CONFIDENCE INTERVALS AND REGIONS FOR QUANTILES USING CONDITIONAL MONTE CARLO AND GENERALIZED LIKELIHOOD RATIOS

Lei Lei

Economics and Business Administration
Chongqing University

Chongqing, 400044, China

Christos Alexopoulos

H. Milton Stewart School
of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0205, USA

Yijie Peng

Guanghua School of Management
Peking University
Beijing, 100871, China

James R. Wilson

Edward P. Fitts Department
of Industrial and Systems Engineering
North Carolina State University
Raleigh, NC 27695–7906, USA

ABSTRACT

This article develops confidence intervals (CIs) and confidence regions (CRs) for quantiles based on independent realizations of a simulation response. The methodology uses a combination of conditional Monte Carlo (CMC) and the generalized likelihood ratio (GLR) method. While batching and sectioning methods partition the sample into nonoverlapping batches, and construct CIs and CRs by estimating the asymptotic variance using sample quantiles from each batch, the proposed techniques directly estimate the underlying probability density function of the response. Numerical results show that the CIs constructed by applying CMC, GLR, and sectioning lead to comparable coverage results, which are closer to the targets compared with batching alone for relatively small samples; and the coverage rates of the CRs constructed by applying CMC and GLR are closer to the targets than both sectioning and batching when the sample size is relatively small and the number of probability levels is relatively large.

1 INTRODUCTION

For a random variable Y with the cumulative distribution function (c.d.f.) $F_Y(\cdot)$, the p-quantile $(0 < p < 1)$ is defined as $y_p = F_Y^{-1}(p) \equiv \inf\{y : F_Y(y) \geq p\}$; and if $F_Y(y)$ is continuous at each $y \in \mathbb{R}$, then $F_Y(y_p) = p$ for each $p \in (0, 1)$. Quantiles, also known as values-at-risk, are used as benchmarks in financial risk management (Jorion 2001), e.g., to regulate capital sufficiency of banks for sustaining losses from their trading activities (Glasserman 2004). Quantiles are also used as performance measures in service systems (Gélinas et al. 1995, Seila 1982) as well as for safety and uncertainty analysis of nuclear power plants. For instance, the U.S. Nuclear Regulatory Commission uses the 95/95 criterion, which requires plant licensees to verify, with 95% confidence, that the 0.95-quantiles of certain performance measures lie below mandated thresholds (U.S. Nuclear Regulatory Commission 2011).

If $\{Y_i : i = 1, \ldots, n\}$ is a sequence of independent and identically distributed (i.i.d.) simulation responses, then y_p can be estimated by $\widehat{y}_p(n) \equiv Y_{(\lceil np \rceil)}$, where $Y_{(1)} \leq \cdots \leq Y_{(n)}$ are the respective order statistics and $\lceil \cdot \rceil$ denotes the ceiling function. If the c.d.f. $F_Y(\cdot)$ is twice differentiable at y_p so that the associated probability density function (p.d.f.) $f_Y(\cdot)$ is differentiable and $f_Y(y_p) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y)\big|_{y=y_p} = F_Y'(y_p) > 0$, then $\widehat{y}_p(n)$ is a consistent estimator and satisfies the following central limit theorem (CLT):

$$n^{1/2}\big(\widehat{y}_p(n) - y_p\big) \Rightarrow \sigma N(0,1) \quad \text{as } n \to \infty, \tag{1}$$

where $\sigma^2 \equiv p(1-p)/f_Y^2(y_p)$ is the asymptotic variance, $N(0,1)$ is the standard normal distribution, and "$\Rightarrow$" denotes convergence in distribution (Serfling 1980, Corollary B, p. 77).

For given $b \geq 2$, the batching methodology forms b nonoverlapping batches of simulation responses, each of size m ($n = bm$), and computes the sample quantiles $\widehat{y}_p(j,m)$ ($j = 1, \ldots, b$) from each batch. The quantile y_p is estimated by the sample average $\overline{\widehat{y}}_p(b,m) = \frac{1}{b}\sum_{j=1}^{b} \widehat{y}_p(j,m)$, and the asymptotic variance is estimated by m times the sample variance $S^2(b,m) \equiv \frac{1}{b-1}\sum_{j=1}^{b}\big[\,\widehat{y}_p(j,m) - \overline{\widehat{y}}_p(b,m)\,\big]^2$. The $100(1-\alpha)\%$ CI for y_p based on the batching method (BM) is

$$\mathscr{C}_{\mathrm{BM}}(m,\alpha) = \Big[\overline{\widehat{y}}_p(b,m) - t_{b-1,1-\alpha/2}S(b,m)\big/\sqrt{b},\ \ \overline{\widehat{y}}_p(b,m) + t_{b-1,1-\alpha/2}S(b,m)\big/\sqrt{b}\Big], \tag{2}$$

where $t_{\nu,\beta}$ denotes the β-quantile of Student's t distribution with ν degrees of freedom. The CLT (1) implies that as $m \to \infty$, the CI (2) is asymptotically valid—i.e., $\lim_{m\to\infty}\Pr\{\,y_p \in \mathscr{C}_{\mathrm{BM}}(m,\alpha)\,\} = 1-\alpha$. The method of sectioning improves upon the batching method by replacing the average batch quantile estimator $\overline{\widehat{y}}_p(b,m)$ in Equation (2) with the overall quantile estimator $\widehat{y}_p(n)$ (Asmussen and Glynn 2007, Nakayama 2014, Dong and Nakayama 2017).

Our approach for constructing a CI for y_p involves direct estimation of the unknown $f_Y(y_p) = F_Y'(y_p)$ based on the fact that the c.d.f. can be expressed as the expectation $F_Y(y_p) = \mathrm{E}[\mathbf{1}(Y \leq y_p)]$, where $\mathbf{1}(Y \leq y_p)$ is the indicator random variable that takes the value 1 when the condition $Y \leq y_p$ is true, and 0 otherwise. Classical infinitesimal perturbation analysis (IPA) and the likelihood ratio (LR) method cannot be applied to estimate $f_Y(y)$ because the random variable $\mathbf{1}(Y \leq y)$ is never a continuous function of y, and the derivative $\frac{\mathrm{d}}{\mathrm{d}y}\mathrm{E}[\mathbf{1}(Y \leq y)]$ is taken with respect to the quantity y, which is not a parameter of the density function $f_Y(y)$ (Peng et al. 2020). Although methods based on finite-differences (FD) and kernel density estimation have been used to construct CIs for y_p (Nakayama 2014), the latter two methods induce bias and require the choice of certain tuning parameters. Moreover, the FD method requires extra simulations and typically leads to large variance. Unlike the FD and kernel methods that rely solely on output-sample information, the conditional Monte Carlo (CMC) method (L'Ecuyer et al. 2019) and the generalized likelihood ratio (GLR) method (Peng et al. 2020) utilize direct information from the underlying simulation model, and yield unbiased density estimators.

Simultaneous estimation of multiple quantiles corresponding to a vector $\boldsymbol{p} \equiv (p_1, \ldots, p_d)$ of probabilities requires the construction of a confidence region. Specifically, we want to construct a region $\mathscr{R}(n,\alpha) \subset \mathbb{R}^d$ such that $\lim_{m\to\infty}\Pr[(y_{p_1}, \ldots, y_{p_d}) \in \mathscr{R}(n,\alpha)] = 1-\alpha$ for given confidence coefficient $\alpha \in (0,1)$ and batch count $b \geq 2$. Ming-hua and Glynn (2002) use a combination of batching and sectioning to construct CRs for stochastic approximation algorithms; but to the best of our knowledge, there is no previous research studying confidence regions for quantiles. In this paper, we discuss the formation of CRs for the four aforementioned methods, batching, sectioning, GLR, and CMC. We will show that CMC and GLR in CRs construction perform better than batching and sectioning methods when d is large and the sample size is small. The performance of batching and sectioning is sensitive to the number of batches, whilst this issue does not affect CMC and GLR. On the other hand, CMC and GLR methods need more information about the structure of the problem and stronger conditions compared to batching and sectioning methods.

The rest of the paper is organized as follows. Section 2 describes the quantile estimation problem and an ideal CI for the conventional quantile point estimator. This result lies at the basis of the CI construction we develop by using CMC and GLRs. Section 2 also discusses CMC and GLR estimators for the density function, which are required for construction of CIs by those methods. Section 3 describes the derivation of CRs for each method under study. Section 4 performs a preliminary evaluation of the four methods based on two examples. The results indicate that the estimated coverages of all the CIs and CRs are close to the target coverage rate when the sample size is large, demonstrating their asymptotic validity; while the

CMC, GLRs and sectioning methods outperform classical batching method when the sample size is small with regard to both construction of CIs and CRs. Finally, Section 5 offers some concluding remarks.

2 CONSTRUCTION OF CONFIDENCE INTERVALS

Assume that the output random variable Y can be expressed as

$$Y = g(X_1, \ldots, X_s)$$

for a given, finite number of inputs s so that $g : \mathbb{R}^s \to \mathbb{R}$; and $X_1, \ldots, X_s$ are continuous random variables so that the random vector $X \equiv (X_1, \ldots, X_s)$ has the joint p.d.f. $f(x)$ for all $x = (x_1, \ldots, x_s) \in \mathbb{R}^s$. Let z_β denote the β-quantile of the $N(0,1)$ distribution. Based on the CLT (1), we see that the ideal $100(1-\alpha)\%$ CI for y_p,

$$\widehat{y}_p(n) \pm z_{1-\alpha/2}\frac{\sqrt{p(1-p)}}{f_Y(y_p)\sqrt{n}}, \tag{3}$$

is asymptotically valid as $n \to \infty$. Of course the latter CI (3) cannot generally be applied in practice because the p.d.f. $f_Y(\cdot)$ is usually unknown.

One way to construct a CI for y_p is to consistently estimate the unknown constant $f_Y(y_p)$. The CI in Equation (2) bypasses this problem by "cancelling" the asymptotic variance $\sigma^2 = p(1-p)/f_Y^2(y_p)$. The literature contains a variety of methods based on finite differences and kernel density estimation (Chu and Nakayama 2012, Nakayama 2014); but both methodologies suffer from trade-offs between bias and variance, i.e., a small perturbation or bandwidth size reduces bias while increasing variance. We propose two unbiased estimators, i.e., CMC and GLR, to estimate $f_Y(y_p)$, and then we construct asymptotically valid CI estimators for the selected quantiles.

Since the p.d.f. $f_Y(y)$ is the derivative of the c.d.f. $F_Y(y)$, a natural approach to estimate $f_Y(y)$ is to express the c.d.f. as the expected value of an indicator random variable:

$$F_Y(y) = \mathrm{E}[\mathbf{1}\{Y \le y\}] = \mathrm{E}[\mathbf{1}\{g(X_1, \ldots, X_s) - y \le 0\}]. \tag{4}$$

For each $y \in \mathbb{R}$ and $\epsilon > 0$, we let $A_y^\epsilon \equiv \{x \in \mathbb{R}^s : y - \epsilon \le g(x) \le y + \epsilon\}$. We introduce the following regularity conditions to derive the GLR estimator for density estimation.

- **Regularity Condition (A.1)** Suppose the components of the input random vector X are independent, i.e., $f(x) = \prod_{i=1}^s f_i(x_i)$, where $f_i(x_i)$ is the marginal p.d.f. of X_i, the joint p.d.f. $f(x)$ is differentiable, and the response function $g(x)$ is twice differentiable at each $x \in \mathbb{R}^s$.
- **Regularity Condition (A.2)** The following uniform convergence condition holds:

  $$\lim_{\epsilon \to 0} \sup_{y \in \mathbb{R}} v(A_y^\epsilon) = 0,$$

  where v denotes Lebesgue measure on $\mathbb{R}$.
- **Regularity Condition (A.3)** The following integrability conditions hold: there exist functions $v_l(\cdot)$, $l = 1, ..., s$, such that for each $i \in \{1, \ldots, s\}$ and $x \in \mathbb{R}^s$, we have $\left|(\partial g(x)/\partial x_i)^{-1}\right| \le \prod_{l=1}^s v_l(x_l)$, and

  $$\lim_{x_i \to \pm\infty} v_i(x_i) f_i(x_i) = 0, \qquad \int_{\mathbb{R}} v_l(x_l) f_l(x_l)\, \mathrm{d}x_l < \infty, \quad l \ne i;$$

  in addition,

  $$\int_{x \in \mathbb{R}^s} \left|\Psi_i(x)\right| f(x)\, \mathrm{d}x < \infty,$$

  where $\Psi_i(x)$ is defined in Equation (6) below.

Under Regularity Conditions (A.1)–(A.3), Peng et al. (2020) showed that

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathrm{E}[\mathbf{1}(Y \le y)] = \mathrm{E}[\mathbf{1}(Y \le y) \cdot \Psi_i], \tag{5}$$

where

$$\Psi_i = \Psi_i(X) = \left[\left(\frac{\partial g(x)}{\partial x_i}\right)^{-1} \frac{\partial \log f_X(x)}{\partial x_i} - \left(\frac{\partial g(x)}{\partial x_i}\right)^{-2}\left(\frac{\partial^2 g(x)}{\partial x_i^2}\right)\right]\Bigg|_{x=X}. \tag{6}$$

We introduce the following regularity condition to derive the CMC estimator for density estimation.

- **Regularity Condition (A.4)** There exists a sigma-field $\mathscr{F}$ such that for all realizations of $\mathscr{F}$, the function $F_Y(y|\mathscr{F}) \equiv \Pr\{Y \le y|\mathscr{F}\}$ is continuous at each $y \in \mathbb{R}$; also the function $F_Y(y|\mathscr{F})$ is differentiable at each $y \in \mathbb{R} \setminus D(\mathscr{F})$, where $D(\mathscr{F}) \subset \mathbb{R}$ is a denumerable set of points. Moreover, there is a random variable Γ defined on the same probability space as $F_Y(y|\mathscr{F})$ such that $\mathrm{E}[\Gamma^2] \le J$ for some constant $J < \infty$, and $\sup\left\{F'_Y(y|\mathscr{F}) \equiv \frac{\mathrm{d}}{\mathrm{d}y}F_Y(y|\mathscr{F}) : y \in \mathbb{R} \setminus D(\mathscr{F})\right\} \le \Gamma$ at every point in the underlying sample space.

Under Regularity Condition (A.4), L'Ecuyer et al. (2019) proved that

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathrm{E}[\mathbf{1}(Y \le y)] = \mathrm{E}[F'_Y(y|\mathscr{F})] \text{ for each } y \in \mathbb{R} \setminus D(\mathscr{F}). \tag{7}$$

Equation (7) is the basis for the CMC estimator of $f_Y(y)$ at each $y \in \mathbb{R}^s \setminus D(\mathscr{F})$; and in practice we have to appropriately choose $\mathscr{F}$ to condition on. For example, if for some $l \in \{1, \ldots, s\}$, the sigma-field defined by $\mathscr{F}_l = (X_1, \cdots, X_{l-1}, X_{l+1}, \cdots, X_s)$ satisfies Regularity Condition (A.4) and the function

$$g_l(x_l|\mathscr{F}_l) \equiv g(X_1, \ldots, X_{l-1}, x_l, X_{l+1}, \ldots, X_s) \text{ for each } x_l \in \mathbb{R}$$

is strictly increasing in x_l for all possible realizations of $\mathscr{F}_l$, then the inverse function $g_l^{-1}(y|\mathscr{F}_l)$ exists at each point in the underlying sample space; and we have

$$F_Y(y|\mathscr{F}) = \Pr\left\{g_l(X_l|\mathscr{F}_l) \le y \,\middle|\, \mathscr{F}_l\right\} = \Pr\left\{X_l \le g_l^{-1}(y|\mathscr{F}_l) \,\middle|\, \mathscr{F}_l\right\} = F_l[\,g_l^{-1}(y|\mathscr{F}_l)\,]\,. \tag{8}$$

From Equation (8) and the change-of-variables formula (Rudin 1964, Theorem 6.33), we have

$$f_Y(y) = \mathrm{E}[F'_Y(y|\mathscr{F})] = \mathrm{E}\left\{f_l\left[\,g_l^{-1}(y|\mathscr{F}_l)\,\right] \cdot \left|\frac{\mathrm{d}}{\mathrm{d}y}g_l^{-1}(y|\mathscr{F}_l)\right|\right\},$$

where f_l is the density function of X_l. Clearly the regularity conditions required by GLR and CMC are stronger than those of batching and sectioning.

Suppose the c.d.f. $F_Y(\cdot)$ and p.d.f. $f_Y(\cdot)$ are differentiable at y_p and $f_Y(y_p) > 0$. With Equations (5) and (7) and a consistent point estimate of y_p, we obtain the GLR and CMC estimators for $f_Y(y_p)$ as follows:

$$\widehat{f}_{\mathrm{GLR},n}\left(\widehat{y}_p\right) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left(Y^{(j)} \le \widehat{y}_p(n)\right) \cdot \Psi_i^{(j)} \tag{9}$$

where the pairs $\left(Y^{(1)}, \Psi_i^{(1)}\right), \ldots, \left(Y^{(n)}, \Psi_i^{(n)}\right)$ are n independent realizations of (Y, Ψ_i), and

$$\widehat{f}_{\mathrm{CMC},n}\left(\widehat{y}_p\right) = \frac{1}{n}\sum_{j=1}^{n} F'\left(\widehat{y}_p \,\middle|\, \mathscr{F}^{(j)}\right) \tag{10}$$

where $\mathscr{F}^{(1)}, \ldots, \mathscr{F}^{(n)}$ are n independent realizations of $\mathscr{F}$. Under an additional moment condition for Ψ_i in Peng et al. (2017), it can be shown that $\widehat{f}_{\mathrm{GLR},n}(\widehat{y}_p) \to f_Y(y_p)$ almost surely (a.s.); further $\widehat{f}_{\mathrm{CMC},n}(\widehat{y}_p) \to f_Y(y_p)$ a.s. when $n \to \infty$.

Then we have the estimated $100(1-\alpha)\%$ CIs

$$\mathscr{C}_{\mathrm{GLR}}(n,\alpha) = \left[\widehat{y}_p(n) - z_{1-\alpha/2}\frac{\sqrt{p(1-p)}}{\widehat{f}_{\mathrm{GLR},\,n}(\widehat{y}_p)\sqrt{n}},\ \widehat{y}_p(n) + z_{1-\alpha/2}\frac{\sqrt{p(1-p)}}{\widehat{f}_{\mathrm{GLR},n}(\widehat{y}_p)\sqrt{n}} \right]$$

based on the GLR method and

$$\mathscr{C}_{\mathrm{CMC}}(n,\alpha) = \left[\widehat{y}_p(n) - z_{1-\alpha/2}\frac{\sqrt{p(1-p)}}{\widehat{f}_{\mathrm{CMC},\,n}(\widehat{y}_p)\sqrt{n}},\ \widehat{y}_p(n) + z_{1-\alpha/2}\frac{\sqrt{p(1-p)}}{\widehat{f}_{\mathrm{CMC},n}(\widehat{y}_p)\sqrt{n}} \right]$$

based on the CMC method, respectively.

3 CONSTRUCTION OF CONFIDENCE REGIONS

In the preceding section, we discussed CI construction for the point-wise quantile at one probability level p. We may also be interested in estimating not only the quantile at one probability level but also simultaneously estimating a vector of quantiles at probability levels $p_1, \ldots, p_d$, in which case a CR is needed to measure the accuracy on the vector of quantile estimates. Specifically, we want to construct a region $\mathscr{R}(n,\alpha) \subset \mathbb{R}^d$ such that

$$\lim_{n\to\infty} \Pr\left[\left(y_{p_1}, \ldots, y_{p_d} \right) \in \mathscr{R}(n,\alpha) \right] = 1-\alpha.$$

For quantile estimators $\widehat{y}_{\boldsymbol{p}}(n) = \left(\widehat{y}_{p_1}(n), \ldots, \widehat{y}_{p_d}(n)\right)$, where $\widehat{y}_{p_i}(n) = Y_{(\lceil np_i \rceil)},\ 1 \le i \le d$, we have the following multivariate analogue of the CLT (1).

Theorem 1 (Serfling 1980, §2.3.3, Theorem B) Let $0 < p_1 < p_2 < \cdots < p_d < 1$. Suppose that $F_Y(y)$ has a density $f_Y(y)$ in neighborhoods of $y_{p_1}, \ldots, y_{p_d}$ and that $f_Y(y)$ is positive and continuous at $y_{p_1}, \ldots, y_{p_d}$. Then $\sqrt{n}\left(\widehat{y}_{\boldsymbol{p}}(n) - y_{\boldsymbol{p}}\right)$ converges in distribution to a multivariate-normal distribution, i.e.,

$$\sqrt{n}\left(\widehat{y}_{\boldsymbol{p}}(n) - y_{\boldsymbol{p}}\right) \Rightarrow N_d(0,\Sigma),$$

as $n \to \infty$, where $N_d(0,\Sigma)$ is a normal d-variate distribution with mean 0 and covariance matrix Σ defined by

$$\Sigma = \begin{pmatrix} \sigma(p_1,p_1) & \sigma(p_1,p_2) & \cdots & \sigma(p_1,p_d) \\ \sigma(p_1,p_2) & \sigma(p_2,p_2) & \cdots & \sigma(p_2,p_d) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma(p_1,p_d) & \sigma(p_2,p_d) & \cdots & \sigma(p_d,p_d) \end{pmatrix}.$$

with

$$\sigma(p_i,p_j) = \frac{\min(p_i,p_j) - p_i p_j}{f_Y(y_{p_i}) f_Y(y_{p_j})},\quad 1 \le i,j \le d. \quad \blacktriangleleft \tag{11}$$

Therefore, the key to constructing an asymptotically valid CR for $y_{\boldsymbol{p}}$ is to estimate Σ consistently, and the form of $\sigma(p_i,p_j)$ in (11) shows that a natural estimation of Σ can be obtained by estimating $f_Y(y_{p_i})$ for $i = 1, \ldots, d$. As discussed in Section 2, we know that $\sigma(p_i,p_j)$ can be estimated consistently by

$$\widehat{\sigma}_{\mathrm{GLR},n}(p_i,p_j) = \frac{\min(p_i,p_j) - p_i p_j}{\widehat{f}_{\mathrm{GLR},n}(\widehat{y}_{p_i}(n))\widehat{f}_{\mathrm{GLR},n}(\widehat{y}_{p_j}(n))} \text{ or } \widehat{\sigma}_{\mathrm{CMC},n}(p_i,p_j) = \frac{\min(p_i,p_j) - p_i p_j}{\widehat{f}_{\mathrm{CMC},n}(\widehat{y}_{p_i}(n))\widehat{f}_{\mathrm{CMC},n}(\widehat{y}_{p_j}(n))}. \tag{12}$$

Under the assumptions of Theorem 1 and the Regularity Conditions (A.1)–(A.4) justifying unbiasedness of GLR and CMC, we have that for $\widehat{\Sigma}_{\mathrm{GLR},n}$ and $\widehat{\Sigma}_{\mathrm{CMC},n}$ defined by (12),

$$n\big(\widehat{y}_{\boldsymbol{p}}(n)-y_{\boldsymbol{p}}\big)\widehat{\Sigma}^{-1}_{\mathrm{GLR},n}\big(\widehat{y}_{\boldsymbol{p}}(n)-y_{\boldsymbol{p}}\big)^{\mathsf{T}} \Rightarrow \chi^2_d,$$

and

$$n\big(\widehat{y}_{\boldsymbol{p}}(n)-y_{\boldsymbol{p}}\big)\widehat{\Sigma}^{-1}_{\mathrm{CMC},n}\big(\widehat{y}_{\boldsymbol{p}}(n)-y_{\boldsymbol{p}}\big)^{\mathsf{T}} \Rightarrow \chi^2_d,$$

as $n \to \infty$, where χ^2_d is chi-square random variable with d degrees of freedom. Therefore, $100(1-\alpha)\%$ asymptotic confidence regions for $y_{\boldsymbol{p}}$ are given by

$$\mathscr{R}_{\mathrm{GLR}}(n,\alpha)=\left\{y\in\mathbb{R}^d : n\big(\widehat{y}_{\boldsymbol{p}}(n)-x\big)\widehat{\Sigma}^{-1}_{\mathrm{GLR},n}(n)\big(\widehat{y}_{\boldsymbol{p}}(n)-y\big)^{\mathsf{T}} \le \chi^2_{(d,\alpha)}\right\}$$

and

$$\mathscr{R}_{\mathrm{CMC}}(n,\alpha)=\left\{y\in\mathbb{R}^d : n\big(\widehat{y}_{\boldsymbol{p}}(n)-x\big)\widehat{\Sigma}^{-1}_{\mathrm{CMC},n}(n)\big(\widehat{y}_{\boldsymbol{p}}(n)-y\big)^{\mathsf{T}} \le \chi^2_{(d,\alpha)}\right\},$$

where $\chi^2_{(d,\alpha)}$ is the $(1-\alpha)$-quantile of chi-square distribution with d degrees of freedom.

Alternatively, we can construct confidence regions via the batching and sectioning methods. Specially, the n samples are split into b nonoverlapping batches, each consisting of m observations. We obtain the jth batch quantile estimator of $y_{\boldsymbol{p}}$ by

$$\widehat{y}_{\boldsymbol{p}}(j,m)=\big(\widehat{y}_{p_1}(j,m),\ldots,\widehat{y}_{p_d}(j,m)\big)=(Y_{j,\lceil mp_1\rceil},\ldots,Y_{j,\lceil mp_d\rceil}),$$

where $Y_{j,1}\le Y_{j,2}\le\cdots\le Y_{j,m}$ are the sample's order statistics within the jth batch for $j=1,\ldots,b$. Then

$$\overline{\widehat{y}}_{\boldsymbol{p}}(b,m)=\frac{1}{b}\left(\sum_{j=1}^{b}\widehat{y}_{p_1}(j,m),\ldots,\sum_{j=1}^{b}\widehat{y}_{p_d}(j,m)\right)$$

is the batching estimator of $y_{\boldsymbol{p}}$ and the sample covariance matrix $\widehat{\Sigma}_b$ is given by

$$\widehat{\Sigma}_b=\frac{1}{b-1}\sum_{i=j}^{b}\big(\widehat{y}_{\boldsymbol{p}}(j,m)-\overline{\widehat{y}}_{\boldsymbol{p}}(b,m)\big)^{\mathsf{T}}\big(\widehat{y}_{\boldsymbol{p}}(j,m)-\overline{\widehat{y}}_{\boldsymbol{p}}(b,m)\big).$$

For batching-based quantile estimation, we need the following result.

Proposition 1 If $b\ge 2$ is fixed and $d>b$, then

$$b\big(\overline{\widehat{y}}_{\boldsymbol{p}}(b,m)-y_{\boldsymbol{p}}\big)\widehat{\Sigma}_b^{-1}\big(\overline{\widehat{y}}_{\boldsymbol{p}}(b,m)-y_{\boldsymbol{p}}\big)^{\mathsf{T}} \Rightarrow \frac{d(b-1)}{(b-d)}F_{(d,b-d)} \text{ as } m\to\infty,$$

where $F_{(d,b-d)}$ has an F distribution with d and $b-d$ degrees of freedom. ◀

Proof For $j=1,\ldots,b$, let $Q_j=m^{1/2}\big[\widehat{y}_{\boldsymbol{p}}(j,m)-y_{\boldsymbol{p}}\big]$ so the $\{Q_j\}$ are i.i.d. For each j, Theorem 1 ensures that $Q_j \Rightarrow Z_j\sim N_d(0,\Sigma)$ as $m\to\infty$; hence the $\{Z_j\}$ constitute a random sample from $N_d(0,\Sigma)$. We take $Q=(Q_1,\ldots,Q_b)$; and we let $\overline{Q}_b=b^{-1}\sum_{j=1}^{b}Q_j$ and $S_Q=(b-1)^{-1}\sum_{j=1}^{b}\big(Q_j-\overline{Q}_b\big)^{\mathsf{T}}\big(Q_j-\overline{Q}_b\big)$ respectively denote the sample mean and sample covariance matrix of the $\{Q_j\}$. Similarly, we define Z, $\overline{Z}_b$, and S_Z from the $\{Z_j\}$. Next we observe that the mapping $\vartheta: Q\mapsto b\,\overline{Q}_b\,S_Q^{-1}\,\overline{Q}_b^{\mathsf{T}}$ is continuous at each point $Q\in\mathbb{R}^b$ such that $\det(S_Q)>0$. Because the $\{Z_j\}$ are i.i.d. $N_d(0,\Sigma)$ with

$$\det(\Sigma)=\frac{p_1\Big[\prod_{i=1}^{d-1}(p_{i+1}-p_i)\Big](1-p_d)}{\prod_{i=1}^{d}f_Y^2\big(y_{p_i}\big)}>0,$$

it follows that $\Pr\{\det(S_Z) > 0\} = 1$ by the theorem of Dykstra (1970). Thus by the continuous-mapping theorem (Whitt 2002, Theorem 3.4.3), we see that

$$b\big(\overline{\widehat{y}}_{\boldsymbol{p}}(b,m) - y_{\boldsymbol{p}}\big)\widehat{\Sigma}_b^{-1}\big(\overline{\widehat{y}}_{\boldsymbol{p}}(b,m) - y_{\boldsymbol{p}}\big)^{\mathsf{T}} = \vartheta(Q) \Rightarrow \vartheta(Z) = b\overline{Z}_b S_Z^{-1}\overline{Z}_b^{\mathsf{T}} \text{ as } m \to \infty;$$

and we have

$$b\overline{Z}_b S_Z^{-1}\overline{Z}_b^{\mathsf{T}} \sim \frac{d(b-1)}{(b-d)}F_{(d,b-d)}$$

by Anderson (2003, Corollary 5.2.1). ∎

Therefore, as $m \to \infty$, an asymptotically valid $100(1-\alpha)\%$ CR for $y_{\boldsymbol{p}}$ based on the batching method is given by

$$\mathscr{R}_{\mathrm{BM}}(n,\alpha) = \left\{y \in \mathbb{R}^d : b\big(\overline{\widehat{y}}_{\boldsymbol{p}}(b,m) - y\big)\widehat{\Sigma}_b^{-1}\big(\overline{\widehat{y}}_{\boldsymbol{p}}(b,m) - y\big)^{\mathsf{T}} \le \frac{d(b-1)}{(b-d)}F_{(d,b-d,\alpha)}\right\}, \tag{13}$$

where $F_{(d,b-d,\alpha)}$ is the $(1-\alpha)$-quantile of the F distribution with d and $b-d$ degrees of freedom.

For the sectioning method, the samples are also split into b batches, but the quantile estimator of $y_{\boldsymbol{p}}$ is obtained as

$$\widehat{y}_{\boldsymbol{p}}(n) = (Y_{(\lceil np_1\rceil)}, \ldots, Y_{(\lceil np_d\rceil)}),$$

and the sample covariance matrix is given by

$$\widetilde{\Sigma}_b = \frac{1}{b-1}\sum_{j=1}^{b}\big(\widehat{y}_{\boldsymbol{p}}(j,m) - \widehat{y}_{\boldsymbol{p}}(n)\big)^{\mathsf{T}}\big(\widehat{y}_{\boldsymbol{p}}(j,m) - \widehat{y}_{\boldsymbol{p}}(n)\big).$$

Then the approximate $100(1-\alpha)\%$ CR for $y_{\boldsymbol{p}}$ based on the sectioning method (SM) is given by

$$\mathscr{R}_{\mathrm{SM}}(n,\alpha) = \left\{y \in \mathbb{R}^d : b(\widehat{y}_{\boldsymbol{p}}(n) - y)\widetilde{\Sigma}_b^{-1}(\widehat{y}_{\boldsymbol{p}}(n) - y)^{\mathsf{T}} \le \frac{d(b-1)}{(b-d)}F_{(d,b-d,\alpha)}\right\}. \tag{14}$$

Although we have found that in practice the empirical coverage probability of the sectioning-based CR (14) is close to that of the batching-based CR (13), we have not been able to prove a convergence property for sectioning that is comparable to Proposition 1.

4 NUMERICAL RESULTS

In this section, we use two examples to test the performance of the BM, SM, CMC, and GLR methods with regard to CI construction (Example 1) and CR construction (Example 2). We estimate the coverage rate of $100(1-\alpha)\%$ CIs (or CRs) by the proportion of the constructed CIs (or CRs) that contain the true quantile (or vector of quantiles) from 10^4 independent experiments.

Example 1 Consider the simple example where $X_1 \sim N(0,1)$ and $X_2 \sim N(0,4)$ are independent, and $Y = X_1 + X_2$. We are interested in estimating the p-quantile y_p of Y and constructing CIs for different values of p via different methods (BM, SM, CMC, GLR). We know that $Y \sim N(0,5)$, so the true value of the p-quantile of Y is $y_p = \sqrt{5}\,z_p$.

We write the density $f_Y(y)$ as

$$\frac{\mathrm{d}}{\mathrm{d}y}\mathrm{E}\big[\mathbf{1}(Y \le y)\big] = \frac{\mathrm{d}}{\mathrm{d}y}\mathrm{E}\big[\mathbf{1}(X_1 + X_2 \le y)\big].$$

The GLR estimator (5) is not unique because we can choose either value of i. The derivatives in the estimator (5) are given by $\frac{\partial g(x)}{\partial x_i} = 1$, $\frac{\partial^2 g(x)}{\partial x_i^2} = 0$ for $i = 1, 2$, $\frac{\partial \log f_X(x)}{\partial x_1} = -x_1$, and $\frac{\partial \log f_X(x)}{\partial x_2} = -x_2/4$. We have two GLR estimators (namely, GLR_1 and GLR_2) for estimating the density f_Y:

$$\widehat{f}_{\text{GLR}_1,n}(\widehat{y}_p) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\big(Y^{(j)} \le \widehat{y}_p(n)\big) \cdot \big(-X_1^{(j)}\big)$$

and

$$\widehat{f}_{\text{GLR}_2,n}(\widehat{y}_p) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\big(Y^{(j)} \le \widehat{y}_p(n)\big) \cdot \big(-X_2^{(j)}/4\big)$$

where $X_i^{(1)}, \ldots, X_i^{(n)}$ are n realizations of X_i for $i = 1, 2$. From Peng et al. (2020), if there are l unbiased GLR estimators $\mathbf{1}(Y \le y)\Psi_{i_\ell}$, $\ell = 1, \ldots, l$, then any linear combination $\sum_{\ell=1}^{l} w_\ell \mathbf{1}(Y \le y)\Psi_{i_\ell}$ with nonnegative weights $\{w_\ell\}$ such that $\sum_{\ell=1}^{l} w_\ell = 1$ is also an unbiased estimator; and the optimal weights for minimizing the variance of the estimator can be obtained from

$$(w_1^*, \ldots, w_l^*) = \underset{w_1,\ldots,w_l}{\arg\min}\ \text{Var}\left[\sum_{\ell=1}^{l} w_\ell \mathbf{1}(Y \le y)\Psi_{i_\ell}\right] \quad \text{s.t.} \quad \sum_{\ell=1}^{l} w_\ell = 1,$$

which has an analytical form:

$$w_i^* = \frac{e_i^T \Sigma_l^{-1} e}{e^T \Sigma_l^{-1} e}, \quad i = 1, \ldots, l,$$

where $e = (1, \ldots, 1)^\intercal$, e_i is a l-dimensional unit vector in ith direction, i.e., $e_1 = (1, 0, 0, \ldots, 0)$, $e_2 = (0, 1, 0, 0, \ldots, 0)$, and Σ_l is the covariance matrix of $(\mathbf{1}(Y \le y)\Psi_{i_1}, \ldots, \mathbf{1}(Y \le y)\Psi_{i_l})$. After some algebra we obtain the optimal assignment $(w_1^*, w_2^*) = (0.2, 0.8)$. Hence the optimal GLR estimator (GLR^*) is

$$\widehat{f}_{\text{GLR}^*,n}(\widehat{y}_p) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}(Y^{(j)} \le \widehat{y}_p(n)) \cdot \big(-w_1^* X_1^{(j)} - w_2^* X_2^{(j)}/4\big).$$

For CMC in this example, we can use either $\mathscr{F} = X_1$ or $\mathscr{F} = X_2$. Assume that we choose $\mathscr{F} = X_2$, then $F(y|\mathscr{F}) = \mathbb{P}(Y \le y|X_2) = \mathbb{P}(X_1 \le y - X_2) = F_1(y - X_2)$ and $f_Y(y_p) = \text{E}[F_1'(y_p|X_2)] = \text{E}[f_1(y_p - X_2)]$. Therefore the density estimator at y_p is

$$\widehat{f}_{\text{CMC},n}(\widehat{y}_p) = \frac{1}{n}\sum_{j=1}^{n} f_1\big(\widehat{y}_p - X_2^{(j)}\big),$$

where F_1 and f_1 are the c.d.f. and the p.d.f. of the standard normal distribution, respectively.

Table 1 displays the coverage rates and average half-widths (AHWs) of the constructed CIs for the BM, SM, CMC, and GLR methods. For large n, the estimated coverages of all the CIs are close to the target coverage rate, demonstrating their asymptotic validity. GLR^* has a slight edge over GLR_1 and GLR_2 in both estimated CI coverage probability and AHW, especially when n is small. In general, SM, CMC, and GLR exhibit better estimated coverage than BM, especially for small n. Finally, CMC, GLR^*, and SM show comparable coverage in different cases.

Example 2 This example has the same setting as Example 1, but we are interested in the experimental evaluation of the CRs obtained by the four competitors (BM, SM, CMC, and GLR). The probabilities p_i are spaced uniformly: $p_i = i/(d+1)$, $i = 1, \ldots, d$. The density estimators of CMC and GLR at p_i are

Table 1: Coverage rates (with parenthesized average CI half-widths) for the BM, SM, GLR, and CMC methods based on Example 1.

n	$b = 16$		$b = 32$		CMC	GLR*	GLR_1	GLR_2
	BM	SM	BM	SM				
$p = 0.9$, $1 - \alpha = 0.9$								
2^{10}	0.885 (0.203)	0.899 (0.206)	0.641 (0.193)	0.897 (0.199)	0.901 (0.197)	0.897 (0.201)	0.882 (0.234)	0.892 (0.202)
2^{12}	0.899 (0.103)	0.904 (0.103)	0.878 (0.101)	0.904 (0.101)	0.899 (0.098)	0.898 (0.098)	0.897 (0.102)	0.897 (0.099)
2^{14}	0.901 (0.051)	0.903 (0.052)	0.8862 (0.050)	0.903 (0.050)	0.899 (0.049)	0.898 (0.049)	0.898 (0.049)	0.897 (0.049)
2^{16}	0.898 (0.026)	0.898 (0.026)	0.903 (0.025)	0.902 (0.025)	0.900 (0.024)	0.897 (0.024)	0.897 (0.024)	0.894 (0.024)
$p = 0.9$, $1 - \alpha = 0.95$								
2^{10}	0.935 (0.246)	0.949 (0.249)	0.740 (0.231)	0.946 (0.238)	0.949 (0.234)	0.942 (0.239)	0.925 (0.232)	0.938 (0.242)
2^{12}	0.950 (0.125)	0.950 (0.126)	0.936 (0.121)	0.949 (0.121)	0.950 (0.117)	0.950 (0.118)	0.943 (0.121)	0.950 (0.118)
2^{14}	0.948 (0.063)	0.949 (0.063)	0.943 (0.060)	0.951 (0.060)	0.948 (0.058)	0.950 (0.058)	0.949 (0.059)	0.950 (0.059)
2^{16}	0.946 (0.031)	0.946 (0.031)	0.946 (0.030)	0.947 (0.030)	0.951 (0.029)	0.951 (0.029)	0.952 (0.029)	0.953 (0.029)

the same as described in Example 1, and we only show the results corresponding to the optimal GLR estimator. The experimental results displayed in Table 2 indicate significant advantages for the CMC and GLR methods with regard to CR construction for quantiles compared to SM and BM, especially when the sample size is relatively small and the dimension of the probability vector is relatively large. Asmussen and Glynn (2007) suggest choosing $b \leq 30$ for both BM and SM in CI construction, but for CR construction, we have to choose b such that $b > d$. The results in Table 2 show cases where $d = 9$, $d = 19$, and $d = 49$. (Entries with "NA" indicate that the respective method is not applicable because $b \leq d$.) Clearly, for fixed relatively small sample sizes, the gap between the estimated coverage rates and nominal rates for BM- and SM-based CRs widens as d increases. It should be clear that, if d is large, b has to be significantly larger to achieve a high accuracy for the target coverage rates of CRs. Since the CMC and GLR methods are immune to this issue, the sample size needed by CMC and GLR to achieve the same accuracy would be much smaller than BM and SM for large d. Notice that the estimated coverage of CMC-based CRs when $n = 2^{12}$ is as accurate as SM-based CRS and better than BM-based CRs when $n = 2^{14}$ and $b = 64$.

5 Conclusion

In this article, we have proposed two new methods for constructing confidence intervals and confidence regions for quantiles in i.i.d. data. The techniques are based on the frameworks of generalized likelihood ratios (GLR) and conditional Monte Carlo (CMC), and their validity has been established under a set of sufficient conditions. Two numerical examples illustrated the potential of the proposed methods over classical counterparts based on batching and sectioning. Future work will focus on additional experimentation in more complex settings and potential extensions to stationary processes.

Table 2: Coverage rates for confidence regions based on the BM, SM, GLR methods and CMC based on Example 2.

n	$b = 16$		$b = 32$		$b = 64$		CMC	GLR*
	BM	SM	BM	SM	BM	SM		
$d = 9$, $1 - \alpha = 0.95$								
2^{12}	0.9299	0.9501	0.8754	0.9480	0.1612	0.9419	0.9521	0.9450
2^{14}	0.9427	0.9470	0.9323	0.9524	0.7637	0.9467	0.9499	0.9482
2^{16}	0.9541	0.9530	0.9476	0.9510	0.9142	0.9493	0.9507	0.9513
$d = 19$, $1 - \alpha = 0.95$								
2^{12}	NA	NA	0.5719	0.9472	0.0087	0.9338	0.9492	0.9461
2^{14}	NA	NA	0.8811	0.9509	0.5597	0.9529	0.9536	0.9483
2^{16}	NA	NA	0.9358	0.9526	0.8792	0.9516	0.9485	0.9513
$d = 49$, $1 - \alpha = 0.95$								
2^{12}	NA	NA	NA	NA	0	0.8955	0.9459	0.9296
2^{14}	NA	NA	NA	NA	0.4684	0.9522	0.9510	0.9434
2^{16}	NA	NA	NA	NA	0.6521	0.9492	0.9516	0.9495

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation of China (NSFC) under Grants 71901003, 71720107003, 71690232, 91846301, and 71790615.

REFERENCES

Anderson, T. W. 2003. *An Inroduction to Multivariate Statistical Analysis*. 3rd ed. Wiley-Interscience.

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*. New York: Springer Science+Business Media.

Chu, F., and M. K. Nakayama. 2012. "Confidence Intervals for Quantiles When Applying Variance-Reduction Techniques". *ACM Transactions on Modeling and Computer Simulation* 22(2):Article 10.

Dong, H., and M. K. Nakayama. 2017. "Quantile Estimation with Latin Hypercube Sampling". *Operations Research* 65(6):1678–1695.

Dykstra, R. L. 1970. "Establishing the Positive Definiteness of the Sample Covariance Matrix". *The Annals of Mathematical Statistics* 41(6):2153–2154.

Gélinas, R., A. Martel, and P. Lefrançois. 1995. "SOS: A Quantile Estimation Procedure for Dynamic Lot-Sizing Problems". *The Journal of the Operational Research Society* 46(11):1337–1351.

Glasserman, P. 2004. *Monte Carlo Methods in Financial Engineering*. New York: Springer-Verlag.

Jorion, P. 2001. *Value at Risk: The New Benchmark for Managing Financial Risk*. 2nd ed. New York: McGraw-Hill.

L'Ecuyer, P., F. Puchhammer, and A. B. Abdellah. 2019. "Monte Carlo and Quasi-Monte Carlo Density Estimation via Conditioning". https://arxiv.org/abs/1906.04607, accessed 20th March 2020.

Ming-hua, H., and P. W. Glynn. 2002. "Confidence Regions for Stochastic Approximation Algorithms". In *Proceedings of the 2002 Winter Simulation Conference*, edited by E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, 370–376. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Nakayama, M. K. 2014. "Confidence Intervals for Quantiles Using Sectioning When Applying Variance-Reduction Techniques". *ACM Transactions on Modeling and Computer Simulation* 24(4):Article 4.

Peng, Y., M. C. Fu, P. W. Glynn, and J. Hu. 2017. "On the Asymptotic Analysis of Quantile Sensitivity Estimation by Monte Carlo Simulation". In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 2336–2347. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Peng, Y., M. C. Fu, B. Heidergott, and H. Lam. 2020. "Maximum Likelihood Estimation by Monte Carlo Simulation: Towards Data-Driven Stochastic Modeling". *Operations Research* forthcoming.

Rudin, W. 1964. *Principles of Mathematical Analysis*. 2nd ed. New York: McGraw-Hill.

Seila, A. F. 1982. "A Batching Approach to Quantile Estimation in Regenerative Simulations". *Management Science* 28(5):573–581.

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

U.S. Nuclear Regulatory Commission 2011. "Applying Statistics". Technical Report NUREG-1475, Rev. 1, U.S. Nuclear Regulatory Commission, Washington, DC.

Whitt, W. 2002. *Stochastic-Process Limits: An Inroduction to Stochastic-Process Limits and Their Application to Queues*. New York: Springer-Verlag.

AUTHOR BIOGRAPHIES

LEI LEI is an assistant professor in Economics and Business Administration in Chongqing University, China. She holds a Ph.D. in Management Science from Fudan University, China. Her research interests include discrete-event stochastic systems, sensitivity analysis and simulation with applications towards supply chain management and financial engineering. Her email address is leilei312@cqu.edu.cn, and her Web page is http://ceba.cqu.edu.cn/info/2823/2587.htm.

CHRISTOS ALEXOPOULOS is a Professor in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Tech. His main research interests are in the areas of applied probability and computer simulation. He is an Associate Editor for *ACM Transactions on Modeling and Computer Simulation* and *Networks*. His e-mail address is christos@gatech.edu, and his Web page is http://www.isye.gatech.edu/~christos.

YIJIE PENG is an Assistant Professor in the Department of Management Science and Information Systems in Guanghua School of Management at Peking University, Beijing, China. His research interests include stochastic modeling and analysis, simulation optimization, machine learning, data analytics, and healthcare. He serves as an Associate Editor of the Asia-Pacific Journal of Operational Research. His email address is pengyijie@pku.edu.cn, and his Web page is http://www.gsm.pku.edu.cn/jsjjxq.jsp?urltype=tree.TreeTempUrl&wbtreeid=1141&user_id=pengyijie.

JAMES R. WILSON is a Professor Emeritus in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. His research interests concern modeling, analysis, and simulation of stochastic systems, especially as applied in healthcare, production, and quality systems engineering. He is a Fellow of INFORMS and IISE. His e-mail address is jwilson@ncsu.edu, and his Web page is http://www.ise.ncsu.edu/jwilson.