

Sample-based quantile function estimation using entropy and self-determined probability weighted moments

Jian Deng

Central South University, Changsha, China

M.D. Pandey

University of Waterloo, Waterloo, Canada

Desheng Gu

Central South University, Changsha, China

ABSTRACT: The method of conventional probability-weighted moments (PWMs) and the principle of minimum cross entropy were combined to derive an analytical form of quantile function [Pandey, 2001]. This method is extended and improved in the present paper by utilizing the concept of self determined probability weighted moments (SD-PWMs). Minimum cross entropy principle and SD-PWMs from sample data were used to evaluate the unknown parameters of the cross entropy quantile function. The performances of the SD-PWM based quantile function were compared with those based on conventional PWMs. The algorithm directly accommodates sample outliers by assigning non-exceedance probabilities accordingly, which is extraordinarily suitable for estimating extreme events from a small sample set. The proposed method would be expected to be used in extreme value and risk analyses of engineering systems.

1 INTRODUCTION

The problem of estimation of extreme quantiles corresponding to small probabilities of exceedance (POE) is commonly encountered in the risk analysis of engineering systems. The first step in quantile estimation involves fitting an analytical probability distribution for adequate representation of sample observations. To achieve this, the distribution type is judged empirically from the available data, and then distribution parameters are suitably estimated using methods such as maximum likelihood method, the method of least squares, and the method of moments. However, the bias and efficiency of quantile estimates remains sensitive to the type of assumed distribution.

An alternative approach to the distribution fitting comes from the modern information theory in which entropy has been developed. Maximum entropy principle (MaxEnt) was presented as a rational approach for choosing the most unbiased probability distribution, amongst all possible distributions, which is consistent with available data and contains minimum spurious information. Later, a rational method was developed to integrate a prior distribution with available data for inference purposes based on the concept of cross entropy. The method is based on the minimum cross-entropy principle (CrossEnt) stating that given a prior distribution and some data, select that posterior distribution which minimizes the cross-entropy. When only moment constraints are specified, the cross-entropy minimization was proved to be a uniquely correct method of probabilistic inference that satisfies all consistency axioms [Shore and Johnson, 1980].

Since probability weighted moments (PWMs) were first defined [Greenwood et al, 1979], it attracts many researchers from various scientific and engineering fields. In contrast with ordinary statistical moments, the main advantage of PWMs is that their higher order values can be accurately estimated from small samples. In a previous paper, Pandey (2001) originally interpreted PWMs as moments of quantile function (QF) and presents a distribution free method for unbiased estimation of the QF of a random

variable using the principle of CrossEnt subject to constraints specified in terms of the first few order conventional PWMs.

The problem of outliers has been always focusly-concerned in statistical data analysis and the literature on outliers is vast. Just take flood frequency analysis as example, the term “outlier” is commonly used to denote large floods in the systematic record or historical floods which lie far above the majority of the floods in the sample. The mere existence of these outliers complicates the frequency analysis procedure. Since PWMs are linear combinations of the observed data values, they are shown to be to some extent insensitive to outliers (extreme observations) in the data, as compared to the ordinary statistical moments, where the data are squared, cubed, etc. However, such conventional PWMs was still accused of being unable accurately and effectively to accommodate outlier in a finite sample data [Haktanir, 1997], because they assign non-exceedance probabilities to sample points based on only their rank number in an ordered series rather than the magnitude of the points themselves.

In this paper, We extended Pandey’s elegant idea by use of first few order probability weighted moments to self-determined probability weighted moments (SD-PWMs). To assess the bias and efficiency of CrossEnt quantile estimates, Monte Carlo simulations were performed.

2 PROBABILITY WEIGHTED MOMENTS

2.1 Probability weighted moments

The probability weighted moment (PWM) of a given random variable was formally defined by Greenwood et al. (1979) as

$$M_{r,s,t} = E[X^r F^s (1-F)^t] = \int_0^1 [x(F)]^r F^s (1-F)^t dF$$

When X denotes the random variable, $F = F(x, \phi_1, \phi_2, \dots, \phi_k)$ is the cumulative distribution function having $\phi_1, \phi_2, \dots, \phi_k$ as parameters, $x(F) = x(F, \lambda_1, \lambda_2, \dots, \lambda_k)$ is the inverse cumulative distribution function having $\lambda_1, \lambda_2, \dots, \lambda_k$ as parameters, and r, s, t are real numbers. The following two forms of PWM are particularly simple and useful:

$$\text{Type 1: } \alpha_t = M_{1,0,t} = \int_0^1 x(F)(1-F)^t dF \quad (1)$$

And

$$\text{Type 2: } \beta_s = M_{1,s,0} = \int_0^1 x(F)F^s dF \quad (2)$$

For an ordered sample, $x_1 \leq x_2 \leq \dots \leq x_n$, the type 1 and type 2 IPWMs can be also estimated as

$$a_t = \frac{1}{n} \sum_{i=1}^n [(1 - P_{nex,i})^t x_i] \quad (3a)$$

$$b_s = \frac{1}{n} \sum_{i=1}^n [(P_{nex,i}^s) x_i] \quad (3b)$$

where $t, s=0,1,\dots,(n-1)$ are non-negative integers, x_i is the i th element in the observed ordered sample, $x_1 \leq x_2 \leq \dots \leq x_n$, $P_{nex,i}$ is the non-exceedance probability of the i th event, which can be estimated by a plotting position technique (PPT) or without a PPT:

(1) U-IPWM: $P_{nex,i}$ in unbiased PPT estimators [Landwehr et al 1979] is taken as

$$P_{nex,i}^j = \frac{(i-1)!(n-j-1)!}{(n-1)!(i-j-1)!}, \quad j = 0,1,2,\dots,n-1 \quad (4)$$

(2) Biased-IPWM: $P_{nex,i}$ in biased PPT estimators is expressed as

$$P_{nex,i}^j = \left(\frac{i-0.35}{n} \right)^j \quad (5)$$

(3) SD-PWM: In self-determined probability-weighted moments (SD-PWMs) [Haktanir,1997], $P_{nex,i}$ was given by the cumulative distribution function of the probability distribution used,

$$P_{nex,i}^j = [F(x_i, \phi_1, \phi_2, \dots, \phi_k)]^j \quad (6)$$

Eq.(3a) requires that the polynomial in $P_{nex,i}$ be expanded prior to use of Eqs.(4-6).

2.2 PWMs as Moments of Quantile Function

For a non-negative random variable, the β_s can be interpreted as moments of the quantile function [Pandey, 2001]. Recall the definition of an ordinary moment of s^{th} order

$$E[X^s] = \int_R x^s f(x) dx = \int_0^1 [x(u)]^s du \quad (7)$$

Where $du = dF(x) = \frac{f(x)dx}{\int f(x)dx}$ is a probability measure, which is a monotonic, continuous and non-negative function with $0 \leq F(x) \leq 1$. Comparing Eqs.(2) and (7), $x(u)$ is analogous to $f(x)$, whereas the probability u is analogous to the random variable x . A probability measure, $dT(u)$, can be introduced using the following normalizing transformation

$$dT(u) = \frac{x(u)du}{\int_0^1 x(u)du} = \frac{x(u)du}{\beta_0} \quad (8)$$

where β_0 ($=\alpha_0$) is the area under the quantile function, and is also equal to the average of the random variable. Thus the PWM can be redefined using a new measure, dT as

$$\beta_k = \beta_0 \int_0^1 u^k dT(u) \quad (9)$$

Comparing Eqs.(7) and (9), (β_k / β_0) can be interpreted as k^{th} moment of the quantile function $x(u)$,

$0 < u < 1$.

3 QUANTILE FUNCTIONS

3.1 Derivation of FPWM based CrossEnt QF

Given a prior distribution, the principle of minimum cross-entropy chooses that distribution from a set of candidate distributions which minimizes the cross-entropy and satisfies other specified constraints, e.g. moments. Given the moment constraints only, Shore and Johnson [1980] proved that the cross-entropy minimization is a uniquely correct method of probabilistic inference that satisfies all the consistency axioms. When the prior distribution is uniform, cross-entropy minimization is equivalent to entropy (or uncertainty) maximization.

Let $x(F)$ represent the true but unknown quantile distribution of random variable X , and $x_M(F)$ be the M -th order estimated model. Considering $y(F)$ as a prior estimate of the exact QF $x(F)$, the cross-entropy of $x(F)$ can be written as

$$S(x) = -\int_0^1 [x(F) \ln \frac{x(F)}{y(F)}] dF \quad (10)$$

and the available information is presented as PWMs

$$\int_0^1 x_M(F) F^k dF = b_k, \quad k = 0, 1, \dots, M \quad (11)$$

where $\{b_i\}_{i=0}^M$ is sample estimate of population PWM, M is the highest order of PWM considered in the analysis.

To account for the constraints in Eq. (11), the entropy function is augmented as

$$\bar{S} = -\int_0^1 [x(F) \ln \frac{x(F)}{y(F)}] dF - (\lambda_0 - 1) \left[\int_0^1 x(F) dF - b_0 \right] - \sum_{k=1}^M \lambda_k \left[\int_0^1 F^k x(F) dF - b_k \right] \quad (12)$$

where λ_k denotes an unknown Lagrangian multiplier. Note that $(\lambda_0 - 1)$ is used as the first multiplier instead of λ_0 as a matter of convenience. To derive the QF, the usual optimization condition is used

$$\frac{\partial \bar{S}}{\partial x(F)} = 0 \quad (13)$$

Substitution from Eq.(12) into Eq. (13) and subsequent simplification leads to the following solution:

$$x_M(F) = y(F) \exp \left[- \sum_{k=0}^M \lambda_k F^k \right] \quad (14)$$

where $\{\lambda_i\}_{i=0}^M$ is Lagrangian multiplier. Eq.(14) is referred to as PWM-based CrossEnt QFs from a sample.

3.2 Parameter estimation of quantile function

Having obtaining the analytical form of maximum entropy quantile function as in Eqs.(14-15), the next step is to estimate the unknown parameters from a sample data. Substitution from Eq.(14) into Eq.(11),

$$\int_0^1 F^k y(F) \exp\left[-\sum_{i=0}^M \lambda_i F^i\right] dF = b_k \quad (k = 0, 1, \dots, M) \quad (16)$$

In Pandey's minimum cross entropy quantile function, b_k in Eq.(15) was directly estimated by use of the unbiased PPT. For a prescribed set of PWM order M , the Lagrangian multiplier are obtained by solving the system of $(M+1)$ nonlinear equations in Eq.(16).

However, as pointed by Haktanir (1997), Any PPT, however suitable it may be, ascribes a non-exceedance probability to any element of the sample series as a function of its rank number in the ordered series only. Using any particular PPT would comprise some ambiguity because: (1) It cannot directly account for a possible outlier present in the sample series since it determines the probability of the maximum element of the series as a simple function of the sample length only; and (2) it assumes a definite order for the elements of the sample series as defined by its simple form in Eq.(4) or Eq.(5).

On the other hand, if b_k in Eq.(16) was estimated by the distribution itself without a PPT, a better and more rigorous approximation would be obtained. Method of SD-PWMs without plotting positions assign non-exceedance probabilities to sample points based on the magnitude of the points themselves rather than only rank-order within the sample data. Such an algorithm directly accounts for outliers by assigning non-exceedance probabilities accordingly, which is extraordinarily suitable for estimating extreme events from a finite sample set. Because SD-PWMs are completely decided by sample data's value and magnitude, no any other hypothesis, SD-PWM CrossEnt QF was defined as sample-based quantile function. Using Haktanir's SD-PWMs, Eq.(16) becomes

$$\int_0^1 F^k y(F) \exp\left[-\sum_{i=0}^M \lambda_i F^i\right] dF = \frac{1}{n} \sum_{j=0}^n [F(x_j)]^k x_j \quad (17)$$

When a specific element x_j ($j = 1, 2, \dots, n$) is considered, Eq.(14) is

$$x_j = y[F(x_j)] \exp\left\{-\sum_{i=0}^M \lambda_i [F(x_j)]^i\right\} \quad (18)$$

Then the term of $F(x_j)$ in right side of Eq.(18) can be obtained by solution of Eq.(19), which can be restated as follows.

$$(\lambda_0 + \ln(x_j)) - \log[y(F_j)] + \lambda_1 F_j + \lambda_2 F_j^2 + \dots + \lambda_M F_j^M = 0 \quad (19)$$

Where $F_j = F(x_j)$. For properly-selected values of λ_i ($i = 0, 1, \dots, M$), there should exist a F_j which corresponds to x_j . This is a nonlinear equation in one unknown, which can be solved a standard iterative algorithm. The initial guess of F_j may be its plotting position. Eq.(19) also shows that $F(x_j) = F(x_j; \lambda_0, \lambda_1, \dots, \lambda_M)$. This results expressions of Eq.(17) in which the unknown parameters λ_i ($i = 0, 1, \dots, M$) appear as both dependent variables and independent variables. Consequently, Eq.(17) constitute a system of simultaneous nonlinear equations. Determination of those particular values of unknown parameters for a given sample data will require some sort of an iterative model.

Since the parameters of sample-based quantile function using SD-PWMs without PPT are determined using iterative techniques, initial guesses on the parameters are required. Because SD-PWMs follow conventional PWMs with PPT, parameters of quantile function based on PWMs with PPT are reasonable

initial guesses for those based on SD-PWMs. Particularly, when $M=2$, the resultant equations are expressed as

$$\int_0^1 y(F) \exp[-\lambda_0 - \lambda_1 F - \lambda_2 F^2] dF = \frac{1}{n} \sum_{j=1}^n x_j \quad (20a)$$

$$\int_0^1 F \cdot y(F) \exp[-\lambda_0 - \lambda_1 F - \lambda_2 F^2] dF = \frac{1}{n} \sum_{j=1}^n [F(x_j)] x_j \quad (20b)$$

$$\int_0^1 F^2 \cdot y(F) \exp[-\lambda_0 - \lambda_1 F - \lambda_2 F^2] dF = \frac{1}{n} \sum_{j=1}^n [F(x_j)]^2 x_j \quad (20c)$$

These nonlinear integral equations can be solved using the software MATLAB[®]. The selection of the type of prior distribution is entirely up to the user. However, the standard exponential distribution is known to be a good choice for describing unbounded tail of a distribution. The cumulative density function (CDF) and QF of standard exponential distribution are as follows:

$$F(y) = 1 - \exp(-y) \quad \text{and} \quad y(F) = -\ln(1 - F) \quad (0 \leq y \leq \infty) \quad (21)$$

4 EXAMPLES

4.1 Example 1

The CDF, QF and PWMs of Weibull distribution are given in Appendix. The Weibull distribution is taken as the parent distribution with a fixed scale parameter as $a = 1$ and varying values of the shape parameter (c), ranging from 1 to 4.

Two experiments were performed. In the first one, 2000 nonoverlapping samples of 20 elements each were used, which were extracted from the original unordered series (40,000 elements). These are unmodified samples. In the second experiment, the 100th element from the top of this long 40000-element sample ranked in descending order (x_{100}) was first found. And the median value of each of the short 20-element sample was searched out and replaced with x_{100} . By doing so, each short sample was converted to a series containing almost an outlier [Haktanir, 1997]. These are modified samples with outliers. With every one of these unmodified and modified sample data, CrossEnt QFs were obtained using biased sample PWMs, unbiased sample PWMs and SD-PWMs of order 0-2 ($M=2$), respectively. And then quantile (POE= 10^{-2}) accuracy was evaluated and compared by using common error metrics such as bias and root-mean square error (RMSE)) being expressed as:

$$\text{Normalized bias} = \frac{E[\hat{z} - z]}{z} = \frac{1}{N} \sum_{k=1}^N \left[\frac{\hat{z}_k}{z} - 1 \right]; \quad \text{RMSE} = \frac{1}{N} \sqrt{E[(\hat{z} - z)^2]}$$

where $\hat{z}$ denotes the sample estimate of a random quantity z (quantile), $\hat{z}_k$ is the k th sample estimate in simulation and N being the number of simulations. The order 10^{-2} of POE generally represent nominal values of various mechanical loads (live, dead, and construction load) and material properties, such as yield strength and fracture toughness. The procedure is shown in Fig.1.

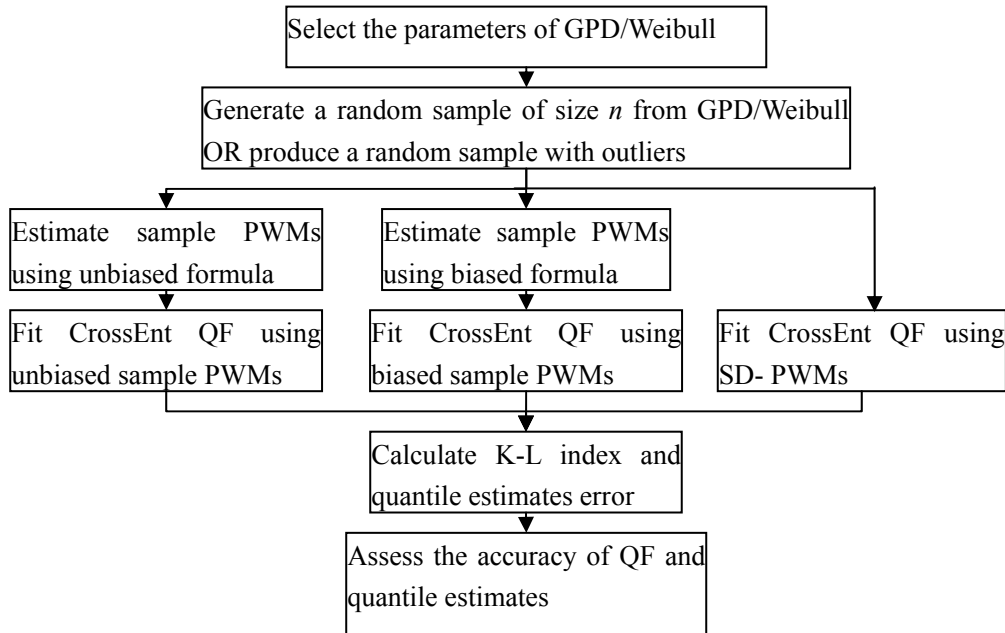


Fig.1 Simulation-based assessment CrossEnt quantile estimates using various PWMs

When unmodified samples or modified samples were used, among crossEnt QFs using biased sample PWMs, unbiased sample PWMs and SD-PWMs, QFs using SD-PWMs show lowest normalized bias (Fig.2 and Fig.4) and lowest normalized RMSE (Fig.3 and Fig.5), except the case of shape parameter equal to 1. This suggests that improvement can be achieved in efficiency and accuracy of quantile estimation by use of SD-PWMs over of conventional PWMs.

4.2 Example 2

The cumulative distribution and quantile function of generalized Pareto distribution (GPD) are given in Appendix. The GPD is taken as the parent distribution with a fixed scale parameter ($d=1.0$) and varying values of the shape parameter (c), ranging from -0.4 to +0.4. Two experiments as Example 1 were performed. In both cases, CrossEnt QFs using SD-PWMs have comparable normalized bias and normalized RMSE as those using biased sample PWMs and unbiased sample PWMs, which are illustrated in Figures 6-9.

5 CONCLUSIONS

This paper presents a sample-based method for extreme quantile estimation from a small sample of data using the minimum cross entropy principle and self-determined probability weighted moments. This method directly accommodates outliers by assigning non-exceedance probabilities accordingly, which is extraordinarily suitable for estimating extreme events from a finite sample set. The proposed method is a further development of a direct method for quantile estimation, which used the probability weighed moments (PWMs) in place of product moments commonly used with CrossEnt principle.

Two examples, with unmodified examples and modified examples respectively, are considered to evaluate the performance of the proposed method: (1) Generalized Pareto distribution used in peaks over threshold method of extreme value analysis, and (2) Weibull distribution, commonly used in reliability analysis. In both cases, the proposed method exhibits significantly higher estimation accuracy. The bias and RMSE, associated with SD-PWM method are smaller or comparable to that obtained from the use of biased PWM method and unbiased PWM method. The results show that some improvement can be achieved in efficiency and accuracy of quantile estimation by use of SD-PWMs over of conventional PWMs, especially sample data containing outliers. The proposed method should be useful in extreme value and reliability analyses of engineering systems [Deng,2006]

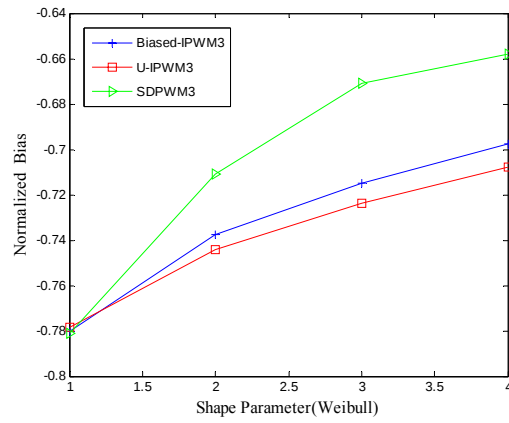


Fig.2 Bias of CrossEnt QFs from samples without outlier using Weibull PWMs

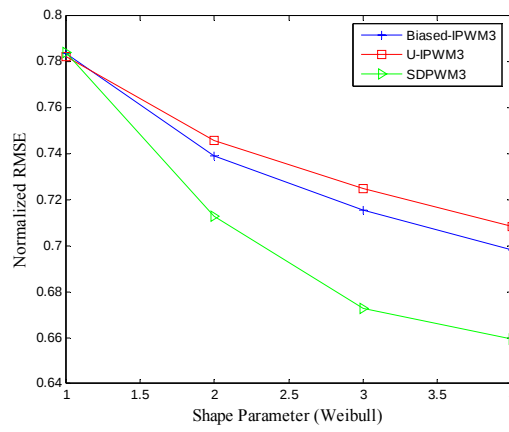


Fig.3 RMSE of CrossEnt QFs from samples without outlier using Weibull PWMs

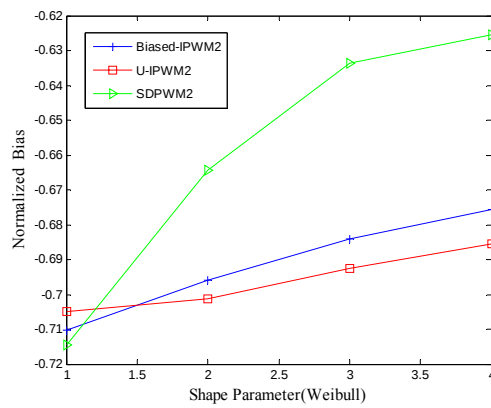


Fig.4 Bias of CrossEnt QFs from samples with outlier using Weibull PWMs

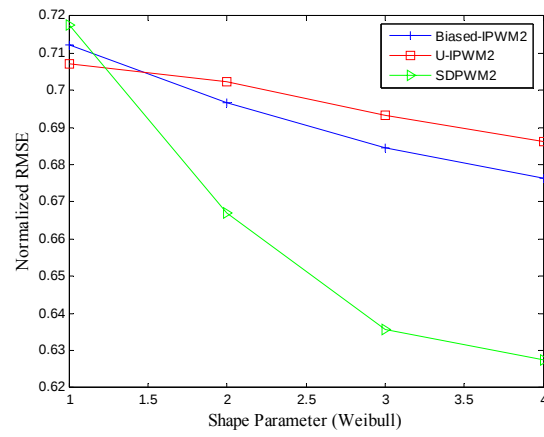


Fig.5 RMSE of CrossEnt QFs from samples with outlier using Weibull PWMs

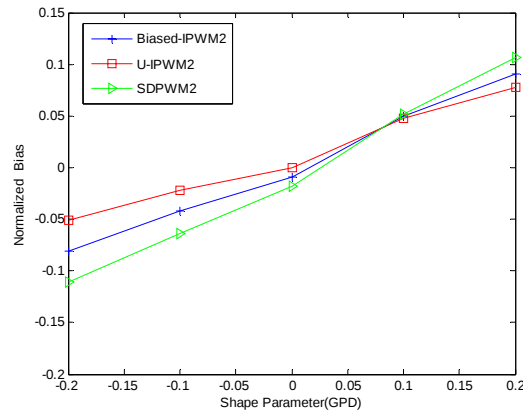


Fig.6 Bias of CrossEnt QFs from samples without outlier using GPD PWMs

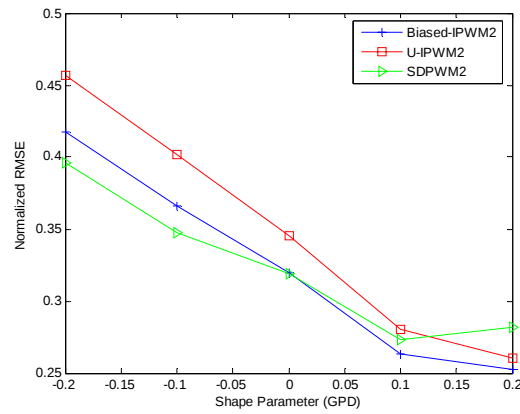


Fig.7 RMSE of CrossEnt QFs from samples without outlier using GPD PWMs

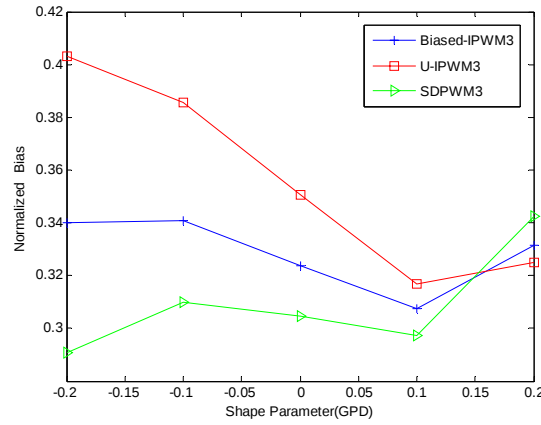


Fig.8 Bias of CrossEnt QFs from samples with outlier using GPD PWMs

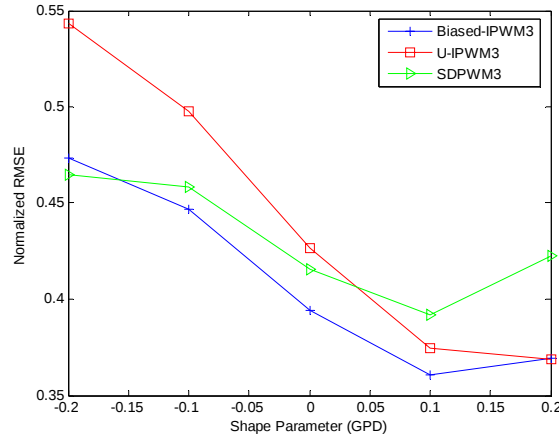


Fig.9 RMSE of CrossEnt QFs from samples with outlier using GPD PWMs

ACKNOWLEDGMENTS

Supports from National Natural Science Foundation of China (Nos 50404010 and 50574098), Hunan Natural Science Foundation (05jj10010), program for NCET in universities and SRF for ROCS, SEM were gratefully acknowledged. The authors thankfully acknowledged the support provided by Science and Engineering Research Canada (NSERC) and University Network of Excellence in Nuclear Engineering (UNENE).

APPENDIX

A.1 Generalized Pareto Distribution

The cumulative distribution function (CDF) of generalized Pareto distribution is defined as

$$F(x) = u = 1 - [1 - cx/d]^{1/c} \quad (c \neq 0) \quad \text{or} \quad F(x) = 1 - \exp(-x/d) \quad (c=0) \quad (\text{A.1})$$

where c and d are known as shape and scale parameters. The range of x is $0 \leq x < \infty$ for $c \leq 0$ and $0 \leq x \leq \frac{c}{d}$ for $c > 0$. The quantile function (QF) can be obtained by inverting (A.1)

$$x(u) = \frac{d}{c} [1 - (1-u)^c] \quad (\text{for } c \neq 0) \quad \text{or} \quad x(u) = -d \log(1-u) \quad (\text{for } c = 0) \quad (\text{A.2})$$

The type 1 IPWMs are given as

$$\alpha_k = \frac{d}{(k+1)(k+1+c)} \quad \text{for } k=0,1,\dots,n \quad (\text{A.3})$$

which exist provided that $c > -1$. The distribution parameter can be estimated as

$$c = \frac{\alpha_0}{\alpha_0 - 2\alpha_1} - 2; \quad d = \frac{2\alpha_0\alpha_1}{\alpha_0 - 2\alpha_1} \quad (\text{A.4})$$

A.2 Weibull Distribution

The Weibull cumulative distribution function is given as

$$F(x) = 1 - \exp\left[-\left(\frac{x}{a}\right)^b\right] \quad (\text{A.5})$$

where $a > 0$ and $b > 0$ are the scale and shape parameters, respectively. The Weibull QF is given as

$$x(F) = a[-\log(1-F)]^{1/b} \quad (\text{A.6})$$

The type 1 IPWMs are given as

$$\alpha_k = \frac{a\Gamma(1+1/b)}{(k+1)^{(1+1/b)}} \quad \text{for } k=0,1,\dots,n \quad (\text{A.7})$$

where $\Gamma[\cdot]$ denotes the gamma function which is defined as

$$\Gamma[z] = \int_0^\infty t^{z-1} e^{-t} dt \quad (\text{A.8})$$

The distribution parameters can be estimated as

$$a = \frac{\alpha_0}{\Gamma[\ln(\alpha_0/\alpha_1)/\ln(2)]}; \quad b = \frac{\ln(2)}{\ln[\alpha_0/(2\alpha_1)]} \quad (\text{A.9})$$

REFERENCES

- Pandey, M.D. (2001) *Extreme quantile estimation using order statistics with minimum cross-entropy Principle. Probabilistic Engineering Mechanics*, 16(1), 31-42
- Shore J.E., Johnson R.W. (1980) *Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Transaction on Information Theory*, 26(1):26-37
- Greenwood, J.A., Landwehr, J.M., Matalas, N.C. and Wallis, J.R.. (1979) *Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form, Water Resources Research*, 15(5): 1049-1054
- Haktan r, T. (1997) *Self-determined probability-weighted moments method and its application to various distributions. Journal of Hydrology*, 194(1): 180-200
- Deng J.(2006) *Structural reliability analysis for implicit performance function using radial basis function network. International Journal of Solids and Structures*, 43(11-12): 3255-3291
- Deng J., Pandey M.D.(2007) *Estimation of the maximum entropy quantile function using fractional probability weighted moments. Structural Safety, (in press), doi:10.1016/j.strusafe.2007.05.005*

