Variable screening based on Gaussian Centered L-moments

Hyowon An[a,*], Kai Zhang[a], Hannu Oja[b], J. S. Marron[a]

[a] *The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA*
[b] *University of Turku, 20500 Turku, FI*

Abstract

An important challenge in big data is identification of important variables. For this purpose, methods of discovering variables with non-standard univariate marginal distributions are proposed. The conventional moments based summary statistics can be well-adopted, but their sensitivity to outliers can lead to selection based on a few outliers rather than distributional shape such as bimodality. To address this type of non-robustness, the L-moments are considered. Using these in practice, however, has a limitation since they do not take zero values at the Gaussian distributions to which the shape of a marginal distribution is most naturally compared. As a remedy, Gaussian Centered L-moments are proposed, which share advantages of the L-moments but have zeros at the Gaussian distributions. The strength of Gaussian Centered L-moments over other conventional moments is shown in theoretical and practical aspects such as their performances in screening important genes in cancer genetics data.

Keywords: Robust statistics; L-moments; L-statistics; skewness; kurtosis

Declarations of interest: none.

1. Introduction

Data quality is an issue that is currently not receiving as much attention as it deserves in the age of big data. Traditional analysis of small data sets involves a study of marginal distributions, which easily finds data quality challenges such as skewness and suggests remedies such as the Box-Cox transformation (Box and Cox, 1964). Direct implementation of this type of operation is challenging with high dimensional data, as there are too many marginal distributions to individually visualize. This hurdle can be overcome by using summary statistics to screen a representative set for visualization and potential remediation.

*Corresponding author. Postal address: 615 Pavonia Avenue, APT 1110, Jersey City, NJ 07306. Email: ahwbest@gmail.com. Tel: 919-360-8713.

Conventional summaries such as the *sample skewness* $\hat{\gamma}_1$ and *sample kurtosis* $\hat{\gamma}_2$ defined as

$$\hat{\gamma}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)^{3/2}}, \quad \hat{\gamma}_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)^{2}} - 3,$$

respectively, can be very useful for this screening process. However, as studied in Brys et al. (2004, 2006) and Kim and White (2004), those have limitations for the purpose, e.g. they can strongly be influenced by outliers. In some situations, outliers are well worth finding. But in other cases, summaries that are dominated by outliers can miss important distributional features such as skewness and bimodality. Variables with such features can be of keen interest in cancer research.

In this paper, the limitation of conventional summary statistics is demonstrated using a modern high dimensional data set from cancer research. These data are part of the TCGA project (Weinstein et al., 2013), and the precise version of the data here was used in Feng et al. (2016). The data include gene expression profiles of 16,615 genes and 817 breast cancer patients, each of whom is classified according to five cancer subtypes of major importance in cancer treatment. While much is known about this data, as discussed in Feng et al. (2016), the sheer data size means that there have only been cursory studies of the marginal, or individual gene, distributions. In this study, we conduct a much deeper search for genes with unexpected marginal structure such as skewness and kurtosis. This can yield relationships between genes and biologically meaningful features such as breast cancer subtypes.

The top two rows of Figure 1 show the marginal distributions of seven variables, i.e. genes, with the smallest conventional sample skewness values. The upper left plot shows the sample quantile curve of these summary values as a function of their ranks. The left arrow in the plot indicates that the seven marginal distribution plots have the smallest sample skewness values. The remaining plots are sorted in an ascending order of the sample skewness values that are given near the top of each plot. Each symbol of a different color represents a breast cancer patient by subtypes; see Table 1. The black solid lines are kernel density estimates of marginal distributions and colored solid lines are sub-densities corresponding to different subtypes. As detailed in Section 5.1 of Marron and Dryden (2021), the combination of data overlays with sub-density estimates provides particularly useful general insights into marginal distributions, and especially here, emphasizes data features such as outliers.

Figure 1 shows that even though the genes with the smallest sample skewness

Subtype	LumA	LumB	Her2	Basal	Normal-like
Symbol	+	×	*	◁	▷

Table 1: The symbols and colors corresponding to the five breast cancer subtypes in the marginal distribution plots.

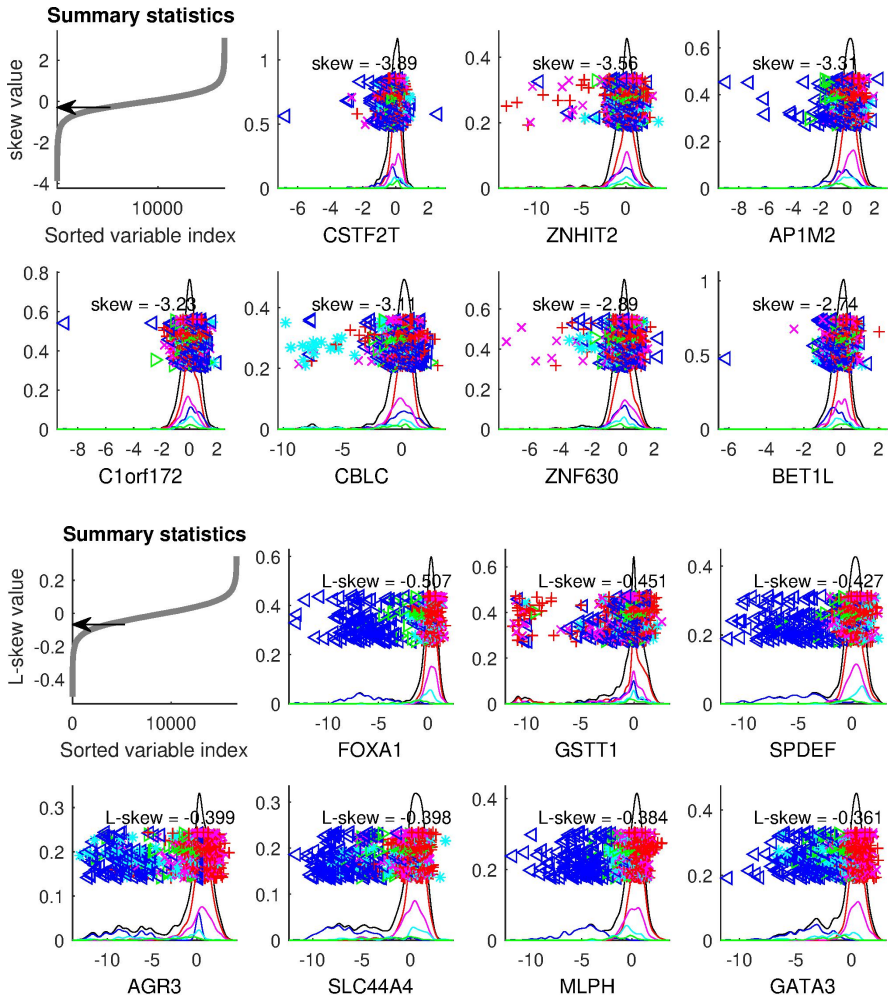

Figure 1: The marginal distribution plots of seven genes with the smallest conventional sample skewness (top two rows) and sample L-skewness values (bottom two rows). Each of the seven genes in the top two rows is driven by a few strong outliers that tend to obscure distributional shapes, while the seven genes in the bottom two rows seem to appear because of their distributional structure.

values were screened, the genes such as CSTF2T, C1orf172 and BET1L have a couple of outliers on their left sides rather than distributional skewness to the left. The sample skewness seems inadequate for effectively screening interesting genes in terms of distributional skewness. This challenge is well addressed using the *sample L-skewness* proposed in Hosking (1990) as shown in the bottom two rows of Figure 1. All the genes screened by the sample L-skewness except GSTT1 have clusters of Basal-type samples (blue triangles). These show that the sample L-skewness screens genes with distributional skewness especially

coming from cancer subtypes, which is desirable since the former is clinically much more important than the latter.

Investigation about distributional shape is often performed relative to the Gaussian distributions, since many data are aggregations of small and independent errors which approximately follow the Gaussian distributions by the Central Limit Theorem. However, as discussed in Hosking (1990), the population *L-moments* have zero values at the uniform distributions, making their signs and absolute values measure directions and magnitude of departure from the uniform distributions, respectively. The paper proposed using the sample L-skewness to perform the goodness-of-fit test for Gaussianity against skewed distributions, but did not consider the L-kurtosis against kurtotic alternative distributions.

These ideas motivate development of improved Gaussian centered versions of the L-moments. The unpublished results Decurninge (2014) discovered a variant of the L-moments called the *Hermite L-moments*. While the Hermite L-moments were shown to have strength in multivariate analysis, no attention was paid to their potential for univariate summaries, nor was their distributional center at the Gaussian distributions mentioned. In this paper, we comprehensively investigate possibilities of the HL-moments as univariate summaries. To further study their usefulness, we investigate their theoretical robustness and consistent estimators.

Another simple approach to shift the center of the L-moments is to subtract the population *L-kurtosis* (Hosking, 1990) value at the Gaussian distributions from itself. However, this can result in the loss of theoretical soundness of the definition of the L-moments since they are no longer differences of expected order statistics as in Equation (2.1) of Hosking (1990). We propose an alternative definition of the L-moments by which the numerical subtraction can be understood as a theoretically principled shift of the zeros of the L-moments to the Gaussian distributions.

These *Gaussian Centered L-moments* are developed in Section 3 as new univariate summaries. Their theoretical properties such as consistency and robustness are studied in Sections 4 and 5, respectively. Their abilities to screen variables with interesting marginal distributions are quantitatively analyzed in Section 6. A computational study that shows strength of a proposed estimator is given in Section 7. Section 8 summarizes our findings and presents potential future considerations.

2. Mathematical preliminaries

Assume that a random variable X follows a cumulative distribution function F with a probability density function f. We denote the distribution of $aX + b$ for $a \neq 0$ and $b \in \mathbb{R}$ by $F_{a,b}$. Also, a random sample $X_1, X_2, \cdots, X_n$ is assumed to be generated from F, and $X_{i:n}$ denotes the i-th order statistic of the random sample. In this paper we consider only an absolutely continuous and strictly increasing cumulative distribution function, in the sense that it is strictly increasing on its support $S(F) = \overline{\{x|0 < F(x) < 1\}}$ where $\overline{A}$ indicates

the closure of a set $A \subset \mathbb{R}$. Let $\mathcal{F}$ be the class of such distribution functions and the quantile function $F^{-1} : (0, 1) \to \mathbb{R}$ be the inverse of $F \in \mathcal{F}$. It is always assumed that the composition of G^{-1} and F, $G^{-1} \circ F$, where $G \in \mathcal{F}$, is defined on $S(F)$. When F is symmetric, we denote its *point of symmetry* (Doksum, 1975) by $m(F)$.

Various kinds of orthogonal polynomials are used throughout this paper. One is the *shifted Legendre polynomials* $\{P_r^*|r = 1, 2, \cdots\}$ which have been comprehensively investigated in Chapter 4 of Szegö (1959). The shifted Legendre polynomials are orthogonal to each other on the unit interval $(0, 1)$ with respect to the weight function $w(x) = 1$. Other orthogonal polynomials are the *Hermite polynomials* $\{H_r|r = 1, 2, \cdots\}$ presented in Chapter 5 of Szegö (1959), which are orthogonal to each other on the real line $\mathbb{R}$ with respect to the weight function $w(x) = e^{-x^2/2}$.

2.1. L-statistics and L-moments

The term *L-statistic* is used to indicate a statistic in the form of a *linear combination of order statistics*, which is generally expressed as

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} c_{n,i} X_{i:n}, \tag{1}$$

where $c_{n,i}$ is a function of the sample size n and the rank i of the order statistic $X_{i:n}$. Section 11.4 of David and Nagaraja (2003) surveyed the literature on various sets of conditions on the coefficients $\{c_{n,i}|n \geq 1, 1 \leq i \leq n\}$ and the distribution function F which ensure that $\hat{\theta}_n$ almost surely converges in the limit as $n \to \infty$ to the quantity

$$\theta_J(F) = \int_{-\infty}^{\infty} x f(x) J(F(x)) \, \mathrm{d}x = \int_0^1 F^{-1}(u) J(u) \mathrm{d}u, \tag{2}$$

where $J : (0, 1) \to \mathbb{R}$ is a measurable function. We call a functional in the form of Equation (2) an *L-functional*, the term used in Necir and Meraghni (2010).

A connection between L-statistics and location, scale, skewness and kurtosis of a distribution has been made by Hosking (1990). The population L-moments were defined in Hosking (1990) as

$$\lambda_r = \int_{-\infty}^{\infty} x f(x) P_{r-1}^*(F(x)) \mathrm{d}x = \int_0^1 F^{-1}(u) P_{r-1}^*(u) \mathrm{d}u, \tag{3}$$

which shows that the L-moments are L-functionals. Regarding computation of these population L-moments, very useful methods were presented in Dutang (2017). The population *L-skewness* and *L-kurtosis* are defined as $\lambda_3^* = \lambda_3/\lambda_2$ and $\lambda_4^* = \lambda_4/\lambda_2$, respectively. Based on the *sample L-moments* $\hat{\lambda}_{n,r}$ in Equation (3.1) of Hosking (1990), the sample L-skewness and *sample L-kurtosis* are defined as $\hat{\lambda}_{n,3}^* = \hat{\lambda}_{n,3}/\hat{\lambda}_{n,2}$ and $\hat{\lambda}_{n,4}^* = \hat{\lambda}_{n,4}/\hat{\lambda}_{n,2}$, respectively.

2.2. Oja's criteria

When defining new measures of location, scale, skewness and kurtosis, a challenge is to ensure that the new measures reflect the intuitive meaning of those distributional properties. This challenge is addressed by Oja (1981) using stochastic dominance ideas, whose third and fourth criteria regarding skewness and kurtosis are presented here.

Definition 2.1 (Oja (1981))**.** The functional $\theta : \mathcal{F} \to \mathbb{R}$ is a

a. *measure of skewness* in $\mathcal{F}$ if $\theta(F_{a,b}) = \text{sign}(a)\theta(F)$ for all $a \neq 0, b \in \mathbb{R}, F \in \mathcal{F}$ and $\theta(F) \leq \theta(G)$ whenever $G^{-1} \circ F$ is convex (of order 2).

b. *measure of kurtosis* in a family of symmetric distributions $\mathcal{F}_s \subset \mathcal{F}$ if $\theta(F_{a,b}) = \theta(F)$ for all $a \neq 0, b \in \mathbb{R}, F \in \mathcal{F}_s$ and $\theta(F) \leq \theta(G)$ whenever $F, G \in \mathcal{F}_s$, $G^{-1} \circ F$ is concave on $\{x | x \leq m(F)\}$ and convex on $\{x | x > m(F)\}$. In this case, we say F *does not have more kurtosis than* G. ■

For the definitions of measures of location and scale, refer to Oja (1981). Note from the paper that distributional kurtosis connects heavy-tailedness and bimodality. More kurtosis implies being peaked at its central region and heavy-tailed, while less kurtosis implies heavy shoulders near the center yielding bimodality.

3. Gaussian Centered L-moments

As mentioned in Section 1, a search for non-Gaussianity is an important subject of exploratory data analysis. This motivates introducing the following definition.

Definition 3.1. A sequence of functionals $\{\theta_r | r = 1, 2, \cdots\}$ is *centered at the family of distributions* $\mathcal{F}$ when it satisfies $\theta_r(F) = 0$ for all $r = 3, 4, \cdots$ and $F \in \mathcal{F}$. ■

Important functionals centered at the Gaussian distributions are the cumulants; see Feller (1968) and Marcinkiewicz (1939). For developing moments centered at the Gaussian distributions, we introduce the following definition.

Definition 3.2. We call functionals $\{\theta_r : \mathcal{F} \to \mathbb{R} | r = 1, 2, \cdots\}$ *Gaussian Centered L-moments* (*GCL-moments*) in $\mathcal{F}$ if they are L-functionals and centered at the Gaussian distributions. ■

The letter 'L' generally referred to the linear combination of expected order statistics, but here it refers to any L-functionals in the form of Equation (2).

3.1. Hermite L-moments

The fact that the L-moments are centered at the uniform distributions results from the orthogonality of the shifted Legendre polynomials. This motivates us to adopt the Hermite polynomials to locate the center of L-functionals at the Gaussian distributions. This results in the Hermite L-moments (*HL-moments*) defined as

$$\eta_r = \int_{-\infty}^{\infty} x f(x) H_{r-1}\left(\Phi^{-1}(F(x))\right)\,\mathrm{d}x = \int_0^1 F^{-1}(u) H_{r-1}\left(\Phi^{-1}(u)\right)\,\mathrm{d}u. \quad (4)$$

Note that η_r are L-functionals and $\eta_r(\Phi(\cdot|\mu,\sigma^2)) = 0$ for all $\mu \in \mathbb{R}, \sigma > 0$ and $r = 3, 4, \cdots$ by the orthogonality of the Hermite polynomials

Recall from Definition 2.1.a and b that a measure of skewness or kurtosis should be invariant under linear transformation of a random variable. This motivates us to define the *Hermite L-moment ratios* defined as $\eta_r^* = \eta_r/\eta_2$ for $r = 3, 4, \cdots$. The *Hermite L-skewness* (*HL-skewness*) and *Hermite L-kurtosis* (*HL-kurtosis*) are defined as η_3^* and η_4^*, respectively. A central issue is whether the HL-skewness and kurtosis actually measure the skewness and kurtosis of a distribution in the sense of Definition 2.1.

Theorem 3.1. The HL-moments-based measures η_1, η_2, η_3^* and η_4^* satisfy Oja's criteria for measures of location, scale, skewness and kurtosis, respectively.

Proof. See Proof of Theorem 2.1 in Page 70 of An (2017). ■

This theorem shows that estimators of the HL-moments can actually be used as univariate summary statistics.

Note that the HL-moments are closely related to the inverse Edgeworth expansion (Hall, 1983)

$$F^{-1}(u) \approx \mu + \sigma\Phi^{-1}(u) + \sigma\sum_{r=3}^{\infty} \frac{EH_r(Y)}{r!} H_{r-1}\big(\Phi^{-1}(u)\big) \quad (5)$$

where μ and σ are the mean and standard deviation of the random variable $X \sim F$ and $Y = (X-\mu)/\sigma$ is a standardized random variable. Note that the terms $H_{r-1}\big(\Phi^{-1}(u)\big)$ in this expansion also appear in the definition of the HL-moments in Equation (4). Based on Equation (5), Brown and Hettmansperger (1996) proposed the following estimators as indicators of departure from Gaussianity

$$\hat{\eta}_{n,r}^{(\mathrm{BH})} = \int_0^1 F_n^{-1}(u) H_{r-1}\left(\Phi^{-1}(u)\right)\,\mathrm{d}u, \quad (6)$$

for $r = 1, 2, \cdots$ where $F_n(x) = \frac{1}{n}\sum_{i=1}^n I\left(X_i \le x\right)$. We call this estimator the *Brown-Hettmansperger* (*BH*) estimator. Replacing F_n by F in Equation (6) yields the population HL-moments η_r, which was not elaborated in Brown and Hettmansperger (1996).

3.2. Rescaled L-moments

Additional insights come from another view of why the L-moments are centered at the uniform distributions. Note that the fourth population L-moment is

$$\lambda_4 = \frac{1}{4}\left\{E\left(X_{4:4} - X_{3:4}\right) - 2E\left(X_{3:4} - X_{2:4}\right) + E\left(X_{2:4} - X_{1:4}\right)\right\}.$$

This expression indicates that if F has equally spaced expected order statistics, then its fourth L-moment is zero. However, for the standard Gaussian distribution, the space between the inner pair of expected order statistics (≈ 0.5940) is smaller than the spaces between the two outer pairs (≈ 0.7324).

This motivates us to rescale the spacing between expected order statistics by the corresponding spacing of the standard Gaussian distribution. The following theorem shows that another expression of the L-moments can be derived in terms of spacing between expected order statistics.

Theorem 3.2. The r-th L-moment λ_r can be expressed as

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-2}(-1)^k \binom{r-2}{k} E\left(X_{(r-k):r} - X_{(r-k-1):r}\right),$$

for $r = 2, 3, \cdots$.

Proof. See Section 1 of the Supplementary Material. ■

We first show that when the standard Gaussian distribution is used for rescaling, the resulting measures of scale, skewness and kurtosis are linear functions of the corresponding measures based on the L-moments. Let $\delta_{i,j:k}(F) = E\left(X_{j:k} - X_{i:k}\right)$ for $1 \le i < j \le k$ be the space between the i-th and j-th expected order statistics of F. Then we define the *Rescaled L-moments* (*RL-moments*) as

$$\rho_r = \frac{1}{r}\sum_{k=0}^{r-2} \frac{(-1)^k}{\delta_{(r-k-1),(r-k):r}(\Phi)} \binom{r-2}{k} E\left(X_{(r-k):r} - X_{(r-k-1):r}\right), \tag{7}$$

for $r = 2, 3, \cdots$ and $\rho_1 = \lambda_1$. The corresponding *Rescaled L-moment ratios* are defined as $\rho_r^* = \rho_r/\rho_2$ for $r = 3, 4, \cdots$.

Theorem 3.3. We have

$$\begin{aligned} &\rho_1 = \lambda_1, && \rho_2 = \tfrac{1}{\delta_{1,2:2}(\Phi)}\lambda_2, \\ &\rho_3^* = \tfrac{\delta_{1,2:2}(\Phi)}{\delta_{1,2:3}(\Phi)}\lambda_3^*, && \rho_4^* = \tfrac{\delta_{1,2:2}(\Phi)}{5}\left\{\tfrac{3}{\delta_{2,3:4}(\Phi)} + \tfrac{2}{\delta_{3,4:4}(\Phi)}\right\}\lambda_4^* \\ &&& \quad -\tfrac{3\delta_{1,2:2}(\Phi)}{5}\left\{\tfrac{1}{\delta_{2,3:4}(\Phi)} - \tfrac{1}{\delta_{3,4:4}(\Phi)}\right\}. \end{aligned} \tag{8}$$

These four measures satisfy Oja's criteria for a measure of location, scale, skewness and kurtosis, respectively.

Proof. See Section 2 of the Supplementary Material. ■

Theorem 3.3 shows that subtracting the L-kurtosis value of the standard Gaussian distribution from the L-kurtosis itself is actually equivalent to rescaling based on the spacing between expected order statistics of the standard Gaussian distribution. It can easily be shown that the RL-moments are L-functionals such that

$$\rho_r = \int_{-\infty}^{\infty} x f(x) R_{r-1}(F(x))\,\mathrm{d}x = \int_0^1 F^{-1}(u) R_{r-1}(u)\,\mathrm{d}u,$$

where $R_r : (0,1) \to \mathbb{R}$ is an r-th degree polynomial that satisfies

$$R_r(u) = \sum_{k=1}^{r} \alpha_{r,k} P_k^*(u), \tag{9}$$

for $0 < u < 1$ with constants $\alpha_{r,j} \in \mathbb{R}$ for $j = 1, 2, \cdots, r$. For example, we have $\alpha_{2,1} = 0$ and $\alpha_{2,2} = 1/\delta_{1,2:2}(\Phi)$. Equation (7) also implies that the RL-moments are centered at the Gaussian distributions. These enable us to use the RL-moments, or equivalently the L-moments, as one of the GCL-moments in Section 6.

The RL-moment ratios of higher orders than 4 are not necessarily linear functions of the L-moment ratios of the same orders. For example, we have $\rho_5^* = \alpha_{5,5}^* \lambda_5^* + \alpha_{5,3}^* \lambda_3^*$ where

$$\alpha_{5,5}^* = \frac{\delta_{1,2:2}(\Phi)}{7}\left\{\frac{6}{\delta_{3,4:5}(\Phi)} + \frac{1}{\delta_{4,5:5}(\Phi)}\right\},$$
$$\alpha_{5,3}^* = -\frac{6\delta_{1,2:2}(\Phi)}{7}\left\{\frac{1}{\delta_{3,4:5}(\Phi)} - \frac{1}{\delta_{4,5:5}(\Phi)}\right\},$$

whose proof follows the same steps of derivation as the proof of Theorem (3.3). This implies that even though the RL-moments have essentially the same first four moments as the L-moments, their high order moments can exhibit different behavior.

4. Estimation of the Gaussian Centered L-moments

This section contains details of derivation of the estimators used in this paper. One of the main strengths of the GCL-moments is their interpretability in terms of departure from the Gaussianity. For such a strength to be effective in real data analysis, estimators of the GCL-moments should converge to their theoretical parallels yielding the desired interpretability.

4.1. Sample Hermite L-moments

There are multiple ways to estimate L-functionals in Equation (2) by L-statistics in Equation (1). For the HL-moments, our estimator starts from the

following approximation

$$\eta_r = E_F\left(H_{r-1}\left(\Phi^{-1}(F(X))\right)X\right) \approx \frac{1}{n}\sum_{i=1}^{n} H_{r-1}\left(\Phi^{-1}\left(F\left(X_i\right)\right)\right)X_i$$
$$= \frac{1}{n}\sum_{i=1}^{n} H_{r-1}\left(\Phi^{-1}\left(F\left(X_{i:n}\right)\right)\right)X_{i:n}, \quad (10)$$

where the approximation can be replaced by the almost sure convergence as $n \to \infty$ when suitable assumptions are made on the distribution F. For the last expression in Equation (10) to actually play the role of an estimator, the terms including F should be estimated. A typical L-statistic can be derived from the following approximation

$$\frac{1}{n}\sum_{i=1}^{n} H_{r-1}\left(\Phi^{-1}\left(\underline{F\left(X_{i:n}\right)}\right)\right)X_{i:n} \approx \frac{1}{n}\sum_{i=1}^{n} H_{r-1}\left(\Phi^{-1}\left(\underline{E\left(F\left(X_{i:n}\right)\right)}\right)\right)X_{i:n}$$
$$= \frac{1}{n}\sum_{i=1}^{n} H_{r-1}\left(\Phi^{-1}\left(\frac{i}{n+1}\right)\right)X_{i:n}, \quad (11)$$

where the underlined expressions present approximated (left) and approximating (right) terms and $U_{i:n}$ is the i-th uniform order statistic. It was seen in Example 1b of Shorack (1972) that the approximation here can be substituted for by the almost sure convergence when X_i's follow the standard Gaussian distribution.

Another estimator can be obtained from an approximation in Equation (11) as

$$\frac{1}{n}\sum_{i=1}^{n} \underline{H_{r-1}\left(\Phi^{-1}\left(F\left(X_{i:n}\right)\right)\right)}X_{i:n} \approx \frac{1}{n}\sum_{i=1}^{n} \underline{E\left(H_{r-1}\left(\Phi^{-1}\left(F\left(X_{i:n}\right)\right)\right)\right)}X_{i:n}$$
$$= \frac{1}{n}\sum_{i=1}^{n} E\left(H_{r-1}\left(Z_{i:n}\right)\right)X_{i:n}, \quad (12)$$

where $Z_{i:n}$ is the i-th standard Gaussian order statistic. The key idea is that careful choice of location of the expectation can increase accuracy of approximation of an L-functional by an L-statistic. Since the quantile function Φ^{-1} is a highly nonlinear function, taking expectation outside Φ^{-1} can yield better approximation to the HL-moments. We call the L-statistic of the last expression the r-th *sample HL-moment*, and denote it by $\hat{\eta}_r$. The sample HL-moments can be understood as inner products between the order statistics $X_{i:n}$ and expected polynomials of the standard Gaussian order statistics.

In the same way as the population HL-moments in Section 3.1, we define the *sample HL-moment ratios* as $\hat{\eta}^*_{n,r} = \hat{\eta}_{n,r}/\hat{\eta}_{n,2}$ for $r = 3, 4, \cdots$. The *sample HL-skewness* and *sample HL-kurtosis* are defined as $\hat{\eta}^*_{n,3}$ and $\hat{\eta}^*_{n,4}$, respectively. Theorem 4.1 shows that the sample HL-moments and their ratios are consistent estimators of their theoretical parallels.

Theorem 4.1. Suppose that $E|X_1|^{1+\epsilon} < \infty$ for some $\epsilon > 0$. Then we have $\hat{\eta}_{n,r} \to^{\text{a.s.}} \eta_r$ as $n \to \infty$ for $r = 1, 2, \cdots$. As a result, we have $\hat{\eta}^*_{n,r} \to^{\text{a.s.}} \eta^*_r$ as $n \to \infty$ for $r = 3, 4, \cdots$.

Proof. See Subsection 5 of the Supplementary Material. ■

4.2. Advanced sample Hermite L-moments

The sample HL-moments can be improved by the linear regression theory in the spirit of LaBrecque (1977). We use the sample HL-skewness as an example. Note that the third sample HL-moment is proportional to the regression coefficient estimator of the following linear regression model,

$$X_{i:n} = \beta_3 E\left(H_2\left(Z_{i:n}\right)\right) + Z_{i:n}^{(\beta_1,\beta_2)}, \tag{13}$$

where $X_{i:n}$ is the response, β_3 is the slope parameter, $E\left(H_2\left(Z_{i:n}\right)\right)$ is the predictor, and $Z_{i:n}^{(\beta_1,\beta_2)}$ is the random error, which is the i-th order statistic of the random sample of size n from $\mathcal{N}\left(\beta_1, \beta_2^2\right)$. The ordinary least squares (OLS) estimator for η_3 in this regression model is

$$\hat{\beta}_{n,3}^{(\text{OLS})} = \frac{1}{\left(\boldsymbol{\xi}_n^{(2)}\right)^T \boldsymbol{\xi}_n^{(2)}} \left(\boldsymbol{\xi}_n^{(2)}\right)^T \boldsymbol{X}_n = \frac{n}{\left(\boldsymbol{\xi}_n^{(2)}\right)^T \boldsymbol{\xi}_n^{(2)}} \hat{\eta}_{n,3}, \tag{14}$$

where $\boldsymbol{\xi}_n^{(r)} = \left(E\left(H_r\left(Z_{1:n}\right)\right), \cdots, E\left(H_r\left(Z_{n:n}\right)\right)\right)^T$ and $\boldsymbol{X}_n = \left(X_{1:n}, \cdots, X_{n:n}\right)^T$ for $r = 1, 2, \cdots$. While this may not be an optimal estimator for β_3 since $Z_{i:n}^{(\beta_1,\beta_2)}$ for $i = 1, 2, \cdots, n$ are not independent nor distributed as the same Gaussian distribution, it is still useful because of the consistency of $\hat{\eta}_{n,r}$.

To improve the estimator $\hat{\beta}_{n,3}^{(\text{OLS})}$, we modify the regression model in Equation (13) so that the random errors have zero means as follows,

$$X_{i:n} = \beta_1 + \beta_2 E\left(Z_{i:n}\right) + \beta_3 E\left(H_{r-1}\left(Z_{i:n}\right)\right) + \left(Z_{i:n}^{(0,\beta_2)} - \beta_2 E\left(Z_{i:n}\right)\right), \tag{15}$$

where β_1 is the intercept, β_2 is the coefficient of the predictor $E\left(Z_{i:n}\right)$ and $\left(Z_{i:n}^{(0,\beta_2)} - \beta_2 E\left(Z_{i:n}\right)\right)$ is the centered random error. Note that these random errors have zero means but have a non-identical variance-covariance matrix. In vector notation, this equation is written as follows,

$$\boldsymbol{X}_n = \beta_1 \mathbf{1} + \beta_2 \boldsymbol{\xi}_n^{(1)} + \beta_3 \boldsymbol{\xi}_n^{(2)} + \left(\boldsymbol{Z}_n^{(0,\beta_2)} - \beta_2 \boldsymbol{\xi}_n^{(1)}\right), \tag{16}$$

where $\mathbf{1} = (1, \cdots, 1)^T$ and $\boldsymbol{Z}_n^{(0,\beta_2)} = \left(Z_{1:n}^{(0,\beta_2)}, \cdots, Z_{n:n}^{(0,\beta_2)}\right)$. A similar equation with this model was shown in Subsection 2.2 of LaBrecque (1977).

The paper LaBrecque (1977) considered the regression model of the Shapiro-Wilk statistic (Shapiro and Wilk, 1965) as the starting point, and developed goodness-of-fit statistics of Gaussianity in the directions of distributional skewness and kurtosis. They included powers of expected standard Gaussian order

statistics as predictors in this model, and applied the Gram-Schmidt orthogonalization and generalized least squares (GLS) method to obtain

$$\hat{\eta}_{n,r}^{(\mathrm{LaB})} = \frac{1}{\left(\boldsymbol{\zeta}_n^{(r-1)*}\right)^T V_n^{-1} \boldsymbol{\zeta}_n^{(r-1)*}} \left(\boldsymbol{\zeta}_n^{(r-1)*}\right)^T V_n^{-1} \boldsymbol{X}_n, \tag{17}$$

where $\boldsymbol{\zeta}_n^{(r)} = \left(H_r\left(E\left(Z_{1:n}\right)\right), \cdots, H_r\left(E\left(Z_{n:n}\right)\right)\right)^T$ for $r = 3, 4$. This inspires us to improve the OLS estimator in Equation (16) in the same way. The paper LaBrecque (1977) took polynomials outside of the expected order statistics since they based their starting point in Shapiro-Wilk's model, and we start from our predictors in Equation (16).

To obtain an optimal estimator by utilizing the covariances among the random errors in Equation (16), we apply the Gram-Schmidt orthogonalization and GLS method to obtain the following residual vector,

$$\boldsymbol{\xi}_n^{(2)*} = \boldsymbol{\xi}_n^{(2)} - \left(\mathbf{1}^T V_n^{-1} \mathbf{1}\right)^{-1} \left(\mathbf{1}^T V_n^{-1} \boldsymbol{\xi}_n^{(2)}\right) \mathbf{1},$$

where V_n is the variance-covariance matrix of $Z_{i:n}$ for $i = 1, 2, \cdots, n$, and the fact that $V_n^{-1}\mathbf{1} = \mathbf{1}$ shown in Lloyd (1952) is used. The third advanced sample HL-moment is obtained by regressing $\boldsymbol{X}_n$ on $\boldsymbol{\xi}_n^{(2)*}$. The fourth sample HL-moment can be improved in the same way, yielding the third and fourth *advanced sample HL-moments* as

$$\begin{aligned}
\hat{\eta}_{n,3}^{(\mathrm{A})} &= \frac{2}{\left(\boldsymbol{\xi}_n^{(2)*}\right)^T V_n^{-1} \boldsymbol{\xi}_n^{(2)*}} \left(\boldsymbol{\xi}_n^{(2)*}\right)^T V_n^{-1} \boldsymbol{X}_n, \\
\hat{\eta}_{n,4}^{(\mathrm{A})} &= \frac{6}{\left(\boldsymbol{\xi}_n^{(3)*}\right)^T V_n^{-1} \boldsymbol{\xi}_n^{(3)*}} \left(\boldsymbol{\xi}_n^{(3)*}\right)^T V_n^{-1} \boldsymbol{X}_n,
\end{aligned} \tag{18}$$

respectively. The numerators 2 and 6 are needed to make these estimators have approximately the same scale with the corresponding sample HL-moments in data analysis.

Note that the meanings of the words 'optimal' and 'advanced' are limited to the Gaussianity assumption. When $F \sim \mathcal{N}(0, 1)$, the standard errors of the third and fourth sample HL-moments, LaBrecque estimators and advanced sample HL-moments are calculated as Table 2. The sample sizes $n = 817$ and 20 in the first column are those of our TCGA data analysis in Section 6 and computational study in Section 7. The second column shows the order of the moments, and the remaining columns show the standard errors. The sample HL-moments used in this table are based on Equation (14). Overall, the advanced sample HL-moments have smaller standard errors than the other two estimators.

Performances of these estimators in real and non-Gaussian data should be carefully analyzed. To this end, we practically show in Section 6.3 that the advanced sample HL-moments outperform other estimators including the sample HL-moments. This presents a different view of the suggestion in Subsection

Size (n)	Order	Sample	LaB	Advanced
817	3rd	1.4362×10^{-2}	1.4417×10^{-2}	1.4357×10^{-2}
817	4th	7.3244×10^{-3}	7.4909×10^{-3}	7.3074×10^{-3}
20	3rd	0.1055	0.1110	0.1052
20	4th	0.0678	0.0810	0.0670

Table 2: The standard errors of the third and fourth HL-moments related estimators in Section 4 under the Gaussianity assumption. Overall, the advanced sample HL-moments have the smallest standard errors and the LaB estimators have the largest errors.

2.2 of Brown and Hettmansperger (1996) that ignoring the covariance matrix V_n has little impacts on estimation of scale. In the directions of skewness and kurtosis, existence of V_n has an impact on performance.

A scale estimator can be obtained by focusing on the estimator of β_2 in Equation (16) by the GLS method. Since the predictors $\mathbf{1}$, $\boldsymbol{\xi}_n^{(1)}$, and $\boldsymbol{\xi}_n^{(2)*}$ satisfy orthogonality properties that $\mathbf{1}V_n^{-1}\boldsymbol{\xi}_n^{(1)} = 0$ and $\boldsymbol{\xi}_n^{(1)}V_n^{-1}\boldsymbol{\xi}_n^{(2)*} = 0$, the scale estimator is obtained as the L-statistic,

$$\hat{\eta}_{n,2}^{(\mathrm{A})} = \frac{1}{\left(\boldsymbol{\xi}_n^{(1)}\right)^T V_n^{-1}\boldsymbol{\xi}_n^{(1)}} \left(\boldsymbol{\xi}_n^{(1)}\right)^T V_n^{-1}\boldsymbol{X}_n.$$

The paper (LaBrecque, 1977) did not specify random errors in their model and used the sample standard deviation of $\boldsymbol{X}_n$ as a scale estimator. Here we notice that the parameter that determines distributional scale of $\boldsymbol{X}_n$ is shared with the coefficient of the predictor $\boldsymbol{\xi}_n^{(1)}$. Hence, the second advanced sample HL-moment is defined as $\hat{\eta}_{n,2}^{(\mathrm{A})}$, which more naturally aligns with the third and fourth advanced sample HL-moments. We call $\hat{\eta}_3^{(\mathrm{A})*} = \hat{\eta}_{n,3}^{(\mathrm{A})}/\hat{\eta}_{n,2}^{(\mathrm{A})}$ and $\hat{\eta}_4^{(\mathrm{A})*} = \hat{\eta}_{n,4}^{(\mathrm{A})}/\hat{\eta}_{n,3}^{(\mathrm{A})}$ the *advanced sample HL-skewness* and *-kurtosis*, respectively.

Note that LaBrecque (1977) did not show the almost sure convergence of their estimator, while Brown and Hettmansperger (1996) showed the asymptotic Gaussianity of their estimators only under the Gaussianity assumption. The paper Leslie (1984) proved that we have

$$\left\| V_n^{-1}\boldsymbol{\xi}_n^{(1)} - 2\boldsymbol{\xi}_n^{(1)} \right\| \to 0,$$

as $n \to \infty$ where $\|\cdot\|$ is the Euclidean norm. If it can be shown that substituting $\boldsymbol{\xi}_n^{(r)}$ for $\boldsymbol{\xi}_n^{(1)}$ still makes the limit valid, the result can be combined with Theorem 4.1 to yield consistency of the advanced sample HL-moments.

4.3. Estimation of the Rescaled L-moments

Since the r-th RL-moment is a linear combination of the L-moments as shown in Subsection 3.2, the r-th *sample RL-moment* is naturally derived as a linear combination of the sample L-moments. From Equation (9), we define the

r-th sample RL-moment as $\hat{\rho}_{n,1} = \hat{\lambda}_{n,1}$ and

$$\hat{\rho}_{n,r} = \sum_{k=1}^{r} \alpha_{k,r} \hat{\lambda}_{n,r},$$

for $r = 2, 3, \cdots$. As a result, the first two sample RL-moments, *sample RL-skewness*, *sample RL-urtosis* are derived as follows,

$$\begin{array}{ll} \hat{\rho}_{n,1}, = \hat{\lambda}_{n,1} & \hat{\rho}_{n,2} = \frac{1}{\delta_{1,2:2}(\Phi)} \hat{\lambda}_{n,2}, \\ \hat{\rho}^*_{n,3} = \frac{\delta_{1,2:2}(\Phi)}{\delta_{1,2:3}(\Phi)} \hat{\lambda}^*_{n,3}, & \hat{\rho}^*_{n,4} = \frac{\delta_{1,2:2}(\Phi)}{5} \left\{ \frac{3}{\delta_{2,3:4}(\Phi)} + \frac{2}{\delta_{3,4:4}(\Phi)} \right\} \hat{\lambda}^*_{n,4} \\ & \quad - \frac{3\delta_{1,2:2}(\Phi)}{5} \left\{ \frac{1}{\delta_{2,3:4}(\Phi)} - \frac{1}{\delta_{3,4:4}(\Phi)} \right\}. \end{array}$$

The fact that these estimators are consistent estimators of the RL-moments easily follow from consistency of the sample L-moments.

5. Robustness

We carefully study the relative robustness of the methods explained in Section 3. We use the *influence function* (Huber and Ronchetti, 2009) as a primary tool for robustness analysis , which is defined as

$$\mathrm{IF}(x; F, \theta) = \lim_{\epsilon \downarrow 0} \frac{\theta\left(F_{\epsilon,x}\right) - \theta(F)}{\epsilon}, \tag{19}$$

where $F_{\epsilon,x} = (1-\epsilon)F + \epsilon\delta_x$ and δ_x is a degenerate distribution putting mass 1 at the point x. The influence function measures the effect of contamination of a distribution by a point x on the functional θ. Hence, if a functional is sensitive to an outlier, its influence function should have large values for large absolute values of x. In this paper, we compare the robustness of various measures of skewness and kurtosis based on their influence functions evaluated at a family of distributions.

The papers Groeneveld (1991) and Ruppert (1987) compared various measures of skewness and kurtosis, respectively, using the influence function. As a criterion of comparison, both the papers compared the degrees of polynomials that are *asymptotic tight bounds* of the influence functions. Suppose that $J_1, J_2 : \mathbb{R} \to \mathbb{R}_+$ are two functions where $\mathbb{R}_+ = \{x \in \mathbb{R} | x \geq 0\}$. We write $J_1(x) = \Theta\left(J_2(x)\right)$ to mean that there exist $a_1, a_2 > 0$ and $x' > 0$ such that

$$a_1 J_2(x) \leq J_1(x) \leq a_2 J_2(x),$$

for all $|x| \geq x'$. This roughly means that asymptotic behavior of both the functions J_1 and J_2 is the same. In those two papers, if two functionals θ_1 and θ_2 satisfy $|\mathrm{IF}\left(x; F, \theta_1\right)| = \Theta(|x|)$ and $|\mathrm{IF}\left(x; F, \theta_2\right)| = \Theta\left(x^2\right)$, then θ_1 was considered to be more robust than θ_2 for the distribution F.

An interesting family of distributions for evaluation of influence functions is *Tukey's g and h distributions* (Martinez and Iglewicz, 1984) which contain all the transformed random variables of the form

$$\left(\frac{e^{gZ}-1}{g}\right)\exp\left[\frac{hZ^2}{2}\right],$$

where $g \in \mathbb{R}, h \geq 0$ and Z is the standard Gaussian random variable. We denote the distribution function of Tukey's g and h distribution by $T_{g,h}$. As shown in Brys et al. (2004, 2006), Φ does not have more kurtosis than $T_{0,h}$ (Definition 2.1.b) if $g = 0$ and $h > 0$. This implies that Tukey's h distributions have heavier tails than the Gaussian distributions in Oja's sense. Note that heavier tails of a distribution indicate a higher chance of existence of extreme outliers. This motivates us to adopt Tukey's h distributions as grounds for comparison between different measures of skewness and kurtosis.

Before we derive the influence functions of the GCL-moments, we introduce the previous results for the conventional skewness and kurtosis defined as

$$\gamma_1 = \frac{E\left(X-EX\right)^3}{\left\{E\left(X-EX\right)^2\right\}^{3/2}}, \quad \gamma_2 = \frac{E\left(X-EX\right)^4}{\left\{E\left(X-EX\right)^2\right\}^2} - 3,$$

respectively. Note that Ruppert (1987) used the notion of *symmetric influence function* defined as

$$\mathrm{SIF}(x;F,\theta) = \lim_{\epsilon \downarrow 0} \frac{\theta\left(\left(F_{\epsilon,x} + F_{\epsilon,-x}\right)/2\right) - \theta(F)}{\epsilon}.$$

The symmetric influence function measures the sensitivity of a functional to symmetric contamination by points $-x$ and x so it is suitable for comparison of kurtosis measures.

Theorem 5.1 (Groeneveld (1991),Ruppert (1987))**.** Suppose that F is a symmetric distribution such that $\mu(F) = 0$ and $\sigma^2(F) = 1$. Then we have

$$\mathrm{IF}\left(x;F,\gamma_1\right) = x^3 - 3x = H_3(x), \; \mathrm{SIF}\left(x;F,\gamma_2\right) = x^4 - 6x^2 + 3 = H_4(x). \quad \blacksquare$$

Now we show the relationships between the influence functions and symmetric influence functions of the Gaussian Centered L-moments.

Theorem 5.2. If F is a symmetric distribution, we have

$$\mathrm{SIF}\left(x;F,\lambda_4^*\right) = \mathrm{IF}\left(x;F,\lambda_4^*\right), \quad \mathrm{SIF}\left(x;F,\rho_4^*\right) = \mathrm{IF}\left(x;F,\rho_4^*\right),$$
$$\mathrm{SIF}\left(x;F,\eta_4^*\right) = \mathrm{IF}\left(x;F,\eta_4^*\right).$$

Proof. See Section 3 of the Supplementary Material. ■

In addition to the conventional and GCL-moments, we consider some quantile based measures, which are known to be robust, as baseline measures. Bowley's skewness measure (Bowley, 1920) is a typically used quantile-based measure of skewness defined as

$$\gamma_p = \frac{F^{-1}(1-p) - F^{-1}(1/2) - \left\{F^{-1}(1/2) - F^{-1}(p)\right\}}{F^{-1}(1-p) - F^{-1}(p)}$$

where p is usually set to 0.25, which results in γ_p being based on quartiles. On the other hand, Ruppert's interfractile range ratio (Ruppert, 1987) is frequently used as a measure of kurtosis and defined as

$$\gamma_{p_1,p_2} = \frac{F^{-1}(1-p_1) - F^{-1}(p_1)}{F^{-1}(1-p_2) - F^{-1}(p_2)}$$

where the parameters p_1 and p_2 were set to 0.1 and 0.3, respectively, in that paper. Note that Bowley's skewness measure is zero at the Gaussian distributions, but Ruppert's kurtosis measure is not zero there. Both Bowley's and Ruppert's measures were shown to satisfy Oja's criteria for measures of skewness and kurtosis, respectively, in those papers. In this paper, we use sample quantile based estimators defined in those papers to estimate the measures.

The main result is given as the following theorem.

Theorem 5.3. We have

$$\begin{aligned}
|\mathrm{IF}(x; T_{0,h}, \gamma_{0.25})| = |\mathrm{IF}(x; T_{0,h}, \gamma_{0.1,0.3})| &= \Theta(1), \\
|\mathrm{IF}(x; T_{0,h}, \lambda_r^*)| = |\mathrm{IF}(x; T_{0,h}, \rho_r^*)| &= \Theta(|x|), \\
|\mathrm{IF}(x; T_{0,h}, \eta_r^*)| &= \Theta\Big(|x| \left\{\log(|x|+1)\right\}^{(r-1)/2}\Big), \\
|\mathrm{IF}(x; T_{0,h}, \gamma_1)| &= \Theta\left(|x|^3\right), \\
|\mathrm{IF}(x; T_{0,h}, \gamma_2)| &= \Theta\left(|x|^4\right),
\end{aligned}$$

for all $h > 0$ and $r = 3, 4, \cdots$.

Proof. See Section 4 of the Supplementary Material. ■

Theorem 5.3 implies that the GCL-moments are more robust than the conventional skewness and kurtosis on Tukey's h distributions. Note that the influence function of η_r^*, $\mathrm{IF}(x; T_{0,h}, \eta_r^*)$, does not depend on the parameter h. This indicates that even slightly heavier tails than the standard Gaussian distribution result in better robustness of the HL-moments than the conventional moments. Note that the L-moments are more robust than the HL-moments even though the uniform center of the L-moments has lighter tails than the Gaussian center of the HL-moments. This shows that heavier tails of the distributional center of L-functionals do not always imply their more robustness. The bounded influence functions of Bowley's and Ruppert's measures, $\gamma_{0.25}$ and $\gamma_{0.1,0.3}$, were obtained from the results of Ruppert (1987) and Groeneveld (1991). This justifies the use of quantile based measures as baseline robust estimators in our TCGA data analysis given in Section 6.

6. TCGA data analysis

The goal of our TCGA data analysis is to discover biologically meaningful genes out of 16,615 genes based on their expression profiles of 817 breast cancer

patients. As mentioned in Section 1, there are many biological features of importance in cancer research and genes with strong such features seem to have non-standard distributions of expression profiles. Hence, we focus on departure from Gaussianity in the directions of skewness and kurtosis as a measure of screening genes in the TCGA data.

Note that skewness and bimodality of a distribution are not totally independent concepts. It was shown in Pearson (1916) and Hosking (1990) that the conventional population moments and L-moments satisfy the following relationships

$$\gamma_1^2 - 2 \leq \gamma_2, \quad \frac{1}{4}\left(5\left(\lambda_3^*\right)^2 - 1\right) \leq \lambda_4^* < 1.$$

This implies that a skewness or kurtosis measure alone is not enough to describe skewness and bimodality of a distribution. Bimodal distributions should have low kurtosis estimates when they are symmetric, but this does not hold for asymmetric bimodal distributions. Hence, we check both of the ranked lists of genes generated by skewness and kurtosis estimators and comprehensively diagnose performances of different estimators. Joint use of skewness and kurtosis estimators such as the Jarque-Bera statistic (Jarque and Bera, 1980) cannot imply whether the departure mainly comes from skewness, kurtosis or a combination of both. Hence, we do not proceed that way.

6.1. Gene Set Enrichment Analysis

We conjecture that genes with non-Gaussian distributional shape can be related to biological features, and confirm this based on *Gene Set Enrichment Analysis* (*GSEA*, Subramanian et al. (2005)). A basic goal of GSEA is to assess goodness of a ranked list of genes in terms of how well the list places biologically meaningful genes near its top or bottom, i.e. how well the list screens meaningful genes. Since a measure of skewness or kurtosis provides an order of the genes, its screening ability can be assessed by applying GSEA to the ranked lists generated by the measure. In our GSEA, we study performances of various measures over gene sets published by the Broad Institute. A collection of 16,107 such gene sets, with a minimum of 15 and maximum of 10,000 genes, is available in the public database MSigDB v.7.0. Each of those gene sets has a specific scientific meaning, so we regard them as biologically meaningful gene sets.

Main output of GSEA for each ranked list of genes is the estimated *False Discovery Rate* (*FDR*), which is the proportion of randomly permuted lists of genes and gene sets with larger enrichment scores than the enrichment score of the given ranked list of genes and gene set. Refer to Subramanian et al. (2005) for detailed discussion of the enrichment score and FDR. One difference of our analysis from the conventional GSEA is that different estimators of skewness and kurtosis have different scales. Based on suggestions of Subramanian et al. (2005), we use the classic scoring scheme given in the paper. Also based on the recommendations of the GSEA User Guide (`http://software.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html`), we perform $1,000$ times of gene permutations and set the FDR level at 0.05, which were recommended for the classic scoring scheme.

6.2. Comparison among skewness and kurtosis estimators

Table 3 shows the results of comparison between different skewness estimators when the FDR level is fixed at 0.05. The title row shows the names of the skewness estimators and the numbers of respective screened gene sets. For example, the ranked list of genes generated by the L-skewness screened 3,295 biologically meaningful gene sets, i.e. the FDR's of those gene sets were less than 0.05. The title column shows the skewness estimators in the same order with the last estimator removed. The other columns show, from left to right, the best to worst performing skewness estimators by the numbers of meaningful gene sets screened by them. We do a pairwise comparison of the performances of estimators in terms of the number of gene sets that are flagged as screened and assess statistical significance using McNemar's test (Section 14.5 of Gibbons and Chakraborti (2010)). McNemar's test assumes that each object is tested by two procedures, and compares success probabilities of the two procedures while considering correlations between each object's paired outcomes. In our GSEA comparison, each gene set is screened by two ranked lists of genes, which justifies use of the test.

The values given in the middle of the table are the p-values. For example, the p-value 0.0097 in the first row and the second column is the significance of the difference in the numbers of screened gene sets by the L-skewness (3,295) and HL-skewness (3,222). Since each pair of skewness estimators needs only one comparison, we color cells corresponding to repetitive comparisons gray. The boldfaced numbers indicate that the corresponding p-values are significant after Bonferroni's correction for the six comparisons in the table.

It can be seen in Table 3 that, generally, the GCL-moments based estimators better capture meaningful departures from Gaussianity in the direction of skewness. The L-skewness performs significantly better than all the other estimators except the HL-skewness as shown in the second row of Table 3. The comparison between the L- and HL-skewness is hard to conclude, since the p-value of the comparison is not significant after considering multiplicity. The HL-skewness performs better than the conventional and Bowley's skewness, and the degrees of the differences are statistically significant. This coincides with our observations made in Section 1 that the conventional skewness is driven by outliers too much to screen subtype-driven genes. Bowley's skewness is not successful in screening meaningful genes. Its inferiority to the other estimators is statistically significant as seen in the last column. The TCGA data have 5 subtypes, and consideration of only the first and third sample quartiles by Bowley's skewness seems to result in a strong loss of subtype information.

Table 4 shows the comparison results among kurtosis estimators. The HL-kurtosis dominates all the other estimators with p-values close to zero. Ruppert's kurtosis performs significantly worse than the other estimators, which strengthens our observation made for Table 3 that Bowley's skewness is not efficient in screening biologically meaningful genes. The conventional kurtosis performs better than the L-kurtosis, but the degree of superiority is not statistically significant (p-value ≈ 0.9766), which implies that the comparison between

Skewness (# of screened genes)	L (3,295)	HL (3,222)	Conventional (3,092)	Bowley's (1,989)
L (3,295)		0.0097	**< 0.0001**	**< 0.0001**
HL (3,222)			**< 0.0001**	**< 0.0001**
Conventional (3,092)				**< 0.0001**

Table 3: The numbers of gene sets screened by different skewness estimators with the FDR level 0.05 and the significances of their differences based on McNemar's test. All differences are statistically significant after Bonferroni's adjustment except the difference between the L- and HL-skewness. The GCL-moments based estimators perform the best and Bowley's skewness performs the worst.

Kurtosis (# of screened genes)	HL (1,998)	Conventional (1,506)	L (1,504)	Ruppert's (762)
HL (1,998)		**< 0.0001**	**< 0.0001**	**< 0.0001**
Conventional (1,506)			0.9766	**< 0.0001**
L (1,504)				**< 0.0001**

Table 4: The numbers of screened gene sets by kurtosis estimators. The HL-kurtosis performs the best and Ruppert's kurtosis performs the worst. The superiority of the HL-kurtosis over the other estimators is statistically significant.

the conventional kurtosis and L-kurtosis is inconclusive. Overall, these results show that the HL-kurtosis is particularly useful in screening meaningful genes in the direction of distributional kurtosis.

To comprehensively compare measures of skewness and kurtosis that are based on different types of statistics, we combine comparison results in Tables 3 and 4. If a gene set is screened by either the HL-skewness or HL-kurtosis, we say that the gene set is screened by the measures based on the HL-moments. Otherwise, we say the gene set is not screened. This joint amount of gene set screening gives a quantitative summary of performance of a pair of skewness and kurtosis measures.

Table 5 shows the comparison results of combined skewness and kurtosis estimators. The abbreviation 'Q' stands for quantiles and indicates the combination of Bowley's skewness and Ruppert's kurtosis. It is seen that all the differences in the table are statistically significant even after Bonferroni's adjustment. The HL-moments based measures dominate all the other measures, and the quantile based measures perform significantly worse than the other measures. These results imply that the HL-moments are particularly good at screening biologically meaningful genes based on departure of their marginal distributions from the Gaussian distributions.

6.3. Comparisons among different HL-moments based estimators

As mentioned in Subsections 3.1, 4.1 and 4.2, various estimators of the HL-moments can be suggested. In this subsection, we compare the performances of those estimators by GSEA. The four HL-moments based estimators considered

Union (# of screened genes)	HL (4,642)	Conventional (4,300)	L (4,160)	Q (2,493)
HL (4,642)		**< 0.0001**	**< 0.0001**	**< 0.0001**
Conventional (4,300)			**0.0004**	**< 0.0001**
L (4,160)				**< 0.0001**

Table 5: The numbers of gene sets screened by different pairs of skewness and kurtosis measures and the significances of their differences. The HL-moment based estimators perform the best, while the sample quantile based estimators perform the worst. All the differences are statistically significant even after Bonferroni's adjustment.

Union (# of screened genes)	Advanced (4,716)	LaB (4,699)	BH (4,693)	Sample (4,642)
Advanced (4,716)		0.1839	0.191	**0.0002**
LaB (4,699)			0.7489	**0.0016**
BH (4,693)				**0.0033**

Table 6: The numbers of gene sets enriched by different HL-moments based estimators at the FDR level 0.05 and the significance of their differences based on McNemar's test. The advanced sample HL-moments perform better than all the other estimators.

herein are the LaB estimators in Equation (17), BH estimators in Equation (6), the advanced sample HL-moments in Equation (18), and the sample HL-moments in Equation (12). Note that the variance-covariance matrix V_n cannot be precisely calculated for $n = 817$. Hence, we use the method of David and Johnson (1954) as in LaBrecque (1977) to approximate the variance-covariance matrix V_{817} to calculate LaB estimators and the advanced sample HL-moments. For calculation of $E\left(H_r\left(Z_{i:n}\right)\right)$, we use numerical integration for both the estimators. The skewness and kurtosis estimators are the ratios of the third and fourth estimators to the scale estimators in the definitions of both the estimators. For the comparisons in this subsection, se only present combined comparison results of the skewness and kurtosis estimators.

The comparison results are given in Table 6. Among the estimators, the advanced sample HL-moments perform the best, but its superiority over the LaB estimator is statistically nonsignificant. This seems to result from similarities between the advanced sample HL-moments and LaB estimators explained in Subsection 4.2. The sample HL-moments perform worse than the other estimators. Especially, significance of the difference between the advanced sample HL-moments and the sample HL-moments is very significant. This supports our claim in Subsection 4.2 that existence of V_n has an impact on performances of skewness and kurtosis estimators. In our GSEA, the advanced HL-moments best achieve a balance between sensitivity to departure from Gaussianity and robustness to outliers.

Abbreviation	Distribution name	Density function
Gumbel	Standard Gumbel	$\exp\left(-\left(x+e^{-x}\right)\right)$
Location MoG	Location-mixture of Gaussian	$0.9\phi(x)+0.1\phi(x-1)$
Scale MoG	Scale-mixture of Gaussian	$0.9\phi(x)+0.1\phi(x/3)$

Table 7: The four alternative hypothetical distributions in the goodness-of-fit test of Gaussianity used in Section 7. Skewness estimators are compared for the Gumbel and Location MoG distributions, and kurtosis estimators are compared for the Scale MoG distributions.

7. Goodness-of-fit test for Gaussianity

Section 6 of Brown and Hettmansperger (1996) performed the goodness-of-fit test of Gaussianity to show superiority of their estimators over the LaB estimators. We perform a similar experiment to compare performances of the BH estimator, LaB estimator, advanced sample HL-moments (*advanced estimator*) and sample L-moments (*L-estimator*), or equivalently sample RL-moments, to gain a deeper insight about their relative performances. The test assumes the following composite null and alternative hypotheses,

$$H_0 : F = \Phi\left(\cdot|\mu,\sigma^2\right) \text{ for some } \mu\in\mathbb{R},\sigma>0, \quad H_1 : \text{not } H_0.$$

The alternative hypothetical distributions used in Brown and Hettmansperger (1996) are summarized in Table 7. The paper pointed out that skewness estimators are most appropriate for the Gumbel and Location MoG distributions and kurtosis estimators are most appropriate for the Cauchy and the Scale MoG distributions. The Cauchy distribution is removed from our study since the simulated p-values of the four estimators on the distribution are mostly zero, which seems to result from the non-finite mean of the Cauchy distribution. The sample HL-moments are removed from our study as well since their relative performances to the other estimators were random and inconclusive. For the LaB and advanced estimators, we perform numerical integration to directly calculate the variance-covariance matrix of the standard Gaussian order statistics.

To compare their estimators with LaBrecque (1977), the paper Brown and Hettmansperger (1996) compared p-value intervals that the BH and LaB estimates belong to. We improve this method by directly simulating p-values from estimates generated by the Monte Carlo method on the null standard Gaussian distribution. To illustrate, we generate $n_0 = 50,000$ random samples, each of which has the size $n = 20$, from the standard Gaussian distribution to obtain n_0 test statistic values called *null values*. Next, we generate $n_a = 10,000$ random samples of the size n from an alternative distribution to obtain n_a values of the test statistic called *alternative values*. For each alternative value, we count the number of greater null values. The ratio of that number to n_a becomes a simulated p-value. For each pair of simulated p-values of two estimators, there are three categories of comparison results; the first or second simulated p-value is lower, or their values are the same. We calculate the final likelihood ratio test (LRT) p-value of comparison between the two estimators based on the n_a trial results of the three-outcome multinomial experiment.

	1st est.	2nd est.	1st better	Ties	2nd better	p-value
Standard Gumbel	LaB	BH	**4,953**	219	4,828	0.2063
	LaB	Adv.	4,057	225	**5,718**	**< 0.0001**
	BH	Adv.	4,259	303	**5,438**	**< 0.0001**
	LaB	L	4,733	99	**5,168**	**< 0.0001**
	BH	L	4,728	80	**5,192**	**< 0.0001**
	Adv.	L	4,843	69	**5,088**	0.0139
Location MoG	LaB	BH	4,936	26	**5,038**	0.3071
	LaB	Adv.	**5,056**	32	4,912	0.1492
	BH	Adv.	**5,055**	16	4,929	0.2073
	LaB	L	4,975	1	**5,024**	0.6241
	BH	L	4,974	1	**5,025**	0.6100
	Adv.	L	4,970	1	**5,029**	0.5552
Scale MoG	LaB	BH	4,074	151	**5,775**	**< 0.0001**
	LaB	Adv.	4,201	144	**5,655**	**< 0.0001**
	BH	Adv.	4,650	292	**4,998**	**0.0004**
	LaB	L	**5,516**	57	4,427	**< 0.0001**
	BH	L	**5,687**	77	4,236	**< 0.0001**
	Adv.	L	**5,620**	81	4,299	**< 0.0001**

Table 8: The numbers of times that one estimator has a lower simulated p-value than the other. The boldfaced counts in the third to fifth columns indicate that the corresponding estimators perform better. The boldfaced LRT p-values in the last column indicate that they are statistically significant after Bonferroni's adjustment.

The comparison results are given in Table 8. There are three sub-tables in the table, each of which presents comparison results for an alternative distribution. The title column shows the abbreviated names of alternative distributions given in Table 7. The first and second columns show the estimator names. The third to fifth columns show the numbers of times that the first estimator's simulated p-values are lower, both of the simulated p-values are the same, and the second estimator's simulated p-values are lower. The boldfaced numbers indicate that the corresponding estimator performs better in comparison, and boldfaced LRT p-values in the last column indicate that the LRT p-values are statistically significant after Bonferroni's adjustment for the same alternative distribution.

It is seen in Table 8 that the BH estimators perform better than the LaB estimators except for the standard Gumbel distribution, which coincides with Brown and Hettmansperger (1996). When compared with the LaB estimators, the advanced estimators perform significantly better on the standard Gumbel and Scale MoG. This result is reversed on the Location MoG, but the difference is not significant. Compared with the BH estimators, the advanced estimators perform significantly better on the standard Gumbel and Scale MoG distributions, but non-significantly worse on the Location MoG. Overall, the advanced estimators overcome the weakness of the LaB estimators against the BH esti-

mators. Comparison between the L-estimators and either of the LaB and BH estimators is inconclusive since the L-estimator performs better on the standard Gaussian, worse on the Scale MoG, and the differences are non-significant on the Location MoG. Comparison between the L-estimators and advanced estimators are inconclusive on the standard Gumbel and Location MoG, but the advanced estimator performs significantly better on the Scale MoG. These imply that the advanced estimators generally perform the best in our computational study.

8. Conclusion

In this paper, we developed robust measures of skewness and kurtosis that are centered at the Gaussian distributions and share the strengths of the L-moments. We showed that our measures indeed represent their desired distributional properties based on Oja's criteria, and analyzed their robustness based on the influence functions. Their consistent estimators were shown to be effective in screening non-Gaussian variables in high dimensional cancer research data, especially in the directions of distributional skewness and kurtosis.

Locating L-functionals at a different family of distributions than the Gaussian has been studied elsewhere. The theory of *trimmed L-moments* (*TL-moments*) was proposed by Elamir and Seheult (2003) as a more robust version of the L-moments, and Hosking (2007) showed that the TL-moments are actually centered at the logistic distributions. Consideration of other important distributions in data analysis such as t-distributions and mixtures of the Gaussian distributions as distributional centers remains as a future research topic.

References

An, H., 2017. Gaussian Centered L-moments. Ph.D. thesis. The University of North Carolina at Chapel Hill.

Bowley, A.L., 1920. Elements of Statistics, 4th Edition. New York: Scribner's.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. (With discussion). J. Roy. Statist. Soc. Ser. B 26, 211–252.

Brown, B.M., Hettmansperger, T.P., 1996. Normal scores, normal plots, and tests for normality. J. Amer. Statist. Assoc. 91, 1668–1675. doi:10.2307/2291594.

Brys, G., Hubert, M., Struyf, A., 2004. A robust measure of skewness. J. Comput. Graph. Statist. 13, 996–1017. doi:10.1198/106186004X12632.

Brys, G., Hubert, M., Struyf, A., 2006. Robust measures of tail weight. Comput. Statist. Data Anal. 50, 733–759. doi:10.1016/j.csda.2004.09.012.

David, F., Johnson, N., 1954. Statistical treatment of censored data part I. fundamental formulae. Biometrika 41, 228–240.

David, H.A., Nagaraja, H.N., 2003. Order statistics. Wiley Series in Probability and Statistics. third ed., Wiley-Interscience John Wiley & Sons, Hoboken, NJ. doi:10.1002/0471722162.

Decurninge, A., 2014. Multivariate quantiles and multivariate L-moments. arXiv preprint arXiv:1409.6013 .

Doksum, K.A., 1975. Measures of location and asymmetry. Scand. J. Statist. 2, 11–22.

Dutang, C., 2017. Theoretical L-moments and TL-moments using combinatorial identities and finite operators. Communications in Statistics-Theory and Methods 46, 3801–3828.

Elamir, E.A., Seheult, A.H., 2003. Trimmed l-moments. Computational Statistics & Data Analysis 43, 299–314.

Feller, W., 1968. An introduction to probability theory and its applications. Vol. I. Third edition, John Wiley & Sons, Inc., New York-London-Sydney.

Feng, Q., Hannig, J., Marron, J.S., 2016. A note on automatic data transformation. Stat 5, 82–87.

Gibbons, J.D., Chakraborti, S., 2010. Nonparametric statistical inference. Fifth edition, Chapman and Hall/CRC. doi:10.1201/9781439896129.

Groeneveld, R.A., 1991. An influence function approach to describing the skewness of a distribution. Amer. Statist. 45, 97–102. doi:10.2307/2684367.

Hall, P., 1983. Inverting an Edgeworth expansion. Ann. Statist. 11, 569–576. doi:10.1214/aos/1176346162.

Hosking, J.R.M., 1990. L-moments: analysis and estimation of distributions using linear combinations of order statistics. J. Roy. Statist. Soc. Ser. B 52, 105–124.

Hosking, J.R.M., 2007. Some theory and practical uses of trimmed L-moments. J. Statist. Plann. Inference 137, 3024–3039. doi:10.1016/j.jspi.2006.12.002.

Huber, P.J., Ronchetti, E.M., 2009. Robust statistics. Wiley Series in Probability and Statistics. second ed., John Wiley & Sons, Inc., Hoboken, NJ. doi:10.1002/9780470434697.

Jarque, C.M., Bera, A.K., 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Econom. Lett. 6, 255–259. doi:10.1016/0165-1765(80)90024-5.

Kim, T.H., White, H., 2004. On more robust estimation of skewness and kurtosis. Finance Research Letters 1, 56–73. doi:https://doi.org/10.1016/S1544-6123(03)00003-5.

LaBrecque, J., 1977. Goodness-of-fit tests based on nonlinearity in probability plots. Technometrics 19, 293–306.

Leslie, J.R., 1984. Asymptotic properties and new approximations for both the covariance matrix of normal order statistics and its inverse, in: Colloquia Mathematica Societatis János Bolyai, Goodness-of-Fit Debrecen„ Hungry.

Lloyd, E., 1952. Least-squares estimation of location and scale parameters using order statistics. Biometrika 39, 88–95.

Marcinkiewicz, J., 1939. Sur une propriété de la loi de Gauß. Math. Z. 44, 612–618. doi:10.1007/BF01210677.

Marron, J.S., Dryden, I.L., 2021. Object Oriented Data Analysis. Chapman and Hall/CRC.

Martinez, J., Iglewicz, B., 1984. Some properties of the Tukey g and h family of distributions. Comm. Statist. A—Theory Methods 13, 353–369. doi:10.1080/03610928408828687.

Necir, A., Meraghni, D., 2010. Estimating L-functionals for heavy-tailed distributions and application. J. Probab. Stat. 2010, Art. ID 707146, 34.

Oja, H., 1981. On location, scale, skewness and kurtosis of univariate distributions. Scand. J. Statist. 8, 154–168.

Pearson, K., 1916. Mathematical contributions to the theory of evolution. XIX. Second supplement to a memoir on skew variation. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 216, 429–457.

Ruppert, D., 1987. What is kurtosis? An influence function approach. Amer. Statist. 41, 1–5. doi:10.2307/2684309.

Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52, 591–611.

Shorack, G.R., 1972. Functions of order statistics. Ann. Math. Statist. 43, 412–427. doi:10.1214/aoms/1177692622.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U.S.A. 102, 15545–15550.

Szegö, G., 1959. Orthogonal polynomials. American Mathematical Society Colloquium Publications, Vol. 23. Revised ed, American Mathematical Society, Providence, R.I.

Weinstein, J.N., ⋯, Stuart, J.M., 2013. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 45, 1113–1120.