Some Convergence Problems On Heavy Tail Estimation Using Upper Order Statistics For Generalized Pareto and Lognormal Distributions

Raul Hernandez-Molinar
Tec de Monterrey, Campus San Luis Potosí.
John Lefante.
Tulane University.

In some applications, the population characteristics of main interest can be found in the tails of the distribution function. The study of risk of extreme events will lead to the use of probability distributions and the scenarios that correspond to the tail of these distributions.

Considering two approaches: parametric and nonparametric, the research emphasizes the assessment of distribution tails, assuming that underlying distributions are heavy tailed.

Two heavy tailed distributions are considered: Generalized Pareto and Lognormal. The Maximum likelihood estimation method, using the complete sample, and using only the upper order statistics provide estimators of the parameters. Measures of Bias and Mean Squared Error of the estimators of the parameters, and the Conditional Mean Exceedence Functions of the distributions, are generated.

The methodology for estimating population parameters, has potential applications in financial markets, quality control, assurance portfolios, monitoring of residual discharges, medical applications, design of environmental policies, or calibration and adjustment of processes and equipment.

The main idea is to present, and analyze the methods used for the estimation, and some convergence problems when these two distribution functions are used in generating scenarios.

I. Introduction

The determination of distribution functions based on censored samples of extreme values is very useful for instance, in the determination of limits of tolerance, or policy formulations. If the determination involves the risk of extreme events or situations with a low probability of occurrence, the analysis can consider those techniques derived under extreme value theory.

The work deals with the comparison of two estimation methods: the classical maximum likelihood estimation method and an asymptotic maximum likelihood estimation method (AEVML), when a censored sample of *k* upper order statistics is used. Two distribution functions are employed as underlying distributions:

Lognormal, Generalized Pareto. These two distributions belong to the domain of attraction of the Gumbel limiting distribution *exp(-exp(-x))*.

Using the joint density function of the set of upper order statistics, normalizing constants depending on the underlying distribution parameters, and normalizing constants proposed by I. Weissman (1978), under the parametric or nonparametric approach, the estimation of the underlying distribution parameters will be achieved. The estimated distribution function parameters are defined in closed form, and estimates of parameters, large quantiles, and confidence intervals of these quantiles will be determined.

Differences between the estimated parameters are analyzed when the entire sample and the sample with the upper order statistics are generated using Monte Carlo simulation processes.

II. Maximum Likelihood Asymptotic Method

This Method was proposed by I. Weissman, it is applied when we have a random sample of size *n*, using only the *k* upper order statistics.

The method is called "asympthotic" given the limiting distribution function is generated taken into account the asymptothical property of the extreme values. The main assumption is that the distribution function belongs to the domain of attraction of the limiting distribution function Gumbel.

The normalization constants a_n, b_n need to be known.

It is possible the approximation of the likelihood function of the joint distribution function for the *k* upper order statistics, which are evaluated on $(y_{n-k+1;n},\cdots,y_{n;n})$.

The method maximizes the asymptotic likelihood function in order to obtain then estimators for the distribution function, from any population sample under study, based on the determination of the normalization constants a_n, b_n, which are required for the standardization of the random variables in the domain of attraction for the limiting distribution.

The asymptotic likelihood function for the *k* upper order statistics, using the Gumbel distribution is:

$$u^k(u_i) = \exp\left\{-\exp\left[-(u_k)\right]\prod_{i=1}^{k}\frac{1}{a_n}\exp\left[-(u_i)\right]\right\}.$$

The log for the corresponding likelihood function is:

$$l\left(u^k(u_i)\right) = -\exp\left[-(u_k)\right] - \sum_{i=1}^{k}(u_i) - k\ln(a_n).$$

III. Estimating the parameters for Lognormal Distribution Function.

The proposed normalizing constants are:

$$a_n^* = \sigma\left(2\ln\frac{n}{2\sqrt{\pi}}\right)^{-1/2}, \text{ y} \tag{1}$$

$$b_n^* = \sigma\left(2\ln\frac{n}{2\sqrt{\pi}}\right)^{1/2} - \left\{\frac{\ln\ln\left(\frac{n}{2\sqrt{\pi}}\right)}{\left(\frac{4}{\sqrt{2}\sigma}\right)\left(\frac{\ln(n)}{2\sqrt{\pi}}\right)^{1/2}}\right\} + \mu. \tag{2}$$

Note that:
$$L_1 = \left(2\ln\frac{n}{2\sqrt{\pi}}\right)^{1/2}. \tag{3}$$

and
$$L_2 = \left\{\frac{\ln\ln\left(\frac{n}{2\sqrt{\pi}}\right)}{\left(\frac{4}{\sqrt{2}}\right)\left(\frac{\ln(n)}{2\sqrt{\pi}}\right)^{1/2}}\right\}. \tag{4}$$

This means the following:

$$a_n = \left(\frac{\sigma}{L_1}\right)\left[\exp(\sigma L_1 - \sigma L_2 + \mu)\right], \text{ y} \tag{5}$$

$$b_n = \exp(\sigma L_1 - \sigma L_2 + \mu). \tag{6}$$

The estimation of the parameters associated to the original distribution, using normalizing constants generated from Weissman´s equations (1) and (2); and solving to determine the parameters:

$$\hat{\sigma} = \left(\frac{\hat{a}_n}{\hat{b}_n}\right)L_1, \tag{7}$$

$$\hat{\sigma} = \left[\frac{\frac{1}{k}\sum_{i=1}^{k} y_{n-i+1;n} - y_{n-k+1;n}}{y_{n-k+1;n} + \hat{a}_n \ln(k)} \right] \left(2\ln \frac{n}{2\sqrt{\pi}} \right)^{1/2}, \text{ and} \tag{8}$$

$$\hat{\mu} = \ln(\hat{b}_n) + \hat{\sigma} L_2 - \hat{\sigma} L_1 . \tag{9}$$

then,

$$\hat{\mu} = \ln\left[y_{n-k+1;n} + \hat{a}_n \ln(k) \right] - \hat{\sigma} \left\{ \frac{\ln\ln\left(\frac{n}{2\sqrt{\pi}} \right)}{\left(\frac{4}{\sqrt{2}} \right)\left(\frac{\ln(n)}{2\sqrt{\pi}} \right)^{1/2}} \right\} - \hat{\sigma} \left(2\ln \frac{n}{2\sqrt{\pi}} \right)^{1/2}. \tag{10}$$

IV. Estimating the parameters for the Generalized Pareto

It is possible to consider the normalizing constants as a function of the α and β parameters for the Generalized Pareto distribution, using the following

$$a_n = \frac{1}{\alpha} \tag{11}$$

$$b_n = \ln\left(\alpha\beta n^{\frac{1}{\alpha}} \right) \tag{12}$$

We can estimate the parameters α and β, using equation (11) :

$$\hat{\alpha} = \frac{1}{\hat{a}_n} . \tag{13}$$

If we employ the sample with the *k* upper order statistics, and the normalizing constants proposed by Weissman:

$$\hat{\alpha} = \left[\frac{1}{k} \sum_{i=1}^{k} y_{n-i+1;n} - y_{n-k+1;n} \right]^{-1}. \tag{14}$$

and with the equation (12)

$$\exp(b_n) = \alpha \beta n^{\frac{1}{\alpha}} \tag{15}$$

and

$$\beta = \frac{\exp(b_n)}{\alpha n^{\frac{1}{\alpha}}}. \tag{16}$$

Using the sample with the *k* upper order statistics, we have:

$$\hat{\beta} = \frac{\exp\left[y_{n-k+1;n} + \hat{a}_n \ln(k) \right]}{\hat{\alpha} n^{\frac{1}{\alpha}}} \tag{17}$$

$$= \frac{\exp\left[y_{n-k+1;n} + \frac{1}{\hat{\alpha}} \ln(k) \right]}{\hat{\alpha} n^{\frac{1}{\alpha}}}. \tag{18}$$

V. Summary of the equations for estimating the distribution functions parameters based on the upper order statistics

Lognormal

$$\hat{\sigma} = \left[\frac{\frac{1}{k}\sum_{i=1}^{k} y_{n-i+1;n} - y_{n-k+1;n}}{y_{n-k+1;n} + \hat{a}_n \ln(k)} \right] \left(2\ln \frac{n}{2\sqrt{\pi}} \right)^{1/2}$$

$$\hat{\mu} = \ln\left[y_{n-k+1;n} + \hat{a}_n \ln(k) \right] - \hat{\sigma} \left\{ \frac{\ln\ln\left(\frac{n}{2\sqrt{\pi}} \right)}{\left(\frac{4}{\sqrt{2}} \right)\left(\frac{\ln(n)}{2\sqrt{\pi}} \right)^{1/2}} \right\} -$$

$$-\hat{\sigma}\left(2\ln \frac{n}{2\sqrt{\pi}} \right)^{1/2}$$

Generalized Pareto

$$\hat{\alpha} = \left[\frac{1}{k}\sum_{i=1}^{k} y_{n-i+1;n} - y_{n-k+1;n} \right]^{-1}$$

$$\hat{\beta} = \frac{\exp\left[y_{n-k+1;n} + \frac{1}{\hat{\alpha}} \ln(k) \right]}{\hat{\alpha} n^{\frac{1}{\alpha}}}$$

VI. Simulating the Process in order to compare two methods (Classical versus Asymptotic)

Monte Carlo simulations were achieved, employing S-Plus. We determine the sample size, the critical values, confidence intervals, and the number of the upper order statistics required.

An important condition was that all the values in the simulations correspond to those values greater than the 95th percentile.

The simulations were achieved 5000 y 10000 repetitions. It was observed that the convergence of the joint distribution for the *k* upper *order statistics* is affected when the ratio $\frac{k}{n}$ tends to increase.

VII. Results.

The parameters of the distribution function were defined in closed form. They were used to generate estimations of the parameters, upper quantiles and confidence intervals.

We estimate the parameters using the two methods (classical maximum likelihood, and the proposed method based on the *k* upper order statistics).

In the comparison, the parameters observed significant differences.

A linear regression model has been employed in order to make an adjustment, reducing the bias. Some results are presented in the following tables and figures.

Table 1. Lognormal Distribution Function: μ=1.0, σ=2.0
Estimatión of μ, σ, and Confidence Intervals for 95% del Percentil
97.5th Percentile= 136.99 $^{\lozenge}$

Method	n	k	$x_{.975}$	μ.hat	σ.hat	LL($x_{.975}$)	UL($x_{.975}$)
CML	50		151.1002	0.999383	1.990654	135.6191	168.4317
AEVML		10	155.858	3.104203	0.85973	144.6066	168.1029
AEVML		15	135.1858	3.148323	0.779081	125.1908	146.0811
AEVML		17	128.3995	3.150791	0.755762	118.7995	138.873
AEVML		20	119.5155	3.145783	0.726985	110.4244	129.4466
CML	100		143.279	1.000702	1.992253	130.365	157.5019
AEVML		10	205.9494	3.309579	0.916164	194.486	218.1611
AEVML		15	186.8276	3.388895	0.837506	176.3729	197.9664
AEVML		17	179.917	3.403894	0.814389	169.7932	190.7061
AEVML		20	170.5077	3.41607	0.78553	160.8188	180.8382
CML	500		138.9284	1.000366	2.000714	130.2017	148.2427
AEVML		10	315.2707	3.67144	1.013223	304.8402	326.0713
AEVML		15	311.6269	3.814709	0.935529	301.6637	321.9316
AEVML		17	308.304	3.851366	0.91245	298.5165	318.4247
AEVML		20	302.546	3.893518	0.882987	293.0059	312.4085
CML	1000		137.3952	1.000022	1.998569	130.0562	145.1493
AEVML		10	369.8256	3.78984	1.046469	360.0961	379.8254
AEVML		15	375.5244	3.956688	0.968441	366.0909	385.208
AEVML		17	374.7261	3.999175	0.946086	365.4121	384.2844
AEVML		20	371.9592	4.051995	0.916252	362.8262	381.3287
CML	2000		137.2012	1.000657	1.998956	131.0037	143.6921
AEVML		10	426.0407	3.894652	1.076785	417.0787	435.1993
AEVML		15	442.1998	4.08059	0.999353	433.393	451.189
AEVML		17	444.7263	4.131756	0.97607	435.9977	453.6331
AEVML		20	446.0707	4.194388	0.945884	437.4646	454.8496
CML	5000		137.1661	1.000455	1.99979	132.2124	142.3056
AEVML		10	508.2382	4.034428	1.105586	500.3182	516.2858
AEVML		15	538.3942	4.235752	1.031588	530.4895	546.418
AEVML		17	546.3654	4.296521	1.007768	538.4927	554.3544
AEVML		20	554.2807	4.367966	0.978381	546.4664	562.2079

$^{\lozenge}$ CML: classical maximum likelihood estimation method.
AEVML: asymptotic extreme value maximum likelihood estimation method.
x.975: 97.5th percentile.
LL(x.975): lower limit of the 95% confidence interval of the 97.5th percentile.
UL(x.975): upper limit of the 95% confidence interval of the 97.5th percentile.

Table 2. Lognormal Distribution Function: μ=1.0, σ=2.0.
Estimatión of μ, σ, and 95% Confidence Intervals for the Percentile
Estimation considering the 2%of the sample
97.5^{th} Percentile= 136.99 ◊

Method: CML

n	x.975	μ.hat	σ.hat	LL(x.975)	UL(x.975)
500	138.4873	1.001506	1.998716	129.7782	147.7835
1000	137.989	0.999796	2.000731	130.6301	145.7633
2000	137.1454	0.999307	1.999385	130.9487	143.6357
3000	137.321	1.000727	1.999909	131.7036	143.178
4000	136.8556	0.998262	1.999664	131.6295	142.2893
5000	137.1894	1.000325	1.999989	132.2349	142.3296
6000	137.0754	1.000817	1.999389	132.3404	141.9799
7000	137.3935	1.000432	2.000839	132.8274	142.1167
8000	137.0333	0.999743	1.999914	132.6214	141.592
9000	136.9886	0.998732	2.00031	132.7029	141.4127

Method: AEVML

n	k	x.975	μ.hat	σ.hat	LL(x.975)	UL(x.975)
500	10	330.4393	3.684161	1.012436	319.8662	341.3822
1000	20	372.3839	4.056589	0.915406	363.2409	381.7633
2000	40	425.2246	4.384709	0.828921	417.3486	433.252
3000	60	453.0681	4.547747	0.784901	445.8897	460.3635
4000	80	474.4482	4.65574	0.757353	467.7159	481.2783
5000	100	500.655	4.746268	0.737836	494.1984	507.1968
6000	120	512.0802	4.802929	0.722165	505.9022	518.3343
7000	140	518.9844	4.847886	0.708789	513.0456	524.9923
8000	160	528.3794	4.886662	0.698838	522.6217	534.201
9000	180	537.1612	4.921627	0.689864	531.5583	542.8234

◊ CML: classical maximum likelihood estimation method.
AEVML: asymptotic extreme value maximum likelihood estimation method.
x.975: 97.5^{th} percentile.
LL(x.975): lower limit of the 95% confidence interval of the 97.5^{th} percentile.
UL(x.975): upper limit of the 95% confidence interval of the 97.5^{th} percentile.

Table 3. Función de Distribución μ=1.0, σ=2.0.
Estimation of μ, σ, and 95% Confidence Intervals for the Percentile
Estimation considering *n=10,000* with *k* increasing.
Percentile 97.5th = 136.99 $^{\lozenge}$

n		Method: CML x.975	μ.hat	σ.hat	LL(x.975)	UL(x.975)
10,000		137.1993	1.000573	2.000187	133.0201	141.5097
		137.0134	0.999097	2.000246	132.8371	141.3211
		137.1566	1.000678	1.99995	132.9783	141.4661
		137.0425	0.999938	1.999906	132.8661	141.3502
		137.1893	0.999633	2.000645	133.0098	141.5001
		137.1877	1.000457	2.000192	133.0087	141.4979
		137.2782	1.000442	2.000535	133.0975	141.5903
		137.1205	1.000955	1.999679	132.9431	141.4292
		136.93	0.999622	1.999662	132.7556	141.2357
		136.9478	0.999823	1.999646	132.7731	141.2538

n	k	Method: AEVML x.975	μ.hat	σ.hat	LL(x.975)	UL(x.975)
10,000	10	579.7007	4.130346	1.125745	572.5354	586.958
	20	644.0885	4.478735	1.003854	636.8888	651.3702
	40	677.9958	4.760767	0.885444	671.0577	685.0062
	60	657.0752	4.854463	0.823368	650.4871	663.7306
	80	639.6383	4.905287	0.784977	633.2879	646.0528
	100	628.8564	4.940787	0.757947	622.6709	635.1037
	120	607.5412	4.950729	0.736388	601.5459	613.5966
	140	593.6539	4.960985	0.719725	587.7946	599.5719
	160	574.4377	4.956677	0.704763	568.7358	580.1972
	180	559.9716	4.957157	0.693063	554.3852	565.6145

$^{\lozenge}$ CML: classical maximum likelihood estimation method.
AEVML: asymptotic extreme value maximum likelihood estimation method.
x.975: 97.5th percentile.
LL(x.975): lower limit of the 95% confidence interval of the 97.5th percentile.
UL(x.975): upper limit of the 95% confidence interval of the 97.5th percentile.

Tabla 4. Generalized Pareto Distribution: α=0.5,β=1.0.
Estimation of α, β, and 95% Confidence Interval for the Percentile.
97.5th Percentile= 10.64 $^{\lozenge}$

Method	n	k	x.975	α.hat	β.hat	LL(x.975)	UL(x.975)
MM	100		10.21411	0.448253	1.049305	7.574498	12.85373
AEVML		10	9.072792	1.927104	0.67707	1.395652	16.74993
AEVML		15	9.641101	1.665762	0.614059	2.492206	16.79
AEVML		17	9.882994	1.595344	0.595928	2.835489	16.9305
AEVML		20	10.281	1.509057	0.572483	3.288849	17.27315
MM	500		10.44153	0.481475	1.014706	9.227232	11.65583
AEVML		10	9.146177	2.271939	0.900852	3.148405	15.14395
AEVML		15	9.07275	2.042759	0.814372	3.789824	14.35568
AEVML		17	9.060889	1.995187	0.793973	4.002802	14.11898
AEVML		20	9.065163	1.926766	0.767022	4.274302	13.85602
MM	1000		10.56122	0.49109	1.006824	9.690577	11.43185
AEVML		10	9.673477	2.31842	1.002401	4.059459	15.28749
AEVML		15	9.406827	2.113953	0.895521	4.370872	14.44278
AEVML		17	9.352905	2.05958	0.867201	4.510687	14.19512
AEVML		20	9.293226	2.003406	0.836414	4.71387	13.87258
MM	2000		10.58521	0.493648	1.005997	9.968819	11.20161
AEVML		10	10.31456	2.399917	1.137711	5.319209	15.30991
AEVML		15	9.823995	2.177263	0.989529	5.206302	14.44169
AEVML		17	9.719389	2.129426	0.956601	5.244032	14.19475
AEVML		20	9.612262	2.072786	0.919219	5.345247	13.87928
MM	5000		10.63911	0.498254	1.001534	10.24651	11.03171
AEVML		10	11.12103	2.416307	1.276549	5.956	16.28605
AEVML		15	10.43398	2.220278	1.099526	6.155651	14.7123
AEVML		17	10.25518	2.169934	1.053808	6.138093	14.37227
AEVML		20	10.06631	2.115599	1.005845	6.132447	14.00018

$^{\lozenge}$ MM: moments estimation method.
AEVML: asymptotic extreme value maximum likelihood estimation method.
x.975: 97.5th percentile.
LL(x.975): lower limit of the 95% confidence interval of the 97.5th percentile.
UL(x.975): upper limit of the 95% confidence interval of the 97.5th percentile.

Figure 1. Lognormal data from random generator. m=1.0, s=2.0.
Correcting the Bias on m: Using only m.aevml as predictor
(n=sample size, k=number of upper order statistics)

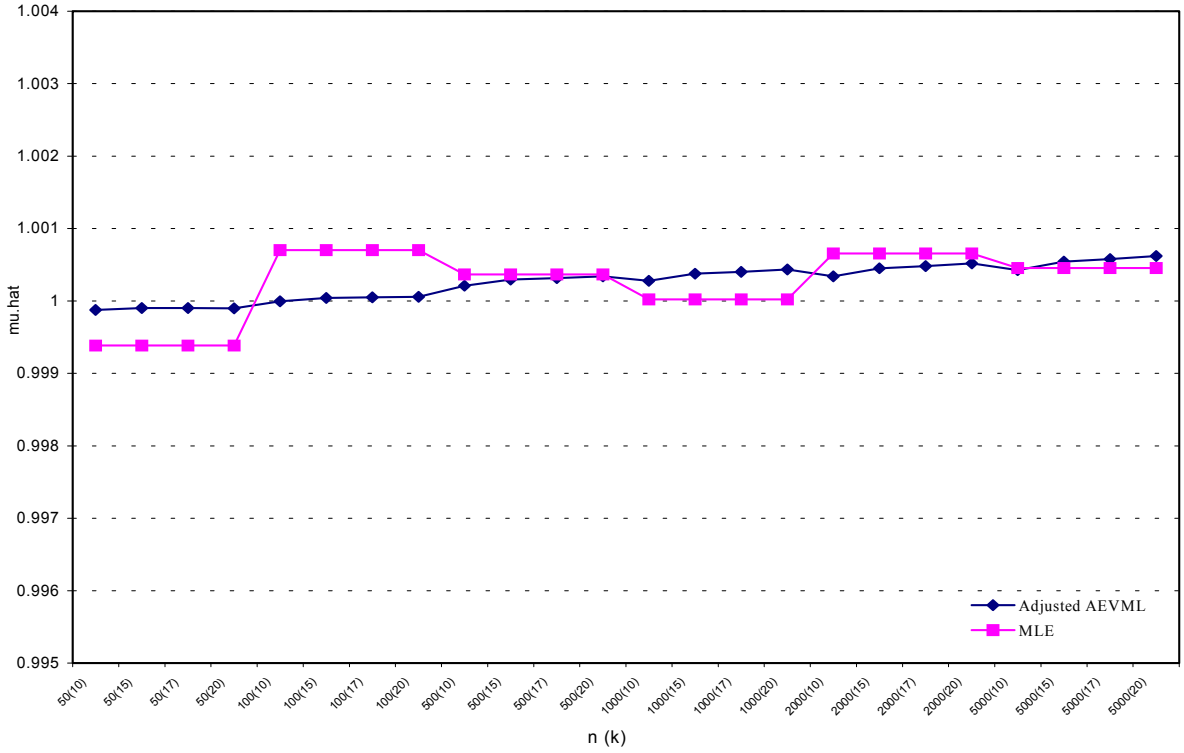

Figure 2. Lognormal data from random generator. μ=1.0, σ=2.0
Correcting the Bias on σ. Using n, k, and σ.aevml as covariates
(n=sample size, k=number of upper order statistics)

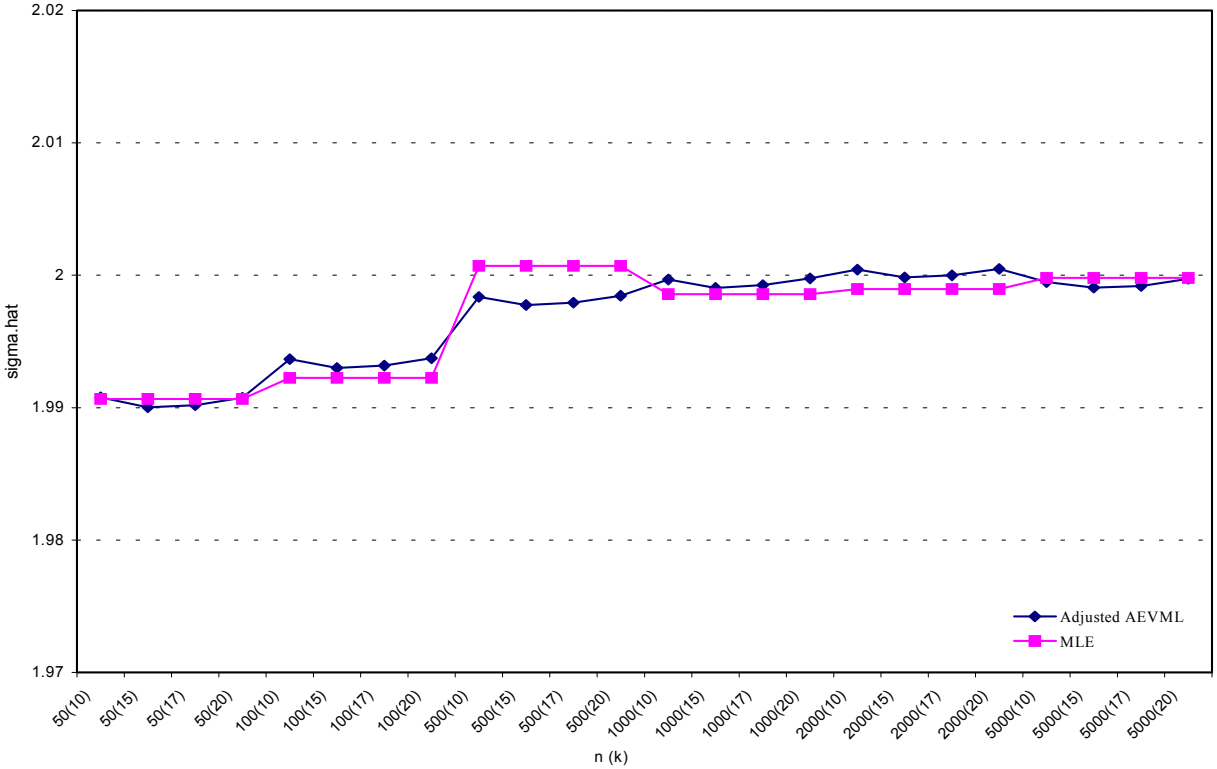

Figure 3. Generalized Pareto data from random generator. α=0.5, β=1.0.
Correcting the Bias on α: Using n, k, and, α.aevml as covariates
(n=sample size, k=number of upper order statistics)

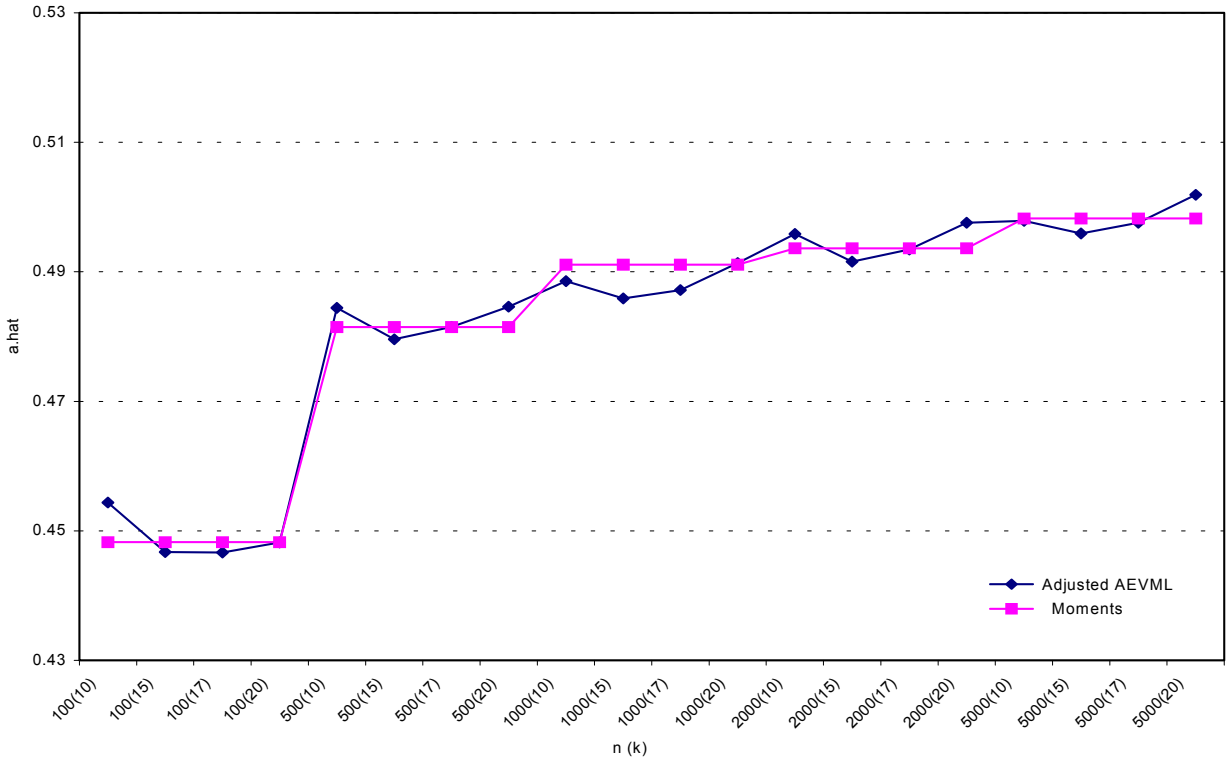

Figure 4 Generalized Pareto data from random generator. α=0.5, β=1.0.
Correcting the Bias on β: Using n, k and β.aevml as covariates
(n=sample size, k=number of upper order statistics)