

Near-consistent robust estimations of moments for unimodal distributions

Tuban Lee^{a,1}

^aInstitute of Biomathematics, Macau SAR 999078, China

This manuscript was compiled on February 21, 2023

Current robust estimators are inconsistent with population moments under non-sample-dependent breakdown points without distributional assumptions. Here, based on the invariant structure of probability distributions, a series of sophisticated yet efficient estimators, robust to both gross errors and departures from parametric assumptions, are proposed for estimating mean and central moments with insignificant asymptotic biases for common continuous unimodal distributions. This article also illuminates the understanding of the common nature of probability distributions and the measures of them.

quantile inequality | invariant | unimodal | adaptive estimation | U -statistics

The asymptotic inconsistencies between sample mean (\bar{x}) and nonparametric robust location estimators in asymmetric distributions on the real line have been noticed for more than two centuries (1) but unsolved. Strictly speaking, it is unsolvable because by trimming, some information about the original distribution is removed, making it impossible to estimate the values of the removed parts without distributional assumptions. Newcomb (1886, 1912) provided the first modern approach to this problem by developing a class of estimators that gives "less weight to the more discordant observations" (2, 3). In 1964, Huber (4) used the minimax procedure to obtain M-estimator for contaminated normal distribution, which has played a pre-eminent role in the later development of robust statistics. As previously demonstrated, under growing asymmetric departures from normality, the bias of the Huber M-estimator increases rapidly. This is a common issue in parametric estimations. For example, He and Fung constructed (5) a robust M-estimator for the two-parameter Weibull distribution. All moments can be calculated from its estimated parameters. As expected, it is inadequate for the gamma, Perato, lognormal, and especially the generalized Gaussian distributions, because the logarithmic function does not produce a result for negative inputs (SI Dataset S1). Another two interesting approaches are the quantile estimator introduced by Seki and Yokoyama (1993) (6) and the median/MAD estimator considered by Oliver (2006) (7). However, the performance of these estimators is even worse than that of the M-estimator when the distributional assumption is violated (SI Dataset S1). Descriptive statistics for parametric models currently rely heavily on the accuracy of distributional assumptions. Even when distributions such as the Weibull and gamma belong to the same larger family, the generalized gamma distribution, a misassumption can still result in substantial biases, rendering the approach ill-suited.

In previous work on semiparametric robust mean estimation, although greatly shrinking the asymptotic biases, binomial mean (BM_ϵ) is still inconsistent for any skewed distribution if $\epsilon > 0$ (if $\epsilon \rightarrow 0$, since the alternating sum of binomial coefficients is zero, $BM \rightarrow \mu$). All robust location estimators

commonly used are symmetric due to the universality of the symmetric distributions. One can construct an asymmetric trimmed mean that is consistent for a semiparametric class of skewed distributions. This approach has been investigated previously, but it is not symmetric and therefore only suitable for some special applications (8). From semiparametric to parametric, an ideal robust location estimator would have a non-sample-dependent breakdown point (defined in Subsection F) and be consistent with any symmetric distribution and a skewed distribution with finite second moments. This is called an invariant mean. Based on the mean-weighted average-median inequality, the robust mean is defined as

$$rm_{d,\epsilon,n} := \lim_{c \rightarrow \infty} \left(\frac{(WA_{\epsilon,n} + c)^{d+1}}{(\text{median} + c)^d} - c \right),$$

where d is for bias correction, $WA_{\epsilon,n}$ is $BM_{\epsilon,n}$ in the first three Subsections, while other symmetric weighted averages can also be used in practice as long as the inequalities hold. The next theorem shows the significance of this composite estimator.

Theorem 1. *If the second moments are finite, $rm_{d \approx 0.375, \epsilon = \frac{1}{8}}$ is a consistent mean estimator for the exponential and any symmetric distributions and the Pareto distribution with quantile function $Q(p) = x_m(1-p)^{-\frac{1}{\alpha}}$, $x_m > 0$, when $\alpha \rightarrow \infty$.*

Proof. Finding d, ϵ values that make $rm_{d,\epsilon}$ a consistent mean estimator is equivalent to finding the solution of $E[rm_{d,\epsilon}] = E[X]$. Rearranging the definition, $rm_{d,\epsilon} = \lim_{c \rightarrow \infty} \left(\frac{(BM_\epsilon + c)^{d+1}}{(\text{median} + c)^d} - c \right) = (d+1)BM_\epsilon - d\text{median} = \mu$. So, $d = \frac{\mu - BM_\epsilon}{BM_\epsilon - \text{median}}$. The pdf of the exponential distribution is $f(x) = \lambda^{-1}e^{-\lambda^{-1}x}$, $\lambda \geq 0$, $x \geq 0$, the cdf is $F(x) = 1 - e^{-\lambda^{-1}x}$, $x \geq 0$. The quantile function is

Significance Statement

Bias, variance, and contamination are the three main errors in statistics. Consistent robust estimation is unattainable without parametric assumptions. Here, based on a paradigm shift inspired by mean-median-mode inequality, Bickel-Lehmann spread, M-estimator, and adaptive estimation, 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.

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

$Q(p) = \ln\left(\frac{1}{1-p}\right) \lambda$. $E[x] = \lambda$. $E[\text{median}] = Q\left(\frac{1}{2}\right) = \ln 2 \lambda$.
 For the exponential distribution, the expectation of $\text{BM}_{\frac{1}{8}}$
 is $E\left[\text{BM}_{\frac{1}{8}}\right] = \lambda \left(1 + \ln\left(\frac{46656}{8575\sqrt{35}}\right)\right)$. Obviously, the scale
 parameter λ can be canceled out, $d \approx 0.375$. The proof
 of the second assertion follows directly from the coincidence
 property. For any symmetric distribution with a finite
 second moment, $E[\text{BM}_\epsilon] = E[\text{median}] = E[X]$. Then
 $E[\text{rm}_{d,\epsilon}] = \lim_{c \rightarrow \infty} \left(\frac{(E[X]+c)^{d+1}}{(E[X]+c)^d} - c\right) = E[X]$. The proof for
 the Pareto distribution is more general. The mean of the
 Pareto distribution is given by $\frac{\alpha x_m}{\alpha-1}$. The d value with two
 unknown percentiles p_1 and p_2 for the Pareto distribution is
 $d_{\text{Pareto}} = \frac{\frac{\alpha x_m}{\alpha-1} - x_m(1-p_1)^{-\frac{1}{\alpha}}}{x_m(1-p_1)^{-\frac{1}{\alpha}} - x_m(1-p_2)^{-\frac{1}{\alpha}}}$. Since any weighted aver-
 age can be expressed as an integral of the quantile function,
 $\lim_{\alpha \rightarrow \infty} \frac{\frac{\alpha}{(1-p_1)^{-1/\alpha} - (1-p_2)^{-1/\alpha}}}{\frac{\alpha}{(1-p_1)^{-1/\alpha} - (1-p_2)^{-1/\alpha}}} = -\frac{\ln(1-p_1)+1}{\ln(1-p_1)-\ln(1-p_2)}$, the d
 value for the Pareto distribution approaches that of the ex-
 ponential distribution as $\alpha \rightarrow \infty$, regardless of the type of
 weighted average used. This completes the demonstration. \square

Theorem 1 implies that for the Weibull, gamma,
 Pareto, lognormal and generalized Gaussian distribution,
 $\text{rm}_{d \approx 0.375, \epsilon = \frac{1}{8}}$ is consistent for at least one particular
 case of these two-parameter distributions. The biases of
 $\text{rm}_{d \approx 0.375, \epsilon = \frac{1}{8}}$ for distributions with skewness between those
 of the exponential and symmetric distributions are tiny (SI
 Dataset S1). Therefore, $\text{rm}_{d \approx 0.375, \epsilon = \frac{1}{8}}$ has excellent perfor-
 mance for all these common unimodal distributions (SI Dataset
 S1).

Besides introducing the concept of invariant mean, the
 purpose of this paper is to demonstrate that the estimation of
 central moments can be transformed into a location estimation
 problem by using U -statistics, the central moment kernel
 distributions have nice properties, and, in light of previous
 works, a series of sophisticated yet efficient robust estimators
 can be constructed whose biases are typically smaller than the
 variances ($n = 5400$, Table 1) for unimodal distributions.

Background and Main Results

A. Invariant mean. It has long been known that a theoreti-
 cal model can be adjusted to fit the first two moments of
 the observed data. A continuous distribution belonging to a
 location-scale family has the form $F(x) = F_0\left(\frac{x-\mu}{\lambda}\right)$, where
 F_0 is a "standard" distribution. Then, $F(x) = Q^{-1}(x) \rightarrow$
 $x = Q(p) = \lambda Q_0(p) + \mu$. So, any weighted average can
 be expressed as $\lambda W_{\text{estimator}}(\epsilon) + \mu$, where $W_{\text{estimator}}(\epsilon)$
 is a function of $Q_0(p)$ according to the definition of the
 weighted average. The simultaneous cancellation of μ and
 λ in $\frac{(\lambda W_{\mu}(\epsilon) + \mu) - (\lambda W_{\text{BM}_\epsilon}(\epsilon) + \mu)}{(\lambda W_{\text{BM}_\epsilon}(\epsilon) + \mu) - (\lambda W_{\text{median}}(\epsilon) + \mu)}$ ensures that d is a con-
 stant. Consequently, the roles of BM_ϵ and median in $\text{rm}_{d,\epsilon}$
 can be replaced by any weighted averages.

The performance in heavy-tailed distributions can be im-
 proved further by constructing the quantile mean as

$$qm_{d,\epsilon,n} := \hat{Q}_n \left(\left(\hat{F}_n(\text{WA}_{\epsilon,n}) - \frac{1}{2} \right) d + \hat{F}_n(\text{WA}_{\epsilon,n}) \right),$$

provided that $\hat{F}_n(\text{WA}_{\epsilon,n}) \geq \frac{1}{2}$, where $\hat{F}_n(x)$ is the empir-
 ical cumulative distribution function of the sample, \hat{Q}_n is
 the sample quantile function. The most popular method

for computing the sample quantile function was proposed
 by Hyndman and Fan in 1996 (9). To minimize the finite
 sample bias, here, $\hat{F}_n(x) := \frac{1}{n} \left(\frac{x - \hat{Q}_n(\frac{sp}{n})}{\hat{Q}_n(\frac{1}{n}(sp+1)) - \hat{Q}_n(\frac{sp}{n})} + sp \right)$,
 where $sp = \sum_{i=1}^n \mathbf{1}_{X_i \leq x}$, $\mathbf{1}_A$ is the indicator of event A . The
 solution of $\hat{F}_n(\text{WA}_{\epsilon,n}) < \frac{1}{2}$ is reversing the percentile by
 $1 - \hat{F}_n(\text{WA}_{\epsilon,n})$, the obtained percentile is also reversed*.
 Moreover, in extreme heavy-tailed distributions, the calcu-
 lated percentile can exceed the breakdown point of WA_ϵ , so
 the percentile will be modified to $1 - \epsilon$ if this happens. The
 quantile mean uses the location-scale invariant in a different
 way, as shown in the following proof.

Theorem A.1. $qm_{d \approx 0.321, \epsilon = \frac{1}{8}}$ is a consistent mean estimator
 for the exponential, Pareto ($\alpha \rightarrow \infty$) and any symmetric
 distributions provided that the second moments are finite.

Proof. Similarly, rearranging the definition, $d = \frac{F(\mu) - F(\text{BM}_\epsilon)}{F(\text{BM}_\epsilon) - \frac{1}{2}}$.

Recall the cdf is $F(x) = 1 - e^{-\lambda^{-1}x}$, $x \geq 0$, the expectation
 of BM_ϵ can be expressed as $\lambda W_{\text{BM}_\epsilon}(\epsilon)$, so $F(\text{BM}_\epsilon)$ is free of
 λ . When $\epsilon = \frac{1}{8}$, $d = \frac{-e^{-1} + e^{-\left(1 + \ln\left(\frac{46656}{8575\sqrt{35}}\right)\right)}}{\frac{1}{2} - e^{-\left(1 + \ln\left(\frac{46656}{8575\sqrt{35}}\right)\right)}} \approx 0.321$.

The proof of the symmetric case is similar. Since for
 any symmetric distribution with a finite second moment,
 $F(E[\text{BM}_\epsilon]) = F(\mu) = \frac{1}{2}$. Then, the expectation of the
 quantile mean is $qm_{d,\epsilon} = F^{-1}\left(\left(F(\mu) - \frac{1}{2}\right)d + F(\mu)\right) =$
 $F^{-1}\left(0 + F(\mu)\right) = \mu$.

For the assertion related to the Pareto distribu-
 tion, the cdf of it is $1 - \left(\frac{x_m}{x}\right)^\alpha$. So, the d
 value with two unknown percentile p_1 and p_2 is

$$\begin{aligned}
 d_{\text{Pareto}} &= \frac{1 - \left(\frac{x_m}{\frac{\alpha x_m}{\alpha-1}}\right)^\alpha - \left(1 - \left(\frac{x_m}{x_m(1-p_1)^{-\frac{1}{\alpha}}}\right)^\alpha\right)}{\left(1 - \left(\frac{x_m}{x_m(1-p_1)^{-\frac{1}{\alpha}}}\right)^\alpha\right) - \left(1 - \left(\frac{x_m}{x_m(1-p_2)^{-\frac{1}{\alpha}}}\right)^\alpha\right)} = \\
 &= \frac{1 - \left(\frac{\alpha-1}{p_1-p_2}\right)^\alpha - p_1}{1 - \left(\frac{\alpha-1}{p_1-p_2}\right)^\alpha - p_1}. \text{ When } \alpha \rightarrow \infty, \left(\frac{\alpha-1}{p_1-p_2}\right)^\alpha = \frac{1}{e}. \text{ The } d \text{ value} \\
 &\text{for the exponential distribution is identical, since } d_{\text{exp}} = \\
 &= \frac{(1-e^{-1}) - \left(1 - e^{-\ln\left(\frac{1}{1-p_1}\right)}\right)}{\left(1 - e^{-\ln\left(\frac{1}{1-p_1}\right)}\right) - \left(1 - e^{-\ln\left(\frac{1}{1-p_2}\right)}\right)} = \frac{1 - \frac{1}{e} - p_1}{p_1 - p_2}. \text{ All results}
 \end{aligned}$$

are now proven. \square

The definitions of location and scale parameters are such
 that they must satisfy $F(x; \lambda, \mu) = F\left(\frac{x-\mu}{\lambda}; 1, 0\right)$. Recall that
 $x = \lambda Q_0(p) + \mu$, so the percentile of any weighted average
 is free of λ and μ , guaranteeing the validity of the quantile
 mean. $qm_{d \approx 0.321, \epsilon = \frac{1}{8}}$ works better in the fat-tail scenarios (SI
 Dataset S1). Theorem 1 and A.1 show that $\text{rm}_{d \approx 0.375, \epsilon = \frac{1}{8}}$
 and $qm_{d \approx 0.321, \epsilon = \frac{1}{8}}$ are both consistent mean estimators for
 any symmetric distribution and a skewed distribution with
 finite second moments. It's obvious that the breakdown points
 of $\text{rm}_{d \approx 0.375, \epsilon = \frac{1}{8}}$ and $qm_{d \approx 0.321, \epsilon = \frac{1}{8}}$ are both $\frac{1}{8}$. Therefore
 they are all invariant means.

It is constructive to consider a symmetric weighted average
 as a mixture of trimmed mean and symmetric quantile average.
 Since the quantile inequality is not a necessary condition of the

*Without loss of generality, in the following discussion, only the $\hat{F}_n(\text{WA}_{\epsilon,n}) \geq \frac{1}{2}$ case will be
 considered.

trimming inequality, the safest choice of symmetric weighted average in rm and qm is the $\frac{1}{4}$ -trimmed mean (SI Dataset S1). Although using a less-biased symmetric weighted average can generally improve performance, there is a higher risk of violation in the semiparametric framework. However, another key factor determining the risk of violation is the skewness of the distribution. Consider there is a near-symmetric distribution S such that SQA_ϵ as a function of ϵ is monotonic increasing from 0 to u and monotonic decreasing from u to $\frac{1}{2}$, and $\mu = m$. Based on the definition, S does not follow the quantile inequality. Depending on the fluctuation degree, $|SQA_u - m|$, the mean-SWA $_\epsilon$ -median inequality is usually not valid for S . Then, the monotonic increase from 0 to u implies that $Q'(\epsilon) \geq Q'(1-\epsilon) \Leftrightarrow f(Q(1-\epsilon)) \geq f(Q(\epsilon))$ always holds for $0 \leq \epsilon \leq u$. Similarly, $Q'(\epsilon) \leq Q'(1-\epsilon) \Leftrightarrow f(Q(\epsilon)) \geq f(Q(1-\epsilon))$ always holds for $u \leq \epsilon \leq \frac{1}{2}$. Transforming S with a function $\chi(x)$ such that $\frac{d^2\chi}{dx^2} \geq 0 \wedge \frac{d\chi}{dx} \geq 0$ over the interval supported will decrease $f(Q(\epsilon))$, and the decrease rate, due to the order, is much smaller than $f(Q(1-\epsilon))$. That means, as the second derivative of $\chi(x)$ increases, eventually, after a point, $f(Q(\epsilon))$ will always be greater than $f(Q(1-\epsilon))$, i.e., the SQA_ϵ function will be monotonic decreasing and S will eventually follow the quantile inequality. $|\mu - m|$ will increase so that the skewness will increase. Even the transformation is not enough to make the SQA function strictly monotonic; suppose it is generally decreasing in $[0, u]$, but increasing in $[u, \frac{1}{2}]$, since $1 - 2\epsilon$ of the symmetric quantile averages will be included in the computation of BM_ϵ , as long as $|u - \frac{1}{2}| \ll 1 - 2\epsilon$, and other parts of the SQA function satisfy the inequality constraints which define the ν th quantile inequality, the mean- BM_ϵ -median inequality will still be valid (as an example, the SQA function is non-monotonic when the shape parameter of the Weibull distribution $\alpha > \frac{1}{1-\ln(2)} \approx 3.259$ as shown in the previous article, yet the mean- $BM_{\frac{1}{8}}$ -median inequality is still valid when $\alpha \leq 3.322$). Accordingly, in a family of distributions that differ by a skewness-increasing transformation, the violation of the mean-WA $_\epsilon$ -median inequality often only occurs when the distribution is near-symmetric, but the over-corrections in rm and qm are dependent on the WA $_\epsilon$ -median difference, which is correlated to the skewness, so the over-correction, if it happens, is often tiny with a moderate d . This qualitative analysis provides another perspective, in addition to the bias bounds (10), that rm and qm based on the mean-WA $_\epsilon$ -median inequality are generally safe.

B. Robust estimations of the central moments. In 1976, Bickel and Lehmann, in their third paper of the landmark series *Descriptive Statistics for Nonparametric Models* (11), generalized a class of estimators called "measures of disperse," which is now often named as Bickel-Lehmann dispersion. As an example, they proposed a first version of the trimmed standard deviation, $\hat{\tau}^2(F; \epsilon) \equiv \tau^2(F; \epsilon)$, for independent and identically distributed random variables X_i with a distribution F , where $\tau^2(F; \epsilon) = \frac{1}{1-2\epsilon} \int_{Q(\epsilon)}^{Q(1-\epsilon)} y dG(y)$, Q is the quantile function of G , G is the distribution of $Y = X^2$. Obviously, when $\epsilon = 0$, the result is equivalent to the second raw moment. In 1979, in the same series (12), they explored another class of estimators called "measures of spread," which "does not require the assumption of symmetry." From that, a popular efficient scale estimator, the Rousseeuw-Croux scale estimator (13), was derived in 1993, but the importance of tackling the

symmetry assumption has been greatly underestimated. In the final section of the paper, they considered another two possible versions of the trimmed standard deviations, which were modified symmetrically here for comparison,

$$\left[n \left(\frac{1}{2} - \epsilon \right) \right]^{-\frac{1}{2}} \left[\sum_{k=\frac{n}{2}}^{n(1-\epsilon)} [X_k - X_{n-k+1}]^2 \right]^{\frac{1}{2}}, \quad [1]$$

and

$$\left[\binom{n}{2} (1 - 2\epsilon) \right]^{-\frac{1}{2}} \left[\sum_{k=\binom{n}{2}\epsilon}^{\binom{n}{2}(1-\epsilon)} (X - X')_k^2 \right]^{\frac{1}{2}}, \quad [2]$$

where $(X - X')_1 \leq \dots \leq (X - X')_{\binom{n}{2}}$ are the order statistics of the "pseudo-sample" $X_i - X_j$, $i < j$. The paper ended with, "We do not know a fortiori which of the measures [1] or [2] is preferable and leave these interesting questions open."

Observe that the kernel of the unbiased estimation of the second central moment by using U -statistic is $\psi_2(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$. If adding the $\frac{1}{2}$ term in [2], as $\epsilon \rightarrow 0$, the result is equivalent to the standard deviation estimated by using U -statistic (also noted by Janssen, Serfling, and Veraverbeke in 1987) (14). In fact, they also implied that, when ϵ is 0, [2] is $\sqrt{2}$ times the standard deviation.

To address their open questions, the nomenclature used in this paper is introduced as follows:

Nomenclature. Given a robust estimator $\hat{\theta}$. The first part of the name of the robust statistic defined in this paper is a prefix that indicates the type of estimator, and the second part is the name of the population parameter θ that the estimator is consistent with as $\epsilon \rightarrow 0$. The abbreviation of the estimator is the initial letter(s) of the first part plus the common abbreviation of the consistent estimator that measures the population parameter. If the estimator is not a U -statistic, the breakdown point, ϵ , is indicated in the subscript of the abbreviation of the estimator. If the estimator is a robustified U -statistic or a composite estimator, the breakdown point of the location estimator is indicated (except the median).

Naturally, the trimmed standard deviation following this nomenclature is $Tsd_{\epsilon, n} := \left[TM_\epsilon \left((\psi_2(X_{N_1}, X_{N_2}))_{N=1}^{\binom{n}{2}} \right) \right]^{-\frac{1}{2}}$, where $TM_\epsilon(Y)$ denotes the ϵ -symmetric trimmed mean with the sequence $(\psi_2(X_{N_1}, X_{N_2}))_{N=1}^{\binom{n}{2}}$ as an input. If the square root is removed, it is named as the trimmed variance ($Tvar_{\epsilon, n}$). It is now very clear that this definition, essentially the same as [2], should be preferable. Not only because it is essentially a trimmed U -statistic for the standard deviation but also because the quantile inequality of the pseudo-sample distribution is ensured by the next exciting theorem.

Theorem B.1. *The second central moment kernel distribution generated from any continuous unimodal distribution follows the quantile inequality.*

Proof. Let $Q(p)$, $0 \leq p \leq 1$, denote the quantile of the continuous unimodal distribution $f_X(x)$. The corresponding probability density is $f(Q(p))$. Generating the distribution of the pair $(Q(p_i), Q(p_j))$, $i < j$, $p_i < p_j$, the corresponding probability density is $f_{X, X}(Q(p_i), Q(p_j)) = 2f(Q(p_i))f(Q(p_j))$.

Transforming the pair $(Q(p_i), Q(p_j))$, $i < j$, by the function $\Phi(x_1, x_2) = x_1 - x_2$, the pairwise difference distribution has a mode that is arbitrary close to $M - M = 0$. The monotonic increasing of the pairwise difference distribution was first implied in its unimodality proof done by Hodges and Lehmann in 1954 (15). Whereas they used induction to get the result, Dharmadhikari and Jogdeo in 1982 (16) gave a modern proof of the unimodality using Khintchine's representation (17). Assuming absolute continuity, Purkayastha (18) introduced a much simpler proof in 1998. Transforming the pairwise difference distribution by squaring and multiplying $\frac{1}{2}$ does not change the monotonicity, making the pdf become monotonically decreasing with mode at zero. In the previous semiparametric robust mean estimation article, it is proven that a distribution with monotonic pdf always follows the quantile inequality, which gives the desired result. \square

Remark. The assumption of continuity of distributions is important for monotonicity because, unlike in the continuous case, it is possible to get pairs with the same value for a discrete distribution. For example, let the probabilities of the singletons $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$ and $\{5\}$ of a probability mass function of a discrete probability distribution be $\frac{1}{11}$, $\frac{4}{11}$, $\frac{3}{11}$, $\frac{2}{11}$, and $\frac{1}{11}$, respectively. This is a unimodal distribution, but the corresponding ψ_2 distribution is non-monotonic, whose singletons $\{0\}$, $\{0.5\}$, $\{2\}$, $\{4.5\}$ and $\{8\}$ have probabilities $\frac{21}{66}$, $\frac{24}{66}$, $\frac{2}{14}$, $\frac{6}{66}$, and $\frac{1}{66}$, respectively.

Previously, it was shown that any symmetric distribution with a finite second moment follows the ν th quantile inequality. That means the quantile inequality does not require unimodality, e.g., for a symmetric bimodal distribution, it also follows the quantile inequality. Examples from the Weibull distribution show that unimodality does not guarantee quantile inequality. Theorem B.1 reveals another profound relationship between unimodality and quantile inequality, which is sufficient for trimming inequality.

In 1928, Fisher constructed k -statistics as unbiased estimators of cumulants (19). Halmos (1946) proved that the functional θ admits an unbiased estimator if and only if it is a regular statistical functional of degree k and showed a relation of symmetry, unbiasedness and minimum variance (20). In 1948, Hoeffding generalized U -statistics (21) which enable the derivation of a minimum-variance unbiased estimator from each unbiased estimator of an estimable parameter. Heffernan (1997) (22) obtained an unbiased estimator of the k th central moment by using U -statistics and demonstrated that it is the minimum variance unbiased estimator for distributions with finite moments (23, 24). In 1984, to study generalized L -statistics, Serfling considered the U -statistic structure (25). Gijbels, Janssen and Veraverbeke generalized the trimmed U -statistics in 1988 (26). Due to the combinatorial explosion, the bootstrap (27), introduced by Efron in 1979, is indispensable in large sample studies. In 1981, Bickel and Freedman (28) showed that the bootstrap is asymptotically valid to approximate the original distribution in a wide range of situations, including U -statistics. After that, the limit laws of bootstrapped U -statistics have been intensively studied and proven by Athreya, Ghosh, Low, and Sen (1984) and Helmers, Janssen, and Veraverbeke (1990) (29, 30). Users can check the accuracy by comparing the unbiased central moments (31) to the bootstrap central moments. The weighted k th central

moment ($k \leq n$) is defined as,

$$Wkm_{\epsilon, n} := WA_{\epsilon, n} \left((\psi_k(X_{N_1}, \dots, X_{N_k}))_{N=1}^n \right),$$

where X_{N_1}, \dots, X_{N_k} are the n choose k elements from X , $\psi_k(x_1, \dots, x_k) = \sum_{j=0}^{k-2} (-1)^j \binom{1}{k-j} \sum (x_1^{k-j} \dots x_{i_{j+1}}) + (-1)^{k-1} (k-1) x_1 \dots x_k$, the second summation is over $i_1, \dots, i_{j+1} = 1$ to k with $i_1 < \dots < i_{j+1}$ (12, 19, 20, 22). Despite the complexity, the structure of the k th central moment kernel distributions can be elucidated by decomposing.

Theorem B.2. For each pair $(Q(p_i), Q(p_j))$ of the original distribution, let $x_1 = Q(p_i)$ and $x_k = Q(p_j)$, $\Delta = Q(p_i) - Q(p_j)$. The k th central moment kernel distribution, $k > 2$, can be seen as a mixture distribution and each of the components has the support $(-\binom{k}{3+\frac{1}{k}})^{-1}(-\Delta)^k, \frac{1}{k}(-\Delta)^k$.

Proof. Generating the distribution of the k -tuple $(Q(p_{i_1}), \dots, Q(p_{i_k}))$, $k > 2$, $i_1 < \dots < i_k$, $p_{i_1} < \dots < p_{i_k}$, the corresponding probability density is $f_{X, \dots, X}(Q(p_{i_1}), \dots, Q(p_{i_k})) = k! f(Q(p_{i_1})) \dots f(Q(p_{i_k}))$. Transforming the distribution of the k -tuple by the function $\psi_k(x_1, \dots, x_k)$, denoting $\bar{\Delta} = \psi_k(Q(p_{i_1}), \dots, Q(p_{i_k}))$. The probability $f_{\Xi_k}(\bar{\Delta}) = \sum_{\bar{\Delta} = \psi_k(Q(p_{i_1}), \dots, Q(p_{i_k}))} f_{X, \dots, X}(Q(p_{i_1}), \dots, Q(p_{i_k}))$ is the summation of the probabilities of all k -tuples such that $\bar{\Delta}$ is equal to $\psi_k(Q(p_{i_1}), \dots, Q(p_{i_k}))$. The following Ξ_k is equivalent.

Ξ_k : Every pair with a difference equal to $\Delta = Q(p_{i_1}) - Q(p_{i_k})$ can generate a pseudodistribution (but the integral is not equal to 1, so "pseudo") such that x_2, \dots, x_{k-1} exhaust all combinations under the inequality constraints, i.e., $Q(p_{i_1}) = x_1 < x_2 < \dots < x_{k-1} < x_k = Q(p_{i_k})$. The combination of all the pseudodistributions with the same Δ is ξ_Δ . The combination of ξ_Δ , i.e., from $\Delta = 0$ to $Q(0) - Q(1)$, is Ξ_k .

The support of ξ_Δ is the extrema of ψ_k subject to the inequality constraints. Using the Lagrange multiplier, one can easily determine the only critical point at $x_1 = \dots = x_k = 0$, where $\psi_k = 0$. Other candidates are within the boundaries, i.e., $\psi_k(x_1 = x_1, x_2 = x_k, \dots, x_k = x_k), \dots, \psi_k(x_1 = x_1, \dots, x_i = x_1, x_{i+1} = x_k, \dots, x_k = x_k), \dots, \psi_k(x_1 = x_1, \dots, x_{k-1} = x_1, x_k = x_k)$. $\psi_k(x_1 = x_1, \dots, x_i = x_1, x_{i+1} = x_k, \dots, x_k = x_k)$ can be divided into k groups. If $\frac{k+1-i}{2} \leq j \leq \frac{k-1}{2}$, from $j+1$ st to $k-j$ th group, the g th group has $i \binom{i-1}{g-j-1} \binom{k-i}{j}$ terms having the form $(-1)^{g+1} \frac{1}{k-g+1} x_1^{k-j} x_k^j$, from $k-j+1$ th to $i+j$ th group, the g th group has $i \binom{i-1}{g-j-1} \binom{k-i}{j-k+g-1} \binom{i}{k-j}$ terms having the form $(-1)^{g+1} \frac{1}{k-g+1} x_1^{k-j} x_k^j$. If $j < \frac{k+1-i}{2}$, from $j+1$ st to $i+j$ th group, the g th group has $i \binom{i-1}{g-j-1} \binom{k-i}{j}$ terms having the form $(-1)^{g+1} \frac{1}{k-g+1} x_1^{k-j} x_k^j$. If $j \geq \frac{k}{2}$, from $k-j+1$ st to j th group, the g th group has $(k-i) \binom{k-i-1}{j-k+g-1} \binom{i}{k-j}$ terms having the form $(-1)^{g+1} \frac{1}{k-g+1} x_1^{k-j} x_k^j$, from $j+1$ th to $j+i$ th group, $i+j < k$, the g th group has $i \binom{i-1}{g-j-1} \binom{k-i}{j} + (k-i) \binom{k-i-1}{j-k+g-1} \binom{i}{k-j}$ terms having the form $(-1)^{g+1} \frac{1}{k-g+1} x_1^{k-j} x_k^j$. The final k th group is the term $(-1)^{k-1} (k-1) x_1^{k-i} x_k^{i-1}$. So, if $i+j = k$, $j \geq \frac{k}{2}$, $i \leq \frac{k}{2}$, the summed coefficient of $x_1^{k-i} x_k^{i-1}$ is $(-1)^{k-1} (k-1) + \sum_{g=i+1}^{k-1} (-1)^{g+1} \frac{1}{k-g+1} (k-i) \binom{k-i-1}{g-i-1} + \sum_{g=k-i+1}^{k-1} (-1)^{g+1} \frac{1}{k-g+1} i \binom{i-1}{g-k+i-1} = (-1)^{k-1} (k-1) +$

343 $(-1)^{k+1} + (k-i)(-1)^k + (-1)^k(i-1) =$
 344 $(-1)^{k+1}$. The summation identities are
 345 $\sum_{g=i+1}^{k-1} (-1)^{g+1} \frac{1}{k-g+1} (k-i) \binom{k-i-1}{g-i-1} =$
 346 $(k-i) \int_0^1 \sum_{g=i+1}^{k-1} (-1)^{g+1} \binom{k-i-1}{g-i-1} t^{k-g} dt =$
 347 $(k-i) \int_0^1 ((-1)^i (t-1)^{k-i-1} - (-1)^{k+1}) dt =$
 348 $(k-i) \left(\frac{(-1)^k}{i-k} + (-1)^k \right) = (-1)^{k+1} +$
 349 $(k-i)(-1)^k \cdot \sum_{g=k-i+1}^{k-1} (-1)^{g+1} \frac{1}{k-g+1} i \binom{i-1}{g-k+i-1} =$
 350 $\int_0^1 \sum_{g=k-i+1}^{k-1} (-1)^{g+1} i \binom{i-1}{g-k+i-1} t^{k-g} dt =$
 351 $\int_0^1 (i(-1)^{k-i} (t-1)^{i-1} - i(-1)^{k+1}) dt = (-1)^k (i-1).$
 352 If $j < \frac{k+1-i}{2}$, $i > k-1$, if $i = k$, $\psi_k = 0$, if $\frac{k+1-i}{2} \leq j \leq \frac{k-1}{2}$,
 353 $\frac{k+1}{2} \leq i \leq k-1$, the summed coefficient of $x_1^i x_k^{k-i}$ is
 354 $(-1)^{k-1} (k-1) + \sum_{g=k-i+1}^{k-1} (-1)^{g+1} \frac{1}{k-g+1} i \binom{i-1}{g-k+i-1} +$
 355 $\sum_{g=i+1}^{k-1} (-1)^{g+1} \frac{1}{k-g+1} (k-i) \binom{k-i-1}{g-i-1}$, the same as above. If
 356 $i+j < k$, since $\binom{i}{k-j} = 0$, the related terms can be ignored, so,
 357 using the binomial theorem and beta function, the summed co-
 358 efficient of $x_1^{k-j} x_k^j$ is $\sum_{g=j+1}^{i+j} (-1)^{g+1} \frac{1}{k-g+1} i \binom{i-1}{g-j-1} \binom{k-i}{j} =$
 359 $i \binom{k-i}{j} \int_0^1 \sum_{g=j+1}^{i+j} (-1)^{g+1} \binom{i-1}{g-j-1} t^{k-g} dt =$
 360 $\binom{k-i}{j} i \int_0^1 ((-1)^j t^{k-j-1} \left(\frac{t}{i-1} \right)^{i-1}) dt =$
 361 $\binom{k-i}{j} i \frac{(-1)^{j+i+1} \Gamma(i) \Gamma(k-j-i+1)}{\Gamma(k-j+1)} = \frac{(-1)^{j+i+1} i! (k-j-i)! (k-i)!}{(k-j)! j! (k-j-i)!} =$
 362 $(-1)^{j+i+1} \frac{i! (k-i)!}{k!} \frac{k!}{(k-j)! j!}$. The coefficient of $x_1^i x_k^{k-i}$ in
 363 $\binom{k}{i}^{-1} (-1)^{1+i} (x_1 - x_k)^k$ is $\binom{k}{i}^{-1} (-1)^{1+i} \binom{k}{i} (-1)^{k-i} =$
 364 $(-1)^{k+1}$, the same as the summed coefficient if
 365 $i+j = k$. If $i+j < k$, the coefficient of $x_1^{k-j} x_k^j$ is
 366 $\binom{k}{i}^{-1} (-1)^{1+i} \binom{k}{j} (-1)^j$, same as the summed coefficient.
 367 Since $\psi_k(x_1 = x_1, \dots, x_i = x_1, x_{i+1} = x_k, \dots, x_k = x_k) =$
 368 $\binom{k}{i}^{-1} (-1)^{1+i} (x_1 - x_k)^k$, the maximum and minimum of
 369 ψ_k follow directly from the properties of the binomial
 370 coefficient. \square

371 ξ_Δ is closely related to $f_\Xi(\Delta)$, which is the pairwise dif-
 372 ference distribution, since the probability density of ξ_Δ is
 373 $f_{\Xi_k}(\bar{\Delta}|\Delta) \cdot \sum_{\bar{\Delta} = -\frac{k}{2}(-\Delta)^k}^{\frac{1}{k}(-\Delta)^k} (-\Delta)^k f_{\Xi_k}(\bar{\Delta}|\Delta) = f_\Xi(\Delta)$. Re-
 374 call that $f_\Xi(\Delta)$ is monotonic increasing with a mode at the
 375 origin if the original distribution is unimodal. Thus, in gen-
 376 eral, ignoring the shape of ξ_Δ , Ξ_k is monotonic left and right
 377 around zero. In fact, the median of Ξ_k is also close to zero,
 378 as it can be cast as a weighted average of the medians of
 379 ξ_Δ . When Δ is small, all values of ξ_Δ are close to zero,
 380 resulting in a median arbitrarily close to zero, When Δ is
 381 large, the median of ξ_Δ depends on its skewness, but the
 382 corresponding weight is much smaller, so even if ξ_Δ is highly
 383 skewed, the median of Ξ_k will only be slightly shifted from
 384 zero (denote the median of Ξ_k as m_{Ξ_k} , for five parametric
 385 distributions here, $|m_{\Xi_k}|$ s are all $\leq 0.1\sigma$ for Ξ_3 and Ξ_4 , SI
 386 Dataset S1). Assuming $m_{\Xi_k} = 0$, for the even ordinal central
 387 moment kernel distribution, the average probability density
 388 on the left side of zero is greater than that on the right side,
 389 since $\frac{\frac{1}{2}}{\binom{k}{2}^{-1}(Q(0)-Q(1))^k} > \frac{\frac{1}{2}}{\frac{1}{k}(Q(0)-Q(1))^k}$. This means that,
 390 on average, the inequality $f(Q(\epsilon)) \geq f(Q(1-\epsilon))$ holds. For
 391 the odd ordinal distribution, the discussion is harder since
 392 it is generally symmetric. Just consider Ξ_3 , let $x_1 = Q(p_i)$
 393 and $x_3 = Q(p_j)$, changing the value of x_2 from $Q(p_i)$ to
 394 $Q(p_j)$ will monotonically change the value of $\psi_3(x_1, x_2, x_3)$,

395 since $\frac{\partial \psi_3(x_1, x_2, x_3)}{\partial x_2} = -\frac{x_2^2}{2} - x_1 x_2 + 2x_1 x_3 + x_2^2 - x_2 x_3 - \frac{x_3^2}{2}$,
 396 $-\frac{3}{4}(x_1 - x_3)^2 \leq \frac{\partial \psi_3(x_1, x_2, x_3)}{\partial x_2} \leq -\frac{1}{2}(x_1 - x_3)^2 \leq 0$. If the
 397 original distribution is right-skewed, ξ_Δ will be left-skewed,
 398 so, for Ξ_3 , the average probability density of the right side of
 399 zero will be greater than that of the left side, which means,
 400 on average, the inequality $f(Q(\epsilon)) \leq f(Q(1-\epsilon))$ holds (the
 401 same result can be inferred from the definition of central mo-
 402 ments, the positive of odd order central moment is directly
 403 related to the left-skewness of the corresponding kernel dis-
 404 tribution). In all, the monotonicity of the pairwise difference
 405 distribution guides the general shape of the k th central mo-
 406 ment kernel distribution, $k > 2$, forcing it to be unimodal-like
 407 with mode and median close to zero, then, the inequality
 408 $f(Q(\epsilon)) \leq f(Q(1-\epsilon))$ or $f(Q(\epsilon)) \geq f(Q(1-\epsilon))$ holds in
 409 general. Although the inequality may be violated in a small
 410 range, as discussed in Subsection A, the mean-SWA $_{\epsilon}$ -median
 411 inequality remains frequently valid for the central moment
 412 kernel distribution.

413 Another key property of the central moment kernel distri-
 414 bution, location invariant, is introduced in the next theorem.

Theorem B.3. $\psi_k(x_1 = \lambda x_1 + \mu, \dots, x_k = \lambda x_k + \mu) = \lambda^k \psi_k(x_1, \dots, x_k)$.

417 *Proof.* ψ_k can be divided into k groups. From 1st to
 418 $k-1$ th group, the g th group has $\binom{k}{g} \binom{g}{1}$ terms having
 419 the form $(-1)^{g+1} \frac{1}{k-g+1} x_{i_1}^{k-g+1} \dots x_{i_g}$. The final k th
 420 group is the term $(-1)^{k-1} (k-1) x_1 \dots x_k$. Let $x_{i_1} = x_1$,
 421 $k \neq g$, the g th group of ψ_k has $\binom{k-1}{g-1}$ terms having
 422 the form $(-1)^{g+1} \frac{1}{k-g+1} x_1^{k-g+1} x_2 \dots x_l x_{i_1} \dots x_{i_{g-l}}$, where
 423 x_1, x_2, \dots, x_l are fixed, $x_{i_1}, \dots, x_{i_{g-l}}$ are selected such that
 424 $i_1, \dots, i_{g-l} \neq 1, 2, \dots, l$. Let $\Psi_k(x_1, x_2, \dots, x_l, x_{i_1}, \dots, x_{i_{g-l}}) =$
 425 $(\lambda x_1 + \mu)^{k-g+1} (\lambda x_2 + \mu) \dots (\lambda x_l + \mu) (\lambda x_{i_1} + \mu) \dots$
 426 $(\lambda x_{i_{g-l}} + \mu)$, the first group of Ψ_k is $\lambda^k x_1 \dots x_l x_{i_1} \dots x_{i_{g-l}}$,
 427 the h th group of Ψ_k , $h > 1$, has $\binom{k-g+1}{k-h-l+2}$ terms
 428 having the form $\lambda^{k-h+1} \mu^{h-1} x_1^{k-h-l+2} x_2 \dots x_l$. Trans-
 429 forming ψ_k by Ψ_k , then combining all terms with
 430 $\lambda^{k-h+1} \mu^{h-1} x_1^{k-h-l+2} x_2 \dots x_l$, $x_1^{k-h-l+2} \neq x_1$, the summed
 431 coefficient is $S_{1l} = \sum_{g=l}^{h+l-1} (-1)^{g+1} \frac{1}{k-g+1} \binom{k-g+1}{k-h-l+2} \binom{k-l}{g-l} =$
 432 $\sum_{g=l}^{h+l-1} (-1)^{g+1} \frac{(k-l)!}{(h+l-g-1)!(k-h-l+2)!(g-l)!} = 0$, since the
 433 summation is starting from l , ending at $h+l-1$, the first
 434 term includes the factor $g-l=0$, the final term includes
 435 the factor $h+l-g-1=0$, the terms in the middle are
 436 also zero due to the factorial property. Another possible
 437 choice is letting one of $x_{i_2} \dots x_{i_g}$ equal to x_1 , the g th
 438 group of ψ_k has $(k-h) \binom{h-1}{g-k+h-1}$ terms having the form
 439 $(-1)^{g+1} \frac{1}{k-g+1} x_1 x_2 \dots x_j^{k-g+1} \dots x_{k-h+1} x_{i_1} \dots x_{i_{g-k+h-1}}$,
 440 provided that $k \neq g$, $2 \leq j \leq k-h+1$, where
 441 x_1, \dots, x_{k-h+1} are fixed, x_j^{k-g+1} and $x_{i_1}, \dots, x_{i_{g-k+h-1}}$
 442 are selected. Transforming these terms by
 443 $\Psi_k(x_1, x_2, \dots, x_j, \dots, x_{k-h+1}, x_{i_1}, \dots, x_{i_{g-k+h-1}}) =$
 444 $(\lambda x_1 + \mu)(\lambda x_2 + \mu) \dots (\lambda x_j + \mu)^{k-g+1} \dots (\lambda x_{k-h+1} + \mu)$
 445 $(\lambda x_{i_1} + \mu) \dots (\lambda x_{i_{g-k+h-1}} + \mu)$, then, there
 446 are $k-g+1$ terms having the form
 447 $\lambda^{k-h+1} \mu^{h-1} x_1 x_2 \dots x_{k-h+1}$. So, the combined result
 448 is $(-1)^{g+1} (k-h) \binom{h-1}{g-k+h-1} \lambda^{k-h+1} \mu^{h-1} x_1 x_2 \dots x_{k-h+1}$.
 449 Transforming the final k th group of ψ_k by
 450 Ψ_k , then, there is one term having the form
 451 $(-1)^{k-1} (k-1) \lambda^{k-h+1} \mu^{h-1} x_1 x_2 \dots x_{k-h+1}$. Another

possible combination is that the g th group of ψ_k contains $(g-k+h-1)\binom{h-1}{g-k+h-1}$ terms having the form $(-1)^{g+1}\frac{1}{k-g+1}x_1x_2\ldots x_{k-h+1}x_{i_1}\ldots x_{i_j}^{k-g+1}\ldots x_{i_{g-k+h-1}}$, there is only one term having the form $\lambda^{k-h+1}\mu^{h-1}x_1x_2\ldots x_{k-h+1}$. The above summation $S1_l$ should also be included, i.e., $x_1^{k-h-l+2} = x_1$, $k = h+l-1$, so, combining all terms with $\lambda^{k-h+1}\mu^{h-1}x_1x_2\ldots x_{k-h+1}$, according to the binomial theorem, the summed coefficient is $S2_l = \sum_{g=k-h+1}^{k-1}(-1)^{g+1}\binom{h-1}{g-k+h-1}(k-h+1+\frac{g-k+h-1}{k-g+1}) + (-1)^{k-1}(k-1) = (-1)^k + (-1)^k(k-h) + (h-2)(-1)^k + (-1)^{k-1}(k-1) = 0$. The result is the same if replacing x_1 with x_i , where i is from 2 to k , and replacing x_l with other x_i . Thus, all terms including μ can be canceled out. The proof is complete by noticing that the remaining part is $\lambda^k\psi_k(x_1, \dots, x_k)$. \square

Consider two continuous distributions belonging to the same location-scale family, their corresponding k th central moment kernel distributions only differ in scaling. So d is invariant, as shown in Subsection A. The robust k th central moment, based on rm , is defined by,

$$rkm_{d,\epsilon,n} := (d+1)Wkm_{\epsilon,n} - dmkm_{\epsilon,n},$$

where $Wkm_{\epsilon,n}$ is using the binomial k th central moment ($Bkm_{\epsilon,n}$) here, $mkm_{\epsilon,n}$ is the median k th central moment. Similarly, the quantile will not change after scaling. The quantile k th central moment is thus defined as

$$qkm_{d,\epsilon,n} := \hat{Q}_n\left(\left(pWkm - \frac{1}{2}\right)d + pWkm\right),$$

where $pWkm = \hat{F}_n(Wkm_{\epsilon,n})$, \hat{F}_n is the empirical cumulative distribution function of the corresponding central moment kernel distribution.

Finally, for standardized moments, quantile skewness and quantile kurtosis are defined to be $qskew_{d,\epsilon,n} := \frac{qtm_{d,\epsilon,n}}{qsd_{d,\epsilon,n}^3}$ and

$qkurt_{d,\epsilon,n} := \frac{qfm_{d,\epsilon,n}}{qsd_{d,\epsilon,n}^4}$. Quantile standard deviation ($qsd_{d,\epsilon,n}$),

robust standard deviation ($rsd_{d,\epsilon,n}$), quantile third central moment ($qtm_{d,\epsilon,n}$), quantile fourth central moment ($qfm_{d,\epsilon,n}$), robust third central moment ($rtm_{d,\epsilon,n}$), robust fourth central moment ($rfm_{d,\epsilon,n}$), robust skewness ($rskew_{d,\epsilon,n}$), and robust kurtosis ($rkurt_{d,\epsilon,n}$) are all defined similarly as above and not repeated here. The transformation to a location problem can also empower related statistical tests. From the better performance of the quantile mean in heavy-tailed distributions, quantile central moments are generally better than robust central moments regarding asymptotic bias.

To avoid confusion, the robust location estimations of the kernel distributions here are very different from Joly and Lugosi (2016) and Laforgue, Clemencon, and Bertail (2019)'s approach (32, 33), which is computing the median of all U -statistics from different blocks based on the median of means technique.

C. Congruent distribution. In the realm of nonparametric statistics, the precise value of a robust location estimator is of secondary importance. What is of primary importance is the relative difference between each group. The statement implicitly assumes that, in the absence of contamination, as the parameters of the distribution vary, the nonparametric robust location estimator will asymptotically change in the same

direction as the sample mean. Otherwise if the results based on trimmed mean are completely different from those based on median, a contradiction arises. A distribution satisfying this property for any symmetric weighted average is called a congruent distribution. If extending to any weighted average, it is strong congruent. A distribution is completely congruent if and only if it is congruent and its all central moment kernel distributions are also congruent. Complete strong congruence is analogous. From the definition, distributions with infinite moments are always not congruent. Also, Chebyshev's inequality implies that, for any probability distribution with finite moments, even if some weighted averages change in a direction different from that of the sample mean, the deviations are bounded. The following theorems show the conditions that a distribution is congruent or strong congruent by classifying distributions through inequalities.

Theorem C.1. *Let the symmetric quantile average function of a parametric distribution be denoted as $SQA(\epsilon, \alpha_1, \dots, \alpha_i, \dots, \alpha_k)$, where α_i represent the parameters of the distribution. This distribution is congruent if and only if the sign of $\frac{\partial SQA(\epsilon, \alpha_i)}{\partial \alpha_i}$ remains the same for all $0 < \epsilon < \frac{1}{2}$. Replacing SQA with Q constitutes a necessary and sufficient condition for the distribution to be considered strong congruent.*

Proof. Asymptotically, any symmetric weighted average can be expressed as an integral of the symmetric quantile average function. Since the sign won't change after integration, from definition, the sign of $\frac{\partial SQA(\epsilon, \alpha_i)}{\partial \alpha_i}$ remains the same for all $0 < \epsilon < \frac{1}{2}$ is equivalent to all symmetric weighted averages also change in the same direction as the sample mean since the sample mean is also a symmetric weighted average and can be expressed as $\int_0^{\frac{1}{2}} SQA(\epsilon) d\epsilon$. The same logic applies to the strong congruence case, as the constancy of the sign of $\frac{\partial Q(p, \alpha_i)}{\partial \alpha_i}$ for all $0 < p < 1$ is equivalent to the statement that all weighted averages also change in the same direction as the sample mean. The proof is finished. \square

Theorem C.2. *If a distribution is strong congruent, then it is congruent.*

Proof. From the definition in the previous article, any symmetric weighted average is also a weighted average. This concludes the proof. \square

Theorem C.3. *A symmetric distribution with a finite second moment is always congruent.*

Proof. For any symmetric distribution with a finite second moment, all symmetric quantile averages coincide. The conclusion follows immediately. \square

Theorem C.4. *A positive define location-scale distribution with a finite second moment is always strong congruent.*

Proof. As shown in discussions in Subsection A, for a location-scale distribution, any weighted average can be expressed as $\lambda W_{estimator}(\epsilon) + \mu$, where $W_{estimator}(\epsilon)$ is a function of $Q_0(p)$ according to the definition of the weighted average. Therefore, the derivatives with respect to the parameters λ or μ are always positive. By application of Theorem C.1, the desired outcome is obtained. \square

Theorem C.5. *The second central moment kernel distribution derived from a continuous location-scale unimodal distribution with a finite second moment is always strong congruent.*

Proof. Theorem B.3 shows that the corresponding central moment kernel distribution is also a location-scale family distribution. Theorem B.1 shows that it is positively defined. Implementing Theorem C.4 yields the desired result. \square

For the Pareto distribution, $\frac{\partial Q(p, \alpha)}{\partial \alpha} = \frac{x_m(1-p)^{-1/\alpha} \ln(1-p)}{\alpha^2}$. Since $\ln(1-p) < 0$ for all $0 < p < 1$, $(1-p)^{-1/\alpha} > 0$ for all $0 < p < 1$ and $\alpha > 0$, the Pareto distribution is strong congruent and therefore congruent. The derivative for the lognormal distribution is $\frac{\partial \text{SQ}(\epsilon, \sigma)}{\partial \sigma} = \frac{-\text{erfc}^{-1}(2\epsilon)e^{\mu-\sqrt{2}\sigma\text{erfc}^{-1}(2\epsilon)} - \text{erfc}^{-1}(2-2\epsilon)e^{\mu-\sqrt{2}\sigma\text{erfc}^{-1}(2-2\epsilon)}}{\sqrt{2}}$. Since the inverse complementary error function is positive when the input is smaller than 1, and negative when the input is larger than 1, $\text{erfc}^{-1}(2\epsilon) = -\text{erfc}^{-1}(2-2\epsilon)$, $e^{\mu-\sqrt{2}\sigma\text{erfc}^{-1}(2-2\epsilon)} > e^{\mu-\sqrt{2}\sigma\text{erfc}^{-1}(2\epsilon)}$, $\frac{\partial \text{SQ}(\epsilon, \sigma)}{\partial \sigma} > 0$, the lognormal distribution is congruent. It is not strong congruent, since the quantile function is $Q(p) = e^{\mu-\sqrt{2}\sigma\text{erfc}^{-1}(2p)}$, the derivative is $\frac{\partial Q(p, \sigma)}{\partial \sigma} = -\sqrt{2}\text{erfc}^{-1}(2p)e^{\mu-\sqrt{2}\sigma\text{erfc}^{-1}(2p)}$. Theorem C.3 implies that the generalized Gaussian distribution is congruent. Although the derivatives of the quantile functions of the gamma and generalized Gaussian distributions are very complex, just consider the Gaussian distribution, when the scale parameter increases, $Q(p)$ for $p > \frac{1}{2}$ will increase, for $p < \frac{1}{2}$ will decrease, therefore, they are not strong congruent distributions. For the Weibull distribution, just consider the median, $E[m] = \lambda \sqrt[3]{\ln(2)}$, $E[\mu] = \lambda \Gamma(1 + \frac{1}{\alpha})$, then, when $\alpha = 1$, $E[m] = \lambda \ln(2) \approx 0.693\lambda$, $E[\mu] = \lambda$, but when $\alpha = \frac{1}{2}$, $E[m] = \lambda \ln^2(2) \approx 0.480\lambda$, $E[\mu] = 2\lambda$, the mean increases, but the median decreases, therefore, it is not congruent. When α changes from 1 to $\frac{1}{2}$, the average probability density on the left side of median increases, since $\frac{\frac{1}{2}}{\lambda \ln(2)} < \frac{\frac{1}{2}}{\lambda \ln^2(2)}$, but the mean increases, meaning that the distribution is more heavy-tailed, the probability density of large values will also increase. The simultaneous increases in both probability densities, one close to zero and the other close to infinity, are the reason for non-congruence. Note that the gamma distribution does not have this issue, it looks to be congruent.

Although many common parametric distributions are not congruent, Theorem C.4 establishes that strong congruence always holds for a positive define location-scale family distribution and thus for the second central moment kernel distribution generated from a continuous location-scale unimodal distribution as shown in Theorem C.5. Theorem B.2 demonstrates that all their central moment kernel distributions are unimodal-like with mode and median close to zero, as long as they are unimodal distributions. This implies, align with Theorem B.3, that different kernel distributions mainly differ in scale and they are, in some senses, reduced to a location-scale family distribution. If $Q(0) - Q(1)$ remains constant, increasing the mean will result in a more heavy-tailed distribution, i.e., the probability density closer to $\frac{1}{k}(-\Delta)^k$ will increase. While the total probability density on either side of zero will remain unchanged as the median is generally close to zero and much less impacted during the mean increasing, the probability density close to zero will decrease. This transformation will also increase other symmetric weighted averages, in the general

sense, due to the heavy tail. As a result, nearly all symmetric weighted averages for all central moment kernel distributions derived from unimodal distributions should change in the same direction as the parameters change.

D. A two-parameter distribution as the consistent distribution.

Up to this point, the consistent robust estimation has been limited to a parametric location-scale distribution. The location parameter is often omitted for simplicity. A distribution specified by a shape parameter (denoted as α here) and a scale parameter (denoted as λ here) is often referred to as a two-parameter distribution. Weibull, gamma, Pareto, lognormal, and generalized Gaussian distributions are all two-parameter unimodal distributions. α can be converted to skewness or kurtosis, e.g., for the gamma distribution, the skewness is $\frac{2}{\sqrt{\alpha}}$, the kurtosis is $\frac{6}{\alpha} + 3$. If α is a constant, the two-parameter distribution is reduced to a single-parameter distribution. The discussion in Subsection A shows that, for a single-parameter distribution as the consistent distribution and a fixed ϵ , there should be a k -tuple (d_{im}, \dots, d_{ikm}) (using a distribution as the consistent distribution means the d values used are calibrated by the distribution or the corresponding kernel distributions generated from this distribution). For a two-parameter distribution, let $D(\text{kurtosis}, [\text{skewness}], k, \text{etype}, \text{dtype}, n) = d_{ikm}$ denote these relations, where the first input is the kurtosis, the second input is the square of the skewness, the third is the order of the central moment (if $k = 1$, the mean), the fourth is the type of estimator, the fifth is the type of consistent distribution, the sixth input is the sample size. For simplicity, the last three inputs will be omitted in the following discussion. Note that specifying d values of d for a two-parameter distribution requires only kurtosis or skewness.

Using a two-parameter distribution as the consistent distribution is a problem of robust estimation of parametric models. The object is to find solutions for the system of equations

$$\begin{cases} rm(WA, \text{median}, D(rkurt, |rskew|, 1)) = \mu \\ rvar(Wvar, mvar, D(rkurt, |rskew|, 2)) = \mu_2 \\ rtm(Wtm, mtm, D(rkurt, |rskew|, 3)) = \mu_3 \\ rfm(Wfm, mfm, D(rkurt, |rskew|, 4)) = \mu_4 \end{cases}, \text{ where } \begin{cases} rskew = \frac{\mu_3}{\mu_2} \\ rkurt = \frac{\mu_4}{\mu_2} \end{cases}$$

μ_3 and μ_4 are the population second, third and fourth central moments. $rkurt$ and $|rskew|$ should be the invariant points of the functions $\varkappa(rkurt) = \frac{rfm(Wfm, mfm, D(rkurt, 4))}{rvar(Wvar, mvar, D(rkurt, 2))^2}$ and

$\varsigma(|rskew|) = \left| \frac{rtm(Wtm, mtm, D(|rskew|, 3))}{rvar(Wvar, mvar, D(|rskew|, 2))^{\frac{3}{2}}} \right|$. Clearly, this is an overdetermined nonlinear system of equations, because the skewness and kurtosis are interrelated for a two-parameter distribution. As an overdetermined system constructed with random coefficients is almost always inconsistent, it is natural to optimize them separately using the fixed-point iteration (see Algorithm 1, only $rkurt$ is provided, others are the same).

The following theorem shows the validity of Algorithm 1.

Theorem D.1. *$rkurt$ and $|rskew|$, defined as the largest attracting fix points of the functions $\varkappa(rkurt)$ and $\varsigma(|rskew|)$, are consistent estimators for a completely congruent two-parameter distribution, as long as they are within the domain of D .*

Proof. Without loss of generality, only $rkurt$ is considered here,

while the logic for $|rskew|$ is the same. Even without knowing any details of D , from the definition, $\lim_{rkurt \rightarrow \infty} \frac{\kappa(rkurt)}{rkurt} =$

$$\lim_{rkurt \rightarrow \infty} \frac{\frac{rkurt \mu_{2cali}^2 - Wfm_{cali}}{Wfm_{cali} - mfm_{cali}} (Wfm - mfm) + Wfm}{rkurt \left(\frac{\mu_{2cali} - Wvar_{cali}}{Wvar_{cali} - mvar_{cali}} (Wvar - mvar) + Wvar \right)^2}.$$

Since Wfm_{cali} and mfm_{cali} are from the same kernel distribution as $\mu_{4cali} = rkurt \mu_{2cali}^2$, so an increase in μ_{4cali} will also result in an increase in Wfm_{cali} and hence $Wfm_{cali} \gg Wfm$. Furthermore, Theorem B.2 and qualitative discussion in Subsection B shows that mfm_{cali} is close to zero, the increases in Wfm_{cali} leads to an increase in $(Wfm_{cali} - mfm_{cali})$. According to the property of invariance, assuming $rkurt = \mu_{4cali}$,

$$\left(\frac{\mu_{2cali} - Wvar_{cali}}{Wvar_{cali} - mvar_{cali}} (Wvar - mvar) + Wvar \right)^2 > 1, \text{ then}$$

$$\lim_{rkurt \rightarrow \infty} \frac{\frac{rkurt - Wfm_{cali}}{Wfm_{cali} - mfm_{cali}} (Wfm - mfm) + Wfm}{rkurt} < 1.$$

As a result, if there is at least one fix point, let the largest one be fix_{max} , then it is attracting since $|\frac{\partial(\kappa(rkurt))}{\partial(rkurt)}| < 1$ for all $rkurt \in [fix_{max}, kurtosis_{max}]$. Asymptotically, consider any $rkurt > \mu_4$, $Wfm_{cali} > Wfm$, $mfm_{cali} > mfm$, $\frac{rkurt - Wfm_{cali}}{Wfm_{cali} - mfm_{cali}} (Wfm - mfm) + Wfm < rkurt$, the same logic applies, a consistent estimator must be the last attracting fix point, fix_{max} is the consistent estimator. \square

Algorithm 1 $rkurt$ for a two-parameter distribution

Input: D ; $Wvar$; Wfm ; $mvar$; mfm ; $maxit$; δ

Output: $rkurt_{i-1}$

$i = 0$

2: $rkurt_i \leftarrow \kappa(kurtosis_{max})$ \triangleright Using the maximum kurtosis available in D as an initial guess.

repeat

4: $i = i + 1$

$rkurt_{i-1} \leftarrow rkurt_i$

6: $rkurt_i \leftarrow \kappa(rkurt_{i-1})$

until $i > maxit$ or $|rkurt_i - rkurt_{i-1}| < \delta$ $\triangleright maxit$ is the maximum number of iterations, δ is a small positive number. Here, $maxit = 100$, $\delta = 10^{-30}$.

As a result of Theorem D.1, assuming continuity and congruence of the central moment kernel distributions, Algorithm 1 converges surely if a fix point exists within the domain of D . At this stage, D can only be approximated through a Monte Carlo study. Using linear interpolation can ensure continuity. A common encountered problem is that D has a domain depending on both the consistent distribution and the Monte Carlo study, and the iteration may halt at the boundary if the fix point is not within the domain. However, by setting a proper maximum number of iterations, the algorithm will return the boundary value which is optimal within D . For quantile moments, the logic is similar, if the percentiles do not exceed the breakdown point. If so, consistent estimation is impossible and the algorithm will stop due to the maximum number of iterations. The fix point iteration is, in principle, similar to the iterative reweighing in M-estimator, but an advantage of this algorithm is that the optimization is only related to the function of d value and is independent of the sample size (except for the quantile moments, which require re-computation of the quantile function, but this operation has a time complexity of $O(1)$ for a sorted sample). Since $|rskew|$

can specify d_{rm} after optimization, this enables the robust estimations of all four moments to reach a near-consistent level for all five unimodal distributions (Table 1, SI Dataset S1), just using the Weibull distribution as the consistent distribution.

E. Variance. As the fundamental theorem in statistics, the central limit theorem states that the standard deviation of the limiting form of the sampling distribution of the sample mean is $\frac{\sigma}{\sqrt{n}}$. The principle was later applied to the sampling distributions of the robust location estimators (2, 34–40) and it was found that the efficiencies of the robust location estimators are sometimes very different from the arithmetic mean. Daniell (1920) stated (34) that the comparison of efficiencies of the various kinds of estimators is useless unless they all tend to coincide asymptotically. Bickel and Lehmann argued, also in the landmark series (38, 39), that meaningful comparisons can be made by studying the standardized variance, asymptotic variances, and sharp lower bounds of these estimators. Here, the scaled standard error (SSE) is proposed to estimate the variances of all estimators, including robust/quantile moments, on a scale similar to that of the sample mean.

Definition E.1 (Scaled standard error). Let $\mathcal{M}_{s_i s_j} \in \mathbb{R}^{i \times j}$ denote the sample-by-statistics matrix, i.e., the first column is the main statistics of interest, $\widehat{\theta}_m$, the second to the j th column are $j - 1$ statistics required to scale, $\widehat{\theta}_{r_1}, \widehat{\theta}_{r_2}, \dots, \widehat{\theta}_{r_{j-1}}$. Then, the scaling factor $\mathcal{S} = \left[1, \frac{\widehat{\theta}_{r_1}}{\widehat{\theta}_m}, \frac{\widehat{\theta}_{r_2}}{\widehat{\theta}_m}, \dots, \frac{\widehat{\theta}_{r_{j-1}}}{\widehat{\theta}_m} \right]^T$ is a $j \times 1$ matrix, which $\bar{\theta}$ is the mean of the column. The normalized matrix is $\mathcal{M}_{s_i s_j}^N = \mathcal{M}_{s_i s_j} \mathcal{S}$. The SSEs are the unbiased standard deviations of the corresponding columns.

Setting the bootstrap moments as the main, the SSEs of all robust estimators proposed here are often between those of the median moments and those of the sample moments (SI Dataset S1). This is because similar monotonic relations between robustness and variance are also very common, e.g., Bickel and Lehmann (39) proved that a bound for the efficiency of TM_e to sample mean is $(1 - 2\epsilon)^2$ and this monotonic bound is valid for any distribution. Lai, Robbins, and Yu (1983) proposed an estimator that adaptively chooses the mean or median in a symmetric distribution and showed that the result is typically as good as the better of sample mean and median regarding variance (41). It may be interpreted as an attempt to use the variance version of mean-SWA-median inequality. While they used bootstrap standard error as the criterion, another approach can be dated back to Laplace (1812) (42) is using a linear combination of median and mean and the weight is deduced to achieve minimum variance; examples for symmetric distributions see Samuel-Cahn, Chan and He, and Damilano and Puig (43–45).

Scaled standard error enables the direct comparison of variances of different location estimators for asymmetric distributions. Here, two invariant means and related weighted averages ($BM_{\frac{1}{8}}, SQM_{\frac{1}{8}}, BM_{\nu=2, \epsilon=\frac{1}{8}}, WM_{\frac{1}{8}}, BWM_{\frac{1}{8}}$, and $TM_{\frac{1}{8}}$ used here) can create twelve possible combinations. Each combination has an SSE for a single-parameter distribution, which can be inferred by a Monte Carlo study. Then, among twelve possible combinations, there is one that has the smallest SSE (if the percentiles of quantile moments exceed the breakdown point, this combination will be excluded). Theoretically, bootstrap is the optimal way to infer the variance-optimal choice without distributional assumption, however, the com-

Table 1. The performance of invariant moments for five common unimodal distributions compared to current popular methods

Errors	TM _{1/8}	H-L	WM _{1/8}	HM	BM _{1/8}	rm _{1/8}	qm _{1/8}	Tsd _{1/8} ²	rvar _{1/8}	qvar _{1/8}	rtm _{1/8}	qtm _{1/8}	rfm _{1/8}	qfm _{1/8}
WAAB	0.128	0.109	0.078	0.102	0.057	0.002	0.004	0.237	0.031	0.013	0.025	0.009	0.071	0.019
WRMSE	0.131	0.111	0.082	0.105	0.061	0.017	0.019	0.235	0.037	0.025	0.030	0.018	0.073	0.027
WAB _{n=5400}	0.128	0.108	0.078	0.102	0.057	0.002	0.005	0.235	0.031	0.013	0.025	0.009	0.070	0.019
WSE ∨ WSSE	0.014	0.014	0.014	0.014	0.015	0.017	0.018	0.015	0.016	0.018	0.012	0.014	0.012	0.016

Errors	m	BWM _{1/8}	SQM _{1/8}	SM _{1/8}	\bar{x}	im _{v, 1/8}	var	ivar _{v, 1/8}	tm	itm _{v, 1/8}	fm	ifm _{v, 1/8}
WAAB	0.205	0.104	0.057	0.070	0.000	0.003	0.000	0.007	0.000	0.009	0.000	0.017
WRMSE	0.208	0.108	0.061	0.074	0.014	0.017	0.017	0.020	0.021	0.020	0.028	0.031
WAB _{n=5400}	0.205	0.104	0.057	0.070	0.000	0.003	0.000	0.006	0.000	0.010	0.001	0.016
WSE ∨ WSSE	0.016	0.014	0.015	0.014	0.014	0.016	0.017	0.018	0.021	0.014	0.026	0.020

Errors	\bar{x}	rm	im _v	var	var _{bs}	rvar	ivar _v	tm	tm _{bs}	rtm	itm _v	fm	fm _{bs}	rfm	ifm _v
RMSE	0.014	0.016	0.016	0.018	0.017	0.019	0.018	0.021	0.019	0.018	0.018	0.027	0.023	0.023	0.022
SE ∨ SSE	0.014	0.016	0.016	0.017	0.017	0.019	0.018	0.020	0.019	0.018	0.018	0.025	0.021	0.022	0.021

The first section uses the exponential distribution as the consistent distribution for five common unimodal distributions, Weibull, gamma, Pareto, lognormal and generalized Gaussian distribution, while the second section uses the Weibull distribution as the consistent distribution. The third section uses the Weibull distribution plus optimization (ikm_v is invariant k th moment, variance-optimized). The fourth section uses the Weibull distribution for the Weibull distribution; the breakdown points are all $\frac{1}{8}$ (not indicated). $BM_{1/8}$ is the weighted average used in robust/quantile moments. The average asymptotic bias (AAB, $n \rightarrow \infty$), root mean square error (RMSE, $n = 5400$), average bias (AB, $n = 5400$) and variance (SE ∨ SSE, $n = 5400$) of the estimators are in the units of the standard deviations of the kernel distributions. bs indicates the bootstrap central moments. W means that the results were weighted by the number of Google Scholar search results on May 30, 2022 (including synonyms). The calibrations of d values and the computations of AAB, AB, and SSE were described in Subsection E, F and SI Methods. The detailed results and related codes are available in SI Dataset S1 and [GitHub](#).

putational cost is very high. Similar to Subsection D, let $I(kurtosis, |skewness|, k, dtype, n) = ikm_{WA}$ denote these relations. Then since $\lim_{rkurt \rightarrow \infty} \frac{I(rkurt, 4)}{I(rkurt, 2)^2 rkurt} < 1$, the same fix point iteration algorithm can be used to choose the variance-optimum combination. The only problem is that unlike D , I is defined to be discontinuous and also simulated by a Monte Carlo study, but linear interpolation can also be used to ensure the continuity. Using this approach, the result is often very close to the optimum choice (SI Datasets S1).

In general, compared to the unbiased sample central moments, the variances of invariant central moments are much smaller (except the second central moment, Table 1). In 1958, Richtmyer proposed quasi-Monte Carlo simulation based on low-discrepancy sequences, which dramatically reduces the computational cost of large sample simulation (46). Quasi-Monte Carlo methods frequently employ Sobol sequences as the favored numerical sets (47). Do and Hall extended the principle to bootstrap in 1991 (48) and found that the quasi-random approach is competitive in terms of variance when compared with other bootstrap Monte Carlo procedures. By using quasi-sampling, the impact of the number of repetitions of the bootstrap, or bootstrap size, on variance is negligible. An estimator based on the quasi-bootstrap approach can be seen as a very complex deterministic estimator which is not only computationally efficient, but also statistical efficient. The only drawback is that a small bootstrap size can produce additional finite sample bias but can be corrected by recalibrating the d values (18 thousand is used here as default as it balances computational cost and finite sample bias, except the asymptotic value calculation).

F. Robustness. The measure of robustness to gross errors used here is the breakdown point proposed by Hampel (49) in 1968. However, the sample-dependent breakdown point has apparently not been defined previously.

Definition F.1 (Sample-dependent breakdown point). An estimator $\hat{\theta}$ has a sample-dependent breakdown point if and only if its asymptotic breakdown point $\epsilon(\hat{\theta}, R, \zeta, v)$ is zero and the empirical influence function of $\hat{\theta}$ is bounded, where R is the measure of badness, ζ is the contaminating processes, v is the uncontaminated process. For a full formal definition of the asymptotic breakdown point, which is the breakdown point when $n \rightarrow \infty$, and the empirical influence function, the reader is referred to Genton and Lucas (2003) and Devlin, Gnanadesikan and Kettenring (1975)'s papers (50, 51).

Bear in mind that it differs from the "infinitesimal robustness" defined by Hampel, which is related to whether the asymptotic influence function is bounded (52–54). The proof of the consistency of MoM assumes that it is an estimator with a sample-dependent breakdown point since its breakdown point is $\frac{\beta}{2n}$, where β is the number of blocks, then $\lim_{n \rightarrow \infty} \left(\frac{\beta}{2n}\right) = 0$, if β is a constant or $\beta \ll n$, and any changes in any one of the points of the sample cannot breakdown this estimator (55–57).

For the robust estimations of central moments or other robustified U -statistics based on a robust location estimator, the asymptotic breakdown points are suggested by the following theorem by extending the method in Donoho and Huber (1983)'s proof of the breakdown point of the Hodges-Lehmann estimator (58).

Theorem F.1. Given n independent random variables

(X_1, \dots, X_n) with the same distribution F and a U -statistic associated with a symmetric kernel of degree k . Then, assuming as $n \rightarrow \infty$, $k \ll n$, the asymptotic breakdown point of the robust location estimation of the distribution of the kernel of the U -statistic is $1 - (1 - \epsilon)^{\frac{1}{k}}$, where ϵ is the breakdown point of the symmetric robust location estimator.

Proof. According to the definition of ϵ -contamination (58), suppose m contaminants are added to the sample. The fraction of bad values in the sample is $\epsilon_U = \frac{m}{n+m}$, while the original n data points are not impacted. That means, in the distribution of the kernel, $\binom{n}{k}$ of total $\binom{n+m}{k}$ points are not corrupted. Then, the breakdown will not occur if the following inequality holds

$$\binom{n}{k} > \left(\frac{1}{\epsilon} - 1\right) \times \left(\binom{n+m}{k} - \binom{n}{k}\right),$$

Since ϵ is the breakdown point of the robust location estimator, $\frac{1}{2} \geq \epsilon \geq 0$,

$$\frac{1}{1-\epsilon} > \frac{\binom{n+m}{k}}{\binom{n}{k}} = \frac{(n+m)(n+m-1)\dots(n+m-k+1)}{n(n-1)\dots(n-k+1)}.$$

For asymptotic breakdown point, assuming $n \rightarrow \infty$, $k \ll n$, $\lim_{n \rightarrow \infty} \left(\frac{n+m-k+1}{n-k+1}\right) = \frac{n+m}{n} = x$, then the above inequality does not hold when $x \geq \left(\frac{1}{1-\epsilon}\right)^{\frac{1}{k}}$. So, the breakdown point of the U -statistic is $\epsilon_U = \frac{m}{n+m} = 1 - \frac{n}{n+m} = 1 - \frac{1}{x} = 1 - (1 - \epsilon)^{\frac{1}{k}}$. \square

Remark. If $k = 1$, $1 - (1 - \epsilon)^{\frac{1}{k}} = \epsilon$, so this formula also holds for the robust location estimator. In addition, the numerical solutions for $k = 2, 3, 4$, $\epsilon = \frac{1}{8}$ are $\approx 0.065, 0.044$, and 0.033 , respectively. When $\epsilon = \frac{1}{2}$, the robustified U -statistic becomes U -quantile, which converges almost surely as proven by Choudhury and Serfling (59) in 1988.

Every statistic is based on certain assumptions. For instance, the sample mean assumes that the second moment of the underlying distribution is finite. If this assumption is violated, the variance of the sample mean becomes infinitely large even the population mean is finite. Therefore, the sample mean not only has zero robustness to gross errors, but also has zero robustness to departures. If departures are unlimited, nearly any estimators can be broken, so posing a constraint on departures is necessary for comparison.

Bias bound (1) is the first approach to study the robustness to departures under regularity conditions, i.e., although all estimators can be biased under departures from the assumptions, but their standardized maximum biases can differ substantially (60, 61). In the previous semiparametric robust mean estimation article, it is shown that another way to qualitatively compare the estimators' robustness to departures from the symmetry assumption is constructing and comparing the corresponding semiparametric models. An estimator based on a smaller model is naturally more robust to asymmetric departures within that model. Although the comparison is limited to the smaller semiparametric model and is not universal, quite surprisingly, the results coincide with those obtained from the bias bound analysis. Bias bound is more universal since it is possible to deduce the bounds for distributions with finite moments without assuming unimodality (60, 61). However, the bias bounds are often hard to deduce for complex estimators.

Also, sometimes there are discrepancies between maximum bias and average bias. For example, the maximum bias of $rm_{d \approx 0.375, \epsilon = \frac{1}{8}}$ is higher than $SQM_{\frac{1}{8}}$, but it has much better average bias (SI Dataset S1). Since the estimators proposed here are all consistent under certain assumptions, measuring their biases is a convenient way of measuring the robustness to departures. Average asymptotic bias is thus defined as follows.

Definition F.2 (Average asymptotic bias). For a single-parameter distribution, the average asymptotic bias (AAB) is just the standardized asymptotic bias $\frac{|\hat{\theta} - \theta|}{\sigma}$, where $\hat{\theta}$ is the estimation of θ and σ is the population standard deviation of the distribution. For a two-parameter distribution, the first step is setting the lower bound of the kurtosis range of interest $\tilde{\mu}_{4l}$. Then, the average asymptotic bias is defined as

$$AAB_{\hat{\theta}} := \frac{1}{C} \sum_{\substack{\delta + \tilde{\mu}_{4l} \leq \tilde{\mu}_4 \leq C\delta + \tilde{\mu}_{4l} \\ \tilde{\mu}_4 \text{ is a multiple of } \delta}} E_{x|\tilde{\mu}_4} \left[\frac{|\hat{\theta} - \theta|}{\sigma} \right]$$

where $\tilde{\mu}_4$ is the kurtosis specifying the two-parameter distribution, $E_{x|\tilde{\mu}_4}$ denotes the expected value over $P(x|\tilde{\mu}_4)$.

Standardization is crucial for comparing the performances of estimators under different distributions. Currently, there are several options available, such as using the root mean square deviation from the mode (as in Gauss (1)), the mean absolute deviation, or standard deviation. The standard deviation is used here because of its central role in standard error estimation. The estimation of central moments based on the location estimations of the kernel distributions also enables the standardization of average biases (ABs, for finite sample scenarios) and average asymptotic biases (AABs) of robustified central moments. The only difference is that the population standard deviation is replaced by the asymptotic standard deviation of the kernel distribution (σ_{km}).

In Table 1, $\delta = 0.1$, $C = 120$. For the Weibull, gamma, lognormal and generalized Gaussian distributions, the kurtosis range is from 3 to 15 (there are two shape parameter solutions for the Weibull distribution, the lower one is used here). For the Pareto distribution, the range is from 9 to 21. To provide a more practical and straightforward illustration, all results from five distributions are further weighted by the number of Google Scholar search results. Interestingly, the asymptotic biases of $TM_{\frac{1}{8}}$ and $WM_{\frac{1}{8}}$, after averaging and weighting, are 0.128σ and 0.078σ , respectively, in line with the sharp bias bounds of $TM_{2,14:15}$ and $WM_{2,14:15}$ (a different subscript is used to indicate a sample size of 15, with the removal of the first and last order statistics.), 0.173σ and 0.126σ , for distributions with finite moments, without assuming unimodality (60, 61). This setting seems arbitrary, however, the quantile inequality ensures that the order among different SWAs remains generally the same as the parameters change, the range of kurtosis is in fact not very important for AAB. It is important if using the maximum biases within the range of kurtosis among all five unimodal distributions as a measure of robustness to departures, because different estimators reach their maximum biases at different parameters. Within the range of kurtosis setting, nearly all SWAs and $SWkms$ proposed here reach or at least close to their maximum biases (SI Dataset S1). The pseudo-maximum bias is thus defined as the maximum value of the biases in the AAB computations for all five unimodal distributions. In most cases, the pseudo-maximum biases of

invariant moments occur in lognormal or generalized Gaussian distributions (SI Dataset S1), since besides unimodality, the Weibull distribution differs entirely from them.

Discussion

Moments, including raw moments, central moments, and standardized moments, are key parameters that determine probability distributions. Central moments are much more popular than raw moments because they are invariant to translation. In 1947, Hsu and Robbins proved that the arithmetic mean converges completely to the population mean provided the second moment is finite (62). The strong law of large numbers (proven by Kolmogorov in 1933) (63) implies that the k th sample central moment is asymptotically unbiased. Recently, fascinating statistical phenomena about Taylor's law for distributions with infinite moments were found by Drton and Xiao (2016) (64), Pillai and Meng (2016) (65), Cohen, Davis, and Samorodnitsky (2020) (66), and Brown, Cohen, Tang, and Yam (2021) (67). Lindquist and Rachev (2021) commented: "What are the proper measures for the location, spread, asymmetry, and dependence (association) for random samples with infinite mean?" (68). This is not the focus of this paper, but it is almost sure that the estimators proposed here will have a place. For example, they (68) suggested using median, interquartile range, and medcouple (69) as the robust versions of the first three moments (70–72). Obviously now, if one wants to preserve the original relationship between each moment while ensuring maximum robustness, the natural choices are median, median variance, and median third central moment. Analogously to the most robust version of L-moment (73) being trimmed L-moment (74), the central moment now also has its standard most robust version based on the congruence of the central moment kernel distributions generated from unimodal distributions.

More generally, parametrics, nonparametrics, and semiparametrics are the current three main branches of statistics. Consistent robust estimation is impossible without specific parametric assumptions. Maximum likelihood was first introduced by Fisher in 1922 (75) in a multinomial model and later generalized by Cramér (1946), Hájek (1970), and Le Cam (1972) (31, 37, 76). Besides Newcomb (2, 3), the general robust estimation of parametric models dates back to 1939, when Wald (77) suggested the use of minimax estimates to solve such problems. Hodges and Lehmann in 1950 (78) expanded upon this concept and obtained minimax estimates for a series of important problems. It was soon clear that a minimax estimator should be a Bayes estimator with regard to the least favorable prior distribution of θ as a minimax estimator is the best in the worst case scenario. Following Huber's seminal work (4), M -statistics have dominated the field of parametric robust statistics for over half a century. In 1984, Bickel addressed the challenge of robustly estimating the parameters of a linear model while acknowledging the possibility that the model may be invalid, but still within the confines of a larger model (79, 80). As the title *Parametric Robustness: Small Biases can be Worthwhile* suggests, biases exists, but by carefully designing the estimators, they can be very small. The study of semiparametric models was initiated by Stein (81) (1956). Estimation of the center of symmetry for an unknown symmetric distribution is an important example in his paper. The adaptive estimation of the center of symmetry was studied

by van Eeden (1970) and Takeuchi (1971) (82, 83). Bickel, in 1982, simplified the general heuristic necessary condition proposed by Stein (81) (1956) and derived sufficient conditions for this type of problem, adaptive estimation (84). As pointed out in Begun, Hall, Huang, and Wellner's paper (1983) and their semiparametrics textbook with Bickel, Klaassen, and Ritov (1993) (85, 86), the two problems, semiparametrics (or adaptive estimation) and parametric robustness, are closely related but different. The paradigm shift here opens up the possibility that by defining a large semiparametric model, constructing estimators simultaneously for two or more very different semiparametric/parametric models within the large semiparametric model, then even for a seemingly "wrong" parametric model belongs to the large semiparametric model but not to the semiparametric/parametric models used for calibration, their performance might still be near-optimal due to the common nature shared by the models used by the estimators. The models can be directly expanded and not limited to a single parametric form. Maybe it can be named as comparametrics. Closely related topics are "mixture model" and "constraint defined model" generalized in (86) and the method of sieves, introduced by Grenander in 1981 (87). Furthermore, it shows that by utilizing the invariant structure of probability distributions and their measures, it is possible to construct consistent robust estimators without considering the residuals. Maybe this class of estimators can be named as I -statistics. As building blocks of statistics, invariant moments provide an option for estimating distribution parameters robustly and near-consistently with moderate variances under mild assumptions. This can improve the consistency of statistical results across studies, particularly when heavy-tailed distributions may be present (88–92).

Data Availability. Data for Table 1 are given in SI Dataset S1. All codes have been deposited in [GitHub](#).

ACKNOWLEDGMENTS. I gratefully acknowledge the constructive comments made by the editor which substantially improved the clarity and quality of this paper.

1. CF Gauss, *Theoria combinationis observationum erroribus minimis obnoxiae*. (Henricus Dieterich), (1823).
2. S Newcomb, A generalized theory of the combination of observations so as to obtain the best result. *Am. Journal Math.* 8, 343–366 (1886).
3. S Newcomb, Researches on the motion of the moon. part ii, the mean motion of the moon and other astronomical elements derived from observations of eclipses and occultations extending from the period of the babylonians until ad 1908. *United States. Naut. Alm. Off. Astron. paper*; v. 9.9, 1 (1912).
4. PJ Huber, Robust estimation of a location parameter. *Ann. Math. Stat.* 35, 73–101 (1964).
5. X He, WK Fung, Method of medians for lifetime data with weibull models. *Stat. medicine* 18, 1993–2009 (1999).
6. T Seki, S Yokoyama, Simple and robust estimation of the weibull parameters. *Microelectron. Reliab.* 33, 45–52 (1993).
7. D Olive, Robust estimators for transformed location scale families. *Unpubl. manuscript available from (www.math.siu.edu/olive/preprints.htm)* (2006).
8. RV Hogg, Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *J. Am. Stat. Assoc.* 69, 909–923 (1974).
9. RJ Hyndman, Y Fan, Sample quantiles in statistical packages. *The Am. Stat.* 50, 361–365 (1996).
10. C Bernard, R Kazzi, S Vanduffel, Range value-at-risk bounds for unimodal distributions under partial information. *Insur. Math. Econ.* 94, 9–24 (2020).
11. PJ Bickel, EL Lehmann, Descriptive statistics for nonparametric models. iii. dispersion in *Selected works of EL Lehmann*. (Springer), pp. 499–518 (2012).
12. PJ Bickel, EL Lehmann, Descriptive statistics for nonparametric models iv. spread in *Selected Works of EL Lehmann*. (Springer), pp. 519–526 (2012).
13. PJ Rousseeuw, C Croux, Alternatives to the median absolute deviation. *J. Am. Stat. association* 88, 1273–1283 (1993).
14. P Janssen, R Serfling, N Veraverbeke, Asymptotic normality of u-statistics based on trimmed samples. *J. statistical planning inference* 16, 63–74 (1987).
15. J Hodges, E Lehmann, Matching in paired comparisons. *The Annals Math. Stat.* 25, 787–791 (1954).

- 1038 16. S Dharmadhikari, K Jogdeo, Unimodal laws and related in *A Festschrift For Erich L. Lehmann*. (CRC Press), p. 131 (1982).
- 1039 17. AY Khintchine, On unimodal distributions. *Izv. Nauchno-Issled. Inst. Mat. Mech.* **2**, 1–7 (1938).
- 1040 18. S Purkayastha, Simple proofs of two results on convolutions of unimodal distributions. *Stat. & probability letters* **39**, 97–100 (1998).
- 1041 19. RA Fisher, Moments and product moments of sampling distributions. *Proc. Lond. Math. Soc.* **2**, 199–238 (1930).
- 1042 20. PR Halmos, The theory of unbiased estimation. *The Annals Math. Stat.* **17**, 34–43 (1946).
- 1043 21. W Hoeffding, A class of statistics with asymptotically normal distribution. *The Annals Math. Stat.* **19**, 293–325 (1948).
- 1044 22. PM Heffernan, Unbiased estimation of central moments by using u-statistics. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.)* **59**, 861–863 (1997).
- 1045 23. D Fraser, Completeness of order statistics. *Can. J. Math.* **6**, 42–45 (1954).
- 1046 24. AJ Lee, *U-statistics: Theory and Practice*. (Routledge), (2019).
- 1047 25. RJ Serfling, Generalized l-, m-, and r-statistics. *The Annals Stat.* **12**, 76–86 (1984).
- 1048 26. I Gijbels, P Janssen, N Veraverbeke, Weak and strong representations for trimmed u-statistics. *Probab. theory related fields* **77**, 179–194 (1988).
- 1049 27. B Efron, Bootstrap methods: Another look at the jackknife. *The Annals Stat.* **7**, 1–26 (1979).
- 1050 28. PJ Bickel, DA Freedman, Some asymptotic theory for the bootstrap. *The annals statistics* **9**, 1196–1217 (1981).
- 1051 29. KB Athreya, M Ghosh, LY Low, PK Sen, Laws of large numbers for bootstrapped u-statistics. *J. statistical planning inference* **9**, 185–194 (1984).
- 1052 30. R Helmers, P Janssen, N Veraverbeke, *Bootstrapping U-quantiles*. (CWI. Department of Operations Research, Statistics, and System Theory [BS]), (1990).
- 1053 31. H Cramér, *Mathematical methods of statistics*. (Princeton university press) Vol. 43, (1999).
- 1054 32. E Joly, G Lugosi, Robust estimation of u-statistics. *Stoch. Process. their Appl.* **126**, 3760–3773 (2016).
- 1055 33. P Laforge, S Cléménçon, P Bertail, On medians of (randomized) pairwise means in *International Conference on Machine Learning*. (PMLR), pp. 1272–1281 (2019).
- 1056 34. P Daniell, Observations weighted according to order. *Am. J. Math.* **42**, 222–236 (1920).
- 1057 35. PJ Bickel, et al., Some contributions to the theory of order statistics in *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability*. Vol. 1, pp. 575–591 (1967).
- 1058 36. H Chernoff, JL Gastwirth, MV Johns, Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation. *The Annals Math. Stat.* **38**, 52–72 (1967).
- 1059 37. L LeCam, On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals Math. Stat.* **41**, 802–828 (1970).
- 1060 38. P Bickel, E Lehmann, Descriptive statistics for nonparametric models i. introduction in *Selected Works of EL Lehmann*. (Springer), pp. 465–471 (2012).
- 1061 39. PJ Bickel, EL Lehmann, Descriptive statistics for nonparametric models ii. location in *selected works of EL Lehmann*. (Springer), pp. 473–497 (2012).
- 1062 40. P Janssen, R Serfling, N Veraverbeke, Asymptotic normality for a general class of statistical functions and applications to measures of spread. *The Annals Stat.* **12**, 1369–1379 (1984).
- 1063 41. T Lai, H Robbins, K Yu, Adaptive choice of mean or median in estimating the center of a symmetric distribution. *Proc. Natl. Acad. Sci.* **80**, 5803–5806 (1983).
- 1064 42. PS Laplace, *Theorie analytique des probabilités*. (1812).
- 1065 43. E Samuel-Cahn, Combining unbiased estimators. *The Am. Stat.* **48**, 34 (1994).
- 1066 44. Y Chan, X He, A simple and competitive estimator of location. *Stat. & Probab. Lett.* **19**, 137–142 (1994).
- 1067 45. G Damilano, P Puig, Efficiency of a linear combination of the median and the sample mean: The double truncated normal distribution. *Scand. J. Stat.* **31**, 629–637 (2004).
- 1068 46. RD Richtmyer, A non-random sampling method, based on congruences, for "monte carlo" problems, (New York Univ., New York. Atomic Energy Commission Computing and Applied ...), Technical report (1958).
- 1069 47. IM Sobol', On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki* **7**, 784–802 (1967).
- 1070 48. KA Do, P Hall, Quasi-random resampling for the bootstrap. *Stat. Comput.* **1**, 13–22 (1991).
- 1071 49. FR Hampel, *Contributions to the theory of robust estimation*. (University of California, Berkeley), (1968).
- 1072 50. MG Genton, A Lucas, Comprehensive definitions of breakdown points for independent and dependent observations. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.)* **65**, 81–94 (2003).
- 1073 51. SJ Devlin, R Gnanadesikan, JR Kettenring, Robust estimation and outlier detection with correlation coefficients. *Biometrika* **62**, 531–545 (1975).
- 1074 52. FR Hampel, A general qualitative definition of robustness. *The annals mathematical statistics* **42**, 1887–1896 (1971).
- 1075 53. FR Hampel, The influence curve and its role in robust estimation. *J. american statistical association* **69**, 383–393 (1974).
- 1076 54. PJ Rousseeuw, FR Hampel, EM Ronchetti, WA Stahel, *Robust statistics: the approach based on influence functions*. (John Wiley & Sons), (2011).
- 1077 55. AS Nemirovskij, DB Yudin, *Problem complexity and method efficiency in optimization*. (Wiley-Interscience), (1983).
- 1078 56. MR Jerrum, LG Valiant, VV Vazirani, Random generation of combinatorial structures from a uniform distribution. *Theor. computer science* **43**, 169–188 (1986).
- 1079 57. N Alon, Y Matias, M Szegedy, The space complexity of approximating the frequency moments in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*. pp. 20–29 (1996).
- 1080 58. DL Donoho, PJ Huber, The notion of breakdown point. *A festschrift for Erich L. Lehmann* **157184** (1983).
- 1081 59. J Choudhury, R Serfling, Generalized order statistics, bahadur representations, and sequential nonparametric fixed-width confidence intervals. *J. Stat. Plan. Inference* **19**, 269–282 (1988).
- 1082 60. M Bieniek, Comparison of the bias of trimmed and winsorized means. *Commun. Stat. Methods* **45**, 6641–6650 (2016).
- 1083 61. K Danielak, T Rychlik, Theory & methods: Exact bounds for the bias of trimmed means. *Aust. & New Zealand J. Stat.* **45**, 83–96 (2003).
- 1084 62. PL Hsu, H Robbins, Complete convergence and the law of large numbers. *Proc. national academy sciences* **33**, 25–31 (1947).
- 1085 63. A Kolmogorov, Sulla determinazione empirica di una legge di distribuzione. *Inst. Ital. Attuari, Giorn.* **4**, 83–91 (1933).
- 1086 64. M Drton, H Xiao, Wald tests of singular hypotheses. *Bernoulli* **22**, 38–59 (2016).
- 1087 65. NS Pillai, XL Meng, An unexpected encounter with cauchy and lévy. *The Annals Stat.* **44**, 2089–2097 (2016).
- 1088 66. JE Cohen, RA Davis, G Samorodnitsky, Heavy-tailed distributions, correlations, kurtosis and Taylor's law of fluctuation scaling. *Proc. Royal Soc. A* **476**, 20200610 (2020).
- 1089 67. M Brown, JE Cohen, CF Tang, SCP Yam, Taylor's law of fluctuation scaling for semivariances and higher moments of heavy-tailed data. *Proc. Natl. Acad. Sci.* **118**, e2108031118 (2021).
- 1090 68. WB Lindquist, ST Rachev, Taylor's law and heavy-tailed distributions. *Proc. Natl. Acad. Sci.* **118**, e2118893118 (2021).
- 1091 69. G Brys, M Hubert, A Struyf, A robust measure of skewness. *J. Comput. Graph. Stat.* **13**, 996–1017 (2004).
- 1092 70. DC Hoaglin, F Mosteller, JW Tukey, *Exploring data tables, trends, and shapes*. (John Wiley & Sons), (2011).
- 1093 71. PJ Huber, Wiley series in probability and mathematics statistics. *Robust statistics* pp. 309–312 (1981).
- 1094 72. RA Maronna, RD Martin, VJ Yohai, M Salibián-Barrera, *Robust statistics: theory and methods (with R)*. (John Wiley & Sons), (2019).
- 1095 73. JR Hosking, L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *J. Royal Stat. Soc. Ser. B (Methodological)* **52**, 105–124 (1990).
- 1096 74. EA Elamir, AH Seheult, Trimmed l-moments. *Comput. Stat. & Data Analysis* **43**, 299–314 (2003).
- 1097 75. RA Fisher, On the mathematical foundations of theoretical statistics. *Philos. transactions Royal Soc. London. Ser. A, containing papers a mathematical or physical character* **222**, 309–368 (1922).
- 1098 76. J Hájek, Local asymptotic minimax and admissibility in estimation in *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*. Vol. 1, pp. 175–194 (1972).
- 1099 77. A Wald, Contributions to the theory of statistical estimation and testing hypotheses. *The Annals Math. Stat.* **10**, 299–326 (1939).
- 1100 78. J Hodges, EL Lehmann, Some problems in minimax point estimation in *Selected Works of EL Lehmann*. (Springer), pp. 15–30 (2012).
- 1101 79. P Bickel, Parametric robustness: small biases can be worthwhile. *The Annals Stat.* **12**, 864–879 (1984).
- 1102 80. PJ Bickel, Robust regression based on infinitesimal neighbourhoods. *The Annals Stat.* pp. 1349–1368 (1984).
- 1103 81. C Stein, et al., Efficient nonparametric testing and estimation in *Proceedings of the third Berkeley symposium on mathematical statistics and probability*. Vol. 1, pp. 187–195 (1956).
- 1104 82. C Van Eeden, Efficiency-robust estimation of location. *The Annals Math. Stat.* **41**, 172–181 (1970).
- 1105 83. K Takeuchi, A uniformly asymptotically efficient estimator of a location parameter. *J. Am. Stat. Assoc.* **66**, 292–301 (1971).
- 1106 84. PJ Bickel, On adaptive estimation. *The Annals Stat.* **10**, 647–671 (1982).
- 1107 85. JM Begun, WJ Hall, WM Huang, JA Wellner, Information and asymptotic efficiency in parametric-nonparametric models. *The Annals Stat.* **11**, 432–452 (1983).
- 1108 86. P Bickel, CA Klaassen, Y Ritov, JA Wellner, *Efficient and adaptive estimation for semiparametric models*. (Springer) Vol. 4, (1993).
- 1109 87. U Grenander, *Abstract Inference*. (1981).
- 1110 88. JT Leek, RD Peng, Reproducible research can still be wrong: adopting a prevention approach. *Proc. Natl. Acad. Sci.* **112**, 1645–1646 (2015).
- 1111 89. B Baribault, et al., Metastudies for robust tests of theory. *Proc. Natl. Acad. Sci.* **115**, 2607–2612 (2018).
- 1112 90. MJ Schuemie, G Hripcsak, PB Ryan, D Madigan, MA Suchard, Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc. Natl. Acad. Sci.* **115**, 2571–2577 (2018).
- 1113 91. P Patil, G Parmigiani, Training replicable predictors in multiple studies. *Proc. Natl. Acad. Sci.* **115**, 2578–2583 (2018).
- 1114 92. E National Academies of Sciences, et al., *Reproducibility and Replicability in Science*. (National Academies Press), (2019).