Why Cannot We Have a Strongly Consistent Family of Skew Normal (and Higher Order) Distributions

Thongchai Dumrongpokaphan and Vladik Kreinovich

Abstract In many practical situations, the only information that we have about the probability distribution is its first few moments. Since many statistical techniques requires us to select a single distribution, it is therefore desirable to select, out of all possible distributions with these moments, a single "most representative" one. When we know the first two moments, a natural idea is to select a normal distribution. This selection is *strongly consistent* in the sense that if a random variable is a sum of several independent ones, then selecting normal distribution for all of the terms in the sum leads to a similar normal distribution for the sum. In situations when we know three moments, there is also a widely used selection – of the so-called skew-normal distribution. However, this selection is not strongly consistent in the above sense. In this paper, we show that this absence of strong consistency is not a fault of a specific selection but a general feature of the problem: for third and higher order moments, no strongly consistent selection is possible.

1 Formulation of the Problem

Need to select a distribution based on the first few moments. In many practical situations, we only have a partial information about the probability distribution. For example, often, all we know is the values of the first few moments.

Most probabilistic and statistical techniques assume that we know the exact form of a probability distribution; see, e.g., [5]. In situations when we only have partial information about the probability distribution, there are many probability distribu-

Thongchai Dumrongpokaphan
Department of Mathematics, Faculty of Science, Chiang Mai University, Thailand,
e-mail: tcd43@hotmail.com

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso, 500 W. University,
El Paso, Texas 79968, USA, e-mail: vladik@utep.edu

tions which are consistent with our knowledge. To use the usual techniques in such a situation, we therefore need to select, from all possible distributions, a single one.

In situations when we know the first two moments, there is a strongly consistent way of selecting a single distribution. In situations when all we know is the first two moments $\mu = \int x \cdot \rho(x)\,dx$ and $M_2 = \int x^2 \cdot \rho(x)\,dx$, a natural idea is to select a distribution for which the entropy (uncertainty) $S = -\int \rho(x) \cdot \ln(\rho(x))\,dx$ is the largest possible; see, e.g., [3].

By applying the Lagrange multiplier techniques, one can easily check that maximizing entropy under the constraints $\int \rho(x)\,dx = 1$, $\mu = \int x \cdot \rho(x)\,dx$ and $M_2 = \int x^2 \cdot \rho(x)\,dx$ leads to the Gaussian (normal) distribution, with probability density

$$\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

where $\sigma^2 = M_2 - \mu^2$.

This selection is *strongly consistent* in the following sense. Often, the random variable of interest has several components. For example, an overall income consists of salaries, pensions, unemployment benefits, interest on bank deposits, etc. Each of these categories, in its turn, can be subdivided into more subcategories. If for each of these categories, we only know the first two moments, then, in principle, we can apply the selection:

- either to the overall sum,
- or separately to each term,
- or we can go down to the next level of granularity and apply the selection to each term on this granularity level, etc.

It seems reasonable to require that whichever granularity level we select, the resulting distribution for the overall sum should be the same. This is indeed true for normal distributions. Indeed, knowing μ and M_2 is equivalent to knowing μ and the variance σ^2, and it is known that for the sum $X = X_1 + X_2$ of two independent random variables, its mean and variance are equal to the sum of the means and variances of the two components: $\mu = \mu_1 + \mu_2$ and $\sigma^2 = \sigma_1^2 + \sigma_2^2$. So:

- If we apply the selection to the sum itself, we then get a normal distribution with mean μ and standard deviation σ.
- Alternatively, if we first apply the selection to each component, we conclude that X_1 is normally distributed with mean μ_1 and standard deviation σ_1 and X_2 is normally distributed with mean μ_2 and standard deviation σ_2.

It is known that the sum of two independent normally distributed random variables is also normally distributed, with the mean $\mu = \mu_1 + \mu_2$ and the variance $\sigma^2 = \sigma_1^2 + \sigma_2^2$. Thus, in both cases, we indeed get the same probability distribution. This *strong consistency* is one of the reasons why selecting a normal distribution is indeed ubiquitous in practical applications.

Selection of a skew normal distribution is not strongly consistent. A natural next case is when, in addition to the first two moments μ and M_2, we also know the

third moment M_3. Alternatively, this can be described as knowing the mean μ, the variance $V = \sigma^2$, and the third central moment

$$m_3 \stackrel{\text{def}}{=} E[(X-\mu)^3].$$

In this case, we can no longer use the Maximum Entropy approach to select a single distribution. Indeed, if we try to formally maximize the entropy under the four constrains corresponding to the condition $\int \rho(x)\,dx = 1$ and to the three moments, then the Lagrange multiplier method leads to the function

$$\rho(x) = \exp(a_0 + a_1 \cdot x + a_2 \cdot x^2 + a_3 \cdot x^3)$$

which does not satisfy the requirement $\int \rho(x)\,dx = 1$:

- when $a_3 > 0$, then $\rho(x) \to \infty$ as $x \to +\infty$, so $\int_{-\infty}^{\infty} \rho(x)\,dx = \infty$; and
- when $a_3 < 0$, then $\rho(x) \to \infty$ as $x \to -\infty$, so also $\int_{-\infty}^{\infty} \rho(x)\,dx = \infty$.

There is a widely used selection, called a *skew normal* distribution (see, e.g., [2, 4, 6]), when we choose a distribution with the probability density function

$$\rho(x) = \frac{1}{2\omega} \cdot \phi\left(\frac{x-\eta}{\omega}\right) \cdot \Phi\left(\alpha \cdot \frac{x-\eta}{\omega}\right),$$

where:

- $\phi(x) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{x^2}{2}\right)$ is the pdf of the standard Gaussian distribution, with mean 0 and standard deviation 1, and
- $\Phi(x)$ is the corresponding cumulative distribution function

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\,dt.$$

For this distribution,

- $\mu = \eta + \omega \cdot \delta \cdot \sqrt{\frac{2}{\pi}}$, where $\delta \stackrel{\text{def}}{=} \frac{\alpha}{\sqrt{1+\alpha^2}}$,
- $\sigma^2 = \omega^2 \cdot \left(1 - \frac{2\delta^2}{\pi}\right)$, and
- $m_3 = \frac{4-\pi}{2} \cdot \sigma^3 \cdot \frac{(\delta \cdot \sqrt{2/\pi})^3}{(1-2\delta^2/\pi)^{3/2}}$.

The skew normal distribution has many applications, but it is *not* strongly consistent in the above sense: in general, the sum of two independent skew normal variables is *not* skew normal, and thus, the result of applying the selection depends on the level of granularity to which this selection is applied.

Natural question. Since the usual selection corresponding to three moments is not strongly consistent, a natural question is:

- is this a fault of this particular selection – and an alternative strongly consistent selection *is* possible,
- or is this a feature of the problem – and for the case of three moments, a strongly consistent selection is not possible?

In this paper, we show that under the reasonable assumption of scale-invariance, for three and more moments, a strongly consistent selection is not possible – and thus, the absence of strong consistency is a feature of the problem and not a limitation of the current selection of skew normal distributions.

2 Analysis of the Problem and the Main Result

Let us formulate the selection problem in precise terms. We want to assign, to each triple (μ, V, m_3) consisting of the mean, the variance, and the central third moment m_3, a probability distribution $\rho(x, \mu, V, m_3)$.

Let us list the natural properties of this assignment.

First property: continuity. Moments are rarely known exactly, we usually know them with some accuracy. It is thus reasonable to require that if the moments change slightly, then the corresponding distribution should not change much. In other words, it is reasonable to require that the function $\rho(x, \mu, V, m_3)$ is a continuous function of μ, V, and m_3.

Comment. As we can see from our proof, to prove the impossibility, it is sufficient to impose an even weaker requirement: that the dependence of $\rho(x, \mu, V, m_3)$ on μ, V, and m_3 is *measurable*.

Second property: strong consistency. We require that if X_1 and X_2 are independent random variables for which:

- X_1 is distributed according to the distribution $\rho(x, \mu_1, V_1, m_{31})$, and
- X_2 is distributed according to the distribution $\rho(x, \mu_2, V_2, m_{32})$,

then the sum $X = X_1 + X_2$ is distributed according to the distribution

$$\rho(x, \mu_1 + \mu_2, V_1 + V_2, m_{31} + m_{32}).$$

Final property: scale-invariance. Numerical values of different quantities depend on the choice of a measuring unit. For example, an income can be described in Baht or – to make comparison with people from other countries easier – in dollars. If we change the measuring unit to a new one which is λ times smaller, then the actual incomes will not change, but the numerical values will change: all numerical values will be multiplied by λ: $x \to x' = \lambda \cdot x$.

If we perform the selection in the original units, then we select a distribution with the probability density function $\rho(x,\mu,V,m_3)$. If we simply re-scale x to $x' = \lambda \cdot x$, then for x', we get a new distribution $\rho'(x') = \dfrac{1}{\lambda} \cdot \rho\left(\dfrac{x'}{\lambda},\mu,V,m_3\right)$.

We should get the exact same distribution if we make a selection *after* the re-scaling to the new units. After this re-scaling:

- the first moment is multiplied by λ: $\mu \to \lambda \cdot \lambda$,
- the variance is multiplied by λ^2: $V \to \lambda^2 \cdot V$, and
- the central third moment is multiplied by λ^3: $m_3 \to \lambda^3 \cdot m_3$.

So, in the new units, we get a probability distribution $\rho(x',\lambda\cdot\mu,\lambda^2\cdot V,\lambda^3\cdot m_3)$. A natural requirement is that the resulting selection should be the same, i.e., that we should have

$$\frac{1}{\lambda} \cdot \rho\left(\frac{x'}{\lambda},\mu,V,m_3\right) = \rho(x',\lambda\cdot\mu,\lambda^2\cdot V,\lambda^3\cdot m_3)$$

for all x, λ, μ, V, and m_3.

Comment. One can easily check that the both the above selection corresponding to skew normal distributions is scale-invariant (and that for the case of two moments, the standard normal-distribution selection is also scale-invariant).

Now, we can formulate the problem in precise terms.

Definition 1.

- *We say that a tuple* (μ,V,m_3) *is* possible *if there exists a probability distribution with mean* μ*, variance* V*, and central third moment* m_3.
- *By a* 3-selection*, we mean a measurable mapping that maps each possible tuple* (μ,V,m_3) *into a probability distribution* $\rho(x,\mu,V,m_3)$.
- *We say that a 3-selection is* strongly consistent *if for every two possible tuples, if* $X_1 \sim \rho(x,\mu_1,V_1,m_{31})$*,* $X_2 \sim \rho(x,\mu_2,V_2,m_{32})$*, and* X_1 *and* X_2 *are independent, then* $X_1+X_2 \sim \rho(x,\mu_1+\mu_2,V_1+V_2,m_{31}+m_{32})$.
- *We say that a 3-selection is* scale-invariant *if for every possible tuple* (μ,V,m_3)*, for every* $\lambda > 0$*, and for all* x'*, we have*

$$\frac{1}{\lambda} \cdot \rho\left(\frac{x'}{\lambda},\mu,V,m_3\right) = \rho(x',\lambda\cdot\mu,\lambda^2\cdot V,\lambda^3\cdot m_3)$$

.

Proposition 1. *No 3-selection is strongly consistent and scale-invariant.*

Comment. A similar result can be formulated – and similarly proven – for the case when we also know higher order moments. In this case, instead of the original moments, we can consider *cumulants* κ_n: terms at $\dfrac{\mathrm{i}^n \cdot t^n}{n!}$ in the Taylor expansion of the corresponding generating function $\ln(E[\exp(\mathrm{i}\cdot t\cdot X)])$. For $n=1$, $n=2$, and

$n = 3$, we get exactly the mean, the variance, and the central third moment. In general, cumulants are additive: if $X = X_1 + X_2$ and X_1 and X_2 are independent, then $\kappa_n(X) = \kappa_n(X_1) + \kappa_n(X_2)$.

Discussion. Since we cannot make a strongly consistent selection, what should we do? One possible idea is to use the fact that, in addition to the sum, min and max are also natural operations in many applications. For example, in econometrics, if there are several ways to invest money with the same level of risk, then an investor selects the one that leads to the largest interest rate.

From this viewpoint, once we have normally distributed random variables, it is also reasonable to consider minima and maxima of normal variables. Interestingly, in some cases, these minima and maxima are distributed according to the skew normal distribution. This may be an additional argument in favor of using these distributions.

Proof. It is known that when we deal with the sum of independent random variables $X = X_1 + X_2$, then, instead of the original probability density functions, it is more convenient to consider *characteristic functions*

$$\chi_{X_1}(\omega) \stackrel{\text{def}}{=} E[\exp(\mathrm{i} \cdot \omega \cdot X_1)], \;\; \chi_{X_2}(\omega) \stackrel{\text{def}}{=} E[\exp(\mathrm{i} \cdot \omega \cdot X_2)],$$

and

$$\chi_X(\omega) \stackrel{\text{def}}{=} E[\exp(\mathrm{i} \cdot \omega \cdot X)].$$

Indeed, for these characteristic functions, we have

$$\chi_X(\omega) = \chi_{X_1}(\omega) \cdot \chi_{X_2}(\omega).$$

Comment. To avoid possible confusion, it is worth noticing that this frequency ω is unrelated to the parameter ω of the skew normal distribution.

Thus, instead of considering the original probability density functions

$$\rho(x, \mu, V, m_3),$$

let us consider the corresponding characteristic functions

$$\chi(\omega, \mu, V, m_3) \stackrel{\text{def}}{=} \int \exp(\mathrm{i} \cdot \omega \cdot x) \cdot \rho(x, \mu, V, m_3)\, dx.$$

Since the original dependence $\rho(x, \mu, V, m_3)$ is measurable, its Fourier transform $\chi(\omega, \mu, V, m_3)$ is measurable as well.

In terms of the characteristic functions, the string consistency requirement takes a simpler form

$$\chi(\omega, \mu_1 + \mu_2, V_1 + V_2, m_{31} + m_{32}) = \chi(\omega, \mu_1, V_1, m_{31}) \cdot \chi(\omega_2, \mu_2, V, m_{32}).$$

This requirement becomes even simpler if we take logarithm of both sides. Then, for the auxiliary functions $\ell(\omega,\mu,V,m_3) \stackrel{\text{def}}{=} \ln(\chi(\omega,\mu,V,m_3))$, we get an even simpler form of the requirement:

$$\ell(\omega,\mu_1+\mu_2,V_1+V_2,m_{31}+m_{32}) = \ell(\omega,\mu_1,V_1,m_{31})+\ell(\omega_2,\mu_2,V,m_{32}).$$

It is known (see, e.g., [1]) that the only measurable functions with this additivity property are linear functions, so

$$\ell(\omega,\mu,V,m_3) = \mu\cdot\ell_1(\omega)+V\cdot\ell_2(\omega)+m_3\cdot\ell_3(\omega)$$

for some functions $\ell_1(\omega)$, $\ell_2(\omega)$, and $\ell_3(\omega)$.

Let us now use the scale invariance requirement. When we re-scale a random variable X, i.e., replace its numerical values x to new numerical values $x' = \lambda\cdot x$, then, for the new random variable $X' = \omega\cdot X$, we have

$$\chi_{X'}(\omega) = E[\exp(\mathrm{i}\cdot\omega\cdot X')] = E[\exp(\mathrm{i}\cdot\omega\cdot(\lambda\cdot X))] = E[\exp(\mathrm{i}\cdot(\omega\cdot\lambda)\cdot X)] = \chi_X(\lambda\cdot\omega).$$

Thus re-scaled characteristic function $\chi(\lambda\cdot\omega,\mu,V,m_3)$ should be equal to the characteristic function obtained when we use re-scaled values of the moments $\chi(\omega,\lambda\cdot\mu,\lambda^2\cdot V,\lambda^3\cdot m_3)$:

$$\chi(\lambda\cdot\omega,\mu,V,m_3) = \chi(\omega,\lambda\cdot\mu,\lambda^2\cdot V,\lambda^3\cdot m_3).$$

Their logarithms should also be equal, so:

$$\ell(\lambda\cdot\omega,\mu,V,m_3) = \ell(\omega,\lambda\cdot\mu,\lambda^2\cdot V,\lambda^3\cdot m_3).$$

Substituting the above linear expression for the function $\ell(\omega,\mu,V,m_3)$ into this equality, we conclude that

$$\mu\cdot\ell_1(\lambda\cdot\omega)+V\cdot\ell_2(\omega\cdot\omega)+m_3\cdot\ell_3(\lambda\cdot\omega) = \lambda\cdot\mu\cdot\ell_1(\omega)+\lambda^2\cdot V\cdot\ell_2(\omega)+\lambda^3\cdot m_3\cdot\ell_3(\omega).$$

This equality must hold for all possible triples (μ,V,m_3). Thus, the coefficient at μ, V, and m_3 on both sides must coincide.

- By equating coefficients at μ, we conclude that $\ell_1(\lambda\cdot\omega) = \lambda\cdot\ell_1(\omega)$. In particular, for $\omega = 1$, we conclude that $\ell_1(\lambda) = \lambda\cdot\ell_1(1)$, i.e., that $\ell_1(\omega) = c_1\cdot\omega$ for some constant c_1.
- By equating coefficients at V, we conclude that $\ell_2(\lambda\cdot\omega) = \lambda^2\cdot\ell_1(\omega)$. In particular, for $\omega = 1$, we conclude that $\ell_2(\lambda) = \lambda^2\cdot\ell_1(1)$, i.e., that $\ell_2(\omega) = c_2\cdot\omega^2$ for some constant c_2.
- By equating coefficients at m_3, we conclude that $\ell_3(\lambda\cdot\omega) = \lambda^3\cdot\ell_3(\omega)$. In particular, for $\omega = 1$, we conclude that $\ell_3(\lambda) = \lambda\cdot\ell_3(1)$, i.e., that $\ell_3(\omega) = c_3\cdot\omega^3$ for some constant c_3.

Thus, we get $\ell(\omega,\mu,V,m_3) = c_1\cdot\mu\cdot\omega+c_2\cdot V\cdot\omega^2+c_3\cdot m_3\cdot\omega^3$. Since, by definition, $\ell(\omega,\mu,V,m_3)$ is the logarithm of the characteristic function, we thus conclude

that the characteristic function has the form

$$\chi(\omega, u, V.m_3) = \exp(c_1 \cdot \mu \cdot \omega + c_2 \cdot V \cdot \omega^2 + c_3 \cdot m_3 \cdot \omega^3).$$

In principle, once we know the characteristic function, we can reconstruct the probability density function by applying the inverse Fourier transform. The problem here is that, as one can easily check by numerical computations, the Fourier transform of the above expression is, in general, *not* an everywhere non-negative function – and thus, cannot serve as a probability density function.

This proves that a strongly consistent selection of a probability distribution is indeed impossible.

Comment. If we only consider two moments, then the above proof leads to the characteristic function $\chi(\omega, \mu, V) = \exp(c_1 \cdot \mu \cdot \omega + c_2 \cdot V \cdot \omega^2)$ that describes the Gaussian distribution. Thus, we have, in effect proven the following auxiliary result:

Definition 2.

- *We say that a tuple* (μ, V) *is* possible *if there exists a probability distribution with mean* μ *and variance* V.
- *By a* 2-selection*, we mean a measurable mapping that maps each possible tuple* (μ, V) *into a probability distribution* $\rho(x, \mu, V)$.
- *We say that a 2-selection is* strongly consistent *if for every two possible tuples, if* $X_1 \sim \rho(x, \mu_1, V_1)$, $X_2 \sim \rho(x, \mu_2, V_2)$, *and* X_1 *and* X_2 *are independent, then* $X_1 + X_2 \sim \rho(x, \mu_1 + \mu_2, V_1 + V_2)$.
- *We say that a 2-selection is* scale-invariant *if for every possible tuple* (μ, V), *for every* $\lambda > 0$, *and for all* x', *we have*

$$\frac{1}{\lambda} \cdot \rho\left(\frac{x'}{\lambda}, \mu, V\right) = \rho(x', \lambda \cdot \mu, \lambda^2 \cdot V)$$

.

Proposition 2. *Every strongly consistent and scale-invariant 2-selection assigns, to each possible tuple* (μ, V), *a Gaussian distribution with mean* μ *and variance* V.

Acknowledgments

This work was supported by Chiang Mai University, Thailand. This work was also supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation.

References

1. J. Aczél and J. Dhombres, *Functional Equations in Several Variables*, Camridge University Press, Cambridge, UK, 2008.
2. A. Azzalini and A. Capitanio, *The Skew-Normal and Related Families*, Cambridge University Press, Cambridge, Massachusetts, 2013.
3. E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.
4. B. Li, D. Shi, and T. Wang, "Some applications of one-sided skew distributions", *International Journal Intelligent Technologies and Applied Statistics*, 2009. Vol. 2, No. 1.
5. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.
6. T. Wang, B. Li, and A. K. Gupta, "Distribution of quadratic forms under skew normal settings", *Journal of Multivariate Analysis*, 2009, Vol. 100, pp. 533–545.