Article

Information–Theoretic Aspects of Location Parameter Estimation under Skew–Normal Settings

Javier E. Contreras-Reyes

Instituto de Estadística, Facultad de Ciencias, Universidad de Valparaíso, Valparaíso 2360102, Chile; jecontrr@uc.cl

Abstract: In several applications, the assumption of normality is often violated in data with some level of skewness, so skewness affects the mean's estimation. The class of skew–normal distributions is considered, given their flexibility for modeling data with asymmetry parameter. In this paper, we considered two location parameter (μ) estimation methods in the skew–normal setting, where the coefficient of variation and the skewness parameter are known. Specifically, the least square estimator (LSE) and the best unbiased estimator (BUE) for μ are considered. The properties for BUE (which dominates LSE) using classic theorems of information theory are explored, which provides a way to measure the uncertainty of location parameter estimations. Specifically, inequalities based on convexity property enable obtaining lower and upper bounds for differential entropy and Fisher information. Some simulations illustrate the behavior of differential entropy and Fisher information bounds.

Keywords: skew–normal distribution; location parameter; skewness; differential entropy; Fisher information; Cramér–Rao bound; convexity

1. Introduction

A typical problem in statistical inference is estimating the parameters from a data sample [1], especially if the data have some level of skewness. Therefore, the estimation of these parameters is affected by asymmetry. Recent research addressed data asymmetry with the class of skew–normal distributions, given their flexibility for modeling data with the skewness (asymmetry/symmetry) parameter [2]. In particular, Ref. [3] focused on estimating location parameter (μ), assuming that the coefficient of variation and skewness parameter are known. Specifically, they presented the least square estimator (LSE) and the best unbiased estimator (BUE) for μ. The precision of the location parameter estimation is directly influenced by skewness [4] and, hence, affects the confidence intervals and sample size [5,6].

Given that complex parametric distributions with several parameters are often considered [2], the information measures (entropies and/or divergences) play an important role in quantifying uncertainty provided by a random process about itself, and it is sufficient to study the reproduction of a marginal process through a noiseless system. One main application is related to the selection of models and detection of the number of clusters [7], or the interpretation of physical phenomena [8,9]. However, the use of entropies and/or divergences is widely considered to compare estimations [1]. For example, Ref. [10] considered the Kullback–Leibler (KL) divergence as a method to compare sample correlation matrices to an application in financial markets, assuming two multivariate normal densities. Using the estimated parameters based on maximum likelihood estimation, Refs. [11–13] considered the KL divergence for an asymptotic test to evaluate the data skewness and/or bimodality.

Given that precision was evaluated with confidence intervals in [5], the quantification of uncertainty for location parameter estimation under skew–normal settings motivated

this study. The properties for MSE and (emphasizing) BUE, using classic theorems and properties of information theory are explored, which enable measuring the uncertainty of location parameter estimations based on differential entropy and Fisher information [1]. The Cramér–Rao inequality [14] linked Fisher information with the variance of an unbiased estimator, which is considered to find a lower bound for Fisher information. In addition, considering a stochastic representation [15] of a skew–normal random variable, the convexity property of Fisher information is also used to find an upper bound for Fisher information.

This paper is organized as follows: some properties and inferential aspects based on information theory are presented in Section 2. In Section 3, the computation and description of information–theoretic theorems related to location parameter estimation of skew–normal distribution are presented. In Section 4, some simulations illustrate the usefulness of the results. Final remarks conclude the paper in Section 5.

2. Information-Theoretic Aspects

In this section, some main theorems and properties of information theory are described. Specifically, these properties are based on differential entropy and Fisher information.

Definition 1. *Let X be a random variable with support in $\mathbb{R}$ and continuous probability density function (pdf), $f(x;\theta)$, which depends on parameter θ. The differential entropy of X [1] is defined by*

$$H(X) = -\mathbb{E}[\log f(X;\theta)] = -\int_{\mathbb{R}} f(x;\theta) \log f(x;\theta)dx,$$

where notation $\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)f(x;\theta)dx$ was used.

Differential entropy depends only on the pdf of the random variable. In the following theorem, the scaling property of differential entropy is presented.

Theorem 1. *For any real constant a, the differential entropy of aX Theorem 8.6.4 of [1] is given by*

$$H(aX) = H(X) + \log|a|.$$

In particular, for two random variables, the following differential entropy bounds hold.

Theorem 2. *Let X and Y be two independent random variables. Suppose $Z \overset{D}{=} X + Y$, where "$\overset{D}{=}$" denotes equality in distribution, then*

(i)

$$\frac{H(X) + H(Y) + \log 2}{2} \leq H(Z) \leq H(X) + H(Y).$$

(ii) For any constant ρ such that $0 \leq \rho \leq 1$,

$$\rho H(X) + (1-\rho)H(Y) \leq H(\sqrt{\rho}X + \sqrt{1-\rho}Y).$$

Proof. For part (i), consider first the general case for $X_1, X_2, \ldots, X_n$ independent and identically distributed (i.i.d.) random variables see Equations (5) and (6) of [16], then

$$H(X_1) \leq H\left(\frac{X_1 + X_2}{\sqrt{2}}\right) \leq \cdots \leq H\left(\frac{1}{\sqrt{n-1}} \sum_{i=1}^{n-1} X_i\right) \leq H\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i\right).$$

Considering the latter inequality for two variables, X and Y, and the scaling property of Theorem 1, we obtain $2H(X+Y) - \log 2 \geq H(X) + H(Y)$, yielding the left side of the inequality. For the right side, see [17]. The inequality of part (ii) is proved in Theorem 7 of [14]. □

The inequality of Theorem 2 (ii) is based on the convexity property, and allows obtaining a lower bound for differential entropy Theorem 8.6.5 of [1].

Theorem 3. *Let X be a random variable with zero mean and finite σ^2, then*

$$H(X) \leq \frac{1}{2} \log(2\pi e \sigma^2),$$

and equality is achieved if, and only if, $X \sim N(0, \sigma^2)$.

Theorem 3, also known as the maximum entropy principle, implies that Gaussian distribution maximizes the differential entropy over all distributions with the same variance. This theorem has several implications for information theory, mainly when the differential entropy of an unknown distribution is hard to obtain. Thus, this upper bound is a good alternative. Another consequence is the relationship between estimation error and differential entropy, which includes the Cramér–Rao bound as described next. First, the Fisher information for continuous densities needs to be defined as follows.

Definition 2. *Let X be a random variable with support in $\mathbb{R}$ and continuous density function $f(x; \theta)$, which depends on parameter θ, so $\int_{\mathbb{R}} f(x; \boldsymbol{\theta}) dx = 1$. The Fisher information of X [1] is defined by*

$$J(X) = \mathbb{E}\left[\left\{\frac{\partial}{\partial x} \log f(x; \theta)\right\}^2\right] = \int_{\mathbb{R}} \left(\frac{\partial}{\partial x} f(x; \theta)\right)^2 \frac{1}{f(x; \theta)} \, dx. \tag{1}$$

The Fisher information is a measure of the minimum error in estimating a parameter θ of a distribution. Classical definitions of Fisher information considered differentiation with respect to θ to define $J(\theta)$; however, by considering a parametric form as $f(x - \theta; \theta)$, differentiation with respect to x is equivalent to differentiation with respect to θ as in Equation (1) [1]. The following inequality links Fisher information and variance.

Theorem 4. *Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)^\top$ be a sample of n random variables drawn i.i.d. $\sim f(x; \theta)$, the mean-squared error of an unbiased estimator $T(\mathbf{X})$ of parameter θ is lower bounded by the reciprocal of the Fisher information Theorem 11.10.1 of [1]:*

$$Var[T(\mathbf{X})] \geq \frac{1}{J(X)}, \tag{2}$$

where $J(X)$ is defined in Equation (1) and, if the inequality is achieved, $T(\mathbf{X})$ is efficient.

Theorem 4, also known as the Cramér–Rao inequality, allows determining the best estimator of θ to obtain a lower bound for Fisher information. The Cramér–Rao inequality was first planned for any estimator $T(\mathbf{X})$ (not necessarily unbiased) of θ in terms of mean-squared error, in this case

$$\mathbb{E}[\{T(\mathbf{X}) - \theta\}^2] \geq \frac{\left(1 + \frac{\partial}{\partial \theta} Bias(\theta)\right)^2}{J(X)} + Bias(\theta)^2,$$

where $Bias(\theta) = \mathbb{E}[T(\mathbf{X}) - \theta]$; see Equation (11).290 of [1]. Clearly, if $T(\mathbf{X})$ is an unbiased estimator of θ, Theorem 4 is a particular case of the latter inequality. Inequality (2) was obtained through the Cauchy–Schwarz inequality on the variance of all unbiased estimators. The following inequality, also known as the Fisher information inequality, is based on the convexity property and is useful to obtain an upper bound for Fisher information.

Theorem 5. *For any two independent random variables X and Y, and any constant ρ such that $0 \leq \rho \leq 1$, then*

$$J(\sqrt{\rho}X + \sqrt{1-\rho}Y) \leq \rho J(X) + (1-\rho)J(Y).$$

Proof. See proof of Theorem 13 in [14]. □

3. Location Parameter Estimation

The skew–normal distribution is an extension of the normal one, allowing for the presence of skewness.

Definition 3. *X is called a skew–normal random variable [15] and denoted as $X \sim SN_1(\mu, \sigma^2, \lambda)$ if it has pdf*

$$f(x; \boldsymbol{\theta}) = \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left[\lambda\left(\frac{x-\mu}{\sigma}\right)\right], \quad x \in \mathbb{R}, \quad \boldsymbol{\theta} = (\mu, \sigma^2, \lambda);$$

with location $\mu \in \mathbb{R}$, scale $\sigma^2 \in \mathbb{R}$, and shape $\lambda \in \mathbb{R}$ parameters. In addition, $\phi(x)$ is the pdf of the standardized normal distribution with 0 mean and variance 1, denoted as $N(0,1)$, and $\Phi(x)$ is the corresponding cumulative distribution function (cdf) of the standardized normal distribution.

Random variable X is represented by the following stochastic representation:

$$X \stackrel{d}{=} \mu + \sigma(\delta|U_0| + \sqrt{1-\delta^2}U), \tag{3}$$

where $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$, and U_0 and $U \sim N(0,1)$ are independently distributed; see Equation (2.14) of [15].

Additionally, X has a representation based on a link between differential entropy and Fisher information, due to *de Bruijn*'s identity. By matching the stochastic representation (3) with Equation (20) of [16], it is possible to assign $Y = \mu + \sigma\delta|U_0|$ with fixed δ. Then,

$$H(Y) = \frac{1}{2}\log(2\pi e) + \int_{-\infty}^{\infty}\lambda\left\{(1+\lambda^2)J(X) - 1\right\}d\lambda,$$

where an approximation for Fisher information $J(X)$ appears in the proof of Proposition 5 below (with $n = 1$ observation).

Definition 4. **X** *is called a multivariate skew–normal random vector [18] and denoted as* **X** $\sim$ *$SN_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ if it has pdf*

$$f_n(\mathbf{x}; \boldsymbol{\theta}) = 2\phi_n(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi[\boldsymbol{\lambda}^\top\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})], \quad x \in \mathbb{R}^n, \quad \boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}),$$

with location vector $\boldsymbol{\mu} \in \mathbb{R}^n$, scale matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$, and skewness vector $\boldsymbol{\lambda} \in \mathbb{R}^n$ parameters. In addition, $\phi_n(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the n-dimensional normal pdf with location parameter $\boldsymbol{\mu}$ and scale parameter $\boldsymbol{\Sigma}$.

Let $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim SN_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, with $\boldsymbol{\mu} = \mathbf{1}_n\mu$, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$, and $\boldsymbol{\lambda} = \mathbf{1}_n\lambda$, where $\mathbf{1}_n = (1, \ldots, 1)^\top \in \mathbb{R}^n$ and $\mathbf{I}_n$ denotes the $n \times n$-identity matrix. Following [3] and Corollary 2.2 of [5], the following properties hold.

Property 1. *$X_i \sim SN_1(\mu, \sigma^2, \lambda_*)$, $i = 1, \ldots, n$, with $\lambda_* = \frac{\lambda}{\sqrt{1+(n-1)\lambda^2}}$.*

Property 1 indicates that $X_1, \ldots, X_n$ is a random sample with identically distributed but random variables not independent from a univariate skew–normal population with location μ, scale σ^2, and shape λ_* parameters.

Property 2. $E[X_i] = \mu + \sigma b \delta_*$ and $Var[X_i] = \sigma^2(1 - nb^2\delta_*^2)$, with $b = \sqrt{\frac{2}{\pi}}$ and $\delta_* = \frac{\lambda}{1+n\lambda^2}$.

Property 3. $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim SN_1(\mu, \frac{\sigma^2}{n}, \sqrt{n}\lambda)$.

Property 4. $(\frac{n-1}{\sigma^2})S^2 \sim \chi_{n-1}^2$ with $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, where χ_{n-1}^2 denotes the chi-square distribution with $n - 1$ degrees of freedom, and sample mean $\overline{X}$ and sample variance S^2 are independent.

3.1. Least Square Estimator

Assuming that the coefficient of variation $\tau = |\sigma/\mu|$ and shape parameter λ are known, Theorem 4.1 of [3] provides the least square estimator for μ and its variance, given by

$$T_{LSE}(\mathbf{X}) = \omega\overline{X}, \tag{4}$$

$$Var[T_{LSE}(\mathbf{X})] = \omega^2(1 - n\delta_n^2)\frac{\sigma^2}{n}, \tag{5}$$

$$\omega = \frac{n}{n + \tau\left(\frac{\tau + n\delta_n}{1 + \delta_n\tau}\right)},$$

where $\delta_n = b\delta_*$, and δ_* is defined in Property 2. The least square estimator for μ was obtained by minimizing the MSE of $nc\overline{X}$ with respect to a constant c. The MSE of $T_{LSE}(\mathbf{X})$ is

$$MSE[T_{LSE}(\mathbf{X})] = \left(\frac{\sigma^2}{n} + \mu^2 + 2\mu\delta_n\sigma\right)\omega^2 - 2\mu\omega(\mu + \sigma\delta_n) + \mu^2. \tag{6}$$

Proposition 1. *Let* $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim SN_n(\mu\mathbf{1}_n, \sigma^2\mathbf{I}_n, \mathbf{1}_n\lambda)$, *with known* τ *and* λ. *Thus,*

(i)

$$H(T_{LSE}(\mathbf{X})) = \frac{1}{2}\log\left(2\pi e\frac{\sigma^2}{n}\omega^2\right) - H_N(\eta),$$
$$H_N(\eta) = \mathbb{E}[\log\{2\Phi(\eta W)\}],$$
$$W \sim SN_1(0, 1, \eta),$$
$$\eta = \sigma\lambda.$$

(ii)

$$H(T_{LSE}(\mathbf{X})) \le \frac{1}{2}\log\left(2\pi e\frac{\sigma^2}{n}\omega^2(1 - n\delta_n^2)\right).$$

Proof. Part (i) follows straightforwardly from Theorem 1, Property 3, (4) and Proposition 2.1 of [19] (for the univariate case). Part (ii) is straightforward from Theorem 3 and (5). □

Differential entropy of $T_{LSE}(\mathbf{X})$ corresponds to the difference of the normal differential entropy and a term called negentropy, $H_N(\eta)$, that depends on σ and λ parameters. Additionally, note that part (ii) yields the upper bound for $H_N(\eta)$ of part (i), $H_N(\eta) \le \frac{1}{2}\log(1 - n\delta_n^2)$.

As a particular case of Proposition 1, it is possible to obtain the differential entropy of sample mean $\overline{X}$ by choosing $\omega = 1$:

$$H(\overline{X}) = \frac{1}{2}\log\left(2\pi e\frac{\sigma^2}{n}\right) - \mathbb{E}[\log\{2\Phi(\eta W)\}]; \tag{7}$$

its respective upper bound

$$H(\overline{X}) \le \frac{1}{2}\log\left(2\pi e\frac{\sigma^2}{n}(1 - n\delta_n^2)\right);$$

and, from Equation (6), its respective MSE

$$MSE[\overline{X}] = \frac{\sigma^2}{n}.$$ (8)

3.2. Best Unbiased Estimator

Assuming that the coefficient of variation $\tau = |\sigma/\mu|$ and shape parameter λ are known, Theorem 5.1 of [3] provides the best unbiased estimator (BUE) for μ, given by

$$T_{BUE}(\mathbf{X}) = (1 - \alpha)d_1(\mathbf{X}) + \alpha d_2(\mathbf{X}),$$ (9)

$$d_1(\mathbf{X}) = \frac{\overline{X}}{1 + \delta_n \tau},$$

$$d_2(\mathbf{X}) = c_n \sqrt{n-1} S,$$

$$c_n = \frac{1}{\sqrt{2\tau^2}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)},$$ (10)

$$\alpha = \frac{1}{(1 + \delta_n \tau)[(n-1)c_n]^2},$$ (11)

where $\Gamma(x)$ denotes the usual gamma function and S is defined in Property 4.

Remark 1. *Equation (10) can be approximated using an asymptotic expression for the gamma function given by $\Gamma(x + a) \approx \sqrt{2\pi} x^{x+a-1/2} e^{-x}$, $a < \infty$, as $|x| \to \infty$ [19]. Then,*

$$c_n \approx \frac{1}{\sqrt{n\tau^2}},$$ (12)

as $n \to \infty$. Since the exact form (10) can be undefined for large samples ($n > 200$), approximation (1) is very useful for these cases. Note that from (11) and (1), $\delta_n, \alpha \to 0$ as $n \to \infty$, which implies that estimator $T_{BUE}(\mathbf{X})$ is only influenced by $d_1(\mathbf{X})$ for large samples.

From Properties 3 and 4, Ref. [3] also proved that

$$d_1(\mathbf{X}) \sim SN_1\left(\frac{\mu}{1 + \delta_n \tau}, \frac{\sigma^2}{n(1 + \delta_n \tau)^2}, \sqrt{n}\lambda\right),$$ (13)

$$Var[d_1(\mathbf{X})] = \frac{(\mu\tau)^2(1 - n\delta_n^2)}{n(1 + \delta_n \tau)^2},$$ (14)

$$Var[d_2(\mathbf{X})] = 2(\mu\tau)^2\left(1 - \frac{1}{2(n-1)(\tau c_n)^2}\right).$$ (15)

Given that $Cov(d_1, d_2) = 0$, from Equations (9), (14) and (15), we obtain

$$Var[T_{BUE}(\mathbf{X})] = (1 - \alpha)^2 Var[d_1(\mathbf{X})] + \alpha^2 Var[d_2(\mathbf{X})].$$ (16)

The following proposition provides two upper bounds of differential entropy for $T_{BUE}(\mathbf{X})$ based on Theorem 3.

Proposition 2. *Let $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim SN_n(\mu\mathbf{1}_n, \sigma^2\mathbf{I}_n, \mathbf{1}_n\lambda)$, with known τ and λ. Thus,*

(i)

$$H(T_{BUE}(\mathbf{X})) \leq \frac{1}{2}\log\left(\frac{4\pi e}{n}\left(\frac{\sigma^2\alpha(1-\alpha)}{1 + \delta_n \tau}\right)^2\left(1 - \frac{1}{2(n-1)(\tau c_n)^2}\right)\right),$$

(ii)

$$H(T_{BUE}(\mathbf{X})) \leq \frac{1}{2}\log(2\pi e Var[T_{BUE}(\mathbf{X})]\}).$$

Proof. From Theorem 3 and Equations (14) and (15), the differential entropies of $d_1(\mathbf{X})$ and $d_2(\mathbf{X})$ are, respectively, upper bounded by

$$H(d_1(\mathbf{X})) \leq \frac{1}{2}\log\left(2\pi e \frac{\sigma^2}{n(1+\delta_n\tau)^2}\right), \tag{17}$$

$$H(d_2(\mathbf{X})) \leq \frac{1}{2}\log\left(2\pi e 2\sigma^2\left(1-\frac{1}{2(n-1)(\tau c_n)^2}\right)\right). \tag{18}$$

Considering the right side on the inequality of Theorem 2(i), with $Z = X + Y$, $X = (1-\alpha)d_1(\mathbf{X})$ and $Y = \alpha d_2(\mathbf{X})$ (thus $Cov(X,Y) = 0$), we obtain

$$H(T_{BUE}(\mathbf{X})) \leq H((1-\alpha)d_1(\mathbf{X})) + H(\alpha d_2(\mathbf{X})) = H(d_1(\mathbf{X})) + H(d_2(\mathbf{X})) + \log|(1-\alpha)\alpha|,$$

where Theorem 1 is applied later. Then, Equations (17) and (18) yield part (i). On the other hand, by considering directly Theorem 3 on $T_{BUE}(\mathbf{X})$, Equation (16) implies part (ii). □

The following proposition provides two lower bounds of differential entropy for $T_{BUE}(\mathbf{X})$.

Proposition 3. *Let* $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim SN_n(\mu\mathbf{1}_n, \sigma^2\mathbf{I}_n, \mathbf{1}_n\lambda)$, *with known* τ $(0 < \tau < 1)$ *and* λ. *Thus*

(i)

$$\frac{H(d_1(\mathbf{X})) + H(d_2(\mathbf{X})) + \log(2\alpha^2(1-\alpha)^2)}{2} \leq H(T_{BUE}(\mathbf{X})),$$

(ii)

$$(1-\alpha^2)H(d_1(\mathbf{X})) + \alpha^2 H(d_2(\mathbf{X})) \leq H(T_{BUE}(\mathbf{X}));$$

where

$$H(d_1(\mathbf{X})) = \frac{1}{2}\log\left(\frac{2\pi e\sigma^2}{n(1+\delta_n\tau)^2}\right) - H_N(\eta_1),$$

$$H_N(\eta_1) = \mathbb{E}[\log\{2\Phi(\eta_1 W_1)\}],$$

$$W_1 \sim SN_1(0,1,\eta_1),$$

$$\eta_1 = \left|\frac{\lambda\sigma}{1+\delta_n\tau}\right|;$$

and

$$H(d_2(\mathbf{X})) = \log\left(\frac{|\alpha|\left[\Gamma\left(\frac{n-1}{2}\right)\right]^n}{2\left[\tau\Gamma\left(\frac{n}{2}\right)\right]^{n-1}}\right) - \frac{(n-2)\Gamma\left(\frac{n-1}{2}\right)}{4(2c_n^2)^{-\left(\frac{n-1}{2}\right)}}\left(\psi\left(\frac{n-1}{2}\right) + \log(2c_n^2)\right)$$

$$+ \frac{\Gamma(n+1)}{c_n^2}\frac{\left[\Gamma\left(\frac{n-1}{2}\right)\right]^{n+2}}{\left[\tau\Gamma\left(\frac{n}{2}\right)\right]^{n+3}}.$$

Proof. Differential entropy of $d_1(\mathbf{X})$ is straightforward from evaluating (13) on Proposition 2.1 of [19] (for the univariate case). Given that distribution of $d_2(\mathbf{X})$ is unknown, Ref. [3] provided its pdf

$$f_{d_2}(x; \mu, \sigma) = 2\frac{\left[\tau\Gamma\left(\frac{n}{2}\right)\right]^{n-1}}{\left[\Gamma\left(\frac{n-1}{2}\right)\right]^n}x^{n-2}e^{-\frac{x^2}{2c_n^2}}. \tag{19}$$

Through Equations (19) and (3.381.4) of [20], the moments of d_2 are given by

$$\mathbb{E}_{d_2}[X^m] = 2\Gamma(m+n-1)\frac{\left[\Gamma\left(\frac{n-1}{2}\right)\right]^{2m+n-2}}{\left[\tau\Gamma\left(\frac{n}{2}\right)\right]^{2m+n-1}}, \quad m = 0, 1, \ldots; \tag{20}$$

and, using Equation (4.352.1) of [20], the moment of $\log x$ is

$$
\begin{aligned}
\mathbb{E}_{d_2}[\log X] &= \int_0^\infty f_{d_2}(x; \mu, \sigma) \log x dx, \\
&= 2\frac{\left[\tau\Gamma\left(\frac{n}{2}\right)\right]^{n-1}}{\left[\Gamma\left(\frac{n-1}{2}\right)\right]^n} \int_0^\infty x^{n-2} e^{-\frac{x^2}{2c_n^2}} \log x dx, \\
&= \frac{(2c_n^2)^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}{4}\left(\psi\left(\frac{n-1}{2}\right) + \log(2c_n^2)\right), \tag{21}
\end{aligned}
$$

where $\psi(x) = \frac{d}{dx}\log\{\Gamma(x)\}$ is the digamma function. Therefore, by definition (1), the differential entropy of $d_2(\mathbf{X})$ is computed as

$$
\begin{aligned}
H(d_2(\mathbf{X})) &= -\int_0^\infty f_{d_2}(x; \mu, \sigma) \log f_{d_2}(x; \mu, \sigma) dx, \\
&= -\log\left(\frac{2\left[\tau\Gamma\left(\frac{n}{2}\right)\right]^{n-1}}{|\alpha|\left[\Gamma\left(\frac{n-1}{2}\right)\right]^n}\right) - (n-2)\underbrace{\int_0^\infty f_{d_2}(x; \mu, \sigma) \log x dx}_{\mathbb{E}_{d_2}[\log X]} + \frac{1}{2c_n^2}\underbrace{\int_0^\infty f_{d_2}(x; \mu, \sigma)x^2 dx}_{\mathbb{E}_{d_2}[X^2]}.
\end{aligned}
$$

Thus, Equations (20) and (21) are evaluated in the latter expression to obtain $H(d_2(\mathbf{X}))$. By assuming $Z = X + Y$ in Theorem 2(i) ($Cov(X, Y) = 0$), with $X = (1-\alpha)d_1(\mathbf{X})$ and $Y = \alpha d_2(\mathbf{X})$, the inequality of part (i) is obtained.

By assuming $Z = X + Y$ in Theorem 2(ii), with $X = d_2(\mathbf{X})$, $Y = d_1(\mathbf{X})$ (thus $Cov(X, Y) = 0$), and $\rho = \alpha^2$, and since $d_1(\mathbf{X})$ and $d_2(\mathbf{X})$ are two unbiased estimators of μ [3], the inequality of part (ii) is obtained. □

The following proposition provides a lower bound for Fisher information of parameter μ based on $T_{BUE}(\mathbf{X})$.

Proposition 4. *Let* $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim SN_n(\mu\mathbf{1}_n, \sigma^2\mathbf{I}_n, \mathbf{1}_n\lambda)$, *with known* τ *and* λ. *Thus,*

$$J(\mu) \geq \left[(1-\alpha)^2\frac{\sigma^2(1-n\delta_n^2)}{n(1+\delta_n\tau)^2} + \alpha^2 2\sigma^2\left(1 - \frac{1}{2(n-1)(\tau c_n)^2}\right)\right]^{-1}.$$

Proof. Considering that $T_{BUE}(\mathbf{X})$ is an unbiased estimator of μ, from the Crámer–Rao inequality of Theorem 4 and Equations (14)–(16), we obtain

$$J(\mu) \geq \frac{1}{(1-\alpha)^2 Var[d_1(\mathbf{X})] + \alpha^2 Var[d_2(\mathbf{X})]},$$

yielding the result. □

The following Proposition provides an upper bound of Fisher information for parameter μ based on $T_{BUE}(\mathbf{X})$ and the convexity property.

Proposition 5. *Let* $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim SN_n(\mu\mathbf{1}_n, \sigma^2\mathbf{I}_n, \mathbf{1}_n\lambda)$, *with known* τ ($0 < \tau < 1$) *and* λ. *Thus*

$$J(\mu) \leq (1-\alpha^2)J(d_1(\mathbf{X})) + \alpha^2 J(d_2(\mathbf{X})),$$

where

$$J(d_1(\mathbf{X})) \approx 1 + \frac{n(b\lambda)^2}{\sqrt{1 + 2nb^4\lambda^2}},$$

$$J(d_2(\mathbf{X})) = \frac{2(n-2)(n^2-2n-2)\Gamma(n-2)}{\left(\frac{1}{2c_n^2}\right)^{n-3}} \frac{\left[\tau\Gamma\left(\frac{n}{2}\right)\right]^{n-1}}{\left[\Gamma\left(\frac{n-1}{2}\right)\right]^n} + \frac{1}{c_n^4},$$

Proof. By assuming $Z = X + Y$ in Theorem 5, with $X = d_2(\mathbf{X})$, $Y = d_1(\mathbf{X})$ (thus $Cov(X, Y) = 0$), and $\rho = \alpha^2$, and since $d_1(\mathbf{X})$ and $d_2(\mathbf{X})$ are two unbiased estimators of μ [3], we obtain $J(\mu) \leq \alpha^2 J(d_2(\mathbf{X})) + (1 - \alpha^2)J(d_1(\mathbf{X}))$. Note that condition $0 < \tau < 1$ ensures that $0 \leq \alpha^2 \leq 1$.

For $J(d_1(\mathbf{X}))$, the steps of Section 3.2 of [9] were considered. By Equations (1) and (13), and the change of variable $z = (x - \mu^*)/\sigma^*$, with $\mu^* = \mu/(1 + \delta_n\tau)$, $\sigma^* = \sigma/\sqrt{n(1 + \delta_n\tau)^2}$ and $\lambda^* = \sqrt{n}\lambda$, $J(d_1(\mathbf{X}))$ can be computed as

$$
\begin{aligned}
J(d_1(\mathbf{X})) &= \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial x} f(x;\theta)\right)^2 \frac{1}{f(x;\theta)} dx \\
&= \int_{-\infty}^{\infty} f(z;\lambda^*)[-z + \lambda^*\zeta(\lambda^*z)]^2 dz \\
&= \int_{-\infty}^{\infty} z^2 f(z;\lambda^*) dz - 2\lambda^* \int_{-\infty}^{\infty} z\zeta(\lambda^*z)f(z;\lambda^*)dz + [\lambda^*]^2 \int_{-\infty}^{\infty} \zeta(\lambda^*z)^2 f(z;\lambda^*)dz \\
&= \int_{-\infty}^{\infty} z^2 f(z;\lambda^*) dz - 4\lambda^* \int_{-\infty}^{\infty} z\phi(\lambda^*z)\phi(z)dz + 2[\lambda^*]^2 \int_{-\infty}^{\infty} \phi(z)\frac{\phi(\lambda^*z)^2}{\Phi(\lambda^*z)}dz,
\end{aligned}
\tag{22}
$$

where $\zeta(x) = \phi(x)/\Phi(x)$ is the *zeta* function. From Equation (22), the first and second terms are the second moment of a standardized skew–normal random variable ($\mathbb{E}[Z^2] = 1$) and the first moment of a standardized normal random variable ($\mathbb{E}[R] = 0$, $R \sim N(0,1)$), respectively. The third term is

$$
\begin{aligned}
\int_{-\infty}^{\infty} \phi(z)\frac{\phi(\lambda^*z)^2}{\Phi(\lambda^*z)}dz &= \int_0^{\infty} \phi(z)\frac{\phi(\lambda^*z)^2}{\Phi(\lambda^*z)}dz + \int_0^{\infty} \phi(z)\frac{\phi(\lambda^*z)^2}{1 - \Phi(\lambda^*z)}dz \\
&= \int_0^{\infty} \phi(z)\frac{\phi(\lambda^*z)^2}{\Phi(\lambda^*z)[1 - \Phi(\lambda^*z)]}dz
\end{aligned}
$$

The following approximation of normal densities (see p. 83 of [15]),

$$\frac{\phi(y)}{\sqrt{\Phi(y)[1 - \Phi(y)]}} \approx \pi b^2 \phi(b^2 y), \quad \forall y \in \mathbb{R},$$

and some basic algebraic operations of normal densities are useful to approximate the third term of Equation (22) as

$$
\begin{aligned}
\int_0^{\infty} \phi(z)\frac{\phi(\lambda^*z)^2}{\Phi(\lambda^*z)[1 - \Phi(\lambda^*z)]}dz &\approx \pi^2 b^4 \int_0^{\infty} \phi(z)\phi(b^2\lambda^*z)^2 dz \\
&= \frac{\pi^2 b^4}{2\pi\sqrt{1 + 2b^4[\lambda^*]^2}} \int_0^{\infty} \phi(z; 0, \{1 + 2b^4[\lambda^*]^2\}^{-1})dz \\
&= \frac{\pi b^4}{4\sqrt{1 + 2b^4[\lambda^*]^2}}.
\end{aligned}
$$

Given that $\frac{\pi b^4}{2} = b^2$ and $\lambda^* = \sqrt{n}\lambda$, we obtain

$$J(d_1(\mathbf{X})) \approx 1 + \frac{n(b\lambda)^2}{\sqrt{1 + 2nb^4\lambda^2}}.$$

Using Equation (19), $J(d_2(\mathbf{X}))$ is computed as

$$
\begin{aligned}
J(d_2(\mathbf{X})) &= \int_0^\infty \left(\frac{\partial}{\partial x} f_{d_2}(x; \mu, \sigma)\right)^2 \frac{1}{f_{d_2}(x; \mu, \sigma)} dx \\
&= \int_0^\infty f_{d_2}(x; \mu, \sigma)[(n-2)x^{-1} - \frac{1}{c_n^2}]^2 dx \\
&= (n-2)^2 \int_0^\infty x^{-2} f_{d_2}(x; \mu, \sigma) dx - 2\frac{(n-2)}{c_n^2} \int_0^\infty x^{-1} f_{d_2}(x; \mu, \sigma) dx + \frac{1}{c_n^4} \\
&= 2(n-2)^2 \frac{\left[\tau\Gamma\left(\frac{n}{2}\right)\right]^{n-1}}{\left[\Gamma\left(\frac{n-1}{2}\right)\right]^n} \underbrace{\int_0^\infty x^{n-4} e^{-\frac{1}{2c_n^2}x^2} dx}_{M_1} - \frac{4(n-2)}{c_n^2} \frac{\left[\tau\Gamma\left(\frac{n}{2}\right)\right]^{n-1}}{\left[\Gamma\left(\frac{n-1}{2}\right)\right]^n} \underbrace{\int_0^\infty x^{n-3} e^{-\frac{1}{2c_n^2}x^2} dx}_{M_2} + \frac{1}{c_n^4} \\
&= \frac{2(n-2)(n^2 - 2n - 2)\Gamma(n-2)}{\left(\frac{1}{2c_n^2}\right)^{n-3}} \frac{\left[\tau\Gamma\left(\frac{n}{2}\right)\right]^{n-1}}{\left[\Gamma\left(\frac{n-1}{2}\right)\right]^n} + \frac{1}{c_n^4},
\end{aligned}
$$

where Equation (3.381.4) of [20] is applied to solve integrals M_1 and M_2. $\quad\square$

Remark 2. *Considering the same argument as in Remark (1), it can be noted that inequalities of Propositions 4 and 5 are only affected by $J(d_1(\mathbf{X}))$, i.e., for large samples, we obtain*

$$\frac{1}{Var[d_1(\mathbf{X})]} \leq J(\mu) \leq J(d_1(\mathbf{X})).$$

4. Simulations

All location parameter estimators, variances, Fisher information and differential entropies were calculated with R software [21]. Samples based on skew–normal random variables were drawn based on stochastic representation (3) and with the `rsn` function of `sn` package. All R codes used in this paper are available upon request from the corresponding author.

In general, τ takes a value between 0 and 1. If τ is close to 0, the sample has low variability, and if it is close to 1, the sample has high variability and mean loss reliability. For example, if $\tau > 0.3$, the mean is less representative of the sample. Sometimes, if μ is close to zero, τ takes high values (high variability) and could exceed unity. Therefore, for illustrative purposes, in all simulations, a coefficient of variation set of $\tau = 0.1, \dots, 1$ is considered. Additionally considered are positive asymmetry parameters $\lambda = 0.1, \dots, 5$, sample sizes $n = 10$ and 250, and theoretical location parameters $\mu = 0.1, 0.5$ and 1. For the computation of information measures, σ is replaced by $\tau|\mu|$ and location parameters μ are evaluated by their respective estimators.

The MSE of $T_{LSE}(\mathbf{X})$ is given in Equation (6), and MSE of $T_{BUE}(\mathbf{X})$ is the variance of the (unbiased) estimator (see Equation (16)). Without loss of generality, $\tau = 1$ is considered in Figure 1 because the same pattern is repeated for values of τ between 0 and 1. Comparing the MSE of both estimators, Figure 1 shows for all cases that differences between MSEs tend to increase for large values of μ, and MSEs turn around a specific value when the sample size increases. Moreover, MSEs of the unbiased estimator are less than those obtained by LSE, i.e., BUE dominates LSE Equation (11.263) of [1]. Therefore, the analysis focuses on BUE in the next section.
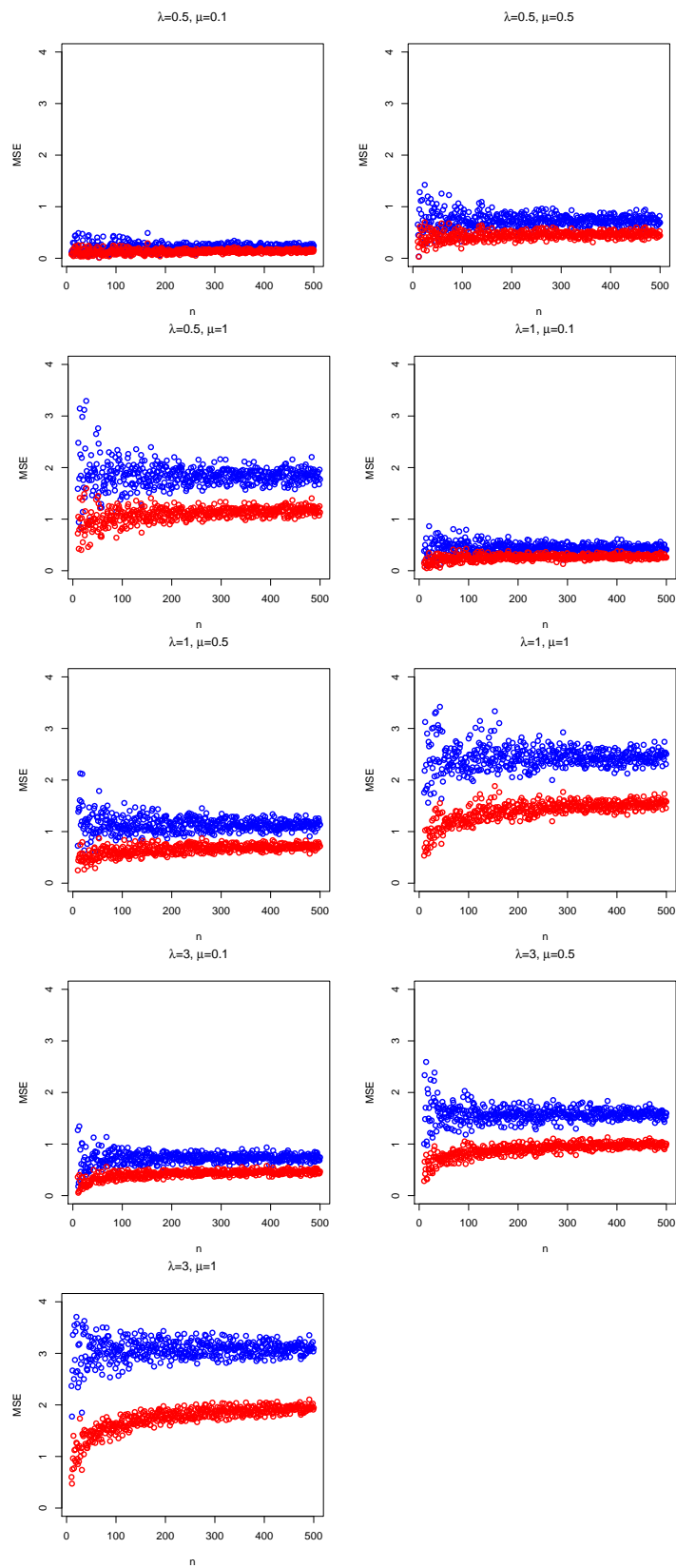
Figure 1. Mean square errors (MSE) for $T_{LSE}(\mathbf{X})$ [blue dots] and $T_{BUE}(\mathbf{X})$ [red dots] considering $\tau = 1$ and several skewness λ and location μ parameters in the simulations.

Behavior of differential entropy bounds given in Propositions 2 and 3 is illustrated in Figure 2 as 3D plots. Without loss of generality, $\tau \in (0,1]$ is considered in Figure 2 because the same pattern is repeated for values of $\tau > 1$, i.e., entropies keep increasing. The upper bound corresponds to the minimum value between bounds given in Proposition 2(i) and (ii), which is the one given in (ii). Thus, the upper bound of $H(T_{BUE}(\mathbf{X}))$ is determined by the variance (or MSE) of the estimator. In contrast, the lower bounds correspond to the maximum value between bounds given in Proposition 3 (i) and (ii) [17].
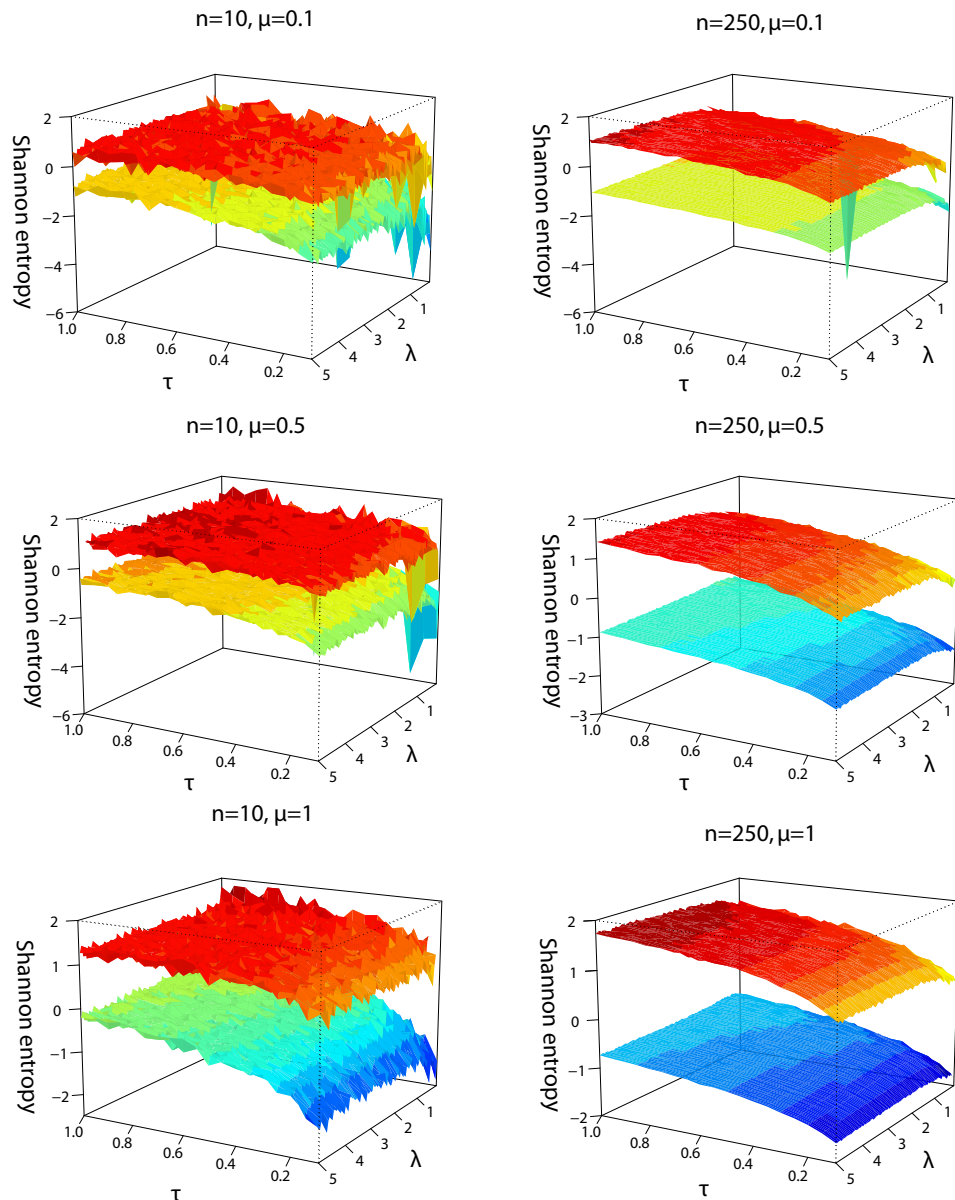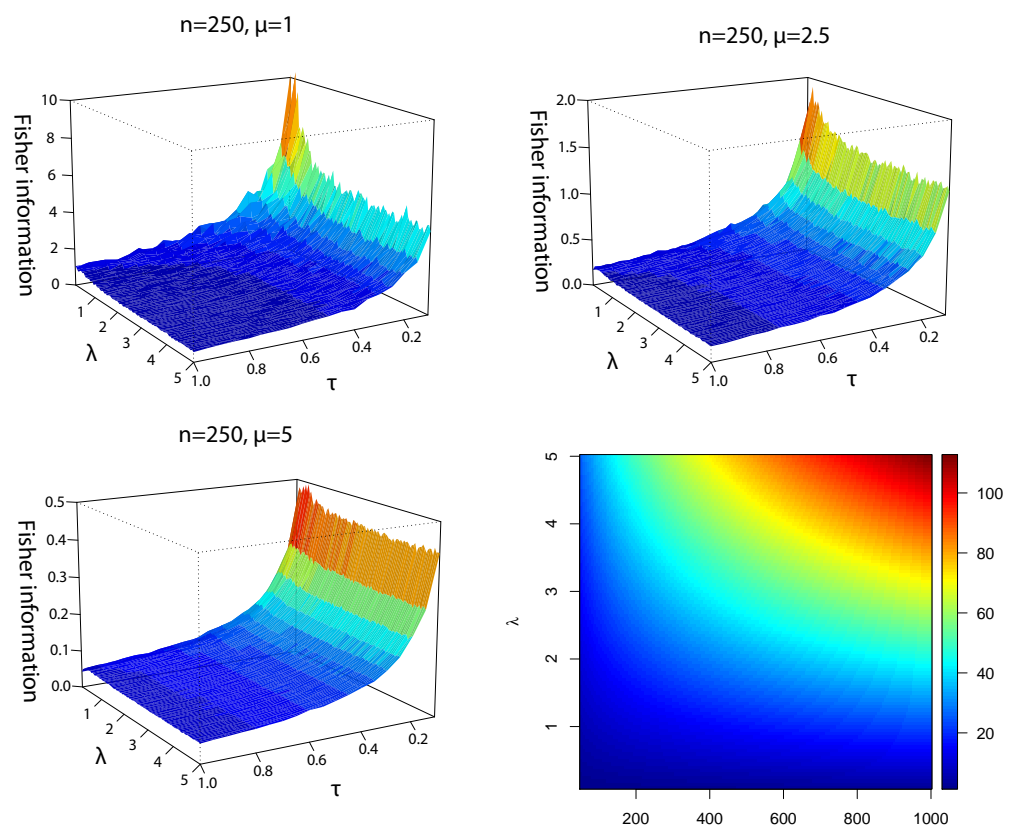

Figure 2. Differential entropy bounds for $T_{BUE}(\mathbf{X})$ considering $n = 10$ and 250, $\mu = 0.1$, 0.5 and 1; and several skewness λ and coefficient of variation τ parameters in the simulations.

Sample sizes ($n = 250$) imply that $\alpha \approx 0$ and lower bounds only depend on $H(d_1(\mathbf{X}))$. For small sample sizes ($n = 10$), α could be an intermediate value of the interval $(0,1)$, thus, lower bounds depend on $H(d_1(\mathbf{X}))$ and $H(d_2(\mathbf{X}))$. For $n = 10$, the surfaces are rough, given the randomness of bounds produced by the small sample. When $\lambda \approx 0$ (symmetry condition), the bounds decay to negative values. This is analogous to considering the skew–normal density as a non-stationary process [15], when λ is near zero, so the Hurst exponent decreases abruptly [8]. On the other hand, for $n = 250$, surfaces are soft and

bounds increase slightly for large λ. For all cases, information increases when τ tends to 1 because it produces more variability in samples.

For practical purposes, the average between bounds can be considered to provide an approximation of differential entropy [7] in similar form to average lengths of the confidence interval [3]. Given that all lower bounds of differential entropy depend on the entropy of $d_1(\mathbf{X})$, which depends on variance and sample size, they could take negative values and tend to zero when τ tends to 1. Therefore, the difference between lower and upper bounds could increase and turn out an inadequate approximation if the lower bound is negative. For the latter reason, the Fisher information considers only positive values, as studied next.

The Fisher information bounds given in Propositions 4 and 5 are illustrated in Figure 3 as 3D and 2D plots, respectively. As in the differential entropy case, and without loss of generality, $\tau \in (0, 1]$ is considered in Figure 3 because the same pattern is repeated for values of $\tau > 1$, i.e., entropies keep decreasing. Following the Cramér–Rao theorem, the variance of BUE corresponds to the reciprocal of the Fisher information. In contrast, the lower bound corresponds to a combination of the Fisher information of $d_1(\mathbf{X})$ and $d_2(\mathbf{X})$.

Figure 3. Fisher information lower bounds for $T_{BUE}(\mathbf{X})$ considering $n = 250$, $\mu = 1, 2.5$ and 5; and several skewness λ and coefficient of variation τ parameters in the simulations. The fourth panel shows the upper bounds for $T_{BUE}(\mathbf{X})$ considering $n = 100, \ldots, 1000$ and several skewness parameters λ.

As in the differential entropy case, large sample sizes ($n = 250$) imply that $\alpha \approx 0$ and lower bounds only depend on $J(d_1(\mathbf{X}))$, as mentioned in Remark 2. For small sample sizes ($n = 10$), α could be an intermediate value of the interval $(0, 1)$, thus lower bounds depend on $J(d_1(\mathbf{X}))$ and $J(d_2(\mathbf{X}))$. When $\tau \approx 0$ (low variability condition), the lower bounds take the highest values. This reciprocal relationship is determined by the Cramér–Rao theorem: more variability, less Fisher information. In addition, the 2D plot shows that the smallest upper bounds of $J(T_{BUE}(\mathbf{X}))$ are produced when $\lambda \approx 0$ [9]. Given that upper bounds do not depend on τ and μ because the skew–normal densities are standardized,

these measures are illustrated with respect to n and λ. In addition, when λ and n increase simultaneously, the upper bounds of Fisher information take the largest values.

5. Concluding Remarks

In this paper, some properties of the best unbiased estimator proposed by [3] were presented, using classic theorems of information theory, which provide a way to measure the uncertainty of location parameter estimations. Given that BUE dominates LSE, this paper focused on this estimator. Inequalities based on differential entropy and Fisher information allowed obtaining lower and upper bounds for these measures. Some simulations illustrated the behavior of differential entropy and Fisher information bounds.

Classical theorems of information theory considered the obtained additional properties of unbiased location parameter estimators. However, these theorems could be applied to other estimators, such as Bayesian [22] (as long as the prior pdf density is known), shrinkage [23], or bootstrap-based [24] ones. The assumption of the sample that came from a multivariate skew–normal distribution is too strong and not always applicable in the real world, so the properties revised here could be extended to more complex densities, for example, those that assess bimodality and heavy tails in data [7,11,13,19]. On the other hand, and given that Fisher information bounds under skew–normal settings were considered in this study, further work could focus on developing time-dependent Fisher information for skew–normal density [25], which could be applied to real data in survival analysis.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares that there is no conflict of interest in the publication of this paper.

References

1. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley & Son, Inc.: New York, NY, USA, 2006.
2. Adcock, C.; Azzalini, A. A selective overview of skew-elliptical and related distributions and of their applications. *Symmetry* **2020**, *12*, 118. [CrossRef]
3. Wang, Z.Y.; Wang, C.; Wang, T.H. Estimation of location parameter in the skew normal setting with known coefficient of variation and skewness. *Int. J. Intel. Technol. Appl. Stat.* **2016**, *9*, 191–208.
4. Trafimow, D.; Wang, T.; Wang, C. From a sampling precision perspective, skewness is a friend and not an enemy! *Educ. Psychol. Meas.* **2019**, *79*, 129–150. [CrossRef] [PubMed]
5. Wang, C.; Wang, T.; Trafimow, D.; Myüz, H.A. Necessary sample sizes for specified closeness and confidence of matched data under the skew normal setting. *Comm. Stat. Simul. Comput.* 2019, *in press*. [CrossRef]
6. Wang, C.; Wang, T.; Trafimow, D.; Talordphop, K. Estimating the location parameter under skew normal settings: Is violating the independence assumption good or bad? *Soft Comput.* **2021**, *25*, 7795–7802. [CrossRef]
7. Abid, S.H.; Quaez, U.J.; Contreras-Reyes, J.E. An information-theoretic approach for multivariate skew-*t* distributions and applications. *Mathematics* **2021**, *9*, 146. [CrossRef]
8. Contreras-Reyes, J.E. Analyzing fish condition factor index through skew-gaussian information theory quantifiers. *Fluct. Noise Lett.* **2016**, *15*, 1650013. [CrossRef]
9. Contreras-Reyes, J.E. Fisher information and uncertainty principle for skew-gaussian random variables. *Fluct. Noise Lett.* **2021**, *20*, 21500395. [CrossRef]
10. Tumminello, M.; Lillo, F.; Mantegna, R.N. Correlation, hierarchies, and networks in financial markets. *J. Econ. Behav. Organ.* **2010**, *75*, 40–58. [CrossRef]
11. Contreras-Reyes, J.E. An asymptotic test for bimodality using the Kullback–Leibler divergence. *Symmetry* **2020**, *12*, 1013. [CrossRef]
12. Contreras-Reyes, J.E.; Kahrari, F.; Cortés, D.D. On the modified skew-normal-Cauchy distribution: Properties, inference and applications. *Comm. Stat. Theor. Meth.* **2021**, *50*, 3615–3631. [CrossRef]

13. Contreras-Reyes, J.E.; Maleki, M.; Cortés, D.D. Skew-Reflected-Gompertz information quantifiers with application to sea surface temperature records. *Mathematics* **2019**, *7*, 403. [CrossRef]

14. Dembo, A.; Cover, T.M.; Thomas, J.A. Information theoretic inequalities. *IEEE Trans. Infor. Theor.* **1991**, *37*, 1501–1518. [CrossRef]

15. Azzalini, A. *The Skew-Normal and Related Families*; Cambridge University Press: Cambridge, UK, 2013; Volume 3.

16. Madiman, M.; Barron, A. Generalized entropy power inequalities and monotonicity properties of information. *IEEE Trans. Infor. Theor.* **2007**, *53*, 2317–2329. [CrossRef]

17. Xie, Y. Sum of Two Independent Random Variables. ECE587, Information Theory. 2012. Available online: https://www2.isye.gatech.edu/~yxie77/ece587/SumRV.pdf (accessed on 15 January 2022).

18. Azzalini, A.; Dalla-Valle, A. The multivariate skew-normal distribution. *Biometrika* **1996**, *83*, 715–726. [CrossRef]

19. Contreras-Reyes, J.E. Asymptotic form of the Kullback–Leibler divergence for multivariate asymmetric heavy-tailed distributions. *Phys. A Stat. Mech. Its Appl.* **2014**, *395*, 200–208. [CrossRef]

20. Gradshteyn, I.S.; Ryzhik, I.M. *Table of Integrals, Series, and Products*, 7th ed.; Academic Press: Cambridge, MA, USA; Elsevier: London, UK, 2007.

21. R Core Team. *A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021. Available online: http://www.R-project.org (accessed on 15 January 2022).

22. Bayes, C.L.; Branco, M.D. Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Braz. J. Prob. Stat.* **2007**, *21*, 141–163.

23. Kubokawa, T.; Strawderman, W.E.; Yuasa, R. Shrinkage estimation of location parameters in a multivariate skew-normal distribution. *Comm. Stat. Theor. Meth.* **2020**, *49*, 2008–2024. [CrossRef]

24. Ye, R.; Fang, B.; Wang, Z.; Luo, K.; Ge, W. Bootstrap inference on the Behrens–Fisher-type problem for the skew-normal population under dependent samples. *Comm. Stat. Theor. Meth.* 2021, *in press*. [CrossRef]

25. Kharazmi, O.; Asadi, M. On the time-dependent Fisher information of a density function. *Braz. J. Prob. Stat.* **2018**, *32*, 795–814. [CrossRef]