Robust estimations of moments for unimodal distributions

Tuban Lee

This manuscript was compiled on June 10, 2023

A. Invariant Moments. All popular robust location estimators, such as the symmetric trimmed mean, symmetric Winsorized mean, Hodges-Lehmann estimator, Huber M-estimator, and median of means, are symmetric. As shown in RSSM I, a γ-weighted Hodges-Lehmann mean ($\mathrm{WHLM}_{k,\epsilon,\gamma}$) can achieve consistency for the population mean in any γ-symmetric distribution with a finite mean. However, it falls considerably short of consistently handling other parametric distributions that are not γ-symmetric. Shifting from semiparametrics to parametrics, consider a robust estimator with a non-sample-dependent breakdown point (defined in Subsection **??**) which is consistent simultaneously for both a semiparametric distribution and a parametric distribution that does not belong to that semiparametric distribution, it is named with the prefix 'invariant' followed by the name of the population parameter it is consistent with. Here, the recombined I-statistic is defined as

$$\mathrm{RI}_{d,\mathbf{k}_1,\mathbf{k}_2,k_1,k_2,\epsilon_1,\epsilon_2,\gamma_1,\gamma_2,n,LU_1,LU_2} := \lim_{c\to\infty}\left(\frac{\left(LU_{1\mathbf{k}_1,k_1,\epsilon_1,\gamma_1,n}+c\right)^{d+1}}{\left(LU_{2\mathbf{k}_2,k_2,\epsilon_2,\gamma_2,n}+c\right)^{d}}-c\right),$$

where d is the key factor for bias correction, $LU_{\mathbf{k},k,\epsilon,\gamma,n}$ is the LU-statistic, $\mathbf{k}$ is the degree of the U-statistic, k is the degree of the LL-statistic, ϵ is the upper asymptotic breakdown point of the LU-statistic. It is assumed in this series that in the subscript of an estimator, if $\mathbf{k}$, k and γ are omitted, $\mathbf{k}=1$, $k=1$, $\gamma=1$ are assumed, if just one γ is indicated, $\gamma_1=\gamma_2$, if n is omitted, only the asymptotic behavior is considered, in the absence of subscripts, no assumptions are made. The subsequent theorem shows the significance of a recombined I-statistic.

Theorem A.1. *Define the recombined mean as* $rm_{d,k_1,k_2,\epsilon_1,\epsilon_2,\gamma_1,\gamma_2,n,WL_1,WL_2} := RI_{d,\mathbf{k}_1=1,\mathbf{k}_2=1,k_1,k_2,\epsilon_1,\epsilon_2,\gamma_1,\gamma_2,n,LU_1=WL_1,LU_2=WL_2}$. *Assuming finite means,* $rm_{d=\frac{\mu-WL_{1k_1,\epsilon_1,\gamma_1}}{WL_{1k_1,\epsilon_1,\gamma_1}-WL_{2k_2,\epsilon_2,\gamma_2}},k_1,k_2,\epsilon_1,\epsilon_2,\gamma_1,\gamma_2,WL_1,WL_2}$ *is a consistent mean estimator for a location-scale distribution, where* μ, $WL_{1k_1,\epsilon_1,\gamma_1}$, *and* $WL_{2k_2,\epsilon_2,\gamma_2}$ *are different location parameters from that location-scale distribution. If* $\gamma_1=\gamma_2$, $WL = WHLM$, rm *is also consistent for any* γ*-symmetric distributions.*

Proof. Finding d that make $rm_{d,k_1,k_2,\epsilon_1,\epsilon_2,\gamma_1,\gamma_2,\mathrm{WL}_1,\mathrm{WL}_2}$ a consistent mean estimator is equivalent to finding the solution of $rm_{d,k_1,k_2,\epsilon_1,\epsilon_2,\gamma_1,\gamma_2,\mathrm{WL}_1,\mathrm{WL}_2}=\mu$. First consider the location-scale distribution. Since $rm_{d,k_1,k_2,\epsilon_1,\epsilon_2,\gamma_1,\gamma_2,\mathrm{WL}_1,\mathrm{WL}_2} = \lim_{c\to\infty}\left(\frac{\left(\mathrm{WL}_{1k_1,\epsilon_1,\gamma_1}+c\right)^{d+1}}{\left(\mathrm{WL}_{2k_2,\epsilon_2,\gamma_2}+c\right)^{d}}-c\right) = (d+1)\,\mathrm{WL}_{1k_1,\epsilon_1,\gamma} - d\mathrm{WL}_{2k_2,\epsilon_2,\gamma} = \mu$. So, $d=\frac{\mu-\mathrm{WL}_{1k_1,\epsilon_1,\gamma_1}}{\mathrm{WL}_{1k_1,\epsilon_1,\gamma_1}-\mathrm{WL}_{2k_2,\epsilon_2,\gamma_2}}$. In RSSM I, it was established that any $\mathrm{WL}(k,\epsilon,\gamma)$ can be expressed as $\lambda\mathrm{WL}_0(k,\epsilon,\gamma)+\mu$ for a location-scale distribution parameterized by a location parameter μ and a scale parameter λ, where $\mathrm{WL}_0(k,\epsilon,\gamma)$ is a function of $Q_0(p)$, the quantile function of a standard distribution without any shifts or scaling, according to the definition of the weighted L-statistic. The simultaneous cancellation of μ and λ in $\frac{(\lambda\mu_0+\mu)-(\lambda\mathrm{WL}_{1_0}(k_1,\epsilon_1,\gamma_1)+\mu)}{(\lambda\mathrm{WL}_{1_0}(k_1,\epsilon_1,\gamma_1)+\mu)-(\lambda\mathrm{WL}_{2_0}(k_2,\epsilon_2,\gamma_2)+\mu)}$ assures that the d in rm is always a constant for a location-scale distribution. The proof of the second assertion follows directly from the coincidence property. According to Theorem 18 in RSSM I, for any γ-symmetric distribution with a finite mean, $\mathrm{WHLM}_{1k_1,\epsilon_1,\gamma}=\mathrm{WHLM}_{2k_2,\epsilon_2,\gamma}=\mu$. Then $rm_{d,k_1,k_2,\epsilon_1,\epsilon_2,\gamma,\mathrm{WHLM}_1,\mathrm{WHLM}_2} = \lim_{c\to\infty}\left(\frac{(\mu+c)^{d+1}}{(\mu+c)^{d}}-c\right)=\mu$. This completes the demonstration. □

For example, the Pareto distribution has a quantile function $Q_{Par}(p)=x_m(1-p)^{-\frac{1}{\alpha}}$, where x_m is the minimum possible value that a random variable following the Pareto distribution can take, serving a scale parameter, α is a shape parameter. The mean of the Pareto distribution is given by $\frac{\alpha x_m}{\alpha-1}$. As $\mathrm{WL}(k,\epsilon,\gamma)$ can be expressed as a function of $Q(p)$, one can set the two $\mathrm{WL}_{k,\epsilon,\gamma}$s in the d value as two arbitrary quantiles $Q_{Par}(p_1)$ and $Q_{Par}(p_2)$. For the Pareto distribution, $d_{Per}=\frac{\mu_{Per}-Q_{Par}(p_1)}{Q_{Par}(p_1)-Q_{Par}(p_2)}=\frac{\frac{\alpha x_m}{\alpha-1}-x_m(1-p_1)^{-\frac{1}{\alpha}}}{x_m(1-p_1)^{-\frac{1}{\alpha}}-x_m(1-p_2)^{-\frac{1}{\alpha}}}$. x_m can be canceled out.

Significance Statement

Bias, variance, and contamination are the three main errors in statistics. Consistent robust estimation is unattainable without parametric assumptions. In this article, invariant moments are proposed as a means of achieving near-consistent and robust estimations of moments, even in scenarios where moderate violations of distributional assumptions occur, while the variances are sometimes smaller than those of the sample moments.

T.L. designed research, performed research, analyzed data, and wrote the paper.

The author declares no competing interest.

[1] To whom correspondence should be addressed. E-mail: tl@biomathematics.org