Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

ON THE ASYMPTOTIC ANALYSIS OF QUANTILE SENSITIVITY ESTIMATION BY MONTE CARLO SIMULATION

Yijie Peng

Systems Engineering & Operations Research
George Mason University
Fairfax, VA 22032, USA

Michael C. Fu

Robert H. Smith School of Business
Institute for Systems Research
University of Maryland
College Park, MD 20742, USA

Peter W. Glynn

Department of Management Science and Engineering
Stanford University
Standford, CA 94305, USA

Jianqiang Hu

School of Management
Fudan University
Shanghai, 200433, CHINA

ABSTRACT

We provide a unified framework to treat the asymptotic analysis for the non-batched quantile sensitivity estimators of Fu et al. (2009), Liu and Hong (2009), and Lei et al. (2017). With only mild differences in regularity conditions and proofs, asymptotic results including strong consistency and a central limit theorem are established for all three estimators. Simulation results substantiate the theoretical analysis.

1 INTRODUCTION

Quantiles play a central role in characterizing quality of service in the service industry and risk in the financial industry. In financial risk management, quantile is also known as value-at-risk (VaR) and has become a standard benchmark that can be translated directly into minimum capital requirements (see Jorion 2007).

In applications, the quantile of a stochastic system is rarely available in closed form. Therefore, statistical sampling (simulation) is commonly used to estimate the quantile. The asymptotic properties of quantile estimators have been studied extensively in a large body of statistics literature, e.g., David and Nagaraja (1970) and Serfling (2009). Simulation techniques to enhance the efficiency of quantile estimation can be found in Jin et al. (2003) and Glasserman (2004).

Recent attention has focused on quantile sensitivity estimation; see Hong (2009), Fu et al. (2009), Liu and Hong (2009), Heidergott and Volk-Makarewicz (2009), Heidergott et al. (2014), Jiang et al. (2014), Jiang and Fu (2015), Volk-Makarewicz and Heidergott (2015), Heidergott and Volk-Makarewicz (2016), and Lei et al. (2017). Extensions to sensitivity estimation of conditional value-at-risk (CVaR), closely related to quantile sensitivity estimation, can be found in Hong and Liu (2009) and Hong et al. (2014). Asymptotic analysis is an important part of most of the cited works.

The quantile sensitivity estimators can be categorized as either batched or non-batched. Batching means n independent and identically distributed (i.i.d.) observations are split into k batches with m observations in each batch ($n \approx km$). For example, the infinitesimal perturbation analysis (IPA) estimator in Hong (2009) and the weak derivative (WD) estimator in Heidergott and Volk-Makarewicz (2016) generally need

batching, although batching can be avoided in certain special cases, e.g. Jiang and Fu (2015). Asymptotic analysis is simpler for the batched estimators, but the convergence rates for these estimators are generally worse than their non-batched counterparts.

Examples of existing non-batched quantile sensitivity estimators include a conditional Monte Carlo (CMC) method in Fu et al. (2009) and a kernel-based (KB) estimator in Liu and Hong (2009). The asymptotic analysis is more challenging for these estimators, because the *same* batch of data is used to estimate both the quantile and the sensitivity; thus, classical statistics results based on i.i.d assumptions do not apply due to the introduced dependence.

Hong and Liu (2009), Fu et al. (2009), and Liu and Hong (2009) used a novel technique to deal with the dependence issue for their estimators, but the analysis is somewhat tedious and to a large extent case dependent. More importantly, the asymptotic results are not as strong as the asymptotic results for classic quantile estimation, as Fu et al. (2009) and Liu and Hong (2009) only proved weak consistency, and the convergence rate of the CMC estimator in Fu et al. (2009) is established by the square-root convergence of the second moment instead of a central limit theorem. Without a central limit theorem, there is no theoretical justification for constructing a confidence interval and doing hypothesis testing.

More recently, Peng et al. (2017) proposed a generalized likelihood ratio (GLR) method for sensitivity estimation of discontinuous sample performances, and Lei et al. (2017) applied GLR to derive a new quantile sensitivity estimator that does not use batching. In this work, we provide a single framework to treat the asymptotic analysis for the non-batched quantile sensitivity estimators in Fu et al. (2009), Liu and Hong (2009), and Lei et al. (2017). With only a slight difference in regularity conditions, asymptotic results including strong consistency and a central limit theorem are established for all three estimators. An alternative proof for the central limit theorem can be found in Glynn et al. (2017).

The rest of the paper is organized as follows. Section 2 introduces the quantile sensitivity estimation problem and the estimator. We provide preliminary background on empirical processes theory and present our main results in Section 3. The last section offers conclusion.

2 QUANTILE SENSITIVITY ESTIMATION

For a random variable (r.v.) $Z(\theta)$ in a parametric family, the quantile at probability level α, denoted by $q_\alpha(\theta)$, is defined as

$$q_\alpha(\theta) \doteq \sup\{y:\ F(y;\theta) \le \alpha\},$$

where F is the distribution function of Z. If $F(\cdot;\theta)$ is continuous, the quantile can be simplified as $q_\alpha(\theta) = F^{-1}(\alpha;\theta)$. Throughout the paper, we assume $Z(\theta)$ is a continuous r.v., i.e., admitting a density $f(\cdot;\theta)$, and the following regularity condition is presumed:

A0 There exists $\varepsilon > 0$ such that the density $f(x;\theta)$ exists on $(q_\theta^\alpha - \varepsilon, q_\theta^\alpha + \varepsilon)$, $f(q_\theta^\alpha;\theta) > 0$ and $f(\cdot;\theta)$ is continuous at q_θ^α.

With condition A0, the quantile sensitivity $\partial q_\alpha(\theta)/\partial\theta$ is well defined, and by using implicit function differentiation on $F(q_\alpha(\theta);\theta) = \alpha$ with respect to θ, we have (see Fu et al. 2009, Heidergott and Volk-Makarewicz 2016)

$$\frac{dq_\alpha(\theta)}{d\theta} = -\left.\frac{\partial F(x;\theta)}{\partial\theta}\right|_{x=q_\alpha(\theta)} \Big/ f(q_\alpha(\theta);\theta)\ . \tag{1}$$

Let $\mathbf{Z} = \{Z_1,\ldots,Z_n\}$ be i.i.d realizations of $Z(\theta)$. The order statistics of $\mathbf{Z}$ will be denoted by

$$Z_{1:n} < Z_{2:n} < \ldots < Z_{\ell:n} < \ldots < Z_{n:n}\ .$$

A classic quantile estimator is the order statistic $Z_{\lceil \alpha n\rceil:n}$, where $\lceil\cdot\rceil$ is the ceiling operator. With condition A0, the following asymptotic results including strong consistency and a central limit theorem for the classic

quantile estimator can be found in statistics textbooks such as Serfling (2009):

$$\begin{aligned} &\lim_{n\to\infty} Z_{\lceil \alpha n\rceil:n} = q_\alpha(\theta) \quad a.s., \\ &\sqrt{n}(Z_{\lceil \alpha n\rceil:n} - q_\alpha(\theta)) \xrightarrow{d} N\left(0, \frac{\alpha(1-\alpha)}{f^2(q_\alpha(\theta);\theta)}\right), \quad n\to\infty, \end{aligned} \tag{2}$$

where a.s. means almost surely and $\xrightarrow{d}$ indicates convergence in distribution.

In the context of Monte Carlo simulation, the distribution function and density of an output r.v., e.g. the average waiting time in an $M/M/1$ queue, usually are not available analytically, so formula (1) cannot be directly applied. On the other hand, a simulation model for the output r.v. is constructed by (see Fu et al. 2009, Heidergott and Volk-Makarewicz 2016, Lei et al. 2017)

$$Z(\theta) = h(X(\theta);\theta) \tag{3}$$

where $h(\cdot;\theta)$ is a measurable function, and $X(\theta)$ is a vector of input r.v.s in a parametric family, e.g. the interarrival times and service times in a queue. We suppress the dependence of X and h on θ henceforth. Monte Carlo simulation is used to estimate $\partial q_\alpha(\theta)/\partial\theta$ through simulating i.i.d. replications $\mathbf{O} = \{(X_1,Z_1),\ldots,(X_n,Z_n)\}$, where $\{X_1,\ldots,X_n\}$ are i.i.d. realizations of X. The (non-batched) quantile sensitivity estimators in Fu et al. (2009), Liu and Hong (2009), and Lei et al. (2017) all have the following form:

$$\widehat{D}_n = -\frac{\sum_{i=1}^n \Phi^{(1)}(X_i, Z_{\lceil \alpha n\rceil:n})}{\sum_{i=1}^n \Phi^{(2)}(X_i, Z_{\lceil \alpha n\rceil:n})}\ . \tag{4}$$

where the numerator and denominator of the quantile sensitivity estimators have the following two forms:

M1 $\Phi^{(j)}(x,\gamma) = G_j(x,\gamma)$, $j=1,2$, where $G_j(x,\gamma)$ is continuous with respect to γ.

M2 $\Phi^{(j)}(x,\gamma) = \Psi_j(x)\mathbf{1}\{h(x)\le\gamma\}$, $j=1,2$, where $h(\cdot)$ is the measurable function in (3) and $\mathbf{1}\{\cdot\}$ denotes the indicator function.

The CMC estimator in Fu et al. (2009) and the KB estimator in Liu and Hong (2009) have the form M1, and the GLR estimator in Lei et al. (2017) has the form M2. To establish asymptotic results for the two types of estimators above, we need two sets of regularity conditions.

Estimators of form M1 require a Lipschitz condition on G_j, and first and second moment conditions on the corresponding Lipschitz functions, which are given by Assumption A1 as follows:

A1.1 suppose for $G_j(\cdot,\gamma)$, $j=1,2$, there exist measurable functions $K_j(\cdot)$ s.t.

$$|G_j(x,\gamma_1) - G_j(x,\gamma_2)| < K_j(x)|\gamma_1-\gamma_2|, \quad j=1,2\ ;$$

A1.2 $E[K_j(X)] < \infty$, $j=1,2$;
A1.2 $E[K_j^2(X)] < \infty$, $j=1,2$.

Estimators of form M2 require first and second moment conditions on Ψ_j, which are given by Assumption A2 as follows:

A2.2 $E[\Psi_j(X)] < \infty$, $j=1,2$;
A2.2 $E[\Psi_j^2(X)] < \infty$, $j=1,2$.

Although the first moment condition is implied by the second moment condition, we will use the first moment condition to establish strong consistency and the second moment condition to establish a central limit theorem. Although weak consistency is implied by a central limit theorem, strong consistency is not.

Our goal is to establish asymptotic results analogous to (2) for the quantile sensitivity estimator $\widehat{D}_n$. Define $\Phi_\alpha^{(j)}(\cdot) \doteq \Phi^{(j)}(\cdot, q_\alpha)$, and further introduce the following regularity condition.

A3 $\partial_\theta F(x;\theta)$ and $f(x;\theta)$ are differentiable with respect to x.

Theorem 1 For quantile sensitivity estimator (4) of form M1 under conditions A1 and A3 and form M2 under conditions A2 and A3, we have

$$\widehat{D}_n \to \frac{dq_\alpha(\theta)}{d\theta} \quad a.s., \quad n \to \infty,$$

and

$$\sqrt{n}\left(\widehat{D}_n - \frac{dq_\alpha(\theta)}{d\theta}\right) \xrightarrow{d} N(0,\sigma^2), \quad n \to \infty,$$

where $\sigma^2 = Var(\Gamma)$, and

$$\begin{aligned}\Gamma \doteq & \frac{1}{f(q_\alpha(\theta);\theta)}\Phi_\alpha^{(1)} - \frac{\partial_\theta F(x;\theta)|_{x=q_\alpha(\theta)}}{f^2(q_\alpha(\theta);\theta)}\Phi_\alpha^{(2)} \\ & + \left\{ \frac{\partial_\theta F(x;\theta)\partial_{xx}^2 F(x;\theta)\big|_{x=q_\alpha(\theta)}}{f^3(q_\alpha(\theta);\theta)} - \frac{\partial_{x\theta}^2 F(x;\theta)\big|_{x=q_\alpha(\theta)}}{f^2(q_\alpha(\theta);\theta)} \right\} \mathbf{1}\{Z(\theta) \le q_\alpha(\theta)\}\end{aligned}$$

Remark. The asymptotic variance in the theorem matches that in Example 3 of Glynn et al. (2017). The proof of the theorem can be found in the next section. The difficulty in directly obtaining these two asymptotic results is due to the fact that without batching, X_i and $Z_{\lceil \alpha n \rceil:n}$ are dependent, because they come from the same batch of observations $\mathbf{O}$, $i = 1,..n$; thus the classic law of large numbers and central limit theorem do not directly apply. The asymptotic variance in the theorem can be used to build confidence intervals, if the (first- and second-order) derivatives of the distribution function can be estimated. For this purpose, we can use the GLR estimators for distribution sensitivities in Fu et al. (2017), which are all of form M2.

3 ASYMPTOTIC ANALYSIS

In this section, we use an empirical processes technique to address the dependence in proving the asymptotic results for the quantile sensitivity estimator (4). We introduce only the minimum theory required to establish our results. Please refer to Van der Vaart (2000) for details.

3.1 Empirical Processes

Let $\xi_1,...,\xi_n$ be i.i.d. r.v.s with distribution P. We denote the empirical distribution defined as the discrete uniform measure on the observations by $\mathbb{P}_n(\cdot) \doteq \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}(\cdot)$, a random measure on $\mathbb{R}$, where $\delta_\xi(\cdot)$ is the distribution that is degenerate at ξ, where ξ is a generic r.v. with distribution P. The empirical expectation and the expectation of of $\varphi(\xi)$ will be denoted respectively by

$$\mathbb{P}_n\varphi \doteq \int \varphi(z)\ \mathbb{P}_n(dz) = \frac{1}{n}\sum_{i=1}^n \varphi(\xi_i),$$

$$P\varphi \doteq \int \varphi(z)\ P(dz) = E[\varphi(\xi)]\ .$$

Definition 1 A class $\mathscr{F}$ of measurable functions is called P-Glivenko-Cantelli if

$$\|\mathbb{P}_n\varphi - P\varphi\|_{\mathscr{F}} \doteq \sup_{\varphi\in\mathscr{F}} |\mathbb{P}_n\varphi - P\varphi| \to 0 \quad a.s., \quad n\to\infty \ .$$

Remark. We can see the a.s. convergence is uniform with respect to functions in a Glivenko-Cantelli class. The Glivenko-Cantelli class is introduced for the proof of strong consistency of the quantile sensitivity estimator.

The empirical process $\{\mathbb{G}_n\varphi : \varphi\in\mathscr{F}\}$ evaluated at φ is defined as

$$\mathbb{G}_n\varphi \doteq \sqrt{n}(\mathbb{P}_n\varphi - P\varphi) = \frac{\sum_{i=1}^n(\varphi(\xi_i) - P\varphi)}{\sqrt{n}} \ .$$

By the multivariate central limit theorem, for any finite set of measurable functions $\{\varphi_i,\ i=1,..,k\}$ s.t. $P\varphi_i^2<\infty$, $i=1,..,k$,

$$(\mathbb{G}_n\varphi_1,..,\mathbb{G}_n\varphi_k) \xrightarrow{d} (\mathbb{G}_P\varphi_1,..,\mathbb{G}_P\varphi_k), \quad n\to\infty,$$

and $\{\mathbb{G}_P\varphi : \varphi\in\mathscr{F}\}$ is a centered Gaussian process with normal finite-dimensional distribution:

$$(\mathbb{G}_P\varphi_1,..,\mathbb{G}_P\varphi_k) \sim N(0,\Sigma),$$

with the covariance matrix $\Sigma = (\Sigma_{i,j})_{k\times k}$ given by

$$\Sigma_{i,j} = P\varphi_i\varphi_j - P\varphi_i P\varphi_j \ .$$

Definition 2 A class $\mathscr{F}$ of measurable functions is called P-Donsker if the sequence of processes $\{\mathbb{G}_n\varphi : \varphi\in\mathscr{F}\}$ converges in distribution to the centered Gaussian process $\{\mathbb{G}_P\varphi : \varphi\in\mathscr{F}\}$ in the space $\ell^\infty(\mathscr{F})$, where $\ell^\infty(\mathscr{F})$ is a collection of bounded real-valued functionals on $\mathscr{F}$, equipped with the uniform (sup) norm $||\cdot||_{\mathscr{F}}$.

Remark. Notice that the definition of a Donsker class requires the convergence in distribution of a stochastic process, which is stronger than the finite-dimensional convergence in distribution of the stochastic process at finite epochs. A sufficient condition to justify the convergence in distribution of the stochastic process can be found in Chapter 18 of Van der Vaart (2000). The convergence in distribution is "uniform" with respect to functions in a Donsker class. The Donsker class is introduced for the proof of the central limit theorem of the quantile sensitivity estimator.

Whether a class of functions is Glivenko-Cantelli or Donsker depends on the "size" of the class. A finite class of integrable functions is always Glivenko-Cantelli and a finite class of square-integrable functions is always Donsker. To provide a sufficient condition for Glivenko-Cantelli or Donsker for an infinite class of functions requires a way to measure the size of a class of functions in terms of entropy. The bracketing entropy is relative to an $L_r(P)$-norm defined as

$$\|\varphi\|_{P,r} = (P|\varphi|^r)^{1/r} \ .$$

For two functions l and u with finite $L_r(P)$-norm (they need not belong to $\mathscr{F}$), the bracket $[l,u]$ is defined as $\{\varphi\in\mathscr{F} :\ l(x)\le\varphi(x)\le u(x),\ \forall\, x\}$. An ε-bracket in $L_r(P)$ is a bracket $[l,u]$ such that $(P(u-l)^r)^{1/r}<\varepsilon$. The bracketing number $N_{[\,]}(\varepsilon,\mathscr{F},L_r(P))$ is the minimal number of ε-brackets needed to cover $\mathscr{F}$, and the bracketing integral is defined as

$$J_{[\,]}(y,\mathscr{F},L_r(P)) = \int_0^y \sqrt{\ln N_{[\,]}(\varepsilon,\mathscr{F},L_2(P))}\ d\varepsilon \ .$$

Lemma 1 (Lemmas 19.13 and 19.14 in Van der Vaart 2000)

(i) A class of measurable functions $\mathscr{F}$ s.t. $\forall\ \varepsilon > 0$, $N_{[\ ]}(\varepsilon, \mathscr{F}, L_1(P)) < \infty$ is P-Glivenko-Cantelli.
(ii) A class of measurable functions $\mathscr{F}$ s.t. $J_{[\ ]}(1, \mathscr{F}, L_2(P)) < \infty$ is P-Donsker.

Then we introduce the notion of a "random function". In our paper, a random function means a measurable function that depends on random observations, i.e. $\widehat{\varphi}_n(\cdot) \doteq \varphi(\cdot; \widehat{\gamma}_n)$, where $\widehat{\gamma}_n$ is estimated by the samples $\xi_1, .., \xi_n$. In particular, an example of random functions are measurable functions in the numerator and denominator of the quantile sensitivity estimator (4), i.e.,

$$\widehat{\Phi}^{(j)}_{n,\alpha}(\cdot) \doteq \Phi^{(j)}(\cdot, Z_{\lceil \alpha n \rceil : n}),\ j = 1, 2,$$

where $Z_{\lceil \alpha n \rceil : n}$ depends on observations $\mathbf{Z} = \{Z_1, \ldots, Z_n\}$, which are functions (3) of n realizations of input r.v.s $\{X_1, \ldots, X_n\}$.

The most critical conditions for establishing strong consistency and a central limit theorem for the random functions $\widehat{\Phi}^{(j)}_{n,\alpha}$, $j = 1, 2$, are that the function classes

$$\mathscr{F}^{(j)}_{\gamma} \doteq \{\Phi^{(j)}(\cdot, \gamma) : \gamma \in \mathbb{R}\}, \quad j = 1, 2,$$

belong to P-Glivenko-Cantelli and P-Donsker, respectively (see Chapter 19.4, Van der Vaart 2000), which requires some additional regularity conditions.

The following lemma establishes a general result for strong consistency and a central limit theorem for a random function.

Lemma 2 (Chapter 19.4 of Van der Vaart 2000)

(i) Suppose $\mathscr{F}$ is a P-Glivenko-Cantelli class and $\widehat{\varphi}_n(\cdot) \in \mathscr{F}$. If there is a function $\varphi \in L_1(P)$ such that $\int (\widehat{\varphi}_n(z) - \varphi_0(z))\ dP(z) \to 0$ a.s., $n \to \infty$, then

$$\mathbb{P}_n \widehat{\varphi}_n \to P\varphi_0 \quad a.s., \quad n \to \infty,$$

where

$$\mathbb{P}_n \widehat{\varphi}_n \doteq \frac{1}{n} \sum_{i=1}^{n} \widehat{\varphi}_n(\xi_i)\ .$$

(ii) Suppose that $\mathscr{F}$ is a P-Donsker class and $\widehat{\varphi}_n(\cdot) \in \mathscr{F}$. If there is a function $\varphi_0 \in L_2(P)$ such that $\int (\widehat{\varphi}_n(z) - \varphi_0(z))^2\ dP(z) \xrightarrow{d} 0$, $n \to \infty$, then

$$\mathbb{G}_n \widehat{\varphi}_n \xrightarrow{d} \mathbb{G}_P \varphi_0, \quad n \to \infty,$$

where

$$\mathbb{G}_n \widehat{\varphi}_n \doteq \sqrt{n} \left(\int \widehat{\varphi}_n(z)\ \mathbb{P}_n(dz) - \int \widehat{\varphi}_n(z)\ P(dz) \right)\ .$$

3.2 Main Results

Now we are ready to present the main results.

Lemma 3 For quantile sensitivity estimator (4) of form M1,

(i) under A1.2, the function classes $\mathscr{F}^{(j)}_{\gamma}$, is P-Glivenko-Cantelli, $j = 1, 2$;
(ii) under A1.3, the function classes $\mathscr{F}^{(j)}_{\gamma}$ is P-Donsker, $j = 1, 2$.

The proof of Lemma 3 can be directly found in Chapter 19.6 (page 271) of Van der Vaart (2000).

Lemma 4 For quantile sensitivity estimator (4) of form M2,

(i) under A2.1, the function classes $\mathscr{F}_\gamma^{(j)}$, is P-Glivenko-Cantelli, $j = 1,2$;

(ii) under A2.2, the function classes $\mathscr{F}_\gamma^{(j)}$ is P-Donsker, $j = 1,2$.

Proof. Let $-\infty = \gamma_0 < \gamma_1 < \cdots < \gamma_k = \infty$ be a partition of $\mathbb{R}$. Consider brackets of the form $[l_i, u_i]$, $i = 1,..,k$, where

$$l_i(x) = \mathbf{1}\{h(x) \le \gamma_{i-1}, \Psi_j(x) \ge 0\}\Psi_j(x) + \mathbf{1}\{h(x) \le \gamma_i, \Psi_j(x) < 0\}\Psi_j(x),$$
$$u_i(x) = \mathbf{1}\{h(x) \le \gamma_i, \Psi_j(x) \ge 0\}\Psi_j(x) + \mathbf{1}\{h(x) \le \gamma_{i-1}, \Psi_j(x) < 0\}\Psi_j(x) \ .$$

We can see $[l_i, u_i]$, $i = 1,..,k$, cover the function class $\mathscr{F}_\gamma^{(j)}$. Under the first moment condition in (i), it is easy to show l_i and u_i has finite $L_1(P)$-norm. For the bracketing entropy relative to the $L_1(P)$-norm,

$$E[u_i(X) - l_i(X)] = E\left[\mathbf{1}\{\gamma_{i-1} \le h(X) \le \gamma_i\}|\Psi_j(X)|\right] \ .$$

Denote $c_1 \doteq E[|\Psi_j(X)|]$, and let $b_1 = \lceil c_1/\varepsilon \rceil + 1$. With appropriate selection of the partition points, we can find b_1 ε-brackets in $L_1(P)$ which cover $\mathscr{F}_\gamma^{(j)}$. Therefore, $N_{[\,]}(\varepsilon, \mathscr{F}_\gamma^{(j)}, L_1(P)) \le b_1 < \infty$. By Lemma 1, $\mathscr{F}_\gamma^{(j)}$ is P-Glivenko-Cantelli.

Under the second moment condition in (ii), it is easy to show l_i and u_i have finite $L_2(P)$-norm. For the bracketing entropy relative to the $L_2(P)$-norm,

$$E^{\frac{1}{2}}[(u_i(X) - l_i(X))^2] = E^{\frac{1}{2}}\left[\mathbf{1}\{\gamma_{i-1} \le h(X) \le \gamma_i\}|\Psi_j(X)|^2\right] \ .$$

Denote $c_2 \doteq E[|\Psi_j(X)|^2]$, and let $b_2 = \lceil c_2/\varepsilon^2 \rceil + 1$. With appropriate selection of the partition points, we can find b_2 ε-brackets in $L_2(P)$ covering $\mathscr{F}_\gamma^{(j)}$, which implies $N_{[\,]}(\varepsilon, \mathscr{F}_\gamma^{(j)}, L_2(P)) \le b_2$. For any $r > 0$, we have $\lim_{\varepsilon \to 0} \varepsilon^r \ln \varepsilon = 0$, therefore

$$J_{[\,]}(1, \mathscr{F}_\gamma^{(j)}, L_r(P)) = \int_0^1 \sqrt{\ln N_{[\,]}(\varepsilon, \mathscr{F}_\gamma^{(j)}, L_2(P))}\ d\varepsilon \le \int_0^1 \sqrt{-2\ln\varepsilon + \ln(c_2+1)}\ d\varepsilon < \infty \ .$$

By Lemma 1, $\mathscr{F}_\gamma^{(j)}$ is P-Donsker. □

Theorem 2 For quantile sensitivity estimator (4) of form M1,

(i) under A1.1 and A1.2, $\mathbb{P}_n\widehat{\Phi}_{n,\alpha}^{(j)} \to P\Phi_\alpha^{(j)}$ a.s. as $n \to \infty$, $j = 1,2$;

(ii) under A1.1 and A1.3, $\mathbb{G}_n\widehat{\Phi}_{n,\alpha}^{(j)} \xrightarrow{d} \mathbb{G}_P\Phi_\alpha^{(j)}$ as $n \to \infty$, $j = 1,2$.

Proof. Since $Z_{\lceil \alpha n \rceil:n} \to q_\alpha$ a.s. as $n \to \infty$, with the Lipschitz condition,

$$\left|\int (\widehat{\Phi}_{n,\alpha}^{(j)}(x) - \Phi_\alpha^{(j)}(x))P(dx)\right| = \left|\int (G_j(x,\gamma) - G_j(x,\gamma'))P(dx)\right|\Big|_{\gamma = Z_{\lceil \alpha n \rceil:n}, \gamma' = q_\alpha}$$
$$\le |Z_{\lceil \alpha n \rceil:n} - q_\alpha| \int K_j(x)P(dx) \to 0 \quad a.s., \quad n \to \infty,$$

by using the first moment condition in (i). From the conclusion (i) of Lemma 4, the conclusion (i) of this theorem is proved by referring to Lemma 2. With the Lipschitz condition,

$$\int (\widehat{\Phi}_{n,\alpha}^{(j)}(x) - \Phi_{\alpha}^{(j)}(x))^2 P(dx) = \left\{ \int (G_j(x,\gamma) - G_j(x,\gamma'))^2 P(dx) \right\} \Big|_{\gamma = Z_{\lceil \alpha n \rceil :n}, \gamma' = q_\alpha}$$
$$\leq |Z_{\lceil \alpha n \rceil :n} - q_\alpha|^2 \int K_j^2(x) P(dx) \to 0 \quad a.s., \quad n \to \infty,$$

which implies convergence in distribution, by using the second moment condition in (ii). From the conclusion (ii) of Lemma 4, the conclusion (ii) of this theorem is proved by referring to Lemma 2. □

Theorem 3 For quantile sensitivity estimator (4) of form M2,

(i) under A2.1, $\mathbb{P}_n \widehat{\Phi}_{n,\alpha}^{(j)} \to P\Phi_{\alpha}^{(j)}$ a.s. as $n \to \infty$, $j = 1,2$;

(ii) under A2.2, $\mathbb{G}_n \widehat{\Phi}_{n,\alpha}^{(j)} \xrightarrow{d} \mathbb{G}_P \Phi_{\alpha}^{(j)}$ as $n \to \infty$, $j = 1,2$.

Proof. Since $Z_{\lceil \alpha n \rceil :n} \to q_\alpha$ a.s. as $n \to \infty$, from the structure of the estimator of form M.2, $\forall\, x$,

$$\widehat{\Phi}_{n,\alpha}^{(j)}(x) = \Psi_j(x) \mathbf{1}\left\{h(x) \leq Z_{\lceil \alpha n \rceil :n}\right\} \to \Psi_j(x) \mathbf{1}\left\{h(x) \leq q_\alpha\right\} = \Phi_{\alpha}^{(j)}(x) \quad a.s. \quad n \to \infty\ .$$

With the first moment condition in (i) and the fact that $\mathbf{1}\{\cdot \leq 0\}$ is bounded, the sequence $\widehat{\Phi}_{n,\alpha}^{(j)}$ is dominated by an integrable function, so conclusion (i) can be proved by similar arguments as in the proof of Theorem 2. In addition,

$$\int (\widehat{\Phi}_{n,\alpha}^{(j)}(x) - \Phi_{\alpha}^{(j)}(x))^2 P(dx)$$
$$= \left\{ \int \mathbf{1}\{\min(\gamma, q_\alpha) \leq h(x) \leq \max(\gamma, q_\alpha)\} \Psi_j^2(x) P(dx) \right\} \Big|_{\gamma = Z_{\lceil \alpha n \rceil :n}}.$$

With condition A0, we know

$$\lim_{n \to \infty} P(\min(q_\alpha, \gamma) \leq h(X) \leq \max(q_\alpha, \gamma))|_{\gamma = Z_{\lceil \alpha n \rceil :n}} = 0 \quad a.s.$$

With the second moment condition in (ii), by the absolute continuity of the Lebesgue integral (or dominated convergence theorem), we have

$$\lim_{n \to \infty} \int (\widehat{\Phi}_{n,\alpha}^{(j)}(x) - \Phi_{\alpha}^{(j)}(x))^2 P(dx) = 0 \quad a.s.,$$

which implies convergence in distribution. From the conclusion (ii) of Lemma 4, the conclusion (ii) of this theorem is proved by referring to Lemma 2. □

Remark. The key idea for the proofs of strong consistency and a central limit theorem for a random function in Theorems 2 and 3 is to use the uniform convergence of the empirical expectation and empirical process in the function class Glivenko-Cantelli and Donsker, respectively. In general, pointwise convergence is not sufficient to guarantee uniform convergence, but together with finite covers for the function class, uniform convergence can be established by checking pointwise convergence.

Generally, it is not possible to obtain a central limit theorem for $\sqrt{n}(\mathbb{P}_n \widehat{\Phi}_{n,\alpha}^{(j)} - P\Phi_{\alpha}^{(j)}) = \mathbb{G}_n \widehat{\Phi}_{n,\alpha}^{(j)} + \sqrt{n}(P\widehat{\Phi}_{n,\alpha}^{(j)} - P\Phi_{\alpha}^{(j)})$, $j = 1,2$, without imposing additional conditions. One such condition is A3.

Theorem 4 For quantile sensitivity estimator (4) of form M1 under conditions A1 and A3 and form M2 under conditions A2 and A3,

$$\sqrt{n}(\mathbb{P}_n\widehat{\Phi}^{(j)}_{n,\alpha} - P\Phi^{(j)}_{\alpha}) \xrightarrow{d} N(0,\sigma_0^2), \quad j=1,2,$$

where

$$\sigma_0^2 = Var(\Phi^{(j)}_{\alpha}) + \frac{(1-\alpha)}{f(q_\alpha)} \left.\frac{\partial P\Phi|_\gamma}{\partial\gamma}\right|_{\gamma=q_\alpha} \left(\frac{\alpha}{f(q_\alpha)} \left.\frac{\partial P\Phi^{(j)}|_\gamma}{\partial\gamma}\right|_{\gamma=q_\alpha} - 2P\Phi^{(j)}_{\alpha} \right),$$

and $Var(\Phi^{(j)}_{\alpha}) \doteq Var(\Phi^{(j)}(X,q_\alpha))$.

Proof. With condition A3, by the delta method and the central limit theorem for the quantile estimator (see Chapters 3 and 21 of Van der Vaart 2000), we have

$$\sqrt{n}(P\widehat{\Phi}^{(j)}_{n,\alpha} - P\Phi^{(j)}_{\alpha}) = \mathbb{G}_n\Pi^{(j)}|_{\gamma=q_\alpha} + o_p(1),$$

where

$$\Pi^{(j)}(x,\gamma) \doteq -\frac{\partial P\Phi^{(j)}|_\gamma}{\partial\gamma}\frac{\mathbf{1}\{h(x)\le\gamma\}-\alpha}{f(\gamma)} \ .$$

Since the property of Glivenko-Cantelli or Donsker for a function class would not change after adding a finite number of functions satisfying the corresponding integrability condition, applying conclusion (ii) of Theorem 3 to $\left(\widehat{\Phi}^{(j)}_{n,\alpha}, \Pi^{(j)}|_{\gamma=q_\alpha}\right)$, we have

$$\mathbb{G}_n\left(\widehat{\Phi}^{(j)}_{n,\alpha}, \Pi^{(j)}|_{\gamma=q_\alpha}\right) \xrightarrow{d} \mathbb{G}_P(\Phi^{(j)}_{\alpha}, \Pi^{(j)}|_{\gamma=q_\alpha}) \ .$$

Notice that

$$\sqrt{n}(\mathbb{P}_n\widehat{\Phi}^{(j)}_{n,\alpha} - P\Phi^{(j)}_{\alpha}) = \mathbb{G}_n\widehat{\Phi}^{(j)}_{n,\alpha} + \mathbb{G}_n\Pi^{(j)}|_{\gamma=q_\alpha} + o_p(1) \ .$$

By the continuous mapping theorem (see Van der Vaart (2000)), the conclusion in the theorem follows immediately. □

Theorem 4 establishes a central limit theorem for the numerator and denominator in the quotient in (4) by using the results in Theorems 2 and 3, together with the central limit theorem for the quantile itself and the delta method. Theorems 2, 3 and 4 imply strong consistency and a central limit theorem for the quotient in (4).

Proof. of **Theorem 1**. From conclusion (i) of Theorems 2 and 3,

$$\mathbb{P}_n\widehat{\Phi}^{(1)}_{n,\alpha} \to P\Phi^{(1)}_{\alpha}, \quad \mathbb{P}_n\widehat{\Phi}^{(2)}_{n,\alpha} \to P\Phi^{(2)}_{\alpha} \quad a.s., \quad n\to\infty \ .$$

With condition A0, we know $f(\cdot)$ is strictly positive at q_α. Therefore, conclusion (i) in the corollary is proved. With the conditions assumed in the theorem,

$$\begin{aligned}
\sqrt{n}\left(\widehat{D}_n - \frac{dq_\alpha(\theta)}{d\theta}\right) &= -\sqrt{n}\left(\frac{\mathbb{P}_n\widehat{\Phi}^{(1)}_{n,\alpha}}{\mathbb{P}_n\widehat{\Phi}^{(2)}_{n,\alpha}} - \frac{P\Phi^{(1)}_{\alpha}}{P\Phi^{(2)}_{\alpha}}\right) \\
&= -\frac{1}{\mathbb{P}_n\widehat{\Phi}^{(2)}_{n,\alpha}P\Phi^{(2)}_{\alpha}}\Big\{P\Phi^{(2)}_{\alpha}(\mathbb{G}_n\widehat{\Phi}^{(1)}_{n,\alpha} + \mathbb{G}_n\Pi^{(1)}|_{\gamma=q_\alpha}) \\
&\qquad\qquad -P\Phi^{(1)}_{\alpha}(\mathbb{G}_n\widehat{\Phi}^{(2)}_{n,\alpha} + \mathbb{G}_n\Pi^{(2)}|_{\gamma=q_\alpha}) + o_p(1)\Big\} \ .
\end{aligned}$$

Applying Theorems 2 and 3 to $\left(\widehat{\Phi}^{(1)}_{n,\alpha}, \Pi^{(1)}|_{\gamma=q_\alpha}, \widehat{\Phi}^{(2)}_{n,\alpha}, \Pi^{(2)}|_{\gamma=q_\alpha}\right)$, conclusion (ii) of the corollary follows immediately by noticing $\Gamma = \eta V'$, where

$$\eta \doteq \left(\frac{1}{P\Phi^{(2)}_\alpha}, \frac{1}{P\Phi^{(2)}_\alpha}, -\frac{P\Phi^{(1)}_\alpha}{P\Phi^{(2)}_\alpha}, -\frac{P\Phi^{(1)}_\alpha}{P\Phi^{(2)}_\alpha} \right),$$

$$V \doteq \left(\Phi^{(1)}_\alpha, -\frac{\partial_\gamma P\Phi^{(1)}_\alpha|_\gamma \mathbf{1}\{Z(\theta) \le q_\alpha(\theta)\}}{P\Phi^{(2)}_\alpha}, \frac{\Phi^{(2)}_\alpha}{P\Phi^{(2)}_\alpha}, -\frac{\partial_\gamma P\Phi^{(2)}_\alpha|_\gamma \mathbf{1}\{Z(\theta) \le q_\alpha(\theta)\}}{\left(P\Phi^{(2)}_\alpha\right)^2} \right) .$$

□

Remark. For the quantile sensitivity estimators in Fu et al. (2009) and Lei et al. (2017), we have $P\Phi^{(1)}|_\gamma = \partial F(\gamma;\theta)/\partial\gamma$ and $P\Phi^{(2)}|_\gamma = f(\gamma)$, so the asymptotic results of CMC and GLR can be directly established by Corollary 1. For the KB estimator in Liu and Hong (2009), there is a bandwidth (tuning parameter) δ in the estimator, and $\lim_{\delta\to 0} P\Phi^{(1)}|_\gamma = \partial F(\gamma;\theta)/\partial\gamma$ and $\lim_{\delta\to 0} P\Phi^{(2)}|_\gamma = f(\gamma)$. Without considering the bias, the asymptotic results in Corollary 1 also directly apply to the estimator in Liu and Hong (2009). To establish asymptotic unbiasedness, Liu and Hong (2009) let δ_n go to zero as n goes to infinity, and impose requirements on the rate at which δ_n decreases with respect to n. With some additional regularity conditions, the asymptotic results in Liu and Hong (2009) can be significantly streamlined based on an analysis similar to Corollary 1.

4 NUMERICAL RESULTS

In the numerical experiment, we test the performances of the CMC, KB, and GLR estimators on a linear Gaussian model, i.e. $Z(\theta) = h(X;\theta) = \eta + \theta\zeta$, where $X = (\eta, \zeta)$ and $\eta, \zeta \sim N(0,1)$ are independent. For this model, the quantile sensitivity can be calculated analytically as $\partial q_\alpha(\theta)/\partial\theta = z_\alpha\theta/\sqrt{\theta^2+1}$, where z_α is the α-quantile of the standard normal distribution.

In this example, the CMC estimator is given by

$$\frac{\sum_{i=1}^n \zeta_i \phi(Z_{\lceil \alpha n\rceil:n} - \theta\zeta_i)}{\sum_{i=1}^n \phi(Z_{\lceil \alpha n\rceil:n} - \theta\zeta_i)},$$

where $\phi(\cdot)$ is the density of the standard normal distribution, the KB estimator is given by

$$\frac{\sum_{i=1}^n \zeta_i \phi\left((Z_{\lceil \alpha n\rceil:n} - Z_i)/\delta_n\right)}{\sum_{i=1}^n \phi\left((Z_{\lceil \alpha n\rceil:n} - Z_i)/\delta_n\right)},$$

and the GLR estimator is given by

$$\frac{\sum_{i=1}^n \mathbf{1}\{Z_i \le Z_{\lceil \alpha n\rceil:n}\}\left(\eta_i\zeta_i + \zeta_i^2 - 1\right)}{\sum_{i=1}^n \mathbf{1}\{Z_i \le Z_{\lceil \alpha n\rceil:n}\}\left(\eta_i + \zeta_i\right)} .$$

GLR estimators for $\partial_\theta F(x;\theta)$, $f(x;\theta)$, $\partial_{xx} F(x;\theta)$, and $\partial_{x\theta} F(x;\theta)$ are, respectively,

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{Z_i \le x\}\zeta_i\eta_i, \quad -\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{Z_i \le x\}\eta_i, \quad \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{Z_i \le x\}(\eta_i^2 - 1), \quad \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{Z_i \le x\}\zeta_i(1-\eta_i^2) .$$

For the numerical results, we estimate the median sensitivity, i.e. $\alpha = 0.5$, where the true value is $\partial q_{0.5}(\theta)/\partial\theta = 0$. For the number of replications in the experiments, we take $n = 10^4, 10^5, 10^6$. The

	$n = 10^4$	$n = 10^5$	$n = 10^6$
CMC	$-1.1 \times 10^{-4} \pm 1.3 \times 10^{-4}$ (90%)	$-1.6 \times 10^{-6} \pm 4.2 \times 10^{-5}$ (90%)	$-1.9 \times 10^{-8} \pm 1.3 \times 10^{-5}$ (90%)
KB	$1.0 \times 10^{-4} \pm 2.1 \times 10^{-4}$ (89%)	$-2.8 \times 10^{-5} \pm 7.6 \times 10^{-5}$ (90%)	$3.7 \times 10^{-6} \pm 3.0 \times 10^{-5}$ (89%)
GLR	$1.1 \times 10^{-4} \pm 2.3 \times 10^{-4}$ (88%)	$-3.7 \times 10^{-5} \pm 7.3 \times 10^{-5}$ (88%)	$3.4 \times 10^{-6} \pm 2.3 \times 10^{-5}$ (89%)

Table 1: 0.5-quantile sensitivity of a linear Gaussian model based on 10^4 independent macro simulations: mean $\pm$ std err (actual coverage of 90% confidence interval).

statistics are estimated based on 10000 macro simulations. For KB, we set the bandwidth to $\delta_n = n^{-1/5}$ recommended in Liu and Hong (2009). In Table 1, we can see that CMC has the lowest variance; the standard deviations of both CMC and GLR decrease at rate $n^{-1/2}$, which matches the theoretical result, while the standard deviation of KB decreases at rate slightly slower than $n^{-1/2}$, which is caused by the decreasing of bandwidth; the actual coverage rates of the confidence intervals of CMC, KB, and GLR are very close to the target 90% level.

5 CONCLUSION

We use an empirical process technique to deal with the dependence issue in asymptotic analysis for quantile sensitivity estimation, which offers both succinct proofs in a single framework and stronger results for various non-batched quantile sensitivity estimators.

Acknowledgments.

This work was supported in part by the National Science Foundation (NSF) under Grants CMMI-1362303 and CMMI-1434419, by the National Science Foundation of China (NSFC) under Grants 71571048, by the Air Force of Scientific Research (AFOSR) under Grant FA9550-15-10050, by the Science and Technology Agency of Sichuan Province under Grant 2014GZX0002, by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institution of Higher Learning, and by the China Postdoctoral Science Foundation under Grant 2015M571495.

REFERENCES

David, H. A., and H. N. Nagaraja. 1970. *Order Statistics*. Wiley Online Library.

Fu, M. C., B. Heidergott, H. Lam, and Y. Peng. 2017. "Maximum likelihood estimation by Monte Carlo simulation". *working paper*.

Fu, M. C., L. J. Hong, and J.-Q. Hu. 2009. "Conditional Monte Carlo estimation of quantile sensitivities". *Management Science* 55 (12): 2019–2027.

Glasserman, P. 2004. *Monte Carlo Methods in Financial Engineering*. Springer.

Glynn, P., L. Fan, M. C. Fu, J.-Q. Hu, and Y. Peng. 2017. "Technical note — Central limit theorems for estimated functions at estimated points". *submitted to Management Science*.

Heidergott, B., and W. Volk-Makarewicz. 2009. "Quantile sensitivity estimation". In *International Conference on Network Control and Optimization*, 16–29. Springer.

Heidergott, B., and W. Volk-Makarewicz. 2016. "A measure-valued differentiation approach to sensitivity analysis of quantiles". *Mathematics of Operations Research* 41 (1): 293–317.

Heidergott, B., W. Volk-Makarewicz, and F. J. Vázquez-Abad. 2014. "Gradient estimation for quantiles of stationary waiting times". In *Proceedings of IEEE International Workshop on Discrete Event Systems*, 241–246.

Hong, L. J. 2009. "Estimating quantile sensitivities". *Operations Research* 57 (1): 118–130.

Hong, L. J., Z. Hu, and G. Liu. 2014. "Monte Carlo methods for value-at-risk and conditional value-at-risk: A review". *ACM Transactions on Modeling and Computer Simulation* 24 (4): 1–37.

Hong, L. J., and G. Liu. 2009. "Simulating sensitivities of conditional value at risk". *Management Science* 55 (2): 281–293.

Jiang, G., and M. C. Fu. 2015. "Technical note — On estimating quantile sensitivities via infinitesimal perturbation analysis". *Operations Research* 63 (2): 435–441.

Jiang, G., M. C. Fu, and C. Xu. 2014. "Bias reduction in estimating quantile sensitivities". *IFAC Proceedings Volumes* 47 (3): 10463–10468.

Jin, X., M. C. Fu, and X. Xiong. 2003. "Probabilistic error bounds for simulation quantile estimators". *Management Science* 49 (2): 230–246.

Jorion, P. 2007. *Value at Risk: The New Benchmark for Managing Financial Risk*, Volume 2. McGraw-Hill, New York.

Lei, L., Y. Peng, M. C. Fu, and J.-Q. Hu. 2017. "Applications of generalized likelihood ratio method to distribution sensitivities and steady-state simulation". *Journal of Discrete Event Dynamic Systems*:accepted.

Liu, G., and L. J. Hong. 2009. "Kernel estimation of quantile sensitivities". *Naval Research Logistics* 56 (6): 511–525.

Peng, Y., M. C. Fu, J.-Q. Hu, and B. Heidergott. 2017. "A new unbiased stochastic derivative estimator for discontinuous sample performances with structural parameters". *submitted to Operations Research*.

Serfling, R. J. 2009. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons.

Van der Vaart, A. W. 2000. *Asymptotic Statistics*, Volume 3. Cambridge University Press.

Volk-Makarewicz, W., and B. Heidergott. 2015. "Sensitivity analysis of ranked data: from order statistics to quantiles". *Discrete Event Dynamic Systems* 25 (4): 453–495.

AUTHOR BIOGRAPHIES

YIJIE PENG is a Research Assistant Professor in the Department of Systems Engineering and Operations Research at George Mason University. His research interests lie in sensitivity analysis, and ranking and selection in simulation optimization field. His email address is ypeng10@gmu.edu.

MICHAEL C. FU holds the Smith Chair of Management Science in the Robert H. Smith School of Business, with a joint appointment in the Institute for Systems Research and affiliate faculty appointment in the Department of Electrical and Computer Engineering, all at the University of Maryland. His research interests include simulation optimization and applied probability, with applications in supply chain management and financial engineering. He served as WSC2011 Program Chair, NSF Operations Research Program Director, *Management Science* Stochastic Models and Simulation Department Editor, and *Operations Research* Simulation Area Editor. He is a Fellow of INFORMS and IEEE. His email address is mfu@umd.edu.

PETER W. GLYNN is the Thomas Ford Professor in the Department of Management Science and Engineering (MS&E) at Stanford University. He is a Fellow of INFORMS and of the Institute of Mathematical Statistics, has been co-winner of Best Publication Awards from the INFORMS Simulation Society in 1993, 2008, and 2016, and was the co-winner of the John von Neumann Theory Prize from INFORMS in 2010. In 2012, he was elected to the National Academy of Engineering. His research interests lie in stochastic simulation, queueing theory, and statistical inference for stochastic processes. His email address is glynn@stanford.edu.

JIAN-QIANG HU is a Professor with the Department of Management Science, School of Management, Fudan University. He was an Associate Professor with the Department of Mechanical Engineering and the Division of Systems Engineering at Boston University before joining Fudan University. His research interests include discrete-event stochastic systems, simulation, queueing network theory, stochastic control theory, with applications towards supply chain management, risk management in financial markets and derivatives, and communication networks. His e-mail addresses is hujq@fudan.edu.cn.