Gaussian Centered L-moments

STOR 893 Object Oriented Data Analysis

Hyowon An

April 5, 2016

Outline

Preliminaries

Notation

- $X \sim F$, $Y \sim G$
- $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$
- Every distribution is absolutely continuous and strictly increasing in the sense that it is strictly increasing on its support

$$S(F) = \{x | 0 < F(x) < 1\}.$$

- F^{-1}: Quantile of F
- $\phi(\cdot|\mu, \sigma^2), \Phi(\cdot|\mu, \sigma^2)$: PDF and CDF of $\mathcal{N}(\mu, \sigma^2)$

Skewness and kurtosis revisited

- $\mu_k = EX^k$: k-th moment
- The first four cumulants are

$$\begin{aligned}
\kappa_1 &= \mu_1, \\
\kappa_2 &= \mu_2 - \mu_1^2, \\
\kappa_3 &= \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3, \\
\kappa_4 &= \mu_4 - 4\mu_3\mu_1 - 3\mu_2^2 + 12\mu_2\mu_1^2 - 6\mu_1^4.
\end{aligned}$$

- The (conventional) skewness γ_1 and (conventional excess) kurtosis γ_2 are

$$\gamma_1 = \frac{E\left(X - EX\right)^3}{\left\{E\left(X - EX\right)^2\right\}^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}, \quad \gamma_2 = \frac{E\left(X - EX\right)^4}{\left\{E\left(X - EX\right)^2\right\}^2} - 3 = \frac{\kappa_4}{\kappa_2^2}.$$

- The 3rd and higher order cumulants are zero for the Gaussian distributions by the Marcinkiewicz theorem.
  - $\Rightarrow$ The conventional skewness and kurtosis are zero for the Gaussian distributions.

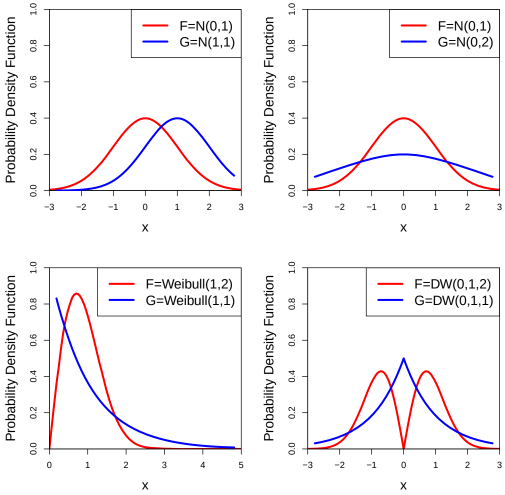


Figure 1: Two distributions with different location, scale, skewness and kurtosis

Motivation

- To define conventional skewness and kurtosis, we need to have

$$E\left(|X|^3\right) < \infty, \; E\left(|X|^4\right) < \infty$$

respectively.

- The Sample (conventional) skewness and Sample (conventional) kurtosis

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

can be highly driven by some observations X_i having large values.

- Robust measures of skewness and kurtosis of a distribution is needed.

L-statistics and the L-moments

L-statistics

- An L-statistic is

$$\tilde{\theta} = \frac{1}{n} \sum_{i=1}^{n} c_{ni} X_{i:n}$$

where c_{ni} is a constant depending on both i and n.

- The sample median is an L-statistic;

$$m = X_{\lfloor 2/n \rfloor : n}.$$

- The sample trimmed mean is an L-statistic;

$$m_\alpha = \frac{1}{n(1-2\alpha)} \left(X_{\lfloor \alpha n + 1 \rfloor : n} + X_{\lfloor \alpha n + 2 \rfloor : n} + \cdots + X_{(n - \lfloor \alpha n \rfloor) : n} \right).$$

- Often, we introduce a continuous function $h : (0,1) \to \mathbb{R}$ yielding

$$\tilde{\theta} = \frac{1}{n} \sum_{i=1}^{n} h\left(\frac{i}{n+1}\right) X_{i:n}$$

- We have

$$\frac{1}{n} \sum_{i=1}^{n} h\left(\frac{i}{n+1}\right) X_{i:n} \overset{\text{a.s.}}{\to} \int_0^1 F^{-1}(u) h(u)\, \mathrm{d}u$$

  when F and h satisfy some of the conditions, e.g. (Serfling, 1980).

L-moments

- The r-th L-moment (Hosking, 1990) is

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k \left(\begin{array}{c} r-1 \\ k \end{array} \right) EX_{r-k:r}.$$

- The first four L-moments are

$$\begin{aligned} \lambda_1 &= EX_{1:1}, \\ \lambda_2 &= \frac{1}{2}E\left(X_{2:2} - X_{1:2}\right), \\ \lambda_3 &= \frac{1}{3}E\left(X_{3:3} - 2X_{2:3} + X_{1:3}\right), \\ \lambda_4 &= \frac{1}{4}E\left(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}\right). \end{aligned}$$

- For a random variable $X \sim F$ such that $E|X| < \infty$,

$$-\infty < \lambda_r(F) < \infty \text{ for all } r = 1, 2, \cdots.$$

$$\lambda_1 = EX_{1:1}$$

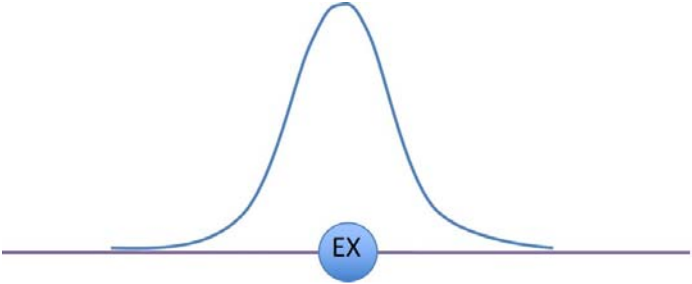


Figure 2: A pictorial description of the first order L-moment (Kimes, 2013)

$$\lambda_2 = \frac{1}{2} E\left(X_{2:2} - X_{1:2}\right)$$

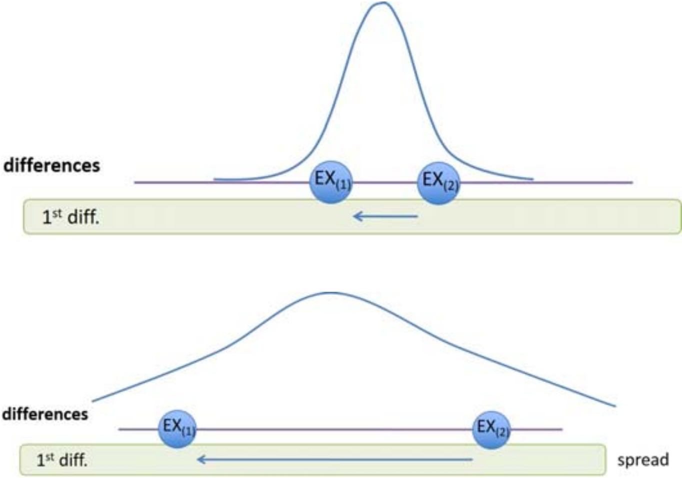


Figure 3: A pictorial description of the second order L-moment (Kimes, 2013)

$$\lambda_3 = \frac{1}{3} E\left(X_{3:3} - 2X_{2:3} + X_{1:3}\right)$$

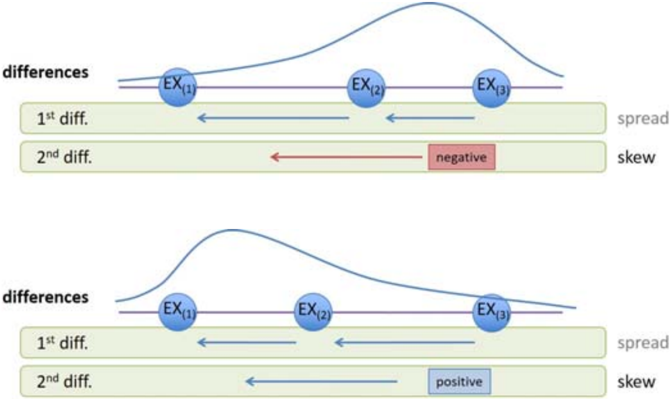


Figure 4: A pictorial description of the third order L-moment (Kimes, 2013)

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

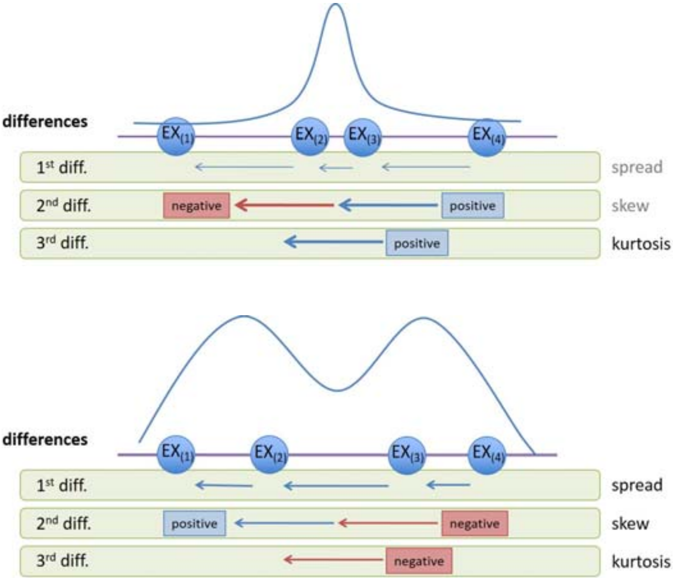


Figure 5: A pictorial description of the fourth order L-moment (Kimes, 2013)

- The r-th L-moment λ_r is alternatively expressed as

$$\lambda_r(F) = \int_{-\infty}^{\infty} x f(x) P^*_{r-1}(F(x)) \mathrm{d}x = \int_0^1 F^{-1}(u) P^*_{r-1}(u) \mathrm{d}u$$

where P^*_r is the r-th order shifted Legendre polynomial.

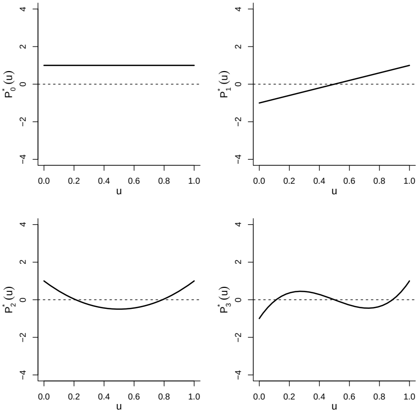


Figure 6: The shifted Legendre polynomials

- The shifted Legendre polynomials P_r^* are orthogonal to each other with respect to the weight function $w : (0,1) \to \mathbb{R}$ such that

$$w(x) = 1$$

for all $0 < x < 1$. That is,

$$\int_0^1 P_{r_1}^*(x) P_{r_2}^*(x) \mathrm{d}x = 0.$$

for all $r_1 \neq r_2$.

- The paper (Hosking, 1990) showed that

$$\tilde{\lambda}_r = \frac{1}{n} \sum_{i=1}^{n} P_{r-1}^* \left(\frac{i}{n+1} \right) X_{i:n} \overset{\text{a.s.}}{\to} \int_0^1 F^{-1}(u) P_{r-1}^*(u) \, \mathrm{d}u = \lambda_r.$$

- $\tilde{\lambda}_r$ is less affected by outliers than $\hat{\gamma}_1$ or $\hat{\gamma}_2$.

Gaussian Centered L-moments

Gaussian Centered L-moments: Motivation

- The cumulants are not robust..
- It can be seen that

$$\lambda_r(\text{Uniform}(a,b)) = \int_0^1 \{(b-a)x + a\} P^*_{r-1}(x)\,\mathrm{d}x = 0$$

for all $-\infty < a < b < \infty$ and $r = 3, 4, \cdots$ by the orthogonality of P^*_r.

$\Rightarrow$ The L-moments are centered at the uniform distributions.

- The signs and absolute values of $\lambda_r(F)$ do not tell us the relationship between F and Φ;

$$\lambda_4(F) \geq 0 \overset{?}{\Rightarrow} F \text{ is more kurtotic than } \Phi.$$

- Data usually follow the Gaussian distributions after suitable transformations.

A variation of the L-moments centered at the Gaussian distributions that

- only needs a random variable X to satisfy $E|X| < \infty$,
- has an strongly consistent L-statistic,
- is centered at the Gaussian distributions.

Gaussian Centered L-moments: Definition

Definition

Functionals $\{\theta_r : \mathcal{F} \to \mathbb{R} | r = 1, 2, \cdots\}$ are called Gaussian Centered L-moments (GCL-moments) of $\mathcal{F}$ if they are

1. L-functionals: $\exists h_r : (0,1) \to \mathbb{R}$ such that

$$\theta_r(F) = \int_0^1 F^{-1}(u) h_r(u)\, \mathrm{d}u.$$

2. Centered at the Gaussian distributions:

$$\theta_r(\Phi(\cdot|\mu, \sigma^2)) = 0 \text{ for all } \mu \in \mathbb{R}, \sigma^2 > 0, r = 3, 4, \cdots .$$

- The L-moments are L-functionals since

$$\lambda_r(F) = \int_0^1 F^{-1}(u) P^*_{r-1}(u) \mathrm{d}u.$$

- Under some conditions on F and h_r,

$$\tilde{\theta}_r = \sum_{i=1}^n h_r\left(\frac{i}{n+1}\right) X_{i:n} \overset{\text{a.s.}}{\to} \int_0^1 F^{-1}(u) h_r(u)\, \mathrm{d}u = \theta_r.$$

Two versions:

- Hermite L-moments (HL-moments)
- Gaussian Rescaled L-moments (GRL-moments)

Hermite L-moments

- The L-moments
$$\lambda_r(F) = \int_{-\infty}^{\infty} x f(x) P^*_{r-1}(F(x)) \mathrm{d}x$$
are centered at the unform distributions since
$$\lambda_r(\text{Uniform}(a,b)) = \int_0^1 \{(b-a)x + a\} \, P^*_{r-1}(x) \, \mathrm{d}x = 0$$
for all $-\infty < a < b < \infty$ and $r = 3, 4, \cdots$.
- The Hermite polynomials H_r are orthogonal to each other with respect to the weight function $w : \mathbb{R} \to \mathbb{R}$ such that
$$w(x) = e^{x^2/2}$$
for all $x \in \mathbb{R}$. That is,
$$\int_{-\infty}^{\infty} e^{x^2/2} H_{r_1}(x) H_{r_2}(x) \mathrm{d}x = 0.$$
for all $r_1 \neq r_2$.

- The following L-functionals

$$\eta_r(F) = \int_{-\infty}^{\infty} x f(x) H_{r-1}\left(\Phi^{-1}(F(x))\right) \mathrm{d}x$$

are centered at the Gaussian distributions since

$$\eta_r(\Phi(\cdot|\mu,\sigma^2)) = \int_{-\infty}^{\infty} (\mu + \sigma x)\phi(x) H_{r-1}(x)\,\mathrm{d}x = 0$$

for all $r = 3, 4, \cdots$, $\mu \in \mathbb{R}$ and $\sigma > 0$ where H_r is the r-th order Hermite polynomial.

- The r-th Hermite L-moment (HL-moment):

$$\eta_r = \int_{-\infty}^{\infty} x f(x) H_{r-1}\left(\Phi^{-1}(F(x))\right) \mathrm{d}x$$

- The r-th sample Hermite L-moment (sample HL-moment):

$$\tilde{\eta}_r = \frac{1}{n} \sum_{i=1}^{n} H_{r-1}\left(\Phi^{-1}\left(\frac{i}{n+1}\right)\right) X_{i:n}$$

- The HL-moments are GCL-moments.
- The first four HL-moments satisfy Oja's criterion.

Hermite L-moments

- Recall that the skewness and kurtosis are defined as

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

- The Hermite L-skewness (HL-skewness) and Hermite L-kurtosis (HL-kurtosis) are defined as

$$\eta_3^* = \frac{\eta_3}{\eta_2}, \quad \eta_4^* = \frac{\eta_4}{\eta_2}.$$

- The sample Hermite L-skewness (sample HL-skewness) and sample Hermite L-kurtosis (sample HL-kurtosis) are defined as

$$\tilde{\eta}_3^* = \frac{\tilde{\eta}_3}{\tilde{\eta}_2}, \quad \tilde{\eta}_4^* = \frac{\tilde{\eta}_4}{\tilde{\eta}_2}.$$

Asymptotic distribution of sample HL-moments

- Recall that

$$\tilde{\theta}_r = \frac{1}{n} \sum_{i=1}^{n} h_r \left(\frac{i}{n+1} \right) X_{i:n} \overset{\text{a.s.}}{\to} \int_0^1 F^{-1}(u) h_r(u) \, \mathrm{d}u = \theta_r$$

under some conditions on F and h_r.

- We showed that

$$\sqrt{n} \left(\begin{pmatrix} \tilde{\eta}_{n,2} \\ \tilde{\eta}_{n,3} \\ \tilde{\eta}_{n,4} \end{pmatrix} - \begin{pmatrix} \eta_2 \\ \eta_3 \\ \eta_4 \end{pmatrix} \right) \overset{d}{\to} \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{22} & \sigma_{23} & \sigma_{34} \\ \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix} \right)$$

as $n \to \infty$ where

$$\sigma_{r_1 r_2} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{F(x \wedge y) - F(x)F(y)\} H_{r_1 - 1}\left(\Phi^{-1}(F(x))\right) H_{r_2 - 1}\left(\Phi^{-1}(F(x))\right) \mathrm{d}x \, \mathrm{d}y$$

where $x \wedge y = \min\{x, y\}$.

- We showed that

$$\sqrt{n}\left(\left(\begin{array}{c}\tilde{\eta}^*_{n,3}\\ \tilde{\eta}^*_{n,4}\end{array}\right)-\left(\begin{array}{c}\eta^*_3\\ \eta^*_4\end{array}\right)\right)\xrightarrow{d}\mathcal{N}\left(\left(\begin{array}{c}0\\ 0\end{array}\right),\Psi\right)$$

as $n\to\infty$ where

$$D=\frac{1}{\eta_2^2}\left(\begin{array}{cc}\left(\eta_3^*\right)^2\sigma_{22}-2\eta_3^*\sigma_{23}+\sigma_{33} & \eta_3^*\eta_4^*\sigma_{22}-\eta_4^*\sigma_{23}-\eta_3^*\sigma_{24}+\sigma_{34}\\ \eta_3^*\eta_4^*\sigma_{22}-\eta_4^*\sigma_{23}-\eta_3^*\sigma_{24}+\sigma_{34} & \left(\eta_4^*\right)^2\sigma_{22}-2\eta_4^*\sigma_{24}+\sigma_{44}\end{array}\right).$$

- We showed that the sample HL-skewness and kurtosis are asymptoticallly independent for the Gaussian distributions, i.e.

$$\lim_{n\to\infty}\mathsf{Cov}_\Phi\left(n^{1/2}\tilde{\eta}^*_{n,r_1},n^{1/2}\tilde{\eta}^*_{n,r_2}\right)=0$$

for all $r_1\neq r_2$.

Gaussian Rescaled L-moments

For the Gaussian distributions,

$$\begin{aligned}\lambda_3 &= 0,\\ \lambda_4 &= \frac{1}{4}E\left(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}\right)\\ &= \frac{1}{4}\left\{E\left(X_{4:4} - X_{1:4}\right) - 3E\left(X_{3:4} - X_{2:4}\right)\right\}\\ &\neq 0.\end{aligned}$$

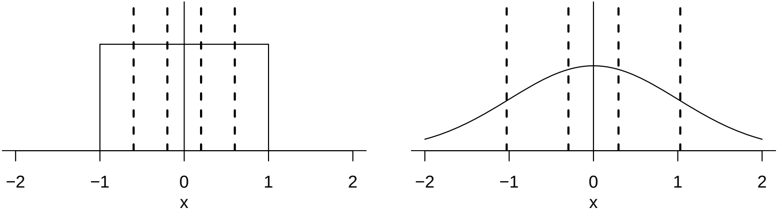


Figure 7: Four expected order statistics of Uniform$(-1, 1)$ and $\mathcal{N}(0, 1)$

- Recall that

$$
\begin{aligned}
\lambda_1 &= EX_{1:1}, \\
\lambda_2 &= \frac{1}{2}E\left(X_{2:2} - X_{1:2}\right), \\
\lambda_3 &= \frac{1}{3}\left\{E\left(X_{3:3} - X_{2:3}\right) - E\left(X_{2:3} - X_{1:3}\right)\right\}, \\
\lambda_4 &= \frac{1}{4}\left\{E\left(X_{4:4} - X_{3:4}\right) - E\left(X_{3:4} - X_{2:4}\right)\right\} \\
&\quad -\frac{1}{4}\left\{E\left(X_{3:4} - X_{2:4}\right) - E\left(X_{2:4} - X_{1:4}\right)\right\}.
\end{aligned}
$$

The first four Gaussian Rescaled L-moments (GRL-moments) are

$$
\begin{aligned}
\rho_1 &= EX_{1:1}, \\
\rho_2 &= \frac{1}{2}\left\{\frac{1}{\delta_{1,2:2}(Z)}E\left(X_{2:2}-X_{1:2}\right)\right\} \\
&\approx 0.8862\lambda_2, \\
\rho_3 &= \frac{1}{3}\left\{\frac{1}{\delta_{2,3:3}(\Phi)}E\left(X_{3:3}-X_{2:3}\right)-\frac{1}{\delta_{1,2:3}(\Phi)}E\left(X_{2:3}-X_{1:3}\right)\right\} \\
&\approx 1.1816\lambda_3, \\
\rho_4 &= \frac{1}{4}\left\{\frac{1}{\delta_{3,4:4}(\Phi)}E\left(X_{4:4}-X_{3:4}\right)-\frac{1}{\delta_{2,3:4}(\Phi)}E\left(X_{3:4}-X_{2:4}\right)\right\} \\
&\quad -\frac{1}{4}\left\{\frac{1}{\delta_{2,3:4}(\Phi)}E\left(X_{3:4}-X_{2:4}\right)-\frac{1}{\delta_{1,2:4}(\Phi)}E\left(X_{2:4}-X_{1:4}\right)\right\}
\end{aligned}
$$

where $\delta_{i,j:k}(\Phi)=E\left(Z_{j:k}-Z_{i:k}\right)$ for $1\le i\le j\le k$ and $Z\sim\Phi$.

- The r-th GRL-moment has the integral representation as follows,

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

where

$$\begin{aligned} R_0(u) &= P_0^*(u) \\ R_1(u) &\approx 0.8862 P_1^*(u) \\ R_2(u) &\approx 1.1816 P_2^*(u) \\ R_3(u) &\approx (6c+2)u^3 - 3(3c+1)u^2 + (3c+3)u - 1 \end{aligned}$$

and $c = 3.4658$.

- The polynomials R_r are not orthogonal to each other with respect to the weight function $w(x) = 1$.

- The r-th sample GRL-moments are

$$\tilde{\rho}_r = \frac{1}{n} \sum_{i=1}^{n} R_{r-1}\left(\frac{i}{n+1}\right) X_{i:n}$$

- The GRL-moments are GCL-moments.
- The first four GRL-moments satisfy Oja's criterion, i.e. those are measures of location, scale, skewness and kurtosis of a distribution.

- The Gaussian Rescaled L-skewness (GRL-skewness) and Gaussian Rescaled L-kurtosis (GRL-kurtosis) are defined as

$$\rho_3^* = \frac{\rho_3}{\rho_2}, \quad \rho_4^* = \frac{\rho_4}{\rho_2}.$$

- The sample Gaussian Rescaled L-skewness (sample GRL-skewness) and sample Gaussian Rescaled L-kurtosis (sample GRL-kurtosis) are defined as

$$\tilde{\rho}_3^* = \frac{\tilde{\rho}_3}{\tilde{\rho}_2}, \quad \tilde{\rho}_4^* = \frac{\tilde{\rho}_4}{\tilde{\rho}_2}.$$

Robustness of GCL-moments

Comparison of robustness of GCL-moments

- Recall that L-functionals are in the form

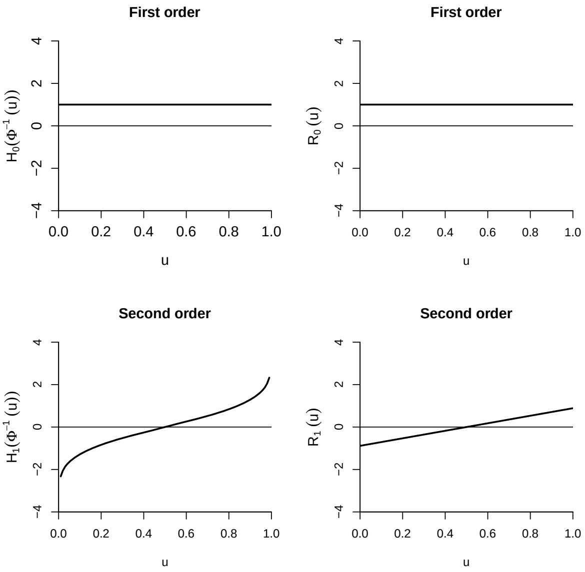
$$\theta(F) = \int_0^1 F^{-1}(u)h(u)\,\mathrm{d}u$$

for some function $h : (0,1) \to \mathbb{R}$.

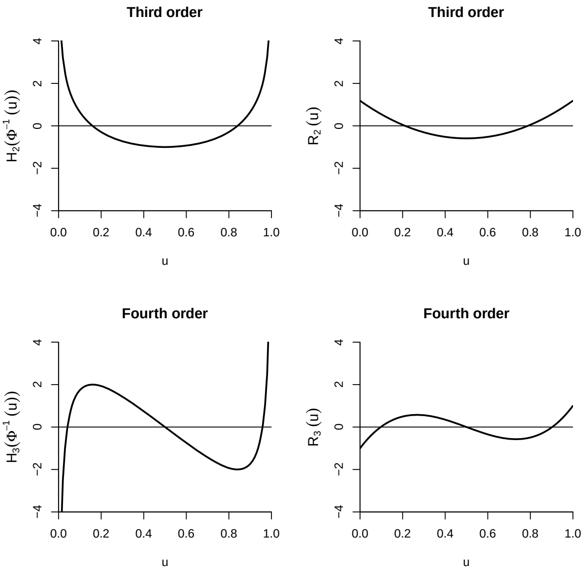
- Note that

$$\begin{aligned} \eta_r &= \int_0^1 F^{-1}(u) H_{r-1}\left(\Phi^{-1}(u)\right)\,\mathrm{d}u, \\ \rho_r &= \int_0^1 F^{-1}(u) R_{r-1}(u)\,\mathrm{d}u. \end{aligned}$$

Comparison of polynomials in GCL-moments



Comparison of polynomials in GCL-moments



Influence function

- $\mathcal{M}$: Class of probability measures
- A functional T is Gâteaux differentiable at $F \in \mathcal{M}$ if there is a linear functional L_F such that for all $G \in \mathcal{M}$

$$\lim_{t \to 0} \frac{T(F_t) - T(F)}{t} = L_F(G - F)$$

  where

$$F_t = (1-t)F + tG$$

- The influence function of a Gâteaux differentiable functional T evaluated at $F \in \mathcal{M}$ is

$$IC(x; F, T) = \lim_{t \to 0} \frac{T(F_t) - T(F)}{t}$$

  where $F_t = (1-t)F + t\delta_x$ and $x \in \mathbb{R}$.

Influence function of conventional moments

- It was shown in (Groeneveld, 1991) that

$$
\begin{aligned}
IC\left(x;F,\gamma_1\right) &= x^3 - 3x \\
&= H_3(x)
\end{aligned}
$$

  for all F which is symmetric and cube integrable.

- It was shown in (Ruppert, 1987) that

$$
\begin{aligned}
IC\left(x;F,\gamma_2\right) &= x^4 - 6x + 3 \\
&= H_4(x)
\end{aligned}
$$

  for all F which is symmetric and fourth power integrable.

Influence functions of GCL-moments

- We showed that

$$\begin{aligned} \mathsf{IC}\left(x;\Phi,\eta_r^*\right) &= \frac{1}{r}H_r(x), \\ \mathsf{IC}\left(x;\Phi,\rho_r^*\right) &= \int_{-\infty}^{0}\Phi(y)R_{r-1}(\Phi(y))\,\mathrm{d}y \\ &\quad -\int_{0}^{\infty}\{1-\Phi(y)\}R_{r-1}(\Phi(y))\,\mathrm{d}y \\ &\quad +\int_{0}^{x}R_{r-1}(\Phi(y))\,\mathrm{d}y \end{aligned}$$

  for $r = 3, 4, \cdots$.

- Note that

$$\begin{aligned} \left|\mathsf{IC}\left(x;\Phi,\eta_r^*\right)\right| &= O\left(|x|^r\right), \\ \left|\mathsf{IC}\left(x;\Phi,\rho_r^*\right)\right| &= O\left(|x|\right) \end{aligned}$$

  for all $r = 1, 2, \cdots$.

Comparison of polynomials in GCL-moments

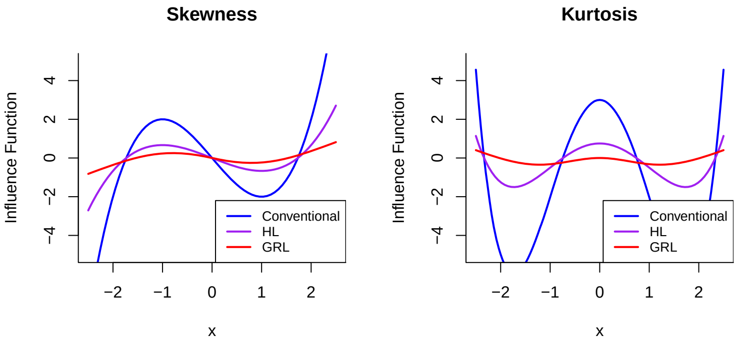


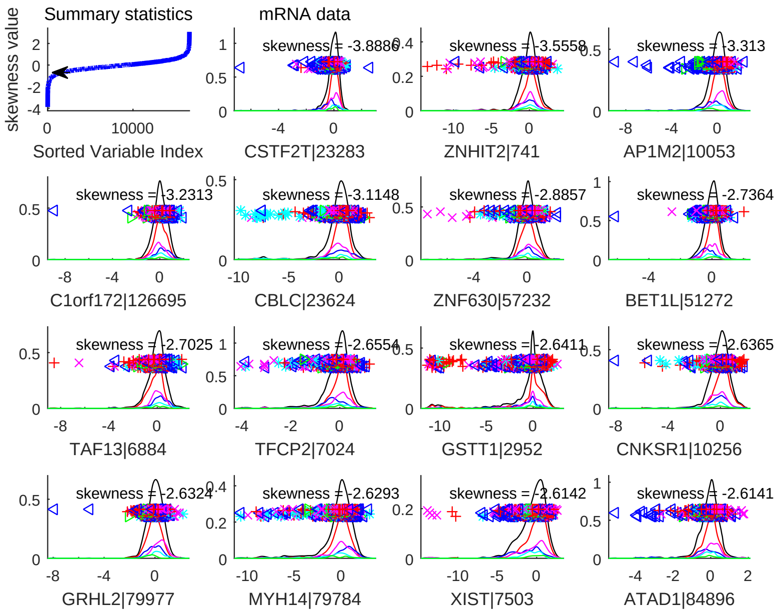
Figure 8: Influence curves of various moments at the standard Gaussian distribution

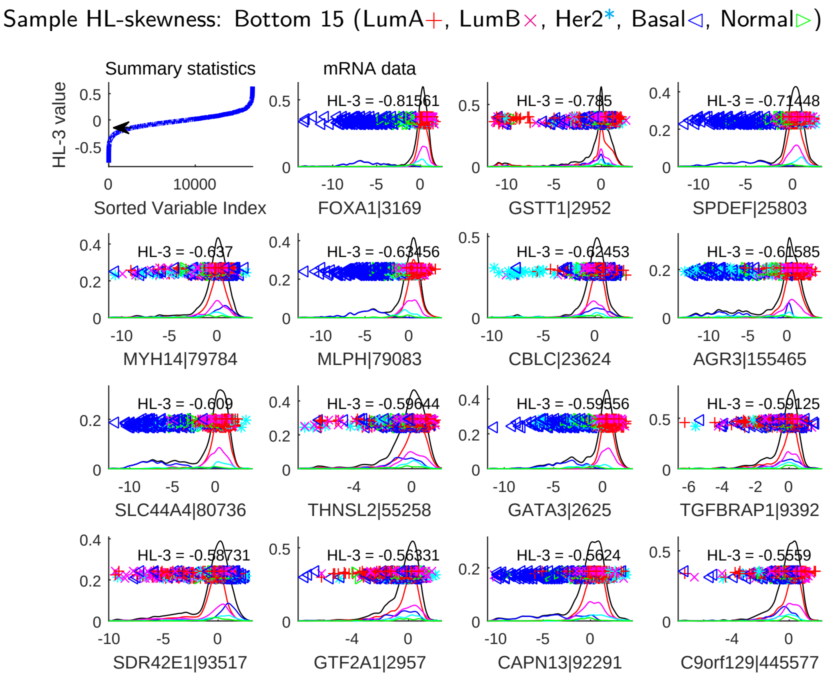
Analysis of TCGA lobular freeze data

TCGA Lobular Freeze data

- Gene expressions of breast cancer patients
- 16,615 genes, 817 cases
- 5 subtypes
  - LumA $+$
  - LumB $\times$
  - Her2 $*$
  - Basal $\triangleleft$
  - Normal-like $\triangleright$
- Looking for genes in which the distributions of different subtypes are best separated from each other.

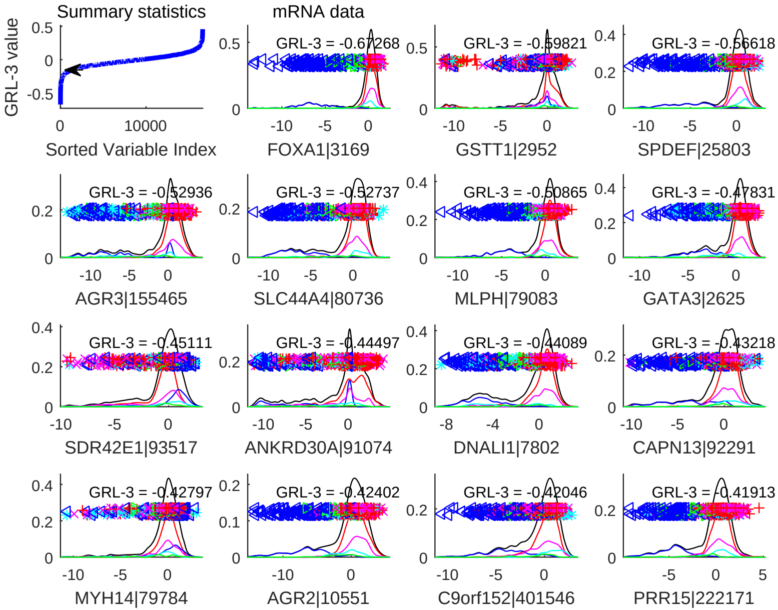
Sample skewness: Bottom 15 (LumA+, LumB×, Her2*, Basal◁, Normal▷)



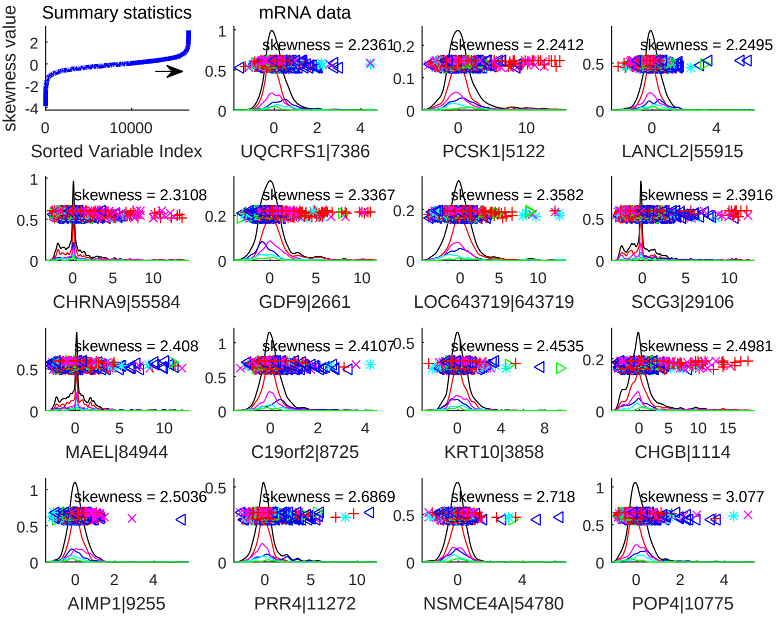


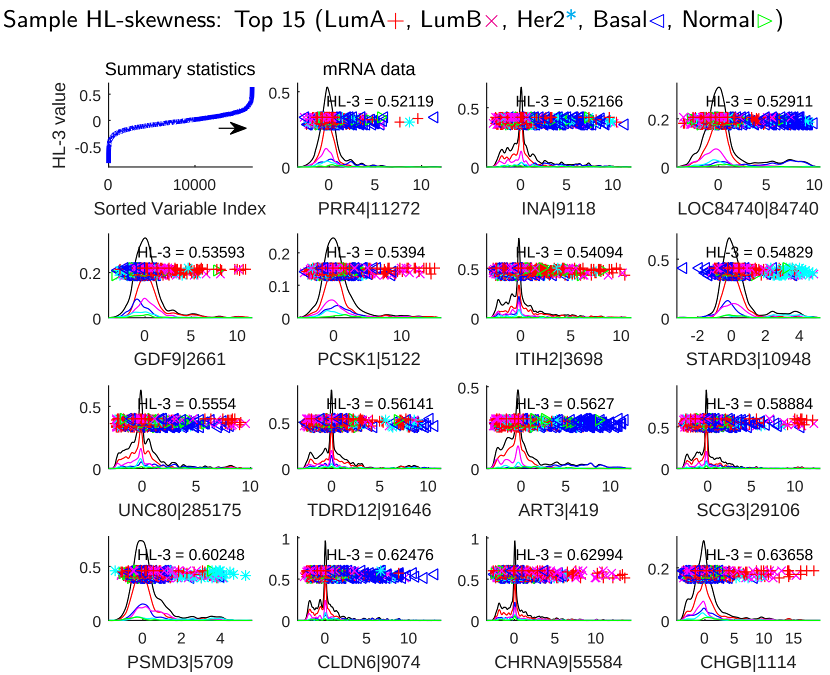
Sample HL-skewness: Bottom 15 (LumA+, LumB×, Her2*, Basal◁, Normal▷)
Summary statistics
mRNA data
HL-3 value
Sorted Variable Index
HL-3 = -0.81561
FOXA1|3169
HL-3 = -0.785
GSTT1|2952
HL-3 = -0.71448
SPDEF|25803
HL-3 = -0.637
MYH14|79784
HL-3 = -0.63456
MLPH|79083
HL-3 = -0.62453
CBLC|23624
HL-3 = -0.61585
AGR3|155465
HL-3 = -0.609
SLC44A4|80736
HL-3 = -0.59644
THNSL2|55258
HL-3 = -0.59556
GATA3|2625
HL-3 = -0.59125
TGFBRAP1|9392
HL-3 = -0.58731
SDR42E1|93517
HL-3 = -0.56331
GTF2A1|2957
HL-3 = -0.5624
CAPN13|92291
HL-3 = -0.5559
C9orf129|445577

Sample GRL-skewness: Bottom 15 (LumA+, LumB×, Her2*, Basal◁, Normal▷)

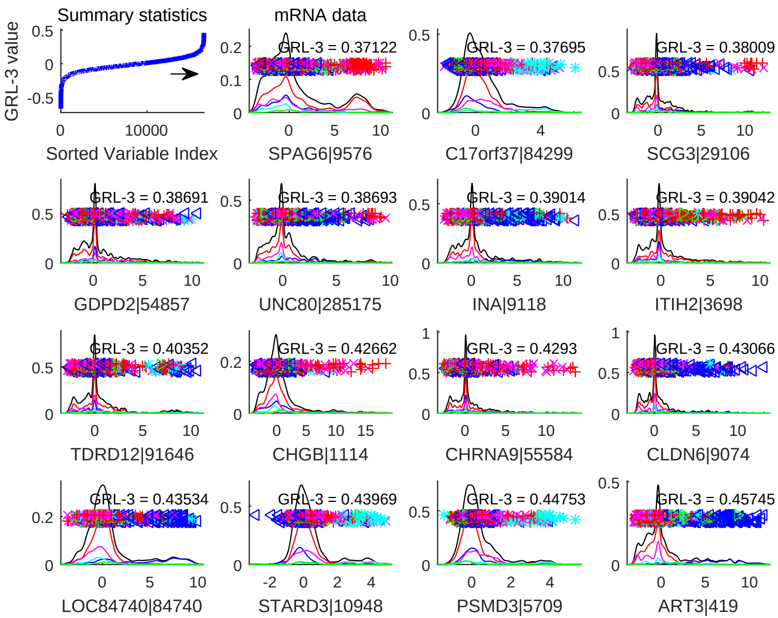


Sample skewness: Top 15 (LumA+, LumB×, Her2*, Basal◁, Normal▷)

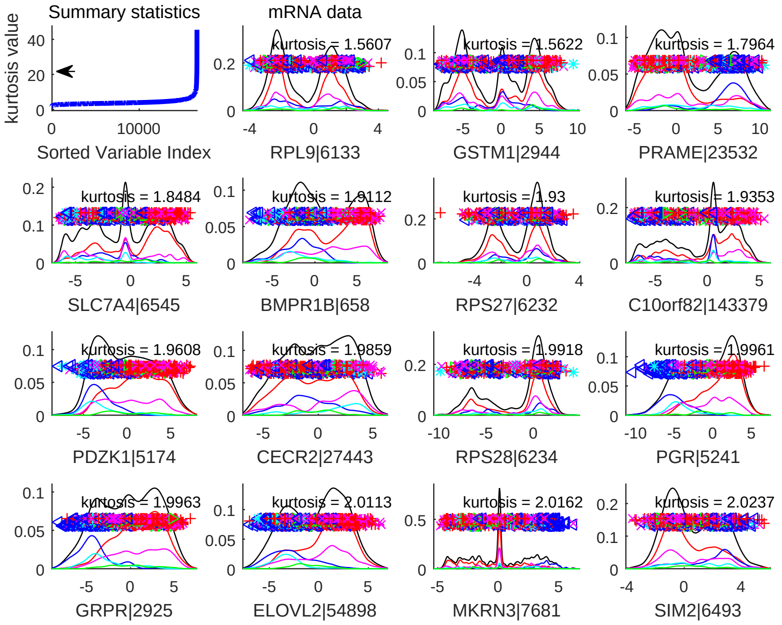


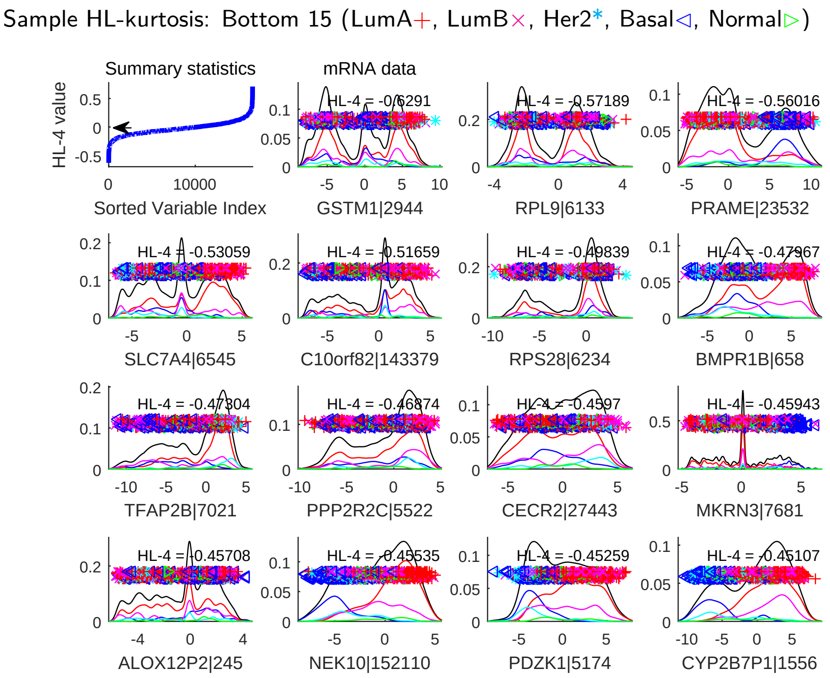

Sample HL-skewness: Top 15 (LumA+, LumB×, Her2*, Basal◁, Normal▷)
Summary statistics
mRNA data
HL-3 value
Sorted Variable Index
HL-3 = 0.52119
PRR4|11272
HL-3 = 0.52166
INA|9118
HL-3 = 0.52911
LOC84740|84740
HL-3 = 0.53593
GDF9|2661
HL-3 = 0.5394
PCSK1|5122
HL-3 = 0.54094
ITIH2|3698
HL-3 = 0.54829
STARD3|10948
HL-3 = 0.5554
UNC80|285175
HL-3 = 0.56141
TDRD12|91646
HL-3 = 0.5627
ART3|419
HL-3 = 0.58884
SCG3|29106
HL-3 = 0.60248
PSMD3|5709
HL-3 = 0.62476
CLDN6|9074
HL-3 = 0.62994
CHRNA9|55584
HL-3 = 0.63658
CHGB|1114

Sample GRL-skewness: Top 15 (LumA+, LumB×, Her2*, Basal◁, Normal▷)



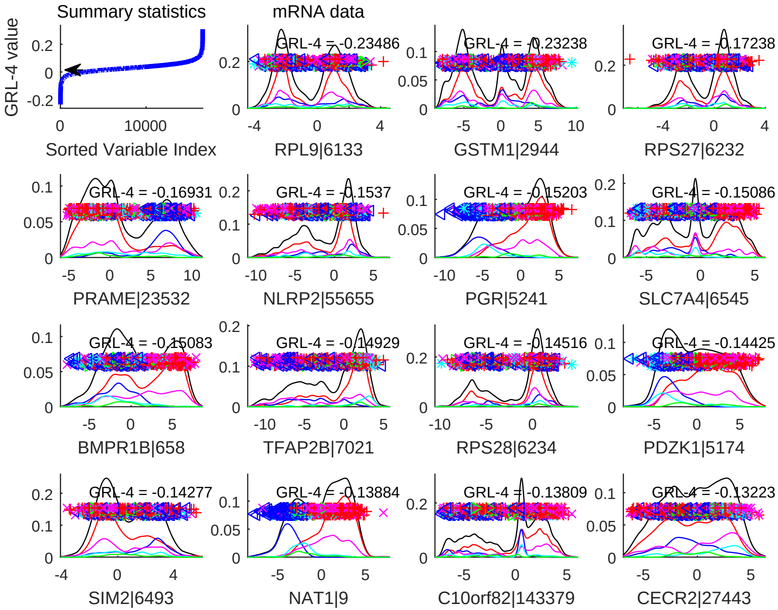
Sample kurtosis: Bottom 15 (LumA+, LumB×, Her2*, Basal◁, Normal▷)



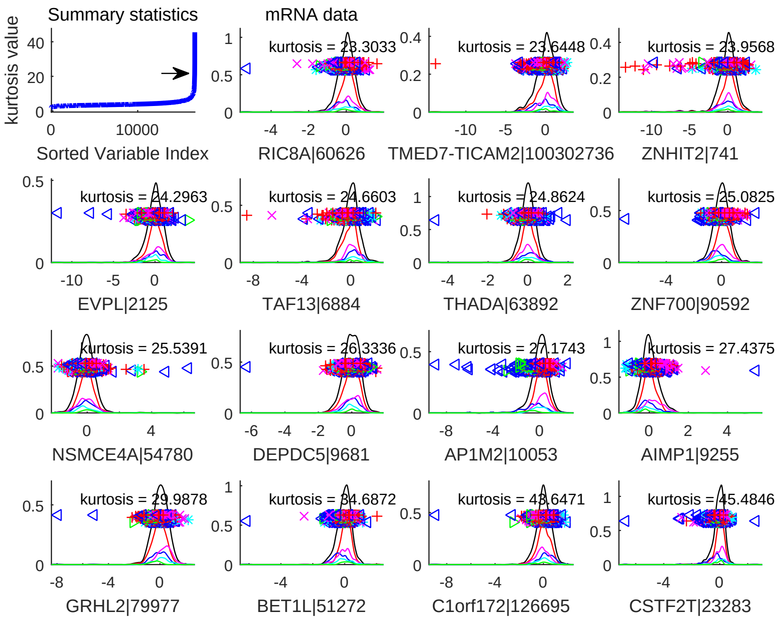


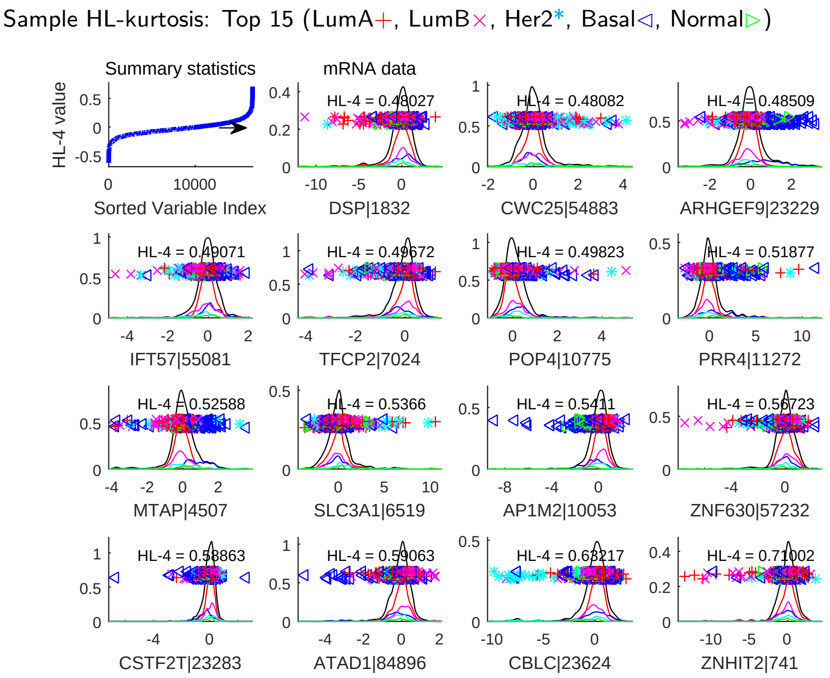
Sample HL-kurtosis: Bottom 15 (LumA+, LumB×, Her2*, Basal◁, Normal▷)
Summary statistics
mRNA data
HL-4 value
Sorted Variable Index
HL-4 = -0.6291
GSTM1|2944
HL-4 = -0.57189
RPL9|6133
HL-4 = -0.56016
PRAME|23532
HL-4 = -0.53059
SLC7A4|6545
HL-4 = -0.51659
C10orf82|143379
HL-4 = -0.49839
RPS28|6234
HL-4 = -0.47967
BMPR1B|658
HL-4 = -0.47304
TFAP2B|7021
HL-4 = -0.46874
PPP2R2C|5522
HL-4 = -0.4597
CECR2|27443
HL-4 = -0.45943
MKRN3|7681
HL-4 = -0.45708
ALOX12P2|245
HL-4 = -0.45535
NEK10|152110
HL-4 = -0.45259
PDZK1|5174
HL-4 = -0.45107
CYP2B7P1|1556

Sample GRL-kurtosis: Bottom 15 (LumA+, LumB×, Her2*, Basal◁, Normal▷)



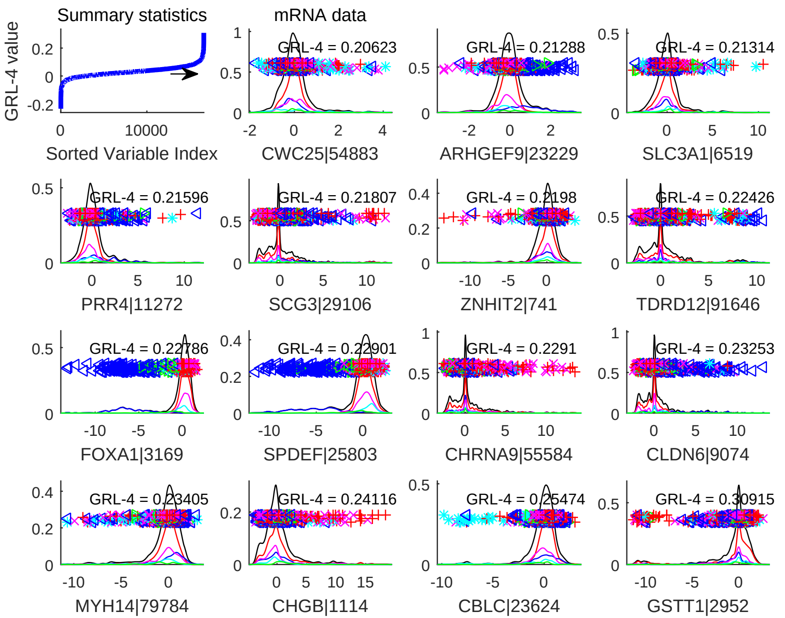
Sample kurtosis: Top 15 (LumA+, LumB×, Her2*, Basal◁, Normal▷)





Sample HL-kurtosis: Top 15 (LumA+, LumB×, Her2*, Basal◁, Normal▷)
Summary statistics
mRNA data
HL-4 value
Sorted Variable Index
HL-4 = 0.48027
DSP|1832
HL-4 = 0.48082
CWC25|54883
HL-4 = 0.48509
ARHGEF9|23229
HL-4 = 0.49071
IFT57|55081
HL-4 = 0.49672
TFCP2|7024
HL-4 = 0.49823
POP4|10775
HL-4 = 0.51877
PRR4|11272
HL-4 = 0.52588
MTAP|4507
HL-4 = 0.5366
SLC3A1|6519
HL-4 = 0.5411
AP1M2|10053
HL-4 = 0.56723
ZNF630|57232
HL-4 = 0.58863
CSTF2T|23283
HL-4 = 0.59063
ATAD1|84896
HL-4 = 0.63217
CBLC|23624
HL-4 = 0.71002
ZNHIT2|741

Sample GRL-kurtosis: Top 15 (LumA+, LumB×, Her2*, Basal◁, Normal▷)



PAM50 genes

- The paper (Tibshirani, 2002) suggested an algorithm called Prediction Analysis of Microarray (PAM) for selecting genes which might best separate different subtypes from each other.
- The PAM50 genes are the genes selected by the PAM algorithm.
- A better measure of sorting will better find the PAM50 genes out of top n genes suggested by the measure.

Precision and recall

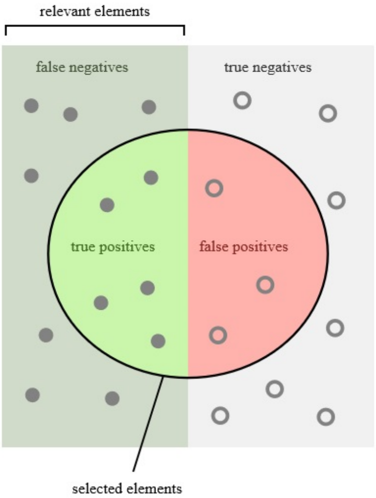


Figure 9: A pictorial description about precision and recall (Wikipedia)

- Precision: How many selected items are relevant?
- Recall: How many relevant items are selected?

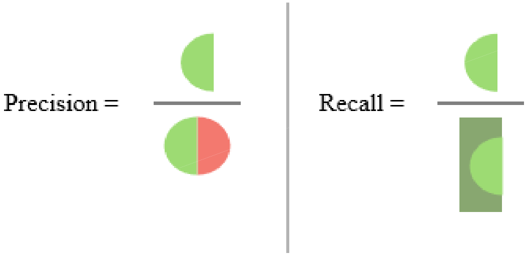


Figure 10: A pictorial description about precision and recall (Wikipedia)

- Suppose that the PAM50 genes are A, B, C.
- If top n genes suggested by a measure is

$$X_1, \cdots, X_{n_1}, A, X_{n_1+2}, \cdots, X_{n_2}, B, X_{n_2+2}, \cdots, X_{n_3}, C,$$

then we have Figure 11.

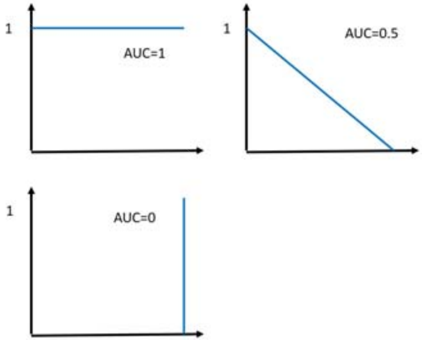


Figure 11: Examples of precision-recall curves

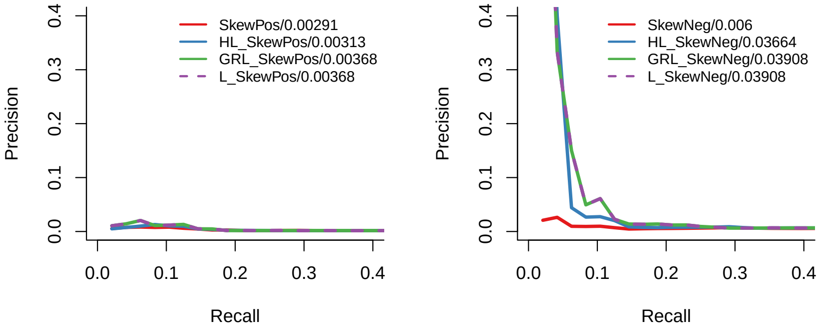


Figure 12: Precision-recall curves of ranks generated by various skewness measures

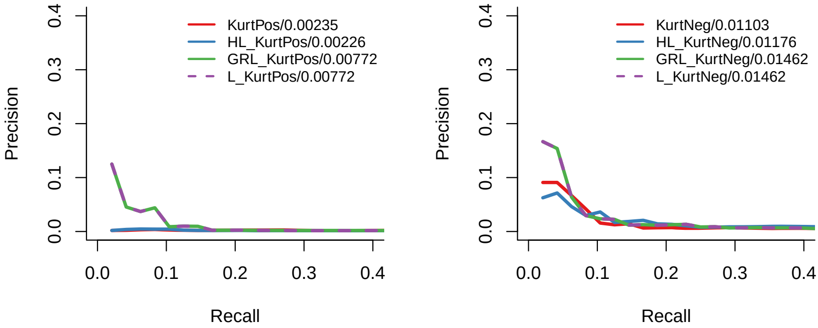


Figure 13: Precision-recall curves of ranks generated by various kurtosis measures

References

Groeneveld, R. A. (1991). An Influence Function Approach to Describing the Skewness of a Distribution. *The American Statistician*, *45(2)*, 97 – 102.

Henderson, A. R. (2006). Testing Experimental Data for Univariate Normality. *Clinica Chimica Acta*, *366*, 112 – 129.

Hosking, J. R. M. (1990). L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *Journal of the Royal Statistical Society. Series B*, *52(1)*, 105 – 124.

Huber, P. J. and Ronchetti, E. M. (2009). Robust Statistics, 2nd Edition. *Wiley, New York*.

Oja, H. (1981). On Location, Scale, Skewness and Kurtosis of Univariate Distributions. *Scandinavian Journal of Statistics*, *8*, 154 – 168.

Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics. *Wiley, New York*.

References

Ruppert, D. (1987). What is Kurtosis? An Influence Function Approach. *The American Statistician*, *41(1)*, 1 – 5.

Tibshirani, R., Hastie, R., Narasimhan B. and Chu, G. (2002). Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression. *Proceedings of the National Academy of Sciences*, *99(10)*, 6567 – 6572.