OPERATIONS RESEARCH
Vol. 57, No. 1, January–February 2009, pp. 118–130
ISSN 0030-364X | EISSN 1526-5463 | 09 | 5701 | 0118

informs

DOI 10.1287/opre.1080.0531

Estimating Quantile Sensitivities

L. Jeff Hong
Department of Industrial Engineering and Logistics Management, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China, hongl@ust.hk

Quantiles of a random performance serve as important alternatives to the usual expected value. They are used in the financial industry as measures of risk and in the service industry as measures of service quality. To manage the quantile of a performance, we need to know how changes in the input parameters affect the output quantiles, which are called quantile sensitivities. In this paper, we show that the quantile sensitivities can be written in the form of conditional expectations. Based on the conditional-expectation form, we first propose an infinitesimal-perturbation-analysis (IPA) estimator. The IPA estimator is asymptotically unbiased, but it is not consistent. We then obtain a consistent estimator by dividing data into batches and averaging the IPA estimates of all batches. The estimator satisfies a central limit theorem for the i.i.d. data, and the rate of convergence is strictly slower than $n^{-1/3}$. The numerical results show that the estimator works well for practical problems.

Subject classifications: quantile; value-at-risk; sensitivity analysis; simulation; statistical analysis.
Area of review: Simulation.
History: Received March 2006; revisions received September 2006, March 2007, July 2007; accepted September 2007. Published online in *Articles in Advance* September 17, 2008.

1. Introduction

Quantiles of a random performance serve as important alternatives to the usual expected value. The α-quantile of a continuous random variable Y is a value q_α such that $\Pr\{Y \leqslant q_\alpha\} = \alpha$ for any prespecified α $(0 < \alpha < 1)$. When $\alpha = 0.5$, the corresponding quantile is the median, which provides information on the location of a distribution; when α is close to zero or one, the corresponding quantile provides tail information of a distribution that is often missed by some other widely used measures, e.g., mean and variance.

Quantiles have been adopted by many industries as major measures of random performance. In the financial industry, quantiles, also known as value-at-risks (VaRs), are widely accepted measures of capital adequacy. For example, the Bank for International Settlement uses the 10-day VaR at the 99% level to measure the adequacy of bank capital (Duffie and Pan 1997). In the service industry, quantiles are often used as measures of service quality. For example, the service quality of an out-of-hospital system is frequently measured by the 90th percentile of the times taken to respond to emergency requests and to transport patients to a hospital (Austin and Schull 2003). Quantiles have also been used as billing measures in some circumstances. For example, some Internet service providers (ISPs) charge their users based on the 95th percentile of the traffic load in a billing cycle (Goldenberg et al. 2004).

To improve or optimize the quantile performance of a system, one needs to understand how changes in the input parameters affect the output quantile performance. These effects are often called *quantile sensitivities*. When parameters vary continuously, sensitivities are essentially partial derivatives. The vector of these partial derivatives is the *quantile gradient*, which plays an important role in optimization problems with quantile objectives or constraints. One example is robust optimization, where parameters in the optimization problems may be noisy. In this example, the random objective and constraints may be substituted by their quantiles to obtain a robust solution (see, for example, Hong and Qi 2007 for a quantile-based robust linear programming). Another example is optimization problems with chance constraints, where each constraint is required to be satisfied with at least a certain probability (Birge and Louveaux 1997). Note that chance constraints can be transformed into quantile constraints.

1.1. Literature Review

Estimating quantile sensitivities is related to two streams of literature: quantile estimation and gradient estimation. Bahadur (1966) provides a quantile estimator for i.i.d. data, and shows that the estimator is strongly consistent and has an asymptotic normal distribution. A more recent and thorough review of quantile estimation for i.i.d. data can be found in Serfling (1980). Sen (1972) extends Bahadur's (1966) estimator, and shows that the strong consistency and the asymptotic normality of the estimator also hold for dependent data that are ϕ-mixing. Heidelberger and Lewis (1984) design new procedures using the maximum transformation to estimate the extreme quantiles of highly positively correlated data. In the simulation literature, a number of variance reduction techniques have been proposed for quantile estimation; for example, Hsu and Nelson

(1990) and Hesterberg and Nelson (1998) use control variates, Glynn (1996) applies importance sampling, and Avramidis and Wilson (1998) employ correlation-induction techniques. Jin et al. (2003) study the probabilistic error bounds for simulation quantile estimators using large deviation techniques, and they provide a new quantile estimator that can be shown to be more efficient than some of the existing quantile estimators. Glasserman et al. (2000, 2002) study the problem of estimating portfolio VaR. They utilize the properties of financial portfolios and apply importance sampling and stratified sampling to reduce the variances of their estimators.

Gradient estimation has also been studied extensively in the simulation literature (see Fu 2006 for an introduction). However, most of the work focuses on estimating gradient for expectations or long-run averages. There are three major approaches: finite-difference approximation, perturbation analysis, and the likelihood ratio/score function (LR/SF) method. For the finite-difference approximation, the key problem is the trade-off between the bias and variance of the estimator. This trade-off is analyzed by Fox and Glynn (1989). In perturbation analysis, the key problem is the interchangeability of differentiation and expectation. The research was originated by Ho and Cao (1983), and is summarized in Glasserman (1991) and Fu and Hu (1997). For the LR/SF method, the interchangeability is often not an issue. However, the variance of the estimator is often high. Related literature includes Reiman and Weiss (1989), Glynn (1990), and Rubinstein and Shapiro (1993).

Although both quantile estimation and gradient estimation have been studied extensively, there are very few papers on the estimation of quantile sensitivities. The only paper that we are aware of is Hong and Qi (2007), who study linear programming with noisy parameters. They propose to minimize the quantile of the noisy objective function. In that paper, they derive a conditional-expectation form of the quantile gradient for the linear function and provide a heuristic estimator. Their numerical results show that their gradient estimator greatly improves the performance of their optimization algorithms when compared to other gradient estimators.

1.2. Contributions and Organization

In this paper, we are interested in the estimation of quantile gradient for general functions, either with or without closed forms. In §2, we define the problem explicitly. In §§3 and 4, we first derive a conditional-expectation form for the gradient of a probability function under some mild conditions. We then show that the quantile gradient can also be written as a conditional expectation based on the relation between probability and quantile. When the conditional expectation can be calculated directly, the quantile gradient can be derived analytically. When it cannot be calculated directly, the conditional-expectation form can also be used to design gradient estimators.

In §5, we propose an estimator using infinitesimal perturbation analysis (IPA). We show that the IPA estimator is asymptotically unbiased but fails to be consistent. In §6, we suggest a new quantile sensitivity estimator by dividing data into batches and averaging the IPA estimates of all batches. We show that this new estimator is consistent if both the number of batches (k) and the number of samples within each batch (m) go to infinity as the total sample size (n) goes to infinity. We also show that the estimator satisfies a central limit theorem when m and k also satisfy $\lim_{n\to\infty}\sqrt{k}/m = 0$, and the rate of convergence is $k^{-1/2}$, which is strictly slower than $n^{-1/3}$.

In §7, we apply our approach to two examples: a portfolio management problem and a production-inventory problem. The numerical results show that the estimator can appropriately estimate quantile sensitivities, and they also provide more insights on the quality of the estimator. In §8, we discuss the estimation of the steady-state quantile sensitivities. The simple numerical example shows that the estimator proposed in this paper also works for dependent data. Finally, the paper is concluded in §9.

2. Problem Definition

Let $h(\theta, X)$ be a function of θ and X, where θ is the parameter with respect to which we differentiate and X is a vector of random variables. In this paper, we assume that θ is one-dimensional and $\theta \in \Theta$, where $\Theta \subset \Re$ is an open set. If θ is multidimensional, we may treat each dimension as a one-dimensional parameter, while fixing other dimensions constants. Because simulation output can be viewed as a function of parameters and random numbers, $h(\theta, X)$ is a general representation of the simulation output. In many cases, there may exist more meaningful representation of θ and X. In portfolio management, for example, θ may be the percentages of the total fund that are allocated to different assets and X may be the (random) annual rates of return of the assets. In Markovian queueing systems, for example, θ may represent the arrival and service rates and X may represent the sequence of exponential random variables with rate 1.

For any $\theta \in \Theta$, let $q_\alpha(\theta)$ be the α-quantile ($0 < \alpha < 1$) of $h(\theta, X)$. If $h(\theta, X)$ has a density in the neighborhood of $q_\alpha(\theta)$, then $\Pr\{h(\theta, X) \leqslant q_\alpha(\theta)\} = \alpha$. Suppose that we have observed $X_1, X_2, \ldots, X_n$. Then, $q_\alpha(\theta)$ can be estimated by the $\lceil n\alpha \rceil$th order statistic of $h(\theta, X)$; see, for example, Serfling (1980) for i.i.d. data and Sen (1972) for dependent data. In this paper, we are interested in estimating $q'_\alpha(\theta) = dq_\alpha(\theta)/d\theta$ using the same data.

3. A Closed Form of Probability Sensitivity

Let $p_a(\theta) = \Pr\{h(\theta, X) \leqslant a\}$ for any real number a. Then, $p_a(\theta)$ is a function of θ. We first derive a closed form of $p'_a(\theta) = dp_a(\theta)/d\theta$ in this section, and then use it to derive a closed form of $q'_\alpha(\theta)$ in §4.

We make the following assumption on $h(\theta, X)$.

ASSUMPTION 1. *The pathwise derivative $\partial_\theta h(\theta, X)$ exists w.p.1 for any $\theta \in \Theta$, and there exists a function $k(X)$ with $\mathrm{E}[k(X)] < \infty$, such that*

$$|h(\theta_2, X) - h(\theta_1, X)| \leqslant k(X)|\theta_2 - \theta_1|$$

for all $\theta_1, \theta_2 \in \Theta$.

Assumption 1 is a typical assumption used in pathwise derivative estimation. Glasserman (1991) develops the commuting conditions for generalized semi-Markov processes under which this assumption holds. Broadie and Glasserman (1996) demonstrate the use of this assumption in estimating price sensitivities of financial derivatives. This assumption guarantees that $\partial_\theta \mathrm{E}[h(\theta, X)] = \mathrm{E}[\partial_\theta h(\theta, X)]$.

Let $F(t; \theta)$ and $f(t; \theta)$ denote the cumulative distribution function and density of $h(\theta, X)$. We make the following assumptions on $f(t; \theta)$ and $F(t; \theta)$.

ASSUMPTION 2. *For any $\theta \in \Theta$, $h(\theta, X)$ has a continuous density $f(t; \theta)$ in a neighborhood of $t = a$, and $\partial_\theta F(t; \theta)$ exists and is continuous with respect to both θ and t at $t = a$.*

Assumption 2 requires that $h(\theta, X)$ is a continuous random variable in a neighborhood of $t = a$. Because $h(\theta, X)$ is typically a continuous function of X and some of X may be continuous random variables, $h(\theta, X)$ is typically a continuous random variable. Furthermore, note that $p_a(\theta) = F(a; \theta)$. Therefore, assuming that $F(a; \theta)$ is differentiable with respect to θ is equivalent to assuming that $p_a(\theta)$ is differentiable with respect to θ.

For any $\theta \in \Theta$, let

$$g(t; \theta) = \mathrm{E}[\partial_\theta h(\theta, X) \mid h(\theta, X) = t].$$

We make the following assumption on $g(t; \theta)$.

ASSUMPTION 3. *For any $\theta \in \Theta$, $g(t; \theta)$ is continuous at $t = a$.*

Note that $h(\theta, X)$ is a continuous random variable in the neighborhood of $t = a$ by Assumption 2. Then, a small change in t often results in a small change in X, and thus a small change in $\partial_\theta h(\theta, X)$. Therefore, $g(t; \theta)$ is typically continuous at $t = a$. Although Assumptions 2 and 3 are weak assumptions, they are typically difficult to verify for practical problems.

To establish the closed form of probability sensitivity, we also need the following lemma that is often used to analyze pathwise derivatives.

LEMMA 1 (BROADIE AND GLASSERMAN 1996). *Let f denote a Lipschitz continuous function and D_f denote the set of points at which f is differentiable. Suppose that Assumption 1 is satisfied and $\Pr\{h(\theta, X) \in D_f\} = 1$ for all $\theta \in \Theta$. Then, at every $\theta \in \Theta$,*

$$\frac{d\mathrm{E}[f(h(\theta, X))]}{d\theta} = \mathrm{E}\left[\frac{\partial f(h(\theta, X))}{\partial \theta}\right].$$

Then, we have the following theorem that gives the closed form of $p_a'(\theta)$.

THEOREM 1. *Suppose that Assumptions 1 to 3 are satisfied. Then,*

$$p_a'(\theta) = -f(a; \theta) \cdot \mathrm{E}[\partial_\theta h(\theta, X) \mid h(\theta, X) = a]. \tag{1}$$

PROOF. By Assumptions 2 and 3, for any $\theta \in \Theta$, we let $(l_\theta, u_\theta) \subset \Theta$ be the neighborhood of a, i.e., $a \in (l_\theta, u_\theta)$, such that $h(\theta, X)$ has a continuous density $f(t; \theta)$ and $g(t; \theta)$ is continuous for any $t \in (l_\theta, u_\theta)$. Let

$$\pi(t; \theta) = \mathrm{E}\big[(t - h(\theta, X)) \cdot 1_{\{h(\theta, X) \leqslant t\}}\big]$$

for any $t \in (l_\theta, u_\theta)$. Then,

$$\begin{aligned}\pi(t; \theta) &= tF(t; \theta) - \mathrm{E}\big[h(\theta, X) \cdot 1_{\{h(\theta, X) \leqslant l_\theta\}}\big] \\ &\quad - \mathrm{E}\big[h(\theta, X) \cdot 1_{\{l_\theta < h(\theta, X) \leqslant t\}}\big] \\ &= tF(t; \theta) - \mathrm{E}\big[h(\theta, X) \cdot 1_{\{h(\theta, X) \leqslant l_\theta\}}\big] - \int_{l_\theta}^{t} vf(v; \theta)\, dv.\end{aligned}$$

Because $f(v; \theta)$ is continuous at $v = t$ by Assumption 2, then

$$\partial_t \pi(t; \theta) = F(t; \theta) + tf(t; \theta) - tf(t; \theta) = F(t; \theta)$$

for any $t \in (l_\theta, u_\theta)$. Then, $\partial_t \pi(a; \theta) = F(a; \theta) = p_a(\theta)$ and $p_a'(\theta) = \partial_\theta \partial_t \pi(a; \theta)$, where we use the (slightly abusive) notation $\partial_t \pi(a; \theta) = \partial_t \pi(t; \theta)|_{t=a}$.

By Assumption 2, $\partial_t \pi(a; \theta)$ is continuous at a and θ. Then, by Marsden and Hoffman (1993, Exercise 24, p. 387),

$$p_a'(\theta) = \partial_\theta \partial_t \pi(a; \theta) = \partial_t \partial_\theta \pi(a; \theta). \tag{2}$$

Let $f(z) = (t - z) \cdot 1_{\{z \leqslant t\}}$. It is Lipschitz continuous and differentiable for any $z \neq t$. Because $\Pr\{L(\theta) = t\} = 0$ for any $t \in (l_\theta, u_\theta)$, then by Assumption 1 and Lemma 1,

$$\begin{aligned}\partial_\theta \pi(t; \theta) &= \frac{d\mathrm{E}[f(h(\theta, X))]}{d\theta} = \mathrm{E}\left[\frac{\partial f(h(\theta, X))}{\partial \theta}\right] \\ &= -\mathrm{E}\big[\partial_\theta h(\theta, X) \cdot 1_{\{h(\theta, X) \leqslant t\}}\big].\end{aligned}$$

Note that for any $t \in (l_\theta, u_\theta)$, we can write

$$\begin{aligned}\partial_\theta \pi(t; \theta) &= -\mathrm{E}\big[\partial_\theta h(\theta, X) \cdot 1_{\{L(\theta) \leqslant t\}}\big] \\ &= -\mathrm{E}\big[\partial_\theta h(\theta, X) \cdot 1_{\{h(\theta, X) \leqslant l_\theta\}}\big] \\ &\quad - \int_{l_\theta}^{t} g(v; \theta) f(v; \theta)\, dv.\end{aligned}$$

Then, by the continuity of $g(t; \theta)$ and $f(t; \theta)$ (Assumptions 2 and 3), $\partial_t \partial_\theta \pi(t; \theta) = -g(t; \theta) f(t; \theta)$. Therefore,

$$\begin{aligned}\partial_t \partial_\theta \pi(a; \theta) &= -g(a; \theta) f(a; \theta) \\ &= -f(a; \theta) \cdot \mathrm{E}[\partial_\theta h(\theta, X) \mid h(\theta, X) = a].\end{aligned}$$

Then, the conclusion of the theorem follows directly from Equation (2). □

REMARKS. (1) Because the indicator function $1_{\{h(\theta, X)\leqslant a\}}$ is not Lipschitz continuous, we cannot apply Lemma 1 to direct differentiate $p_a(\theta) = \mathrm{E}[1_{\{h(\theta, X)\leqslant a\}}]$. This is the major difficulty of estimating probability sensitivity using pathwise derivatives. In the proof of Theorem 1, we develop an approach to overcome this difficulty. We first construct $\pi(t; \theta)$ and show that $\partial_t \pi(a; \theta) = p_a(\theta)$. Then, $p'_a(\theta) = \partial_\theta \partial_t \pi(a; \theta)$. Using the fact that $\partial_\theta \partial_t \pi(a; \theta) = \partial_t \partial_\theta \pi(a; \theta)$, we interchange the order of differentiations to avoid differentiating an indicator function. This approach allows us to obtain a closed form of $p'_a(\theta)$.

(2) Suppose that $h(\theta, U) = F^{-1}(U; \theta)$, where U is a uniform random variable. Then, a well-known result of perturbation analysis (e.g., Suri and Zazanis 1988, Glasserman 1991, and Fu and Hu 1997) states that

$$\partial_\theta h(\theta, U) = -\frac{\partial_\theta F(h(\theta, U); \theta)}{\partial_t F(h(\theta, U); \theta)}.$$

if $\partial_t F(h(\theta, U); \theta) \neq 0$. If we let U satisfy $h(\theta; U) = a$, then

$$\partial_\theta h(\theta, U)|_{h(\theta, U)=a} = -\frac{\partial_\theta F(a; \theta)}{\partial_t F(a; \theta)} = -\frac{p'_a(\theta)}{f(a; \theta)}.$$

Therefore,

$$p'_a(\theta) = -f(a; \theta) \cdot \partial_\theta h(\theta, U)|_{h(\theta, U)=a}. \tag{3}$$

Note that U is uniquely determined by $h(\theta, U) = F^{-1}(U; \theta) = a$. Then, $\partial_\theta h(\theta, U)|_{h(\theta, U)=a}$ is no longer a random variable. It is a constant for any fixed θ. Then,

$$\partial_\theta h(\theta, U)|_{h(\theta, U)=a} = \mathrm{E}[\partial_\theta h(\theta, U) \mid h(\theta, U) = a].$$

Therefore, Equation (3) is a special case of Theorem 1.

4. From Probability Sensitivity to Quantile Sensitivity

Note that

$$F(a, \theta) = p_a(\theta), \quad \partial_\theta F(a, \theta) = p'_a(\theta), \quad \text{and}$$
$$\partial_t F(a, \theta) = f(a; \theta). \tag{4}$$

Now we can state and prove the following result on $q'_\alpha(\theta)$.

THEOREM 2. *Suppose that Assumptions* 1 *to* 3 *are satisfied at* $a = q_\alpha(\theta)$. *Then,*

$$q'_\alpha(\theta) = \mathrm{E}[\partial_\theta h(\theta, X) \mid h(\theta, X) = q_\alpha(\theta)]. \tag{5}$$

PROOF. Because $h(\theta, X)$ is a continuous random variable in the neighborhood of $q_\alpha(\theta)$ by Assumption 2, then we have

$$F(q_\alpha(\theta), \theta) = \alpha. \tag{6}$$

By differentiating with respect to θ on both sides of Equation (6), we have

$$\partial_t F(q_\alpha(\theta), \theta) \cdot q'_\alpha(\theta) + \partial_\theta F(a, \theta)|_{a=q_\alpha(\theta)} = 0.$$

Then, by Theorem 1 and Equation (4),

$$\begin{aligned} q'_\alpha(\theta) &= -\frac{1}{f(q_\alpha(\theta); \theta)} p'_a(\theta)|_{a=q_\alpha(\theta)} \\ &= \mathrm{E}[\partial_\theta h(\theta, X) \mid h(\theta, X) = q_\alpha(\theta)]. \end{aligned}$$

This concludes the proof of the theorem. □

Although the quantile gradient is of the form of a conditional expectation, direct estimation of it is not easy. The event $\{h(\theta, X) = q_\alpha(\theta)\}$ is of probability zero. Therefore, the chance of observing any samples satisfying $h(\theta, X) = q_\alpha(\theta)$ is zero for any finite number of samples. In the next two sections, we focus on estimating quantile sensitivities using an i.i.d. sample of X.

5. An IPA Estimator

In the rest of this paper, we are interested in estimating $q'_\alpha(\theta)$ using an i.i.d. sample $\{X_1, X_2, \ldots, X_n\}$, and we further assume that $h(\theta, X)$ is a continuous random variable for simplicity.

Given n i.i.d. observations $h(\theta, X_1), h(\theta, X_2), \ldots, h(\theta, X_n)$, an estimator of $q_\alpha(\theta)$ is $\hat{q}^n_\alpha(\theta) = h(\theta, X_{(\lceil n\alpha \rceil)})$, where $X_{(k)}$, $k = 1, 2, \ldots, n$, satisfy $h(\theta, X_{(1)}) \leqslant h(\theta, X_{(2)}) \leqslant \cdots \leqslant h(\theta, X_{(n)})$. Note that $X_{(k)}$ is *not* the kth order statistic of X. Because we assume that $h(\theta, X)$ is a continuous random variable, then $h(\theta, X_k) \neq h(\theta, X_l)$ w.p.1 for any $k \neq l$. Therefore,

$$h(\theta, X_{(1)}) < h(\theta, X_{(2)}) < \cdots < h(\theta, X_{(n)}) \quad \text{w.p.1.} \tag{7}$$

Serfling (1980) shows that $\hat{q}^n_\alpha(\theta) \to q_\alpha(\theta)$ w.p.1 as $n \to \infty$ for i.i.d. observations.

A natural method to estimate $q'_\alpha(\theta)$ is to use the IPA method. Because $h(\theta, X)$ is continuous in θ, then by Equation (7),

$$h(\theta + \delta, X_{(1)}) < h(\theta + \delta, X_{(2)}) < \cdots < h(\theta + \delta, X_{(n)}) \text{ w.p.1}$$

when $|\delta|$ is small enough. Because $h(\theta, X)$ is differentiable in θ, then the IPA estimator of $q'_\alpha(\theta)$ can be defined as

$$\begin{aligned} D_n &= \lim_{\delta \to 0} \frac{\hat{q}^n_\alpha(\theta + \delta) - \hat{q}^n_\alpha(\theta)}{\delta} \\ &= \lim_{\delta \to 0} \frac{h(\theta + \delta, X_{(\lceil n\alpha \rceil)}) - h(\theta, X_{(\lceil n\alpha \rceil)})}{\delta} \\ &= \partial_\theta h(\theta, X_{(\lceil n\alpha \rceil)}). \end{aligned} \tag{8}$$

We may write $D_n = \partial_\theta h(\theta, X)|_{h(\theta, X)=\hat{q}^n_\alpha(\theta)}$. Under some technical conditions, we can show that D_n converges weakly to $\partial_\theta h(\theta, X)|_{h(\theta, X)=q_\alpha(\theta)}$ as $n \to \infty$. Note that

$h(\theta, X) = q_\alpha(\theta)$ may not uniquely determine X, e.g., when X is of multidimension. Then, $\partial_\theta h(\theta, X)|_{h(\theta,X)=q_\alpha(\theta)}$ is often a random variable which cannot be $q'_\alpha(\theta)$ because $q'_\alpha(\theta)$ is a constant. Therefore, the IPA estimator D_n is not consistent.

In the next theorem, we show that the expectation of the IPA estimator converges to the quantile sensitivity. Therefore, D_n is asymptotically unbiased.

THEOREM 3. *Suppose that Assumptions* 1 *to* 3 *are satisfied at* $a = q_\alpha(\theta)$ *and* $\sup_n \mathrm{E}(D_n^2) < \infty$. *Then,* $\mathrm{E}(D_n) \to q'_\alpha(\theta)$ *as* $n \to \infty$.

PROOF. Let $F_{\hat{q}}(\cdot)$ denote the c.d.f. of $\hat{q}^n_\alpha(\theta)$ and $\Omega_{\hat{q}}$ denote the set of values that $\hat{q}^n_\alpha(\theta)$ may take. Then, for any set $u \in \Re$,

$$\begin{aligned}
&\mathrm{E}(D_n)\\
&\quad = \mathrm{E}[\partial_\theta h(\theta, X_{(\lceil n\alpha \rceil)})]\\
&\quad = \int_{\Omega_{\hat{q}}} \mathrm{E}[\partial_\theta h(\theta, X_{(\lceil n\alpha \rceil)}) \mid \hat{q}^n_\alpha(\theta) = t]\, dF_{\hat{q}}(t)\\
&\quad = \int_{\Omega_{\hat{q}}} \mathrm{E}\big[\partial_\theta h(\theta, X_{(\lceil n\alpha \rceil)})\\
&\qquad\qquad \mid h(\theta, X_{(\lceil n\alpha \rceil)}) = t, \hat{q}^n_\alpha(\theta) = t\big]\, dF_{\hat{q}}(t). \qquad (9)
\end{aligned}$$

Note that the event $\{h(\theta, X_{(\lceil n\alpha \rceil)}) = t, \hat{q}^n_\alpha(\theta) = t\}$ is equivalent to the event that one of $h(\theta, X_i)$, $i = 1, 2, \ldots, n$, is equal to t and $\lceil n\alpha \rceil - 1$ of them are less than t and the rest of them are greater than t w.p.1. Without loss of generality, we let the last observation X_n satisfy $h(\theta, X_n) = t$. Then,

$$\begin{aligned}
&\big\{h(\theta, X_{(\lceil n\alpha \rceil)}) = t, \hat{q}^n_\alpha(\theta) = t\big\}\\
&\quad = \Big\{h(\theta, X_n) = t, \sum_{i=1}^{n-1} 1_{\{h(\theta, X_i) < t\}} = \lceil n\alpha \rceil - 1,\\
&\qquad \sum_{i=1}^{n-1} 1_{\{h(\theta, X_i) > t\}} = n - 1 - \lceil n\alpha \rceil \Big\} \quad \text{w.p.1.}
\end{aligned}$$

Because X_n is independent of $X_1, X_2, \ldots, X_{n-1}$, then by Equation (9),

$$\begin{aligned}
\mathrm{E}(D_n) &= \int_{\Omega_{\hat{q}}} \mathrm{E}[\partial_\theta h(\theta, X_n) \mid h(\theta, X_n) = t]\, dF_{\hat{q}}(t)\\
&= \int_{\Omega_{\hat{q}}} \mathrm{E}[\partial_\theta h(\theta, X) \mid h(\theta, X) = t]\, dF_{\hat{q}}(t)\\
&= \int_{\Omega_{\hat{q}}} g(t; \theta)\, dF_{\hat{q}}(t) = \mathrm{E}[g(\hat{q}^n_\alpha(\theta); \theta)]. \qquad (10)
\end{aligned}$$

Because $g(t; \theta)$ is continuous at $t = q_\alpha(\theta)$ and $\hat{q}^n_\alpha(\theta) \to q_\alpha(\theta)$ w.p.1 as $n \to \infty$, then by the continuous mapping theorem (Durrett 1996), $g(\hat{q}^n_\alpha(\theta); \theta) \Rightarrow g(q_\alpha(\theta); \theta)$ as $n \to \infty$. Because $q'_\alpha(\theta) = g(q_\alpha(\theta); \theta)$ by Theorem 2, then we only need to prove that $\{g(\hat{q}^n_\alpha(\theta); \theta)\}$ is uniformly integrable.

Note that

$$\begin{aligned}
\mathrm{E}(D_n^2) &= \mathrm{E}\{[\partial_\theta h(\theta, X_{(\lceil n\alpha \rceil)})]^2\}\\
&= \int_{\Omega_{\hat{q}}} \mathrm{E}\{[\partial_\theta h(\theta, X)]^2 \mid h(\theta, X) = t\}\, dF_{\hat{q}}(t) \qquad (11)\\
&\geqslant \int_{\Omega_{\hat{q}}} \mathrm{E}^2[\partial_\theta h(\theta, X) \mid h(\theta, X) = t]\, dF_{\hat{q}}(t) \qquad (12)\\
&= \int_{\Omega_{\hat{q}}} g^2(t; \theta)\, dF_{\hat{q}}(t) = \mathrm{E}[g^2(\hat{q}^n_\alpha(\theta); \theta)],
\end{aligned}$$

where Equation (11) follows from the derivation of Equation (10) and Equation (12) follows from Jensen's inequality (Durrett 1996). Because $\sup_n \mathrm{E}(D_n^2) < \infty$, then $\sup_n \mathrm{E}[g^2(\hat{q}^n_\alpha(\theta); \theta)] < \infty$. Therefore, $\{g(\hat{q}^n_\alpha(\theta); \theta)\}$ is uniformly integrable. This concludes the proof of the theorem. □

The IPA estimator is asymptotically unbiased, but it is not consistent. Therefore, it is not an appropriate estimator of the quantile sensitivities. To illustrate this important issue, we consider the following example. Let $h(\theta, X) = \theta X_1 + X_2$, where X_1 and X_2 are independent standard normal random variables. Then, $h(\theta, X)$ follows a normal distribution with mean zero and variance $\theta^2 + 1$. Let z_α denote the α quantile of a standard normal distribution. Then, $q_\alpha(\theta) = z_\alpha\sqrt{\theta^2 + 1}$ and $q'_\alpha(\theta) = z_\alpha\theta/\sqrt{\theta^2 + 1}$. The IPA estimator converges to $X_1 \mid [\theta X_1 + X_2 = q_\alpha(\theta)]$, which follows a normal distribution with mean $z_\alpha\theta/\sqrt{\theta^2 + 1}$ and variance $1/(\theta^2 + 1)$. It is a random variable. However, $\mathrm{E}[X_1 \mid \theta X_1 + X_2 = q_\alpha(\theta)] = z_\alpha\theta/\sqrt{\theta^2 + 1} = q'_\alpha(\theta)$. This example shows that the IPA estimator is not consistent, but it is asymptotically unbiased.

6. A Consistent Estimator

Suppose that there exist positive integers m and k such that $m \times k = n$. Then, we divide the n i.i.d. observations into k batches and each batch has m observations. For each batch, we calculate the IPA estimator D_m. Then, we have k observations of D_m, denoted as $D_{m1}, D_{m2}, \ldots, D_{mk}$. Let $\bar{D}_{mk} = (1/k)\sum_{l=1}^k D_{ml}$. In the next theorem, we show that $\bar{D}_{mk}$ is a consistent estimator of $q'_\alpha(\theta)$. In the rest of the paper, we use $\xrightarrow{P}$ to denote "converges in probability."

THEOREM 4. *Suppose that Assumptions* 1–3 *are satisfied at* $a = q_\alpha(\theta)$, $\sup_m \mathrm{E}(D_m^2) < \infty$, *and* $m \to \infty$ *and* $k \to \infty$ *as* $n \to \infty$. *Then,*

$$\bar{D}_{mk} \xrightarrow{P} q'_\alpha(\theta) \quad \text{as } n = mk \to \infty.$$

PROOF. By Theorem 3,

$$\mathrm{E}(D_m) \to q'_\alpha(\theta) \quad \text{as } m \to \infty. \qquad (13)$$

For any $\epsilon > 0$, by the Chebyshev's inequality,

$$\Pr\{|\bar{D}_{mk} - \mathrm{E}(D_m)| \geqslant \epsilon\} \leqslant \frac{\mathrm{Var}(D_m)}{k\epsilon^2} \leqslant \frac{\sup_m \mathrm{E}(D_m^2)}{k\epsilon^2}.$$

Then, $\Pr\{|\bar{D}_{mk} - \mathrm{E}(D_m)| \geqslant \epsilon\} \to 0$ as $k \to \infty$. Therefore,

$$\bar{D}_{mk} - \mathrm{E}(D_m) \xrightarrow{P} 0 \quad \text{as } k \to \infty. \tag{14}$$

Because $n \to \infty$ implies both $m \to \infty$ and $k \to \infty$, then the conclusion of the theorem follows directly from Equations (13) and (14). □

From the proof of Theorem 4, we see that the estimation error comes from two parts: the within-batch error, $\mathrm{E}(D_m) - q'_\alpha(\theta)$, and the across-batch error, $\bar{D}_{mk} - \mathrm{E}(D_m)$. The within-batch error is the bias of the estimator, and the across-batch error is caused by the variance of $\bar{D}_{mk}$. Theorem 4 shows that both the bias and variance go to zero as $n \to \infty$. Therefore, $\bar{D}_{mk}$ is consistent.

In the rest of this section, we show that $\bar{D}_{mk}$ follows a central limit theorem under some conditions. Then, we may construct asymptotically valid confidence intervals based on the theorem. We also discuss the relation between m and k to balance the bias and variance.

To study the asymptotic normality of $\bar{D}_{mk}$, we first prove the following lemmas on the rates of convergence of the bias and the mean square error (MSE) of $\hat{q}^m_\alpha(\theta)$.

Lemma 2. *Suppose that $f(t;\theta)$ is continuously differentiable at $t = q_\alpha(\theta)$ and $f(q_\alpha(\theta);\theta) > 0$. Then, both $\mathrm{E}[\hat{q}^m_\alpha(\theta) - q_\alpha(\theta)]$ and $\mathrm{E}[|\hat{q}^m_\alpha(\theta) - q_\alpha(\theta)|^2]$ are of $O(m^{-1})$.*

Proof. Let $F^{-1}(\cdot;\theta)$ be the inverse c.d.f. of $h(\theta, X)$. Then, by Equations (4.6.3) and (4.6.4) of David (1981),

$$\mathrm{E}[\hat{q}^m_\alpha(\theta)] = F^{-1}(p_m;\theta) + \frac{p_m(1-p_m)}{2(m+2)}(F^{-1})''(p_m;\theta) + o(m^{-1}), \tag{15}$$

$$\mathrm{Var}[\hat{q}^m_\alpha(\theta)] = \frac{p_m(1-p_m)}{m+1}[(F^{-1})'(p_m;\theta)]^2 + o(m^{-1}), \tag{16}$$

where $p_m = \lceil m\alpha \rceil/(m+1)$. Note that

$$(F^{-1})'(p_m;\theta) = \frac{1}{f(F^{-1}(p_m;\theta);\theta)} \quad \text{and}$$

$$(F^{-1})''(p_m;\theta) = -\frac{f'(F^{-1}(p_m;\theta);\theta)}{f^3(F^{-1}(p_m;\theta);\theta)}.$$

Because $f(t;\theta)$ exists in a neighborhood of $t = q_\alpha(\theta)$, then $F^{-1}(y;\theta)$ is continuous in a neighborhood of $y = \alpha$. Furthermore, because

$$|p_m - \alpha| = \left|\frac{\lceil m\alpha \rceil}{m+1} - \alpha\right| \leqslant \frac{1}{m+1},$$

then when m is sufficiently large, $F^{-1}(p_m;\theta)$ can be in any neighborhood of α. Therefore, when $f(t;\theta)$ is continuously differentiable at $t = q_\alpha(\theta)$ and $f(q_\alpha(\theta);\theta) > 0$, $(F^{-1})'(p_m;\theta)$ and $(F^{-1})''(p_m;\theta)$ exist and are bounded when m is sufficiently large.

Because $q_\alpha(\theta) = F^{-1}(\alpha;\theta)$, then by Equation (15),

$$\mathrm{E}[\hat{q}^m_\alpha(\theta) - q_\alpha(\theta)] = F^{-1}(p_m;\theta) - F^{-1}(\alpha;\theta) + \frac{p_m(1-p_m)}{2(m+2)}(F^{-1})''(p_m;\theta) + o(m^{-1}). \tag{17}$$

By Taylor's theorem,

$$F^{-1}(p_m;\theta) - F^{-1}(\alpha;\theta) = (F^{-1})'(\alpha;\theta)(p_m - \alpha) + o(|p_m - \alpha|).$$

Because $|p_m - \alpha| \leqslant 1/(m+1)$, then $F^{-1}(p_m;\theta) - F^{-1}(\alpha;\theta)$ is of $O(m^{-1})$. By Equation (17), $\mathrm{E}[\hat{q}^m_\alpha(\theta) - q_\alpha(\theta)]$ is also of $O(m^{-1})$.

Note that

$$\mathrm{E}[|\hat{q}^m_\alpha(\theta) - q_\alpha(\theta)|^2] = (\mathrm{E}[\hat{q}^m_\alpha(\theta) - q_\alpha(\theta)])^2 + \mathrm{Var}[\hat{q}^m_\alpha(\theta)].$$

Because $\mathrm{E}[\hat{q}^m_\alpha(\theta) - q_\alpha(\theta)]$ is of $O(m^{-1})$, then $(\mathrm{E}[\hat{q}^m_\alpha(\theta) - q_\alpha(\theta)])^2$ is of $o(m^{-1})$. By Equation (16), $\mathrm{Var}[\hat{q}^m_\alpha(\theta)]$ is of $O(m^{-1})$. Therefore, $\mathrm{E}[|\hat{q}^m_\alpha(\theta) - q_\alpha(\theta)|^2]$ is of $O(m^{-1})$. This concludes the proof of the lemma. □

From Lemma 2, we see that both the bias and variance of the quantile estimator converge in the order of m^{-1}. The numerical experiments also show that both the bias and variance do not go to zero monotonically as the sample size increases, and they vary with different α values. To illustrate the behaviors of the bias and variance, we compute the bias and variance of the quantile estimator of the standard normal distribution and plot them in Figures 1 and 2. From the figures, we see that both the bias and variance display oscillatory behaviors as the sample size increases, and they also increase drastically when α approaches zero or one.

In the following lemma, we study the rate of convergence of the bias of $\bar{D}_{mk}$.

Lemma 3. *Suppose that $g(t;\theta)$ is twice differentiable with respect to t and $|\partial_t^2 g(t;\theta)| \leqslant M$ for some $M > 0$. Then, $\mathrm{E}(D_m) - q'_\alpha(\theta)$ is of $O(m^{-1})$.*

Proof. By Equation (10), we have $\mathrm{E}(D_m) = \mathrm{E}[g(\hat{q}^m_\alpha(\theta);\theta)]$. Then, it suffices to prove that $\mathrm{E}[g(\hat{q}^m_\alpha(\theta);\theta)] - q'_\alpha(\theta)$ is of $O(m^{-1})$ as $m \to \infty$.

By Taylor's theorem,

$$g(\hat{q}^m_\alpha(\theta);\theta) = g(q_\alpha(\theta);\theta) + \partial_t g(q_\alpha(\theta);\theta)[\hat{q}^m_\alpha(\theta) - q_\alpha(\theta)] + \partial_t^2 g(Y;\theta)|\hat{q}^m_\alpha(\theta) - q_\alpha(\theta)|^2$$

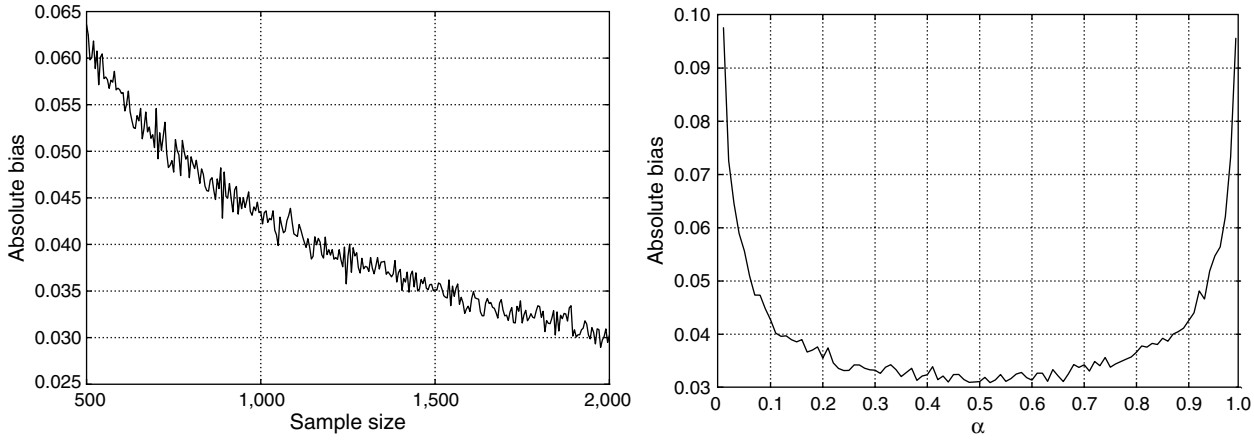
for some random variable Y. Then,

$$\begin{aligned}|\mathrm{E}(D_m) - q'_\alpha(\theta)| &= |\mathrm{E}[g(\hat{q}^m_\alpha(\theta);\theta) - g(q_\alpha(\theta);\theta)]| \\ &\leqslant |\partial_t g(q_\alpha(\theta);\theta)| \cdot |\mathrm{E}[\hat{q}^m_\alpha(\theta) - q_\alpha(\theta)]| \\ &\quad + M \cdot \mathrm{E}[|\hat{q}^m_\alpha(\theta) - q_\alpha(\theta)|^2].\end{aligned} \tag{18}$$

By Lemma 2, the conclusion of this lemma holds. □

Let $\sigma_m^2 = \mathrm{Var}(D_m)$. Then, we have the following theorem that characterizes the asymptotic distribution of $\bar{D}_{mk}$.

Figure 1. Bias of the quantile estimator.



THEOREM 5. *Suppose that Assumptions* 1 *to* 3 *are satisfied at* $a = q_\alpha(\theta)$, *the conditions of Lemmas* 2 *and* 3 *hold, and* $\sup_m \mathrm{E}(|D_m|^{2+\gamma}) < \infty$ *for some* $\gamma > 0$. *If both* $k \to \infty$ *and* $m \to \infty$ *as* $n \to \infty$ *and* $\lim_{n\to\infty} \sqrt{k}/m = 0$, *and* $\sigma_m > 0$ *for any* $m > 0$, *then*

$$\frac{\sqrt{k}}{\sigma_m}[\bar{D}_{mk} - q'_\alpha(\theta)] \;\Rightarrow\; N(0,1) \quad \text{as } n = mk \to \infty.$$

PROOF. Because $\sup_m \mathrm{E}(|D_m|^{2+\gamma}) < \infty$ for some $\gamma > 0$, by Lyapounov's central limit theorem (Theorem 27.3 of Billingsley 1995),

$$\frac{\sqrt{k}}{\sigma_m}[\bar{D}_{mk} - \mathrm{E}(D_m)] \;\Rightarrow\; N(0,1) \quad \text{as } n \to \infty.$$

Because $\lim_{n\to\infty} \sqrt{k}/m = 0$ and $\sigma_m > 0$, then by Lemma 3,

$$\frac{\sqrt{k}}{\sigma_m}[\mathrm{E}(D_m) - q'_\alpha(\theta)] = \frac{\sqrt{k}}{m\sigma_m} \cdot m[\mathrm{E}(D_m) - q'_\alpha(\theta)] \to 0 \quad \text{as } n \to \infty.$$

Therefore,

$$\frac{\sqrt{k}}{\sigma_m}[\bar{D}_{mk} - q'_\alpha(\theta)] = \frac{\sqrt{k}}{\sigma_m}[\bar{D}_{mk} - \mathrm{E}(D_m)] + \frac{\sqrt{k}}{\sigma_m}[\mathrm{E}(D_m) - q'_\alpha(\theta)] \;\Rightarrow\; N(0,1)$$

as $n = mk \to \infty$. This concludes the proof of the theorem. □

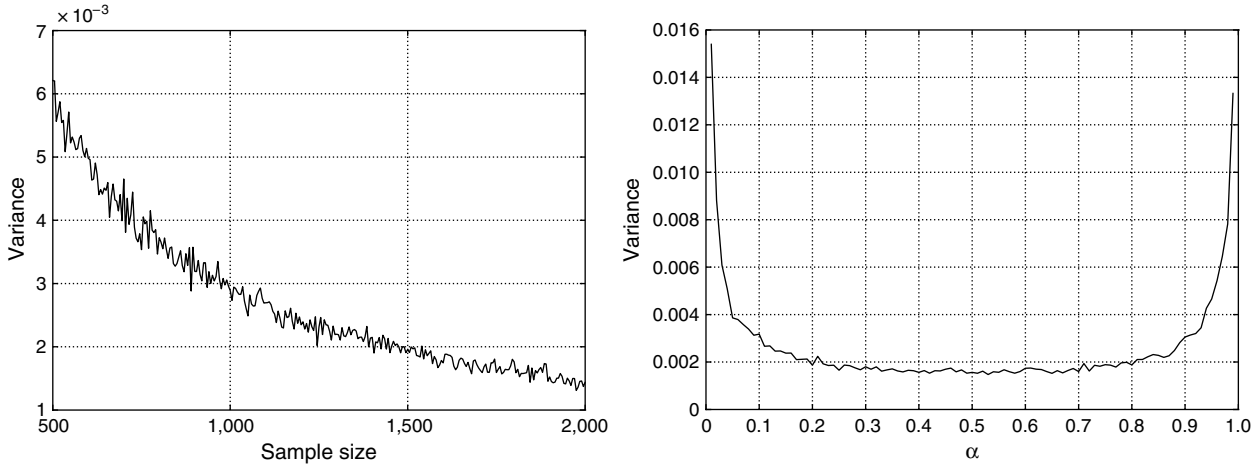
Because σ_m^2 is typically unknown, we can use the sample variance

$$S_{mk}^2 = \frac{1}{k-1}\sum_{i=1}^{k}(D_{mi} - \bar{D}_{mk})^2$$

to estimate σ_m^2. In the appendix, we show that $S_{mk}^2/\sigma_m^2 \to 1$ in probability under certain conditions. Therefore,

$$\frac{\sqrt{k}}{S_{mk}}[\bar{D}_{mk} - q'_\alpha(\theta)] \;\Rightarrow\; N(0,1) \quad \text{as } n \to \infty.$$

Figure 2. Variance of the quantile estimator.



Then, an asymptotically valid $100(1-\beta)\%$ confidence interval of $q'_\alpha(\theta)$ is

$$(\bar{D}_{mk} - z_{1-\beta/2}S_{mk}/\sqrt{k}, \quad \bar{D}_{mk} + z_{1-\beta/2}S_{mk}/\sqrt{k}), \tag{19}$$

where $z_{1-\beta/2}$ is the $1-\beta/2$ quantile of the standard normal distribution.

From Lemma 3 and Theorem 5, we see that the variance of $\bar{D}_{mk}$ converges in the rate of $1/k$ and the bias of $\bar{D}_{mk}$ goes to zero in the rate of $1/m$. Therefore, the rate of convergence of $\bar{D}_{mk}$ is $k^{-1/2}$, which is always strictly slower than $n^{-1/3}$. In practice, one might set $k = n^{2/3-\delta}$ and $m = n^{1/3+\delta}$ for some $0 < \delta < 2/3$. We suggest to set $\delta = 1/6$, then $k = m = \sqrt{n}$. The numerical results reported in §7 show that it is often a good and robust choice for both the estimator and confidence intervals. To minimize the asymptotic MSE of the estimator, however, we may choose δ that is close to zero. When δ is close to zero, the bias of $(\sqrt{k}/\sigma_m)[\bar{D}_{mk} - \mathrm{E}(D_m)]$ is often significantly different from zero when n is not large enough. Then, the actual coverage probability of the confidence interval is often less than the nominal coverage probability. Furthermore, by Equation (18), the bias of $\bar{D}_{mk}$ is related to the bias and variance of the quantile estimator, which both increase drastically when α approaches zero or one (see Figures 1 and 2). Therefore, m often needs to be large to ensure a low bias when α is close to zero or one.

7. Numerical Study

In this section, we study the performances of the quantile sensitivity estimator through two examples: a portfolio management problem and a production-inventory problem. For the first example, the analytical quantile sensitivity can be derived. Then, we use the MSE of the point estimator and the coverage probability of the 90% confidence interval to evaluate the performances of our estimator. For the second example, the analytical quantile sensitivity is not available. Then, we compare the values of the point estimators and the half widths of the confidence intervals of our method and the finite-difference method.

7.1. A Portfolio Management Problem

A portfolio is composed of three assets. The annual rates of return of the assets are denoted as X_1, X_2, and X_3, and the percentages of the total fund allocated to the assets are denoted as θ_1, θ_2, and θ_3. Suppose that $X = (X_1, X_2, X_3)'$ follows a multivariate normal distribution with mean vector $\mu = (0.06, 0.15, 0.25)'$ and variance-covariance matrix

$$\Sigma = \begin{pmatrix} 0.02 & & \\ & 0.10 & \\ & & 0.22 \end{pmatrix} \begin{pmatrix} 1 & -0.3 & -0.2 \\ -0.3 & 1 & 0.2 \\ -0.2 & 0.2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0.02 & & \\ & 0.10 & \\ & & 0.22 \end{pmatrix}.$$

Then, the portfolio annual rate of return is

$$h(\theta, X) = \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3,$$

which follows a normal distribution with mean $\theta'\mu$ and variance $\theta'\Sigma\theta$. Then, the quantile and quantile sensitivities of $h(\theta, X)$ can be calculated analytically. Suppose that we are interested in the quantile sensitivity with respect to θ_3, with $\theta = (0.2, 0.3, 0.5)'$. Then,

$$\partial_{\theta_3} q_\alpha(\theta) = 0.25 + 0.2135 z_\alpha,$$

where z_α is the α quantile of the standard normal distribution. In this subsection, we use the method developed in this paper to estimate the quantile sensitivity, and compare the estimates to the actual value.

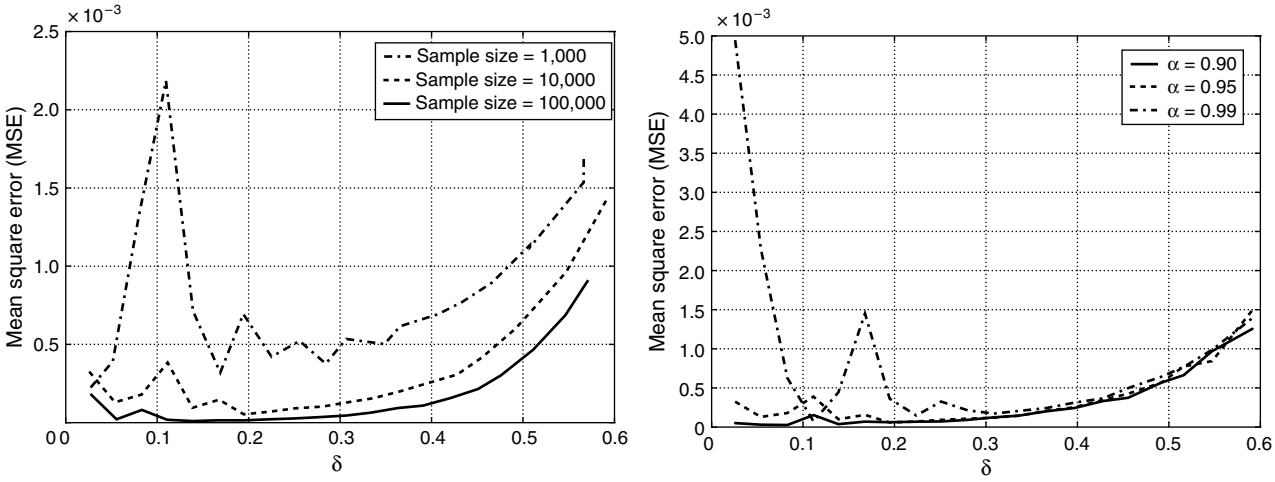
Because $\partial_{\theta_3} h(\theta, X) = X_3$, we can calculate the IPA estimate for each batch. Then, we can calculate the point estimator $\bar{D}_{mk}$. In all experiments reported in this subsection, the results are based on 1,000 independent replications.

We first study the rate of convergence of the estimator. We fix $\alpha = 0.9$, let $k = n^{2/3-\delta}$ and $m = n^{1/3+\delta}$, and plot the relations between δ and the MSE for different sample sizes (see the left panel of Figure 3). Although a smaller δ value corresponds to a better asymptotic rate of convergence, it often has a larger bias when the sample size is not large enough. We also fix sample size $n = 10{,}000$, and plot the relations between δ and the MSE for different α values (see the right panel of Figure 3). When α becomes closer to one, the bias becomes larger (as in Figure 1). Therefore, the MSE also becomes larger. In the rest of this section, we let $m = k = \sqrt{n}$ to study the behaviors of the estimator and confidence interval. However, when α becomes close to zero or one, we suggest to use larger m.

To verify the consistency and asymptotic normality, we let $\alpha = 0.9$ and increase the sample size from 1,000 to 100,000. The MSEs and coverage probabilities are reported in Figure 4. Note that the oscillatory behaviors of both figures are caused by the oscillatory behaviors of the bias and variance of the quantile estimator (see Figures 1 and 2). The left panel of Figure 4 shows that the MSE becomes smaller and less oscillatory as the sample size increases. With 50,000 samples, the square root of the MSE is already below 1% of the actual quantile sensitivity. Therefore, our point estimator appears to be consistent. The right panel of Figure 4 shows that the coverage probability becomes closer to 90% and less oscillatory as the sample size increases. Therefore, the confidence interval that we propose appears to be asymptotically valid.

We also fix the sample size to 40,000 and vary the α values to study the effect of α to the performances of the point estimator and the confidence interval. The results are reported in Figure 5. From the experiments, we see that the point estimator and the confidence interval work reasonably well for nonextreme α values. However, when α approaches zero or one, the qualities of the point estimator and the confidence interval reduce drastically. This is due to the bias caused by the inaccurate estimation of the extreme quantiles (see Figures 1 and 2).

Figure 3. The analysis of the rate of convergence.



7.2. A Production-Inventory Problem

A capacitated production system operates under a base-stock inventory policy. It has a base-stock level $s > 0$, and it has a capacity of producing maximum c units per period. Within each period, production of the last period first arrives. Then, the demands of the period occur, and they are filled or backlogged based on the available inventory. At the end of the period, the production amount is determined. Let I_i be the inventory minus backlog in period i, and D_i and R_i be the demand and production amount in period i, respectively. Then, the system evolves as follows (Glasserman and Tayur 1995):

$$I_{i+1} = I_i - D_i + R_{i-1},$$

$$R_i = \min\{c, [s + D_i - (I_i + R_{i-1})]^+\},$$

where $a^+ = \max\{a, 0\}$.

In this example, we further assume that there are linear holding and backorder costs. The holding cost is h per unit per period and the backorder cost is b per unit per period. Let c_i be the cost of period i. Then,

$$c_i = h(R_{i-1} + I_i^+) + bI_i^-,$$

where $a^- = -\min\{a, 0\}$. The performance measure we are interested in is

$$h(s, D) = \sum_{i=1}^{n} c_i,$$

which is the total cost over the first n periods and $D = (D_1, D_2, \ldots, D_n)$. Because the total cost is a random variable and the decision maker is risk averse, we are interested in the α-quantile of the total cost with $\alpha \geqslant 0.5$. To find an optimal base-stock level s, we assume that s is a continuous decision variable and we are interested in finding the quantile sensitivity with respect to it.

Glasserman and Tayur (1995) have studied this problem under a more general setting. However, they are interested

Figure 4. Performances of the point estimator and the 90% confidence interval for different sample sizes.

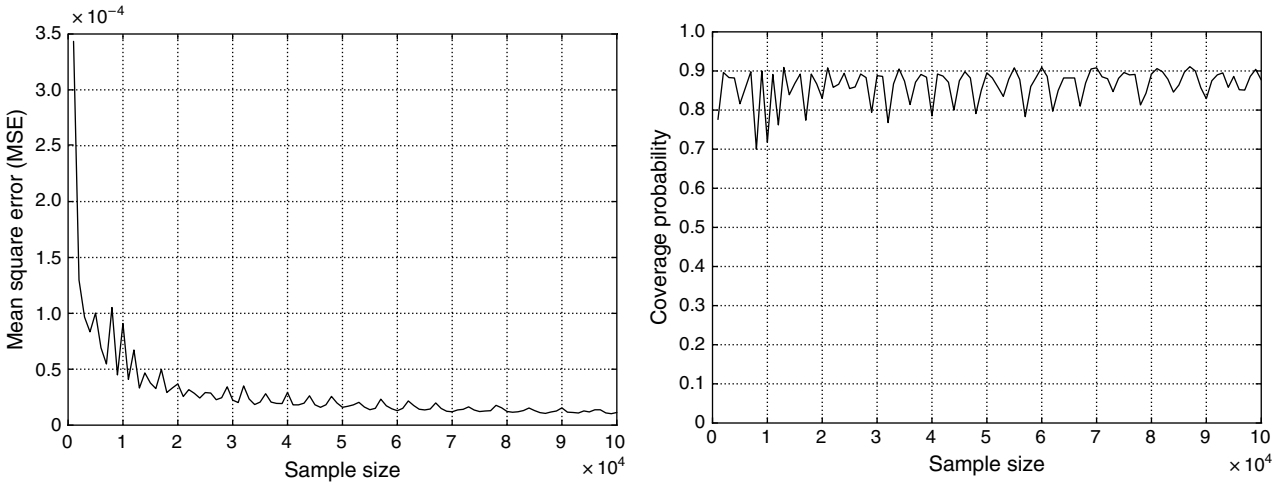
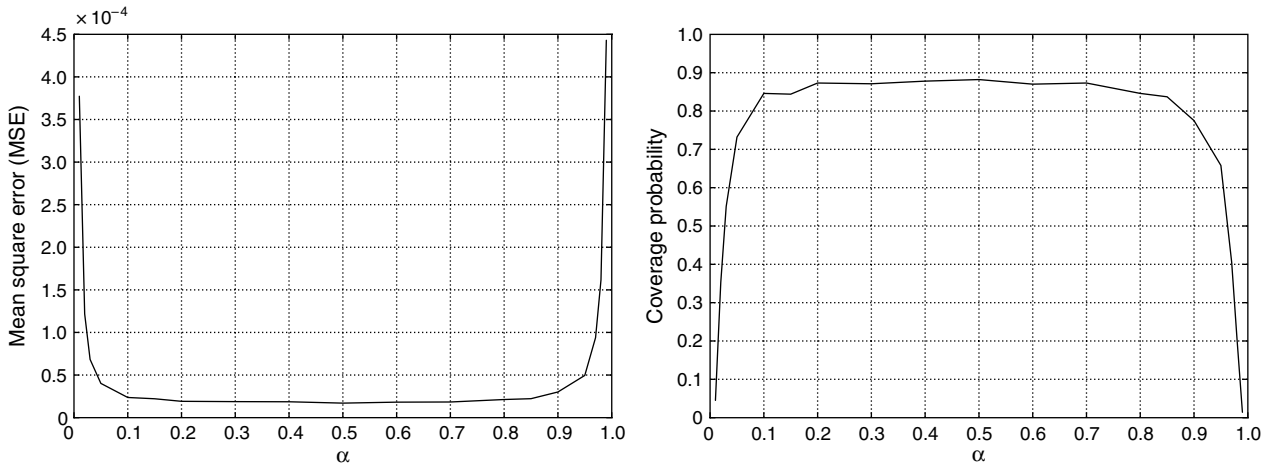


Figure 5. Performances of the point estimator and the 90% confidence interval for different α values.



in the sensitivity of the expected total cost. In the rest of this subsection, we combine the IPA estimator proposed by Glasserman and Tayur (1995) and the consistent estimator of this paper to estimate the quantile sensitivity. We set $s = 1.5$, $c = 0.5$, $h = 0.1$, $b = 0.2$, $n = 20$, $I_1 = s$, and $R_0 = 0$. We further let D_i follow an exponential distribution with rate 1 for all n periods.

To compute $\partial_s h(s, D) = \sum_{i=1}^{n} \partial_s c_i$, we need to know how to calculate $\partial_s c_i$. Note that

$$\partial_s c_i = \partial_s R_{i-1}\ h + 1_{\{I_i>0\}}\partial_s I_i h - 1_{\{I_i<0\}}\partial_s I_i b,$$

where $1_{\{\cdot\}}$ is the indicator function. From the recursive relations of I_i and R_i, we can show that $\partial_s I_i = 1$ and $\partial_s R_i = 0$ for all $i = 1, 2, \ldots, n$. Here, we give an intuitive explanation of the results: if we increase the base-stock level by one, the inventory level I_i will be one unit higher for all periods because the unsatisfied inventory will be backlogged. Therefore, $\partial_s I_i = 1$. Because the production R_i tries to bring the inventory level to the base-stock level if the capacity is allowed, it does not depend on the base-stock level—it only depends on the inventory consumption (which is the demand) and capacity. Therefore, $\partial_s R_i = 0$. Then,

$$\partial_s c_i = 1_{\{I_i>0\}} h - 1_{\{I_i<0\}} b.$$

We let $m = k = 1{,}000$. Then, the total sample size $n = 1 \times 10^6$. We apply our method to estimate the quantile sensitivities with $\alpha = 0.5, 0.6, \ldots, 0.9$. The estimates and the half widths of the 90% confidence intervals are reported in Table 1, and compared to the finite-difference estimator calculated by setting $s_1 = 1.4975$ and $s_2 = 1.5025$, each with 1×10^{10} observations. In the finite-difference estimation, we use independent observations for s_1 and s_2. From Table 1, we see that our estimators are statistically indifferent from the finite-difference estimators. They achieve good precisions with a reasonable sample size.

8. Discussions on Estimating Steady-State Quantile Sensitivities

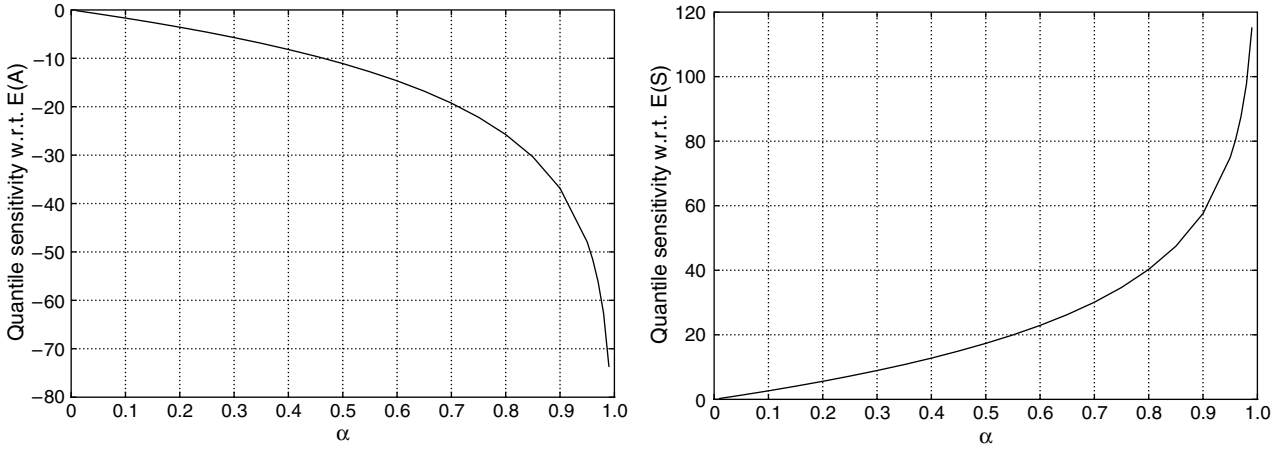
So far in the paper we assume that $h(\theta, X_1), h(\theta, X_2), \ldots, h(\theta, X_n)$ are a sequence of i.i.d. data. In many simulation studies, especially the simulation of queueing systems, we are often interested in the steady-state behavior of the system. In such cases, $h(\theta, X_1), h(\theta, X_2), \ldots, h(\theta, X_n)$ are often a sequence of dependent data. For example, $h(\theta, X_l)$, $l = 1, 2, \ldots,$ may be the waiting time of the lth customer. Then, it is certainly correlated to $h(\theta, X_{l-1})$. Sen (1972) shows that, under some conditions, $\hat{q}_\alpha^n(\theta) = h(\theta, X_{(\lceil n\alpha \rceil)})$ is still a strongly consistent estimator of $q_\alpha(\theta)$ and $\sqrt{n}[\hat{q}_\alpha^n(\theta) - q_\alpha(\theta)]$ converges in distribution to a normal random variable as well.

To estimate the quantile sensitivities using dependent data, we also suggest to use the estimator developed in this paper. Note that, by Sen (1972), the within-batch dependence does not affect the property of D_m, i.e., D_m still converges to the conditional random variable $\partial_\theta h(\theta, X) \mid [h(\theta, X) = q_\alpha(\theta)]$. The across-batch dependence certainly affects the quality of $\bar{D}_{mk}$. However, as $m \to \infty$, the dependences among the batches often become negligible, as studied in the batch-means literature (Law and Kelton 2000). Therefore, it appears possible to prove that $\bar{D}_{mk}$ is also consistent as $n \to \infty$ for the dependent data under certain conditions. Because $\hat{q}_\alpha^n(\theta)$ has an asymptotic normal

Table 1. Quantile sensitivity estimators for the production-inventory problem.

	Finite difference (2×10^{10} obs.)		Our method (1×10^6 obs.)	
α	Estimate	Half width	Estimate	Half width
0.5	−2.7005	0.0512	−2.7598	0.0321
0.6	−2.8455	0.0554	−2.8864	0.0291
0.7	−2.9645	0.0617	−2.9431	0.0292
0.8	−3.0478	0.0720	−3.0709	0.0246
0.9	−3.1645	0.0947	−3.2071	0.0215

Figure 6. Actual values of the quantile sensitivities for the $M/M/1$ queueing model.



distribution for the dependent data, it may also be possible to prove that $\bar{D}_{mk}$ also has an asymptotic normal distribution under certain conditions.

Establishing the consistency and asymptotic normality of the quantile sensitivity estimator for the steady-state simulation is beyond the scope of this paper. However, in this section, we use a simple $M/M/1$ queueing model to show that our estimator may still work for the dependent sequences. Let $\theta = (\mathrm{E}(A), \mathrm{E}(S))'$, where $\mathrm{E}(A)$ and $\mathrm{E}(S)$ are the mean interarrival time and mean service time, respectively, let X be the sequence of random variables that are exponentially distributed with rate 1, and let $h(\theta, X)$ be the customer's waiting time in the system, including both waiting time in the queue and service time, in the steady state of an $M/M/1$ queue. We are interested in the estimation of $\partial q_\alpha(\theta)/\partial \mathrm{E}(A)$ and $\partial q_\alpha(\theta)/\partial \mathrm{E}(S)$, where $q_\alpha(\theta)$ is the α quantile of $h(\theta, X)$.

When the queue is stable, i.e., $\mathrm{E}(A) > \mathrm{E}(S)$, $h(\theta, X)$ is exponentially distributed with rate $(1/\mathrm{E}(S)) - (1/\mathrm{E}(A))$ (Ross 1996). Therefore, for any $0 < \alpha < 1$,

$$q_\alpha(\theta) = -\frac{\mathrm{E}(A)\mathrm{E}(S)}{\mathrm{E}(A) - \mathrm{E}(S)} \log(1-\alpha),$$

$$\frac{\partial q_\alpha(\theta)}{\partial \mathrm{E}(A)} = \left[\frac{\mathrm{E}(S)}{\mathrm{E}(A) - \mathrm{E}(S)}\right]^2 \log(1-\alpha), \tag{20}$$

$$\frac{\partial q_\alpha(\theta)}{\partial \mathrm{E}(S)} = -\left[\frac{\mathrm{E}(A)}{\mathrm{E}(A) - \mathrm{E}(S)}\right]^2 \log(1-\alpha). \tag{21}$$

To obtain the point estimator $\bar{D}_{mk}$, we need $\partial h(\theta, X)/\partial E(A)$ and $\partial h(\theta, X)/\partial E(S)$, which can be computed through the IPA method (Glasserman 1991).

Let $\mathrm{E}(A) = 10$ and $\mathrm{E}(S) = 8$. Then, the actual quantile sensitivities can be computed using Equations (20) and (21). We plot them against different α values in Figure 6, with the left panel being the sensitivities with respect to $\mathrm{E}(A)$ and the right panel being the sensitivities with respect to $\mathrm{E}(S)$. To estimate the quantile sensitivities, we use a total sample size 1×10^6 for various α values. Because quantiles are more difficult to estimate for dependent data than for i.i.d. data, we divide the 1×10^6 observations into 100 batches and each batch has 10,000 observations.

In Figure 7, we report the MSE of the estimators. Although the MSEs increase as α increases, the actual sensitivities also increase as α increases (see Figure 6). With

Figure 7. MSEs of the sensitivity estimators for the $M/M/1$ queueing model.

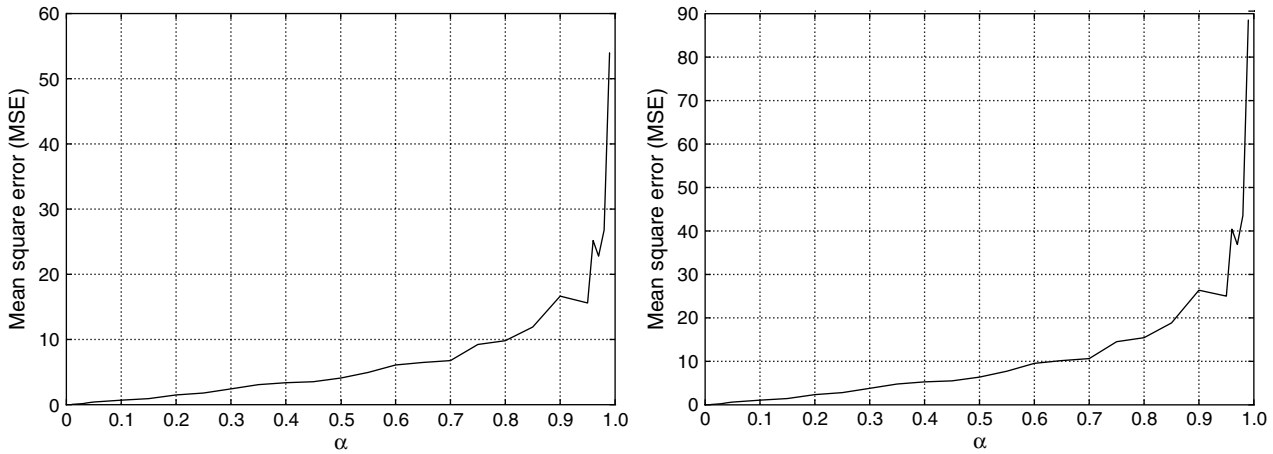
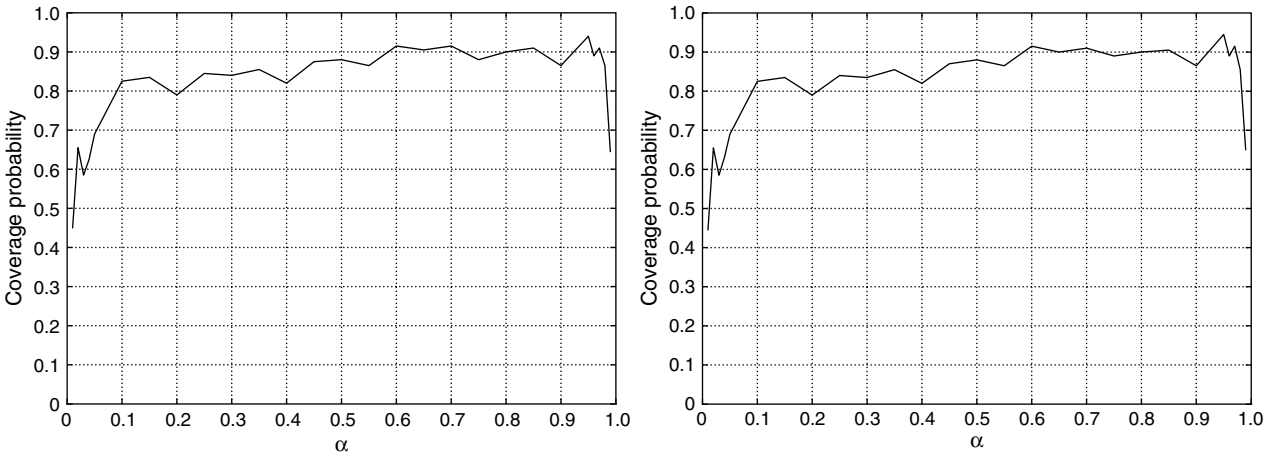


Figure 8. Coverage probabilities of the 90% confidence intervals for the $M/M/1$ queueing model.



$\alpha \geqslant 0.90$, the square roots of the MSEs are less than 10% of the actual sensitivities. Therefore, the estimators can appropriately estimate the steady-state quantile sensitivities for the $M/M/1$ queue. In Figure 8, we report the coverage probabilities of the 90% confidence intervals. When α is not close to zero or one, the coverage probabilities are close to 0.9. When α is close to zero or one, it appears that more within-batch samples are needed to keep the biases low.

9. Conclusions

In this paper, we study the estimation of quantile sensitivities. We first show that the quantile sensitivity can be written in the form of a conditional expectation. Based on this conditional-expectation form, we propose an estimator that is consistent and follows a central limit theorem for the i.i.d. data. The numerical results show that the estimator works well for the test problems. Even though the theory is established for i.i.d. data, the numerical experiments on the steady-state behavior of the $M/M/1$ queue show that our estimator also appears to work for dependent data.

Appendix

In this appendix, we prove that $S_{mk}^2/\sigma_m^2 \xrightarrow{P} 1$ as $n \to \infty$ if the conditions of Theorem 5 hold for $\gamma = 2$.

Proof. Note that

$$S_{mk}^2 = \frac{1}{k-1}\sum_{i=1}^{k}(D_{mi} - \bar{D}_{mk})^2 = \frac{k}{k-1}\left\{\frac{1}{k}\sum_{i=1}^{k}[D_{mi} - \mathrm{E}(D_m)]^2 - [\bar{D}_{mk} - \mathrm{E}(D_m)]^2\right\}.$$

By Equation (14) and the continuous mapping theorem,

$$[\bar{D}_{mk} - \mathrm{E}(D_m)]^2 \xrightarrow{P} 0.$$

Then, it suffices to prove that, as $n \to \infty$,

$$\frac{1}{k}\sum_{i=1}^{k}[D_{mi} - \mathrm{E}(D_m)]^2 - \sigma_m^2 \xrightarrow{P} 0. \tag{22}$$

Note that

$$\mathrm{E}\left\{\frac{1}{k}\sum_{i=1}^{k}[D_{mi} - \mathrm{E}(D_m)]^2\right\} = \mathrm{Var}(D_m) = \sigma_m^2$$

and

$$\mathrm{Var}\left\{\frac{1}{k}\sum_{i=1}^{k}[D_{mi} - \mathrm{E}(D_m)]^2\right\} = \frac{1}{k}\mathrm{Var}\{[D_m - \mathrm{E}(D_m)]^2\} \leqslant \frac{1}{k}\mathrm{E}\{[D_m - \mathrm{E}(D_m)]^4\}.$$

Note that $\sup_m \mathrm{E}(D_m^4) < \infty$ implies that $\sup_m \mathrm{E}\{[D_m - \mathrm{E}(D_m)]^4\} < \infty$. Then, by Cheybeshev's inequality, as $k \to \infty$,

$$\frac{1}{k}\sum_{i=1}^{k}[D_{mi} - \mathrm{E}(D_m)]^2 - \sigma_m^2 \xrightarrow{P} 0.$$

This concludes the proof. □

Acknowledgments

The author is grateful to HKUST postgraduate student Guangwu Liu for conducting the production-inventory and $M/M/1$ experiments reported in §§7.2 and 8 and for useful discussions that led to a simpler proof of Theorem 1. The author thanks the area editor, the associate editor, and two anonymous referees for their insightful comments that improved the paper in numerous ways. This research was partially supported by Hong Kong Research Grants Council grants CERG 613305 and 613706.

References

Austin, P., M. Schull. 2003. Quantile regression: A statistical tool for out-of-patient research. *Acad. Emergency Medicine* **10** 789–797.

Avramidis, A. N., J. R. Wilson. 1998. Correlation-induction techniques for estimating quantiles in simulation experiments. *Oper. Res.* **46** 574–591.

Bahadur, R. R. 1966. A note on quantiles in large samples. *Ann. Math. Statist.* **37** 577–580.

Billingsley, P. 1995. *Probability and Measure*, 3rd ed. Wiley, New York.

Birge, J. R., F. Louveaux. 1997. *Introduction to Stochastic Programming*. Springer-Verlag, New York.

Broadie, M., P. Glasserman. 1996. Estimating security price derivatives using simulation. *Management Sci.* **42** 269–285.

David, H. 1981. *Order Statistics*, 2nd ed. Wiley, New York.

Duffie, D., J. Pan. 1997. An overview of value at risk. *J. Derivatives* **4** 7–49.

Durret, R. 1996. *Probability: Theory and Examples*, 2nd ed. Duxury Press, Belmont, MA.

Fox, B. L., P. W. Glynn. 1989. Replication schemes for limiting expectations. *Probab. Engrg. Inform. Sci.* **3** 299–318.

Fu, M. C. 2006. Gradient estimation. S. G. Henderson, B. L. Nelson, eds. *Handbooks in Operations Research and Management Science: Simulation*, Chapter 19. Elsevier, Amsterdam, 575–616.

Fu, M. C., J. Q. Hu. 1997. *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*. Kluwer Academic Publishers, Norwell, MA.

Glasserman, P. 1991. *Gradient Estimation via Perturbation Analysis*. Kluwer Academic Publishers, Norwell, MA.

Glasserman, P., S. Tayur. 1995. Sensitivity analysis for base-stock levels in multiechelon production-inventory systems. *Management Sci.* **41** 263–281.

Glasserman, P., P. Heidelberger, P. Shahabuddin. 2000. Variance reduction techniques for estimating value-at-risk. *Management Sci.* **46** 1349–1364.

Glasserman, P., P. Heidelberger, P. Shahabuddin. 2002. Portfolio value-at-risk with heavy-tailed risk factors. *Math. Finance* **12** 239–269.

Glynn, P. W. 1990. Likelihood ratio gradient estimation for stochastic systems. *Comm. ACM* **33** 75–84.

Glynn, P. W. 1996. Importance sampling for Monte Carlo estimation of quantiles. *Proc. Second Internat. Workshop Math. Methods in Stochastic Simulation and Experimental Design*. St. Petersburg, FL, 180–185.

Goldenberg, D. K., L. Qiu, H. Xie, Y. R. Yang, Y. Zhang. 2004. Optimizing cost and performance for multihoming. SIGCOMM, Portland, OR, 79–92.

Heidelberger, P., P. A. W. Lewis. 1984. Quantile estimation in dependent sequences. *Oper. Res.* **32** 185–209.

Hesterberg, T. C., B. L. Nelson. 1998. Control variates for probability and quantile estimation. *Management Sci.* **44** 1295–1312.

Ho, Y. C., X.-R. Cao. 1983. Optimization and perturbation analysis of queueing networks. *J. Optim. Theory Appl.* **40** 559–582.

Hong, L. J., X. Qi. 2007. Stochastic linear programming and value-at-risk: An efficient Monte-Carlo approach. Technical report, Department of Industrial Engineering and Logistics Management, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong. http://www.ielm.ust.hk/dfaculty/hongl.

Hsu, J. C., B. L. Nelson. 1990. Control variates for quantile estimation. *Management Sci.* **36** 835–851.

Jin, X., M. C. Fu, X. Xiong. 2003. Probabilistic error bounds for simulation quantile estimators. *Management Sci.* **49** 230–246.

Law, A. M., W. D. Kelton. 2000. *Simulation Modeling and Analysis*, 3rd ed. McGraw-Hill, Singapore.

Marsden, J. E., M. J. Hoffman. 1993. *Elementary Classical Analysis*, 2nd ed. Freeman and Company, New York.

Reiman, M., A. Weiss. 1989. Sensitivity analysis for simulations via likelihood ratios. *Oper. Res.* **37** 830–844.

Ross, S. M. 1996. *Stochastic Processes*, 2nd ed. Wiley, New York.

Rubinstein, R. Y., A. Shapiro. 1993. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization*. Wiley, New York.

Sen, P. K. 1972. On the Bahadur's representation of sample quantiles for sequences of ϕ-mixing random variables. *J. Multivariate Anal.* **2** 77–95.

Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Suri, R., M. Zazanis. 1988. Perturbation analysis gives strongly consistent sensitivity estimates for the M/G/1 queue. *Management Sci.* **34** 39–64.