

LAMBERT W RANDOM VARIABLES—A NEW FAMILY OF GENERALIZED SKEWED DISTRIBUTIONS WITH APPLICATIONS TO RISK ESTIMATION

BY GEORG M. GOERG

Carnegie Mellon University

Originating from a system theory and an input/output point of view, I introduce a new class of generalized distributions. A parametric nonlinear transformation converts a random variable X into a so-called Lambert W random variable Y , which allows a very flexible approach to model skewed data. Its shape depends on the shape of X and a skewness parameter γ . In particular, for symmetric X and nonzero γ the output Y is skewed. Its distribution and density function are particular variants of their input counterparts. Maximum likelihood and method of moments estimators are presented, and simulations show that in the symmetric case additional estimation of γ does not affect the quality of other parameter estimates. Applications in finance and biomedicine show the relevance of this class of distributions, which is particularly useful for slightly skewed data. A practical by-result of the Lambert W framework: data can be “unskewed.”

The R package `LambertW` developed by the author is publicly available (CRAN).

1. Introduction. Exploratory data analysis regarding asymmetry in data is usually based on histograms and nonparametric density estimates, and statements such as “this data set looks almost Gaussian, but it is skewed to the right” or “these asset returns have heavy tails, but they are too skewed that a student- t would make sense” are fairly common. It is therefore natural to generalize symmetric distributions to allow for asymmetry.

A prominent generalization is the skew-normal distribution [Azzalini (1985)], which includes the Gaussian as a special case. A skew-normal random variable (RV) is defined by having the probability density function (p.d.f.) $f(x) = 2\phi(x)\Phi(\alpha x)$, where $\Phi(\cdot)$ is the cumulative distribution function (c.d.f.) of a standard Gaussian, and $\alpha \in \mathbb{R}$ is the skew parameter. This approach to skewness has not only led to substantial research in the skew-normal case [Azzalini and Capitanio (1999), Arellano-Valle and Azzalini (2006)], but the same concept has also been used for student- t [Azzalini and Capitanio (2003)] and Cauchy distributions [Arnold and Beaver (2000), Behboodian, Jamalizadeh and Balakrishnan (2006)],

Received February 2010; revised January 2011.

Key words and phrases. Family of skewed distributions, skewness, transformation of random variables, Lambert W , latent variables, stylized facts of asset returns, value at risk, GARCH.

among others. In all these cases, a parametric manipulation of the original symmetric p.d.f. introduces skewness.

Notwithstanding the huge success of this approach to model skewed data, manipulating the p.d.f. to introduce skewness seems like putting the cart before the horse: densities are skewed, because the random variable is—not the other way around. Also, applied research starts with data, not with histograms.

Motivated by this data-driven view on skewness, I propose a novel approach to asymmetry: Lambert $W \times F_X$ distributions emerge naturally by modeling the observable RV Y as the output of a system \mathcal{S} driven by random input X with c.d.f. $F_X(x)$. Here \mathcal{S} can either be a real chemical, physical, or biological system, or refer to any kind of mechanism in a broader sense. In statistical modeling such a system is simply represented by transformations of RVs. As there are no restrictions on $F_X(x)$, this is a very general framework that can be analyzed in detail for a particularly chosen input c.d.f. Figure 1 illustrates the methodology.

For instance, consider \mathcal{S} being the stock market, where people buy and sell an asset according to its expected success in the future. Asset returns, that is, the percentage change in price, typically exhibit negative skewness and positive excess kurtosis—so-called *stylized facts* [Yan (2005), Cont (2001)]. The left panel of Figure 2 shows daily log-returns (in percent) $\mathbf{y} := \{y_t \mid t = 1, \dots, 1,413\}$ of

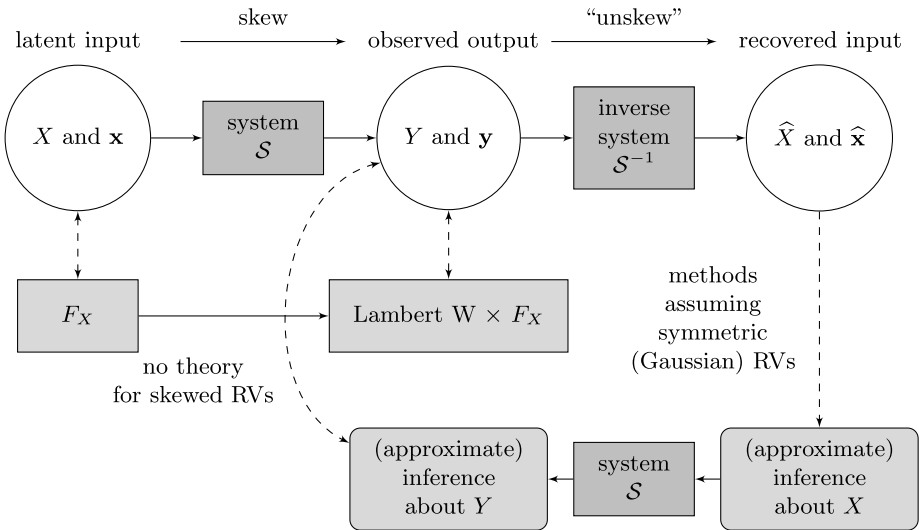


FIG. 1. Schematic view of the Lambert W approach to asymmetry: (left) an input/output system \mathcal{S} transforms (solid arrows) input $X \sim F_X$ to output $Y \sim \text{Lambert } W \times F_X$ and herewith introduces skewness; (right) inference about skewed data $\mathbf{y} = (y_1, \dots, y_N)$: (1) unskew observed \mathbf{y} to latent symmetric data $\hat{\mathbf{x}}$, (2) use methods of your choice (regression, time series models, quantile estimation, hypothesis tests, etc.) for statistical inference based on $\hat{\mathbf{x}}$, and (3) convert results back to the “skewed world” of \mathbf{y} .

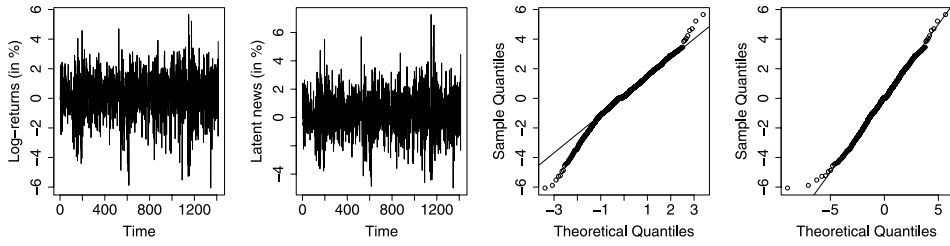


FIG. 2. Daily equity fund returns (LATAM): (left) observed returns \mathbf{y} and estimated latent news $\hat{\mathbf{x}}_{\text{TMLE}}$; (right) Gaussian QQ plot and Lambert $W \times t$ QQ plot of \mathbf{y} .

an equity fund investing in Latin America (LATAM¹). Also, these returns are clearly non-Gaussian given their excess kurtosis (1.201) and large negative skewness (−0.433)—see Table 1. The excess kurtosis is typically addressed by a student t -distribution, but here a Kolmogorov–Smirnov test still rejects $\mathbf{y} \sim t_{6,22}$ on a 5% level (even for the estimated ν), as the empirical skewness is too large. Thus, to model the probabilistic properties of such data, asymmetric distributions must be used.

Using Lambert $W \times F$ RVs to model the asymmetry in asset returns is perfectly suitable not only given empirical evidence of “almost student- t , but a little skewed data,” but also by a more fundamental viewpoint. Price changes are commonly considered as the result of bad and good news hitting the market: bad news,

TABLE 1
LATAM returns \mathbf{y} and back-transformed series $\hat{\mathbf{x}}_{\hat{\nu}}$: (top) summary statistics; (bottom) Shapiro–Wilk (SW) & Jarque–Bera (JB) normality tests, and Kolmogorov–Smirnov (KS) test for student- t with ν degrees of freedom

LATAM	\mathbf{y}	$\hat{\mathbf{x}}_{\hat{\nu}\text{IGMM}}$	$\hat{\mathbf{x}}_{\hat{\nu}\text{TMLE}}$
Min	−6.064	−5.073	−4.985
Max	5.660	7.036	7.268
Mean	0.121	0.190	0.198
Median	0.138	0.138	0.138
St. dev.	1.468	1.456	1.457
Skewness	−0.433	0.000	0.053
Kurtosis	1.201	1.100	1.159
$\hat{\nu}$	6.220	7.093	7.048
SW	0.000	0.000	0.000
JB	0.000	0.000	0.000
KS ($t_{\hat{\nu}}$)	0.028	0.088	0.102

¹Data from January 1, 2002 until May 31, 2007: R package `fEcofin`, data set `equityFunds`, series `LATAM`.

negative returns; good news, positive returns. The empirical evidence of negative skewness evokes the following question: why should news per se be negatively skewed? Or put in other words: do really bad things happen more often than really good things?

In the Lambert W framework this news \leftrightarrow return relation is modeled under the assumption that the probability of getting negative news is about the same as of getting positive news, but typically people react far more drastically facing negative than positive ones. Thus, news $X \sim F_X(x)$ are symmetrically distributed, the market \mathcal{S} acts as an asymmetric filter, and the measurable/observable outcome is a skewed RV Y /data \mathbf{y} .

Last, the right part of Figure 1 also illustrates a very pragmatic, yet useful way to exploit the Lambert W framework for (slightly) skewed data. If a certain statistical procedure or model assumes a symmetric—a Gaussian, as often is the case—distribution and no skewed implementation of this method is available, then instead of applying it to the skewed \mathbf{y} , it is advisable to work with the “symmetrized” data $\hat{\mathbf{x}}$, make statistical inference about X based on $\hat{\mathbf{x}}$, and then transform the obtained results back to the “skewed world” of Y . Although this is only an approximation to the truth, at least this approach takes skewness into consideration instead of ignoring it.

Section 2 defines Lambert W RVs and their basic properties are studied. Section 3 presents analytic expressions of the c.d.f. $G_Y(y)$ and p.d.f. $g_Y(y)$, which are particular variants of their input counterparts. After studying Gaussian input in Section 4, various estimators for the parameter vector of Lambert $W \times F$ RVs are introduced in Section 5. Section 6 compares their finite sample properties and shows that additional estimation of the skewness parameter γ does not affect the quality of other parameter estimates. This new class of distribution functions is particularly useful for data with slightly negative skewness, thus, Section 7 demonstrates its adequacy on an Australian athletes data set and on the LATAM return series.

In particular, Section 7.2 shows that the input-output system (Figure 1) with student- t input X is a proper model for these returns. A detailed comparison of quantile estimates, which are essential to get appropriate risk measures of an asset, confirms the aptness of Lambert W distributions (see Lambert W QQ plot in Figure 2). Empirical evidence for the significance of conditional heteroskedastic time series models using Lambert $W \times F$ innovations concludes Section 7.2.

Finally, Section 8 establishes a direct link of this new class of distributions to the existing statistics literature, noting that the square of a RV having Tukey’s h distribution [Tukey (1977)] has a Lambert $W \times \chi_1^2$ distribution.

Computations, figures and simulations were realized with the open-source statistics package R [R Development Core Team (2008)]. Functions used in the analysis are available as the R package `LambertW`, which provides many other methods to perform Lambert W inference in practice.

2. Lambert W random variables. The general notion of a system \mathcal{S} with random input and output as shown in Figure 1 translates to a variable transformation in statistical terminology.

DEFINITION 2.1 (Noncentral, nonscaled Lambert $W \times F$ RV). Let U be a continuous RV with c.d.f.

$$(2.1) \quad F_U(u) = \mathbb{P}(U \leq u), \quad u \in \mathbb{R},$$

and p.d.f. $f_U(u)$. Then

$$(2.2) \quad Z := U \exp(\gamma U), \quad \gamma \in \mathbb{R},$$

is a *noncentral, nonscaled Lambert $W \times F$ RV* with skewness parameter γ .

If U is from a parametric family $F_U(u | \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ parametrizes the F_U , then Z is a noncentral, nonscaled Lambert $W \times F$ RV with parameter vector $\theta = (\boldsymbol{\beta}, \gamma)$.

The key of this family of RVs is γ , which can take any value on the real line. As $\exp(\cdot)$ is always positive, U and Z have the same sign. For readability let $H_\gamma(u) := u \exp(\gamma u)$. For $\gamma = 0$ transformation (2.2) reduces to the identity $Z \equiv U$; thus, Z possesses the exact same properties as U . By continuity of $H_\gamma(\cdot)$, one can expect for $\gamma \neq 0$ but close, also $Z \neq U$ but close.

Transformation (2.2) indeed describes a system \mathcal{S} with an asymmetry property: let $U \sim F_U(u)$ be a symmetric zero-mean RV, then Z is a skewed version of U —depending on the sign of γ . For $\gamma < 0$ negative U are amplified by the factor $\exp(\gamma U) > 1$ and positive U are damped by $0 < \exp(\gamma U) < 1$: Z is skewed to the left. For $\gamma > 0$ the same reasoning shows that Z is a positively.

The noncentral moments $\mathbb{E}(Z^n)$ equal

$$(2.3) \quad \psi_{(n)} := \int u^n e^{\gamma un} f_U(u) du.$$

If the moment-generating function $M_U(t) := \mathbb{E}e^{tU}$ exists for $t = \gamma n$, then (2.3) can be rewritten to get a more tractable formula. As

$$\frac{\partial^n}{\partial \gamma^n} e^{\gamma Un} = (Un)^n e^{\gamma Un},$$

interchanging differentiation and the integral sign yields

$$(2.4) \quad \psi_{(n)} = n^{-n} \frac{\partial^n}{\partial \gamma^n} M_U(\gamma n).$$

If $M_U(t)$ does not exist (e.g., for student- t U), then (2.3) must be calculated explicitly.

2.1. *Scale family input.* In a typical input/output system \mathcal{S} such as a microphone/loudspeaker setting, the loudspeaker will be louder if speakers raise their voice. In this sense it is stable with respect to scaling: doubling the volume of the input doubles the volume of the loudspeakers—the signal is not affected in any other way. Viewing this system as a Lambert $W \times F$ RV system (where the signal is considered as a RV), multiplying X by a factor κ , should—*ceteris paribus*—only affect the output Y by multiplying by κ ; other properties, such as skewness or kurtosis, should not be altered.

Transformation (2.2), however, does not have this scaling property of U . Hence, to allow a comparable system characterization via γ among different scalable data sets define a scaled Lambert W RV.

DEFINITION 2.2 (Scale Lambert $W \times F$ RV). Let $U := X/\sigma_x$ be the unit-variance version of a continuous RV X from a scale family $F_X(x | \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is the parameter (vector) of F_X and σ_x the standard deviation of X . Then

$$(2.5) \quad Y := \{U \exp(\gamma U)\}\sigma_x = X \exp(\gamma X/\sigma_x), \quad \gamma \in \mathbb{R}, \sigma_x > 0,$$

is a *scale Lambert $W \times F$ RV* with parameter vector $\theta = (\boldsymbol{\beta}, \gamma)$.

Transformation (2.5) is invariant to scaling of the input, for example, a different measurement unit for the input does not modify the asymmetry property of the system, but just scales the output accordingly.

Here σ_x is a function of $\boldsymbol{\beta}$: for an exponentially distributed input $X \sim \exp(\lambda)$, $\boldsymbol{\beta} = \lambda$ and $\sigma_x(\boldsymbol{\beta}) = \lambda^{-1}$; an input X having a Gamma distribution with shape α and rate β gives $\boldsymbol{\beta} = (\alpha, \beta)$ and $\sigma_x(\boldsymbol{\beta}) = \sqrt{\alpha}/\beta$.

2.2. *Location-scale family input.* The focus of this work lies in introducing skewness to symmetric RVs with support on $(-\infty, \infty)$, such as a Gaussian or student- t . These distributions are not only scale, but also shift invariant, a property Lambert $W \times F$ distribution should also have for location-family input. However, transformation (2.5) is not shift-invariant. For example, consider a zero-mean and unit variance input RV $U_0: \Omega \rightarrow \mathbb{R}$, $U_{10} := U_0 + 10$, and let $\gamma = 0.1$. If $U_0(\omega)$ is close to 0, then the shifted $U_{10}(\omega)$ will be close to 10. For the corresponding $Z_0(\omega)$ and $Z_{10}(\omega)$ this does not hold: $Z_0(\omega)$ is close to 0, but $Z_{10}(\omega)$ will not be shifted by 10, but lies close to $10 \exp(1) \approx 27.183$.

DEFINITION 2.3 (Location-scale Lambert $W \times F$ RV). Let X be a RV from a location-scale family with c.d.f. $F_X(x | \boldsymbol{\beta})$ with mean μ_x and standard deviation σ_x ; again $\boldsymbol{\beta}$ parametrizes F_X . Let $U = (X - \mu_x)/\sigma_x$ be the zero-mean, unit-variance version of X . Then

$$(2.6) \quad Y := \{U \exp(\gamma U)\}\sigma_x + \mu_x, \quad \gamma \in \mathbb{R},$$

is a *location-scale Lambert $W \times F$ RV* with parameter vector $\theta = (\boldsymbol{\beta}, \gamma)$.

As before, the parameter γ regulates the closeness between X and its skewed version Y .

For a full parametrization of a Lambert $W \times F$ distribution it is necessary to know $\theta = (\boldsymbol{\beta}, \gamma)$; viewing (2.6) only as a transformation from X to Y , it is more natural—and in practice more useful—to only consider μ_x, σ_x and γ , ignoring the particular structure of X given its parametrization by $\boldsymbol{\beta}$. In order to distinguish these two cases in the remaining part of this work let $\tau := (\mu_x, \sigma_x, \gamma) \in T = \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$. Clearly, τ can be computed from θ , $\tau = \tau(\theta)$, but not necessarily vice-versa.

For example, for a Gaussian X $\tau \equiv \theta$ since $\mu_x(\boldsymbol{\beta}) = \mu_x$ and $\sigma_x(\boldsymbol{\beta}) = \sigma_x$. In contrast, for a location-scale student- t input with $\boldsymbol{\beta} = (c, s, \nu)$ —where c is the location, s the scale and ν the degrees of freedom parameter— $\tau \neq \theta$: $\mu_x(\boldsymbol{\beta}) = c$ and $\sigma_x(\boldsymbol{\beta}) = s\sqrt{\frac{\nu}{\nu-2}}$ if $\nu > 2$.

Thus, below I use either θ if the full parametrization is important or τ if it is sufficient to consider (2.6) only as a transformation rather than a fully specified parametric distribution.

NOTATION 2.4 (Lambert $W \times F$ RV). For simplicity I will refer to all Y in Definitions 2.1, 2.2 and 2.3 as a Lambert $W \times F$ RV. Which one of the three transformations (2.2), (2.5) or (2.6) is used to generate Y will be clear from the type of input X . For example, since a χ_k^2 distribution does not have location or scale parameters, a Lambert $W \times \chi_k^2$ RV refers to Y in Definition 2.1; the exponential distribution is a scale family, thus, a Lambert $W \times \exp(\lambda)$ RV Y is defined in Definition 2.2; and for Gaussian input X , the corresponding Lambert $W \times$ Gaussian Y refers to Definition 2.3.²

2.3. *Latent variables.* So far attention has been drawn to Y and its properties given X and θ (or τ). Now consider the inverse problem: given Y and θ (or only τ), what does X look like?

This is not only interesting for a latent variable interpretation of X , but the inverse of a transformation is essential to derive the c.d.f. of the transformed variable. Before analyzing transformation (2.6), consider the nonlinear transformation $H: \mathbb{C} \rightarrow \mathbb{C}, u \mapsto u \exp u =: z$ [Figure 3 shows $H(u)$ only for $u \in \mathbb{R}$]. For positive u the function is bijective and resembles $\exp(u)$ very closely. For negative u , however, $H(u)$ is quite different from $\exp(u)$: it takes on negative values, its minimum value equals $-\frac{1}{e}$, and—most importantly—it is nonbijective.

Although $H(u)$ has no analytical inverse [Rosenlicht (1969)], its implicitly defined inverse function is well known in mathematics and physics.

²Although technically not correct, one can think of a scale Lambert $W \times F$ transformation having $\tau = (0, \sigma_x, \gamma)$, and a noncentral, nonscaled Lambert $W \times F$ transformation having $\tau = (0, 1, \gamma)$. This is especially useful for empirical work and implementation of the methods involving Lambert $W \times F$ RVs.

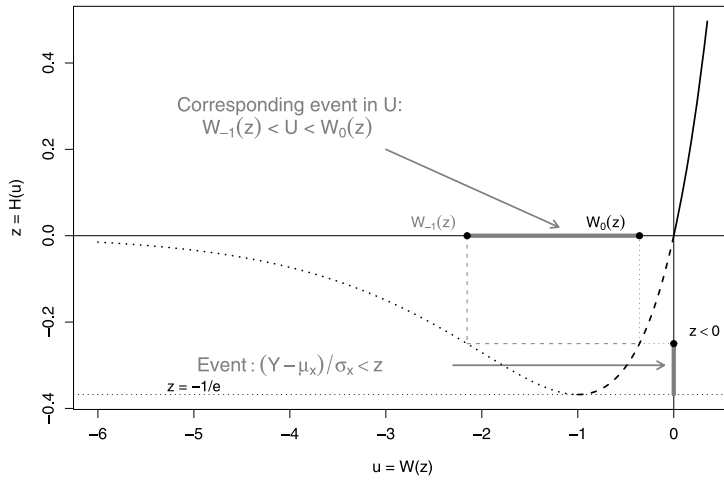


FIG. 3. Lambert W function: transformation $H(u)$ and the two inverse branches of $W(z)$ for $z < 0$: principal branch (dashed curve) and nonprincipal branch (dotted curve).

DEFINITION 2.5 (Lambert W function). The many-valued function $W(z)$ is the root of

$$(2.7) \quad W(z)e^{W(z)} = z, \quad z \in \mathbb{C},$$

and is commonly denoted as the *Lambert W function*.

Generally the Lambert W function is defined for any $z \in \mathbb{C}$. Since Lambert W RVs are only defined for real-valued outcomes, in this work the domain and image of the Lambert W function is restricted to the reals. For $z \in [-\infty, -1/e)$ no real solution exists; for $z \in [-1/e, \infty)$ $W(z)$ is a real-valued function. If $z \in [-1/e, 0)$, there are two real solutions: the principal branch $W_0(z) \geq -1$ and the nonprincipal branch $W_{-1}(z) \leq -1$; for $z \in [0, \infty)$ only one real-valued solution exists, $W_0(z) = W_{-1}(z)$ (see Figure 3).

For a detailed review including useful properties and functional identities of $W(z)$ see Corless et al. (1996), Valluri, Jeffrey and Corless (2000) and the references therein.

Figure 3 also shows how skewness is introduced via transformation (2.6). Symmetric input X (x -axis) is mapped to asymmetric output Y (y -axis) due to the curvature of $H(u)$. Analogously, mapping values from the y -axis to the x -axis “unskews” them. Figure 3 shows $H_\gamma(u) : u \mapsto z$ for $\gamma = 1$, thus, its inverse is Lambert’s W function ($W(z) : z \mapsto u$). The curvature of $H_\gamma(u) = u \exp(\gamma u)$ depends on the skewness parameter: for $\gamma = 0$ no curvature is present [$H_0(u) = u$]; higher γ results in more curvature, and thus more skewness in Y .

It can be easily verified that $W_\gamma(Z) := W(\gamma Z)/\gamma$ is the inverse function of transformation (2.2). Hence, given Y and τ , the unobservable input X can be re-

covered via

$$(2.8) \quad W_\gamma \left(\frac{Y - \mu_x}{\sigma_x} \right) \sigma_x + \mu_x = U \sigma_x + \mu_x = X.$$

For empirical work it is important to point out that (2.8) does not require specific knowledge about F_X or β ; μ_x and σ_x (and γ) suffice. This will become especially useful for estimating the optimal inverse transformation—see Section 5.2.

REMARK 2.6 (Nonprincipal branch). The Lambert W function has two branches on the negative real line (Figure 3), so transformation (2.8) is not unique. For example, consider $z = -0.25$ and $\gamma = 1$. The two real-valued solutions are $W_0(-0.25) \approx -0.357$ and $W_{-1}(-0.25) \approx -2.153$. Assuming a stable input/output system, only the principal branch makes sense³—denoted by $W_{\gamma,0}(\cdot)$. If the nonprincipal solution is required, $W_{\gamma,-1}(\cdot)$ will be used.

The probability p_{-1} that the observed value $Z(\omega)$ was indeed caused by the nonprincipal solution is at most $\mathbb{P}\{U < -1/|\gamma|\}$, since $H_\gamma(u)$ changes its monotonicity at $u = -1/\gamma$. For Gaussian X and $\gamma = 0.1$ —a very large value given empirical evidence—this probability equals $7.62 \cdot 10^{-24}$. For an input with student t -distribution and $\nu = 4$ degrees of freedom $p_{-1} \approx 7.26 \cdot 10^{-5}$. Hence, ignoring the nonprincipal root to obtain unique latent data should not matter too much in practice.

Algorithm 1 describes the empirical version of (2.8). The so obtained

$$(2.9) \quad x_n = W_{\gamma,0} \left(\frac{y_n - \mu_x}{\sigma_x} \right) \sigma_x + \mu_x, \quad n = 1, \dots, N,$$

is the input data $\widehat{\mathbf{x}}_\tau$ generating the observed \mathbf{y} and should have c.d.f. $F_X(x)$. Here $\widehat{(\cdot)}$ does not stand for an estimate of τ , but since $W_{\gamma,0}(\cdot)$ ignores the nonprincipal branch, Algorithm 1 need not return the “true” input data \mathbf{x} —even if τ is known.⁴

Algorithm 1 Get input $\widehat{\mathbf{x}}_\tau$; function `get.input(\cdot)` in the `LambertW` package.

Input: data vector \mathbf{y} ; parameter vector $\tau = (\mu_x, \sigma_x, \gamma)$.

Output: input vector $\widehat{\mathbf{x}}_\tau$.

- 1: $\mathbf{z} = (\mathbf{y} - \mu_x) / \sigma_x$.
 - 2: back-transform \mathbf{z} via the principal branch to $\mathbf{u} = W_{\gamma,0}(\mathbf{z})$.
 - 3: **return** $\widehat{\mathbf{x}}_\tau = \mathbf{u} \sigma_x + \mu_x$.
-

³The output is assumed to be similar to the input, but skewed. Therefore, the input values causing the output should lie close to them: observing $z = -0.25$, it is more reasonable to assume that this corresponds to the close input of $W_0(-0.25) \approx -0.357$ rather than the very extreme $W_{-1}(-0.25) \approx -2.153$.

⁴This only applies if $p_{-1} = \mathbb{P}(U < -\frac{1}{|\gamma|}) > 0$, as otherwise the back-transformation $W_\gamma(z) = W_{\gamma,0}(z)$ is bijective. In particular, if $U \geq 0$ —for example, for scale family input $X \geq 0$ —then $\mathbf{x}_\tau \equiv \mathbf{x}$, not just an approximation. See also Corollary 3.3.

For small γ , $\widehat{\mathbf{x}}_{\tau}$ will most likely equal the true \mathbf{x} for all n ; for large γ some y_j 's might be falsely assigned to the principal x_j 's, although these y_j 's were actually caused by nonprincipal x_j 's. For an estimate $\widehat{\tau}$ the notation $\widehat{\mathbf{x}}_{\widehat{\tau}}$ will be used, which itself is an approximation to $\widehat{\mathbf{x}}_{\tau}$.

3. Distribution and density function. For ease of notation and readability let

$$(3.1) \quad \begin{aligned} z &:= \frac{y - \mu_x}{\sigma_x}, & u_0 &:= W_{\gamma,0}(z), & u_{-1} &:= W_{\gamma,-1}(z), \\ x_0 &:= u_0\sigma_x + \mu_x, & x_{-1} &:= u_{-1}\sigma_x + \mu_x. \end{aligned}$$

By definition,

$$\begin{aligned} G_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(\{U \exp(\gamma U)\}\sigma_x + \mu_x \leq y) \\ &= \mathbb{P}(U \exp \gamma U \leq z). \end{aligned}$$

The transformation $H_{\gamma}(u)$ changes its monotonicity at $u = -1/\gamma$ and its inverse $W_{\gamma,0}(z)$ at $z = -1/(\gamma e)$. Consequently, the event $\{Y < (>) y\}$ [for $\gamma > (<) 0$] has to be split up into separate events in U to derive the distribution of Y .

THEOREM 3.1 (Distribution of a location-scale Y). *The c.d.f. of a location-scale Lambert $W \times F$ RV Y equals (for $\gamma > 0$)*

$$(3.2) \quad G_Y(y | \boldsymbol{\beta}, \gamma) = \begin{cases} 0, & \text{if } y < -\frac{\sigma_x}{\gamma e} + \mu_x, \\ F_X(x_0 | \boldsymbol{\beta}) - F_X(x_{-1} | \boldsymbol{\beta}), & \text{if } -\frac{\sigma_x}{\gamma e} + \mu_x \leq y \leq \mu_x, \\ F_X(x_0 | \boldsymbol{\beta}), & \text{if } y \geq \mu_x. \end{cases}$$

The case $\gamma < 0$ can be obtained analogously and for $\gamma = 0$ it is clear that $G_Y(y | \boldsymbol{\beta}, 0) = F_X(y | \boldsymbol{\beta})$.

PROOF. Follows by matching the events in Z with the corresponding events in U [Glen, Leemis and Drew (1997)]; see Figure 3. \square

For $z = -1/(\gamma e)$ both branches of $W(z)$ coincide, thus, $u_0 = u_{-1}$. Therefore, $F_X(u_0\sigma_x + \mu_x) - F_X(u_{-1}\sigma_x + \mu_x) \equiv 0$ at $z = -1/(\gamma e)$, which implies continuity of $G_Y(y)$ at $y = \mu_x - \frac{\sigma_x}{\gamma e}$; the same reasoning shows continuity at $y = \mu_x$ ($z = 0$).

THEOREM 3.2 (Density of a location-scale Y). *The p.d.f. of a location-scale Lambert $W \times F$ RV Y equals (for $\gamma > 0$)*

$$(3.3) \quad g_Y(y | \boldsymbol{\beta}, \gamma) = \begin{cases} 0, & \text{if } y < -\frac{\sigma_x}{\gamma e} + \mu_x, \\ f_X(x_0 | \boldsymbol{\beta}) \cdot W'_0(\gamma z) - f_X(x_{-1} | \boldsymbol{\beta}) \cdot W'_{-1}(\gamma z), & \text{if } -\frac{\sigma_x}{\gamma e} + \mu_x \leq y \leq \mu_x, \\ f_X(x_0 | \boldsymbol{\beta}) \cdot W'_0(\gamma z), & \text{if } y \geq \mu_x. \end{cases}$$

Again, $\gamma < 0$ can be obtained analogously, and $g_Y(y | \beta, 0) = f_X(y | \beta)$.

PROOF. Using that $\frac{d}{dz} W_\gamma(z) = W'(\gamma z)$, the first derivative of $G_Y(y)$ with respect to y equals (3.3). The same arguments as for $G_Y(y)$ show that $g_Y(y)$ is continuous at $y = -\sigma_x/(\gamma e) + \mu_x$ and $y = \mu_x$. \square

In general, the support of Y depends on $\tau = \tau(\theta) \in T$ if $\gamma \neq 0$. However, restricting τ to the subspace $S_c := \{\tau \in T \mid -\sigma_x/(\gamma e) + \mu_x = c\}$ gives the same support $[c, \infty)$ for all $\tau \in S_c \subseteq T$ [or $(-\infty, c]$ for $\gamma < 0$]. Of particular empirical importance are

$$(3.4) \quad S_0 := \{\theta \in \Theta \mid \mu_x = \sigma_x/(\gamma e)\} \quad \text{and} \quad S_{\pm\infty} = \{\theta \in \Theta \mid \gamma = 0\}.$$

For (a scale family) $X \sim F_X(x | \beta)$ taking values in $[0, \infty)$ and $\gamma \geq 0$, the support of the corresponding (scale) Lambert $W \times F$ RV Y does not depend on τ but always equals $[0, \infty)$.

COROLLARY 3.3 (C.d.f. and p.d.f. of a scale Lambert $W \times F$ RV). *If X is a nonnegative RV taking values in $[0, \infty)$ and $\gamma \geq 0$, then the inverse transformation $W_\gamma(y/\sigma_x)$ is unique. Hence, the c.d.f. and p.d.f. of a scale Lambert $W \times F$ RV Y equal*

$$(3.5) \quad G_Y(y | \beta, \gamma) = F_X\left(W_{\gamma,0}\left(\frac{y}{\sigma_x}\right)\sigma_x \mid \beta\right)$$

and

$$(3.6) \quad g_Y(y | \beta, \gamma) = f_X\left(W_{\gamma,0}\left(\frac{y}{\sigma_x}\right)\sigma_x \mid \beta\right) \cdot W'_0\left(\gamma \frac{y}{\sigma_x}\right).$$

PROOF. Follows by setting $\mu_x = 0$ in (3.1), (3.2) and (3.3), and noting that the case $u < 0$ does not exist since $X \geq 0$. \square

For the c.d.f. and p.d.f. of a noncentral, nonscaled Lambert $W \times F$ RV Y (Definition 2.1) with X taking values in $[0, \infty)$ set $\sigma_x = 1$ in (3.5) and (3.6).

Theorems 3.1 and 3.2 demonstrate the great flexibility of the Lambert W setting, since the closed form expressions for $G_Y(y | \beta, \gamma)$ and $g_Y(y | \beta, \gamma)$ hold for any well-defined input $F_X(x | \beta)$ and $f_X(x | \beta)$, respectively. Thus, researchers can easily create Lambert W variants of their favorite distribution F_X , by simply plugging F_X and f_X in (3.2) and (3.3). Figure 4 shows the p.d.f. and c.d.f. of the three Lambert $\times F$ RVs discussed in Notation 2.4 for four degrees of skewness, $\gamma = (0, 0.1, 0.2, 0.4)$. For $\gamma = 0$ the output Y equals the input X , thus, also their p.d.f.s/c.d.f.s coincide (solid black lines). With increasing γ , the RV Y —and thus its distribution and density—become more and more skewed to the right (since $\gamma > 0$).

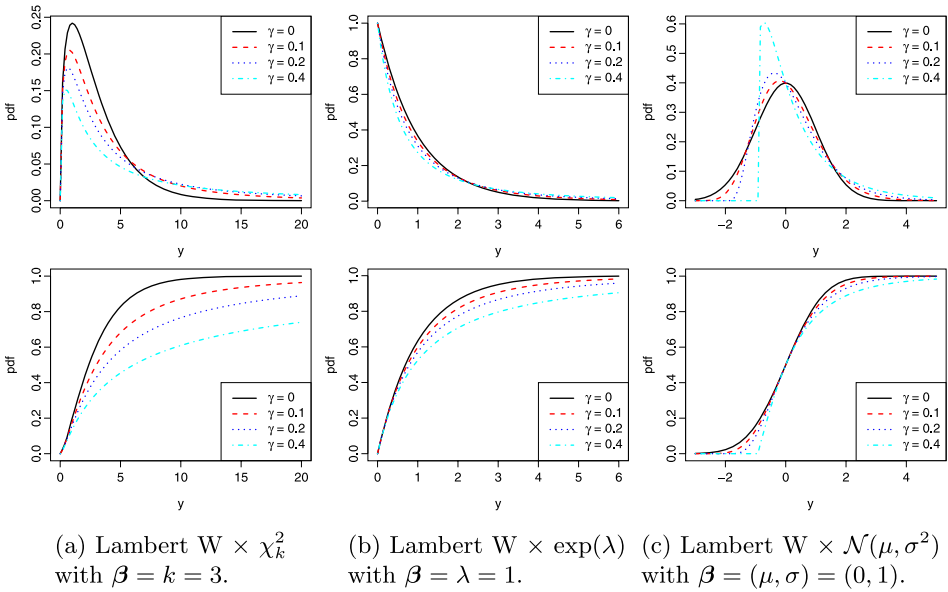


FIG. 4. The p.d.f. and c.d.f. of (a) a “noncentral, nonscaled,” (b) a “scale” and (c) a “location-scale” Lambert $W \times RV Y$ for different degrees of skewness.

Although Lambert W RVs are defined by transformation (2.6), they can be also considered as a particular variant of an arbitrary F_X —independent of this transformation. Sometimes the input/output aspect might be more insightful (e.g., stock returns), whereas otherwise solely the generalized distribution suffices to analyze a given data set. Especially, if the latent variable X does not have any suitable interpretation (see BMI data in Section 7), one can concentrate on the probabilistic properties of Y , ignoring the input X .

3.1. *Quantile function.* Equation (3.2) and an inspection of Figure 3 directly relate μ_x to Y .

COROLLARY 3.4 (Median of Y). For a location-scale Lambert W RV Y ,

$$\mathbb{P}(Y \leq \mu_x) = \mathbb{P}(X \leq \mu_x) \quad \text{for all } \gamma \in \mathbb{R}.$$

In particular, μ_x equals the median of Y , if X is symmetric.

PROOF. The transformation $H_\gamma(u) = u \exp(\gamma u)$ passes through $(0, 0)$ for all $\gamma \in \mathbb{R}$. Furthermore, $\exp(\gamma u) > 0$ for all γ and all $u \in \mathbb{R}$. Therefore,

$$\begin{aligned} \mathbb{P}(Y \leq \mu_x) &= \mathbb{P}(Z \leq 0) = \mathbb{P}(U \exp(\gamma U) \leq 0) = \mathbb{P}(U \leq 0) \\ &= \mathbb{P}(X \leq \mu_x). \end{aligned}$$

For symmetric input $\mathbb{P}(X \leq \mu_x) = \frac{1}{2}$, therefore, μ_x is the median of Y . \square

Corollary 3.4 not only gives a meaningful interpretation of the parameter μ_x for symmetric input, but the sample median of \mathbf{y} also yields a robust estimate of μ_x .

In general, the α -quantile y_α of Y satisfies

$$\begin{aligned} \alpha &\stackrel{!}{=} \mathbb{P}(Y \leq y_\alpha) = \mathbb{P}\left(U \exp(\gamma U) \leq \frac{y_\alpha - \mu_x}{\sigma_x}\right) \\ &= \mathbb{P}(U \exp(\gamma U) \leq z_\alpha). \end{aligned}$$

For $\gamma > 0$ ($\gamma < 0$ analogously) and $z_\alpha > 0$ the function $W_{\gamma,0}(\cdot)$ is bijective. Thus,

$$\mathbb{P}(U \exp(\gamma U) \leq z_\alpha) = \mathbb{P}(U \leq W_{\gamma,0}(z_\alpha)) = \mathbb{P}(X \leq W_{\gamma,0}(z_\alpha)\sigma_x + \mu_x)$$

and by definition of the α -quantile of X ,

$$(3.7) \quad x_\alpha = W_{\gamma,0}(z_\alpha)\sigma_x + \mu_x \quad \Leftrightarrow \quad z_\alpha = u_\alpha e^{\gamma u_\alpha},$$

where $u_\alpha = \frac{x_\alpha - \mu_x}{\sigma_x}$.

For $z_\alpha < 0$ and $\gamma > 0$, however, $W_\gamma(\cdot)$ is not bijective, thus, z_α cannot be computed explicitly as in (3.7), but must be obtained by solving the implicit equation

$$\alpha \stackrel{!}{=} F_X(W_{\gamma,0}(z_\alpha)\sigma_x + \mu_x) - F_X(W_{\gamma,-1}(z_\alpha)\sigma_x + \mu_x).$$

In either case, the α -quantile of Y equals $y_\alpha = z_\alpha\sigma_x + \mu_x$.

4. Gaussian input. The results so far hold for any continuous input RV. To get a better insight consider Gaussian input $U \sim \mathcal{N}(\mu_u, \sigma_u^2)$ as a special case; here $\beta = (\mu_u, \sigma_u)$. Its moment generating function $M_U(t)$ equals

$$\mathbb{E}(e^{tU}) = e^{t\mu_u + t^2/2\sigma_u^2} \quad \text{for all } t \in \mathbb{R}.$$

Therefore, noncentral moments of Z can be computed explicitly [see (2.4)] by

$$\psi_{(n)} = n^{-n} \frac{\partial^n}{\partial \gamma^n} \exp\left(\gamma n \mu_u + \gamma^2 \frac{n^2 \sigma_u^2}{2}\right).$$

In particular,

$$\begin{aligned} \mu_z &= (\mu_u + \gamma \sigma_u^2) e^{\gamma \mu_u + (\gamma^2/2)\sigma_u^2}, \\ \sigma_z^2 &= 2^{-2} (e^{2\gamma \mu_u + 2\gamma^2 \sigma_u^2} ((4\gamma \sigma_u^2 + 2\mu_u)^2 + 4\sigma_u^2)) - (\mu_u + \gamma \sigma_u^2)^2 e^{2\gamma \mu_u + \gamma^2 \sigma_u^2} \\ &= e^{2\gamma \mu_u + 2\gamma^2 \sigma_u^2} ((2\gamma \sigma_u^2 + \mu_u)^2 + \sigma_u^2) - (\mu_u + \gamma \sigma_u^2)^2 e^{2\gamma \mu_u + \gamma^2 \sigma_u^2}. \end{aligned}$$

As already mentioned in Section 2, this is an unstable system, in the sense that a small perturbation in (μ_u, σ_u) results in a completely different (μ_z, σ_z) for $\gamma \neq 0$.

In contrast, the central moments of a location-scale Lambert $W \times$ Gaussian RV Y with input $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ have a much simpler and stable form

$$(4.1) \quad \mu_y = \mu_x + \sigma_x \mathbb{E}(U e^{\gamma U}) = \mu_x + \sigma_x \gamma e^{\gamma^2/2},$$

since $U = (X - \mu_x)/\sigma_x \sim \mathcal{N}(0, 1)$. Using (4.1), the k th central moment of Y can be expressed by the k th central moment of $U e^{\gamma U}$,

$$\mathbb{E}(Y - \mu_y)^k = \sigma_x^k \mathbb{E}(U e^{\gamma U} - \gamma e^{\gamma^2/2})^k.$$

In particular,

$$(4.2) \quad \sigma_y^2 = \sigma_x^2 e^{\gamma^2} ((4\gamma^2 + 1)e^{\gamma^2} - \gamma^2),$$

which only depends on the input variance and the skewness parameter γ .

The main motive to introduce Lambert W RVs is to accurately model skewed data. The skewness coefficient of Y is defined as $\gamma_1(Y) := (\mathbb{E}(Y - \mu_y)^3)/\sigma_y^3$. Analogously, the kurtosis equals $\gamma_2(Y) := (\mathbb{E}(Y - \mu_y)^4)/\sigma_y^4$ and measures the thickness of tails of Y .

LEMMA 4.1. *For a location-scale Lambert $W \times$ Gaussian RV with input $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$,*

$$(4.3) \quad \gamma_1(\gamma) = \gamma \left[\frac{e^{3\gamma^2}(9 + 27\gamma^2) - e^{\gamma^2}(3 + 12\gamma^2) + 5\gamma^2}{(e^{\gamma^2}(1 + 4\gamma^2) - \gamma^2)^{3/2}} \right]$$

and

$$(4.4) \quad \gamma_2(\gamma) = \frac{e^{6\gamma^2}(3 + 96\gamma^2 + 256\gamma^4) - e^{3\gamma^2}(30\gamma^2 + 60\gamma^4 - 96\gamma^6) - 3\gamma^4}{(e^{\gamma^2}(1 + 4\gamma^2) - \gamma^2)^2}.$$

PROOF. Dividing the third and fourth derivative of the moment generating functions for a standard Gaussian U at $t = \gamma n$ with respect to γ by n^n gives

$$\frac{1}{3^3} \frac{d^3}{d\gamma^3} e^{9\gamma^2/2} = 9\gamma e^{9\gamma^2/2} + 27\gamma^3 e^{9\gamma^2/2} = 9\gamma e^{9\gamma^2/2} (1 + 3\gamma^2),$$

$$\frac{1}{4^4} \frac{d^4}{d\gamma^4} e^{8\gamma^2} = 3e^{8\gamma^2} + 96\gamma^2 e^{8\gamma^2} + 256\gamma^4 e^{8\gamma^2} = e^{8\gamma^2} (3 + 96\gamma^2 + 256\gamma^4).$$

The rest follows by expanding $\mathbb{E}(U e^{\gamma U} - \mathbb{E}U e^{\gamma U})^3$ and $\mathbb{E}(U e^{\gamma U} - \mathbb{E}U e^{\gamma U})^4$ via the binomial formula and using the above expressions. \square

As expected, the skewness coefficient is an odd function in γ with the same sign as γ . On the contrary, $\gamma_2(\gamma)$ is even. A first and second order Taylor approximation around $\gamma = 0$ yields $\gamma_1(\gamma) = 6\gamma + \mathcal{O}(\gamma^3)$ and $\gamma_2(\gamma) = 3 + 60\gamma^2 + \mathcal{O}(\gamma^4)$, respectively. Although γ can take any value in \mathbb{R} , in practice, it rarely exceeds

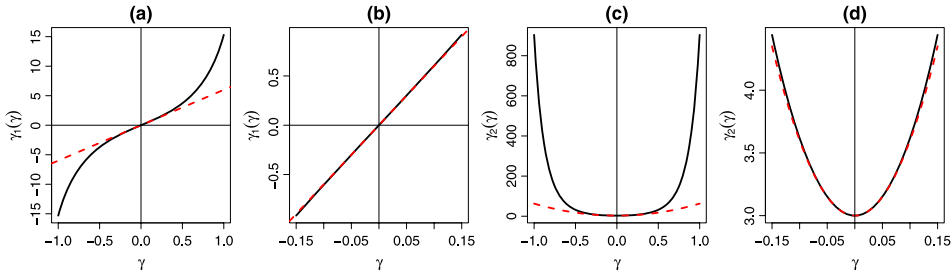


FIG. 5. Pearson skewness (a) and kurtosis (c) coefficient for $\gamma \in [-1, 1]$ and its first order Taylor approximation (dashed line); (b) and (d): zoom to the interval $[-0.15, 0.15]$.

0.15 in absolute value. In this interval the Taylor approximation is almost indistinguishable from the true function (Figure 5).

This first order approximation to $\gamma_1(\gamma)$ offers a rule of thumb

$$(4.5) \quad \widehat{\gamma}_{\text{Taylor}} := \frac{\widehat{\gamma}_1(\mathbf{y})}{6},$$

which can be used as a starting value for better algorithms such as IGMM and MLE (see Section 5).

COROLLARY 4.2. *The skewness and kurtosis coefficient are unbounded for $\gamma \rightarrow \pm\infty$, that is,*

$$\lim_{\gamma \rightarrow \pm\infty} \gamma_1(\gamma) = \pm\infty \quad \text{and} \quad \lim_{\gamma \rightarrow \pm\infty} \gamma_2(\gamma) = \infty.$$

PROOF. Omitting $-\gamma^2$ in the denominator and $5\gamma^2$ in the numerator of the skewness coefficient can be bounded from below

$$\begin{aligned} \frac{9e^{3\gamma^2} + 27\gamma^2 e^{3\gamma^2} - 3e^{\gamma^2} - 12\gamma^2 e^{\gamma^2} + 5\gamma^2}{(e^{\gamma^2} + 4\gamma^2 e^{\gamma^2} - \gamma^2)^{3/2}} &\geq \frac{e^{3\gamma^2}(9 + 27\gamma^2) - e^{\gamma^2}(3 + 12\gamma^2)}{e^{(3/2)\gamma^2}(1 + 4\gamma^2)^{3/2}} \\ &= e^{(3/2)\gamma^2} \frac{9 + 27\gamma^2}{(1 + 4\gamma^2)^{3/2}} \\ &\quad - e^{-\gamma^2/2} \frac{3 + 12\gamma^2}{(1 + 4\gamma^2)^{3/2}}. \end{aligned}$$

As the exponential function dominates rational functions, the first term tends to ∞ , whereas the second one goes to 0 for γ to ∞ .

In case of the kurtosis coefficient, the term $e^{6\gamma^2}$ in the numerator dominates all other terms for large γ and thus determines the asymptotic behavior of $\gamma_2(\gamma)$ for γ to $\pm\infty$. \square

This result shows that the Lambert $W \times$ Gaussian distributions can be used to model a larger variety of skewed data than a skew-normal distribution, since its skewness coefficient is restricted to the interval $(-0.995, 0.995)$ [Azzalini (1985)].

5. Parameter estimation. For a sample of N independent identically distributed (i.i.d.) observations $\mathbf{y} = (y_1, \dots, y_N)$, which presumably originates from transformation (2.6), $\theta = (\boldsymbol{\beta}, \gamma)$ has to be estimated from the data. In addition to the commonly used maximum likelihood estimator (MLE) for θ , I also present a method of moments estimator for τ that builds on the input/output relation in Figure 1.

5.1. *Maximum likelihood estimation.* The log-likelihood function in the i.i.d. case equals

$$(5.1) \quad \ell(\theta \mid \mathbf{y}) = \sum_{i=1}^N \log g_Y(y_i \mid \theta),$$

where $g_Y(\cdot \mid \theta)$ is the p.d.f. of Y —see (3.3). The MLE is that $\theta = (\boldsymbol{\beta}, \gamma)$ which maximizes the log-likelihood

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \ell(\theta \mid \mathbf{y}).$$

Since $g_Y(y_i \mid \theta)$ is a function of $f_X(x_i \mid \boldsymbol{\beta})$, the MLE depends on the specification of the input density. In general, this multivariate, nonlinear optimization problem must be carried out by numerical methods, as the two branches of $W(z)$ for $y \leq \mu_x$ do not allow any further simplification.

For (scale) Lambert $W \times F_X$ with support in $(0, \infty)$ and $\gamma \geq 0$, however, $g_Y(y \mid \boldsymbol{\beta}, \gamma) = f_X(W_{\gamma,0}(\frac{y}{\sigma_x}) \cdot \sigma_x \mid \boldsymbol{\beta}) \cdot W'_0(\gamma \frac{y}{\sigma_x})$ (Corollary 3.3). Thus, (5.1) can be rewritten as

$$(5.2) \quad \ell(\boldsymbol{\beta}, \gamma \mid \mathbf{y}) = \ell(\boldsymbol{\beta} \mid \mathbf{x}_{(0,\sigma_x,\gamma)}) + \sum_{i=1}^N \log W'_0\left(\gamma \frac{y_i}{\sigma_x}\right),$$

where

$$(5.3) \quad \ell(\boldsymbol{\beta} \mid \mathbf{x}_{(0,\sigma_x,\gamma)}) = \sum_{i=1}^N \log f_X\left(W_{\gamma,0}\left(\frac{y_i}{\sigma_x}\right) \cdot \sigma_x \mid \boldsymbol{\beta}\right)$$

is the log-likelihood of the back-transformed data $\mathbf{x}_{(0,\sigma_x,\gamma)} = W_{\gamma,0}(\frac{\mathbf{y}}{\sigma_x}) \cdot \sigma_x$ [no (\cdot) since the inverse is unique in this case]. Note that $W'_0(\gamma \frac{y_i}{\sigma_x})$ only depends on $\sigma_x(\boldsymbol{\beta})$ (and γ), but not necessarily on every coordinate of $\boldsymbol{\beta}$.

The equivalence (5.2) shows the relation between the exact MLE $(\hat{\boldsymbol{\beta}}, \hat{\gamma})$ based on \mathbf{y} and the approximate MLE $\hat{\boldsymbol{\beta}}$ based on $\mathbf{x}_{(0,\sigma_x,\gamma)}$: if we would know σ_x and γ beforehand, then we could just back-transform \mathbf{y} to $\mathbf{x}_{(0,\sigma_x,\gamma)}$ and compute $\hat{\boldsymbol{\beta}}_{\text{MLE}}$

based on $\mathbf{x}_{(0, \sigma_x, \gamma)}$ [maximize (5.3)]; however, in practice, σ_x and γ have to be estimated from \mathbf{y} and this uncertainty enters the log-likelihood (5.2) by the additional term $\sum_{i=1}^n \log W_0'(\gamma \frac{y_i}{\sigma_x})$.

For $z > 0$ it can be easily shown that $W'(z) = \frac{W(z)}{z(1+W(z))} > 0$ as well as $W'(z) < 1$ since $W'(0) = 1$ and $W''(z) = -W'(z) \exp(-W(z)) \frac{W(z)+2}{(W(z)+1)^2} < 0$. Hence, $\sum_{i=1}^n \log W_0'(\gamma \frac{y_i}{\sigma_x}) < 0$ for $\gamma > 0$ and can be thought of as a penalty for transforming \mathbf{y} to the “nicer” $\mathbf{x}_{(0, \hat{\sigma}_x, \hat{\gamma})}$ with estimated parameters: the larger γ , the bigger the penalty on the log-likelihood $\ell(\boldsymbol{\beta} \mid \mathbf{x}_{(0, \sigma_x, \gamma)})$ of the “nice” back-transformed data, since $W''(z) < 0$.

Parameter-dependent support. For location-scale Lambert $W \times F$ RVs the support of $g_Y(y)$ depends on $\tau = \tau(\theta)$ and therefore violates a crucial assumption of most results related to (asymptotic) properties of the MLE. Only for $\gamma = 0$ the support of $g_Y(y) = f_X(y)$ does not depend on θ . For $X \sim \mathcal{N}(0, 1)$ it can be shown that the Fisher information matrix $I_N(\gamma) = -\mathbb{E}(\frac{d^2}{d\gamma^2} \ell(\theta \mid \mathbf{y})) = 8N$. Hence, for the symmetric Gaussian case $\sqrt{N} \hat{\gamma}_{MLE} \rightarrow \mathcal{N}(0, \frac{1}{8})$. Simulations in Section 6 confirm this asymptotic result and suggest that also for the general Gaussian case $\hat{\theta}_{MLE}$ is well behaved, that is, it is \sqrt{N} -consistent and asymptotically efficient.

A theoretical analysis of the asymptotic behavior of the MLE for $\gamma \neq 0$ is beyond the scope of this study, but simulations show that also for parameter dependent support $\hat{\theta}_{MLE}$ is an unbiased estimator with root mean square errors comparable to the $\gamma = 0$ case.

5.2. Iterative generalized method of moments (IGMM). A disadvantage of the MLE is the mandatory a-priori specification of the input distribution. In practice, however, it is rarely known what kind of distribution is a good fit to the data, even more so if the data is transformed via a nonlinear transformation. Thus, here I present an iterative method to estimate the optimal inverse-transformation (2.8) by estimating τ directly, instead of estimating θ and then computing $\tau(\hat{\theta})$. This method builds on the input/output aspect and only relies upon the specification of the theoretical skewness of X .

The proposed estimator for τ works as follows (see below for a more detailed discussion):

- (1) set starting values $\tau^{(0)} = \tau_0$. Set $k = 0$;
- (2) assume $\mu_x^{(k)}$ and $\sigma_x^{(k)}$ are known and estimate γ from $\mathbf{z}^{(k)} = \frac{\mathbf{y} - \mu_x^{(k)}}{\sigma_x^{(k)}}$ to obtain $\gamma^{(k+1)}$ (Algorithm 2);
- (3) assume $\gamma^{(k+1)}$ is known and get estimates $\mu_x^{(k+1)}$ and $\sigma_x^{(k+1)}$ from the back-transformed data $\hat{\mathbf{x}}_{(\mu_x^{(k)}, \sigma_x^{(k)}, \gamma^{(k+1)})}$ (Algorithm 3). Set $k = k + 1$;
- (4) iterate between (5.2) and (5.2) until convergence of the sequence $\tau^{(k)}$.

Algorithm 2 Find optimal γ : function `gamma_GMM(\cdot)` in the `LambertW` package.

Input: standardized data vector \mathbf{z} ; theoretical skewness $\gamma_1(X)$.

Output: $\hat{\gamma}_{\text{GMM}}$ as in (5.4).

- 1: Compute lower and upper bound for γ : $lb = -\frac{1}{\exp(1)\max(\mathbf{z})}$ and $ub = -\frac{1}{\exp(1)\min(\mathbf{z})}$.
 - 2: $\hat{\gamma}_{\text{GMM}} = \arg \min_{\gamma} \|\hat{\gamma}_1(\hat{\mathbf{u}}) - \gamma_1(X)\|$ where $\hat{\mathbf{u}} = W_{\gamma}(\mathbf{z})$ subject to $\gamma \in [lb, ub]$.
 - 3: **return** $\hat{\gamma}_{\text{GMM}}$.
-

For a moment assume that μ_x and σ_x are known and only γ has to be estimated. Since μ_x and σ_x are known, we can consider $\mathbf{z} = \frac{\mathbf{y} - \mu_x}{\sigma_x}$. A natural choice for γ is the one that results in back-transformed data $\hat{\mathbf{u}}_{\gamma} = W_{\gamma}(\mathbf{z})$ with sample skewness equal to the theoretical skewness of U , which equals the theoretical skewness of X . Formally,

$$(5.4) \quad \hat{\gamma}_{\text{GMM}} = \arg \min_{\gamma} \|\gamma_1(X) - \hat{\gamma}_1(\hat{\mathbf{u}}_{\gamma})\|,$$

where $\|\cdot\|$ is a proper norm in \mathbb{R} , for example, $\|s\| = s^2$ or $\|s\| = |s|$.

Discussion of Algorithm 2. For example, let \mathbf{y} be positively skewed data, $\hat{\gamma}_1(\mathbf{y}) > 0$, and the input \mathbf{x} causing the observed \mathbf{y} is assumed/known to be symmet-

Algorithm 3 Iterative generalized method of moments: function `IGMM(\cdot)` in the `LambertW` package.

Input: data vector \mathbf{y} ; tolerance level tol ; theoretical skewness $\gamma_1(X)$.

Output: IGMM parameter estimate $\hat{\tau}_{\text{IGMM}} = (\hat{\mu}_x, \hat{\sigma}_x, \hat{\gamma})$.

- 1: Set $\tau^{(-1)} = (0, 0, 0)$.
 - 2: Starting values: $\tau^{(0)} = (\mu_x^{(0)}, \sigma_x^{(0)}, \gamma^{(0)})$, where $\mu_x^{(0)} = \tilde{\mathbf{y}}$ and $\sigma_x^{(0)} = \bar{\sigma}_y$ are the sample median and standard deviation of \mathbf{y} , respectively. $\gamma^{(0)} = \frac{\hat{\gamma}_1(\mathbf{y}) - \gamma_1(X)}{6} \rightarrow$ see (4.5) for details.
 - 3: $k = 0$.
 - 4: **while** $\|\tau^{(k)} - \tau^{(k-1)}\| > tol$ **do**
 - 5: $\mathbf{z}^{(k)} = (\mathbf{y} - \mu_x^{(k)})/\sigma_x^{(k)}$,
 - 6: Pass $\mathbf{z}^{(k)}$ to Algorithm 2 $\rightarrow \gamma^{(k+1)}$,
 - 7: back-transform $\mathbf{z}^{(k)}$ to $\mathbf{u}^{(k+1)} = W_{\gamma^{(k+1)}}(\mathbf{z}^{(k)})$; compute $\mathbf{x}^{(k+1)} = \mathbf{u}^{(k+1)}\sigma_x^{(k)} + \mu_x^{(k)}$,
 - 8: update parameters: $\mu_x^{(k+1)} = \bar{\mathbf{x}}_{k+1}$ and $\sigma_x^{(k+1)} = \hat{\sigma}_{x_{k+1}}$,
 - 9: $\tau^{(k+1)} = (\mu_x^{(k+1)}, \sigma_x^{(k+1)}, \gamma^{(k+1)})$,
 - 10: $k = k + 1$.
 - 11: **return** $\tau_{\text{IGMM}} = \tau^{(k)}$.
-

ric, thus, $\gamma_1(X) = 0$. By the nature of transformation $H_\gamma(u)$, the skewness parameter γ must be also positive and the Taylor approximation of $\gamma_1(\gamma)$ for Gaussian input [see (4.5)] gives a good initial estimate $\gamma_0 = \widehat{\gamma}_1(\mathbf{y})/6 > 0$. In the same way as the mapping $u \mapsto u \exp(\gamma u)$ introduces skewness, the inverse transformation $W_\gamma(\mathbf{z})$ results in less positively skewed $\widehat{\mathbf{u}}_\gamma$ due to the curvature in $W_\gamma(\cdot)$ (see Figure 3). As the initial guess γ_0 rarely gives exactly symmetric input, Algorithm 2 searches for a γ such that the empirical skewness of $\widehat{\mathbf{u}}_\gamma$ is as close as possible to the “true” skewness $\gamma_1(X)$.

There are natural bounds for γ to guarantee the observability of \mathbf{y} , for example, a γ too large makes large negative observations in \mathbf{y} impossible (due to the minimum at $z = -1/e$; see Figure 3). However, since \mathbf{y} has actually been observed, the search space for γ must be limited to the interval $O_{\mathbf{z}} := [-\frac{1}{\exp(1)\max(\mathbf{z})}, -\frac{1}{\exp(1)\min(\mathbf{z})}]$. If there exists a $\tilde{\gamma} \in O_{\mathbf{z}}$ such that $\widehat{\gamma}_1(\widehat{\mathbf{u}}_{\tilde{\gamma}}) = \gamma_1(X)$, then Algorithm 2 will return $\widehat{\gamma} = \tilde{\gamma}$ due to the monotonically increasing curvature of $H_\gamma(u)$ and $W_\gamma(z)$ respectively; if there is no such $\tilde{\gamma} \in O_{\mathbf{z}}$, then Algorithm 2 returns either the lower or upper bound of $O_{\mathbf{z}}$, depending on whether \mathbf{z} is negatively or positively skewed.

This univariate minimization problem with constraints can be carried out by standard optimization algorithms.

In practice, μ_x and σ_x are rarely known but also have to be estimated from the data. As \mathbf{y} is shifted and scaled *ahead of* the back-transformation $W_{\gamma,0}(\cdot)$, the initial choice of μ_x and σ_x affects the optimal choice of γ . Therefore, the optimal triple $(\widehat{\mu}_x, \widehat{\sigma}_x, \widehat{\gamma})$ must be obtained iteratively.

Discussion of Algorithm 3. Algorithm 3 first computes $\mathbf{z}^{(k)} = (\mathbf{y} - \mu_x^{(k)})/\sigma_x^{(k)}$ using $\mu_x^{(k)}$ and $\sigma_x^{(k)}$ from the previous step. This normalized output can then be passed to Algorithm 2 to obtain an updated $\gamma^{(k+1)} := \widehat{\gamma}_{\text{GMM}}$. Using this new $\gamma^{(k+1)}$, one can back-transform $\mathbf{z}^{(k)}$ to the presumably zero-mean, unit-variance input $\mathbf{u}^{(k+1)} = W_{\gamma^{(k+1)}}(\mathbf{z}^{(k)})$. Herewith we can obtain a better approximation to the “true” latent \mathbf{x} by $\mathbf{x}^{(k+1)} = \mathbf{u}^{(k+1)}\sigma_x^{(k)} + \mu_x^{(k)}$. However, $\gamma^{(k+1)}$ —and therefore $\mathbf{x}^{(k+1)}$ —has been obtained using $\mu_x^{(k)}$ and $\sigma_x^{(k)}$ which are not necessarily the most accurate estimates in light of the updated approximation $\widehat{\mathbf{x}}_{(\mu_x^{(k)}, \sigma_x^{(k)}, \gamma^{(k+1)})}$. Thus, Algorithm 3 computes new estimates $\mu_x^{(k+1)}$ and $\sigma_x^{(k+1)}$ by the sample mean and standard deviation of $\widehat{\mathbf{x}}_{(\mu_x^{(k)}, \sigma_x^{(k)}, \gamma^{(k+1)})}$, and starts another iteration by passing the updated normalized output $\mathbf{z}^{(k+1)} = \frac{\mathbf{y} - \mu_x^{(k+1)}}{\sigma_x^{(k+1)}}$ to Algorithm 2 to obtain a new $\gamma^{(k+2)}$.

The algorithm returns the optimal $\widehat{\tau}_{\text{IGMM}}$ once the estimated parameter triple does not change anymore from one iteration to the next, that is, if $\|\tau^{(k)} - \tau^{(k+1)}\| < \text{tol}$.

A great advantage of the IGMM estimator is that it does not require any further specification of the input except its skewness. For example, no matter if the input

is normally, student-*t*, Laplace or uniformly distributed, the IGMM estimator finds a τ that gives symmetric $\widehat{\mathbf{x}}_{\widehat{\tau}}$ independent of the particular choice of (symmetric) $F_X(\cdot)$.

A disadvantage of IGMM from a probabilistic point of view is its determination. In general, Algorithm 3 will lead to back-transformed data with sample skewness identical to $\gamma_1(X)$ and so no stochastic element remains in the nature of the estimator.⁵ Note that IGMM does not provide an estimate of $\boldsymbol{\beta}$ (except for Gaussian input); if necessary, an estimate of $\boldsymbol{\beta}$ must be obtained in a separate step, for example, by estimating $\boldsymbol{\beta}$ from the back-transformed data $\widehat{\mathbf{x}}_{\widehat{\tau}}$. However, in general, $\widehat{\boldsymbol{\beta}}_{\text{MLE}}$ estimated only from $\widehat{\mathbf{x}}_{\widehat{\tau}}$ is (slightly) different from $\widehat{\boldsymbol{\beta}}_{\text{MLE}}$ using Lambert *W* MLE on the original data \mathbf{y} : in the first case $\widehat{\tau}$ is assumed to be known and fixed, whereas in the second case $\boldsymbol{\beta}$ and τ are estimated jointly [see (5.2)].

The underlying input data $\mathbf{x} = (x_1, \dots, x_N)$ can be approximated via Algorithm 1 using $\widehat{\tau}_{\text{IGMM}}$. The so obtained $\widehat{\mathbf{x}}_{\widehat{\tau}_{\text{IGMM}}}$ may then be used to check if X has characteristics of a known parametric distribution $F_X(x | \boldsymbol{\beta})$, and thus is an easy, but heuristic check if \mathbf{y} follows a particular Lambert *W* $\times F_X$ distribution. However, such a test can only serve as a rule of thumb for various reasons: (i) $\widehat{\tau} \neq \tau$, thus tests are too optimistic as $\widehat{\mathbf{x}}_{\widehat{\tau}}$ will have “nicer” properties regarding F_X than the true \mathbf{x} would have; (ii) ignoring the nonprincipal branch alters the sample distribution of the input—putting no observations to the far left (or right): not so much of a problem for small γ , the distribution can be truncated considerably for large γ . For Gaussian input various tests are available [Jarque–Bera, Shapiro–Wilk, among others; see Thode (2002)], for other distributions a Kolmogorov–Smirnov test can be used.⁶

5.2.1. *Gaussian IGMM.* For Gaussian X the system of equations

$$(5.5) \quad \mu_y(\gamma) = \mu_x + \sigma_x \gamma e^{\gamma^2/2},$$

$$(5.6) \quad \sigma_y^2(\gamma) = \sigma_x^2 e^{\gamma^2} ((4\gamma^2 + 1)e^{\gamma^2} - \gamma^2)$$

has a unique solution for (μ_x, σ_x) . Given $\widehat{\gamma}_{\text{IGMM}}$ and the sample moments $\overline{\mu}_y$ and $\overline{\sigma}_y$, the input parameters μ_x and σ_x^2 can be obtained by

$$(5.7) \quad \widehat{\sigma}_x^2(\widehat{\gamma}_{\text{IGMM}}) = \frac{\overline{\sigma}_y^2}{e^{\widehat{\gamma}_{\text{IGMM}}^2} ((4\widehat{\gamma}_{\text{IGMM}}^2 + 1)e^{\widehat{\gamma}_{\text{IGMM}}^2} - \widehat{\gamma}_{\text{IGMM}}^2)},$$

$$(5.8) \quad \widehat{\mu}_x(\widehat{\gamma}_{\text{IGMM}}) = \overline{\mu}_y - \widehat{\sigma}_x^2(\widehat{\gamma}_{\text{IGMM}}) \widehat{\gamma}_{\text{IGMM}} e^{\widehat{\gamma}_{\text{IGMM}}^2/2}.$$

⁵If $\gamma_1(X)$ depends on one or more parameters of the distribution of X (e.g., Gamma), then the IGMM algorithm must be adapted to this very problem.

⁶If the data does not represent an independent sample (as usual for financial data), then critical values of several test statistics need not be valid anymore and adapted tests should be used [see Weiss (1978)].

Hence, line 8 of Algorithm 3 can be altered to

$$(5.9) \quad \text{8b: } \mu_x^{(k+1)} = \hat{\mu}_x(\bar{\mu}_y, \bar{\sigma}_y, \gamma_{k+1}) \text{ and } \sigma_x^{(k+1)} = \hat{\sigma}_x(\bar{\sigma}_y, \gamma_{k+1}), \text{ given by (5.7) and (5.8).}$$

Even though this simplification would lead to a faster estimation of τ , it is mostly of theoretical interest, as it cannot be guaranteed that X indeed is Gaussian; the more general Algorithm 3 should be used in practice.⁷

6. Simulations. Although the c.d.f., p.d.f. and moments of a Lambert W RVs are nontrivial expressions, their simulation is straightforward (Algorithm 4).

This section explores the finite-sample properties of estimators for $\theta = (\mu_x, \sigma_x, \gamma)$ under Gaussian input $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$.⁸ In particular, conventional Gaussian MLE (estimation of μ_y and σ_y only; $\gamma \equiv 0$), IGMM and Lambert $W \times$ Gaussian MLE, and—for a skew competitor—the skew-normal MLE⁹ are studied. Whereas a comparison of accuracy and efficiency in $\hat{\gamma}$ does not make sense, it is meaningful to analyze $\hat{\mu}_y$ and $\hat{\sigma}_y$ of skew-normal versus Lambert $W \times$ Gaussian MLE.

Scenarios. Each estimator is applied to 3 kinds of simulated data sets for 4 different sample sizes of $N = 50, 100, 250$ and $1,000$:

$\gamma = 0$: Data is sampled from a symmetric RV $Y = X \sim \mathcal{N}(0, 1)$. Does additional estimation of γ affect the properties of $\hat{\mu}_y$ or $\hat{\sigma}_y$?

$\gamma = -0.05$: A typical value for financial data, such as the LATAM returns introduced in Section 1.

Algorithm 4 Random sample generation: function `rLambertW(\cdot)` in the `LambertW` package.

Input: number of observations n ; parameter vector β ; specification of the input distribution $F_X(x | \beta)$; skewness parameter γ .

Output: random sample (y_1, \dots, y_n) of a Lambert $W \times F$ RV.

- 1: Simulate n samples $\mathbf{x} = (x_1, \dots, x_n) \sim F_X(x | \beta)$.
 - 2: Compute $\mu_x(\beta)$ and $\sigma_x(\beta)$ given the type of Lambert $W \times F$ distribution (noncentral, nonscale; scale; location-scale).
 - 3: $\mathbf{u} = (\mathbf{x} - \mu_x(\beta)) / \sigma_x(\beta)$.
 - 4: $\mathbf{z} = \mathbf{u} \exp(\gamma \mathbf{u})$.
 - 5: **return** $\mathbf{y} = \mathbf{z} \sigma_x(\beta) + \mu_x(\beta)$.
-

⁷All numerical estimates $\hat{\tau}_{\text{IGMM}}$ reported in Section 6 were obtained using the more general algorithm with line 8, not 8b.

⁸For the special case of Gaussian input $\tau \equiv \theta$, thus, IGMM estimates $\hat{\tau}_{\text{IGMM}} = \hat{\theta}_{\text{IGMM}}$ can be compared directly to $\hat{\theta}_{\text{MLE}}$.

⁹Function `sn.mle` in the R package `sn`.

$\gamma = 0.3$: This large value reveals the importance of the two branches of the Lambert W function. How does the skew-normal MLE handle extremely skewed data [$\gamma(0.3) = 1.9397$]?

Simulations are based on $n = 1,000$ replications. The input mean μ_x and standard deviation σ_x are chosen such that the observed RV has $\mu_y(\gamma) = 0$ and $\sigma_y(\gamma) = 1$ for all γ . These functional relations can be obtained by (5.7) and (5.8). For IGMM the tolerance level was set to $tol = 10^{-6}$ and the Euclidean norm was used.

REMARK 6.1. The Gaussian and skew-normal MLE estimate the mean and standard deviation of Y . Both Lambert W methods estimate the mean and standard deviation of the latent variable X plus the skewness parameter γ . Thus, for a meaningful comparison the implied estimates $\hat{\sigma}_y(\hat{\mu}_x, \hat{\gamma})$ and $\hat{\mu}_y(\hat{\mu}_x, \hat{\sigma}_x, \hat{\gamma})$ given by (5.5) and (5.6) are reported below.

6.1. *Symmetric data: $\gamma = 0$.* This parameter choice investigates if imposing the Lambert W framework, even though its use is superfluous, causes a quality loss in the estimation. Furthermore, critical values can be obtained for the finite sample behavior of $\hat{\gamma}$ under the null hypothesis of a symmetric distribution.

Table 2 displays the bias and root mean square error (RMSE) of $\hat{\theta}$. Not only are all estimators unbiased, but they also have essentially equal RMSE for $\hat{\mu}_y$ and $\hat{\sigma}_y$.

TABLE 2
Bias and RMSE of $\hat{\theta}$ for $\gamma = 0$ and $X \sim N(0, 1)$

	N	Bias			RMSE $\cdot \sqrt{N}$		
		$\gamma = 0$	$\mu_y = 0$	$\sigma_y = 1$	$\gamma = 0$	$\mu_y = 0$	$\sigma_y = 1$
Gaussian ML	50	0.0000	0.0054	-0.0175	0.0000	0.9943	0.7053
	100	0.0000	0.0016	-0.0084	0.0000	0.9812	0.7410
	250	0.0000	-0.0029	-0.0009	0.0000	0.9997	0.6917
	1,000	0.0000	0.0005	-0.0013	0.0000	0.9788	0.7105
IGMM	50	-0.0015	0.0054	-0.0060	0.4567	0.9945	0.7059
	100	-0.0012	0.0015	-0.0030	0.4368	0.9813	0.7405
	250	0.0001	-0.0017	-0.0009	0.4210	0.9997	0.6919
	1,000	0.0003	0.0005	-0.0008	0.4014	0.9788	0.7102
Lambert W ML	50	-0.0013	0.0054	-0.0126	0.5144	0.9951	0.7210
	100	-0.0013	0.0016	-0.0072	0.4670	0.9813	0.7407
	250	0.0002	-0.0017	-0.0027	0.4333	0.9997	0.6922
	1,000	0.0003	0.0005	-0.0012	0.4039	0.9788	0.7106
Skew-normal ML	50	NA	0.0052	-0.0135	NA	0.9928	0.7149
	100	NA	0.0015	-0.0073	NA	0.9821	0.7409
	250	NA	-0.0018	-0.0027	NA	1.0004	0.6925
	1,000	NA	0.0000	-0.0013	NA	0.9788	0.7105

It is well known that the Gaussian MLE of σ_x is only asymptotically unbiased, but for small samples it underestimates the standard deviation, whereas a method of moments estimator such as IGMM does not have that problem (see $N = 50$). For $\hat{\gamma}$ the IGMM estimator has slightly smaller RMSE than MLE for small N ; for large N the difference disappears. This can also be explained by an only asymptotically unbiased MLE for σ_x , and the functional relation (4.2) of γ , σ_x and σ_y .

Overall, estimating γ has no effect on the quality of the remaining parameter estimates, if the data comes from a truly (symmetric) Gaussian distribution. A Shapiro Wilk Gaussianity test on the $n = 1,000$ estimates of $\hat{\gamma}_{IGMM}$ and $\hat{\gamma}_{MLE}$ gives p -values of 68.91% and 68.25%, respectively ($N = 1,000$), and thus confirms the asymptotic normality of $\hat{\gamma}$ as stated in Section 5.1.

6.2. *Slightly skewed data: $\gamma = -0.05$.* This choice of γ is motivated by real world data—in particular, asset returns typically exhibit slightly negative skewness [$\gamma_1(-0.05) = -0.30063$].

Table 3 presents the effect of ignoring small asymmetry in data. Gaussian MLE is by definition biased for γ , but $\hat{\mu}_y$ and $\hat{\sigma}_y$ are still good estimates. Neither IGMM nor Lambert W MLE gives biased $\hat{\theta}$, but the RMSE of $\hat{\sigma}_y$ increases for all estimators and all sample sizes. Again IGMM presents smaller RMSE for $\hat{\gamma}$ than MLE for small N , but not for large N —for the same reason as in the $\gamma = 0$ case. Notably,

TABLE 3
Bias and RMSE of $\hat{\theta}$ for $\gamma = -0.05$ and $X \sim N(\mu_x(\gamma), \sigma_x^2(\gamma))$

	N	Bias			RMSE $\cdot \sqrt{N}$		
		$\gamma = -0.05$	$\mu_y = 0$	$\sigma_y = 1$	$\gamma = -0.05$	$\mu_y = 0$	$\sigma_y = 1$
Gaussian ML	50	0.0500	0.0057	-0.0176	0.3536	0.9952	0.7350
	100	0.0500	0.0016	-0.0083	0.5000	0.9826	0.7741
	250	0.0500	-0.0029	-0.0009	0.7906	0.9981	0.7095
	1,000	0.0500	0.0005	-0.0014	1.5811	0.9781	0.7281
IGMM	50	-0.0008	0.0046	-0.0057	0.4582	0.9954	0.7410
	100	-0.0008	0.0011	-0.0026	0.4389	0.9828	0.7753
	250	0.0002	-0.0019	-0.0007	0.4189	0.9982	0.7102
	1,000	0.0003	0.0005	-0.0009	0.3986	0.9780	0.7276
Lambert W ML	50	-0.0043	0.0052	-0.0116	0.5113	0.9961	0.7570
	100	-0.0029	0.0015	-0.0062	0.4701	0.9829	0.7802
	250	-0.0006	-0.0017	-0.0024	0.4282	0.9981	0.7114
	1,000	0.0001	0.0005	-0.0013	0.3992	0.9781	0.7284
Skew-normal ML	50	NA	0.0073	-0.0136	NA	1.0011	0.7490
	100	NA	0.0026	-0.0067	NA	0.9834	0.7811
	250	NA	-0.0014	-0.0025	NA	0.9990	0.7109
	1,000	NA	0.0000	-0.0012	NA	0.9796	0.7281

the skew-normal MLE for μ_y and σ_y is also unbiased and has the same RMSE as the Lambert W and Gaussian competitors, even though the true distribution is a Lambert $W \times$ Gaussian, not a skew-normal.

6.3. *Extremely skewed data: $\gamma = 0.3$.* In this case, the Lambert W MLE should work better than the skew-normal MLE, since the skewness coefficient $\gamma_1(0.3) = 1.9397$ lies outside the theoretically possible values of skew-normal distributions. Furthermore, the nonprincipal branch of the Lambert W function becomes more important as $p_{-1} \approx 4.29 \cdot 10^{-4}$, so the Lambert W MLE should also outperform IGMM, which ignores the nonprincipal solution.

Only the skew-normal MLE fails to provide accurate estimates of location and scale for heavily skewed data sets; all other estimators are practically unbiased (Table 4). The RMSE for $\hat{\sigma}_y$ almost doubled compared to the symmetric case, and for Gaussian as well as skew-normal MLE it is increasing with sample size instead of decreasing. While $\hat{\gamma}_{\text{IGMM}}$ has less bias, $\hat{\gamma}_{\text{MLE}}$ has a much smaller RMSE: not ignoring the nonprincipal branch more than compensates the finite sample bias in $\hat{\sigma}_x$. Surprisingly, the RMSE for $\hat{\gamma}$ has diminished by about 35% over all sample sizes compared to the symmetric case.

Discussion. Estimation of μ_y is unaffected by the value of γ ; the quality of $\hat{\sigma}_y$, however, depends on γ : the larger γ , the greater the RMSE of $\hat{\sigma}_y$. For $\gamma = 0$ the

TABLE 4
Bias and RMSE of $\hat{\theta}$ for $\gamma = 0.3$ and $X \sim N(\mu_x(\gamma), \sigma_x^2(\gamma))$

	N	Bias			RMSE $\cdot \sqrt{N}$		
		$\gamma = 0.3$	$\mu_y = 0$	$\sigma_y = 1$	$\gamma = 0.3$	$\mu_y = 0$	$\sigma_y = 1$
Gaussian ML	50	-0.3000	0.0029	-0.0336	2.1213	0.9851	1.2941
	100	-0.3000	0.0006	-0.0194	3.0000	0.9863	1.3957
	250	-0.3000	-0.0076	-0.0056	4.7434	1.0045	1.4499
	1,000	-0.3000	0.0002	-0.0013	9.4868	0.9883	1.4954
IGMM	50	-0.0076	0.0081	-0.0057	0.4374	0.9917	1.2417
	100	-0.0055	0.0028	-0.0063	0.4005	0.9853	1.2440
	250	-0.0032	-0.0012	-0.0056	0.3647	1.0009	1.2204
	1,000	-0.0026	-0.0003	-0.0049	0.3197	0.9820	1.1992
Lambert W ML	50	0.0180	0.0221	0.0266	0.3844	1.0152	1.2385
	100	0.0115	0.0131	0.0168	0.3241	1.0095	1.2218
	250	0.0055	0.0053	0.0054	0.2747	1.0102	1.1535
	1,000	0.0000	0.0021	-0.0021	0.2349	0.9818	1.1383
Skew-normal ML	50	NA	0.0695	-0.0938	NA	1.3638	1.1965
	100	NA	0.0558	-0.0834	NA	1.3508	1.3182
	250	NA	0.0520	-0.0748	NA	1.4865	1.5577
	1,000	NA	0.0560	-0.0704	NA	2.1588	2.4585

TABLE 5

Average number of iterations ($tol = 10^{-6}$): (top) IGMM Algorithm 3 including the iterations in Algorithm 2; (bottom) IGMM only (not counting iterations in Algorithm 2). (left) Gaussian input; (right) student- t input with $\nu = 4$ degrees of freedom. Based on $n = 1,000$ replications

N	γ					
	0	-0.05	0.3	0, $\nu = 4$	-0.05, $\nu = 4$	0.3, $\nu = 4$
50	8.39	10.24	34.65	15.82	16.76	26.91
100	6.05	8.16	35.01	17.52	19.12	21.45
250	4.37	6.45	27.51	17.83	21.34	13.23
1,000	3.58	4.96	18.43	17.20	24.58	6.49
50	4.43	4.60	6.24	4.56	4.64	6.23
100	4.10	4.44	6.44	4.46	4.44	5.78
250	3.90	4.26	5.92	4.15	4.19	5.45
1,000	3.58	4.11	5.91	3.78	4.04	5.36

Lambert W methods perform equally well as Gaussian MLE, whereas for nonzero γ Gaussian and—to some extent—skew-normal MLE have inferior qualities compared to the Lambert W alternatives. In particular, the RMSE for $\hat{\sigma}_y$ increases with sample size.

Hence, there is no gain restricting analysis to the (symmetric) Gaussian case, as the Lambert W framework extends this distribution to a broader class, without losing the good properties of Gaussian MLE. For little asymmetry in the data, both the Lambert W and the skew-normal approach give accurate and precise estimates of location, scale and skewness. Yet for heavily skewed data (skewness greater than 0.995 in absolute value), the skew-normal framework fails not only in theory, but also in practice to provide a good approximation.

Table 5 shows the average number of iterations the IGMM algorithm needed to converge: for increasing sample size it needs less iterations—sample moments can be estimated more accurately; more iterations are needed for larger γ —as the starting value for γ is based on the Taylor expansion around $\gamma = 0$ and moving away from the origin makes the initial estimate $\gamma^{(0)} := \hat{\gamma}_{\text{Taylor}}$ less precise.

A closer look at the two sub-tables (top and bottom) shows that finding the optimal γ (Algorithm 2) becomes much more difficult for increasing γ and sample size N than finding the optimal μ_x and σ_x given the optimal $\hat{\gamma}_{\text{GMM}}$ (Algorithm 3). For $\gamma = 0$ and large N there is almost no difference between the total number of iterations (top) and the number of iterations in Algorithm 3 only (bottom). For large γ , however, the total number of iterations is approximately 5 times as large. The right panel shows the values for simulations of a Lambert $W \times t$ RV with $\nu = 4$ degrees of freedom. For small γ , finding $\hat{\gamma}_{\text{GMM}}$ takes much longer than for Gaussian input; surprisingly, for large γ convergence is reached faster. This is probably a result of the constrained optimization: due to more extreme values for a

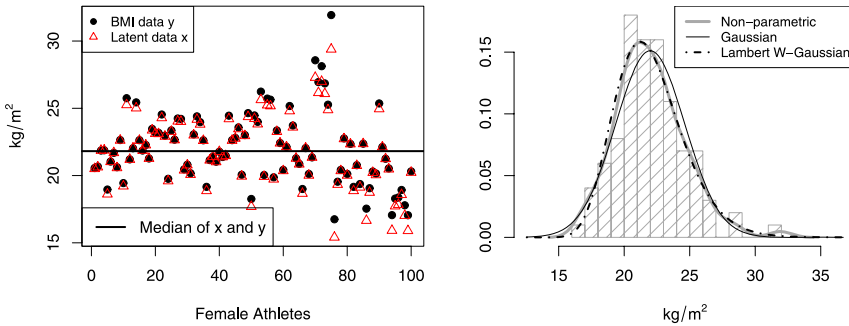


FIG. 6. Australian Athletes BMI: (left) observed data y (dots) and back-transformed data $\hat{x}_{\tau_{MLE}}$ (triangles); (right) histogram plus density estimates.

t -distribution, Algorithm 2 often returns one of the two boundary values for $\hat{\gamma}_{GMM}$ without even starting the optimization process.

Given its good empirical properties, fairly general assumptions about the input variable X , and its fast computation time, the IGMM algorithm can be used as a quick Lambert W check. For a particular Lambert $W \times F$ distribution, the Lambert $W \times F$ MLE gives more accurate results, especially for heavily skewed data.

7. Applications. This section demonstrates the usefulness of the presented methodology on real world data. In the first example I analyze parts of the Australian Athletes data set¹⁰ which can be typically found in the literature on modeling skewed data [Genton (2005), Azzalini and Dalla-Valle (1996)].

The second example reexamines the LATAM returns introduced in Section 1. A Lambert $W \times t$ -distribution is found to give an appropriate fit, both for the raw data as well as the standardized residuals of an auto-regressive conditional heteroskedastic time series model (see Section 7.2.1 for details). In particular, a comparison of risk estimators (Value at Risk) demonstrates the suitability of the Lambert $W \times F$ distributions to model financial data.

7.1. BMI of Australian athletes. Figure 6 shows the Body Mass Index (BMI) of 100 female Australian athletes (dots) and Table 6 lists several statistical properties (column 1). Although the data appear fairly Gaussian, its large positive skewness makes both tests reject normality on a 5% level.

After 5 iterations $\hat{\tau}_{IGMM} = (21.735, 2.570, 0.099)$, which implies $\hat{\mu}_y = 21.992$, $\hat{\sigma}_y = 2.633$, and $\gamma_1(\hat{\gamma}_{IGMM}) = 0.601$, assuming Gaussian input.

The BMI data set consists of exactly $n = 100$ i.i.d. samples and Table 2 lists finite sample properties of $\hat{\gamma}_{IGMM}$ for this case.¹¹ Thus, if $Y = BMI$ was Gaussian,

¹⁰R package LambertW, data set AA.

¹¹Although y is clearly not $\mathcal{N}(0, 1)$, the location-scale invariance of Lambert $W \times$ Gaussian RVs makes this difference to scenario 1 in the simulations [$Y \equiv X \sim \mathcal{N}(0, 1)$] irrelevant; finite sample

TABLE 6
BMI (y) and back-transformed data $\hat{x}_{\hat{\tau}_{IGMM}}$ and $\hat{x}_{\hat{\tau}_{MLE}}$: (top) summary statistics; (bottom) Shapiro–Wilk (SW), Jarque–Bera (JB) normality tests

BMI	y	$\hat{x}_{\hat{\tau}_{IGMM}}$	$\hat{x}_{\hat{\tau}_{MLE}}$
Min	16.750	15.356	15.406
Max	31.930	29.335	29.384
Mean	21.989	21.735	21.742
Median	21.815	21.815	21.815
St. dev.	2.640	2.570	2.569
Skewness	0.683	0.000	0.017
Kurtosis	1.093	0.186	0.187
SW	0.035	0.958	0.959
JB	0.001	0.877	0.874

then

$$(7.1) \quad \frac{\sqrt{100}\hat{\gamma}_{IGMM}}{0.4368} \sim \mathcal{N}(0, 1).$$

Plugging $\hat{\gamma}_{IGMM} = 0.099$ into (7.1) gives 2.279 and a corresponding *p*-value of 0.0113. Thus, $\hat{\gamma}_{IGMM}$ is significant on a 5% level, yielding an indeed positively skewed distribution for the BMI data **y**.

As both tests cannot reject Gaussianity for $\hat{x}_{\hat{\tau}_{IGMM}}$, a Lambert *W* × Gaussian approach seems reasonable. Table 7 shows that all estimates are highly significant, where standard errors are obtained by numerical evaluation of the Hessian at the optimum. As not one single test can reject normality of $\hat{x}_{\hat{\tau}_{MLE}}$ (triangles in Figure 6), an adequate model to capture the statistical properties of the BMI data is

$$BMI = (Ue^{0.099U})2.556 + 21.742, \quad U = \frac{X - 21.742}{2.556} \sim \mathcal{N}(0, 1).$$

TABLE 7
Lambert W × Gaussian MLE for the BMI data

	Estimate	Std. error	<i>t</i> value	Pr(> <i>t</i>)
μ_x	21.742	0.274	79.494	0.000
σ_x	2.556	0.188	13.618	0.000
γ	0.096	0.039	2.481	0.013

properties of γ do not change between $X \sim \mathcal{N}(0, 1)$ and general $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$, since in both cases μ_x and σ_x are also estimated.

For $\hat{\theta}_{MLE}$ the support of *BMI* lies in the half-open interval $[11.967, \infty)$. As all observations lie within these boundaries, $\hat{\theta}_{MLE}$ is indeed a (local) maximum. Figure 6 shows the closeness of the Lambert $W \times$ Gaussian density to the histogram and kernel density estimate, whereas the best Gaussian is apparently an improper approximation.

Although a more detailed study of athlete type and other health indicators might explain the prevalent skewness, the Lambert W results at least support common sense: the human body has a natural physiological lower bound¹² for the BMI, whereas outliers on the right tail—albeit, in principle, also having an upper bound—are more likely.

7.2. *Asset returns.* A lot of financial data, also the LATAM return series introduced in Section 1 (Table 1 and Figure 2), display negative skewness and excess kurtosis. These so-called *stylized facts* are well known and typically addressed via (generalized) auto-regressive conditional heteroskedastic (GARCH) [Engle (1982), Bollerslev (1986)] or stochastic volatility (SV) models [Melino and Turnbull (1990), Deo, Hurvich and Lu (2006)]. A theoretical analysis of Lambert $W \times F$ time series models, however, is far beyond the scope and focus of this work. For empirical evidence regarding the usefulness and significance of Lambert $W \times F$ distributions in GARCH models and possible future research directions see Section 7.2.1. It is also worth noting that the Lambert $W \times F$ transformation (2.2) resembles SV models very closely, and connections between the two can be made in future work.

Based on the news \leftrightarrow return interpretation in a stock market \mathcal{S} , it makes sense to assume a symmetric input distribution $F_X(x)$ for the latent news RV X . Without specifying the symmetric $F_X(x)$ any further, the IGMM algorithm gives a robust estimate for τ : here $\hat{\tau}_{IGMM} = (-0.048, 0.190, 1.456)$. Column 2 of Table 1 shows that the unskewed data $\hat{\mathbf{x}}_{IGMM}$ —here interpreted as news hitting the market—is non-Gaussian, but a t -distribution cannot be rejected. In consequence, Y is modeled as a Lambert $W \times$ location-scale t -distribution with $\beta = (c, s, \nu)$, where c is the location, s the scale and ν the degrees of freedom parameter. Table 8 shows

TABLE 8
Lambert $W \times t$ MLE for the LATAM series

	Estimate	Std. error	t value	$\Pr(> t)$
c	0.197	0.037	5.270	0.000
s	1.240	0.057	21.854	0.000
ν	7.047	2.196	3.208	0.001
γ	-0.053	0.014	-3.860	0.000

¹²The lower truncation of the BMI at 11.967 corresponds to a 180 cm tall athlete only weighing 38.88 kg.

that all coefficients of $\hat{\theta}_{MLE}$ are highly significant; in particular, $\hat{\gamma}$ increased substantially (in absolute value), as γ now solely addresses asymmetry in the data, and ν can capture excess kurtosis. Thus, the prevalent negative skewness in the LATAM daily returns is not an artifact of large outliers in the left tail of an otherwise symmetric distribution, but a significant characteristic of the data.

In order to check if the Lambert $W \times t$ -distribution is indeed an appropriate model for \mathbf{y} , it is useful to study the back-transformed data $\hat{\mathbf{x}}_{\hat{\tau}_{MLE}}$; here $\hat{\tau}_{MLE} := \tau(\hat{\theta}_{MLE}) = (0.197, 1.465, -0.053)$. Not surprisingly, the skewness of $\hat{\mathbf{x}}_{\hat{\tau}_{MLE}}$ reduced to almost 0 (column 3 of Table 1). As a Kolmogorov–Smirnov test cannot reject a student t -distribution for $\hat{\mathbf{x}}_{\hat{\tau}_{MLE}}$, the Lambert $W \times t$ -distribution

$$Y = (Ue^{-0.053U})1.465 + 0.197,$$

$$U = \frac{X - 0.197}{1.465}, \quad U \sqrt{\frac{7.047}{7.047 - 2}} \sim t_{\nu=7.047}$$

is an adequate unconditional probabilistic model for the LATAM returns \mathbf{y} .

The effect of news x_t in the market \mathcal{S} is clearly shown in a scatter plot of $\hat{\mathbf{x}}_{\hat{\tau}_{MLE}}$ versus \mathbf{y} . For example, consider the lower-left point $(x_{1346}, y_{1346}) \approx (-4.8, -6.1)$ in Figure 7. Here, the observed negative return equals -6.1% , but

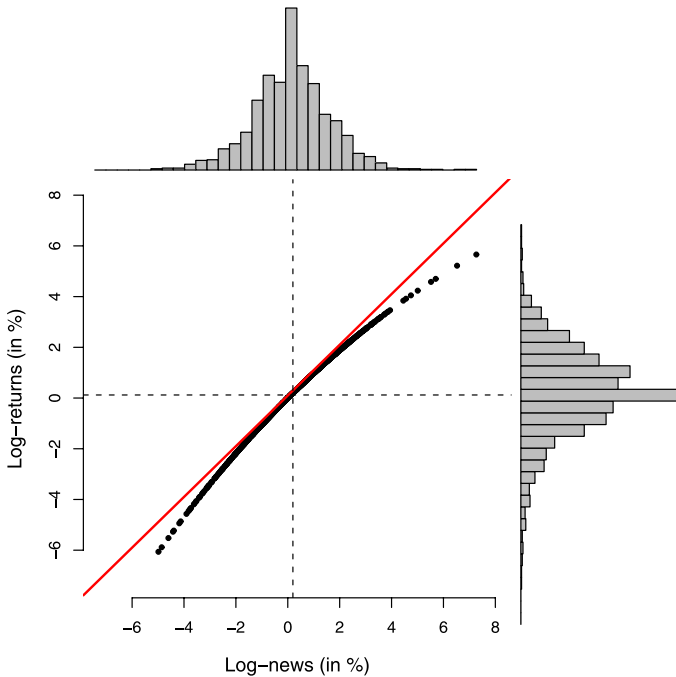


FIG. 7. News $\hat{\mathbf{x}}_{\hat{\tau}_{MLE}} \leftrightarrow$ return \mathbf{y} scatter plot plus histograms; solid 45° line: $\gamma = 0$. Dashed vertical and horizontal lines represent the sample mean of $\hat{\mathbf{x}}_{\hat{\tau}_{MLE}}$ and \mathbf{y} , respectively.

TABLE 9
VaR comparison for the LATAM series

Method	α						
	0.005	0.01	0.05	0.5	0.95	0.99	0.995
empirical	-4.562	-4.078	-2.478	0.138	2.344	3.192	3.818
Gaussian	-3.660	-3.294	-2.293	0.121	2.535	3.535	3.901
t	-4.297	-3.634	-2.214	0.121	2.455	3.875	4.538
Lambert $W \times t$	-4.871	-4.049	-2.358	0.197	2.351	3.437	3.893
Skew- t	-4.715	-3.973	-2.364	0.201	2.346	3.465	3.957

as $\hat{\gamma} = -0.053 < 0$, this outcome was an overreaction to bad news that was only “worth” -4.8% . For location-scale Lambert W RVs the skewness parameter γ is a powerful, yet easy way to characterize different markets/assets. The negative $\hat{\gamma}$ shows that this specific market (system) is exaggerating bad news, and devalues positive news.

Value at risk (VaR). The VAR is a popular measure in financial statistics to estimate the potential loss for an investment in an asset over a fixed time period. That is, the maximum percentage an investor can expect to lose—with a confidence of $1 - \alpha$ —over a fixed time period. Statistically this corresponds to the α -quantile of the distribution. The VAR can be obtained in various ways: the simplest are empirical and theoretical quantiles given the estimated parameter vector of a parametric distribution (which are sufficient for comparative purposes).

As expected, a Gaussian distribution underestimates both the low and high quantiles, as it lacks the capability to capture excess kurtosis (see Table 9). The t -distribution with $\hat{\nu}_{MLE} = 6.22$ degrees of freedom has heavier tails, but underestimates low and overestimates high quantiles: clearly an indication of the prevalent skewness in the data. The Lambert $W \times t$ and the skew t -distribution¹³ are the best approximation to the empirical quantiles: both heavy tails and negative skewness are captured (see also the Lambert $W \times t$ QQ plot in Figure 2). There is no clear “winner” between the two skewed distributions: skew- t quantiles are closer to the empirical ones for small α , Lambert $W \times t$ quantiles are closer for large α . Around the median ($\alpha = 0.5$) both skewed distributions are far away from the true value: the reason being a high concentration of close to 0 returns in financial assets, so-called “inliers” [see Breidt and Carriquiry (1995)].

7.2.1. Nonindependence of financial data. It is well known that financial return series y_t typically exhibit positive auto-correlation in their squares y_t^2 , which

¹³MLE estimates are (0.917, 1.422, -0.799 , 7.156) for the location, scale, shape and degrees of freedom parameter respectively; function `st.mle` in the `sn` package.

violates the independence assumption of the MLE presented in Section 5.1. A standard parametric way to capture this dependence is a GARCH model [Bollerslev (1986), Engle (1982)], which models the variance at time t , σ_t^2 , as a function of its own past. A simple, yet very successful model for an uncorrelated y_t is a GARCH(1, 1),

$$\begin{aligned} y_t &= \mu + \varepsilon_t \sigma_t, \\ \sigma_t^2 &= \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2, \end{aligned}$$

where ε_t is a zero-mean, unit-variance i.i.d. sequence [for technical details see Nelson (1990), Engle (1982)]. Typically, $\varepsilon_t \sim \mathcal{N}(0, 1)$, but also student t - or skew t -distributions are used for more flexibility in the conditional distribution of ε_t given the information set Ω_{t-1} available at time $t - 1$ [Bauwens and Laurent (2005)]. French, Schwert and Stambaugh (1987) also found that the standardized residuals $(y_t - \hat{\mu})/\hat{\sigma}_t$ —which can be considered as an i.i.d. sequence—still exhibit negative skewness after fitting a Gaussian GARCH model to S&P 500 returns.

After fitting a student- t GARCH(1, 1) model¹⁴ to the LATAM return series \mathbf{y} , the Lambert $W \times t$ MLE fit for the standardized residuals—which are approximately i.i.d. and thus do not violate the MLE assumptions—still gives a highly significant $\hat{\gamma} = -0.048$ with a p -value of 0.000113 (other estimates are not shown here).

While I will not study Lambert $W \times$ student- t GARCH models in detail, this example and the great flexibility of Lambert $W \times F$ distribution combined with the possibility to symmetrize skewed data suggest that Lambert $W \times F$ GARCH (and SV) models are a promising area of future research.

This analysis confirms previous findings that negative skewness is an important feature of asset returns. For example, optimal portfolio models based on skewed distributions lead to better suited decision rules to react to asymmetric price movements. It also shows that Lambert W distributions model the characteristics of financial returns as well as skew t -distributions, with the additional option to recover symmetric latent data, which is not possible for RVs based on a manipulation of the p.d.f. rather than a variable transformation.

8. Relation to Tukey's h distribution. During the final review process, Professor Andrew F. Siegel suggested possible connections of Lambert W distributions to Tukey's g - h distribution [Tukey (1977)]

$$(8.1) \quad Z = \frac{\exp(gU) - 1}{g} \exp\left(\frac{h}{2}U^2\right), \quad h \geq 0,$$

where $U \sim \mathcal{N}(0, 1)$. Here g is the skew parameter and h controls the tail behavior of Z .

¹⁴Function `garchFit(·)` in the `fGarch` package.

Although the underlying idea to introduce skewness is the same, the specific transformations to get the skewness effects are different, and so are the properties of the transformed RVs.

For $g \rightarrow 0$,

$$(8.2) \quad Z = U \exp\left(\frac{h}{2}U^2\right)$$

becomes symmetric. The RV Z has Tukey's h distribution and is commonly used to model heavy-tails [Fischer (2006), Field (2004)]. Equation (8.2) reveals a close link of Lambert $W \times F$ RVs to the existing statistics literature by noting that if $Z \sim h$, then $Z^2 = U^2 e^{hU^2}$ has a noncentral, nonscaled Lambert $W \times \chi_1^2$ distribution with $\gamma = h$.

For further important connections between the Lambert W function and Tukey's h distribution see Goerg (2011).

9. Discussion and outlook. Whereas the Lambert W function plays an important role in mathematics, physics, chemistry, biology and other fields, it has not yet been used in statistics. Here I introduce it in an input/output setting to skew and “unskew” RVs and data, respectively.

Successful application to biomedical and financial data together with the great flexibility with respect to the type of input RV X of Lambert $W \times F$ RVs promise a wide range of applications as well as theoretical studies for particularly chosen input distributions.

Last but not least, a very pragmatic advantage of the transformation-based Lambert $W \times F$ RVs compared to other approaches to asymmetry: data can be “unskewed” using Lambert's W function.

Acknowledgments. I am grateful to Professor Wilfredo Palma for giving me the opportunity to work at the Department of Statistics, Pontificia Universidad Católica de Chile, Santiago, where I completed important parts of this study.

Furthermore, I want to thank Professor Reinaldo Arellano-Valle, Professor Cosma Shalizi, the Editor Professor Stephen Fienberg and two anonymous referees for helpful comments and suggestions on the manuscript.

REFERENCES

- ARELLANO-VALLE, R. B. and AZZALINI, A. (2006). On the unification of families of skew normal distributions. *Scand. J. Stat.* **33** 561–574. [MR2298065](#)
- ARNOLD, B. C. and BEAVER, R. J. (2000). The skew-Cauchy distribution. *Statist. Probab. Lett.* **49** 285–290. [MR1794746](#)
- AZZALINI, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Stat.* **12** 171–178. [MR0808153](#)
- AZZALINI, A. and CAPITANIO, A. (1999). Statistical applications of the multivariate skew normal distributions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 579–602. [MR1707862](#)

- AZZALINI, A. and CAPITANIO, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *J. Roy. Statist. Soc. Ser. B* **65** 367–389. [MR1983753](#)
- AZZALINI, A. and DALLA-VALLE, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83** 715–726. [MR1440039](#)
- BAUWENS, L. and LAURENT, S. (2005). A new class of multivariate skew densities, with application to generalized autoregressive conditional heteroscedasticity models. *J. Bus. Econom. Statist.* **23** 346–354. [MR2159684](#)
- BEHBOODIAN, J., JAMALIZADEH, A. and BALAKRISHNAN, N. (2006). A new class of skew-Cauchy distributions. *Statist. Probab. Lett.* **76** 1488–1493. [MR2245569](#)
- BOLLERSLEV, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **31** 307–327. [MR0853051](#)
- BREIDT, J. and CARRIQUIRY, A. L. (1995). Improved quasi-maximum likelihood estimation for stochastic volatility models. In *Modelling and Prediction: Honoring Seymour Geisser*. Springer, New York.
- CONT, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quant. Finance* **1** 223–236.
- CORLESS, R. M., GONNET, G. H., HARE, D. E. G. and JEFFREY, D. J. (1996). On the Lambert W function. *Adv. Comput. Math.* **5** 329–359. [MR1414285](#)
- DEO, R., HURVICH, C. and LU, Y. (2006). Forecasting realized volatility using a long memory stochastic volatility model: Estimation, prediction and seasonal adjustment. *J. Econometrics* **131** 29–58. [MR2275995](#)
- ENGLE, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50** 987–1007. [MR0666121](#)
- FIELD, C. A. (2004). Using the gh distribution to model extreme wind speeds. *J. Statist. Plann. Inference* **122** 15–22. [MR2057911](#)
- FISCHER, M. (2006). Generalized Tukey-type distributions with application to financial and teletraffic data. Available at <http://econpapers.repec.org/RePEc:zbw:faucse:722006>.
- FRENCH, K. R., SCHWERT, G. W. and STAMBAUGH, R. F. (1987). Expected stock returns and volatility. *Journal of Financial Economics* **19** 3–29.
- GENTON, M. G. (2005). Discussion of “The skew-normal.” *Scand. J. Statist.* **32** 189–198.
- GLEN, A. G., LEEMIS, L. and DREW, J. H. (1997). A generalized univariate change-of-variable transformation technique. *INFORMS J. Comput.* **9** 288–295.
- GOERG, G. M. (2011). The Lambert Way to Gaussianize skewed, heavy tailed data with the inverse of Tukey’s h transformation as a special case. Unpublished manuscript. Available at <http://arxiv.org/abs/1010.2265>.
- MELINO, A. and TURNBULL, S. M. (1990). Pricing foreign currency options with stochastic volatility. *J. Econometrics* **45** 239–265.
- NELSON, D. B. (1990). Stationarity and persistence in the GARCH(1, 1) model. *Econometric Theory* **6** 318–334. [MR1085577](#)
- R DEVELOPMENT CORE TEAM (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at <http://www.R-project.org>.
- ROSENBLICHT, M. (1969). On the explicit solvability of certain transcendental equations. *Inst. Hautes Études Sci. Publ. Math.* **36** 15–22. [MR0258808](#)
- THODE, H. C., JR. (2002). *Testing for Normality. Statistics: Textbooks and Monographs* **164**. Dekker, New York. [MR1989476](#)
- TUKEY, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading.
- VALLURI, S. R., JEFFREY, D. J. and CORLESS, R. M. (2000). Some applications of the Lambert W function to physics. *Canad. J. Phys.* **78** 823–831.

- WEISS, M. S. (1978). Modification of the Kolmogorov–Smirnov statistic for use with correlated data. *J. Amer. Statist. Assoc.* **73** 872–875.
- YAN, J. (2005). Asymmetry, fat-tail, and autoregressive conditional density in financial return data with systems of frequency curves. Available at <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.76.2741>.

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: gmg@stat.cmu.edu