Gradient Estimation of Quantiles for Stationary Waiting Times

Bernd Heidergott\textsuperscript{1}, Warren Volk-Makarewicz\textsuperscript{1}, Felisa Vázquez-Abad\textsuperscript{2}

\textsuperscript{1}Department of Econometrics and Operations Research, Vrije Universiteit, Amsterdam

\textsuperscript{2}Department of Computer Science, City University, New York

WODES, 31st August 2010
Introduction

Abstract

- Problem: quantiles cannot be expressed as an expectation, so determining the sensitivity is not immediate.
  
  Methods for solving such problems has been recently developed.

What we will show:

- Quantile sensitivity estimation for stationary (continuous) waiting time sequence, \( W = [W_\theta(n)] \).

\[
q'_\alpha(\theta) := \frac{\partial}{\partial \theta} q_\alpha(\theta) = \frac{\partial}{\partial \theta} F_\infty(\theta) \frac{f_\infty(q_\alpha(\theta))}{f_\infty(q_\alpha(\theta))},
\]

for queuing systems: \( M(\theta)/M(\mu)/1, M(\theta)/G/1/T \).

- Structure: Development of expressions for numerator and denominator, Simulations.
Quantiles

• Ubiquitous. Provides a value, (statistically) ensuring that probability of event that is more extreme to be no more than $\alpha$.

• Bahadur (1966), $X\left(\left[\alpha n\right]\right) \overset{a.s.}{\to} q_\alpha(\theta)$, $X$ iid.

• Hesse (1990), Above result for a large class of stationary linear process $X(n) = \sum_{i=0}^{\infty} \delta(i)\epsilon(n-i)$, $\epsilon(i)$ are innovations.

• David & Nagaraja (2003), has a through review on order statistics.

• Koenker (2005), recent edition in the quantile regression literature.
Sensitivity of Quantiles

Effect of change in $\alpha$-level quantile, $q_\alpha(\theta)$, due to change in parameter $\theta$. Applications:

- QoS: Communication Network, $q_\alpha(\theta) \leq M$, $M$ maximal sojourn time of information packet.
- VaR, CVaR: Determining change in loss in banking, insurance, or investment decisions.

Review:

- Hong (2009), Hong and Liu (2009), Fu et al. (2009), Sensitivity estimation of quantiles, CVaR, via PA for static problems.
- Heidergott, Volk-Makarewicz (2009), Sensitivity estimation via MVD.
Quantile Sensitivity Expression: $G(\theta)/G/1$ Queue

- Continuous RV, $F_{\theta}(q_{\alpha}(\theta)) = \alpha$.

Quantile Sensitivity, $\alpha$-level, for $G(\theta)/G/1$ Queue.

$q'_{\alpha}(\theta) = -\frac{\partial}{\partial \theta} F_{\theta}^{\infty}(q_{\alpha}(\theta)) \frac{f_{\theta}^{\infty}(q_{\alpha}(\theta))}{f_{\theta}^{\infty}(q_{\alpha}(\theta))}$

$$= c_{\theta} E_{\theta} \left[ \sum_{n=1}^{\tau_{\theta}^{\pm}} \left( 1_{\{W_{\theta}^{+}(n,W_{\theta}(0)) \leq q_{\alpha}(\theta)\}} - 1_{\{W_{\theta}^{-}(n,W_{\theta}(0)) \leq q_{\alpha}(\theta)\}} \right) \right]$$

$$= E_{\theta} \left[ f_{A_{\theta}}(W_{\theta} + S - q_{\alpha}(\theta)) 1_{\{W_{\theta} + S - q_{\alpha}(\theta) > 0\}} \right]$$

provided that $\sup_{\theta \in \Theta} E_{\theta}[S(1) - A_{\theta}(1)] < 0$. 
Markov Chain Analysis

- DES $X_\theta = [X_\theta(n)]$: a Harris recurrent sequence with atom $a$.
- Governed by stochastic recursive sequence

$$X_\theta(n + 1) = h(X_\theta(n), Y_\theta(n + 1)),$$

where $h$ is a measurable function, $Y_\theta(n + 1)$ is a vector random variable independent from $X_\theta(n)$.

- $G/G/1$: Stochastic sequence, $[W_\theta(n)]$, waiting time of a packet/customer with atom 0.

$$Y_\theta(m) = (A_\theta(m), S(m)),$$

inter-arrival time between, and service time for a packet.

$m$ represents the $m^{th}$ 'customer' in the queue arriving from outside (and not necessarily admitted).
Markov Chain Analysis

- Markov kernel $P_\theta(s; \cdot)$;
  Lebesgue density $f_\theta(s; \cdot)$; Stationary distribution $\pi_\infty$, density $f_\infty$.
- Continuous RV: $f_\infty(x) = \mathbb{E}_\infty \left[ f_\theta(X_\theta; x) \right]$.
  Proof: via $\pi_\infty = P_\theta \pi_\infty$.
- $G/G/1$: More care is required.
  $W_\theta(n + 1) = \max(W_\theta(n) + S(n) - A_\theta(n + 1), 0)$.
  Condition on $W_\theta(n) = u$ then $S$, observing all conditions for $(W_\theta(n), W_\theta(n + 1))$.
- $F_\infty(x | u)$ for $W_\theta(n + 1) = x > 0$, $W_\theta(n) = u$

$$F_\infty(x | u) = \mathbb{E}_\theta \left[ F_{A_\theta}(u + S) - F_{A_\theta}(u + S - x) 1_{\{u+S-x>0\}} \right]$$

or

$$f_\infty(x) = \mathbb{E}_\theta \left[ f_{A_\theta}(W_\theta + S - x) 1_{\{W_\theta+S-x>0\}} \right].$$
Measure Valued Differentiation

Pflug (1996), Heidergott et al. (2010)

- Distributional approach to sensitivity analysis.
- Markov Kernel, $P_\theta(s, A)$, mapping $(S, S) \rightarrow (T, T)$.
- Following certain conditions, for $\theta \in \Theta \subset \mathbb{R}$.

Measure Valued Derivative

$$\frac{\partial}{\partial \theta} P_\theta(s, \cdot) = c_\theta(s)(P^+_{\theta}(s, \cdot) - P^-_{\theta}(s, \cdot))$$

where $c_\theta(s)P^\pm_{\theta}(s, \cdot) = \mu^\pm(s, \cdot)$.

Note that $P^\pm_{\theta}(s, \cdot)$ need not be independent.
MVD: Exponential Distribution

Given density, $A_\theta \sim \exp(\theta)$

$$f_\theta(x) = \theta e^{-\theta x}1(x > 0)$$

we obtain

$$\frac{\partial}{\partial \theta} f_\theta(x) = (1 - \theta x)e^{-\theta x}1(x > 0)$$

$$= \frac{1}{\theta} (\theta e^{-\theta x}1(x > 0) - \theta^2 xe^{-\theta x}1(x > 0))$$

- $A^+_\theta \sim \exp(\theta)$, $A^-_\theta \sim \gamma(2, \theta)$. 
Heidergott et. al. (2006)

- MVD of $[W_\theta(n)]$, construct two sequences, $[W_{\theta}^{\pm}(n)]$, called 'phantoms'.
- Given $W_\theta(0)$, transition to $W_{\theta}^{\pm}(1)$ governed by $P_{\theta}^{\pm}$:
  \[W_{\theta}^{\pm}(1) = \max ( W_{\theta}(0) + S(1) - A_{\theta}^{\pm}(1), 0 ) .\]
- Subsequent transitions for $[W_{\theta}^{\pm}(n)]_{n \geq 2}$ are according to $P_{\theta}$:
  \[n \geq 2, W_{\theta}^{\pm}(n) = \max ( W_{\theta}^{\pm}(n - 1) + S(n) - A_{\theta}(n), 0 ) .\]
- Stopping time, $\tau_{\theta}^{\pm}( W_{\theta}(0) )$, for phantom MVD processes is when both processes coalesce:
  \[\tau_{\theta}^{\pm}(s) = \inf \{ n \geq 1 : W_{\theta}^{\pm}(n) = 0, W_{\theta}(0) = s \} .\]
- In simulation, stopping time is the criteria to terminate MVD calculation.
Numerator of Quantile Sensitivity

- Value of MVD in DES is the cumulative difference in the two phantom processes.
- Performance function, $g(u) = \mathbf{1}_{\{u \leq x\}}$, $\mathbb{E}_{\theta}[g(X)] = F_{\theta}(x)$.

$$\frac{\partial}{\partial \theta} F_{\theta}(x) = c_\theta \mathbb{E}_{\theta} \left[ \sum_{n=1}^{\tau_{\theta}^\pm(W_\theta(0))} \left( \mathbf{1}_{\{W^+_\theta(n,W_\theta(0)) \leq x\}} - \mathbf{1}_{\{W^-_\theta(n,W_\theta(0)) \leq x\}} \right) \right]$$
Quantile Sensitivity Expression: $G(\theta)/G/1$ Queue

- Continuous RV, $F_\theta(q_\alpha(\theta)) = \alpha$.

Quantile Sensitivity, $\alpha$-level, for $G(\theta)/G/1$ Queue.

$$q'_\alpha(\theta) = -\frac{\frac{\partial}{\partial \theta} F_\theta^\infty(q_\alpha(\theta))}{f_\theta^\infty(q_\alpha(\theta))}$$

$$= -\frac{c_\theta \mathbb{E}_\theta \left[ \sum_{n=1}^{\tau^\pm_\theta(W_\theta(0))} \left( 1\{W_\theta^+(n, W_\theta(0)) \leq q_\alpha(\theta)\} - 1\{W_\theta^-(n, W_\theta(0)) \leq q_\alpha(\theta)\} \right) \right]}{\mathbb{E}_\theta \left[ f_A_\theta \left( W_\theta + S - q_\alpha(\theta) \right) 1\{W_\theta + S - q_\alpha(\theta) > 0\} \right]}$$

provided that $\sup_{\theta \in \Theta} \mathbb{E}_\theta[S(1) - A_\theta(1)] < 0$. 
Quantile Sensitivity Estimator: $G(\theta)/G/1$ queue

- Computational budget: $N = km$ stationary waiting times.

$$\hat{q}_\alpha'(\theta) = -c_\theta \frac{\sum_{i=1}^{k} N_{i,m}}{\sum_{i=1}^{k} D_{i,m}},$$

where

$$D_{i,m} = \frac{1}{m} \sum_{n=1}^{m} f_A\left( W_{\theta,i}(n) + S_i(n) - X_i(m) \right) \mathbf{1}_{\{ W_{\theta,i}(n) + S_i(n) - X_i(m) > 0 \}}$$

- $i = \{1, 2, \ldots, k\}$.
- $X_i(m) = W_{\theta,i}\left( \left\lfloor \alpha m \right\rfloor \right)$, (if known) $q_\alpha(\theta)$. 
Quantile Sensitivity Estimator: $G(\theta)/G/1$ queue

- **MVD component**

\[
N_{i,m} = \frac{1}{\nu_{\tau_{\theta,i}^{\pm},(m)}} \sum_{n=1}^{\tau_{\theta,i}^{\pm},(m)} \left( \mathbf{1}_{\{W_{\theta,i}^+(n,W_{\theta,i}(0)) \leq X_i(m)\}} - \mathbf{1}_{\{W_{\theta,i}^-(n,W_{\theta,i}(0)) \leq X_i(m)\}} \right)
\]

- $\tau_{\theta,i}^{\pm,(m)} = \inf \{ n \geq m : W_{\theta,i}^+(n,W_{\theta,i}(0)) = W_{\theta,i}^-(n,W_{\theta,i}(0)) = 0 \}$.
- $\nu_{\tau_{\theta,i}^{\pm},(m)} = \text{card}\{1 \leq n \leq \tau_{\theta,i}^{\pm,(m)} : W_{\theta,i}^+(n,W_{\theta,i}(0)) = W_{\theta,i}^-(n,W_{\theta,i}(0)) = 0 \}$.
- $X_i(m) = W_{\theta,i}(\lceil \alpha m \rceil), \; q_{\alpha}(\theta)$. 

Volk-Makarewicz (VU)
Example 1: $M(\theta)/M(\mu)/1$ Queue

- Inter-arrival and service times of customer are exponential:
  - $A_{\theta,i}(n) \sim \exp(\theta)$, $S_i(n) \sim \exp(\mu)$;
  - $\mathbb{E}_\theta[A_{\theta,i}(1)] = 1/\theta$, $\mathbb{E}_\theta[S_i(1)] = 1/\mu$; for $0 < \rho = \theta/\mu < 1$.
  - $A_{\theta,i}^+(n) \sim \exp(\theta)$, $A_{\theta,i}^-(n) \sim \gamma(2, \theta)$.
  - $A_{\theta,i}^+(n) = a_{\theta,i}(n)$, $A_{\theta,i}^-(n) = a_{\theta,i}(n) + A'_{\theta,i}$. $A'_{\theta,i}$ is an iid copy.

- For this DES:
  - $W_{\theta,i}^+(1, W_{\theta,i}(0)) \geq W_{\theta,i}^-(1, W_{\theta,i}(0))$ stochastically.
  - $\tau_{\theta,i}(W_{\theta,i}(0)) = \inf \{ n \geq 1 : W_{\theta,i}^+(1, W_{\theta,i}(0)) = 0 \}$.
  - $X_i(m) = q_\alpha(\theta)$, $W_{\theta,i}((\lceil \alpha m \rceil))$. 

Volk-Makarewicz (VU)
Quantile Sensitivity Estimation
WODES 2010 15 / 26
Example 1: $M(\theta)/M(\mu)/1$ Queue

- Stationary Distribution: mass at 0 w.p $1 - \rho$, $W_\theta|W_\theta > 0 \sim \exp(\mu - \theta)$ w.p. $\rho$.

- Sensitivity of quantile

\[ q'_\alpha(\theta) = -\frac{1}{(\mu - \theta)^2} \left\{ \ln \left( \frac{\mu}{\theta} \right) + \ln(1 - \alpha) \right\} + \frac{1}{\theta(\mu - \theta)} \]

for $\alpha > 1 - \rho$. 
Parameters

Parameter values of $\hat{q}'_\alpha(\theta)$ for the $M(\theta)/M(\mu)/1$ queue

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$\theta$</th>
<th>$\mu$</th>
<th>$(k, m)$</th>
<th>$N_{est}$</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.45</td>
<td>4.50</td>
<td>$\left(2^7, 2^7\right)$</td>
<td>200</td>
<td>0.90, 0.95</td>
</tr>
<tr>
<td>0.50</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table: $N_{est}$ = Number of estimates.
## Results

Estimates of $\hat{q}_{\alpha}(\theta)$ for the stationary waiting time for an $M(\theta)/M(\mu)/1$ queue. $k = m = 2^7$. $X_i(m) = q_{\alpha}(\theta)$.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$\alpha = 0.90$</th>
<th>$\alpha = 0.95$</th>
<th>$\alpha = 0.90$</th>
<th>$\alpha = 0.95$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.5548</td>
<td>0.6331</td>
<td>0.5478</td>
<td>0.5576</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>0.0211</td>
<td>0.0421</td>
<td>0.0242</td>
<td>0.0402</td>
</tr>
<tr>
<td>True Value</td>
<td>0.5487</td>
<td>0.5910</td>
<td>0.5487</td>
<td>0.5910</td>
</tr>
<tr>
<td>$\rho = 0.50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>2.2593</td>
<td>3.2291</td>
<td>2.1681</td>
<td>2.9924</td>
</tr>
<tr>
<td>$\rho = 0.90$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1544.2</td>
<td>2217.6</td>
<td>1135.2</td>
<td>1188.5</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>387.58</td>
<td>905.85</td>
<td>266.61</td>
<td>340.36</td>
</tr>
<tr>
<td>True Value</td>
<td>923.33</td>
<td>1200.6</td>
<td>923.33</td>
<td>1200.6</td>
</tr>
</tbody>
</table>
Example 2: $M(\theta)/G/1/T$ Queue

- $M/G/1/T$: Arrivals are only accepted/enter the queue if the sojourn time is $\leq T$.
- $A_{i,\theta}(n) \sim \exp(\theta)$, $A_{i,\theta}^+ = a_{i,\theta}(n)$, $A_{i,\theta}^- = a_{i,\theta}(n) + A'_{i,\theta}$.
- Service Times: $S_i(1) \sim \gamma(1 + Pn(\mu_1), \mu_2)$, $\mathbb{E}[S_i(1)] = (1 + \mu_1)/\mu_2$.
- Virtual waiting time, $V_{\theta,i}(n)$:
  \[ V_{\theta,i}(n + 1) = \max \left( V_{\theta,i}(n) + S_i(n) \mathbf{1}_{\{V_{\theta,i}(n) + S_i(n) \leq T\}} - A_{\theta,i}(n + 1), 0 \right). \]
- $X_i(m) = W_{\theta,i}(\lceil \alpha m \rceil)$. 

Volk-Makarewicz (VU)  
Quantile Sensitivity Estimation 
WODES 2010
## Parameters

Parameter values of $q'_\alpha(\theta)$ for the $M(\theta)/G/1/T$ queue

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$\theta$</th>
<th>$\mu_1$</th>
<th>$\mu_2$</th>
<th>$T$</th>
<th>$(k, m)$</th>
<th>$N_{est}$</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.45</td>
<td>6</td>
<td>1/3</td>
<td></td>
<td>${5, 10}$</td>
<td></td>
<td>200</td>
</tr>
<tr>
<td>0.50</td>
<td>3</td>
<td>7/3</td>
<td></td>
<td></td>
<td>$(2^7, 2^7)$</td>
<td></td>
<td>0.90, 0.95</td>
</tr>
<tr>
<td>0.90</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Results

Estimates of $\hat{q}_{\alpha}(\theta)$ for the stationary waiting time for an $M(\theta)/G/1/T$ queue. $k = m = 2^7$.

<table>
<thead>
<tr>
<th></th>
<th>$T = 5$</th>
<th>$T = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha = 0.90$</td>
<td>$\alpha = 0.95$</td>
</tr>
<tr>
<td>$\rho = 0.10$</td>
<td>4.4505</td>
<td>6.3583</td>
</tr>
<tr>
<td></td>
<td>0.3983</td>
<td>0.5969</td>
</tr>
<tr>
<td>$\rho = 0.50$</td>
<td>58.297</td>
<td>78.421</td>
</tr>
<tr>
<td></td>
<td>3.2257</td>
<td>4.3263</td>
</tr>
<tr>
<td>$\rho = 0.90$</td>
<td>69.236</td>
<td>78.105</td>
</tr>
<tr>
<td></td>
<td>3.2693</td>
<td>3.7904</td>
</tr>
</tbody>
</table>
Wrapping things up

- Sources of variance:
  Large $\rho$: cycle length; small $\rho$: estimation of the denominator.
  Quotient of two estimators.
- Sweet spot: $k = m$; choice of least variance.
- Biasedness:
  Sample averaging, method of implementation.
- Right-tailedness before becoming a symmetric estimator for larger $m$, since distribution of waiting times are right tailed.
- Thoughts?


References


- Discrete time ergodic Markov Chain, $P$. State $X(0) = x$.
- Construct alternative sequence, $[X'(n)]$.
- Transition from $X(0)$ to $X'(1)$ governed by $P(\delta) = P + \delta Q$.
  - $Qe = 0$, $\delta > 0$, $e = (1, \ldots, 1)^T$, such that previous matrix remains stochastic.
- Realization factor $D = [d_{ij}]$, performance function $g$
  
  $$d_{ij} = \mathbb{E} \left[ \sum_{n=1}^{\infty} (g(X'(n)) - g(X(n))) \bigg| X(0) = i, X'(0) = j \right]$$

- Two approaches are related where $Q = C(P^+ - P^-)$, $C = \text{diag}(c_i)$. 
Loynes scheme

- Context: Method of perfectly sampling $W_{\theta}(0)$ from the stationary distribution where $Y_{\theta}(n) = (A_{\theta}(n), S(n))$ are iid RV.
  - Given event $\{A_{\theta}(n') \geq T\}$, know that $n' = \eta_{\theta}$, $W_{\theta}(n') = 0$, $\forall \omega \in \Omega$.
- Via a backward construction from $n = 0$, we can determine $\eta_{\theta}$.
  - $\eta_{\theta} \sim Geo(p_{\theta})$, $p_{\theta} = P_{\theta}(A_{\theta}(1) \geq T)$.
- $(W_{\theta}(\eta_{\theta} + n))_{n>0} \sim W_{\theta}(\eta_{\theta})$.
  - MVD: All rows in derivative calculation stop at $n = m$. 