Thailand Statistician
October 2025; 23(4): 771-785
http://statassoc.or.th
Contributed paper

A Non-Parametric Estimator of the Probability Weighted Moments for Large Datasets

Toufik Guermah [a] and Abdelaziz Rassoul [b]*
[a] LMA Laboratory/Departement of Mathematics/Hassiba Benbouali University, Chlef, Algeria
[b] GEE Laboratory/National Higher School for Hydraulics, Blida, Algeria
*Corresponding author; e-mail: a.rassoul@ensh.dz

Received: 5 February 2022
Revised: 2 July 2023
Accepted: 23 July 2023

Abstract

In this paper, we introduces a nonparametric median-of-means (MoM) estimator for Probability Weighted Moments (PWM) specifically designed for large datasets. Our approach draws inspiration from the data grouping method, a widely utilized technique in various domains including economics, hydrology, finance, and insurance. We establish the consistency and asymptotic normality of the proposed estimator under reasonable assumptions regarding the increasing number of subgroups. Additionally, we present a novel approach for testing hypotheses related to Probability Weighted Moments (PWM) using the Empirical Likelihood method (EL) specifically tailored for the median. Notably, our method circumvents the need for prior estimation of the variance structure associated with the estimator, a task that can be challenging and prone to inaccuracies. We conducted numerical simulations to assess the performance of our proposed estimator. The results clearly indicate that our estimator showcases remarkable robustness, particularly when confronted with outliers.

Keywords: PWM, median-of-means, large datasets, empirical likelihood, hypothesis test.

1. Introduction and Motivation

Probability-weighted moments (PWM) generalize the concept of moments of a probability distribution. It is generally used to estimate the parameters of extreme distributions of natural phenomena. Particularly, in the fields of hydrology and Climatology, researchers use PWM for the estimation of parameters of the distributions related to water discharges and maxima of temperatures. Greenwood et al. (1979) proposed the concept of probability weighted moments to estimate the parameters involved in the models of extremes of natural phenomena. The PWM of a random variable X with distribution function $F(.)$ is defined as

$$M_{p,r,s} = \mathbb{E}\left(X^p F(x)^r (1 - F(x))^s\right),$$

where p, r and s are any real numbers. Hosking et al. (1985) studied PWM of the form given by

$$\beta_r = \mathbb{E}\left(XF(x)^r\right), \tag{1}$$

to characterize various distributional properties such as the assessment of scale parameter, skewness of the distribution, and L-moments. In this article we discuss the empirical likelihood inference of β_r proposed by Hosking et al. (1985) given by Equation (1). Given a random sample

$X_1, X_2, ..., X_N$ of size N from an unknown distribution function F, let $X_{(i)}, i = 1, 2, \ldots, N$ be the i-th order statistic. David and Nagaraja (2003) proposed an estimator (D-N estimator) for β_r by replacing the unknown distribution function F in the definition of β_r with its empirical counterpart $F_N(x) = \frac{1}{N}\sum_{i=1}^{N} 1_{\{X_i \leq x\}}$, where 1_A denotes the indicator function in the A. Their estimator is given by

$$\bar{\beta}_{r,N} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{i}{N}\right)^{r} X_{(i)}. \tag{2}$$

To develop an empirical likelihood inference for β_r, Vexler et al. (2017) proposed an estimator (Vexler's estimator) given by

$$\hat{\beta}_{r,N} = \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{i}{N}\right)^{r+1} - \left(\frac{i-1}{N}\right)^{r+1}\right] X_{(i)}, \tag{3}$$

and showed that asymptotic behavior of both $\bar{\beta}_{r,N}$ and $\hat{\beta}_{r,N}$ are the same.

Recently, Vexler et al. (2017) developed an EL-based inference for PWM and showed that the limiting distribution of log EL ratio statistic is χ^2 distribution with one degree of freedom. Using EL approach, they constructed confidence intervals and likelihood ratio tests for β_r. Vexler and Zou (2018) explained the procedure for obtaining probability weights in the EL method for PWM. To find probability weights, the Lagrange multiplier method is used with multipliers λ_1 and λ_2. Jing et al. (2009) introduced the jackknife empirical likelihood (JEL) inference, which combines two of the popular non-parametric approaches namely, the jackknife and the EL approach. It is essential that the convex hull of the estimating equation should have a zero vector as an interior point for the constrained maximization problem of the profile EL function. Chen et al. (2008) mentioned that there are two drawbacks in assigning -∞ to the log-EL statistic suggested by Owen (2001) to ensure the solution existence of the required computational problem; first, it is numerically difficult to determine that there is no solution; second, no information is provided on the relative plausibility of the parameter values where the likelihood is set to zero. To fix these drawbacks of EL method, Chen et al. (2008) proposed the concept of adjusted empirical likelihood (AEL) which preserves the asymptotic properties of EL. Zhao et al. (2015) combined the concepts of AEL and JEL such that the resulting technique, adjusted jackknife empirical likelihood (AJEL) is well-defined for all the values in the parameter space.

This paper mainly focuses on the probabilistic and statistical methods for this issue. Due to the complexity of the asymptotic variance of PWM, the non-parametric confidence intervals constructed via estimating the asymptotic variance are usually inaccurate. Bhati et al. (2021) develop a jackknife empirical likelihood (JEL) and adjusted jackknife empirical likelihood (AJEL) based inference for finding confidence intervals for (PWM), to improve the inference on β_r which avoids the prior estimation of variance. More recently, and motivated by the very poor performance of AJEL in Bhati et al. (2021), Jiang and Zhao (2022) reformulated AJEL of PWMs and improved the performance in terms of coverage probability for a confidence interval.

In this paper, we introduce a more straightforward method based on the idea of random grouping and the usual empirical likelihood method for the median to study PWM. Our approach can be classified as one of the so-called divide-conquer methods. More precisely we divide the dataset into several groups and then obtain the interesting statistic within each group as a first step. In the second step of "conquer", considering robustness, we take the median, instead of mean, of the resulting statistics as our final estimator. It works well, especially in the case of massive data such as high-frequency data in financial markets. In a world full of big data, we believe that we have developed one effective and robust inference approach to reducing the computational burden arising from the analysis of massive data. The so-called median-of-means estimator was proposed, independently, by Blair (1985), Jerrum et al. (1986), Alon et al. (2002). For properties, applications, and extensions of the median-of-means estimator, we refer to Bubeck et al. (2013), Devroye et al. (2016), Hsu and Sabato (2013), Lerasle and Oliveira (2011), Minsker (2015) and Audibert and Catoni (2011).

The rest of the paper is organized as follows: In Section 2 we present our proposed estimator for β_r and its asymptotic properties. Section 3 is devoted to an empirical likelihood approach to testing β_r. Section 4 contains some criteria about the choices of blocks. Some results of simulations are given in Section 5. The proofs of different results are postponed to Appendix.

2. Median-of-means estimate for β_r

In this paper, we utilize the median-of-means method to develop a novel estimator for the PWM. For a more comprehensive understanding of this method, we recommend referring to Alon et al. (2002) for more detailed information.

To fix the idea, we divide the N observations $X_1, ..., X_N$ into K blocks randomly. Assume that each block contains n data points for simplicity. In block $B_j, j = 1, 2, ..., K$, we estimate β_r by

$$\hat{\beta}^{(j)}_{r,n} = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{i}{n}\right)^{r+1} - \left(\frac{i-1}{n}\right)^{r+1}\right] X_{(i)}. \tag{4}$$

Next, we define the median-of-means estimator of β_r as

$$\tilde{\beta}^{MoM}_r := Median\left\{\hat{\beta}^{(1)}_{r,n}, \hat{\beta}^{(2)}_{r,n}, ..., \hat{\beta}^{(K)}_{r,n}\right\}. \tag{5}$$

The asymptotic properties of $\tilde{\beta}^{MoM}_r$ are presented in the following theorems.

Theorem 1 *Assume that $E(|X|^3) < \infty$ and $\sigma^2(F) > 0$, where*

$$\sigma^2(F) := \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (F(x \wedge y) - F(x)F(y))\, F^r(x)\, F^r(y)\, dxdy \tag{6}$$

If F has a strictly positive, continuous density function f, then for any fixed $x > 0$,

$$P\left(\left|\tilde{\beta}^{MoM}_r - \beta_r\right| \geq x\right) \leq \frac{C}{(N/K)^{K/5}} \tag{7}$$

holds for some constant $C := C(x) > 0$ and any positive integer K.

Remark 1

- Note that the constant C at the right-hand side of (7) is not uniform in x.
- Theorem 1 directly implies that the convergence of $\tilde{\beta}^{MoM}_r$ towards to β_r is almost surely by Borel-Cantelli lemma.

Theorem 2

1. *Suppose K is fixed. Let $\Theta_1, \Theta_2, ..., \Theta_K$ be independent and identically distributed standard normal random variables. Then as $N \to \infty$,*

$$\frac{\sqrt{N}}{\sigma(F)}\left(\tilde{\beta}^{MoM}_r - \beta_r\right) \overset{\mathcal{D}}{\to} Median\{\Theta_1, \Theta_2, ..., \Theta_K\}, \tag{8}$$

where "$\overset{\mathcal{D}}{\to}$" means convergence in distribution.

2. *Suppose $N/K^2 \to \infty$ as $K \to \infty$. Then the following asymptotic normality holds,*

$$\frac{\sqrt{N}}{\sigma(F)}\left(\tilde{\beta}^{MoM}_r - \beta_r\right) \overset{\mathcal{D}}{\to} \sqrt{\frac{\pi}{2}}\mathcal{N}(0,1). \tag{9}$$

3. Empirical Likelihood Test Hypothesis

In this section, we based on empirical likelihood (EL, Owen (1990) and Ma et al. (2022)) to consider the testing hypothesis problem of whether the PWM equals a given value. As commented at the end of the section 2, we will not use $\tilde{\beta}_r^{MoM}$ to construct the statistic test since it involves the unknown $\sigma(F)$.

Since different blocks are disjoint, we have $\hat{\beta}_{r,n}^{(1)}, \hat{\beta}_{r,n}^{(2)}, ..., \hat{\beta}_{r,n}^{(K)}$ are independent and share the same distribution, we can regard them as one sample and apply EL.
We denote $Z_{n,k} := I\left(\hat{\beta}_{r,n}^{(k)} \le \beta_r\right)$, for each k, obviously, $\mathbb{E}\left(Z_{n,k}\right) \approx 0.5$, hence the empirical likelihood ratio for β_r is given by

$$\mathcal{R}\left(\beta_r\right) = \max\left\{\prod_{k=1}^{K} Kw_k \mid \sum_{k=1}^{K} w_k Z_{n,k} = 0.5, w_k \ge 0, \sum_{k=1}^{K} w_k = 1\right\}. \tag{10}$$

By the Lagrange multiplier method, the maximum point is given by:

$$w_k = \frac{1}{K}\frac{1}{1+\lambda\left(Z_{n,k}-0.5\right)}, \tag{11}$$

where $\lambda = \lambda\left(\beta_r\right)$ satisfies the following equation

$$\frac{1}{K}\sum_{k=1}^{K}\frac{Z_{n,k}}{1+\lambda\left(Z_{n,k}-0.5\right)} = 0. \tag{12}$$

Similar to the arguments as in Owen (1990), we can get the following theorem.

Theorem 3 *Under the same assumptions in Theorem 2, we have*

$$-2\log\mathcal{R}\left(\beta_r\right) \xrightarrow{\mathcal{D}} \chi_1^2, \text{ as } K, n \to \infty. \tag{13}$$

By the use of Theorem 3, the rejection region for the hypothesis with significance level
$\alpha\,(0 < \alpha < 1)$

$$H_0 : \beta_r = \varpi \text{ vs. } H_1 : \beta_r \ne \varpi \tag{14}$$

can be constructed as

$$\mathcal{R} = \left\{-2\log\mathcal{R}\left(\beta_r\right) \ge \chi_1^2\left(\beta\right)\right\}, \tag{15}$$

where $\chi_1^2\left(\alpha\right)$ is the upper α-th quantile of χ_1^2.

4. Selecting the Number of Block

In this analysis, we explore the optimal choice of the parameter K within the framework of the Empirical Likelihood (EL) method. Our investigation, guided by Equation (8), reveals that, when considering a fixed sample size N, smaller values of K generally lead to superior outcomes. This preference arises because the median, which effectively represents the central tendency of the dataset in practical scenarios, proves to be a robust statistic with a breakdown point of 50%, indicating its resilience against outliers. Consequently, when dealing with contaminated data, it is advisable to opt for a larger value of K.

However, it is crucial to emphasize that the performance of the Empirical Likelihood (EL) method deteriorates when the parameter K is set to small values. Therefore, to guarantee accurate estimation and inference, we employ separate approaches to select the appropriate value of K for both the estimation and inference stages.

If we are interested in the point estimator (8), we can proceed as follows. When the data are not contaminated, we adopt the suggestion by Lugosi and Mendelson (2019), $K = 8\lceil\log(1/\delta)\rceil$ $(0 <$

$\delta < 1$), where $\lceil a \rceil$ is the largest integer not greater than a. In practice, δ comes from the uniform distribution on $(0, 1)$. To eliminate the random effect of K, we replicate 500 times and get their mean as our final choice of K. When the data are contaminated, we set $K = \lceil 0.04N \rceil$.

If we are interested in inference (14), we can proceed as follows. We note that the accuracy of the estimator in each block increases as n increases. However, the power of EL increases as K becomes larger. Hence, we propose one information criteria, AAIC which is analogous to adjusted AIC (AAIC) [Akaike (1998)],

$$AAIC = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\beta}_{r,n}^{(k)} - \beta_r\right)^2 + \frac{m}{K},$$

where $m = \lceil N/K \rceil$. In this paper the above information criteria are minimized over $K \in [K_{low}, K_{upp}]$. Here we set K_{low} to be 30, which is the usual smallest sample size for well EL performing, and $K_{upp} = \lceil N/S_X \rceil$, where $S_X = S \max\{1, KU_X/3\}$ with KU_X an estimator of kurtosis. The adjusted factor 3 is the kurtosis of normal distribution. Here S is a specific constant, such as 50 or 100. When KU_X is higher, we get larger m to improve the accuracy of each block estimation. This is consistent with the belief that each block should contain more data if the skewness and kurtosis of distribution are bigger. K_{opt} is obtained by minimizing $AAIC$.

5. Simulation Study

In this section, we conduct a finite sample analysis to evaluate the empirical performance of our method. To do so, we generate data from two separate distributions according to the following procedure:

1. Gamma distribution: $Gamma(3, 1)$,
2. Log-normal distribution: $Log - \mathcal{N}(0, 1)$.

We conduct 500 Monte Carlo simulations to ensure reliable results. The findings of this subsection are presented through three illustrative examples.

Example 1 This example is used to estimate β_r. We compare the method in Section 2 (MoM, median of PWM) given by formula (5), with the traditional method (TM, i.e, the empirical version of β_r defined by formula (2) with using full data) by the average square error (ASE) criterion:

$$ASE = \frac{1}{500}\sum_{j=1}^{500}\left(\beta_{r,n}^{(j)} - \beta_r\right)^2,$$

where $\beta_{r,n}^{(j)}$ is the estimator of β_r based on the j-th sample. To analyze the sensitivity of two methods against outliers, we contaminate each sample by adding $\rho\%$ of χ^2_{100} observations, where $\rho \in \{0.5, 1\}$. We set $N \in \{600, 1200, 1800, 2400, 3000\}$. Table 1 presents the results.

We have the following comments.

1. The ASEs of two methods are small as the sample size increases.
2. When the data are contaminated, TM does not work since its ASEs are bigger, which implies that TM is very sensitive to outliers. But, MoM has good performance. Hence, our proposed method is better than TM.

Table 1 ASE for Gamma and Log-normal distributions in Example 1

		Gamma				Log-normal			
		$\rho = 0.5\%$		$\rho = 1\%$		$\rho = 0.5\%$		$\rho = 1\%$	
N	r	TM	MoM	TM	MoM	TM	MoM	TM	MoM
600	1	0.4794	0.0188	0.9566	0.0177	0.4861	0.0527	0.9702	0.0508
	2	0.4751	0.0200	0.9453	0.0192	0.4828	0.0564	0.9598	0.0558
	3	0.4723	0.0206	0.9385	0.0189	0.4781	0.0594	0.9494	0.0586
	4	0.4698	0.0204	0.9311	0.0199	0.4745	0.0609	0.9426	0.0592
1200	1	0.4790	0.0182	0.9561	0.0170	0.4864	0.0526	0.9707	0.0505
	2	0.4751	0.0195	0.9465	0.0186	0.4825	0.0558	0.9595	0.0545
	3	0.4724	0.0197	0.9381	0.0194	0.4782	0.0592	0.9504	0.0578
	4	0.4696	0.0200	0.9306	0.0194	0.4751	0.0605	0.9423	0.0594
1800	1	0.4791	0.0157	0.9563	0.0147	0.4861	0.0468	0.9701	0.0454
	2	0.4755	0.0170	0.9462	0.0163	0.4818	0.0507	0.9795	0.0495
	3	0.4724	0.0173	0.9384	0.0165	0.4782	0.0533	0.9503	0.0520
	4	0.4697	0.0179	0.9308	0.0173	0.4754	0.5488	0.9417	0.0535
2400	1	0.4794	0.0115	0.9562	0.0107	0.4864	0.0395	0.9702	0.0348
	2	0.4755	0.0125	0.9463	0.0120	0.4818	0.0397	0.9598	0.0379
	3	0.4723	0.0131	0.9378	0.0126	0.4782	0.0417	0.9504	0.0401
	4	0.4698	0.0131	0.9307	0.0129	0.4751	0.0429	0.9416	0.0417
3000	1	0.4791	0.0116	0.9566	0.0105	0.4866	0.0294	0.9706	0.0281
	2	0.4753	0.0126	0.9464	0.0119	0.4821	0.0323	0.9595	0.0313
	3	0.4722	0.0128	0.9381	0.0123	0.4786	0.0341	0.9502	0.0330
	4	0.4696	0.0133	0.9309	0.0125	0.4753	0.0352	0.9419	0.0342

Example 2 This example is for inference on β_r. We set $K_{low} = 30$, choose sample sizes $N \in \{2000, 3000, ..., 7000\}$ for $S = 50$, and $N \in \{3000, 4000, ..., 7000\}$ for $S = 100$. The nominal significance level is 0.05. We compared our proposed method with the normal approximation method. However, the traditional approximation method performs badly even when we use the known $\sigma^2(F)$, not its estimator. Hence, we only report the empirical size and power of our proposed method. Furthermore, we also report ASE and average K (AK) in the empirical size. For the power, we consider $\beta_r + \theta$ with $\theta \in \{0.1, 0.3, 0.5\}$ as the alternative hypothesis. The simulations are displayed in Tables 2-3. The results with $S = 100$ are better than $S = 50$ since its ASE is slightly smaller. The size of the proposed test is closer to 0.05 as N increases.

Table 2 Empirical size, AK and ASE for Gamma distribution in Example 2

		S=50			S=100		
N	r	Size	AK	ASE	Size	AK	ASE
2000	1	0.063	31.25	0.0177			
	2	0.063	32.58	0.0179			
	3	0.065	32.89	0.0179			
	4	0.070	33.04	0.0178			
3000	1	0.058	34.17	0.0171	0.072	34.25	0.0148
	2	0.059	34.68	0.0177	0.079	34.58	0.0152
	3	0.062	35.25	0.0179	0.078	34.89	0.0154
	4	0.065	35.78	0.0181	0.077	35.47	0.0155
4000	1	0.056	37.76	0.0167	0.076	38.27	0.0144
	2	0.058	38.25	0.0169	0.078	38.58	0.0148
	3	0.059	38.58	0.0171	0.079	38.75	0.0152
	4	0.062	38.98	0.0178	0.081	39.07	0.0153
5000	1	0.053	41.27	0.0163	0.073	41.41	0.0142
	2	0.055	42.24	0.0165	0.074	41.87	0.0145
	3	0.059	43.87	0.0166	0.076	42.17	0.0147
	4	0.059	43.98	0.0171	0.079	43.11	0.0149
6000	1	0.050	44.61	0.0150	0.069	45.14	0.0141
	2	0.053	45.25	0.0157	0.071	44.58	0.0143
	3	0.055	45.98	0.0158	0.072	45.67	0.0144
	4	0.058	46.51	0.0161	0.075	47.27	0.0145
7000	1	0.051	45.69	0.0148	0.066	49.35	0.0140
	2	0.055	46.56	0.0152	0.068	51.25	0.0139
	3	0.057	47.67	0.0154	0.068	53.24	0.0138
	4	0.052	48.35	0.0158	0.071	55.54	0.0137

Table 3 Empirical size, AK and ASE for Log-Normal distribution in Example 2

		S=50			S=100		
N	r	Size	AK	ASE	Size	AK	ASE
2000	1	0.063	33.25	0.037			
	2	0.065	34.21	0.038			
	3	0.064	34.98	0.039			
	4	0.067	35.47	0.041			
3000	1	0.059	34.87	0.035	0.069	31.25	0.032
	2	0.063	36.01	0.037	0.074	32.27	0.033
	3	0.059	37.25	0.038	0.068	33.98	0.032
	4	0.069	38.54	0.040	0.071	34.17	0.034
4000	1	0.067	36.46	0.033	0.077	33.21	0.031
	2	0.071	37.13	0.035	0.079	34.25	0.033
	3	0.074	38.51	0.036	0.081	35.67	0.034
	4	0.072	39.21	0.038	0.078	36.69	0.035
5000	1	0.066	39.77	0.031	0.075	36.86	0.029
	2	0.072	40.07	0.034	0.078	38.25	0.032
	3	0.071	41.25	0.035	0.078	39.46	0.033
	4	0.074	42.21	0.036	0.076	40.09	0.034
6000	1	0.074	41.61	0.031	0.089	38.14	0.028
	2	0.077	42.58	0.033	0.091	39.98	0.030
	3	0.078	44.05	0.035	0.092	41.25	0.032
	4	0.079	45.78	0.035	0.091	43.37	0.033
7000	1	0.078	43.69	0.030	0.088	40.25	0.027
	2	0.079	44.62	0.033	0.092	41.35	0.029
	3	0.075	45.25	0.034	0.090	42.47	0.030
	4	0.077	47.04	0.034	0.093	44.17	0.031

Table 4 Empirical power for Gamma distribution in Example 2

		S=50			S=100		
N	r	0.1	0.3	0.5	0.1	0.3	0.5
2000	1	0.312	0.575	0.859			
	2	0.317	0.605	0.874			
	3	0.324	0.632	0.905			
	4	0.326	0.652	0.911			
3000	1	0.345	0.678	0.942	0.368	0.665	0.925
	2	0.352	0.695	0.961	0.355	0.685	0.958
	3	0.358	0.725	0.974	0.357	0.704	0.975
	4	0.367	0.774	0.983	0.364	0.745	0.988
4000	1	0.472	0.741	0.989	0.773	0.851	0.978
	2	0.513	0.754	0.998	0.796	0.856	0.997
	3	0.547	0.759	1.00	0.812	0.862	1.00
	4	0.559	0.783	1.00	0.784	0.884	1.00
5000	1	0.668	0.795	1.00	0.845	0.921	1.00
	2	0.727	0.825	1.00	0.879	0.935	1.00
	3	0.718	0.847	1.00	0.914	0.941	1.00
	4	0.746	0.867	1.00	0.921	0.947	1.00
6000	1	0.745	0.846	1.00	0.897	0.975	1.00
	2	0.774	0.869	1.00	0.914	0.989	1.00
	3	0.781	0.872	1.00	0.925	1.00	1.00
	4	0.793	0.897	1.00	0.932	1.00	1.00
7000	1	0.785	0.916	1.00	0.951	1.00	1.00
	2	0.792	0.934	1.00	0.962	1.00	1.00
	3	0.752	0.947	1.00	0.971	1.00	1.00
	4	0.779	0.978	1.00	0.945	1.00	1.00

Table 5 Empirical power for Log-Normal distribution in Example 2

		S=50			S=100		
N	r	0.1	0.3	0.5	0.1	0.3	0.5
2000	1	0.347	0.575	0.858			
	2	0.365	0.605	0.863			
	3	0.384	0.632	0.911			
	4	0.395	0.652	0.913			
3000	1	0.411	0.678	0.944	0.804	0.945	0.984
	2	0.452	0.695	0.963	0.825	0.957	0.997
	3	0.458	0.725	0.971	0.836	0.961	1.00
	4	0.472	0.774	0.975	0.847	0.975	1.00
4000	1	0.545	0.768	0.979	0.875	0.973	1.00
	2	0.558	0.764	0.989	0.895	0.984	1.00
	3	0.529	0.774	1.00	0.914	0.998	1.00
	4	0.573	0.788	1.00	0.919	1.00	1.00
5000	1	0.621	0.811	1.00	0.945	1.00	1.00
	2	0.662	0.829	1.00	0.958	1.00	1.00
	3	0.674	0.848	1.00	0.978	1.00	1.00
	4	0.074	0.862	1.00	0.981	1.00	1.00
6000	1	0.774	0.849	1.00	0.984	1.00	1.00
	2	0.787	0.859	1.00	0.998	1.00	1.00
	3	0.785	0.869	1.00	1.00	1.00	1.00
	4	0.795	0.887	1.00	1.00	1.00	1.00
7000	1	0.878	0.927	1.00	1.00	1.00	1.00
	2	0.868	0.947	1.00	1.00	1.00	1.00
	3	0.814	0.958	1.00	1.00	1.00	1.00
	4	0.845	0.977	1.00	1.00	1.00	1.00

Example 3 For testing the value of β_r we fix the sample size as $\{10^4, 10^5, 10^6\}$. For simplify of calculations, we fix $K \in \{30, 60\}$. We report ASE and empirical size. We conserve all other settings as those in Example 2. From Table 6, the results with $K = 30$ are better than $K = 60$ since smaller K produces a large sample size for each block, which makes the block estimator more accurate. On the other hand, the block size $K = 30$ is enough to make EL perform satisfactorily.

Table 6 Empirical size for Example 3

			K=30		K=60	
Distribution	N	r	ASE	Size	ASE	Size
Gamma	10^4	1	0.0101	0.059	0.0121	0.065
		2	0.0113	0.055	0.0123	0.068
		3	0.0118	0.057	0.0125	0.062
		4	0.0121	0.056	0.0128	0.063
	10^5	1	0.0097	0.044	0.0119	0.059
		2	0.0110	0.047	0.0117	0.057
		3	0.0114	0.049	0.0121	0.055
		4	0.0117	0.048	0.0124	0.057
	10^6	1	0.0096	0.047	0.0119	0.051
		2	0.0109	0.049	0.0121	0.054
		3	0.0111	0.044	0.0123	0.052
		4	0.0119	0.051	0.0126	0.055
Log-Normal	10^4	1	0.0264	0.051	0.0281	0.059
		2	0.0311	0.054	0.0318	0.057
		3	0.0327	0.053	0.0339	0.055
		4	0.0331	0.049	0.0345	0.052
	10^5	1	0.0261	0.053	0.0278	0.048
		2	0.0309	0.057	0.0315	0.046
		3	0.0330	0.055	0.0337	0.049
		4	0.0328	0.052	0.0342	0.047
	10^6	1	0.0259	0.047	0.0275	0.051
		2	0.0307	0.051	0.0314	0.052
		3	0.0326	0.049	0.0336	0.055
		4	0.0327	0.050	0.0338	0.049

6. Conclusion

The present paper proposed a new and robust non-parametric estimator for The PWM based on a grouping strategy. The concept of grouping (or Clustering) can be applied, for example, to customer segmentation according to purchase history, interests, or activity monitoring; to Insurance Fraud Detection where it is possible to isolate new claims based on their proximity to clusters that indicate fraudulent patterns, furthermore, Clustering time series data into a specified number of groups based on their similarity is an initial step for further analysis in water management analytic. The asymptotic properties, including consistency and asymptotic normality of the proposed estimator, are obtained. Due to the complexity of the variance term in the normal approximation of the proposed estimator, we constructed a new test for PWM based on the empirical likelihood method for the median. Numerical simulations confirmed that our new proposed estimator is quite robust with respect to outliers. With the rise in prominence of high dimensional data, it might be interesting to generalize the MoM in high dimension (Geometric median or L_1 median). Another possible generalization of the proposed estimator is to consider a bootstrap version where several random splits are performed.

Acknowledgments

We would like to thank the anonymous referees for their helpful suggestions and comments that resulted in a number of clarifications and additions to the original submission.

References

Akaike H. Information theory and an extension of the maximum likelihood principle. In Selected papers of hirotugu akaike. New York: Springer; 1998. pp. 199-213.

Alon N, Matias Y, Szegedy M. The space complexity of approximating the frequency moments. In Proceedings of the twenty-eighth annual ACM symposium on Theory of computing; 1996. p. 20-29.

Audibert JY, Catoni, O. Robust linear least squares regression. Ann Stat. 2011; 39(5): 2766-2794.

Bhati D, Kattumannil SK, Sreelakshmi N. Jackknife empirical likelihood based inference for probability weighted moments. J Korean Stat Soc. 2021; 50, 98-116.

Blair C. (1985). Problem complexity and method efficiency in optimization (as nemirovsky and db yudin). Siam Review. 1985; 27(2): 264.

Bubeck S, Cesa-Bianchi N, Lugosi G. Bandits with heavy tail. IEEE Trans Inf Theory. 2013; 59(11), 7711-7717.

Chen J, Variyath AM, Abraham B. Adjusted empirical likelihood and its properties. J Comput Graph Stat. 2008; 17(2): 426-443.

David HA, Nagaraja HN. Order Statistics. 3rd ed. New York: John Wiley and Sons; 2003.

Devroye L, Lerasle M, Lugosi G, Oliveira RI. Sub-Gaussian mean estimators. Ann Stat. 2016; 44(6): 2695-2725.

Greenwood JA, Landwehr JM, Matalas NC, Wallis JR. Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. Water Resour Res. 1979; 15(5): 1049-1054.

Hoeffding W. A class of statistics with asymptotically normal distribution. In: Breakthroughs in Statistics: Foundations and Basic Theory. Springer; 1992. p. 308-334.

Hosking JRM, Wallis JR, Wood EF. Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. Technometrics. 1985; 27(3): 251-261.

Hsu D, Sabato S. (2016). Loss minimization and parameter estimation with heavy tails. J Mach Learn Res. 2016; 17(1): 543-582.

Jerrum MR, Valiant LG, Vazirani VV. Random generation of combinatorial structures from a uniform distribution. Theor Comput Sci. 1986; 43: 169-188.

Jiang H, Zhao Y. Transformed jackknife empirical likelihood for probability weighted moments. J Stat Comput Simul. 2022; 92(8): 1618-1639.

Jing BY, Yuan J, Zhou W. Jackknife empirical likelihood. J Am Stat Assoc. 2009; 104(487): 1224-1232.

Mohamed L, Abdelaziz R. A robust estimator of the S-Gini index for massive data. Commun Stat Simulat. 2023; 52(8): 3502-3519.

Lerasle M, Oliveira RI. Robust empirical mean estimators. arXiv preprint arXiv:1112.3914. 2011.

Lugosi G, Mendelson S. (2019). Sub-Gaussian estimators of the mean of a random vector. Ann Stat. 2019; 47(2): 783-794.

Ma X, Wang S, Zhou W. (2022). Statistical inference in massive datasets by empirical likelihood. Comput Stat. 2021; 37: 1143-1164.

Minsker S. Geometric median and robust estimation in Banach spaces. Bernoulli. 2015; 21(4): 2308-2335.

Owen A. Empirical likelihood ratio confidence regions. Ann Stat. 1990; 18(1): 90-120.

Owen AB. (2001). Empirical likelihood. Chapman & Hall: CRC press; 2001.

Vexler A, Zou L, Hutson AD. An extension to empirical likelihood for evaluating probability weighted moments. J Stat Plan Infer. 2017; 182: 50-60.

Vexler A, Zou L. Empirical likelihood ratio tests with power one. Stat Probab Lett. 2018; 140: 160-166.

Zhao Y, Meng X, Yang H. Jackknife empirical likelihood inference for the mean absolute deviation. Comput Stat Data Anal. 2015; 91: 92-101.

Appendix: Proofs

Usefull Lemma

The following Lemma, known as Berry-Esseen bound (see, Vexler et al. (2017) and Laidi and Rassoul (2021)), is used in the proof of Theorems 1, 2 and 3.

Lemma 1 *We assume that $E(|X|^3) < \infty$ and $\sigma^2(F) > 0$ which is defined in (6). If F has a strictly positive, continuous density function f on $[-\eta, \eta]$ for some $\eta > 0$, then there exists a constant $C > 0$ such that, for*

$$\sup_{x\in\mathbb{R}} \left\{ P \left| \left(\frac{\sqrt{N}}{\sigma(F)} \left(\hat{\beta}_{r,N} - \beta_r \right) \leq x \right) - \Phi\left(x\right) \right| \right\} \leq \frac{C}{\sqrt{N}} \tag{16}$$

where Φ is the cumulative function of a standard normal random variable.

We now prove one after the other theorems 1, 2 and 3.

Proof: [Proof of theorem 1] Let $(\eta_{n,j})_{j=1,...,K}$ the random variables given by:

$$\eta_{n,j} := \frac{\sqrt{n}}{\sigma(F)} \left(\hat{\beta}_{r,n}^{(j)} - \beta_r \right), j = 1, ..., K \tag{17}$$

For each $j = 1, ..., K$, we have by inequality (16)

$$\sup_{x\in\mathbb{R}} \left\{ \left| \left(\eta_{n,j} \leq x \right) - \Phi\left(x\right) \right| \right\} \leq \frac{C}{\sqrt{n}}.$$

Let $x = \sqrt{n}z/\sigma(F)$, for all $z > 0$, we get

$$P \left(\left(\hat{\beta}_{r,n}^{(j)} - \beta_r \right) \geq z \right) \leq \frac{C}{\sqrt{n}} + 1 - \Phi\left(\frac{\sqrt{n}z}{\sigma(F)} \right)$$

Using the inequality

$$1 - \Phi\left(\frac{\sqrt{n}z}{\sigma(F)} \right) \leq e^{-\frac{nz^2}{2\sigma^2(F)}},$$

which becomes insignificant relative to $(1/\sqrt{n})$ as n approaches infinity and fixed $z > 0$. Consequently

$$P \left(\hat{\beta}_{r,n}^{(j)} - \beta_r \geq z \right) \leq \frac{C}{2\sqrt{n}},$$

and by the same way we have

$$P \left(\hat{\beta}_{r,n}^{(j)} - \beta_r \leq -z \right) \leq \frac{C}{2\sqrt{n}},$$

where C is a constant which depends only on z. Therefore, we can write

$$P \left(\left| \beta_{r,n}^{(j)} - \beta_r \right| \geq z \right) \leq \frac{C}{\sqrt{n}}. \tag{18}$$

We assert that

$$\left| \tilde{\beta}_r^{MoM} - \beta_r \right| \leq Median \left\{ \left| \beta_{r,n}^{(j)} - \beta_r \right|, j = 1, 2, ..., K \right\}. \tag{19}$$

In fact,

$$\tilde{\beta}_r^{MoM} - \beta_r = Median\left\{\hat{\beta}_{r,n}^{(j)} - \beta_r, j = 1, 2, ..., K\right\}.$$

knowing that

$$Median\left\{\hat{\beta}_{r,n}^{(j)} - \beta_r, j = 1, 2, ..., K\right\} \leq Median\left\{\left|\hat{\beta}_{r,n}^{(j)} - \beta_r\right|, j = 1, 2, ..., K\right\}$$

yields

$$\tilde{\beta}_r^{MoM} - \beta_r \leq Median\left\{\left|\hat{\beta}_{r,n}^{(j)} - \beta_r\right|, j = 1, 2, ..., K\right\}.$$

by the same manner, it can be proved that

$$-\left(\tilde{\beta}_r^{MoM} - \beta_r\right) \leq Median\left\{\left|\hat{\beta}_{r,n}^{(j)} - \beta_r\right|, j = 1, 2, ..., K\right\}.$$

This gives proof to (19). Thus, we have:

$$\begin{aligned} P\left(\left|\tilde{\beta}_r^{MoM} - \beta_r\right| \geq z\right) &\leq P\left(Median\left\{\left|\hat{\beta}_{r,n}^{(j)} - \beta_r\right|, j = 1, 2, ..., K\right\} \geq z\right) \\ &: \quad = P(E). \end{aligned}$$

Let $(\theta_j)_{j=1,2,...,K}$ the Bernoulli random variables given by

$$\theta_j = I\left(\left|\hat{\beta}_{r,n}^{(j)} - \beta_r\right| \geq z\right), j = 1, 2, ..., K$$

where $E(\theta_j) \leq C/\sqrt{n}$. It is evident that the event E occurs if and only if $\sum_{j=1}^{K} \theta_j$ is greater than $K/2$. Using Chernoff's inequality gives

$$P(E) = P\left(\sum_{j=1}^{K} \theta_j \geq \frac{K}{2}\right) \leq e^{-KE(\theta_1)}\left(\frac{2eKC}{K\sqrt{n}}\right)^{K/2} \leq \frac{C}{n^{K/5}},$$

This concludes the proof of Theorem 1.

The next Lemma (2) is the a Central Limit Theorem version for partial sums. Before stating the Lemma, let us define, for arbitrary and fixed x, the independent and identically distributed Bernoulli random variables

$$\xi_{n,j}(x) := I(\eta_{n,j} \leq x), j = 1, 2, ..., K$$

and let $p_n(x) = P(\eta_{n,j} \leq x)$. By equation 16, we can write

$$|p_n(x) - \Phi(x)| = O\left(1/\sqrt{n}\right)$$

for all $x \in \mathbf{R}$.

Lemma 2 *Assume that $n/K \to \infty$ as $K \to \infty$. Then, for any fixed real x, we have*

$$\sqrt{K}\left(\frac{1}{K}\sum_{j=1}^{K} \xi_{n,j}(x) - \Phi(x)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Phi(x)(1 - \Phi(x))).$$

In Particular, as $K \to \infty$, we have also

$$\sqrt{K}\left(\frac{1}{K}\sum_{j=1}^{K} \xi_{n,j}\left(x/\sqrt{K}\right) - \frac{1}{2} - \frac{x}{\sqrt{2\pi K}}\Phi(x)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1/4). \quad (20)$$

Proof: For arbitrary real t, and $i^2 = -1$, the independence hypothesis gives

$$E\left(e^{it\sqrt{K}\left(\frac{1}{K}\sum_{j=1}^{K}\xi_{n,j}(x)-\Phi(x)\right)}\right) = \left(Ee^{it\frac{1}{\sqrt{K}}|\xi_{n,j}(x)-\Phi(x)|}\right)^{K},$$

then and using the fact that$|p_n - \Phi(x)| = O\left(1/\sqrt{n}\right)$, $n/K \to \infty$, $K \to \infty$ and

$$\left|\frac{p_n}{\sqrt{K}}\left(1-\Phi(x)\right) + \frac{1-p_n}{\sqrt{K}}\Phi(x)\right| = \left|\frac{p_n-\Phi(x)}{\sqrt{K}}\right| = o\left(1/K\right)$$

and by Taylor's expansion, we can write

$$\begin{aligned} Ee^{it\frac{1}{\sqrt{K}}(\xi_{n,j}(x)-\Phi(x))} &= p_n e^{it\frac{1}{\sqrt{K}}(1-\Phi(x))} + (1-p_n)\, e^{-it\frac{1}{\sqrt{K}}\Phi(x)} \\ &= 1 + it\frac{p_n}{\sqrt{K}}\left(1-\Phi(x)\right) - it\frac{(1-p_n)}{\sqrt{K}}\Phi(x) \\ &\quad -\frac{p_n}{2K}\left[t\left(1-\Phi(x)\right)\right]^2 - \frac{(1-p_n)}{2K}\left[t\Phi(x)\right]^2 + o\left(K^{-1}\right) \\ &= 1 - \frac{p_n}{2K}\left[\Phi(x)\left(1-\Phi(x)\right)\right] + o\left(K^{-1}\right), \end{aligned} \tag{21}$$

(21) leads to the first result of the Lemma.

Using the same reasoning as above, the last result of Lemma can be concluded by replacing x with $x/\sqrt{K}$ and remark that

$$\Phi\left(x/\sqrt{K}\right) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_0^{x/\sqrt{K}} e^{-u^2/2}du = \frac{1}{2} + \frac{x}{\sqrt{2\pi K}} + o\left(K^{-1/2}\right).$$

Slutsky's theorem finalize the proof.

Proof: [Proof of Theorem 2]

1. The first result follows from equation 16, the continuous mapping theorem and the fact that the Median function is continuous.

2. We know that

$$\begin{aligned} \frac{\sqrt{N}}{\sigma(F)}\left(\tilde{\beta}_r^{MoM} - \beta_r\right) &= \sqrt{K}\frac{\sqrt{n}}{\sigma(F)}\left(\tilde{\beta}_r^{MoM} - \beta_r\right) \\ &= \sqrt{K}Median\left\{\eta_{n,j}, j=1,...,K\right\}. \end{aligned} \tag{22}$$

Now for an arbitrary real x and assuming that K is an odd integer, one have

$$\begin{aligned} &P\left(\sqrt{K}Median\left\{\eta_{n,j}, j=1,...,K\right\} \le x\right) \\ =\ &P\left(\sum_{j=1}^{K} I\left(\eta_{n,j} \le \frac{x}{\sqrt{K}}\right) \ge \frac{K+1}{2}\right) \\ =\ &P\left(\sum_{j=1}^{K}\xi_{n,j}\left(\frac{x}{\sqrt{K}}\right) \ge \frac{K+1}{2}\right) \\ =\ &P\left(\sqrt{K}\left\{\frac{1}{K}\sum_{j=1}^{K}\xi_{n,j}\left(\frac{x}{\sqrt{K}}\right) - \frac{1}{2} - \frac{x}{\sqrt{2\pi K}}\right\} \ge -\frac{x}{\sqrt{2\pi}} + O\left(\frac{1}{\sqrt{K}}\right)\right) \end{aligned}$$

which goes to $\Phi\left(\sqrt{2/\pi x}\right)$ according to the above lemma 2. Else if K is an even number, then

$$P\left(\sqrt{K}Median\left\{\eta_{n,j}, j=1,...,K\right\} \le x\right) \ge P\left(\sum_{j=1}^{K} I\left(\eta_{n,j} \le \frac{x}{\sqrt{K}}\right) \ge \frac{K}{2}+1\right)$$

and

$$P\left(\sqrt{K}Median\left\{\eta_{n,j}, j=1,...,K\right\} \le x\right) \le P\left(\sum_{j=1}^{K} I\left(\eta_{n,j} \le \frac{x}{\sqrt{K}}\right) \ge \frac{K}{2}\right)$$

The fact that the right-hand sides of the last two inequalities tend to $\Phi\left(\sqrt{2/\pi x}\right)$ when $K \to \infty$ conclude the proof of Theorem 2.

Proof: [Proof of Theorem 3] First, we know that

$$Z_{n,k} = I\left(\hat{\beta}_{r,n}^{(k)} - \beta_r\right), \text{ for } k = 1, 2, ..., K.$$

(11) can be written as

$$f(\lambda) = \frac{1}{K}\sum_{j=1}^{K} \frac{Z_{n,k} - 0.5}{1 + \lambda\left(Z_{n,k} - 0.5\right)} = 0. \tag{23}$$

By (23), we obtain that

$$\bar{Z}_{n,k} - 0.5 = \lambda\overline{S}, \tag{24}$$

where

$$U_{n,k} = \lambda\left(Z_{n,k} - 0.5\right),$$

$$\overline{S} = \frac{1}{K}\sum_{k=1}^{K} \frac{\left(Z_{n,k} - 0.5\right)^2}{1 + U_{n,k}}$$

and

$$\bar{Z}_{n,k} = \frac{1}{K}\sum_{k=1}^{K} Z_{n,k}.$$

Furthermore

$$S = \frac{1}{K}\sum_{k=1}^{K} \left(Z_{n,k} - 0.5\right)^2 = 0.25,$$

and

$$Z_K = \max_{1\le k\le K} \left|Z_{n,k} - 0.5\right| = 0.5.$$

Under the condition $\omega_i > 0$, we can deduce that $1 + U_{n,k} > 0$, and

$$\lambda S \le \lambda\overline{S}\left(1 + \max_{1\le k\le K} U_{n,k}\right) \le \lambda\overline{S}\left(1 + \lambda Z_K\right) = \left(\bar{Z}_{n,k} - 0.5\right)\left(1 + \lambda Z_K\right)$$

The last equality follows by (24). Hence,

$$\lambda\left[S - \left(\bar{Z}_{n,k} - 0.5\right)Z_K\right] \le \bar{Z}_{n,k} - 0.5.$$

According to Lemma 2, $\bar{Z}_{n,k} - 0.5 = O_p\left(1/\sqrt{K}\right)$. Then

$$\lambda\left[0.25 - O_p\left(1/\sqrt{K}\right)\right] = O_p\left(1/\sqrt{K}\right).$$

That is $\lambda = O_p\left(1/\sqrt{K}\right)$. Furthermore, we have

$$\max_{1\leq k\leq K} |U_{n,k}| = O_p\left(1/\sqrt{K}\right) = o_p(1).$$

The expansion of (23) leads to

$$\begin{aligned}
0 &= \frac{1}{K}\sum_{k=1}^{K}\frac{(Z_{n,k}-0.5)}{1+U_{n,k}} \\
&= \frac{1}{K}\sum_{k=1}^{K}(Z_{n,k}-0.5)\left(1-U_{n,k}+\frac{U_{n,k}^2}{1+U_{n,k}}\right) \\
&= \left(\bar{Z}_{n,k}-0.5\right)-\lambda S+\frac{1}{K}\sum_{k=1}^{K}\frac{(Z_{n,k}-0.5)}{1+U_{n,k}}U_{n,k}^2 \qquad (25)
\end{aligned}$$

whereas the norm of the last term in (25) is bounded by

$$\frac{1}{K}\sum_{k=1}^{K}\lambda^2\frac{|Z_{n,k}-0.5|^3}{1+U_{n,k}} = O(1)\,O_p(1/K)\,O_p(1) = o_p\left(1/\sqrt{K}\right).$$

Solving (25) for λ gives
$\lambda = S^{-1}\left(\bar{Z}_{n,k}-0.5\right)+\beta = 4\left(\bar{Z}_{n,k}-0.5\right)+\beta$, with $\beta = o_p\left(1/\sqrt{K}\right)$.
Applying Taylor expansion in (25) gives us

$$\log(1+U_{n,k}) = U_{n,k}-\frac{1}{2}U_{n,k}^2+\eta_k$$

and for some finite $B > 0, 1 \leq k \leq K$,

$$P\left(|\eta_k| \leq B\,|U_{n,k}|^3\right) \to 1, \text{ as } K\to\infty \text{ and } m\to\infty.$$

Simple calculation leads to

$$\begin{aligned}
-2\log\mathcal{R}(\beta_r) &= 2\sum_{k=1}^{K}\log(1+U_{n,k}) \\
&= 2\sum_{k=1}^{K}\left(U_{n,k}-\frac{1}{2}U_{n,k}^2+\eta_k\right) \\
&= 4K\left(\bar{Z}_{n,k}-0.5\right)^2-4K\beta^2+2\sum_{k=1}^{K}\eta_k
\end{aligned}$$

Lemma 2 ensures the convergence $4K\left(\bar{Z}_{n,k}-0.5\right)^2 \overset{\mathcal{D}}{\to} \chi_1^2$.
Given that

$$4K\beta^2 = 4Ko_p(1/K) = o_p(1),$$

and

$$\left|\sum_{k=1}^{K}\eta_k\right| \leq B\,|\lambda|^3\sum_{k=1}^{K}|Z_{n,k}-0.5|^3 = O_p\left(1/\sqrt{K^3}\right)O(1) = o_p(1).$$

Theorem 3 is thus proved.