrow{d} \quad \sum_{i=1}^{k} \lambda_i \psi_\delta^{\leftrightarrow}\left(\lambda_i^{-1/2} W_i\right) \\ &= \quad \sum_{i=1}^{k} \lambda_i^{1-\delta/2} \psi_\delta^{\leftrightarrow}(W_i) \end{aligned}$$

using the scaling condition (17). When $\psi_\delta^{\leftrightarrow}$ is an odd function (that is, $c_+ = c_-$ in (13) or $a_+ = a_-$ in (14)) then $E[\psi_\delta^{\leftrightarrow}(W_i)] = 0$, in which case, we can say that $\widehat{q}_n(\alpha)$ (and $\bar{q}_n(\alpha)$) are asymptotically unbiased to first order (that is, to order $O_p(a_n^{-1})$); otherwise, the first order asymptotic bias of $\widehat{q}_n(\alpha)$ is $\mu = E[\psi_\delta^{\leftrightarrow}(W)] = E[\psi_\delta^{\leftrightarrow}(W_i)]$. In this case, we have

$$E\left[\sum_{i=1}^{k} \lambda_i^{1-\delta/2} \psi_\delta^{\leftrightarrow}(W_i)\right] = \mu \sum_{i=1}^{k} \lambda_i^{1-\delta/2}$$

and so the asymptotic bias of $\bar{q}_n(\alpha)$ is worse than that of $\widehat{q}_n(\alpha)$ unless $\mu = 0$ or $\delta = 0$. It follows that

$$\text{Var}\left(\sum_{i=1}^{k} \lambda_i^{1-\delta/2} \psi_\delta^{\leftrightarrow}(W_i)\right) = \sum_{i=1}^{k} \lambda_i^{2-\delta} \text{Var}(\psi_\delta^{\leftrightarrow}(W_i)) \quad \begin{cases} < \ \text{Var}(\psi_\delta^{\leftrightarrow}(W)) & \text{if } \delta < 1 \\ > \ \text{Var}(\psi_\delta^{\leftrightarrow}(W)) & \text{if } \delta > 1. \end{cases}$$

The fact that the variance of the limiting distribution of $a_n(\bar{q}_n(\alpha) - q(\alpha))$ tends to 0 with k when $\delta < 1$ suggests that we could improve on the sample quantile by averaging a large number of subsample quantiles; this is true provided that the bias of each subsample quantile is not too severe.

Similar results hold for bagged and subagged estimators. In the case of the bagged estimator, we have

$$a_n(\widetilde{q}_n(\alpha) - q(\alpha)) \xrightarrow{d} \int_{-\infty}^{\infty} \psi_\delta^{\leftrightarrow}(W + u)\phi_\alpha(u)\, du$$

while for subagging (with fraction λ) we have

$$a_n(\widetilde{q}_n(\alpha) - q(\alpha)) \xrightarrow{d} \frac{\lambda^{1/2}}{(1-\lambda)^{1/2}} \int_{-\infty}^{\infty} \psi_\delta^{\leftrightarrow}(W + u)\phi_\alpha\left(\frac{\lambda^{1/2}u}{(1-\lambda)^{1/2}}\right) du = h_{\lambda,\delta}(W).$$

In this case, we have a first order asymptotic equivalence between bagging and subagging with $\lambda = 1/2$ and note that as $\lambda \uparrow 1$, we have

$$h_{\lambda,\delta}(W) \to \psi_\delta^{\leftrightarrow}(W),$$

which is the limiting distribution of the sample quantile. Likewise, if $\widetilde{q}_n(\alpha)$ is defined as the L-estimator in (11) with $\{\nu_n\}$ converging weakly to ν, we have

$$a_n(\widetilde{q}_n(\alpha) - q(\alpha)) \xrightarrow{d} \int \psi_\delta^{\leftrightarrow}(W + t)\, \nu(dt).$$

It is interesting to look at the case where ψ_δ is an odd function. Identifying $h_{1,\delta}(W) = \psi_\delta^{\leftrightarrow}(W)$, it is possible to show that for $0 < \delta < 1$ and $0 < \lambda < 1$,

$$0 \le h'_{\lambda,\delta}(W) < h'_{1,\delta}(W)$$

while for $\delta > 1$ and $0 < \lambda < 1$, we have

$$h'_{\lambda,\delta}(W) > h'_{1,\delta}(W)$$

where $h'_{\lambda,\delta}$ is the derivative of $h_{\lambda,\delta}$. Since $h_{\lambda,\delta}(0) = 0$ for all λ and δ, this implies that bagging and subagging provide (asymptotically) a contraction of the limiting distribution (compared to not using bagging or subagging) when $\delta < 1$ and an "expansion" when $\delta > 1$; likewise, when $\delta = 0$, it's easy to show that $|h_{\lambda,0}(W)| \le |\psi_0^{\leftrightarrow}(W)|$. In particular, for any non-negative "bowl shaped" function ℓ (that is, the set $\{x : \ell(x) \le y\}$ is symmetric and convex for each $y > 0$), we have for $0 < \lambda < 1$

$$\begin{aligned} E\left[\ell\left(h_{\lambda,\delta}(W)\right)\right] &< E\left[\ell\left(\psi_\delta^{\leftrightarrow}(W)\right)\right] \quad \text{if } \delta < 1, \text{ and} \\ E\left[\ell\left(h_{\lambda,\delta}(W)\right)\right] &> E\left[\ell\left(\psi_\delta^{\leftrightarrow}(W)\right)\right] \quad \text{if } \delta > 1. \end{aligned}$$

This confirms the observation made for estimators obtained by combining subsamples; when $\delta < 1$ (for example, if $f(q(\alpha)) = 0$) then the efficiency of the sample quantile can be improved by combining estimators from subsamples while when $\delta > 1$, this approach does not produce a more efficient estimator.

6 Final observations

In this paper, we have studied the asymptotic properties of quantile estimators constructed by combining estimators from subsamples. There are several possible motivations for using such estimators, one of which is computational; in very large datasets, it may be impossible to compute an exact sample quantile so approximation using subsample quantiles seems a reasonable thing to do. Another motivation is more statistically oriented; given the non-smoothness of sample quantiles, we may wish to consider asymptotically equivalent estimators that are smoother in the sense that they are less sensitive to small changes in the data as well as having "nicer" statistical properties, for example, better asymptotic linearity. Finally, in some studies, the only available data is summary data on quantiles as the raw data may be unavailable for confidentiality or other reasons. This paper was motivated, in part, by Bassett *et al.* (2002), in which regression quantile estimation of ACT scores for Illinois high school students is compared to least squares estimation of conditional quantiles using within school quantiles of the ACT scores.

References

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

Bassett, G.W., Tam, M.-Y. and Knight, K. (2002) Quantile models and estimators for data analysis. *Metrika.* **55**, 17-26.

Breiman, L. (1996) Bagging predictors. *Machine Learning.* **24**, 123-140.

Bühlmann, P. (2003) Bagging, subagging and bragging for improving prediction algorithms. In Recent Advances and Trends in Nonparametric Statistics (eds. Akritas, M.G. and Politis, D.N.), pp. 19-34. Elsevier.

Bühlmann, P. and Yu, B. (2002) Analyzing bagging. *Annals of Statistics.* **30**, 927-961.

Buja, A. and Stuetzle, W. (2002) Observations on bagging. (unpublished manuscript)

Chen, S.X. and Hall, P. (2001) Effects of bagging and bias correction on estimators defined by estimating equations. (unpublished manuscript)

Dor, D. and Zwick, U. (1999) Selecting the median. *SIAM Journal on Computing.* **28**, 1722-1758.

Duttweiler, D.L. (1973) The mean-square error of Bahadur's order-statistic approximation. *Annals of Statistics.* **1**, 446-453.

Efron, B. (1982) *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: SIAM.

Friedman, J. and Hall, P. (2000) On bagging and nonlinear estimation. (unpublished manuscript)

Grandvalet, Y. (2004) Bagging equalizes influence. *Machine Learning.* to appear.

Hall, P. and Martin, M.A. (1988) Exact convergence rate of bootstrap quantile variance estimator. *Probability Theory and Related Fields.* **80**, 261-288.

Hurley, C. and Modarres, R. (1995) Low-storage quantile estimation. *Computational Statistics.* **10**, 311-325.

Kiefer, J. (1967) On Bahadur's representation of sample quantiles. *Annals of Mathematical Statistics.* **38**, 1323-1342.

Knight, K. (1998) A delta method approach to Bahadur-Kiefer theorems. *Scandinavian Journal of Statistics.* **25**, 555-568.

Knight, K. (2002) What are the limiting distributions of quantile estimators? In *Statistical Data Analysis Based on the L_1-Norm and Related Methods.* (Y. Dodge, editor) Basel: Birkhäuser, 47-66.

Rousseeuw, P.J. and Bassett, G.W. (1990) The remedian: a robust averaging method for large data sets. *Journal of the American Statistical Association.* **85**, 97-104.

Smirnov, N.V. (1952) Limit distributions for the terms of a variational series. *American Mathematical Society Translations.* no. 67.

Tukey, J.W. (1978) The ninther: a technique for low-effort robust (resistant) location in large samples. In *Contributions to Survey Sampling and Applied Statistics in Honor of H.O. Hartley*, ed. H.A. David. New York: Academic Press, 251-257.