On quantile optimization problem with censored data

Petr Volf[1]

Abstract. The stochastic optimization problem is, as a rule, formulated in terms of expected cost function. However, criterion based on averaging does not take in account possible variability of involved random variables. That is why the criterion considered in the present contribution uses selected quantiles. Moreover, it is assumed that the stochastic characteristics of optimized system are estimated from the data, in non-parametric setting, and that the data may be randomly right-censored. Therefore, certain theoretical results concerning estimators of distribution functions and quantiles under censoring are recalled and then utilized to prove consistency of solution based on estimates. Behavior of solutions for finite data sizes is studied with the aid of randomly generated example.

Keywords: optimization, censored data, product-limit estimator, empirical quantile.

JEL classification: C41, J64
AMS classification: 62N02, 62P25

1 Introduction

Let us consider an optimization problem with utility function $\varphi(y, v)$, where v are input variables from certain feasibility set $\boldsymbol{V}$ and values y are random, results of a random variable (or vector) Y with distribution function F. Standardly, corresponding stochastic optimization problem can be formulated as $\sup_v E_F\varphi(Y, v)$, where E_F stands for the expectation w.r. to F. If F is known, we actually deal with a "deterministic" optimization case. However, criterion based on averaging does not take in account possible variability of r.v. Y and is actually reasonable for optimizing over long time period. Even then the variability of solution can be large. That is why the present paper is devoted to optimization of quantiles of random criterion $Z(v) = \varphi(Y, v)$. Alternatively, we can be interested in a kind of multi-objective optimization, simultaneously reducing also variability of solution (measured by variance, or certain inter-quantile range).

Further, our information on probability distribution could be non-complete. Either, known distribution type depends on unknown parameters. Or, we have to employ nonparametric estimates of F. Then, as a rule, the estimates are plugged into objective function. Hence, we have to analyze both possible bias and increased variability of obtained solution (compared to an ideal solution when F is known). An investigation of usage of empirical (estimated) characteristics in stochastic optimization problems started already in 70-ties. A number of papers has dealt with these problems, let us mention here just Kaňková (2010) with an overview and a number of other references.

In the present paper we consider even more complicated case when distribution function F should be estimated from the data censored randomly from the right side. Such situation is quite frequent in the analysis of demographic, survival or insurance data. The lack of information leads to higher variability of estimates and, consequently, to higher uncertainty of optimal solutions. The approaches to statistical data analysis in cases when the data are censored or even truncated are provided by a number of authors. The most of results were derived in the framework of statistical survival analysis and collected in several monographs (cf. Kalbfleisch and Prentice, 2002, or Andersen et al, 1993).

The main objective of the present paper is to study the increase of uncertainty of results of quantile optimization problem when the censoring is causing growing variability of non-parametric estimate of F. Therefore, in the next section, certain theoretical properties of estimates under random right censoring will be recalled. We shall consider the product-limit estimator as a generalization of the empirical distribution

[1]Institute of Information Theory and Automation, Prague 8, volf@utia.cas.cz

function, and corresponding estimator of quantiles. Their properties in cases with and without censoring will be compared. In Section 3 the consistency of solution employing estimated quantiles is proven. Finally, in Section 4 a simple example deals with optimal maintenance schedule, properties of obtained 'sub-optimal' solution are illustrated with the aid of simulations.

2 Estimators of distribution and quantile functions

Let us consider a continuous-type random variable Y characterizing for instance a random time to certain event. Let another continuous random variable U be a censoring variable, both be positive, continuous and mutually independent. Further, let $f(y)$, $g(u)$, $F(y)$, $G(u)$, $\overline{F}(y) = 1 - F(y)$, $\overline{G}(u) = 1 - G(u)$ denote the density, distribution and survival functions of both variables. It is assumed that we observe just $X = \min(Y, U)$ and $\delta = 1[Y \leq U]$, i. e. δ indicates whether Y is observed or censored from right side. The data are then given as random sample $(X_i, \delta_i, i = 1, \ldots, N)$. Notice that the case without censoring is obtained when $G(t) \equiv 0$ on region where $F(t) < 1$. Let us remark here that in some cases we can deal, for instance, with the logarithm of time. Then the domain of data can be the whole R_1.

A generalization of empirical distribution function is the well known Kaplan–Meier "Product Limit Estimate" (PLE) of survival function. Let us first sort (re-index) the data in increasing order, $X_1 \leq X_2 \leq \cdots \leq X_N$, then the PLE of $\overline{F}(t)$ has the form

$$\overline{F}_N(t) = \prod_{i=1}^{N} \left(\frac{N-i}{N-i+1} \right)^{\delta_i \cdot 1[X_i \leq t]}. \tag{1}$$

Again, notice that when all $\delta_i = 1$, we obtain the empirical survival function. The following proposition is due to Breslow and Crowley (1974):

Proposition 1. *Let $T > 0$ be such that still $\overline{F}(T) \cdot \overline{G}(T) > 0$. Then the random process*

$$V_N(t) = \sqrt{N} \left(\frac{\overline{F}_N(t)}{\overline{F}(t)} - 1 \right) = \sqrt{N}\, \frac{F(t) - F_N(t)}{\overline{F}(t)} \tag{2}$$

converges, on $[0, T]$, when $N \to \infty$, to Gaussian martingale with zero mean and variance function

$$C(t) = \int_0^t \frac{\mathrm{d}F(s)}{\overline{F}(s)^2\, \overline{G}(s)}. \tag{3}$$

Here, $F_N(t) = 1 - \overline{F}_N(t)$. In other words, $V_N(t)$ converges in distribution on $[0, T]$ to the process $W(C(t))$, where $W(\cdot)$ denotes the Wiener process. The asymptotic variance function can be estimated by its empirical version:

$$C_N(t) = \sum_{i=1}^{N} \frac{N\delta_i}{(N-i+1)^2} \cdot 1[X_i \leq t],$$

which is consistent in probability, uniformly w.r. to $t \in [0, T]$ (see again Breslow and Crowley, 1974).

Let us now recall also properties of empirical quantiles. 'True' p-quantile, for any $p \in (0, 1)$, is defined as $Q(p) = \min\{x : F(x) \geq p\}$, and is obtained as a unique solution of equation $F(x) = p$ provided F is strictly increasing. Empirical quantile is then defined as $Q_N(p) = \min\{x : F_N(x) \geq p\}$. Let now $p \in (0, F^{-1}(T))$, where T is from Proposition 1. Notice that $Q_N(p)$ is well defined only if $F_N(x) \geq p$ for some x, therefore with probability tending to 1 when $N \to \infty$. The following statement can be found for instance in Andersen et al (1993), Ch.IV.3 .

Proposition 2. *Let $f(x) > 0$ in the neighborhood of $Q(p)$. Then the empirical quantile $Q_N(p)$ is P-consistent and asymptotically normal, namely, for each $c < 1/2$*

$$N^c \cdot (Q_N(p) - Q(p)) \to^P 0, \quad \sqrt{N}(Q_N(p) - Q(p)) \to^d N(0, S(p))$$

and asymptotic variance equals

$$S(p) = \frac{(1-p)^2 \cdot C(Q(p))}{f(Q(p))^2}.$$

It follows that the variance of $(Q_N(p) - Q(p))$ can be estimated by

$$\frac{S_N(p)}{N} = \frac{(1-p)^2 \cdot C_N(Q_N(p))}{N \cdot f_N(Q_N(p))^2}, \tag{4}$$

which is complicated by inevitable estimation of density function, as a rule with the aid of kernel method.

If we denote $D_N(t) = V_N(t)/(1 + C(t))$, then for the case without censoring we obtain that $C(t) = F(t)/\overline{F}(t)$ and $D_N(t) = \sqrt{N}(F(t) - F_N(t))$ leading to standard Kolmogorov–Smirnov statistics. Notice also that then we obtain a well known result $asvar[\sqrt{N}(Q_N(p) - Q(p))] = \frac{p(1-p)}{f(Q(p))^2}$.

Further, from (3) it is also seen that the variance in the case with censoring (when $\overline{G}(t) \leq 1$) is larger than without it (i.e. when $\overline{G}(t) = 1$ on whole $[0, T]$).

3 Criterion based on quantiles, consistency

Let the optimization problem be now formulated as maximization of a p-quantile of distribution of random variable $Z(v) = \varphi(Y, v)$, for some selected $p \in (0, 1)$. If function $\varphi(y, v)$ is monotone increasing in y for each v, i.e. when there exists its inverse function $\varphi^{-1}(z, v)$, also increasing in z, then the distribution function of $Z(v)$, for fixed v, is $F_Z(z, v) = F(\varphi^{-1}(z, v))$. Therefore also quantiles of $Z(v)$ can be expressed as function of quantiles of Y, namely $Q_Z(p, v) = \varphi(Q(p), v)$ and optimal v depends directly on $Q(p)$. In general, however, connection between distribution and quantiles of variables Y and $Z(v)$ is not so straightforward and has to be analyzed, for instance with the aid of simulation.

Let the following assumptions hold:

A1. Let $p \in (0, F^{-1}(T))$ and let $f(x) > 0$ in a neighborhood of $Q(p)$.

A2. Function $\varphi(y, v)$ is bounded, increasing and continuous in a neighborhood of $y = Q(p)$, uniformly w.r. to $v \in \boldsymbol{V}$.

A3. $\boldsymbol{V}$ is compact and $\varphi(Q(p), v)$ is continuous in $v \in \boldsymbol{V}$. Hence, it is continuous uniformly in $\boldsymbol{V}$.

Further, denote $v^* = \arg\max_v \varphi(Q(p), v)$, $\varphi^* = \varphi(Q(p), v^*)$, $v_N^* = \arg\max_v \varphi(Q_N(p), v)$, $\varphi_N^* = \varphi(Q_N(p), v_N^*)$. When $\boldsymbol{V}$ is compact, at least one v^* exists, while v_N^* is defined with probability tending to 1.

Proposition 3. *When assumptions A1, A2, A3 hold, then*

1. $\varphi_N^* \to \varphi^*$ *in probability,*
2. *there exists a.s. a (random) subsequence* $N(k) \subset \{N\}$ *and* $\bar{v} \in \{v^*\}$ *such that* $v_{N(k)}^* \to \bar{v}$ *a.s.*

Proof.

i) From notation above, it follows that a.s. $\varphi_N^* \geq \varphi(Q_N(p), v^*)$ and $\varphi^* \geq \varphi(Q(p), v_N^*)$.

ii) Further, P-consistency of $Q_N(p)$ and A2 imply that, in probability, $\varphi(Q_N(p), v^*) \to \varphi^*$ and also $\varphi_N^* - \varphi(Q(p), v_N^*) \to 0$.

From i) and ii) assertion 1 follows, namely $\varphi_N^* \to \varphi^*$ in P. Moreover, it is also seen that

iii) $\varphi(Q(p), v_N^*) \to \varphi^*$ in P.

iv) The existence of converging subsequence $v_{N(k)}^*$ follows from compactness of $\boldsymbol{V}$. Then the uniform continuity of $\varphi(Q(p), v)$ ensures that $\varphi(Q(p), v_{N(k)}^*) \to \varphi(Q(p), \bar{v})$ in P.

This, together with iii), yields that $\varphi(Q(p), \bar{v}) = \varphi^*$ a.s.

Thus, except convergence of optimal values we showed also existence of a random sequence of solutions converging towards the set of optimal solutions $\{v^*\}$. If v^* is unique, then $\bar{v} = v^*$ a.s.

4 Example

Let us consider the following rather simple example of optimization problem (see also Volf, 2012): A component of a machine has its time to failure Y given (modeled) by a continuous-type probability distribution with distribution function, density, survival function F, f, $\bar{F} = 1 - F$, respectively. The cost of repair after failure is C_1, the cost of preventive repair is $C_2 < C_1$. For the simplicity we assume that only complete repairs, 'renewals', are provided, i.e. after each repair the component is new (exchanged) or as new. Let τ be the time from renewal to preventive repair, we wish to select an optimal value of τ.

Let us, as a criterion function, consider the proportion of component availability time to the unit of cost, namely

$$\varphi(y,\tau) = \frac{y}{C_1} \quad \text{if} \quad y \le \tau, \quad \varphi(y,\tau) = \frac{\tau}{C_2} \quad \text{if} \quad y > \tau.$$

4.1 Optimization of the mean

Let us first search for an optimal τ, from a reasonable closed interval $\boldsymbol{T}$, maximizing the mean

$$\phi_F(\tau) = E_F\varphi(Y,\tau) = \int_0^\tau \frac{y}{C_1}\mathrm{d}F(y) + \frac{\tau}{C_2}\overline{F}(\tau).$$

Optimal solution can be found directly, by solving equation $\mathrm{d}\phi_F(\tau)/\mathrm{d}\tau = 0$. In our case

$$\frac{\mathrm{d}\phi_F(\tau)}{\mathrm{d}\tau} = \frac{\tau}{C_1}f(\tau) + \frac{1}{C_2}\Big(\overline{F}(\tau) - \tau f(\tau)\Big).$$

Assume that the distribution of Y is Weibull, with parameters $a = 100, b = 2$, i.e. its survival function is $\overline{F}(t) = \exp\left(-\left(\frac{t}{a}\right)^b\right)$, corresponding numerical characteristics are $\mathrm{E}Y \sim 89$, $\mathrm{sd}(Y) \sim 46$, $\mathrm{median}(Y) \sim 83$. Further, let the costs be $C_1 = 10$, $C_2 = 1$. When the distribution function F is known, there exists an unique optimal solution with

$$\tau^* = a\Big(\frac{C_1}{(C_1 - C_2)b}\Big)^{1/b} = 74.5356$$

and maximal mean of working time per cost unit $\phi_F(\tau^*) = 44.7644$.

We can also compute directly the distribution of random variable $Z(\tau) = \varphi(Y,\tau)$, with property $Z(\tau) = \frac{Y}{C_1}$ if $Y \le \tau$ and $Z(\tau) = \frac{\tau}{C_2}$ if $Y > \tau$. If $Y \sim \mathrm{Weibull}(a,b)$ and certain τ is selected, the conditional distribution of random variable $Z(\tau)|Z(\tau) \le \frac{\tau}{C_1}$ has distribution function

$$F_Z(z) = \frac{1 - \exp(-(\frac{z \cdot C_1}{a})^b)}{(1 - \exp(-(\frac{\tau}{a})^b)}, \tag{5}$$

i.e. $Z(\tau)$ has on interval $(0, \tau/C_1)$ Weibull distribution with parameters $(a/C_1, b)$ and $Z_\tau = \tau/C_2$ with probability $P(Y > \tau)$. Therefore, we can compute variance and standard deviation. Namely, for optimal τ^* $\mathrm{sd}(Z(\tau^*)) = 34.5$. If we wish to reduce it, we have to accept certain trade-off. For instance, $\tau = 52$ yields approximately $\mathrm{E}Z(\tau) = 40.4$, so that just by 10% smaller value than the maximum, while standard deviation is reduced to $\mathrm{sd}(Z(\tau)) = 20.7$.

4.2 Optimization of quantiles

Let us return to the criterion based on a quantile. Above, in (5), the distribution of Z_τ has been derived, the quantiles $Q_Z(p,\tau)$ follow immediately from it. Figure 1 shows their form, from two different points of view. It follows that if we wish to maximize certain α-quantile of Z_τ, optimal $\tau^*(\alpha)$ should be such that $P(Y > \tau^*(\alpha)) = 1 - \alpha$, i.e. $\tau^*(\alpha)$ is the α quantile of the distribution of Y. Guaranteed value reached by Z with probability $1 - \alpha$ is then $\tau^*(\alpha)/C_2$. It is a consequence of the form of function $\varphi(y,\tau)$.

For instance for $\alpha = 0.1$ and $Y \sim$ Weibull(100,2) we obtain that $\tau^*(\alpha) = 32.4593$ and 90% guaranteed value of Z is $\tau^*(\alpha)/C_2 = 32.4593$, too, because $C_2 = 1$. If we wish to achieve a higher value of Z with sufficiently large probability, we can select for instance $\tau = 45$. As it corresponds roughly to 18%-quantile of Y, such a choice guarantees that $P(Z(\tau) = 45) \sim 0.82$. On the other hand, if we take τ^* maximizing the mean $E_F\varphi(Y,\tau)$, it guarantees that $P(Z(\tau) = 74.5) \sim 0.57$ and $P(Z(\tau) < 7.45) \sim 0.43$. It is seen that even here a kind of trade-off, with the use of multi-criteria approach, is reasonable.

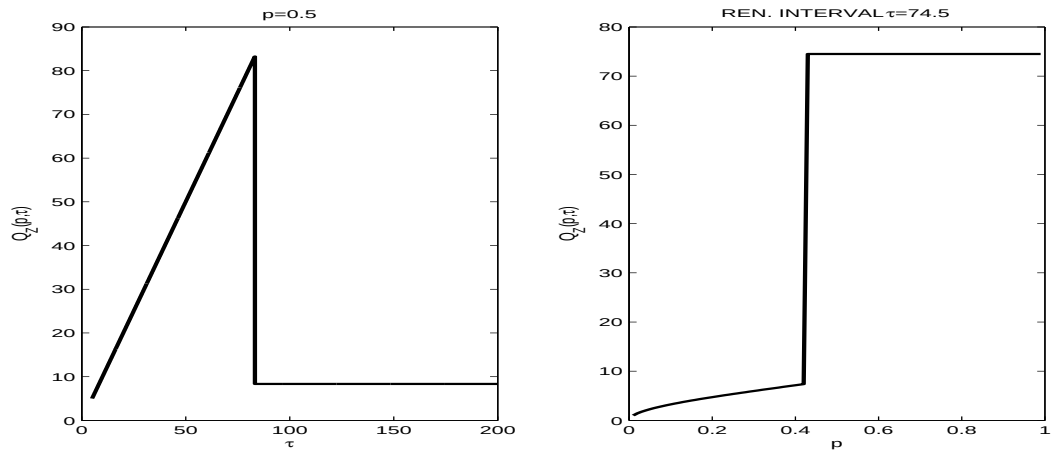

Figure 1 Quantiles $Q_Z(p,\tau)$ of variable $Z(\tau)$, left as a function of τ for given $p = 0.5$, right as a function of p for given $\tau = \tau^*$

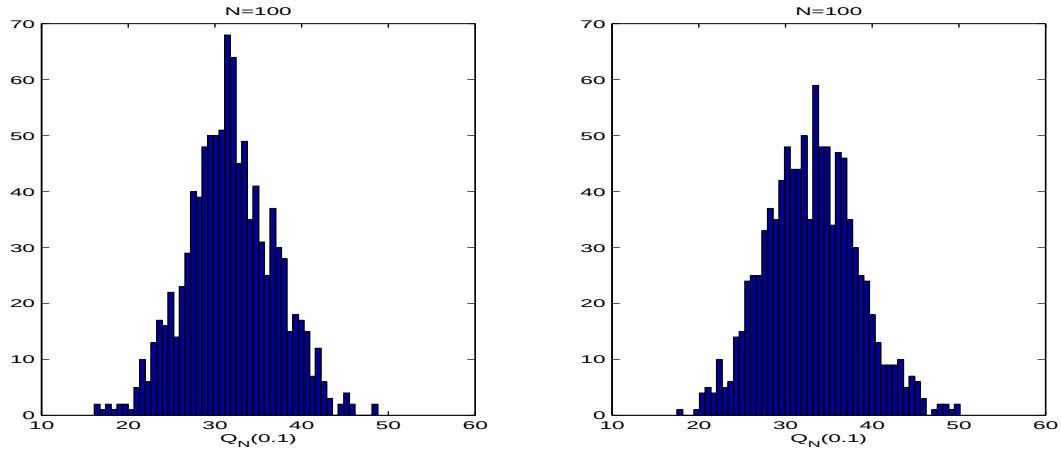

Figure 2 Histogram of generated sample quantiles $Q_N(p)$, for $N = 100$ and $p = 0.1$, computed from non-censored (left) and censored (right) data cases

4.3 A numerical study

In this part we provide a numerical study where it is assumed that the distribution of variable Y is estimated from data. In both considered cases (without or with censoring) $M = 1000$ samples of $N = 100$ and $N = 300$ observations Y_i are generated from the Weibull distribution specified above. In cases with censoring, censoring variables U_i have uniform distribution on $[0, 250]$, hence with survival function $\overline{G}(u) = (250 - u)/250$ (value 250 corresponds roughly to 0.998 quantile of distribution of Y). The rate of censoring is then about $36\,\% \sim EY/250$.

It is assumed that the type of distribution of Y is not known and therefore F is estimated non-parametrically with the aid of the product-limit estimator (i.e. as the empirical distribution function in the case without censoring). Thus, M estimates $F_N^{(m)}, m = 1, ..., M$ are obtained, from each the empirical quantile is computed, for given p. Figure 2 displays histograms of these M estimated quantiles, for $N = 100$, $p = 0.1$, the cases without censoring are plotted in the left subplot, the right subplot shows estimates obtained from censored data. Incorrect specification of the quantile shifts slightly both the guaranteed value and its probability, either decreases the value and increases probability, if $Q_N(p) < Q(p)$, or, in the opposite case, increases the value and decreases its probability.

It is well seen how the variability in the right subplot has increased due to censoring. Table 1 compares sample means and standard deviations computed from M=1000 values $Q_N^{(m)}(p)$ with true $Q(p)$ and standard deviations (denoted 'as-std') obtained as square roots of approximate variances $S(p)/N$, with $S(p)$ from Proposition 2. It is well seen how sample characteristics approach theoretical values.

p=0.1 N	Q(p)	non-cens.: as-std	sample mean	sample std		censored: as-std	sample mean	sample std
100	32.46	5.135	32.11	5.077		5.378	33.16	5.283
300	32.46	2.962	32.96	2.926		3.105	32.61	3.090
p=0.5								
100	83.26	6.006	82.82	5.826		6.907	83.53	6.990
300	83.26	3.467	83.34	3.466		3.988	83.34	3.961

Table 1 Comparison of theoretical and sample-based characteristics of empirical quantiles for $p = 0.1$ and $p = 0.5$, $N = 100$ and $N = 300$

5 Conclusion

We have studied the impact of variability of statistical estimates to uncertainty of solution in a stochastic optimization problem formulated via certain quantiles of utility function. We compared two cases, namely that the stochastic characteristics of the problem were estimated, in a non-parametric way, from fully observed or from randomly right-censored data. Therefore, theoretical properties of estimators of distribution function and quantiles from censored data were recalled, in order to compare them with the behavior of estimates in real situations. Such a comparison was performed with the aid of a simple optimization problem example and randomly generated data. Simultaneously, the convergence of solutions based on estimated quantiles to optimal solution was proven.

Acknowledgements

The research is supported by the project of GA CR No 13-14445S.

References

[1] Andersen, P., Borgan, O., Gill, R., and Keiding, N.: *Models Based on Counting Processes*, Springer, New York, 1993.

[2] Breslow, N. and Crowley, J. E.: A large sample study of the life table and product limit estimates under random censorship, *Ann. Statist.* **2** (1974), 437–453.

[3] Kalbfleisch, J. D. and Prentice, R. L.: *The Statistical Analysis of Failure Time Data (2-nd Edition)*, Wiley, New York, 2002.

[4] Kaňková, V.: Empirical estimates in stochastic optimization via distribution tails, *Kybernetika* **46** (2010), 459–471.

[5] Volf, P.: On precision of optimization in the case of incomplete information, *Bulletin of the Czech Econometric Society,* **19** (2012), 170–184.