I use a parametric, bijective transformation to generate heavy tail versions $Y$ of an arbitrary random variable ($RV$) $X \sim F_X$, by similar concepts as in Goerg (2011) for skewed RVs. The tail behaviour of the so-called heavy tail Lambert $W \times F_X RVY$ depends on a tail parameter $\delta \geq 0$; for $\delta = 0$, $Y \equiv X$, for $\delta > 0 Y$ has heavier tails than $X$. For $X$ being Gaussian, this meta-family of heavy-tailed distributions reduces to Tukey’s $h$ distribution. Lambert’s $W$ function provides an explicit inverse transformation, which can be used to remove skewness and heavy tails from data and then apply standard methods and models to this so obtained “nice” (Gaussianised) data. The optimal inverse transformation can be estimated by maximum likelihood. This transformation based approach to heavy tails also yields analytical, concise and simple expressions for the cumulative distribution (cdf) $G_Y(y)$ and probability density function (pdf) $g_Y(y)$. As a special case, I present explicit expressions for Tukey’s $h$ pdf and cdf - to the authors knowledge for the first time in the literature. Applications to a simulated Cauchy sample, S&P 500 log-returns, as well as solar flares data demonstrate the usefulness of the introduced methodology.

The R package LambertW contains a wide range of methods presented here and is publicly available at CRAN.

Keywords: Gaussianising, heavy tails, power law, Tukey’s $h$ distribution, Lambert W, kurtosis, transformation of random variables; latent variables.
# Contents

1 Introduction 3

2 Modelling heavy tails using transformations 6
  2.1 Tukey’s $h$ distribution ........................................ 6
  2.2 Heavy tail Lambert W Random Variables .......................... 7
  2.3 Inverse transformation: “Gaussianise” heavy-tailed data .......... 8
  2.4 Distribution and Density Function ............................... 10
  2.5 Quantile Function ............................................. 11

3 Tukey’s $h$ Distribution: Gaussian Input 12
  3.1 Tukey’s $h$ versus student’s t ................................. 13

4 Parameter Estimation 13
  4.1 Maximum Likelihood Estimation .................................. 14
  4.1.1 Properties of the MLE for the heavy tail parameter .......... 16
  4.2 Iterative Generalized Method of Moments (IGMM) ............... 17

5 Simulations 19
  5.1 Estimating $\delta$ only ......................................... 19
  5.2 Estimating all parameters jointly .............................. 20
  5.3 Discussion of the simulation study ............................. 23

6 Applications 23
  6.1 Estimating the location of a Cauchy using the sample mean ...... 24
  6.2 Modelling heavy tails in financial return series: S&P 500 case study 24
  6.2.1 Gaussian fit to observed data ................................ 25
  6.2.2 Heavy tail fit to observed data ................................ 25
  6.2.3 “Gaussianising” the observed data ........................... 26
  6.2.4 Gaussian MLE for Gaussianised data .......................... 27
  6.3 Removing power laws: analysis of solar flares ..................... 28

7 Discussion and Outlook 30

References 31

A Auxiliary results 35
  A.1 Properties of $W_\delta(z)$ ...................................... 35
  A.2 Properties of the penalty term $\log R(\delta \mid z_i)$ for standard Gaussian input 37
  A.3 Properties of the Gaussian log-likelihood evaluated at $W_\delta(z)$ ........ 37

B Proofs of the results in the main text 38
  B.1 Inverse transformation ......................................... 38
  B.2 Cdf and pdf .................................................. 38
  B.3 MLE for $\delta$ ................................................ 39
1 Introduction

Both theory and statistical practice are tightly linked to Gaussianity. In theory, many statistical models require the data, noise, or parameters to have a (multivariate) Gaussian distribution: i) t and F-tests in regression rely on the assumption of (approximately) Gaussian errors; ii) pattern recognition or de-noising methods for images often model the additive noise as a Gaussian random field (Achim, Tsakalides, and Bezerianos, 2003); iii) non-standard distributions can be modelled by a mixture of Gaussians (Blum, Zhang, Sadler, and Kozick, 1999); iv) the Kalman filter is known to be optimal - in its basic form - for Gaussian errors (Friedland, 2005); v) many time series models are based on the building block of a Gaussian white noise sequence \( \epsilon_t \) (Brockwell and Davis, 1998; Engle, 1982; Granger and Joyeux, 2001). In all the above cases, the model \( M_N \), parameter estimates \( \hat{\theta} \) and their standard errors, and many other statistical properties, are then studied - all based on the ideal(istic) assumption of Gaussianity.

In practice, however, data/noise is rarely Gaussian but often exhibits asymmetry and/or heavy tails; for example wind speed data (Field, 2004), human dynamics (Vázquez, Oliveira, Dezső, Goh, Kondor, and Barabási, 2006), or Internet traffic data (Gidlund and Debernardi, 2009). Particularly notable examples are financial data (Cont, 2001; Kim and White, 2003) and speech signals (Aysal and Barner, 2006) which almost exclusively exhibit heavy tails. Thus a model \( M_N \) developed for the Gaussian case does not necessarily provide accurate inference anymore.

One way to overcome this shortcoming is to replace the Gaussian distribution in \( M_N \) with a heavy-tailed distribution \( G \) and study properties of the new model \( M_G \): i) regression with Cauchy errors (Smith, 1973); ii) image denoising for \( \alpha \)-stable noise (Achim et al., 2003); iii) non-Gaussian mixture models to approximate the distribution of \( \alpha \)-stable processes (Swami, 2000); iv) Kalman filtering for Cauchy (Idan and Speyer, 2010; Nezafat and Amindavar, 2001) or Lévy noise (Ahn and Feldman, 1999); v) forecasting long memory processes with heavy tail or general non-Gaussian innovations (Ilow, 2000; Palma and Zevallos, 2011), or ARMA Modelling of electricity loads with hyperbolic noise (Nowicka-Zagrajek and Weron, 2002).

While this fundamental approach to solve each problem is attractive from a theoretical perspective, it can become unsatisfactory from a practical viewpoint. Many successful models in the literature assume at some point Gaussianity, their theory for the Normal case is very well understood, many algorithms are implemented for the simple Gaussian case and not for the very particular setting of a Cauchy/Lévy/\( \alpha \)-stable model, thus developing models based on a completely unrelated distribution \( G \) is like throwing out the (Gaussian) baby with the bathwater.

It would be very useful if we could transform a Gaussian RV \( X \) to a (skewed) heavy-tailed RV \( Y \) and vice versa, and thus rely on the knowledge - and software - for the well-understood Gaussian case, while still modelling the skewness and excess kurtosis in the data. Optimally such a transformation should: a) be bijective, so we can go back and forth between the skewed/heavy-tailed distribution \( G_\tau(y) \) and the Gaussian \( F_X(x) \); b) include Normality as a special case, so we can test for skewness/heavy tails; and c) be parametric \( (\tau), Y = H_\tau(X) \), to estimate the optimal transformation efficiently from the sample \( y = (y_1, \ldots, y_N) \).

Liu, Lafferty, and Wasserman (2009) introduce a semi-parametric approach, where \( Y \) has a non-paranormal distribution if \( f(Y) \sim N(\mu, \sigma^2) \) and \( f(\cdot) \) is an increasing smooth function. It is semi-parametric in the sense that \( f(\cdot) \) is estimated non-parametrically. This leads to a greater flexibility in distribution shapes for \( Y \) than
any fixed parametric transformation, but it suffers from two drawbacks: i) non-parametric methods have slower convergence rates compared to parametric techniques, and ii) one identifiability condition for \( f(\cdot) \) is that \( E f(Y) \equiv E Y \) and \( V f(Y) \equiv V Y \). While the first point is the inherent cost for non-parametric generality, the second requires \( Y \) to have finite mean and variance, which is especially limiting for heavy-tailed data where this condition is often not met. Thus here we will study parametric transformations which are not as general as a non-parametric variant, but do not rely on such a restrictive identifiability condition and also work well for small sample sizes.

Figure 1 illustrates this pragmatic approach to analyse heavy-tailed data: applied researchers can make their observed data \( y \) as most Gaussian as possible (\( x_\tau \)) before making inference based on their favorite Gaussian model \( \mathcal{M}_N \). This avoids the development of or the data analysts waiting for - a whole new theory of \( \mathcal{M}_G \) and new software implementations based on a particular heavy-tailed distribution \( G \), while still improving statistical inference on skewed/heavy-tailed data \( y \). For example, consider the simulated \( y = (y_1, \ldots, y_{500}) \) from a standard Cauchy distribution \( C(0,1) \) in Fig. 2a: by Modelling the heavy tails by
Random sample $y \sim C(0, 1)$ from a standard Cauchy ($N = 500$)

Gaussianised Cauchy $x_{\hat{T}, MLE}$ with $\hat{T} = (0.030, 1.054, 0.861)$

Cumulative sample average, $\tau^{(n)}$ estimated for each fixed $n = 5, \ldots, 500$, before Gaussianising the data ($y_1, \ldots, y_n$).

(a) Gaussianising a Cauchy sample.

- a transformation based method - rather than a particular density shape - it is possible to Gaussianise even this Cauchy sample (Fig. 2b). This “nice” data can then be used for subsequent statistical analysis, e.g. estimating the location by the sample average (Fig. 2c) which is a bad choice for the Cauchy sample, but a good choice for the Gaussianised version. For a more detailed analysis of this data see Section 6.1.

The main contributions of this work are three-fold: a) following the recently introduced skewed Lambert $W \times F$ distributions (Goerg, 2011) I introduce a meta-family of heavy tail Lambert $W \times F$ distributions which include Tukey’s $h$ distribution (Tukey, 1977) as a special case; b) I derive the analytic inverse and thus get a bijective transformation to “Gaussianise” heavy-tailed data (Section 2); c) I also provide simple expressions for the cumulative distribution function (cdf) $G_Y(y)$ and probability density function (pdf) $g_Y(y)$ - including Tukey’s $h$ and $hh$ distribution-, which can be easily implemented in standard statistics package (Section 2.4). To the author’s knowledge analytic expressions for Tukey’s $h$ cdf and pdf are presented here (Section 3) for the first time in the literature. Section 4 introduces a methods of moments estimator for the optimal inverse transformation and studies properties of the maximum likelihood estimator (MLE). Section 5 shows their finite sample properties.

As has been shown in many case studies, Tukey’s $h$ distribution (heavy tail Lambert $W \times$ Gaussian) is useful to model data with unimodal, heavy-tailed densities. Section 6 not only confirms this finding for S&P 500 log-returns, but also demonstrates the benefits of removing heavy tails from data for exploratory data analysis: in particular, Gaussianising $\gamma$-ray intensity data reveals a truly bimodal density, which even non-parametric estimators fail to detect if heavy tails are not removed.

Computations, figures, and simulations were done with the open-source statistics package R (R Development Core Team, 2010). Functions used in the analysis and many other methods are available in the R package LambertW, which provides necessary tools to perform Lambert W inference in practice.
2 Modelling heavy tails using transformations

As discussed in the previous section, many statistical methods have to be adapted in presence of heavy tails in the data. Random variables exhibit heavy tails if more mass than for a Gaussian RV lies at the outer end of the density support. One common definition\footnote{There are various similar, but not exactly equivalent definitions of heavy-tailed RVs / distributions; for the context of this work these differences are not essential.} for a RV $Z$ to have a heavy tail with tail index $a$ is that its cdf satisfies $1 - F_Z(z) \sim L(z)z^{-a}$, where $L(z)$ is a slowly varying function at infinity, i.e. $\lim_{z \to \infty} \frac{L(tz)}{L(z)} = 1$ for all $t > 0$ (Baek and Pipiras, 2010). The heavy tail index $a$ is an important characteristic of $Z$; for example, only moments up to order $a$ exist.

2.1 Tukey’s $h$ distribution

A transformation based approach to heavy tails as discussed in the Introduction is the basis of Tukey’s $h$ RVs (Tukey, 1977)

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

where $U$ is standard Normal and $h$ is the heavy tail parameter. Tukey’s $h$ RVs are parametric heavy tail versions of a Gaussian RV with tail parameter $a = 1/h$ (Morgenthaler and Tukey, 2000), which reduce to the Gaussian for $h = 0$. Morgenthaler and Tukey (2000) extend the $h$ distribution to the skewed, heavy-tailed family of $hh$ RVs

$$Z = \begin{cases} U \exp \left( \delta_U U^2 \right), & \text{if } U \leq 0, \\ U \exp \left( \delta_R U^2 \right), & \text{if } U > 0, \end{cases}$$

where again $U \sim N(0,1)$. Here $\delta_U \geq 0$ and $\delta_R \geq 0$ shape the left and right tail of $Z$, respectively; thus transformation (2) can model skewed and heavy-tailed data - see Fig. 3a.

Their use in statistical practice is limited however, as the inverse of (1) or (2) have not been available in explicit, closed form. Consequently, no closed-form expressions for the cdf or pdf are available. Although Morgenthaler and Tukey (2000) express the pdf of (1) as

$$g_Z(z) = \frac{f_U \left( H_\delta^{-1}(z) \right)}{H_\delta'(H_\delta^{-1}(z))},$$

they fall short of explicitly specifying $H_\delta^{-1}(z)$. So far this inverse has been considered analytically intractable Field (2004), or only possible to approximate numerically (Fischer, 2010; Todd C. Headrick and Sheng, 2008).

Thus parameter estimation and inference relies on matching empirical and theoretical quantiles (Field, 2004; Morgenthaler and Tukey, 2000), or by the method of moments (Todd C. Headrick and Sheng, 2008). Only recently Todd C. Headrick and Sheng (2008) provided a numerical approximation to the cdf and pdf. Numerical approximations slow down the estimation of any statistical model - let it be in a frequentist or Bayesian setting. Hence, a closed form, analytically tractable pdf that can be computed efficiently is essential for a wide-spread use of Tukey’s $h$ (& variants) distribution.

Here I provide this inverse transformation and thus also an easily computable cdf and pdf, which can be implemented in standard statistics packages. For ease of notation and concision main results are shown in
detail for the symmetric case $\delta_\ell = \delta_r = \delta$; analogous results for $\delta_\ell \neq \delta_r$ will be stated without a detailed analysis.

2.2 Heavy tail Lambert W Random Variables

Tukey’s $h$ transformation (1) is strongly related to the approach taken by Goerg (2011) to introduce skewness in continuous RVs $X \sim F_X(x)$. It even holds that if $Z \sim$ Tukey’s $h$, then $Z^2$ has a skewed Lambert $W \times \chi^2_1$ distribution with skew parameter $\gamma = h$.

Adapting the skew Lambert $W \times F_X$ input/output idea (see Fig. 1), Tukey’s $h$ RVs can be generalized to heavy-tailed Lambert $W \times F_X$ RVs.

**Definition 2.1** (Non-central, Non-scaled Heavy-tailed Lambert $W \times F_X$ Random Variable). Let $U$ be a continuous RV with cdf $F_U(u \mid \beta)$ and pdf $f_U(u \mid \beta)$, where $\beta$ is a possible parameter vector of $F_X(u)$. Then,

$$Z := U \exp \left( \frac{\delta}{2} U^2 \right), \quad \delta \in \mathbb{R},$$  

(4)

is a non-central, non-scaled heavy tail Lambert $W \times F_X$ RV with parameter vector $\theta = (\beta, \delta)$, where $\delta$ is the tail parameter.

Tukey’s $h$ distribution results for $U$ being a standard Gaussian $\mathcal{N}(0, 1)$. For location-scale input, e.g. a general Gaussian RV it is necessary to extend Definition 2.1.

**Definition 2.2** (Location-scale Heavy-tailed Lambert $W \times F_X$ Random Variable). For a continuous location-scale family input $X \sim F_X(x \mid \beta)$ define a location-scale heavy-tailed Lambert $W \times F_X$ RV

$$Y := \left\{ U \exp \left( \frac{\delta}{2} U^2 \right) \right\} \sigma_x + \mu_x, \quad \delta \in \mathbb{R},$$  

(5)

with parameter vector $\theta = (\beta, \delta)$, where $U = (X - \mu_x)/\sigma_x$ is the zero-mean, unit variance version of $X$.

The input is not necessarily Gaussian but can be any other location-scale continuous RV, e.g. uniform $X \sim U(a, b)$. For scale family input - such as $X \sim \Gamma(a, b)$ - the following definition will be used.

**Definition 2.3** (Scaled Heavy-tailed Lambert $W \times F_X$ Random Variable). Let $X$ be a continuous RV from a scale-family distribution $F_X(x/s \mid \beta)$. Let $\sigma_x$ be the standard deviation of $X$ and $U = X/\sigma_x$. Then,

$$Y := X \exp \left( \frac{\delta}{2} U^2 \right), \quad \delta \in \mathbb{R},$$  

(6)

is a scaled heavy-tailed Lambert $W \times F_X$ RV with parameter vector $\theta = (\beta, \delta)$.

Let $\tau := (\mu_x(\beta), \sigma_x(\beta), \delta)$ be the parameter vector$^3$ that defines transformation (5). For simplicity and readability let $H_\delta(u) := u \exp \left( \frac{\delta}{2} u^2 \right)$.

$^2$Most concepts, terminology, and methods of the skew Lambert $W$ case relate one-to-one to the heavy tail Lambert $W$ RVs presented here. Thus for the sake of concision I refer to Goerg (2011) for a detailed account and background information of the Lambert $W$ framework.

$^3$For non-central, non-scale input set $\tau = (0, 1, \delta)$ and for scale-family input let $\tau = (0, \sigma_x, \delta)$. 

7
The shape parameter $\delta$ (\(=\) Tukey’s $h$) governs the tail behaviour of the transformed RV $Y$: for $\delta > 0$ values further away from $\mu_x$ are increasingly emphasized, leading to a heavy-tailed version of $F_X(x)$; for $\delta = 0$ the output $Y = X$ input; and for $\delta < 0$ values far away from the mean are mapped back again to values closer to $\mu_x$. Thus heavy tail Lambert $W \times F_X$ RVs generalize $X \sim F_X(x)$ to a new class of heavy-tailed versions $Y \sim G_Y(y)$ of itself with a reduction to the original for $\delta = 0$.

The Lambert $W$ formulation of heavy tail Modelling is more general than Tukey’s $h$ distribution in the sense that $X$ can have any distribution $F_X(x)$, not necessarily Gaussian (Fig. 4).

**Remark 2.4** (Only non-negative $\delta$). Although $\delta < 0$ leads to interesting properties of $Y$ I will not discuss it any further: $\delta < 0$ results in a non-bijective transformation and consequently to parameter-dependent support and non-unique input. Thus for the rest of this study I will tacitly assume $\delta \geq 0$, unless stated otherwise.

In this case, if $X$ has support on $(-\infty, \infty)$, then for all $\delta \geq 0$ also the location-scale $Y \in (-\infty, \infty)$. For a scale family $X \in [0, \infty)$ also the scale Lambert $W \times F_X$ RV $Y \in [0, \infty)$.

### 2.3 Inverse transformation: “Gaussianise” heavy-tailed data

For $\delta \geq 0$ transformation (5) is bijective and its inverse can be obtained via Lambert’s $W$ function, which is defined as the inverse of $z = u \exp(u)$, i.e. that function that satisfies $W(z) \exp(W(z)) = z$. Lambert’s $W$ has been studied extensively in mathematics, physics, and other areas of science (Corless, Gonnet, Hare, and Jeffrey, 1993; Rosenlicht, 1969; Valluri, Jeffrey, and Corless, 2000), and is implemented in several standard software packages via the GNU Scientific Library (gsl) (Galassi, Davies, Theiler, Gough, Jungman, Alken, Booth, and Rossi, 2011). Only very recently it received attention in the statistics literature (Goerg, 2011; Jodrá, 2009; Pakes, 2011; Rathie and Silva, 2011). It has several useful properties (see Appendix A and for more details Corless et al. (1993)), in particular $W(z)$ is bijective for $z \geq 0$.

**Lemma 2.5.** The inverse transformation of (5) is

$$W_\tau(Y) := W_\delta \left( \frac{Y - \mu_x}{\sigma_x} \right) \sigma_x + \mu_x = U \sigma_x + \mu_x = X,$$

where

$$W_\delta(z) := \text{sgn}(z) \left( \frac{W(\delta z^2)}{\delta} \right)^{1/2},$$

and $\text{sgn}(z)$ is the sign of $z$. The function $W_\delta(z)$ is bijective for all $\delta \geq 0$ and all $z \in \mathbb{R}$.

**Proof.** See Appendix B.

Lemma 2.5 gives for the first time an analytic, bijective inverse of Tukey’s $h$ transformation: $H^{-1}_\delta(y)$ of (Morgenthaler and Tukey, 2000) is now analytically available as (7). Bijectivity implies that for a given dataset $y$ and parameter vector $\tau$, the exact corresponding input $x_\tau = W_\tau(y)$ with cdf $F_X(x)$ can be obtained.

In view of the importance and popularity of Gaussianity, we clearly want to back-transform skewed, heavy-tailed data to something Gaussian rather than yet another heavy-tailed distribution. Typically tail behaviour of RVs are compared by their fourth central standardized moment $\gamma_2(X) = E(X - \mu)^4/\sigma^4_x$ - i.e. their kurtosis; for a Gaussian RV $\gamma_2(X) = 3$. Hence it is natural to set 3 as the reference value, and for the future when we “normalize the data $y$” we not only subtract the mean, and divide by the standard deviation, but also
transform it to data \( x \tau \) with \( \hat{\gamma}^2 (x\tau) = 3 - \) a “Normalization” in the true sense of the word (see Fig. 2b).

This data-driven view of the Lambert W framework can also be useful for non-parametric density estimation, where multivariate data is often pre-scaled to unit-variance, so a kernel density estimator (KDE) can use the same bandwidth \( h \) in each dimension (Hwang, Lay, and Lippman, 1994; Wasserman, 2007). “Normalizing” the data the Lambert Way does not only pre-scale the data, but also improves KDEs for heavy-tailed data (see also Maiboroda and Markovich, 2004; Markovich, 2005; Takada, 2001).

**Corollary 2.6** (Inverse transformation for Tukey’s \( hh \)). The inverse transformation of (2) is

\[
W_{\delta_\ell, \delta_r}(z) = \begin{cases} 
W_{\delta_\ell}(z), & \text{if } z \leq 0, \\
W_{\delta_r}(z), & \text{if } z > 0.
\end{cases}
\]  

Figure 3b shows \( W_{\delta_\ell, \delta_r}(z) \) for \( \delta_\ell = 0 \) and \( \delta_r = 1/10 \). The original transformation in Fig. 3a generates a right heavy tail version of the input \( U \) (x-axis) as it stretches the positive axis (y-axis), but leaves the negative axis the same. By definition \( W_{\delta_\ell, \delta_r}(z) \) does the opposite and removes the heavier right tail in \( Z \) (positive y-axis) back to the original \( U \). Fig. 3c shows how \( W_\delta(z) \) operates for various degrees of heavy tails and \( z \in [0, 3] \). If \( \delta \) is close to zero, then also \( W_\delta(z) \approx z \); for larger \( \delta \), the inverse maps \( z \) to smaller values, in particular if \( z \) is also large.

**Remark 2.7** (Generalized transformation). Transformation (1) can be generalized to

\[
Z = U \exp \left( \frac{\delta}{2\alpha} U^{2\alpha} \right), \text{ where } U^{2\alpha} = (U^2)\alpha, \alpha > 0.
\]

The inner term \( U^2 \) guarantees that the transformation is bijective for all \( \alpha > 0 \). The inverse transformation is

\[
W_{\delta, \alpha}(z) := \text{sgn}(z) \left( W \left( \frac{\delta z^{2\alpha}}{\delta} \right) \right)^{1/\alpha}.
\]  

For the sake of comparison with Tukey’s \( h \) distribution I will here consider the \( \alpha = 1 \) case only. For \( \alpha = 1/2 \) it is closely related to the skewed Lambert \( W \times F_X \) distributions. Future research can study the general case of heavy-tail Lambert \( W \times F_X \) distributions for arbitrary \( \alpha > 0 \).
2.4 Distribution and Density Function

For ease of notation let

\[ z = \frac{y - \mu_x}{\sigma_x}, \quad u = W_\delta(z), \quad x = W_\tau(y) = u\sigma_x + \mu_x. \] (12)

**Theorem 2.8** (Distribution and Density of \( Y \)). The cdf and pdf of a location-scale heavy tail Lambert \( W \times F_X \) RV equal

\[ G_Y(y \mid \beta, \delta) = F_X(W_\delta(z)\sigma_x + \mu_x \mid \beta) \] (13)

and

\[ g_Y(y \mid \beta, \delta) = f_X \left( W_\delta \left( \frac{y - \mu_x}{\sigma_x} \right) \sigma_x + \mu_x \mid \beta \right) \cdot \frac{W_\delta \left( \frac{y - \mu_x}{\sigma_x} \right)}{1 + W \left( \delta \left( \frac{y - \mu_x}{\sigma_x} \right)^2 \right)} \] (14)

\[ = f_X \left( W_\delta \left( \frac{y - \mu_x}{\sigma_x} \right) \sigma_x + \mu_x \mid \beta \right) \cdot e^{-\frac{1}{2}W_\delta(z)^2} \frac{1}{1 + W \left( \delta \left( \frac{y - \mu_x}{\sigma_x} \right)^2 \right)}, \] (15)

where the second equality follows as by definition \( \frac{W_\delta(z)}{z} = e^{-\frac{1}{2}W_\delta(z)^2} \).

Clearly, \( G_Y(y \mid \beta, \delta = 0) = F_X(y \mid \beta) \) and \( g_Y(y \mid \beta, \delta = 0) = f_X(y \mid \beta) \), since \( \lim_{\delta \to 0} W_\delta(z) = z \) and \( \lim_{\delta \to 0} W(\delta z^2) = 0 \) for all \( z \in \mathbb{R} \).

**Proof.** See Appendix B.

For the cdf and pdf of scale family or non-central, non-scale input set \( \mu_x = 0 \) or \( \mu_x = 0, \sigma_x = 1 \) in Theorem 2.8.

The explicit formula (14) allows a fast computation and theoretical analysis of the likelihood for any input \( f_X(\cdot) \), which is essential for any kind of - either frequentist or Bayesian - statistical analysis. A more detailed analysis of the functional form in (14) and its properties is given in Section 4.1 on maximum likelihood estimation.

Figure 4 shows (13) and (14) for various \( \delta \geq 0 \) with for four different input \( X \sim F_X(x \mid \beta) \): for \( \delta = h = 0 \) the input equals the output (solid black); for larger \( \delta \) the tails of \( G_Y(y \mid \theta) \) and \( g_Y(y \mid \theta) \) get heavier (dashed colored).

**Corollary 2.9** (Cdf and pdf of the double-tail (hh) Lambert \( W \times F_X \) RV). The cdf and pdf of \( Z \) in (2) equal

\[ G_Z(z \mid \beta, \delta_l, \delta_r) = \begin{cases} G_Z(z \mid \beta, \delta_l), & \text{if } z \leq 0, \\ G_Z(z \mid \beta, \delta_r), & \text{if } z > 0, \end{cases} \] (16)
(a) Lambert $W \times \chi^2_k$ with $\beta = k = 1$.
(b) Lambert $W \times \Gamma(s, r)$ with $\beta = (s, r) = (3, 1)$.
(c) Lambert $W \times \mathcal{N}(\mu, \sigma^2)$ with $\beta = (\mu, \sigma) = (0, 1)$.
(d) Lambert $W \times U(a, b)$ with $\beta = (a, b) = (-1, 1)$.

Figure 4: The pdf (top) and cdf (bottom) of a heavy-tail (a) “non-central, non-scaled”, (b) “scale”, and (c and d) “location-scale” Lambert $W \times F_X$ RV $Y$ for various degrees of heavy tails (colored, dashed lines).

and

$$g_Z(z \mid \beta, \delta) = \begin{cases} g_Z(z \mid \beta, \delta_L), & \text{if } z \leq 0, \\ g_Z(z \mid \beta, \delta_R), & \text{if } z > 0. \end{cases}$$ (17)

2.5 Quantile Function

The quantile function Tukey’s $h$ RV $Y$ has been very important in statistical practice, as quantile fitting has been the standard procedure to estimate $\mu_x$, $\sigma_x$, and $\delta$ (or $\delta_L$ and $\delta_R$). In particular, the median of $Y$ equals the median of $X$. Thus for symmetric, location-scale family input the sample median of $y$ is a robust estimate for $\mu_x$ for any $\delta \geq 0$ (see also Section 5). General quantiles can be computed via (Tukey, 1977)

$$y_\alpha = u_\alpha \exp \left( \frac{\delta}{2} u_\alpha^2 \right) \sigma_x + \mu_x,$$ (18)

where $u_\alpha := W_\delta(z_\alpha)$ are the $\alpha$-quantiles of the input distribution $F_U(u)$. As quantiles of $U$ are typically tabulated, or easily available in software packages, (18) can be computed very efficiently using $u_\alpha$ and $\tau$.

This simple calculation is especially useful for statistical education: teaching heavy-tailed statistics in introductory courses will soon become too difficult using Cauchy, Lévy or $\alpha$-stable distributions. Yet, back-transforming data via Lambert’s $W$, using previously learnt methods for the Gaussian case, and then transforming the inference back to the “heavy-tailed world” - e.g. transforming quantiles via using (18) - is straightforward. Thus the Lambert $W \times F_X$ framework can promote heavy-tailed statistics with enduring value in introductory statistics courses.
3 Tukey’s h Distribution: Gaussian Input

For Gaussian input Lambert $W \times F_X$ equals Tukey’s $h$ distribution, which has been studied extensively in the literature. Dutta and Babbel (2002) show that

$$\mathbb{E}Z^n = \begin{cases} 
0, & \text{if } n \text{ is odd and } n < \frac{1}{\delta}, \\
\frac{n!(1-n\delta)^{-\frac{(n+1)}{2}}}{2^{n/2}(n/2)!}, & \text{if } n \text{ is even and } n < \frac{1}{\delta}, \\
\#_n, & \text{if } n > \frac{1}{\delta},
\end{cases}$$

(19)

which in particular implies (Todd C. Headrick and Sheng, 2008)

$$\mathbb{E}Z = \mathbb{E}Z^3 = 0 \text{ if } \delta < 1 \text{ and } 1/3, \text{ respectively}$$

(20)

and

$$\mathbb{E}Z^2 = \frac{1}{(1-2\delta)^{3/2}} \text{ if } \delta < \frac{1}{2}, \quad \mathbb{E}Z^4 = 3 \frac{1}{(1-4\delta)^{5/2}} \text{ if } \delta < \frac{1}{4}. \quad (21)$$

Thus the kurtosis of a heavy tail Lambert $W \times$ Gaussian RV $Y$ equals (see Fig. 5)

$$\gamma_2(\delta) = 3 \frac{(1-2\delta)^3}{(1-4\delta)^{5/2}} \text{ for } \delta < 1/4. \quad (22)$$

For $\delta = 0$ expressions (21) and (22) reduce to the familiar Gaussian values. Expanding (22) around this Gaussian point $\delta = 0$ yields

$$\gamma_2(\delta) = 3 + 12\delta + 66\delta^2 + O(\delta^3). \quad (23)$$

Ignoring $O(\delta^3)$ and solving the quadratic equation (taking the positive root) gives a rule of thumb estimate

$$\hat{\delta}_{Taylor} = \frac{1}{66} \left[ \sqrt{66 \hat{\gamma}_2(y)} - 162 - 6 \right]_+, \quad (24)$$

where $\hat{\gamma}_2(y)$ is the sample kurtosis estimate from data $y$ and $[a]_+ = \max(a, 0)$.

**Corollary 3.1** (Cdf and pdf of Tukey’s h distribution). The cdf of Tukey’s h distribution equals

$$G_Y(y \mid \mu_x, \sigma_x, \delta) = \Phi \left( \frac{W_x(y) - \mu_x}{\sigma_x} \right), \quad (25)$$

where $\Phi(u)$ is the cdf of a standard Normal. The pdf equals (for $\delta > 0$)

$$g_Y(y \mid \mu_x, \sigma_x, \delta) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1 + \delta}{2} W_\delta \left( \frac{y - \mu_x}{\sigma_x} \right)^2 \right) \frac{1}{1 + \sqrt{\delta} \left( \frac{y - \mu_x}{\sigma_x} \right)^2} \quad (26)$$

**Proof.** Take $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ in Theorem 2.8. \qed

Section 4.1 studies functional properties of $g_Y(y \mid \theta)$ in more detail.
3.1 Tukey’s h versus student’s t

Student’s $t_{\nu}$ distribution with $\nu$ degrees of is often used to model heavy-tailed data (Wong, Chan, and Kam, 2009; Yan, 2005), as its tail index equals $\nu$. Thus the $n$th moment of a student $t$ RV $T$ exists if $n < \nu$. In particular,

\[ \mathbb{E}T = \mathbb{E}T^3 = 0 \text{ if } \nu < 1 \text{ or } 3, \quad \mathbb{E}T^2 = \frac{\nu}{\nu - 2} \text{ if } \frac{1}{\nu} < \frac{1}{2}, \]

and kurtosis

\[ \gamma_2(\nu) = 3 \left( \frac{\nu - 2}{\nu - 4} \right) = 3 \left( \frac{\nu - \frac{1}{2}}{\nu - \frac{1}{4}} \right) \text{ if } \frac{1}{\nu} < \frac{1}{4}. \]

Comparing the second expression of (28) with (22) shows a natural association between $1/\nu$ and $\delta$ and also a close similarity between student’s $t$ and Tukey’s $h$ distributions in terms of their first four moments (Fig. 5). By continuity and monotonicity of variance and kurtosis as functions of $1/\nu$ and $\delta$, it is clear that the first four moments of a location-scale $t$ distribution can always be perfectly matched by a corresponding location-scale heavy-tail Lambert $W \times$ Gaussian distributions. Thus if student’s $t$ is used to model heavy tails, and not as the true distribution of a test-statistics, it might be worthwhile to also fit heavy tail Lambert $W \times$ Gaussian distributions for an equally valuable “second opinion” on properties of the data. For example, a parallel analysis of S&P 500 log-returns in Section 6.2 leads to divergent inference regarding the existence of fourth moments. Additionally, the Lambert $W$ approach allows to Gaussianise the data and thus might reveal hidden patterns in the data - which can be easily overseen in presence of heavy tails (Section 6.3).

4 Parameter Estimation

For a sample of $N$ independent identically distributed (i.i.d.) observations $y = (y_1, \ldots, y_N)$, which presumably originates from transformation (5), $\theta = (\beta, \delta)$ has to be estimated from the data. Due to the lack of a closed form pdf of $Y$, this has been typically done by matching quantiles or a method of moments estimator (Field, 2004; Morgenthaler and Tukey, 2000; Todd C. Headrick and Sheng, 2008). Using the pdf (14) these
4.1 Maximum Likelihood Estimation

For an i.i.d. sample \( y \sim g_Y(y \mid \beta, \delta) \) the log-likelihood function equals

\[
\ell(\theta \mid y) = \sum_{i=1}^{N} \log g_Y(y_i \mid \beta, \delta).
\] (29)

The MLE is that \( \theta = (\beta, \delta) \) which maximizes (29), i.e.

\[
\hat{\theta}_{MLE} = \left( \hat{\beta}, \hat{\delta} \right) = \arg \max_{\beta, \delta} \ell(\beta, \delta \mid y).
\]

Since \( g_Y(y_i \mid \beta, \delta) \) is a function of \( f_X(x_i \mid \beta) \), the MLE depends on the specification of the input density.

Due to the multiplicative term in (14), expression (29) can be decomposed in two additive terms

\[
\ell(\beta, \delta \mid y) = \ell(\beta \mid x_\tau) + R(\tau \mid y),
\] (30)

where

\[
\ell(\beta \mid x_\tau) = \sum_{i=1}^{N} \log f_X \left( W_\delta \left( \frac{y_i - \mu_x}{\sigma_x} \right) \sigma_x + \mu_x \right)
\]

\[
= \sum_{i=1}^{N} \log f_X (x_\tau \mid \beta)
\] (31)

is the log-likelihood of the back-transformed data \( x_\tau \) and

\[
R(\tau \mid y) = \sum_{i=1}^{n} \log R(\mu_x, \sigma_x, \delta \mid y_i),
\] (32)

where

\[
R(\mu_x, \sigma_x, \delta \mid y_i) = \frac{W_\delta \left( \frac{y_i - \mu_x}{\sigma_x} \right)}{1 + W_\delta \left( \frac{y_i - \mu_x}{\sigma_x} \right)^2}.
\] (33)

Note that \( R(\mu_x, \sigma_x, \delta \mid y_i) \) only depends on \( \mu_x(\beta) \) and \( \sigma_x(\beta) \) (and \( \delta \)), but not necessarily on every coordinate of \( \beta \).

Decomposition (30) shows the difference between the exact MLE \( \left( \hat{\beta}, \hat{\delta} \right) \) based on \( y \) and the approximate MLE \( \hat{\beta} \) based on back-transformed data \( x_\tau \): if we knew \( \tau = (\mu_x, \sigma_x, \delta) \) beforehand, then we could back-transform \( y \) to \( x_\tau \) (no \( \hat{\tau} \) since the inverse transformation is assumed to be known) and compute \( \hat{\beta}_{MLE} \) based on \( x_\tau \) (maximize (31) with respect to \( \beta \)). In practice, however, \( \tau \) must also be estimated and this enters the likelihood via the additive term \( R(\tau \mid y) \). It can be shown that for any \( y_i \in \mathbb{R} \) the expression

\[
\log R(\mu_x, \sigma_x, \delta \mid y_i) \leq 0 \text{ if } \delta \geq 0,
\]

with equality if and only if \( \delta = 0 \). Thus \( R(\tau \mid y) \) can be interpreted as a penalty for transforming the data. Maximizing (30) faces a trade-off between transforming the data to follow \( f_X(x \mid \beta) \) (and thus increasing \( \ell(\beta \mid x_\tau) \)) versus the penalty of doing a more extreme transformation (and thus decreasing \( R(\tau \mid y) \)).

Figure 6a shows the contour plot of \( R(\mu_x = 0, \sigma_x = 1, \delta \mid y) \) as a function of \( \delta \) and \( y = z \). The penalty for transforming the data increases (in absolute value) either if \( \delta \) gets larger (for any fixed \( y \)) or for larger \( y \) (for
(a) Penalty term $R(\mu_x, \sigma_x, \delta | y_i)$ (33) in the full likelihood $\ell(\beta, \delta | y)$ as a function of $\delta$ and $y$. Input set to $\mu_x = 0$ and $\sigma_x = 1$.

(b) (left) random sample $z$ of Tukey’s $h$ with $U \sim N(0, 1)$ and $h = \delta = 1/3$; (right) additive decomposition of the log-likelihood $\ell(\theta | y)$ (solid, green) in input log-likelihood (dashed blue) and penalty term (red dashed). Vertical lines show true $\delta = 1/3$ and $\hat{\delta}_{MLE} \approx 0.37$.

Figure 6: The log-likelihood $\ell(\theta | y)$ decomposition for Lambert $W \times F_X$ distributions.

any fixed $\delta$). In both cases, increasing $\delta$ will make the transformed data $W_\delta(z)$ get closer to $0 = \mu_x$, which in turn increases its input likelihood. For $\delta = 0$ there is no (multiplicative) penalty since input equals output; for $y = 0$ there is no penalty since $W_\delta(0) = 0$ for all $\delta$.

Figure 6b shows (left) a random sample ($N = 1000$) $z \sim$ Tukey’s $h$ with $\delta = 1/3$ and the decomposition of the log-likelihood as in (30). Since $\beta = (0, 1)$ is known, likelihood functions and penalty terms are only functions of $\delta$ (and the data $z$). The monotonicity of the penalty term (decreasing) and the input likelihood (increasing) as a function of $\delta$ is not a property of this particular sample, but holds true in general (see Theorem 4.1 below). By this monotonicity in each component it follows that their sum (green line) has a unique maximum; here $\hat{\delta}_{MLE} = 0.37$ (red dotted vertical line).

The maximization of (30) can be carried out numerically. Here I show existence and uniqueness of $\hat{\delta}_{MLE}$ assuming that $\mu_x$ and $\sigma_x$ are known. Theoretical results for $\hat{\theta}_{MLE}$ remain a task for future work. Given the “nice” form of $g_Y(y)$ - continuous, twice differentiable, 4 support does not depend on the parameter, etc. - the MLE for $\theta = (\beta, \delta)$ should have the standard optimality properties (Lehmann and Casella, 1998).

4 Assuming that $f_X(\cdot)$ is twice differentiable.
4.1.1 Properties of the MLE for the heavy tail parameter

Without loss of generality (and for better readability) let $\mu_x = 0$ and $\sigma_x = 1$. In this case the likelihood function simplifies to

$$
\ell(\delta \mid z) \propto -\frac{1}{2} \sum_{i=1}^{N} [W_\delta(z_i)]^2 + \sum_{i=1}^{N} \log W_\delta(z_i) - \log \left(1 + \delta [W_\delta(z_i)]^2\right)
$$

(34)

$$
= -\frac{1 + \delta}{2} \sum_{i=1}^{N} [W_\delta(z_i)]^2 - \sum_{i=1}^{N} \log \left(1 + \delta [W_\delta(z_i)]^2\right).
$$

(35)

**Theorem 4.1 (Unique MLE for $\delta$).** Let $Z$ have a Lambert $W \times$ Gaussian distribution, where $\mu_x = 0$ and $\sigma_x = 1$ are assumed to be known and fixed. Also consider only the case $\delta \in [0, \infty)$.

a) If

$$
\sum_{i=1}^{n} z_i^4 \sum_{i=1}^{m} z_i^2 \leq 3,
$$

(36)

then $\hat{\delta}_{MLE} = 0$.

b) If (36) does not hold, then $\hat{\delta}_{MLE} > 0$ exists and is a positive solution to

$$
\sum_{i=1}^{N} z_i^2 W'(\delta z_i^2) \left(\frac{1}{2} W_\delta(z_i)^2 - \left(\frac{1}{2} + \frac{1}{1 + W(\delta z_i^2)}\right)\right) = 0.
$$

(37)

c) There is only one such $\delta$ satisfying (37), i.e. $\hat{\delta}_{MLE}$ is unique.

**Proof sketch.** See Appendix B for a detailed proof.

a) If condition (36) holds, then $D(\delta \mid z) := \frac{\partial}{\partial \delta} \ell(\delta \mid z)$ is negative at $\delta = 0$ and stays negative for all $\delta > 0$. Hence the maximizer occurs at $\delta = 0$.

b) If (36) does not hold, then $D(\delta = 0 \mid z) > 0$, decreases in $\delta$ and crosses the zero line (one candidate for $\hat{\delta}_{MLE}$ occurs here).

c) As $\delta$ gets larger, $D(\delta \mid z)$ reaches a minimum (negative value) and starts increasing again. However, for $\delta \to \infty$ the derivative approaches zero from below and never equals zero again; thus $\hat{\delta}_{MLE}$ is unique.

Condition (36) says that the MLE only yields positive estimates if the data is heavy-tailed enough. Points b) and c) guarantee that there is no ambiguity in the heavy tail estimate. This is an advantage over student’s $t$ distribution, for example, which has numerical problems and local maxima for unknown (and small) $\nu$.

---

5While for some samples $z$ the MLE also exists allowing all $\delta \in \mathbb{R}$, it can not be guaranteed for all $z$. The reason lies again in special properties of the Lambert W function. If $\delta < 0$ (and $z \neq 0$), then $W_\delta(z)$ is either not unique in $\mathbb{R}$ (principal and non-principal branch) or may not even have a real-valued solution.
Algorithm 1: Find optimal $\delta$: function \texttt{delta\_GMM}(\cdot) in the LambertW package.

\begin{algorithmic}
\Input{standardized data vector $z$; theoretical kurtosis $\gamma_2(X)$}
\Output{$\hat{\delta}_{\text{GMM}}$ as in (38)}
\begin{enumerate}
\item $\hat{\delta}_{\text{GMM}} = \arg\min_{\delta} ||\hat{\gamma}_2(u_\delta) - \gamma_2(X)||$, where $u_\delta = W_\delta(z)$ subject to $\delta \geq 0$
\item \Return $\hat{\delta}_{\text{GMM}}$
\end{enumerate}
\end{algorithmic}

The global maximum property of $\hat{\delta}_{\text{MLE}}$ holds for any $\delta \geq 0$.

For future theoretical analysis regarding the MLE it is worthwhile to point out that the log-likelihood and its gradient depend on $\delta$ and $z$ only via $W_\delta(z)$. Given the heavy tails in $z$ (for $\delta > 0$) we might expect that larger $\delta$ lead to difficulties in the evaluation of integrals (e.g. expected log-likelihood, Fisher information). However, $W_\delta(Z) \sim \mathcal{N}(0, 1)$ for the true $\delta \geq 0$, and close to a standard Gaussian if $\hat{\delta}_{\text{MLE}} \approx \delta$. Thus the performance of the MLE should not get worse for large $\delta$ as long as the initial estimate is close enough to the truth. Simulation studies in the next section support this conjecture, even for the joint MLE $\hat{\theta}_{\text{MLE}}$.

A disadvantage of the MLE is the mandatory a-priori specification of the input distribution. Especially for heavy-tailed data the eye is a bad judgement to choose a particular parametric form of the input distribution. It would be useful to estimate $\tau$ directly from the data, without the intermediate step of estimating $\theta$ first (and thus no distributional assumption for the input is necessary).

4.2 Iterative Generalized Method of Moments (IGMM)

Here I present an iterative method to obtain $\hat{\tau}$, which builds on the input/output aspect and relies upon theoretical properties of the input $X$. For example, if a random variable should be exponentially distributed (e.g. waiting times), but the observed data shows heavier tails then it is natural to estimate $\sigma_x = \lambda^{-1}$ and $\delta$ such that the back-transformed data has skewness 2, as this is a particular property of exponential RVs - independent of the rate parameter $\lambda$; to remove heavy tails in $y$ we should choose $\tau$ such that the back-transformed data $x_\tau$ has sample kurtosis 3; or for uniform input, we can try to find a $\tau$ such that $x_\tau$ has a flat density estimate.

Here I describe the estimator to remove heavy-tails in location-scale data, in the sense that the kurtosis of the input equals 3. It can be easily adapted to match other properties of the input as outlined above.

For a moment assume that $\mu_x = \mu_x^{(0)}$ and $\sigma_x = \sigma_x^{(0)}$ are known and fixed; only $\delta$ has to be estimated. A natural choice for $\delta$ is the one that results in back transformed data $x_\tau$ ($\tau = (\mu_x^{(0)}, \sigma_x^{(0)}, \delta)$) with sample kurtosis $\hat{\gamma}_2(x_\tau)$ equal to the theoretical kurtosis $\gamma_2(X)$. Formally,

\begin{equation}
\hat{\delta}_{\text{GMM}} = \arg\min_{\delta} ||\gamma_2(X) - \hat{\gamma}_2(x_\tau)||, \tag{38}
\end{equation}

where $||\cdot||$ is a proper norm in $\mathbb{R}$.

While the concept of this estimator is identical to its skewed version (Goerg, 2011), it has one important advantage: the inverse transformation is bijective. Thus here we do not have to consider “lost” data points when applying the inverse transformation.
Discussion of Algorithm 1: The kurtosis of Y as a function of δ is continuous and monotonically increasing (see (22)). Also u = W_δ(z) has a smaller slope than the identity u = z, and the slope is decreasing as δ is increasing. Thus if the kurtosis of the original data is larger than the objective kurtosis of the back-transformed data, \( \hat{\gamma}_2(y) > \gamma_2(X) \), then there always exists a \( \hat{\delta}(x) \) that achieves \( \hat{\gamma}_2(x_\cdot) \equiv \gamma_2(X) \) for the back-transformed data.

By re-parametrization to \( \hat{\delta} = \log \delta \) the bounded optimization problem can be turned into an unbounded one, and solved by standard optimization algorithms.

In practice, \( \mu_x \) and \( \sigma_x \) are rarely known but also have to be estimated from the data. As \( y \) is shifted and scaled ahead of the back-transformation \( W_\delta(\cdot) \), the initial choice of \( \mu_x \) and \( \sigma_x \) affects the optimal choice of \( \delta \). Therefore the optimal triple \( \hat{\tau} = (\hat{\mu}_x, \hat{\sigma}_x, \hat{\delta}) \) must be obtained iteratively.

Discussion of Algorithm 2: Algorithm 2 first computes \( z^{(k)} = (y - \mu_x^{(k)})/\sigma_x^{(k)} \) using \( \mu_x^{(k)} \) and \( \sigma_x^{(k)} \) from the previous step. This normalized output can then be passed to Algorithm 1 to obtain an updated \( \hat{\delta}^{(k+1)} := \hat{\delta}_{GMM} \). Using this new \( \hat{\delta}^{(k+1)} \) one can back-transform \( z^{(k)} \) to \( u^{(k+1)} = W_{\hat{\delta}^{(k+1)}}(z^{(k)}) \), and consequently obtain a better approximation to the “true” latent \( x \) by \( x^{(k+1)} = u^{(k+1)}\sigma_x^{(k)} + \mu_x^{(k)} \). However, \( \hat{\delta}^{(k+1)} \) - and therefore \( x^{(k+1)} \) - 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 \( \hat{x}_{(\mu_x^{(k)}, \sigma_x^{(k)}, \delta^{(k+1)})} \). Thus Algorithm 2 computes new estimates \( \mu_x^{(k+1)} \) and \( \sigma_x^{(k+1)} \) by the sample mean and standard deviation of \( \hat{x}_{(\mu_x^{(k)}, \sigma_x^{(k)}, \delta^{(k+1)})} \), and starts another iteration by passing the updated normalized output \( z^{(k+1)} = (y - \mu_x^{(k+1)})/\sigma_x^{(k+1)} \) to Algorithm 1 to obtain a new \( \hat{\delta}^{(k+2)} \).

It returns the optimal \( \hat{\tau}_{GMM} \) once the estimated parameter triple does not change anymore from one iteration to the next, i.e. if \( \| \tau^{(k)} - \tau^{(k+1)} \| < tol \).

An advantage of the IGMM estimator is that it requires less specific knowledge about the input, and can
be just seen as a data transformation rather than an accurate, “true” statistical model for the data. Usually, it is also faster than the MLE. Once $\hat{\delta}_{\text{IGMM}}$ has been obtained, the latent input data $\mathbf{x}$ can be recovered via the inverse transformation (7). This new data $\mathbf{x}_{\text{IGMM}}$ can then be used to check if $X$ has characteristics of a known parametric distribution $F_X(x \mid \beta)$, and thus is an easy test if $\mathbf{y}$ follows a particular heavy-tail Lambert $W \times F_X$ distribution. It must be noted that tests are too optimistic as $\mathbf{x}_{\text{GMM}}$ will have “nicer” properties regarding $F_X$ than the true $\mathbf{x}$ would have. However, estimating the transformation requires only three parameters and for a large enough sample, losing three degrees of freedom of the test-statistics should not matter in practice.

**Remark 4.2** (IGMM for double-tail Lambert $W \times F_X$). For a double-tail fit the one-dimensional optimization in Algorithm 1 has to be replaced with a two-dimensional optimization

$$
\left( \delta_l, \delta_r \right)_{\text{GMM}} = \arg \min_{\delta_l, \delta_r} h \left( \gamma_2(X) - \hat{\gamma}_2(\mathbf{x}(\mu^*_\tau, \sigma^*_\tau, \delta_l, \delta_r)) \right).
$$

(39)

Algorithm 2 remains unchanged.

5 Simulations

Since this class of distributions is based on transformations of RVs rather than on a manipulation of the pdf or cdf, generating random samples is straightforward (Algorithm 3). This section explores finite sample properties of estimators for $\theta = (\mu_x, \sigma_x, \delta)$ and $(\mu_y, \sigma_y)$ under Gaussian input $X \sim \mathcal{N}(\mu_x, \sigma^2_x)$. In particular, it compares Gaussian MLE (estimation of $\mu_y$ and $\sigma_y$ only), IGMM and Lambert $W \times$ Gaussian MLE, and - for a heavy tail competitor - the median. For IGMM, optimization over the heavy-tail parameter $\delta$ was restricted to $[0, 10]$ as larger values resulted in a numerical overflow in the package.

All results shown below are based on $n = 1,000$ replications.

5.1 Estimating $\delta$ only

Here I show finite sample properties of $\hat{\delta}_{\text{MLE}}$ assuming standard Gaussian input $U \sim \mathcal{N}(0, 1)$, i.e. $\mu_x = 0$ and $\sigma_x = 1$ are fixed and only $\delta$ is estimated given the sample $(z_1, \ldots, z_N) = \mathbf{z} = \mathbf{u} \cdot e^2 u^2$. Theorem 4.1 shows that the MLE of $\delta$ is unique: either a boundary extremum at $\delta = 0$ or the globally optimal solution to (37). The results reported in Table 1 were obtained via numerical optimization restricted to $\delta \geq 0$.\footnote{This restricted optimization problem was again transformed into an unrestricted one, using $\delta = \log \delta$. Then the function \texttt{nlopt} in R was used to obtain the global optimum $\delta^*$, which then was transformed back to $\delta^* = e^{\delta^*}$.}
\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
$N$ & $\delta = 0$ & $\delta = 1/10$ & $\delta = 1/3$ & $\delta = 1/2$ \\
\hline
10  & 0.025 & 0.191 & -0.017 & 0.394 & -0.042 & 0.915 & -0.082 & 1.167 \\
50  & 0.013 & 0.187 & -0.010 & 0.492 & -0.018 & 0.931 & -0.016 & 1.156 \\
100 & 0.010 & 0.200 & -0.010 & 0.513 & -0.009 & 0.914 & -0.006 & 1.225 \\
400 & 0.005 & 0.186 & -0.003 & 0.528 & 0.000 & 0.927 & -0.004 & 1.211 \\
1000 & 0.003 & 0.197 & 0.000 & 0.532 & -0.001 & 0.928 & -0.001 & 1.203 \\
2000 & 0.003 & 0.217 & -0.001 & 0.523 & 0.000 & 0.935 & -0.001 & 1.130 \\
\hline
$N$ & $\delta = 1$ & $\delta = 2$ & $\delta = 5$ \\
\hline
10  & -0.054 & 1.987 & -0.104 & 3.384 & -0.050 & 7.601 \\
50  & -0.017 & 1.948 & -0.009 & 3.529 & 0.014 & 7.942 \\
100 & -0.014 & 2.024 & -0.001 & 3.294 & 0.011 & 7.798 \\
400 & 0.001 & 1.919 & -0.002 & 3.433 & 0.001 & 7.855 \\
1000 & 0.001 & 1.955 & 0.001 & 3.553 & -0.001 & 7.409 \\
2000 & 0.001 & 1.896 & 0.000 & 3.508 & -0.001 & 7.578 \\
\hline
\end{tabular}
\caption{Finite sample properties of $\hat{\theta}_{MLE}$. For each $N$, $\delta$ was estimated $n = 1,000$ times from a random sample $z \sim$ Tukey’s $h$. The left column for each $\delta$ shows the average bias, $\bar{x}_\text{MLE} - \delta$; each right column shows the root mean square error (RMSE) times $\sqrt{N}$.}
\end{table}

I consider various sample sizes $N = 50, 100, 400, 1000, \text{and} 2000$ as well as a wide range of $\delta \in \{0, 1/10, 1/3, 1/2, 1, 2, 5\}$. Table 1 shows that the MLE gives unbiased results for every $\delta$ and settles down (about $N = 100$) to an asymptotic variance, which is increasing with $\delta$. Assuming $\mu_x$ and $\sigma_x$ to be known is unrealistic and thus these finite sample properties are only an indication of the behaviour of the full MLE, $\hat{\theta}_{MLE}$. Nevertheless the results are very remarkable for the extremely heavy-tailed data ($\delta > 1$), for which typically statistical methods break down. One explanation in this behaviour lies in the particular form of the likelihood function (34) and its gradient (37) (Theorem 4.1). Although both depend on $z$, they only do so via $W_\delta(z) = u \sim N(0, 1)$. Hence as long as $\delta$ is in a sufficiently small neighborhood of the true $\delta$ (34) and (37) are functions of almost Gaussian RVs and thus standard asymptotic results still apply.

### 5.2 Estimating all parameters jointly

Here we consider the more realistic scenario where also $\mu_x$ and $\sigma_x$ are unknown. Similarly to the previous section, we consider various sample sizes ($N = 50, 100, 250$, and 1000) and different degrees of heavy tails, $\delta \in \{0, 1/10, 1/3, 1, 1.5\}$, each one representing a particularly interesting situation: i) Gaussian data (does additional - superfluous - estimation of $\delta$ affect other estimates?), ii) slightly heavy-tailed data, all moments $< 10$ exist, iii) fourth moments do not exist anymore, iv) non-finite variance, v) non-existing mean, vi) extremely heavy-tailed data (does the MLE provide useful estimates at all?).

**Remark 5.1** (Computational problems with Lambert’s $W$ function). *For the joint estimation, numerical overflow problems became much more frequent in the estimation if the true $\delta$ was set to larger values than 2. Hence the largest $\delta$ considered here is $\delta = 1.5$. While this issue has been reported to the authors of the *gsl* package, it still remains unresolved.*

Again optimization is restricted to $\delta \geq 0$ for Lambert $W$ MLE and IGMM. Due to numerical problems in IGMM, it is also restricted to a search in $\delta \leq 10$. 

20
| $\delta$ | median | Gaussian MLE | IGMM | Lambert W MLE | NA |
|---|---|---|---|---|---|
| 1 | $\mu_y$ | $\sigma_y$ | $\mu_y$ | $\sigma_y$ | $\mu_y$ | $\sigma_y$ | $\mu_y$ | $\sigma_y$ | $\mu_y$ | $\sigma_y$ | $\mu_y$ | $\sigma_y$ |
| 50 | 0.00 | 0.99 | 0.00 | 0.99 | 0.00 | 0.99 | 0.00 | 0.99 | 0.00 | 0.99 | 0.00 | 0.99 |
| 100 | 0.00 | 0.99 | 0.00 | 0.99 | 0.00 | 0.99 | 0.00 | 0.99 | 0.00 | 0.99 | 0.00 | 0.99 |
| 250 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 |
| 1000 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 |
| 50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 100 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 250 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 1000 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |

(a) Truly Gaussian data: $\delta = 0$

| $\delta = 1/10$ | median | Gaussian MLE | IGMM | Lambert W MLE | NA |
|---|---|---|---|---|---|
| 50 | -0.02 | -0.02 | -0.02 | -0.02 | -0.09 | 0 |
| 100 | 0.00 | 0.00 | 1.02 | 0.08 | 1.18 | 0 |
| 250 | 0.00 | 0.00 | 0.00 | 1.18 | 0.00 | 1.00 |
| 1000 | 0.00 | 0.00 | 0.00 | 1.18 | 0.00 | 1.00 |
| 50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 100 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 250 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 1000 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |

(b) Slightly heavy-tailed data: $\delta = 1/10$

| $\delta = 1/3$ | median | Gaussian MLE | IGMM | Lambert W MLE | NA |
|---|---|---|---|---|---|
| 50 | 0.00 | 0.00 | 1.00 | 0.00 | 1.00 | 0 |
| 100 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 1.00 |
| 250 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 1.00 |
| 1000 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 1.00 |
| 50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 100 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 250 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 1000 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |

(c) No fourth moments: $\delta = 1/3$

| $\delta = 1$ | median | Gaussian MLE | IGMM | Lambert W MLE | NA |
|---|---|---|---|---|---|
| 50 | 0.00 | -0.10 | 24.2 | -0.01 | 1.18 | 0.90 | 0.00 | 1.01 | 0.99 | 0.00 |
| 100 | 0.00 | -0.10 | 24.4 | -0.00 | 1.09 | 0.99 | -0.01 | 1.12 | 1.14 | 0.00 |
| 250 | 0.00 | -0.10 | 24.6 | -0.00 | 1.09 | 0.99 | -0.01 | 1.12 | 1.14 | 0.00 |
| 1000 | 0.00 | -0.10 | 24.8 | -0.00 | 1.09 | 0.99 | -0.01 | 1.12 | 1.14 | 0.00 |
| 50 | 0.50 | 0.50 | 5.2 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 100 | 0.50 | 0.50 | 5.2 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 250 | 0.50 | 0.50 | 5.2 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 1000 | 0.50 | 0.50 | 5.2 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |

(d) Non-existing mean: $\delta = 1$

| $\delta = 1.5$ | median | Gaussian MLE | IGMM | Lambert W MLE | NA |
|---|---|---|---|---|---|
| 50 | -0.02 | -0.02 | -0.02 | -0.02 | -0.09 | 0 |
| 100 | -0.02 | -0.02 | -0.02 | -0.02 | -0.09 | 0 |
| 250 | -0.02 | -0.02 | -0.02 | -0.02 | -0.09 | 0 |
| 1000 | -0.02 | -0.02 | -0.02 | -0.02 | -0.09 | 0 |
| 50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 100 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 250 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 1000 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |

(e) Extreme heavy tails: $\delta = 1.5$

For IGMM the tolerance level was set to $tol = 1.22 \cdot 10^{-4}$ and the Euclidean norm was used. Table 2 summarizes the simulation. Each sub-table is organized as follows: the columns correspond to the parameter

Table 2: Based on $n = 1,000$ replications each. In each sub-table: (first rows) average, (middle rows) proportion of estimates below true value, (bottom rows) empirical standard deviation times $\sqrt{n}$.
estimates of each method; the three main rows are the average over \( n = 1,000 \) replications (top), the proportion of values below the true parameter (middle), and the empirical standard deviation of the parameter estimate around its empirical average times \( \sqrt{N} \) - not around the truth (bottom). Note that the classic Gaussian MLE estimates \( \sigma_y \) directly, while IGMM and the Lambert \( W \times \) Gaussian MLE give estimates for \( \delta \) and \( \sigma_x \), which implicitly give an estimate of \( \sigma_y \) through the functional relation \( \sigma_y(\delta, \sigma_x) = \sigma_x \cdot \frac{1}{\sqrt{(1-2\delta)^{3/2}}} \) (see (21)). For a fair comparison between Gaussian MLE and Lambert \( W \) estimators each sub-table also includes a column for this implied Lambert \( W \) estimate \( \hat{\sigma}_y = \hat{\sigma}_x \cdot \frac{1}{\sqrt{(1-2\delta)^{3/2}}} \). Some entries in the \( \sigma_y \) column contain “\( \infty \)”, even for \( \delta \) which still imply finite variance. This can occur if at least one of the 1,000 estimates for \( \delta \) is greater or equal to 1/2. In this case the implied estimate \( \hat{\sigma}_y = \infty \) and thus the average over all \( n = 1,000 \) estimates is also “\( \infty \)”. The location parameter is the same for the input as for the output for any \( \delta < 1 \), \( \mu_x = \mu_y \), thus they can be directly compared. For \( \delta \geq 1 \), the mean does not exist; the values reported in each sub-table for these \( \delta \) interpret \( \mu_y \) as the median.

**Gaussian data**: \( \delta = 0 \). This parameter choice investigates if imposing the Lambert \( W \) framework, even though its use is superfluous, causes a quality loss in the estimation of the mean \( \mu_y = \mu_x \) or standard deviation \( \sigma_y = \sigma_x \). Furthermore, critical values can be obtained for the null hypothesis \( H_0 : \delta = 0 \) (Gaussian tails). Table 2a shows that under Gaussianity all estimators are unbiased and quickly tend to a large-sample variance. The last four rows show that the sample median can not outperform the sample mean or the Lambert \( W \times \) Gaussian estimators to estimate the location parameter \( \mu_x \), as it has a much larger standard deviation. Additional estimation of \( \delta \) does not affect the efficiency of \( \hat{\mu}_x \) compared to estimating \( \mu \) only (both for IGMM and Lambert \( W \times \) Gaussian MLE). Estimating \( \sigma_y \) directly by Gaussian MLE does not give better results than estimating it via the Lambert \( W \times \) Gaussian MLE: both are unbiased and have similar standard deviation.

**Slightly heavy-tailed**: \( \delta = 1/10 \). Here the RV \( Y \) has slight excess kurtosis \((3 + 2.5082)\) and \( \sigma_y(\delta, \sigma_x = 1) = 1.182 \). The Lambert \( W \) estimators provide unbiased \( \hat{\tau} \), and already for this small degree of heavy tails have smaller empirical standard deviation for the location parameter than the Gaussian MLE or the median. Also using Lambert \( W \) estimators does not give worse estimates for \( \sigma_y \).

**No fourth moment**: \( \delta = 1/3 \). The true standard deviation of the output equals \( \sigma_y(\delta, \sigma_x = 1) = 2.2795 \). For this degree of heavy tails fourth moments do not exist anymore, which reflects in an increasing empirical standard deviation of \( \hat{\sigma}_y \) as \( N \) grows. In contrast, estimates for \( \sigma_x \) are not drifting off. In presence of these large heavy tails the median outperforms the Gaussian MLE and IGMM as it is much less variable than the latter two. However, it is not as good as the Lambert \( W \times \) Gaussian MLE for \( \mu_x \).

**Non-existing mean**: \( \delta = 1 \). Here not only the standard deviation but also the mean is non-finite. Thus both sample moments diverge, and their empirical standard deviation is also growing very quickly with \( \sqrt{N} \). The median still provides a very good estimate for the location, but is again inferior to both Lambert \( W \) estimators, which are unbiased and seem to converge to an asymptotic variance at rate \( \sqrt{N} \).
Extreme heavy tails: $\delta = 1.5$. As in Section 5.1, IGMM and Lambert W MLE still provide unbiased estimates, even though the data is extremely heavy-tailed. The Lambert W MLE is not only unbiased but also has the smallest empirical standard deviation amongst all alternatives. In particular, the Lambert W MLE for the location has an approximately 20% lower standard deviation than the median.

The last column shows that for some $N$ about 1% of the $n = 1,000$ simulations generated invalid likelihood values (NA and $\infty$) since the search for the optimal $\delta$ lead into regions which lead to a numerical overflow in the evaluation of $W_\delta(z)$. For an informative summary table, these few cases were omitted and new simulations added until a full $n = 1,000$ finite estimates were found. Since this only happened in 1% of the cases and also such heavy-tailed data is rarely encountered in practice, this numerical issue is not a real limitation for statistical practice.

5.3 Discussion of the simulation study

This simulation study confirms well-known facts about the sample mean, standard deviation, and median and compares them to finite sample properties of the two Lambert W estimators. The median is known to be a robust estimate of location, which shows here as its quality does not depend on the thickness of the tails.

IGMM is unbiased estimator for $\tau$ independent of the value of $\delta$. As expected the Lambert W MLE for $\theta$ has the best properties: it is unbiased for all $\delta$, and has the same empirical standard deviation as the Gaussian MLE for small $\delta$, and lower empirical standard deviation than the median for large $\delta$. For $\delta = 0$ it performs as well as the classic sample mean and standard deviation. Hence additional estimation of $\delta$ does not affect the quality of the estimates for location and scale if $\delta$ is small, but greatly improves the inference for large $\delta$.

Thus the only advantage of estimating $\mu_y$ and $\sigma_y$ by the sample mean and standard deviation of $y$ is its speed; otherwise the Lambert W MLE is at least as good as Gaussian MLE and clearly outperforms it for heavy-tailed data.

6 Applications

Tukey’s $h$ distribution has already proven useful to model heavy-tailed data, but estimation has been limited to quantile fitting or methods of moments estimators (Field, 2004; Fischer, 2010; Todd C. Headrick and Sheng, 2008). Theorem 2.8 puts us in the position to compute the likelihood of the data in terms of $\theta = (\beta, \delta, \delta_r)$ and estimate $\theta$ by ML.

This section shows the usefulness of the presented methodology on simulated as well as real world data: Section 6.1 demonstrates the Gaussianizing$^7$ capabilities on a Cauchy random sample from the Introduction, Section 6.2 shows that heavy tail Lambert $W \times$ Gaussian distributions provide an excellent fit to daily S&P 500 log-return series, and finally Section 6.3 shows that removing heavy tails reveals hidden patterns in power-law type data.

$^7$Function Gaussianise in the LambertW package.
6.1 Estimating the location of a Cauchy using the sample mean

It is well-known that the sample mean $\bar{y}$ is a poor estimate of the location parameter of a Cauchy distribution, since the sampling distribution of $\bar{y}$ is again a Cauchy; in particular, its variance does not go to 0 for $n \to \infty$.

Heavy-tailed Lambert $W \times$ Gaussian distributions have similar properties to a Cauchy for $\delta \approx 1$. The mean of $X$ equals the location of $Y$, due to symmetry around $\mu_x$ (for all $\delta \geq 0$) and $c$, respectively. Thus we can estimate $\tau$ from the Cauchy sample $y$ and transform it to $x_{\tau}$, estimate $\mu_x$ from $x_{\tau}$, and thus automatically get an estimate of the location $c$ from the Cauchy distribution we started with.

The data $y \sim C(0, 1)$ in Fig. 2a shows heavy tails with two very extreme (positive) samples. A maximum likelihood (ML) fit using a Cauchy distribution gives $\hat{c} = 0.028(0.055)$ and scale estimate $\hat{s} = 0.864(0.053)$, where standard errors are given in parenthesis. A Lambert $W \times$ Gaussian MLE gives $\hat{\mu}_x = 0.030(0.0547)$, $\hat{\sigma}_x = 1.054(0.0717)$, and $\hat{\delta} = 0.861(0.0819)$. Thus both fits correctly fail to reject $\mu_x = c = 0$. Table 3a shows summary statistics on both samples. Since the Cauchy distribution does not have a well-defined mean, $\bar{y} = 2.304(2.101)$ is not a useful estimate. However, the sample average $\bar{x}_{\tau,MLE} = 0.033(0.0472)$ correctly fails to reject a zero location for $y$. The Gaussianised version $x_{\tau,MLE}$ also features additional characteristics of a Gaussian sample (symmetric, no excess kurtosis), and even the null hypothesis of Normality cannot be rejected (p-value $\geq 0.5$).

Although being a toy example, it shows that removing (strong) heavy tails from data works and provides new “nice” data which can then be used for more refined models.

6.2 Modelling heavy tails in financial return series: S&P 500 case study

A lot of financial data displays negative skewness and excess kurtosis. Since financial data is in general not i.i.d. it is often modelled with a (skew) student-t distribution underlying a (generalized) auto-regressive conditional heteroskedastic (GARCH) (Bollerslev, 1986; Engle, 1982) or a stochastic volatility (SV) model (Deo, Hurvich, and Lu, 2006; Melino and Turnbull, 1990). Using the Lambert $W$ approach we can build upon the knowledge and implications of Gaussianity (and avoid deriving properties of a GARCH or SV model with heavy-tailed, skewed innovations), and simply “Gaussianise” the data $y$ before fitting more complex models (e.g. GARCH or SV models). Time series models based on Lambert $W \times$ Gaussian white noise are far beyond the scope and focus of this work, but can be a direction of future research. For this exploratory analysis I consider the unconditional distribution only.

---

8 Function fitdistr in R package MASS.
Table 3: Summary statistics for observed (heavy-tailed) \( y \) and back-transformed (Gaussianised) data \( x_{\hat{\tau} MLE} \).

|       | \( y \) | \( x_{\hat{\tau} MLE} \) | \( y \) | \( x_{\hat{\tau} MLE} \) | \( y \) | \( x_{\hat{\tau} MLE} \) |
|-------|--------|--------------------------|--------|--------------------------|--------|--------------------------|
| Min   | -161.59 | -3.16                    | Min    | 20                       | Min    | 20                       |
| Max   | 952.95  | 3.81                     | Max    | 231300                   | Max    | 157                      |
| Mean  | 2.30    | 0.03                     | Mean   | 689.4                    | Mean   | 89.0                     |
| Median| 0.04    | 0.04                     | Median | 87                       | Median | 87                       |
| Stdev | 46.98   | 1.06                     | Stdev  | 6520.6                   | Stdev  | 27.0                     |
| Skewness | 17.43 | 0.12                     | Skewness | -0.296                   | Skewness | -0.039                   |
| Kurtosis | 343.34 | 3.21                     | Kurtosis | 582.1                    | Kurtosis | 1.9                      |
| Shapiro-Wilk | * 0.71 | * 0.241                   | Shapiro-Wilk | ** 0.51 | Shapiro-Wilk | ** 0.51 |
| Anderson-Darling | ** 0.51 | * 0.181 | Anderson-Darling | ** 0.51 | Anderson-Darling | ** 0.51 |

(a) \( y \sim C(0,1) \) (Section 6.1)  
(b) \( y = \text{S&P 500} \) (Section 6.2)  
(c) \( y = \text{solar flares} \) (Section 6.3)

** stands for \(< 10^{-16}\); * for \(< 2.2 \cdot 10^{-16}\).

Figure 7a shows the S&P 500 log-returns with a total of \( N = 2,780 \) daily observations.\(^9\) Table 3c confirms the heavy tails (sample kurtosis 7.70), whereas it indicates negative skewness (−0.296). As the sample skewness \( \hat{\gamma}_1(y) \) is very sensitive to outliers in the tails, we should test for symmetry by fitting a skewed distribution and testing its skewness parameter(s) for zero. In case of the double-tail Lambert \( W \times \) Gaussian this means to test \( H_0 : \delta_l = \delta_r \) versus \( H_1 : \delta_l \neq \delta_r \). Since the likelihood can now be computed by (30), we can do this using a likelihood ratio test with one degree of freedom (3 versus 4 parameters). The log-likelihood of the double-tail Lambert \( W \times \) Gaussian fit (Table 4a) equals \(-3606.005 = -2972.276 + (-633.7287) \) (input + penalty) and using only one \( \delta \) for both sides it equals \(-3606.554 = -2972.276 + (-633.7287) \) (input + penalty). Comparing twice their difference to a \( \chi^2 \) distribution with one degree of freedom gives a p-value of 0.29. Additionally, I fit\(^10\) a skew-t distribution (Azzalini and Capitanio, 2003), with location \( c \), scale \( s \), shape \( \alpha \), and degrees of freedom \( \nu \). Also here \( \hat{\alpha} \) is not significantly different from zero (Table 4b). Thus in both cases we can not reject symmetry.

Assume we have to make a decision if we should trade a certificate replicating the S&P 500. Since we can either buy or sell, it is not important if the average return is positive or negative, as long as it is significantly different from zero.

### 6.2.1 Gaussian fit to observed data

If we ignore heavy tails and estimate \((\mu_y, \sigma_y)\) by Gaussian MLE, \( \hat{\mu}_y = 0 \) can not be rejected on a \( \alpha = 1\% \) level (Table 4e). As OLS over-estimates the variance in presence of heavy tails, this test conclusion should not be trusted.

### 6.2.2 Heavy tail fit to observed data

If we account for heavy tails, we can impose any heavy-tailed location-scale distribution, estimate its parameters jointly, and then test the location parameter being equal to zero. The standard errors can be computed

---

\(^9\)R package \texttt{MASS}, dataset \texttt{SP500}.

\(^{10}\)Function \texttt{st.mle} in the R package \texttt{sn}.
by the inverse of the numerically obtained Hessian of the log-likelihood. Both the heavy tail Lambert W × Gaussian (Table 4c) and student-t fit (Table 4d) lead to a rejection of the zero mean null (p-values, 0.0001 and 0.00003, respectively). The standard errors for the location parameter are essentially the same, which supports the claim of a “true” standard error of 3.65 for the location parameter.

Although location and scale estimates are almost identical, the estimates describing the tails lead to very different conclusions: while for \( \hat{\nu} = 3.71 \) only moments up to order 3 exist, in the Lambert W × Gaussian case moments up to order 5 exist (1/0.172 = 5.81). This is especially noteworthy as many theoretical results in the (financial) time series literature rely on the assumption of finite fourth moments (Mantegna and Stanley, 1998; Zadrozny, 2005); consequently empirical studies try to test if financial data actually satisfy this assumption (Cont, 2001; Huisman, Koedijk, Kool, and Palm, 2001). For this particular dataset student’s t and Tukey’s h distribution give different empirical answers to the same question. Since empirical analysis are often based on student’s t distribution (Wong et al., 2009), it might be worthwhile to re-examine their findings in light of heavy tail Lambert W × Gaussian distributions.

6.2.3 “Gaussianising” the observed data

A typical statistical analysis to estimate distribution parameters would conclude here. Using Lambert’s W function we can analyse the input \( x_{\tau_{MLE}} \) to test if a Lambert W × Gaussian distribution is indeed appropriate. Figure 7b shows that the back-transformed data \( x_{\tau_{MLE}} \) is indistinguishable from a Gaussian sample. Not even one Normality test on \( x_{\tau_{MLE}} \) (Anderson Darling, Cramer-von-Mises, Shapiro-Francia, Shapiro-Wilk; see Thode (2002)) can reject Gaussianity: p-values are 0.181, 0.184, 0.311, and 0.241, respectively. Table 3c also shows that Lambert W “Gaussianiziation” was successful: empirical kurtosis is close to 3, and although the sample skewness is still negative, a value of −0.039 is within the typical variation for a Gaussian sample.
Thus the heavy-tailed Lambert $W \times$ Gaussian (= Tukey’s $hh$) distribution

\[ Y = \left( U e^{0.172 U^2} \right) 0.705 + 0.055, \quad U = \frac{X - 0.055}{0.705}, \quad U \sim N(0, 1) \]  

is an adequate (unconditional) model for the S&P 500 log-returns $y$. For the trading decision this means that the expected return is significantly different from zero, and thus a trading certificate should be bought ($\hat{\mu}_x = 0.055 > 0$).

6.2.4 Gaussian MLE for Gaussianised data

For $\delta_l = \delta_r = \delta < 1$, the expectation of the input equals the expectation of the output. We can therefore replace the original test of $\mu_y = 0$ versus $\mu_y \neq 0$ for a non-Gaussian sample $y$, with the very well understood hypothesis test $\mu_x = 0$ versus $\mu_x \neq 0$ for the Gaussian sample $x_{\hat{\tau} MLE}$. In particular, standard errors based on $\hat{\sigma} / \sqrt{N}$ - and thus t and p-values - should be closer to the “truth” (Table 4c and 4d) than a Gaussian MLE on the non-Gaussian $y$ (Table 4e). Table 4f shows that the standard errors for $\hat{\mu}_x$ are indeed much closer; they are even a little bit too small compared to the heavy-tailed versions. Since the “Gaussianising” transformation was estimated, treating $x_{\hat{\tau} MLE}$ as if it was original data is too optimistic regarding its Gaussianity (compare to the penalty term (32) in the overall likelihood (30)).

This example confirms that if a model and its theoretical properties are based on Gaussianity, but the observed data is heavy-tailed, then Gaussianising the data first gives more reliable inference than applying the Gaussian methods to the original, heavy-tailed data (Fig. 1). Clearly, a joint estimation of the model parameters based on Lambert $W \times$ Gaussian errors (or any other heavy-tailed distribution) would be optimal, but theoretical properties and estimation procedures may not have been derived or implemented yet, or are simply not known to those applied researchers who are non-experts in the field of skewed, heavy-tailed
statistics. The Lambert Way to Gaussianise data thus provides a pragmatic method to improve statistical inference on heavy-tailed data, while still preserving the ease of usage and interpretation of Gaussian models.

6.3 Removing power laws: analysis of solar flares

The previous section focused on Lambert W × FX distributions as a “true” model for the data y. Here I consider it merely as a data transformation to remove heavy tails. In the same way as scaling y to zero-mean, unit-variance data \((y - \bar{y})/\sigma_y\) does not necessarily mean we believe the underlying process is Gaussian, we can also convert y to \(x_\tau = W_\tau(y)\) without assuming that the data is actually Lambert W × Gaussian. While \(x_\tau\) might not be as easily interpretable as the original data, it can be helpful for exploratory data analysis (EDA), as the eye is a bad judgement to detect regularities corrupted by heavy tails. Removing them can reveal hidden patterns and thus greatly improve the accuracy of statistical models for y.

Figure 8: Peak X-ray count rates of solar flares.
Here I study solar flare X-ray count rates (Clauset et al., 2009; Newman, 2005). The data\textsuperscript{11} were collected approximately four times a day from Feb. 1980 until Nov. 1989 giving \( T = 12,773 \) observations. See Dennis, Orwig, Kennard, Labow, Schwartz, Shaver, and Tolbert (1991) for a detailed description of this dataset and its scientific background.

The X-ray count rates exhibit a strong right heavy tail (Fig. 8a), which makes more detailed visual inspection as well as simple exploratory analysis hard. A zoom to \( y_i \leq 400 \) in Fig. 8b shows that a lot of values lie between 50 and 100 and this level drops off at the end of the observation cycle. The drop is not an intrinsic property of the data but due to a decreasing sensitivity of the X-ray detectors over the course of 10 years (Dennis et al., 1991).

For the sake of comparison with (Clauset et al., 2009; Newman, 2005) most estimates reported here are based on all \( T = 12,773 \) observations. Figures 8b, 8d, and 8e show a separate density estimate for the first 4,000 and last 2,273 observations, and while the estimates change, the qualitative findings do not.

Clauset et al. (2009) find that a power-law (\( \tilde{a} = 1.79(0.02) \)) with cut-off (\( \tilde{y}_{\min} = 323(89) \)) gives the best fit amongst various alternatives. However, this first EDA might not be complete: not only visually heavy tails can obscure underlying non-trivial structure, but also estimates - such as the power law fit or non-parametric density estimates (Fig. 8b and 8c) - are affected by the heavy right tail. Here I show that Gaussianising this data reveals new insights for the data-generating process, with a new interpretation for the optimality of the cut-off value.

An MLE fit gives \( \tilde{r} = (\tilde{\mu}_x, \tilde{\sigma}_x, \tilde{\delta}_l, \tilde{\delta}_r) = (86.97, 26.80, 0, 2.37) \), which shows that only the right tail needs a Gaussianising transformation.\textsuperscript{12} The second row of Fig. 8 shows EDA for the Gaussianised data. Removing the heavy right tail (estimated tail index \( 1/2.373 = 0.421 \)) reveals a bimodal structure. As the inverse transformation \( W_\tau(y) \) is monotonic in \( y \) for any fixed \( \tau \), the bimodal density is an intrinsic property of the data, not an artefact from the transformation.

The bimodal structure of the input also gives an additional meaning to \( \tilde{y}_{\min} = 323 \). The Gaussianised cut-off value equals \( W_\tau(323) = 121.16 \) with the transformed standard deviation interval \( [117.74, 123.37] \) (corresponding to \( 323 \pm 89 \)). Fitting a two component Gaussian mixture model to \( x_\tau \) yields

\[
\hat{\lambda} \mathcal{N}_1(67.10, 14.04^2) + (1 - \hat{\lambda}) \mathcal{N}_2(113.12, 14.27^2), \quad \text{with} \quad \hat{\lambda} = 0.52, \tag{41}
\]

and corresponding optimal decision boundary between the classes of 90.48. The mean of the larger component, 113.12, lies within one standard deviation of the optimal Gaussianised cut-off 121.16: for lower cut-offs the left-tail of the larger component - or for much lower cut-offs even the first component - would counteract the power-law decay of the upper X-ray count rates.

As already mentioned above, this analysis is not intended to describe the true underlying distribution of the X-ray data and parameter fits are not meant to be used for inference for the physics of X-rays; it should rather show new insights that can be gained by Gaussianizing the solar flare dataset. Future research based on this additional information can lead to new physical interpretations of the statistical properties X-ray

\textsuperscript{11}Dataset SolarFlares in the LambertW package. Also available at \url{http://tuvalu.santafe.edu/~aaronc/powerlaws/data.htm} and \url{http://umbra.nascom.nasa.gov/smrm/hrbbs.html}.

\textsuperscript{12}For comparison Fig. 8d also shows the back-transformed data \( x_\tau \) using the same \( \delta \) on each tail (\( \tilde{\gamma}_1 = (74.46, 26.32, 1.53) \)). However, due to the clear right heavy tail I will continue with the (\( \delta_l, \delta_r \)) transformation.
count rates, see for example Aschwanden (2010).

7 Discussion and Outlook

In this work I adapt the skewed Lambert W input/output framework to introduce heavy tails in continuous RVs $X \sim F_X(x)$ and provide closed-form expressions of the cdf and pdf. For Gaussian input this not only contributes to the existing literature on Tukey’s $h$ distribution, but also gives convincing empirical results: skewed, unimodal data with heavy tails can be transformed to behave like Gaussian data/RVs. Properties of a Gaussian model $M_N$ on the back-transformed data mimic the features of the “true” skewed, heavy-tailed model $M_G$ very closely.

Since Gaussianity is the single most typical, and necessary, assumption in many areas of statistics, machine learning, and signal processing, future research can take many directions. From a theoretical perspective the properties of Lambert W $\times F_X$ distributions viewed as a generalization of already well-known distributions $F$ can be studied - ignoring the possibility to actually back-transform the data. This area will profit from the immense literature on the Lambert W function, which has been discovered only recently by the statistics community. Empirical analysis can focus on actually transforming the data and compare the performance of the approximation versus the correct, joint analysis. The simple comparisons here showed that doing the approximate inference is comparable with the direct heavy tail Modelling, and so provides an easy tool to improve inference for heavy-tailed data in statistical practice.

It is also worth pointing out the educational value of this transformation based approach, as it can be taught in introductory classes and thus already early on promote heavy-tailed statistics with enduring value.

I also provide the R package LambertW, publicly available at CRAN, to facilitate the use of Lambert W $\times F_X$ distributions in practice.

Acknowledgment

I want to thank Andrew F. Siegel who brought Tukey’s $h$ distributions to my attention, and Brian R. Dennis who gave detailed background information and suggestions on the analysis of the solar flares dataset.
References

Achim, A., P. Tsakalides, and A. Bezerianos (2003). SAR image denoising via Bayesian wavelet shrinkage based on heavy-tailed modeling. *Geoscience and Remote Sensing, IEEE Transactions on* 41(8), 1773 – 1784.

Ahn, H. and R. E. Feldman (1999). Optimal Filtering of a Gaussian Signal in the Presence of Levy Noise. *SIAM Journal on Applied Mathematics* 60(1), pp. 359–369.

Aschwanden, M. J. (2010). The state of self-organized criticality of the sun during the last three solar cycles. ii. theoretical model.

Aysal, T. C. and K. E. Barner (2006). Second-order heavy-tailed distributions and tail analysis. *IEEE Transactions on Signal Processing* 54(7), 2827–2832.

Azzalini, A. and A. Capitanio (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistical Society ser B* 65, 367–389.

Baek, C. and V. Pipiras (2010). Estimation of parameters in heavy-tailed distribution when its second order tail parameter is known. *Journal of Statistical Planning and Inference* 140(7), 1957 – 1967.

Blum, R. S., Y. Zhang, B. M. Sadler, and R. J. Kozick (1999). On the approximation of correlated non-gaussian noise pdfs using gaussian mixture models. In American University, Washington DC.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307 – 327.

Breusch, T. S., J. C. Robertson, and A. H. Welsh (1997). The emperor’s new clothes: a critique of the multivariate t regression model. *Statistica Neerlandica* 51(3), 269–286.

Brockwell, P. J. and R. A. Davis (1998, September). *Time Series: Theory and Methods (Springer Series in Statistics)*. Springer.

Clauset, A., C. R. Shalizi, and M. E. J. Newman (2009). Power-law distributions in empirical data. *SIAM Review* 51, 661–703.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance* 1, 223–236.

Corless, R. M., G. H. Gonnet, D. E. G. Hare, and D. J. Jeffrey (1993). On Lambert’s W function. preprint: citeseer.ist.psu.edu/corless96lambert.html.

Dennis, B. R., L. E. Orwig, G. S. Kennard, G. J. Labow, R. A. Schwartz, A. R. Shaver, and A. K. Tolbert (1991). The complete Hard X Ray Burst Spectrometer event list, 1980-1989. Available http://adsabs.harvard.edu/abs/1991chxb.book....D and http://umbra.nascom.nasa.gov/smm/hxrbss.html.

Deo, R., C. Hurvich, and Y. Lu (2006). Forecasting Realized Volatility Using a Long Memory Stochastic VolatilityModel: Estimation, Prediction and Seasonal Adjustment. *Journal of Econometrics* 127, 29 – 58.
Dutta, K. K. and D. Babbel (2002). On Measuring Skewness and Kurtosis in Short Rate Distributions: The Case of the US Dollar London Inter Bank Offer Rates. Center for financial institutions working papers, Wharton School Center for Financial Institutions, University of Pennsylvania.

Engle, R. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica* 50, 987 – 1008.

Fernandez, C. and M. F. J. Steel (1999). Multivariate student-t regression models: Pitfalls and inference. *Biometrika* 86, 153–167.

Field, C. A. (2004). Using the gh distribution to model extreme wind speeds. *Journal of Statistical Planning and Inference* 122(1-2), 15 – 22.

Fischer, M. (2010). Generalized Tukey-type distributions with application to financial and teletraffic data. *Statistical Papers* 51, 41–56. 10.1007/s00362-007-0114-z.

Friedland, B. (2005). *Control System Design: An Introduction to State-Space Methods*. Dover Publications, Incorporated. (Dover Books on Engineering).

Galassi, M., J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, and F. Rossi (2011). *GNU Scientific Library Reference Manual* (3rd ed.). ISBN 0954612078; [http://www.gnu.org/software/gsl/](http://www.gnu.org/software/gsl/).

Gidlund, M. and N. Debernardi (2009). Scheduling performance of heavy-tailed data traffic in wireless high-speed shared channels. In *Proceedings of the 2009 IEEE conference on Wireless Communications & Networking Conference*, WCNC'09, Piscataway, NJ, USA, pp. 1818–1823. IEEE Press.

Goerg, G. M. (2011). Lambert W Random Variables - A New Family of Generalized Skewed Distributions with Applications to Risk Estimation. *Forthcoming in The Annals of Applied Statistics* 5(-), p. —. [arxiv.org/abs/0912.4554](http://arxiv.org/abs/0912.4554).

Granger, C. W. J. and R. Joyeux (2001). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* 1, 15 – 30.

Huisman, R., K. G. Koedijk, C. J. M. Kool, and F. Palm (2001). Tail-index estimates in small samples. *Journal of Business & Economic Statistics* 19(2), 208–16.

Hwang, J., S. Lay, and A. Lippman (1994). Nonparametric multivariate density estimation: A comparative study. *IEEE Trans. Signal Processing* 42, 2795–2810.

Idan, M. and J. L. Speyer (2010, June). Cauchy Estimation for Linear Scalar Systems. *IEEE Transactions on Automatic Control* 55(6), 1329 – 1342.

Ilow, J. (2000). Forecasting network traffic using farima models with heavy tailed innovations. *Acoustics, Speech, and Signal Processing, IEEE International Conference on* 6, 3814–3817.

Jodr´a, P. (2009). A closed-form expression for the quantile function of the Gompertz-Makeham distribution. *Math. Comput. Simul.* 79, 3069–3075.
Kim, T.-H. and H. White (2003). On more robust estimation of skewness and kurtosis: Simulation and application to the sp500 index.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation* (2nd ed.). Springer Texts in Statistics.

Liu, C. and D. B. Rubin (1995). MI estimation of the t distribution using em and its extensions, ecm and ecm. *Statistica Sinica* 5, 19–39.

Liu, H., J. Lafferty, and L. Wasserman (2009). The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. *Journal of Machine Learning Research* 10, 2295–2328.

Maiboroda, R. and N. Markovich (2004). Estimation of heavy-tailed probability density function with application to web data. *Computational Statistics* 19, 569–592. 10.1007/BF02753913.

Mantegna, R. N. and H. E. Stanley (1998). Modeling of financial data: Comparison of the truncated Lévy flight and the ARCH(1) and GARCH(1,1) processes. *Physica A: Statistical and Theoretical Physics* 254(1-2), 77 – 84.

Markovich, N. M. (2005, February). Accuracy of transformed kernel density estimates for a heavy-tailed distribution. *Autom. Remote Control* 66, 217–232.

Melino, A. and S. M. Turnbull (1990). Pricing foreign currency options with stochastic volatility. *Journal of Econometrics* 45(1-2), 239 – 265.

Morgenthaler, S. and J. W. Tukey (2000). Fitting quantiles: Doubling, hr, hq, and hhh distributions. *Journal of Computational and Graphical Statistics* 9(1), pp. 180–195.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf’s law. *Contemporary Physics* 46, 323–351.

Nezafat, M. and H. Amindavar (2001). A new enhanced morphological filter and signal recovery. In EUROCON’2001, *Trends in Communications, International Conference on*, Volume 1, pp. 91 –94 vol.1.

Nowicka-Zagrajek, J. and R. Weron (2002). Modeling electricity loads in California: ARMA models with hyperbolic noise. *Signal Process.* 82, 1903–1915.

Pakes, A. G. (2011, February). Lambert’s w, infinite divisibility and poisson mixtures. *Journal of Mathematical Analysis and Applications*.

Palma, W. and M. Zevallos (2011). Fitting non-gaussian persistent data. *Applied Stochastic Models in Business and Industry* 27(1), 23–36.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rathie, R. N. and P. Silva (2011). Applications of Lambert W Function. *International Journal of Applied Mathematics & Statistics* 23(D11), –.

Rosenlicht, M. (1969). On the explicit solvability of certain transcendental equations. *Pub. Math. Institut des Hautes Etudes Scientifiques* 36, 15 – 22.
Smith, V. K. (1973). Least Squares Regression with Cauchy Errors. *Oxford Bulletin of Economics and Statistics* 35(3), 223–31.

Swami, A. (2000). Non-gaussian mixture models for detection and estimation in heavy-tailed noise. *Acoustics, Speech, and Signal Processing, IEEE International Conference on 6*, 3802–3805.

Takada, T. (2001). Nonparametric density estimation: A comparative study. *Economics Bulletin* 3(16), 1–10.

Thode, Jr., H. C. (2002). *Testing for Normality*. CRC Press.

Todd C. Headrick, R. K. K. and Y. Sheng (2008). Parametric Probability Densities and Distribution Functions for Tukey g-and-h Transformations and their Use for Fitting Data. *Applied Mathematical Sciences* 2(9), 449 – 462.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Valluri, S. R., D. J. Jeffrey, and R. M. Corless (2000). Some Applications of the Lambert W Function to Physics. *Canadian Journal of Physics* 78, 823 – 831.

Vázquez, A., J. G. Oliveira, Z. Dezső, K.-I. Goh, I. Kondor, and A.-L. Barabási (2006). Modeling bursts and heavy tails in human dynamics. *Phys. Rev. E* 73(3), 036127.

Wasserman, L. (2007, May). *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer.

Wong, C. S., W. S. Chan, and P. L. Kam (2009). A Student t-mixture autoregressive model with applications to heavy-tailed financial data. *Biometrika* 96(3), 751–760.

Yan, J. (2005). Asymmetry, fat-tail, and autoregressive conditional density in financial return data with systems of frequency curves. [http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.76.2741](http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.76.2741).

Zadrozny, P. A. (2005). *Necessary and Sufficient Restrictions for Existence of a Unique Fourth Moment of a Univariate GARCH(p,q) Process*, Volume 20 of *Advances in Econometrics*, Chapter –, pp. 365–379. Emerald Group Publishing Limited. [http://ideas.repec.org/p/ces/ceswps/_1505.html](http://ideas.repec.org/p/ces/ceswps/_1505.html).
A Auxiliary results

A.1 Properties of $W_\delta(z)$

The function $W_\delta(z)$ is the building block of Lambert $W \times F_X$ distributions. This section lists useful properties of $W_\delta(z)$ as a function of $z$ as well as a function of $\delta$.

Properties A.1. For $\delta = 0$,

$$W_\delta(z_i) |_{\delta=0} = z_i, \quad W'(\delta z^2_i) |_{\delta=0} = z_i^2, \quad \text{and} \quad W(\delta z^2_i) |_{\delta=0} = 0. \quad (42)$$

By definition it also holds

$$\frac{W_\delta(z)}{z} = e^{-\frac{1}{2} W_\delta(z)^2}, \quad (43)$$

and therefore

$$\log \frac{W_\delta(z)}{z} = -\frac{\delta}{2} W_\delta(z)^2 = -\frac{W(\delta z^2)}{2}. \quad (44)$$

Lemma A.2 (Derivative of $W_\delta(z)$ with respect to $z$). It holds

$$\frac{d}{dz} W_\delta(z) = -\frac{W_\delta(z)}{z} \cdot \frac{1}{1 + \delta W_\delta(z)^2} \cdot \frac{1}{1 + W(\delta z^2)} \quad (45)$$

Proof. One of the many interesting properties of the Lambert W function relates to its derivative which satisfies

$$W'(z) = \frac{W(z)}{z(1+W(z))} = \frac{1}{e^{W(z)}(1+W(z))}, \quad z \neq 0, -1/e. \quad (46)$$

Hence,

$$\frac{d}{dz} W(\delta z^2) = W'(\delta z^2) \cdot 2z = \frac{W(\delta z^2)}{\delta z^2 (1+W(\delta z^2))} = \frac{2W(\delta z^2)}{\delta (1+W(\delta z^2))} \quad (47)$$

Therefore,

$$\frac{d}{dz} W_\delta(z) = \frac{1}{2} \left( \frac{1}{\delta} W(\delta z^2) \right)^{-1/2} \cdot \frac{d}{dz} \frac{W(\delta z^2)}{\delta} \quad (48)$$

$$= \frac{1}{2} \left( \frac{1}{\delta} W(\delta z^2) \right)^{-1/2} \cdot \frac{2W(\delta z^2)}{\delta z (1+W(\delta z^2))} \quad (49)$$

$$= \frac{1}{\delta^{1/2}} (W(\delta z^2))^{-1/2} \cdot \frac{W(\delta z^2)}{z (1+W(\delta z^2))} \quad (50)$$

As $W(\delta z^2) = \delta u^2$ the last line simplifies to

$$\frac{1}{\delta^{1/2} \frac{1}{\delta^{1/2}}} \cdot \frac{\delta u^2}{z (1+\delta u^2)} = \frac{u}{z (1+\delta u^2)}. \quad (51)$$

Now use again $u = W_\delta(z)$. 

\[\square\]
Lemma A.3 (Derivative of $W_\delta(z)^2$ with respect to $\delta$). For all $z \in \mathbb{R}$

$$\frac{\partial}{\partial \delta} [W_\delta(z)]^2 = -\frac{1}{1 + W(\delta z^2)} W_\delta(z)^4 \leq 0. \tag{52}$$

Proof. By definition $[W_\delta(z)]^2 = \frac{W(\delta z^2)}{\delta}$. Thus

$$\frac{\partial}{\partial \delta} \frac{W(\delta z^2)}{\delta} = \frac{\delta \frac{\partial}{\partial z} W(\delta z^2) - W(\delta z^2) \cdot 1}{\delta^2} \tag{53}$$

$$= \frac{\delta W'(\delta z^2) z^2 - W(\delta z^2)}{\delta^2} \tag{54}$$

$$= \frac{\delta W(\delta z^2)}{\delta^2} - \frac{W(\delta z^2)}{\delta^2} z^2 - W(\delta z^2) \tag{55}$$

$$= \frac{W(\delta z^2)}{1 + W(\delta z^2)} - W(\delta z^2) \tag{56}$$

$$= -\frac{W(\delta z^2)^2}{\delta^2} \tag{57}$$

$$= -\frac{1}{1 + W(\delta z^2)} [W_\delta(z)]^4 \tag{58}$$

Since both terms are non-negative for all $z \in \mathbb{R}$ the result follows. \hfill \Box

That is $W_\delta(z)^2$ is a decreasing function in $\delta$ for every $z \in \mathbb{R}$, i.e. the more we remove heavy tails the more $z$ gets shrunk (non-linearly) towards $0 = \lim_{\delta \to \infty} W_\delta(z)$. In particular, $[W_\delta(z)]^2 < z^2 \iff \frac{W_\delta(z)}{z} < 1$ and $\frac{W_{\delta+z}(z)}{z} < \frac{W_\delta(z)}{z}$ for $\delta \geq 0$ and $\epsilon > 0$.

Lemma A.4 (Derivative of $W_\delta(z)$ with respect to $\delta$). It holds

$$\frac{\partial W_\delta(z)}{\partial \delta} = -\frac{1}{2} \frac{1}{1 + W(\delta z^2)} W_\delta(z)^3. \tag{59}$$

Proof.

$$\frac{\partial W_\delta(z)}{\partial \delta} = \operatorname{sgn}(z) \frac{\partial}{\partial \delta} \left( \frac{W(\delta z^2)}{\delta} \right)^{1/2} \tag{60}$$

$$= \operatorname{sgn}(z) \frac{1}{2} \left( \frac{W(\delta z^2)}{\delta} \right)^{-1/2} \frac{\partial}{\partial \delta} \frac{W(\delta z^2)}{\delta} \tag{61}$$

$$= \frac{1}{2} \frac{1}{W_\delta(z)} \frac{\partial}{\partial \delta} [W_\delta(z)]^2 \tag{62}$$

$$= -\frac{1}{2} \frac{1}{1 + W(\delta z^2)} W_\delta(z)^3, \tag{63}$$

where the last line follows by Lemma A.3. \hfill \Box
A.2 Properties of the penalty term \( \log R(\delta \mid z_i) \) for standard Gaussian input

For \( \mu_x = 0 \) and \( \sigma_x = 1 \) the penalty term equals \( (y_i = z_i) \)

\[
R(\delta \mid z_i) = \frac{W_\delta(z_i)}{z_i[1 + \delta^2(W_\delta(z_i))^2]} = \frac{W_\delta(z_i)}{z_i[1 + W(\delta z_i^2)]} \quad \text{(64)}
\]

and \( \log R(\delta \mid z_i) = \log \frac{W_\delta(z_i)}{z_i} - \log [1 + W(\delta z_i^2)] \)

\[
= -\frac{W_\delta(z_i)}{2} - \log [1 + W(\delta z_i^2)] \quad \text{(65)}
\]

**Lemma A.5** (Derivative of \( \log R(\delta \mid z) \) with respect to \( \delta \)). For all \( \delta \geq 0 \) and all \( z \in \mathbb{R} \)

\[
\frac{\partial \log R(\delta \mid z)}{\partial \delta} = -z^2 W'(\delta z^2) \left( \frac{1}{2} + \frac{1}{1 + W(\delta z^2)} \right) \leq 0. \quad \text{(67)}
\]

**Proof.** We have

\[
\frac{\partial \log R(\delta \mid z)}{\partial \delta} = \frac{1}{W_\delta(z)} \frac{\partial W_\delta(z)}{\partial \delta} - \frac{1}{1 + W(\delta z^2)} W'(\delta z^2) z^2
\]

\[
= \frac{1}{W_\delta(z)} \left( -\frac{1}{2} \frac{1}{1 + W(\delta z^2)} W_\delta(z)^2 \right) - \frac{1}{1 + W(\delta z^2)} W'(\delta z^2) z^2 \quad \text{(68)}
\]

\[
= -\frac{1}{1 + W(\delta z^2)} \left( \frac{1}{2} W_\delta(z)^2 + W'(\delta z^2) z^2 \right) \quad \text{(69)}
\]

Using \( W'(\delta z^2) = \frac{W(\delta z^2)}{\delta z^2 W'(\delta z^2)} \) and re-factorising gives (67).

A.3 Properties of the Gaussian log-likelihood evaluated at \( W_\delta(z) \)

The next lemma shows that increasing \( \delta \) always increases the input log-likelihood \( \ell(\delta \mid u_\delta = W_\delta(z)) \) - see also Fig. 6b. For \( \delta \to \infty \) the Gaussianised \( u_\delta \) goes to \( 0 \), which clearly is the maximizer of the Gaussian likelihood if \( \mu = 0 \).

**Lemma A.6** (Derivative of the Gaussian log-likelihood at \( W_\delta(z) \)). For all \( z \in \mathbb{R} \) and for \( \delta \geq 0 \)

\[
\frac{\partial}{\partial \delta} \ell(\mu_x = 0, \sigma_x = 1 \mid W_\delta(z)) = \frac{1}{2} \frac{1}{1 + W(\delta z^2)} [W_\delta(z)]^4 \geq 0. \quad \text{(71)}
\]

**Proof.** The log of the standard Gaussian pdf evaluated at the Gaussianised data \( W_\delta(z) \) simplifies to

\[
\log \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} [W_\delta(z)]^2} = \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2} [W_\delta(z)]^2. \quad \text{(72)}
\]

The rest follows by Lemma A.3.
B Proofs of the results in the main text

B.1 Inverse transformation

Proof of Lemma 2.5. Without loss of generality assume that $\mu_x = 0$ and $\sigma_x = 1$ (otherwise standardize $X$ first). Squaring (1) and multiplying by $\delta$ yields

$$\delta Z^2 = \delta U^2 \exp(\delta U^2)$$

(73)

The inverse of (73) is by definition Lambert’s $W(z)$ function (Rosenlicht, 1969)

$$W(z) \exp W(z) = z, \quad z \in \mathbb{C}.$$  

(74)

$W(z)$ is bijective for $z \geq 0$. Since $\delta U^2 \geq 0$ for all $\delta \geq 0$, applying $W(\cdot)$ to (73), dividing by $\delta$, and taking the square root gives

$$U = \pm \sqrt{\frac{W(\delta Z^2)}{\delta}}$$

(75)

Since $\exp(\frac{1}{2}U^2) > 0$ for all $\delta \in \mathbb{R}$ and all $U$, it follows that $Z = U \exp(\delta/2U^2)$ and $U$ must have the same sign, which concludes the proof.

B.2 Cdf and pdf

Proof of Theorem 2.8. By definition,

$$G_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}\left( \left\{ U \exp \left( \frac{\delta}{2} U^2 \right) \right\} \sigma_x + \mu_x \leq y \right)$$

$$= \mathbb{P}\left( U \exp \left( \frac{\delta}{2} U^2 \right) \leq z \right) = \mathbb{P}(U \leq W_\delta(z))$$

$$= F_U(U \leq W_\delta(z)).$$

(76) (77) (78)

Taking the derivative with respect to $y$ gives

$$\frac{d}{dy}G_Y(y \mid \beta, \delta) = f_X(W_\delta(z)\sigma_x + \mu_x \mid \beta) \cdot \sigma_x \frac{d}{dy} W_\delta \left( \frac{y - \mu_x}{\sigma_x} \right)$$

$$= f_U(W_\delta(z) \mid \beta) \cdot \sigma_x \frac{1}{\sigma_x} \frac{d}{dz} W_\delta \left( \frac{y - \mu_x}{\sigma_x} \right)$$

$$= f_U(W_\delta(z) \mid \beta) \cdot \frac{d}{dz} W_\delta(z).$$

(79) (80) (81)

Using Lemma A.2 yields (14).
B.3 MLE for $\delta$

**Lemma B.1** (Derivative of the Lambert $W \times$ Gaussian log-likelihood). We have

$$D(\delta | z) := \frac{\partial}{\partial \delta} \ell(\delta | z) = \sum_{i=1}^{N} z_i^2 W'(\delta z_i^2) \left( \frac{1}{2} W_{\delta}(z_i^2) - \frac{1}{2} \frac{1}{1 + \frac{1}{1 + W(\delta z_i^2)}} \right)$$  \hspace{1cm} (82)

$$= \frac{1}{2} \sum_{i=1}^{N} \left( \frac{W_{\delta}(z_i^2)}{1 + \delta W_{\delta}(z_i^2)} - \frac{W_{\delta}(z_i^2)}{1 + \delta W_{\delta}(z_i^2)} \right) \left( \frac{1}{2} \frac{1}{1 + \frac{1}{1 + W(\delta z_i^2)}} \right)$$  \hspace{1cm} (83)

$$= \frac{1}{2} \sum_{i=1}^{N} \frac{W_{\delta}(z_i^2)}{1 + W(\delta z_i^2)} - \sum_{i=1}^{N} \frac{W_{\delta}(z_i^2)}{1 + W(\delta z_i^2)} \left( \frac{1}{2} \frac{1}{1 + \frac{1}{1 + W(\delta z_i^2)}} \right).$$  \hspace{1cm} (84)

**Proof.** Follows by applying Lemmas A.5 and A.6 to $\frac{\partial}{\partial \delta} \ell(\delta | z) = \frac{\partial}{\partial \delta} \log R(\delta | z) + \frac{\partial}{\partial \delta} \ell(\mu_x = 0, \sigma_x = 1 | W_{\delta}(z)).$

**Proof of Theorem 4.1.** a) The log-likelihood is increasing at $\delta = 0$ if and only if (set $\delta = 0$ in (84) and use Properties A.1)

$$\sum_{i=1}^{N} z_i^4 > 3 \sum_{i=1}^{N} z_i^2,$$  \hspace{1cm} (85)

i.e. actually transforming the data (choosing $\delta > 0$) increases the overall likelihood only if the data is heavy-tailed enough. It is important to point out that the sum of squares is not squared again. Hence the condition is not equivalent for the data having empirical kurtosis larger than 3.

b) If (85) does not hold, then $\hat{\delta}_{MLE}$ must satisfy $D(\delta | z) \big|_{\delta=\hat{\delta}_{MLE}} = 0$ from (82) in Lemma B.1. It remains to be shown that this equation has (at least) one positive solution.

i) Since $\lim_{\delta \to \infty} W_{\delta}(z) = 0$ for all $z \in \mathbb{R}$, (84) is also true in the limit; however, we can ignore this solution as we require $\hat{\delta}_{MLE} \in \mathbb{R}$. 

ii) By continuity and $\lim_{\delta \to \infty} W_{\delta}(z) = 0$ it also follows that for sufficiently large $\delta_M$, $W_{\delta_M}(z_i) < 1$ for all $z_i \in \mathbb{R}$. Hence $W_{\delta_M}(z_i)^4 < W_{\delta_M}(z_i)^2$ and therefore

$$\frac{1}{2} \sum_{i=1}^{N} \frac{W_{\delta}(z_i^2)}{1 + \delta W_{\delta}(z_i^2)} < \frac{1}{2} \sum_{i=1}^{N} \frac{W_{\delta}(z_i^2)}{1 + \delta W_{\delta}(z_i^2)}$$

$$< \sum_{i=1}^{N} \frac{W_{\delta}(z_i^2)}{1 + \delta W_{\delta}(z_i^2)} \left( \frac{1}{2} \frac{1}{1 + \frac{1}{1 + W(\delta z_i^2)}} \right)$$  \hspace{1cm} (86)

$$\text{for } \delta \geq \delta_M,$$  \hspace{1cm} (87)

which shows that $D(\delta | z) \big|_{\delta \geq \delta_M} < 0$. That is $D(\delta | z)$ is approaching 0 from below for $\delta \to \infty$.

iii) By continuity and since $D(\delta | z) \big|_{\delta=0} > 0$ (if (85) does not hold), it must cross the $D(\delta | z) = 0$-line at least once in the interval $(0, \delta_M)$, proving the existence of $\hat{\delta}_{MLE}$. 

39
c) The log-likelihood can be decomposed in

\[
\ell(\delta | \mathbf{z}) \propto -\frac{1}{2} \sum_{i=1}^{N} [W_\delta(z_i)]^2 + \sum_{i=1}^{N} \log \frac{W_\delta(z_i)}{z_i} - \log \left[ 1 + W(\delta z_i^2) \right].
\]

(88)

Lemmas A.5 and A.6 show that \( R(\delta | \mathbf{z}) \) is monotonically decreasing and \( \ell(\mu_x = 0, \sigma_x = 1 | W_\delta(\mathbf{z})) \) is monotonically increasing in \( \delta \).

Furthermore, \( \lim_{\delta \to \infty} \ell(\mu_x = 0, \sigma_x = 1 | W_\delta(\mathbf{z})) = 0 \), that is the input likelihood is monotonically increasing but bounded from above (by \( 0 = \log 1 \)). On the other hand \( \lim_{\delta \to \infty} R(\delta | \mathbf{z}) = -\infty \) showing that the penalty term is decreasing in an unbounded manner. Thus their sum has a global maximum either at the unique mode of \( \ell(\delta | \mathbf{z}) \) or at the boundary \( \delta = 0 \) - see also Fig. 6b.