12

# The Lambert Way to Gaussianize heavy tailed data with the inverse of Tukey's h transformation as a special case

Georg M. Goerg

Department of Statistics, Carnegie Mellon University Pittsburgh, PA 15213, USA gmg@stat.cmu.edu, www.stat.cmu.edu/~gmg

January 1, 2013

I present a parametric, bijective transformation to generate heavy tail versions Y of arbitrary RVs  $X \sim F_X$ . The tail behavior of the so-called heavy tail Lambert  $W \times F_X$  RV Y 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 Y distribution. Lambert's Y function provides an explicit inverse transformation, which can be estimated by maximum likelihood. This inverse can remove heavy tails from data, and also provide analytical expressions for the cumulative distribution (cdf) and probability density function (pdf). As a special case, these yield explicit formulas for Tukey's Y pdf and cdf - to the author's knowledge for the first time in the literature. Simulations and applications to S&P 500 log-returns and solar flares data demonstrate the usefulness of the introduced methodology.

The R package LambertW implementing the presented methodology is publicly available at CRAN.

# 14 Contents

15	1	Introduction	1
16		1.1 Multivariate Extensions	4
17		1.2 Box-Cox Transformation to Remove Heavy Tails	5
18	2	Generating Heavy Tails Using Transformations	5
19		2.1 Tukey's $h$ Distribution	6
20		2.2 Heavy Tail Lambert W Random Variables	7
21		2.3 Inverse Transformation: "Gaussianize" Heavy-Tailed Data	8
22		2.4 Distribution and Density	10
23		2.5 Quantile Function	12
24	3	Tukey's h distribution: Gaussian input	13
25		3.1 Tukey's h versus student's t	14
26	4	Parameter Estimation	15
27		4.1 Maximum Likelihood Estimation (MLE)	$^{-1}_{15}$
28		4.1.1 Properties of The MLE For The Heavy Tail Parameter	17
29		4.2 Iterative Generalized Method of Moments (IGMM)	18
30	5	Simulations	19
31		5.1 Estimating $\delta$ Only	19
32		5.2 Estimating All Parameters Jointly	20
33		5.3 Discussion of the Simulations	23
34	6	Applications	23
35		6.1 Estimating Location of a Cauchy With The Sample Mean	24
36		6.2 Heavy Tails in Finance: S&P 500 Case Study	25
37		6.2.1 Gaussian Fit to Returns	26
38		6.2.2 Heavy Tail Fit to Returns	27
39		6.2.3 "Gaussianizing" Returns	28
40		6.2.4 Gaussian MLE for Gaussianized Data	28
41		6.3 Removing Power Law From Solar Flare Counts	29
42	7	Discussion and Outlook	31
43	$\mathbf{R}_{0}$	eferences	33
44	Δ	Auxiliary Results and Properties	39
45	4 1	A.1 Inverse Transformation $W_{\delta}(z)$	39
46		A.2 Penalty $\log R(\delta \mid z_i)$ for Standard Gaussian Input	41
47		A.3 Gaussian log-Likelihood at $W_{\delta}(z)$	42

48	$\mathbf{B}$	$\mathbf{Pro}$	ofs	
49		B.1	Inverse transformation	42
50		B.2	Cdf and pdf	43
51		B.3	MLE for $\delta$	44
52	$\mathbf{C}$	Det	ails on IGMM	46
53	D	Sim	ulation Details	49

#### <sub>54</sub> 1 Introduction

```
Statistical theory and practice are both tightly linked to Gaussianity. In theory, many meth-
   ods require Gaussian data or noise: i) regression often assumes Gaussian errors; ii) pattern
   recognition for images often model noise as a Gaussian random field (Achim, Tsakalides,
   and Bezerianos, 2003); iii) many time series models are based on Gaussian white noise
58
   (Brockwell and Davis, 1998; Engle, 1982; Granger and Joyeux, 2001).
59
      In all these cases, a model \mathcal{M}_{\mathcal{N}}, parameter estimates and their standard errors, and other
   properties, are then studied – all based on the ideal(istic) assumption of Gaussianity.
61
62
      In practice, however, data/noise often exhibits asymmetry and heavy tails; for example
63
   wind speed data (Field, 2004), human dynamics
65
      (Vázquez, Oliveira, Dezsö, Goh, Kondor, and Barabási, 2006), or Internet traffic data
66
   (Gidlund and Debernardi, 2009) – just to a name few. Particularly notable examples are
   financial data (Cont, 2001; Kim and White, 2003) and speech signals (Aysal and Barner,
68
   2006), which almost exclusively exhibit heavy tails. Thus a model \mathcal{M}_{\mathcal{N}} developed for the
69
   Gaussian case does not necessarily provide accurate inference anymore.
      One way to overcome this shortcoming is to replace \mathcal{M}_{\mathcal{N}} with a new model \mathcal{M}_{G}, where
71
   G is a heavy tail distribution: i) regression with Cauchy errors (Smith, 1973); ii) image
   denoising for \alpha-stable noise (Achim et al., 2003); iii) forecasting long memory processes
   with heavy tail innovations (Ilow, 2000; Palma and Zevallos, 2011), or ARMA modeling of
74
   electricity loads with hyperbolic noise (Nowicka-Zagrajek and Weron, 2002).
75
      While such fundamental approaches are attractive from a theoretical perspective, they
76
   can become unsatisfactory from a practical viewpoint. Many successful statistical models
   assume Gaussianity, their theory is very well understood, and many algorithms are imple-
78
   mented for the simple – and often much faster – Gaussian case. Thus developing models
79
   based on an entirely unrelated distribution G is like throwing out the (Gaussian) baby with
   the bathwater.
81
82
```

It would be very useful to transform a Gaussian RV X to a heavy-tailed RV Y and vice

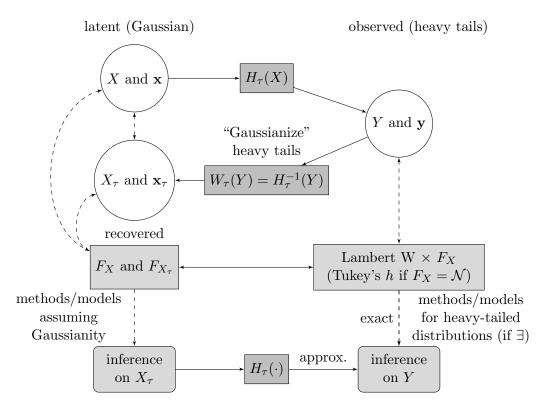


Figure 1: Schematic view of the heavy tail Lambert W ×  $F_X$  framework. (left) Latent input  $X \sim F_X$ :  $H_{\tau}(X)$  from (6) transforms (solid arrows) X to  $Y \sim$  Lambert W ×  $F_X$  and generates heavy tails. (right) Observed heavy tail world Y and  $\mathbf{y}$ : (1) use  $W_{\tau}(\cdot)$  to back-transform  $\mathbf{y}$  to latent "Normal"  $\mathbf{x}_{\tau}$ , (2) use model  $\mathcal{M}_{\mathcal{N}}$  of your choice (regression, time series models, hypothesis testing, etc.) for inference on  $\mathbf{x}_{\tau}$ , and (3) convert results back to the original "heavy-tailed world" of  $\mathbf{y}$ .

versa, and thus rely on knowledge - and software - for the well-understood Gaussian case,

84

while still capturing heavy tails in the data. Optimally such a transformation should: a) be bijective; b) include Normality as a special case for hypothesis testing; and c) be parametric so the optimal transformation can be estimated efficiently.

Figure 1 illustrates this pragmatic approach: researchers can make their observations  $\mathbf{y}$  as Gaussian as possible  $(\mathbf{x}_{\tau})$  before making inference based on their favorite Gaussian model  $\mathcal{M}_{\mathcal{N}}$ . This avoids the development of - or the data analysts waiting for - a whole new theory of  $\mathcal{M}_{G}$  and new implementations based on a particular heavy-tailed distribution G, while still improving statistical inference on heavy-tailed data  $\mathbf{y}$ . For example, consider  $\mathbf{y} = (y_1, \dots, y_{500})$  from a standard Cauchy distribution  $\mathcal{C}(0,1)$  in Fig. 2a: modeling heavy tails by a transformation makes it even possible to Gaussianize this Cauchy sample (Fig.

2c). This "nice" data  $\mathbf{x}_{\tau}$  can then be subsequently analyzed with common techniques. For example, the location can now be estimated using the sample average (Fig. 2d). For details see Section 6.1.

98

109

110

111

112

113

115

116

117

118

119

Liu, Lafferty, and Wasserman (2009) use a semi-parametric approach, where Y has a 99 nonparanormal distribution if  $f(Y) \sim \mathcal{N}(\mu, \sigma^2)$  where  $f(\cdot)$  is an increasing smooth func-100 tion; they estimate  $f(\cdot)$  using non-parametric methods. This leads to a greater flexibility 101 in the distribution of Y, but it suffers from two drawbacks: i) non-parametric methods 102 have slower convergence rates and thus need large samples, and ii) for identifiability of 103  $f(\cdot), \mathbb{E}f(Y) \equiv \mathbb{E}Y$  and  $\mathbb{V}f(Y) \equiv \mathbb{V}Y$  must hold. While i) is inherent to non-parametric methods, point ii) requires Y to have finite mean and variance, which is especially limiting 105 for heavy-tailed data where this condition is often not met. Thus here we use parametric 106 transformations which do not rely on restrictive identifiability conditions and also work well 107 for small sample sizes. 108

The main contributions of this work are three-fold: a) following Goerg (2011) I introduce a meta-family of heavy tail Lambert W  $\times$   $F_X$  distributions with Tukey's h (Hoaglin, 2006) as a special case; b) I present a bijective transformation to "Gaussianize" heavy-tailed data (Section 2); and c) I also provide simple expressions for the cumulative distribution function (cdf)  $G_Y(y)$  and probability density function (pdf)  $g_Y(y)$  - also for Tukey's h -, which can be easily implemented in statistics software (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 and studies 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 × 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 for exploratory data analysis: Gaussianizing  $\gamma$ -ray intensity data reveals a bimodal density, which even non-parametric estimators fail to detect if heavy tails

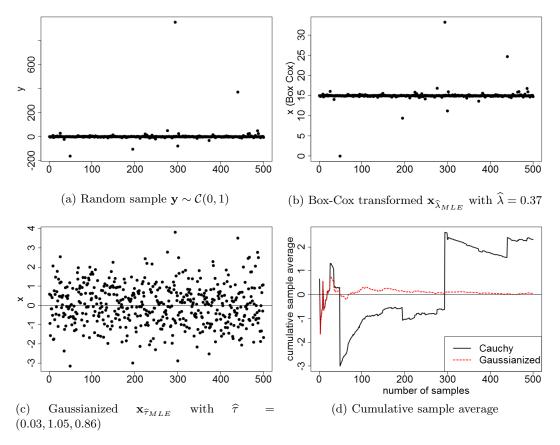


Figure 2: Gaussianizing a standard Cauchy sample. For (d)  $\tau^{(n)}$  was estimated for each fixed  $n = 5, \ldots, 500$ , before Gaussianizing  $(y_1, \ldots, y_n)$ .

are not removed. Finally, we discuss the new methodology and future work in Section 7.

All proofs are given in the Supplementary Material, Appendix B.

Computations, figures, and simulations were done in R (R Development Core Team,

2010). The R package LambertW is publicly available on CRAN.

#### 1.1 Multivariate Extensions

129

While this work focuses on univariate case, multivariate extensions of the presented methods can be defined component-wise – analogously to the multivariate version of Tukey's h distribution (Field and Genton, 2006). While this may not make the transformed RVs
jointly Gaussian, it still provides a good starting point for more well-behaved multivariate
modeling.

#### <sub>15</sub> 1.2 Box-Cox Transformation to Remove Heavy Tails

A popular method to deal with skewed, high variance data is the Box-Cox transformation

$$\mathbf{y}_{\lambda} = \begin{cases} \frac{\mathbf{y}^{\lambda} - 1}{\lambda} & \text{if } \lambda > 0\\ \log \mathbf{y} & \text{if } \lambda = 0. \end{cases}$$
 (1)

The parameter  $\lambda$  can be chosen by MLE. However one major limitation of (1) is the non-136 negativity constraint on y, which prohibits its use in many applications. To avoid this 137 limitation it is common to shift the data,  $\tilde{\mathbf{y}} = \mathbf{y} + |\min(\mathbf{y})| \ge 0$ . However, as Fig. 2b shows 138 applying the Box-Cox transformation to the Cauchy sample completely fails. Furthermore, 139 this restricts Y to a half-open interval  $[c,\infty)$  and is not desirable if the underlying process 140 can occur on the entire real line, since it undermines statistical inference for yet unobserved 141 data. See Sakia (1992) for a more detailed discussion and the Box-Cox transformation in 142 general. 143

144

145

146

147

149

150

151

Furthermore, the main purpose of the Box-Cox transformation is to stabilize variance (Blaylock, Salathe, and Green, 1980; Lawrance, 1987; Tsiotas, 2007) and remove right tail skewness (Goncalves and Meddahi, 2011); a lower empirical kurtosis is merely a by-result of the variance stabilization. In contrast, the Lambert W framework is designed to model heavy-tailed RVs and remove heavy tails from data, and has no difficulties with negative values.

# 2 Generating Heavy Tails Using Transformations

Random variables exhibit heavy tails if more mass than for a Gaussian RV lies at the outer end of the density support. A RV Z has a tail index a if 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.

<sup>&</sup>lt;sup>1</sup>We use  $\tilde{\mathbf{y}} = \mathbf{y} + |\min(\mathbf{y})| + 1$  and use boxcox from the MASS R package;  $\hat{\lambda} = 0.37$ .

<sup>&</sup>lt;sup>2</sup>There are various similar definitions of heavy/fat/long tails; for this work these differences are not essential.

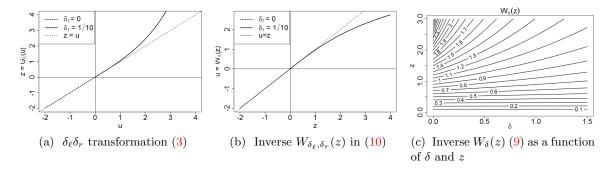


Figure 3: Transformation and inverse transformation for  $\delta_{\ell} = 0$  and  $\delta_{r} = 1/10$ : identity on the left (same tail behavior) and a heavy-tailed transformation in the right tail of input U.

#### 157 2.1 Tukey's h Distribution

A parametric transformation is the basis of Tukey's h RVs (Hoaglin, 2006)

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

where U is standard Normal RV and h is the heavy tail parameter. The RV Z has tail parameter a=1/h (Morgenthaler and Tukey, 2000) and reduces to the Gaussian for h=0. Morgenthaler and Tukey (2000) extend the h distribution to the skewed, heavy-tailed family of hh RVs

$$Z = \begin{cases} U \exp\left(\frac{\delta_{\ell}}{2}U^{2}\right), & \text{if } U \leq 0, \\ U \exp\left(\frac{\delta_{r}}{2}U^{2}\right), & \text{if } U > 0, \end{cases}$$
 (3)

where again  $U \sim \mathcal{N}(0,1)$ . Here  $\delta_{\ell} \geq 0$  and  $\delta_{r} \geq 0$  shape the left and right tail of Z, respectively; thus transformation (3) can model skewed and heavy-tailed data - see Fig. 3a. For simplicity let  $H_{\delta}(u) := u \exp\left(\frac{\delta}{2}u^{2}\right)$ .

Despite their great flexibility they are not popular in statistical practice, because the inverse of (2) or (3) has not been found. Consequently, no closed-form expressions for the cdf or pdf are available. Although Morgenthaler and Tukey (2000) express the pdf of (2) as  $(h \equiv \delta)$ 

$$g_Z(z) = \frac{f_U\left(H_\delta^{-1}(z)\right)}{H_\delta'\left(H_\delta^{-1}(z)\right)},\tag{4}$$

they fall short of explicitly specifying  $H_{\delta}^{-1}(z)$ . So far this inverse has been considered an-

alytically intractable (Field, 2004), or was only numerically approximated (Fischer, 2010;
Headrick, Kowalchuk, and Sheng, 2008). Thus parameter inference relied on matching empirical and theoretical quantiles (Field, 2004; Morgenthaler and Tukey, 2000), or by the
method of moments (Headrick et al., 2008). Only recently Headrick et al. (2008) provided
numerical approximations. Hence, a closed form, analytically tractable pdf that can be
computed efficiently is essential for a wide-spread use of Tukey's h (& variants).

168

180

In this work I present this long sought explicit inverse, which is readily available in standard statistics software. For ease of notation and concision main results are shown for  $\delta_{\ell} = \delta_r = \delta$ ; analogous results for  $\delta_{\ell} \neq \delta_r$  will be stated without details.

#### 172 2.2 Heavy Tail Lambert W Random Variables

Tukey's h transformation (2) is strongly related to the approach taken by Goerg (2011) to introduce skewness in continuous RVs  $X \sim F_X(x)$ . In particular, if  $Z \sim$  Tukey's h, then  $Z^2 \sim$  skewed Lambert W  $\times \chi_1^2$  with skew parameter  $\gamma = h$ .

Adapting the skew Lambert W ×  $F_X$  input/output idea<sup>3</sup> (see Fig. 1), Tukey's h RVs can be generalized to heavy-tailed Lambert W ×  $F_X$  RVs.

**Definition 2.1.** Let U be a continuous RV with  $cdf F_U(u \mid \beta)$ ,  $pdf f_U(u \mid \beta)$ , and parameter vector  $\beta$ . Then,

$$Z = U \exp\left(\frac{\delta}{2}U^2\right), \quad \delta \in \mathbb{R},$$
 (5)

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

Tukey's h distribution results for U being a standard Gaussian  $\mathcal{N}(0,1)$ .

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

$$Y = \left\{ U \exp\left(\frac{\delta}{2}U^2\right) \right\} \sigma_x + \mu_x, \quad \delta \in \mathbb{R}, \tag{6}$$

 $<sup>^3</sup>$ Most concepts and methods from the skew Lambert W  $\times F_X$  case transfer one-to-one to the heavy tail Lambert W RVs presented here. Thus for the sake of concision I refer to Goerg (2011) for details and background information on the Lambert W framework.

with parameter vector  $\theta = (\boldsymbol{\beta}, \delta)$ , where  $U = (X - \mu_x)/\sigma_x$ .

The input is not necessarily Gaussian but can be any other location-scale continuous RV, e.g., from a uniform distribution:  $X \sim U(a,b)$ .

**Definition 2.3.** Let  $X \sim F_X(x/s \mid \beta)$  be a continuous scale-family RV, with standard deviation  $\sigma_x$ ; let  $U = X/\sigma_x$ . Then,

$$Y = X \exp\left(\frac{\delta}{2}U^2\right), \quad \delta \in \mathbb{R},$$
 (7)

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

Let  $\tau := (\mu_x(\boldsymbol{\beta}), \sigma_x(\boldsymbol{\beta}), \delta)$  define transformation (6). If  $X \in (-\infty, \infty)$ , 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  $Y \in [0, \infty)$ .

188

200

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

The Lambert W formulation of heavy tail modeling is more general than Tukey's h distribution as 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, it yields a non-bijective transformation and thus to parameter-dependent support and non-unique input. Thus for the remainder of this work I tacitly assume  $\delta \geq 0$ , unless stated otherwise.

#### 2.3 Inverse Transformation: "Gaussianize" Heavy-Tailed Data

Transformation (6) is bijective and its inverse can be obtained via Lambert's W function, which is the inverse of  $z = u \exp(u)$ , i.e., that function which satisfies  $W(z) \exp(W(z)) = z$ .

<sup>&</sup>lt;sup>4</sup>For non-central, non-scale input set  $\tau = (0, 1, \delta)$ ; for scale-family input  $\tau = (0, \sigma_x, \delta)$ .

Lambert's W has been studied extensively in mathematics, physics, and other areas of science (Corless, Gonnet, Hare, and Jeffrey, 1996; Rosenlicht, 1969; Valluri, Jeffrey, and Corless, 2000), and is implemented in 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 many useful properties (see Appendix A and Corless et al. (1996)), in particular W(z) is bijective for  $z \geq 0$ .

#### **Lemma 2.5.** The inverse transformation of (6) is

$$W_{\tau}(Y) := W_{\delta} \left( \frac{Y - \mu_x}{\sigma_x} \right) \sigma_x + \mu_x = U \sigma_x + \mu_x = X, \tag{8}$$

where

$$W_{\delta}(z) := \operatorname{sgn}(z) \left( \frac{W(\delta z^2)}{\delta} \right)^{1/2}, \tag{9}$$

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

Lemma 2.5 gives for the first time an analytic, bijective inverse of Tukey's h transforma-211 tion:  $H_{\delta}^{-1}(y)$  of Morgenthaler and Tukey (2000) is now analytically available as (8). Bijec-212 tivity implies that for any data  $\mathbf{y}$  and parameter  $\tau$ , the exact input  $\mathbf{x}_{\tau} = W_{\tau}(\mathbf{y}) \sim F_X(x)$ 213 can be obtained. In view of the importance and popularity of Gaussianity, we clearly want to back-215 transform heavy-tailed data to a Gaussian rather than yet another heavy-tailed distribution. 216 Typically tail behavior of RVs are compared by their kurtosis  $\gamma_2(X) = \mathbb{E}(X - \mu_x)^4 / \sigma_x^4$ , which 217 for a Gaussian RV equals 3. Hence for the future when we "normalize y" we can not only 218 subtract the mean, and divide by the standard deviation, but also transform it to  $\mathbf{x}_{\tau}$  with 219  $\hat{\gamma}_2(\mathbf{x}_{\tau}) = 3 - a$  "Normalization" in the true sense of the word (see Fig. 2c). 220 This data-driven view of the Lambert W framework can also be useful for kernel density 221 estimation (KDE), where multivariate data is often pre-scaled to unit-variance, so the same 222 bandwidth can be used in each dimension (Hwang, Lay, and Lippman, 1994; Wasserman, 223 2007). Thus "normalizing" the Lambert Way might likely also improve KDE for heavytailed data (see also Maiboroda and Markovich, 2004; Markovich, 2005). 225

Corollary 2.6 (Inverse transformation for Tukey's hh). The inverse transformation of (3) is

$$W_{\delta_{\ell},\delta_{r}}(z) = \begin{cases} W_{\delta_{\ell}}(z), & \text{if } z \leq 0, \\ W_{\delta_{r}}(z), & \text{if } z > 0. \end{cases}$$
 (10)

Figure 3b shows  $W_{\delta_{\ell},\delta_r}(z)$  for  $\delta_l=0$  and  $\delta_r=1/10$ . The transformation in Fig. 3a generates a right heavy tail version of U (x-axis) by stretching only the positive axis (y-axis). By definition  $W_{\delta_{\ell},\delta_r}(z)$  removes the heavier right tail in Z (positive y-axis). Figure 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 (much) smaller u.

Remark 2.7 (Generalized transformation). Transformation (2) can be generalized to

$$Z = U \exp\left(\frac{\delta}{2\alpha} \left(U^2\right)^{\alpha}\right), \quad \alpha > 0.$$
 (11)

The inner term  $U^2$  guarantees bijectivity for all  $\alpha > 0$ . The inverse is

$$W_{\delta,\alpha}(z) := \operatorname{sgn}(z) \left( W \left( \frac{\delta z^{2\alpha}}{\delta} \right) \right)^{\frac{1}{2\alpha}}. \tag{12}$$

For comparison with Tukey's h I consider  $\alpha = 1$  only. For  $\alpha = 1/2$  transformation (11) is closely related to skewed Lambert  $W \times F_X$  distributions.

#### 233 2.4 Distribution and Density

For ease of notation let

$$z = \frac{y - \mu_x}{\sigma_x}$$
,  $u = W_{\delta}(z)$ , and  $x = W_{\tau}(y) = u\sigma_x + \mu_x$ . (13)

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

$$G_Y(y \mid \beta, \delta) = F_X(W_\delta(z)\sigma_x + \mu_x \mid \beta)$$
(14)

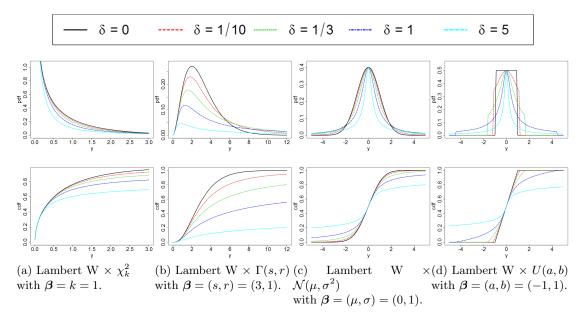


Figure 4: Pdf (top) and cdf (bottom) of a heavy-tail (a) "non-central, non-scaled", (b) "scale", and (c and d) "location-scale" Lambert W  $\times$   $F_X$  RV Y for various degrees of heavy tails (color, dashed lines).

and

242

$$g_{Y}(y \mid \boldsymbol{\beta}, \delta) = f_{X}\left(W_{\delta}\left(\frac{y - \mu_{x}}{\sigma_{x}}\right)\sigma_{x} + \mu_{x} \mid \boldsymbol{\beta}\right) \cdot \frac{W_{\delta}\left(\frac{y - \mu_{x}}{\sigma_{x}}\right)}{\frac{y - \mu_{x}}{\sigma_{x}}\left[1 + W\left(\delta\left(\frac{y - \mu_{x}}{\sigma_{x}}\right)^{2}\right)\right]}.$$
 (15)

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

For scale family or non-central, non-scale input set  $\mu_x = 0$  or  $\mu_x = 0$ ,  $\sigma_x = 1$ .

The explicit formula (15) allows a fast computation and theoretical analysis of the likelihood, which is essential for – either frequentist or Bayesian – statistical inference. Detailed properties of (15) are given in Section 4.1.

Figure 4 shows (14) and (15) 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. The cdf and pdf of Z in (3) equal

$$G_{Z}(z \mid \boldsymbol{\beta}, \delta_{\ell}, \delta_{r}) = \begin{cases} G_{Z}(z \mid \boldsymbol{\beta}, \delta_{\ell}), & \text{if } z \leq 0, \\ G_{Z}(z \mid \boldsymbol{\beta}, \delta_{r}), & \text{if } z > 0, \end{cases}$$

$$(16)$$

and

$$g_{Z}(z \mid \boldsymbol{\beta}, \delta) = \begin{cases} g_{Z}(z \mid \boldsymbol{\beta}, \delta_{\ell}), & \text{if } z \leq 0, \\ g_{Z}(z \mid \boldsymbol{\beta}, \delta_{r}), & \text{if } z > 0. \end{cases}$$

$$(17)$$

#### 246 2.5 Quantile Function

Quantile fitting has been the standard technique to estimate  $\mu_x$ ,  $\sigma_x$ , and  $\delta$  of Tukey's h. In particular, the median of Y and X are equal. Thus for symmetric, location-scale family input the sample median of  $\mathbf{y}$  is a robust estimate for  $\mu_x$  for any  $\delta \geq 0$  (see also Section 5). General quantiles can be computed via (Hoaglin, 2006)

$$y_{\alpha} = u_{\alpha} \exp\left(\frac{\delta}{2}u_{\alpha}^{2}\right)\sigma_{x} + \mu_{x}, \tag{18}$$

where  $u_{\alpha} = W_{\delta}(z_{\alpha})$  are the  $\alpha$ -quantiles of  $F_U(u)$ . 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 conversion can be especially useful for education: teaching heavy-tailed statistics in introductory courses soon becomes too difficult using e.g., Cauchy or  $\alpha$ -stable distributions. Yet, transforming data via Lambert's W, using previously learned methods for the Gaussian case, and then transforming the inference back to the "heavy-tailed world" - e.g., transforming quantiles using (18) - is straightforward. Thus the Lambert W  $\times$   $F_X$ framework can promote heavy-tailed statistics in introductory courses.

## $_{56}$ 3 Tukey's h distribution: Gaussian input

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

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

$$\not\exists, \qquad \text{if } n > \frac{1}{\delta},$$

$$(19)$$

which in particular implies (Headrick et al., 2008)

$$\mathbb{E}Z = \mathbb{E}Z^3 = 0$$
, if  $\delta < 1$  and  $1/3$ , respectively (20)

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

Thus the kurtosis of Y equals (see Fig. 5)

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

For  $\delta = 0$ , (21) and (22) reduce to the familiar Gaussian values.

Corollary 3.1. The cdf of Tukey's h equals

$$G_Y(y \mid \mu_x, \sigma_x, \delta) = \Phi\left(\frac{W_\tau(y) - \mu_x}{\sigma_x}\right), \tag{23}$$

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

$$g_Y(y \mid \mu_x, \sigma_x, \delta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1+\delta}{2} W_\delta \left(\frac{y-\mu_x}{\sigma_x}\right)^2\right) \cdot \frac{1}{1+W\left(\delta \left(\frac{y-\mu_x}{\sigma_x}\right)^2\right)}$$
(24)

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

Section 4.1 studies functional properties of (24) in more detail.

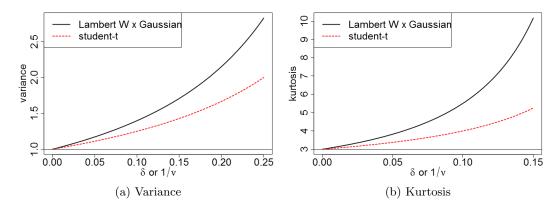


Figure 5: Comparing moments of Lambert W  $\times$  Gaussian and student's t.

#### 3.1 Tukey's h versus student's t

Student's  $t_{\nu}$  distribution with  $\nu$  degrees of freedom 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,

$$\mathbb{E}T = \mathbb{E}T^3 = 0 \text{ if } \nu < 1 \text{ or } < 3, \quad \mathbb{E}T^2 = \frac{\nu}{\nu - 2} = \frac{1}{1 - \frac{2}{\nu}} \text{ if } \frac{1}{\nu} < \frac{1}{2},$$
 (25)

and kurtosis

$$\gamma_2(\nu) = 3\frac{\nu - 2}{\nu - 4} = 3\frac{1 - 2\frac{1}{\nu}}{1 - 4\frac{1}{\nu}} \text{ if } \frac{1}{\nu} < \frac{1}{4}.$$
 (26)

Comparing (26) and (21) with (22) and (25) shows a natural association between  $1/\nu$  and  $\delta$  and a close similarity between the first four moments of student's t and Tukey's h (Fig. 5). By continuity and monotonicity, the first four moments of a location-scale t distribution can always be exactly matched by a corresponding location-scale Lambert W × Gaussian. Thus if student's t is used to model heavy tails, and not as the true distribution of a test statistic, it might be worthwhile to also fit heavy tail Lambert W × Gaussian distributions for an equally valuable "second opinion". For example, a parallel analysis on 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 Gaussianize and thus reveal hidden patterns in the data; patterns that 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  $\mathbf{y} = (y_1, \dots, y_N)$ 273 from transformation (6),  $\theta = (\beta, \delta)$  has to be estimated from the data. Due to the lack of a 274 closed form pdf of Y, this has been typically done by matching quantiles or a method of mo-275 ments estimator (Field, 2004; Headrick et al., 2008; Morgenthaler and Tukey, 2000). These 276 inefficient methods can now be replaced by the – fast and usually efficient – maximum like-277 lihood estimator (MLE) using the pdf in (15). Rayner and MacGillivray (2002) introduce a numerical MLE procedure based on quantile functions, but they conclude that "sample 279 sizes significantly larger than 100 should be used to obtain reliable estimates through max-280 imum likelihood". Simulations in Section 5 show that log-likelihood maximization with the 281 Lambert W methodology converges quickly and is accurate even for sample sizes as small 282 as N = 10. 283

#### 84 4.1 Maximum Likelihood Estimation (MLE)

For an i.i.d. sample  $\mathbf{y} \sim g_Y\left(y \mid \boldsymbol{\beta}, \delta\right)$  the log-likelihood function equals

$$\ell(\theta \mid \mathbf{y}) = \sum_{i=1}^{N} \log g_Y(y_i \mid \boldsymbol{\beta}, \delta). \tag{27}$$

The MLE is that  $\theta = (\beta, \delta)$  which maximizes (27), i.e.

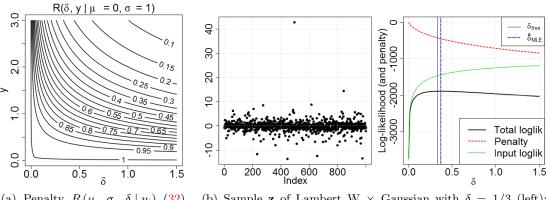
$$\widehat{\theta}_{MLE} = \left(\widehat{\boldsymbol{\beta}}, \widehat{\delta}\right)_{MLE} = \arg\max_{\boldsymbol{\beta}, \delta} \ell\left(\boldsymbol{\beta}, \delta \mid \mathbf{y}\right). \tag{28}$$

Since  $g_Y(y_i \mid \boldsymbol{\beta}, \delta)$  is a function of  $f_X(x_i \mid \boldsymbol{\beta})$ , the MLE depends on the specification of the input density. Eq. (27) can be decomposed as

$$\ell(\beta, \delta \mid \mathbf{y}) = \ell(\beta \mid \mathbf{x}_{\tau}) + \mathcal{R}(\tau \mid \mathbf{y}), \qquad (29)$$

where

$$\ell\left(\boldsymbol{\beta}\mid\mathbf{x}_{\tau}\right) = \sum_{i=1}^{N} \log f_{X}\left(W_{\delta}\left(\frac{y_{i} - \mu_{x}}{\sigma_{x}}\right)\sigma_{x} + \mu_{x}\mid\boldsymbol{\beta}\right) = \sum_{i=1}^{N} \log f_{X}\left(\mathbf{x}_{\tau}\mid\boldsymbol{\beta}\right)$$
(30)



(a) Penalty  $R(\mu_x, \sigma_x, \delta \mid y_i)$  (32) as a function of  $\delta$  and  $y(\mu_x = 0)$  and  $\sigma_x = 1$ .

(b) Sample **z** of Lambert W × Gaussian with  $\delta = 1/3$  (left); Log-likelihood  $\ell(\theta \mid \mathbf{y})$  (solid, black) decomposes in input log-likelihood (dotted, green) and penalty (dashed, red).

Figure 6: Log-likelihood decomposition for Lambert W  $\times$   $F_X$  distributions.

is the log-likelihood of the back-transformed data  $\mathbf{x}_{\tau}$  and

$$\mathcal{R}\left(\tau \mid \mathbf{y}\right) = \sum_{i=1}^{n} \log R\left(\mu_{x}, \sigma_{x}, \delta \mid y_{i}\right), \tag{31}$$

where

$$R(\mu_{x}, \sigma_{x}, \delta \mid y_{i}) = \frac{W_{\delta}\left(\frac{y_{i} - \mu_{x}}{\sigma_{x}}\right)}{\frac{y_{i} - \mu_{x}}{\sigma_{x}}\left[1 + \delta\left(W_{\delta}\left(\frac{y_{i} - \mu_{x}}{\sigma_{x}}\right)\right)^{2}\right]}.$$
(32)

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

Decomposition (29) shows the difference between the exact MLE  $(\hat{\boldsymbol{\beta}}, \hat{\delta})$  based on  $\mathbf{y}$  and the approximate MLE  $\hat{\boldsymbol{\beta}}_{\mathbf{x}_{\tau}}$  based on  $\mathbf{x}_{\tau}$  alone: if we knew  $\tau = (\mu_x, \sigma_x, \delta)$  beforehand, then we could back-transform  $\mathbf{y}$  to  $\mathbf{x}_{\tau}$  and estimate  $\hat{\boldsymbol{\beta}}_{\mathbf{x}_{\tau}}$  from  $\mathbf{x}_{\tau}$  (maximize (30) with respect to  $\boldsymbol{\beta}$ ). In practice, however,  $\tau$  must also be estimated and this enters the likelihood via the additive term  $\mathcal{R}(\tau \mid \mathbf{y})$ . A little calculation shows that for any  $y_i \in \mathbb{R}$ ,  $\log R(\mu_x, \sigma_x, \delta \mid y_i) \leq 0$  if  $\delta \geq 0$ , with equality if and only if  $\delta = 0$ . Thus  $\mathcal{R}(\tau \mid \mathbf{y})$  can be interpreted as a penalty for transforming the data. Maximizing (29) faces a trade-off between transforming the data to follow  $f_X(x \mid \boldsymbol{\beta})$  (and thus increasing  $\ell(\boldsymbol{\beta} \mid \mathbf{x}_{\widehat{\tau}})$ ) versus the penalty of a more extreme transformation (and thus decreasing  $\mathcal{R}(\tau \mid \mathbf{y})$ ) – see Fig. 6b.

Figure 6a shows a 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 fixed y) or for larger y (for fixed  $\delta$ ). In both cases, increasing  $\delta$  makes the transformed data  $W_{\delta}(z)$  get closer to  $0 = \mu_x$ , which in turn increases its input likelihood. For  $\delta = 0$ , the penalty disappears since input equals output; for y = 0 there is no penalty since  $W_{\delta}(0) = 0$  for all  $\delta$ .

Figure 6b shows a random sample (N=1000)  $\mathbf{z} \sim \text{Lambert W} \times \text{Gaussian with } \delta = 1/3$  and the decomposition of the log-likelihood as in (29). Since  $\boldsymbol{\beta} = (0,1)$  is known, the likelihood and penalty are only functions of  $\delta$ . The monotonicity of the penalty (decreasing, red) and the input likelihood (increasing, green) as a function of  $\delta$  is not particular to this sample, but holds true in general (see Theorem 4.1 below). This monotonicity in each component implies that their sum (black line) has a unique maximum; here  $\hat{\delta}_{MLE} = 0.37$  (blue, dashed vertical line).

The maximization of (29) 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 for future work. Given the "nice" form of  $g_Y(y)$  - continuous, twice differentiable, its support does not depend on the parameter, etc. - the MLE for  $\theta = (\beta, \delta)$  should have the usual optimality properties (Lehmann and Casella, 1998).

#### 4.1.1 Properties of The MLE For The Heavy Tail Parameter

Without loss of generality let  $\mu_x = 0$  and  $\sigma_x = 1$ . In this case

$$\ell\left(\delta \mid \mathbf{z}\right) \propto -\frac{1}{2} \sum_{i=1}^{N} \left[W_{\delta}(z_i)\right]^2 + \sum_{i=1}^{N} \log \frac{W_{\delta}(z_i)}{z_i} - \log \left(1 + \delta \left[W_{\delta}(z_i)\right]^2\right)$$
(33)

$$= -\frac{1+\delta}{2} \sum_{i=1}^{N} [W_{\delta}(z_i)]^2 - \sum_{i=1}^{N} \log \left(1 + \delta [W_{\delta}(z_i)]^2\right). \tag{34}$$

Theorem 4.1 (Unique MLE for  $\delta$ ). Let Z have a Lambert W × Gaussian distribution,

where  $\mu_x = 0$  and  $\sigma_x = 1$  are assumed to be known and fixed. Also consider only the case

Assuming that  $f_X(\cdot)$  is twice differentiable.

 $\delta \in [0, \infty).^{6}$ 

$$\frac{\sum_{i=1}^{n} z_i^4}{\sum_{i=1}^{n} z_i^2} \le 3,\tag{35}$$

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

322 If (35) does not hold, then

b)  $\hat{\delta}_{MLE} > 0$  exists and is a positive solution to

$$\sum_{i=1}^{N} z_i^2 W'(\delta z_i^2) \left( \frac{1}{2} W_{\delta}(z_i)^2 - \left( \frac{1}{2} + \frac{1}{1 + W(\delta z_i^2)} \right) \right) = 0.$$
 (36)

 $_{23}$  c) There is only one such  $\delta$  satisfying (36), i.e.  $\widehat{\delta}_{MLE}$  is unique.

Condition (35) says that  $\hat{\delta}_{MLE} > 0$  only 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)  $\nu$  ( $\leftrightarrow$  large  $\delta$ ) (see also Fernandez and Steel, 1999; Liu and Rubin, 1995). On the contrary,  $\hat{\delta}_{MLE}$  is always a global maximum.

329

330

331

332

333

334

335

The log-likelihood and its gradient depend on  $\delta$  and  $\mathbf{z}$  only via  $W_{\delta}(\mathbf{z})$ . Given the heavy tails in  $\mathbf{z}$  (for  $\delta > 0$ ) one might expect convergence issues for larger  $\delta$  (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  $\widehat{\delta}_{MLE} \approx \delta$ . 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. Simulations in Section 5 support this conjecture, even for  $\widehat{\theta}_{MLE}$ .

# 336 4.2 Iterative Generalized Method of Moments (IGMM)

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

<sup>&</sup>lt;sup>6</sup>While for some samples **z** the MLE also exists for  $\delta < 0$ , it can not be guaranteed for all **z**. 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.

 $f_X(x \mid \beta)$ . It would be useful to directly estimate  $\tau$ , without the intermediate step of estimating  $\theta$  first (and thus no distributional assumption for the input is necessary).

Goerg (2011) presented an estimator for  $\tau$  based on iterative generalized methods of moments (IGMM). The idea of IGMM is to find a  $\tau$  such that the back-transformed data  $\mathbf{x}_{\tau}$  has desired properties, e.g., is symmetric or has kurtosis 3. An estimator for  $\mu_x$ ,  $\sigma_x$ , and  $\delta$  can be constructed completely analogously to the skewed IGMM, with the advantage that the heavy tail transformation is bijective (the skewed transformation is not). Since the algorithm is entirely analogous to the skewed case, details are given in the Supplementary Material, Appendix  $\mathbb{C}$ .

348

341

342

343

344

345

346

347

An advantage of IGMM is that it requires less specific knowledge about the input distribution. Usually, it is also faster than the MLE. Once  $\hat{\tau}_{IGMM}$  has been obtained, the back-transformed  $\mathbf{x}_{\hat{\tau}_{IGMM}}$  can be used to check if X has characteristics of a known parametric distribution  $F_X(x \mid \boldsymbol{\beta})$ . It must be noted though that testing for a particular distribution  $F_X$  are too optimistic as  $\mathbf{x}_{\hat{\tau}}$  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 should not matter for all practical purposes.

#### 5 Simulations

This section explores finite sample properties of estimators for  $\theta = (\mu_x, \sigma_x, \delta)$  and  $(\mu_y, \sigma_y)$  under Gaussian input  $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ . In particular, it compares Gaussian MLE (estimation of  $\mu_y$  and  $\sigma_y$  only), IGMM and Lambert W × Gaussian MLE, and - for a heavy tail competitor – the median. All results below are based on n = 1,000 replications.

#### $_{361}$ 5.1 Estimating $\delta$ Only

Here I show finite sample properties of  $\widehat{\delta}_{MLE}$  for  $U \sim \mathcal{N}(0,1)$ , where  $\mu_x = 0$  and  $\sigma_x = 1$  are known and fixed. Theorem 4.1 shows that  $\widehat{\delta}_{MLE}$  is unique: either at the boundary  $\delta = 0$  or at the globally optimal solution to (36). Results in Table 1 were obtained by numerical optimization restricted to  $\delta \geq 0$  ( $\Leftrightarrow \log \delta \in \mathbb{R}$ ) using the nlm function in R.

<sup>&</sup>lt;sup>7</sup>For IGMM, optimization was restricted to  $\delta \in [0, 10]$ .

Table 1: Finite sample properties of  $\hat{\delta}_{MLE}$ . For each N,  $\delta$  was estimated n=1,000 times from a random sample  $\mathbf{z} \sim \text{Tukey's } h$ . The left column for each  $\delta$  shows bias,  $\overline{\hat{\delta}}_{MLE} - \delta$ ; each right column shows the root mean square error (RMSE) times  $\sqrt{N}$ .

N	$\delta = 0$			$\delta = 1$	/10	$\delta = 1$	1/3		$\delta = 1/2$		
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	
N	$\delta =$	1	•	$\delta = 2$		$\delta = 5$					
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				

Table 1 shows that the MLE is unbiased for every  $\delta$  and settles down (about N=100) 366 to an asymptotic variance, which is increasing with  $\delta$ . Assuming  $\mu_x$  and  $\sigma_x$  to be known 367 is unrealistic and thus these finite sample properties are only an indication of the behavior 368 of the joint MLE,  $\hat{\theta}_{MLE}$ . Nevertheless they are very remarkable for extremely heavy-tailed 369 data ( $\delta > 1$ ), where standard statistical methods typically break down. One explanation in 370 this behavior lies in the particular form of the likelihood (33) and its gradient (36) (Theorem 371 4.1). Although both depend on  $\mathbf{z}$ , they only do so through  $W_{\delta}(\mathbf{z}) = \mathbf{u} \sim \mathcal{N}(0,1)$ . Hence 372 as long as  $\widehat{\delta}_{MLE}$  is sufficiently close to the true  $\delta$ , (33) and (36) are functions of almost 373 Gaussian RVs and standard asymptotic results should still apply. 374

#### 5.2 Estimating All Parameters Jointly

375

Here we consider the realistic scenario where  $\mu_x$  and  $\sigma_x$  are also unknown. We consider various sample sizes ( $N=50,\ 100,\$ and 1000) and different degrees of heavy tails,  $\delta\in$  $\{0,1/3,1,1.5\}$ , each one representing a particularly interesting situation: i) Gaussian data (does additional - superfluous - estimation of  $\delta$  affect other estimates?), ii) fourth moments do not exist anymore, iii) non-existing mean, iv) extremely heavy-tailed data – can we get

#### useful estimates at all?

387

Table 2: In each subtable: (first rows) average, (middle rows) proportion of estimates below truth, (bottom rows) empirical standard deviation times  $\sqrt{N}$ .

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

$\delta = 0$	median	Gaussia	an MLE		IGI	ИM		I	Lambert	W MLI	3	NA
N		$\mu_y$	$\sigma_y$	$\mu_x$	$\sigma_x$	δ	$\sigma_y$	$\mu_x$	$\sigma_x$	δ	$\sigma_y$	ratio
50	0.00	0.00	0.98	0.00	0.97	0.02	0.99	0.00	0.96	0.02	0.98	0
100	0.00	0.00	0.99	0.00	0.98	0.01	1.00	0.00	0.97	0.01	0.99	0
1000	0.00	0.00	1.00	0.00	0.99	0.00	1.00	0.00	0.99	0.00	1.00	0
50	0.50	0.50	0.57	0.51	0.60	0.66	0.54	0.51	0.65	0.66	0.56	0
100	0.50	0.51	0.56	0.51	0.62	0.62	0.53	0.52	0.65	0.62	0.56	0
1000	0.50	0.49	0.52	0.49	0.62	0.56	0.52	0.49	0.63	0.56	0.52	0
50	1.24	1.01	0.72	1.01	0.76	0.21	0.73	1.02	0.78	0.26	0.72	0
100	1.25	1.02	0.70	1.02	0.76	0.23	0.70	1.03	0.78	0.26	0.70	0
1000	1.26	0.98	0.73	0.98	0.79	0.22	0.73	0.98	0.79	0.22	0.73	0

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

$\delta = 1/3$	median	Gaussia	an MLE		IG	MM			NA			
N		$\mu_y$	$\sigma_y$	$\mu_x$	$\sigma_x$	δ	$\sigma_y$	$\mu_x$	$\sigma_x$	δ	$\sigma_y$	ratio
50	0.00	0.00	1.98	0.00	1.07	0.29	$\infty$	0.00	1.01	0.33	$\infty$	0
100	0.00	0.00	2.03	0.00	1.04	0.31	$\infty$	0.00	1.00	0.33	$\infty$	0
1000	0.00	0.00	2.18	0.00	1.00	0.33	2.34	0.00	1.00	0.33	2.34	0
50	0.50	0.51	0.78	0.50	0.38	0.63	0.60	0.50	0.52	0.54	0.54	0
100	0.50	0.51	0.78	0.51	0.42	0.61	0.60	0.50	0.51	0.54	0.54	0
1000	0.48	0.51	0.77	0.51	0.47	0.56	0.55	0.51	0.50	0.53	0.52	0
50	1.27	2.21	6.56	1.44	1.45	1.10	NA	1.23	1.35	1.14	NA	0
100	1.30	2.33	11.28	1.43	1.42	1.12	NA	1.19	1.34	1.09	NA	0
1000	1.23	2.25	16.76	1.39	1.45	1.20	15.97	1.17	1.33	1.08	12.30	0

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

$\delta = 1$	median	Gaussi	an MLE		IGMM				Lambert W MLE			
N		$\mu_y$ $\sigma_y$		$\mu_x$	$\sigma_x$	δ	$\sigma_y$	$\mu_x$	$\sigma_x$	δ	$\sigma_y$	ratio
50	0.00	-0.10	24.6	-0.01	1.18	0.90	$\infty$	0.00	1.01	0.99	$\infty$	0
100	0.00	0.74	72.4	0.00	1.09	0.95	$\infty$	0.00	1.01	0.99	∞	0
1000	0.00	3.84	348.1	0.00	1.01	1.00	$\infty$	0.00	1.00	1.00	∞	0
50	0.53	0.52	1.0	0.51	0.34	0.65	1	0.51	0.52	0.52	1	0
100	0.50	0.52	1.0	0.51	0.38	0.63	1	0.50	0.53	0.53	1	0
1000	0.49	0.52	1.0	0.51	0.48	0.53	1	0.49	0.51	0.51	1	0
50	1.27	65.85	424.3	2.10	2.50	2.32	NA	1.19	1.70	2.16	NA	0
100	1.30	410.75	4050.2	2.01	2.28	2.59	NA	1.17	1.74	2.25	NA	0
1000	1.26	3307.58	104052.7	1.93	2.21	2.81	NA	1.11	1.64	2.18	NA	0

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

$\delta = 1.5$	median	Gaussiai	n MLE		IGMM				Lambert W MLE				
N		$\mu_y$	$\sigma_y$	$\mu_x$	$\sigma_x$	δ	$\sigma_y$	$\mu_x$	$\sigma_x$	δ	$\sigma_y$	ratio	
50	-0.02	6.84	309	-0.02	1.23	1.37	$\infty$	-0.01	1.00	1.49	$\infty$	0.01	
100	0.00	-51.16	3080	-0.01	1.12	1.44	$\infty$	0.00	1.01	1.50	$\infty$	0.00	
1000	0.00	176.13	14251	0.00	1.01	1.49	$\infty$	0.00	1.00	1.50	$\infty$	0.00	
50	0.53	0.48	1	0.51	0.34	0.64	1	0.53	0.53	0.54	1	0.01	
100	0.51	0.53	1	0.54	0.37	0.61	1	0.52	0.51	0.51	1	0.00	
1000	0.50	0.50	1	0.50	0.47	0.54	1	0.49	0.53	0.52	1	0.00	
50	1.32	1347.71	9261	2.57	3.20	3.12	NA	1.15	1.86	2.76	NA	0.01	
100	1.33	42156.28	418435	2.39	2.87	3.44	NA	1.12	1.78	2.84	NA	0.00	
1000	1.26	124462.82	3903629	2.18	2.66	3.67	NA	1.11	1.80	2.85	NA	0.00	

The convergence tolerance for IGMM was set to  $tol = 1.22 \cdot 10^{-4}$ . Table 5 summarizes the simulation. Each sub-table is organized as follows: columns represent parameter estimates; the three main rows are the average over n = 1,000 replications (top), the proportion of estimates below the true value (middle), and the empirical standard deviation around the empirical average times  $\sqrt{N}$  – not around the truth (bottom).

The Gaussian MLE estimates  $\sigma_y$  directly, while IGMM and the Lambert W  $\times$  Gaussian

MLE estimates  $\delta$  and  $\sigma_x$ , which implicitly give  $\widehat{\sigma}_y$  through  $\sigma_y(\delta, \sigma_x) = \sigma_x \cdot \frac{1}{\sqrt{(1-2\delta)^{3/2}}}$  if  $\delta < 1/2$  (see (21)). For a fair comparison each sub-table also includes a column for  $\widehat{\sigma}_y = \widehat{\sigma}_x \cdot \frac{1}{\sqrt{(1-2\widehat{\delta})^{3/2}}}$ . Some of these entries contain " $\infty$ ", even for  $\delta < 1/2$ ; this occurs if at least one  $\widehat{\delta} \geq 1/2$ .

For any  $\delta < 1$ ,  $\mu_x = \mu_y$ , thus they can be directly compared. For  $\delta \ge 1$ , the mean does not exist; each sub-table for these  $\delta$  interprets  $\mu_y$  as the median.

Gaussian data:  $\delta = 0$  This setting checks if imposing the Lambert W framework, even 394 though its use is superfluous, causes a quality loss in the estimation of  $\mu_y = \mu_x$  or  $\sigma_y = \sigma_x$ . 395 Furthermore, critical values for  $H_0: \delta = 0$  (Gaussian tails) can be obtained. Table 2a shows 396 that all estimators are unbiased and quickly tend to a large-sample variance. Additional 397 estimation of  $\delta$  does not affect the efficiency of  $\widehat{\mu}_x$  compared to estimating solely  $\mu$  (both 398 for IGMM and Lambert W  $\times$  Gaussian MLE). Estimating  $\sigma_y$  directly by Gaussian MLE 399 does not give better results than the Lambert W × Gaussian MLE: both are unbiased and 400 have similar standard deviation. 401

No fourth moment:  $\delta = 1/3$  Here  $\sigma_y(\delta, \sigma_x = 1) = 2.28$ , but fourth moments do not exist anymore. This results 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 is much less variable than Gaussian MLE and IGMM. Yet, Lambert W × Gaussian MLE for  $\mu_x$  even outperforms the median.

Non-existing mean:  $\delta=1$  Here the mean is non-finite. Thus both sample moments diverge, and their standard errors are also growing quickly. 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 continue to be unbiased even though the data is extremely heavy-tailed. Moreover, Lambert W MLE also has the smallest empirical standard deviation overall. In particular, the Lambert W MLE for  $\mu_x$  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  $\infty$ ). Here the search for the optimal  $\delta$  lead into regions with a numerical overflow in the evaluation of  $W_{\delta}(z)$ . For a comparable summary, 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 in statistical practice.

#### 5.3 Discussion of the Simulations

422

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

IGMM is unbiased for  $\tau$  independent of the magnitude of  $\delta$ . As expected the Lambert W

MLE for  $\theta$  has the best properties: it is unbiased for all  $\delta$ , and for  $\delta = 0$  it performs as well

as the classic sample mean and standard deviation. For small  $\delta$  it has the same empirical

standard deviation as the Gaussian MLE, but a lower one than the median for large  $\delta$ .

Hence the only advantage of estimating  $\mu_y$  and  $\sigma_y$  by sample moments of  $\mathbf{y}$  is speed; otherwise the Lambert W × Gaussian MLE is at least as good as the Gaussian MLE and clearly outperforms it for heavy-tailed data.

# 434 6 Applications

Tukey's h distribution has already proven useful to model heavy-tailed data, but parametric inference was limited to quantile fitting or methods of moments estimation (Field, 2004; Fischer, 2010; Headrick et al., 2008). Theorem 2.8 allows us to estimate θ by ML.

This section shows the usefulness of the presented methodology on simulated as well as real world data: i) Section 6.1 demonstrates Gaussianizing on the Cauchy sample from the Introduction; ii) Section 6.2 shows that heavy tail Lambert W × Gaussian distributions provide an excellent fit to daily S&P 500 log-return series; and iii) Section 6.3 shows how removing heavy tails reveals hidden patterns in power-law type data.

#### 6.1 Estimating Location of a Cauchy With The Sample Mean

It is well-known that the sample mean  $\overline{\mathbf{y}}$  is a poor estimate of the location parameter of a Cauchy distribution, since the sampling distribution of  $\overline{y}$  is again a Cauchy; in particular, 445 its variance does not go to 0 for  $n \to \infty$ . 446 Heavy-tailed Lambert W × Gaussian distributions have similar properties to a Cauchy 447 for  $\delta \approx 1$ . The mean of X equals the location of Y, due to symmetry around  $\mu_x$  (for all 448  $\delta \geq 0$ ) and c, respectively. Thus we can estimate  $\tau$  from the Cauchy sample y, transform y 449 to  $\mathbf{x}_{\widehat{\tau}}$ , estimate  $\mu_x$  from  $\mathbf{x}_{\widehat{\tau}} = W_{\widehat{\tau}}(\mathbf{y})$ , and thus obtain an estimate of c. 450 451 The data  $\mathbf{y} \sim \mathcal{C}(0,1)$  in Fig. 2a has heavy tails with two extreme (positive) samples. A 452 Cauchy ML fit gives  $\hat{c} = 0.03(0.055)$  and  $\hat{s} = 0.86(0.053)$  (standard errors in parenthesis). A 453 Lambert W × Gaussian MLE gives  $\hat{\mu}_x = 0.03(0.055)$ ,  $\hat{\sigma}_x = 1.05(0.072)$ , and  $\hat{\delta} = 0.86(0.082)$ . Thus both fits correctly fail to reject  $\mu_x = c = 0$ . Table 3a shows summary statistics on both 455 samples. Since the Cauchy distribution does not have a well-defined mean,  $\overline{y} = 2.304(2.101)$ 456 is not meaningful. However,  $\mathbf{x}_{\widehat{\tau}_{MLE}}$  is approximately Gaussian and we use the sample av-457 erage to do inference:  $\bar{\mathbf{x}}_{\hat{\tau}_{MLE}} = 0.033(0.0472)$  correctly fails to reject a zero location for 458 y. The transformed  $\mathbf{x}_{\widehat{\tau}_{MLE}}$  features additional Gaussian characteristics (symmetric, no ex-459 cess kurtosis), and even the null hypothesis of Normality cannot be rejected (p-value  $\geq 0.5$ ). 460 461 Figure 2d shows the running sample average for the original sample and its Gaussianized 462 version. For a fair comparison  $\hat{\tau}_{MLE}^{(n)}$  was re-estimated cumulatively for each  $n=5,\ldots,500,$ 463 and then used to compute  $(x_1, \ldots, x_n)$ . Even for small n the transformation works extremely well: the highly influential point around  $n \approx 50$  greatly affects  $\overline{y}$ , but has no relevant ef-465 fect on  $\overline{\mathbf{x}}_{\widehat{\tau}_{MLE}^{(n)}}$ . Overall, the sample average of the Gaussianized data has the usual good 466 properties. And even for very small n it is already clear that the location of the underlying 467 Cauchy distribution is approximately zero. 468 469 Although a toy example, it shows that removing (strong) heavy tails from data works 470

and provides new, "nice" data which can then be used for more refined methods.

Table 3: Summary statistics for observed (heavy-tailed)  $\mathbf{y}$  and back-transformed (Gaussianized) data  $\mathbf{x}_{\widehat{\tau}_{MLE}}$ . \*\* stands for  $< 10^{-16}$ ; \* for  $< 2.2 \cdot 10^{-16}$ .

(a) $\mathbf{y} \sim \mathcal{C}(0,1)$	(b) $\mathbf{y} = S\&P$	500(c) <b>y</b> = solar flares
(Section $6.1$ )	(Section 6.2)	(Section $6.3$ )

	y	$\mathbf{x}_{\widehat{\tau}}$	$\mathbf{x}_{\widehat{\lambda}}$	y	$\mathbf{X}_{\widehat{\mathcal{T}}}$	$\mathbf{y}$	$\mathbf{x}_{\widehat{\tau}}$
Min	-161.59	-3.16	0	-7.11	-2.42	20	20
Max	952.95	3.81	33.18	4.99	2.23	231300	157
Mean	2.30	0.03	14.98	0.05	0.05	689.4	89.0
Median	0.04	0.04	14.96	0.04	0.04	87	87
$\operatorname{Stdev}$	46.980	1.06	1.20	0.95	0.71	6520.6	27.0
Skewness	17.43	0.12	3.90	-0.30	-0.04	22.2	0.1
Kurtosis	343.34	3.21	161.75	7.70	2.93	582.1	1.9
$\overline{\text{SW}}$	*	0.71	**	*	0.24	**	**
AD	**	0.51	**	*	0.18	**	**
	I I			1		l I	

#### 472 6.2 Heavy Tails in Finance: 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 modeled 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 innovations), and simply "Gaussianize" the reutns before fitting more complex – GARCH or SV – models.

Remark 6.1. Time series models with Lambert  $W \times Gaussian$  white noise are far beyond the scope of this work, but can be a direction of future research. Here I only consider the unconditional distribution.

Figure 7a shows the S&P 500 log-returns with a total of N=2,780 daily observations. Table 3b confirms the heavy tails (sample kurtosis 7.70), but also indicates negative skewness (-0.296). As the sample skewness (-0.296) is very sensitive to outliers, 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 × Gaussian this means to test  $H_0: \delta_\ell = \delta_r = \delta$  versus  $H_1: \delta_\ell \neq \delta_r$ . Since the likelihood can now be computed by (29), we can use a likelihood

<sup>&</sup>lt;sup>8</sup>R package MASS, dataset SP500.

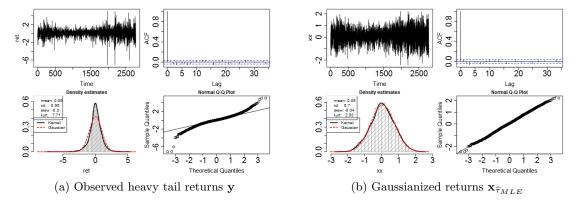


Figure 7: Lambert W Gaussianization of S&P 500 log-returns:  $\hat{\tau} = (0.05, 0.70, 0.17)$ . In (a) and (b): data (top left); autocorrelation function (ACF) (top right); histogram, Gaussian fit, and KDE (bottom left); Normal QQ plot (bottom right).

ratio test with one degree of freedom (3 versus 4 parameters). The log-likelihood of the double-tail Lambert W × Gaussian fit (Table 4a) equals -3606.0 = -2972.27 + (-633.73) (input + penalty), while the one  $\delta$  fit gives -3606.56 = -2971.47 + (-635.09). Here the double tails pay a lower penalty for transforming the data, but in turn give less Gaussian transformed sample. Comparing twice their difference to a  $\chi_1^2$  distribution gives a p-value of 0.29. For comparison, a skew-t fit (Azzalini and Capitanio, 2003), with location c, scale s, shape a, and  $\nu$  degrees of freedom also yields a non-significant a (Table 4b). Thus both fits cannot 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 Returns

498

499

500

501

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). However, a plain sample average over-estimates the variance in presence of heavy tails, and thus adds bias to the test statistic.

<sup>&</sup>lt;sup>9</sup>Function st.mle in the R package sn.

Table 4: MLE fits to S&P 500 y (a, b, c, d, e) and the Gaussianized data  $\mathbf{x}_{\widehat{\tau}_{MLE}}$  (f).

double-tail Lambert W  $\times$  Gaussian = Tukey's hh (S&P 500) (b) skew t (S&P 500)  $\Pr(> |t|)$  $\Pr(> |t|)$ Est. se t Est. se t 0.015 3.66 0.061 1.65 0.06 0.00 0.10 0.10c $\mu_x$ 0.710.01644.00 0.00 0.670.017 38.47 0.00s $\sigma_x$  $\delta_{\ell}$ 0.190.021 8.99 0.00 -0.080.101-0.770.44 $\alpha$ 0.160.019 8.24 0.003.730.29712.570.00  $\delta_r$ ν (c) Lambert W  $\times$  Gaussian = Tukey's h (S&P 500) student-t (S&P 500) (d)  $\Pr(> \mid t \mid)$ Est. se  $\Pr(> |t|)$ Est. se 3.65 0.000 0.06 0.015 3.65 0.00 0.060.015c $\mu_x$ 0.710.01643.950.0000.670.01739.51 0.00s $\sigma_x$ δ 0.170.01611.05 0.000 3.720.29512.61 0.00(e) Gaussian (S&P 500) Gaussian  $(\mathbf{x}_{\widehat{\tau}_{MLE}})$  $\Pr(> |t|)$ Est.  $\Pr(> |t|)$ Est. se t se 0.050.018 2.550.01 0.050.013 3.81 0.00  $\mu_y$  $\mu_{x_{\widehat{\tau}}}$  $\sigma_y$ 0.95 74.570.0074.570.00

0.71

 $\sigma_{x_{\widehat{\tau}}}$ 

0.009

#### 6.2.2Heavy Tail Fit to Returns 506

510

511

512

513

514

515

516

517

518

519

520

0.013

Both a heavy tail Lambert W × Gaussian (Table 4c) and student-t fit (Table 4d) reject 507 the zero mean null (p-values,  $10^{-4}$  and  $3 \cdot 10^{-5}$ , respectively). The standard errors for the 508 location parameter are essentially the same. 509

While location and scale estimates are almost identical, the tail estimates lead to very different conclusions: while for  $\hat{\nu} = 3.71$  only moments up to order 3 exist, in the Lambert W  $\times$  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 finite fourth moments (Mantegna and Stanley, 1998; Zadrozny, 2005); consequently many empirical studies test if financial data actually satisfy this assumption (Cont, 2001; Huisman, Koedijk, Kool, and Palm, 2001). For this particular dataset student's t and a Lambert W  $\times$  Gaussian fit give different answers to the same question. Since previous empirical studies often use student's t as a baseline (Wong et al., 2009), it might be worthwhile to re-examine their findings in light of heavy tail Lambert  $W \times Gaussian$  distributions.

#### 1 6.2.3 "Gaussianizing" Returns

A typical parameter inference study would conclude here. Using Lambert's W function we can analyze the back-transformed  $\mathbf{x}_{\widehat{\tau}_{MLE}}$  to test if a Lambert W × Gaussian distribution is indeed appropriate. Figure 7b shows that  $\mathbf{x}_{\widehat{\tau}_{MLE}}$  is indistinguishable from a Gaussian sample. Not even one Normality test can reject Gaussianity: p-values are 0.18, 0.18, 0.31, and 0.24, respectively (Anderson Darling, Cramer-von-Mises, Shapiro-Francia, Shapiro-Wilk; see Thode (2002)). Table 3b also shows that Lambert W "Gaussianiziation" was successful:  $\widehat{\gamma}_2(\mathbf{x}_{\widehat{\tau}}) = 2.93$  and  $\widehat{\gamma}_2(\mathbf{x}_{\widehat{\tau}}) = -0.039$  are within the typical variation for a Gaussian sample. Thus

$$Y = \left(Ue^{\frac{0.172}{2}U^2}\right)0.705 + 0.055, \quad U = \frac{X - 0.055}{0.705}, \quad U \sim \mathcal{N}(0, 1)$$
 (37)

is an adequate (unconditional) Lambert W × Gaussian model for the S&P 500 log-returns y. For trading, this means that the expected return is significantly larger than zero ( $\hat{\mu}_x = 0.055 > 0$ ), and thus replicating certificates should be bought.

#### 525 6.2.4 Gaussian MLE for Gaussianized Data

534

535

536

537

538

For  $\delta_l = \delta_r \equiv \delta < 1$ , also  $\mu_x \equiv \mu_y$ . We can therefore replace testing  $\mu_y = 0$  versus  $\mu_y \neq 0$  for a non-Gaussian  $\mathbf{y}$ , with the very well understood hypothesis test  $\mu_x = 0$  versus  $\mu_x \neq 0$  for the Gaussian  $\mathbf{x}_{\widehat{\tau}_{MLE}}$ . In particular, standard errors based on  $\frac{\widehat{\sigma}}{\sqrt{N}}$  - and thus t and p-values - should be closer to the "truth" (Table 4c and 4d) than a Gaussian MLE on the non-Gaussian  $\mathbf{y}$  (Table 4e). Table 4f shows that standard errors for  $\widehat{\mu}_{\mathbf{x}}$  are even a bit too small compared to the heavy-tailed versions. Since the "Gaussianizing" transformation was estimated, treating  $\mathbf{x}_{\widehat{\tau}_{MLE}}$  as if it was original data is too optimistic regarding its Gaussianity (recall the penalty (31) in the total likelihood (29)).

This example confirms that if a model and its theoretical properties are based on Gaussianity, but the observed data is heavy-tailed, then Gaussianizing 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 × Gaussianizing the data first gives more

sian errors (or any other heavy-tailed distribution) would be optimal. However, theoretical properties and estimation techniques may not have been developed and implemented yet, or are simply not known to researchers who are non-experts in heavy-tailed statistics. The Lambert Way to Gaussianize data thus is a pragmatic method to improve statistical inference on heavy-tailed data, while preserving the ease of usage and interpretation of Gaussian models.

#### 6.3 Removing Power Law From Solar Flare Counts

The previous section focused on Lambert W ×  $F_X$  distributions as a "true" model for the data  $\mathbf{y}$ . Here I consider it merely as a data transformation to remove heavy tails. In the same way as scaling  $\mathbf{y}$  to zero-mean, unit-variance data,  $(\mathbf{y} - \overline{\mathbf{y}})/\widehat{\sigma}_y$ , does not necessarily mean we believe the underlying process is Gaussian, we can also convert  $\mathbf{y}$  to  $\mathbf{x}_{\tau} = W_{\tau}(\mathbf{y})$  without assuming that  $\mathbf{y}$  is actually Lambert W × Gaussian. While  $\mathbf{x}_{\widehat{\tau}}$  might lose the interpretability of the observed data (e.g. units become distorted), it can be helpful for exploratory data analysis (EDA), as the eye is a bad judgment to detect regularities corrupted by heavy tails. Removing them can reveal hidden patterns and thus greatly improve the accuracy of statistical inference for  $\mathbf{y}$ .

Here I study solar flare gamma-ray count rates (Clauset, Shalizi, and Newman, 2009; Newman, 2005). The data<sup>10</sup> 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 details and scientific background.

The gamma-ray count rates exhibit a strong right heavy tail (Fig. 8a), which makes more detailed visual inspection as well as simple EDA difficult. A zoom to  $y_i \leq 400$  in Fig. 8d shows that a lot of counts lie between 50 and 100 and this level drops off at the end of the observation cycle. This drop is not an intrinsic characteristic of solar flares but due to a decreasing sensitivity of the X-ray detectors over time (Dennis et al., 1991). For the sake of comparison with Clauset et al. (2009); Newman (2005) most estimates are based on all

<sup>&</sup>lt;sup>10</sup>Dataset SolarFlares in the LambertW package.

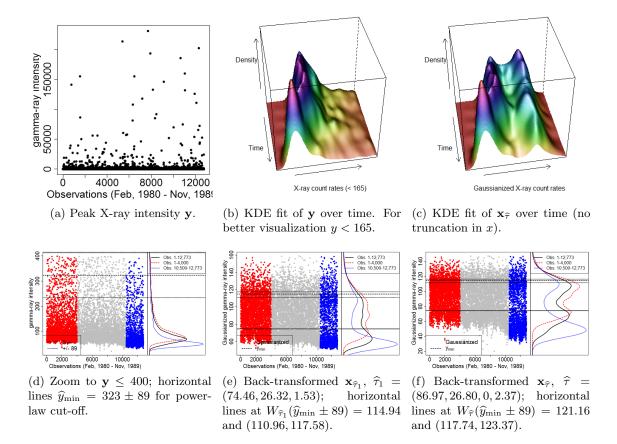


Figure 8: Peak X-ray count rates of solar flares.

T=12,773 observations. Figures 8d, 8e, and 8f also show separate density estimates 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 ( $\hat{a} = 1.79(0.02)$ ) with cut-off ( $\hat{y}_{min} = 323(89)$ ) 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. 8d and 8b) - are affected by the heavy right tail. Here I show that Gaussianizing this data reveals new insights for the data-generating process, with a new interpretation for the optimality of the cut-off.

A Lambert W × Gaussian MLE fit  $\hat{\theta} = (\hat{\mu}_x, \hat{\sigma}_x, \hat{\delta_\ell}, \hat{\delta_r}) = (86.97, 26.80, 0, 2.37)$  confirms

that only the right tail (tail index 1/2.373 = 0.421) needs a Gaussianizing transformation.<sup>11</sup> 579 The last column of Fig. 8 shows EDA for the Gaussianized data. Removing the heavy 580 right tail reveals a bimodal structure, which gives additional meaning to  $\hat{y}_{\min} = 323$ . The 581 Gaussianized cut-off value equals  $W_{\hat{\tau}}(323) = 121.16$  with the transformed standard devia-582 tion interval [117.74, 123.37] (corresponding to  $323\pm89$ ). Fitting a two component Gaussian 583 mixture model to  $\mathbf{x}_{\widehat{\tau}}$  yields  $\widehat{\lambda} \mathcal{N}_1(67.10, 14.04^2) + (1 - \widehat{\lambda}) \mathcal{N}_2(113.12, 14.27^2)$  with  $\widehat{\lambda} = 0.52$ 584 and optimal decision boundary between classes of 90.48. The mean of the larger component, 585 113.12, lies within one standard deviation of the optimal Gaussianized cut-off 121.16: for 586 lower cut-offs the left-tail of the larger component – or for much lower cut-offs even the 587 smaller component – would counteract the power-law decay of the upper gamma-ray count rates. 589

590

591

592

593

594

595

As mentioned above, this analysis is not intended to describe the underlying process of solar flare gamma rays; it should rather show new insights that can be gained by Gaussianizing. Future research based on these new findings might lead to new physical interpretations of the statistical properties gamma-ray count rates, see for example Aschwanden (2011).

#### 7 Discussion and Outlook

I adapt the skewed Lambert W input / output framework to introduce heavy tails in continuous RVs  $X \sim F_X(x)$ . For Gaussian input this not only contributes to existing work on Tukey's h distribution, but also gives convincing empirical results: unimodal data with heavy tails can be transformed to Gaussian data/RVs. Properties of a Gaussian model  $\mathcal{M}_N$  on the back-transformed data mimic the features of the "true" skewed, heavy-tailed model  $\mathcal{M}_G$  very closely.

Since Gaussianity is the single most typical, and often required, assumption in many areas of statistics, machine learning, and signal processing, future research can take many directions. From a theoretical perspective properties of Lambert W  $\times$   $F_X$  distributions viewed as a generalization of already well-known distributions  $F_X$  can be studied. This area

<sup>&</sup>lt;sup>11</sup>For comparison Fig. 8e also shows the back-transformed data  $\mathbf{x}_{\hat{\tau}_1}$  using the same  $\delta$  on each tail ( $\hat{\tau}_1 = (74.46, 26.32, 1.53)$ ). However, due to the clear right heavy tail I will continue with the  $(\delta_l, \delta_r)$  transformation.

will profit from existing literature on the Lambert W function, which has been discovered only recently by the statistics community. Empirical work can focus on transforming the data and compare performances of approximate Gaussian versus joint heavy-tail analysis. The comparisons in this work showed that approximate inference for Gaussianized data is comparable with the direct heavy tail modeling, and so provides an easy tool to improve inference for heavy-tailed data in statistical practice.

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

#### 614 Acknowledgments

I want to thank Andrew F. Siegel who brought Tukey's h distribution to my attention, and Brian R. Dennis who gave detailed background information and suggestions on the solar flares dataset.

#### 618 References

- 619 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.
- Aschwanden, M. J. (2011). The State of Self-Organized Criticality of the Sun During the
- Last Three Solar Cycles. II. Theoretical Model. Solar Physics 274, 119–129.
- 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.
- 626 Azzalini, A. and A. Capitanio (2003). Distributions generated by perturbation of symmetry
- 627 with emphasis on a multivariate skew t distribution. Journal of the Royal Statistical
- Society ser B 65, 367–389.
- 629 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 In-
- ference 140(7), 1957 1967.
- 632 Blaylock, J. R., L. E. Salathe, and R. D. Green (1980). A note on the Box-Cox trans-
- formation under heteroskedasticity. Western Journal of Agricultural Economics 05(02),
- 634 129–136.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of
- Econometrics 31, 307 327.
- <sup>637</sup> Brockwell, P. J. and R. A. Davis (1998). Time Series: Theory and Methods. Springer Series
- in Statistics.
- <sup>639</sup> Clauset, A., C. R. Shalizi, and M. E. J. Newman (2009). Power-law distributions in empirical
- data. SIAM Review 51, 661–703.
- 641 Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues.
- 642 Quantitative Finance 1, 223–236.

- Corless, R. M., G. H. Gonnet, D. E. G. Hare, and D. J. Jeffrey (1996). On the Lambert W
   function. Advances in Computational Mathematics 5, 329–359.
- 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-
- 647 1989. Available http://adsabs.harvard.edu/abs/1991chxb.book.....D and http:
- //umbra.nascom.nasa.gov/smm/hxrbs.html.
- 649 Deo, R., C. Hurvich, and Y. Lu (2006). Forecasting Realized Volatility Using a Long
- 650 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. Technical
- report, Wharton School Center for Financial Institutions, University of Pennsylvania.
- Engle, R. (1982). Autoregressive conditional heteroskedasticity with estimates of the vari-
- ance of U.K. inflation. Econometrica 50, 987 1008.
- <sup>657</sup> Fernandez, C. and M. F. J. Steel (1999). Multivariate Student-t Regression Models: Pitfalls
- and Inference. Biometrika 86, 153–167.
- Field, C. and M. G. Genton (2006). The Multivariate g-and-h Distribution. Technomet-
- rics 48(1), 104–111.
- 661 Field, C. A. (2004). Using the gh distribution to model extreme wind speeds. Journal of
- Statistical Planning and Inference 122(1-2), 15-22.
- 663 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.
- 665 Galassi, M., J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, and F. Rossi
- 666 (2011). GNU Scientific Library Reference Manual (3rd ed.). ISBN 0954612078; www.gnu.
- org/software/gsl.
- 668 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.
- 672 Goerg, G. M. (2011). Lambert W Random Variables A New Family of Generalized
- 673 Skewed Distributions with Applications to Risk Estimation. The Annals of Applied Statis-
- tics 5(3), p. 2197 2230. arxiv.org/abs/0912.4554.
- 675 Goncalves, S. and N. Meddahi (2011). Box-Cox transforms for realized volatility. Journal
- of Econometrics 160(1), 129-144.
- 677 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.
- 679 Headrick, T. C., R. K. Kowalchuk, 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.
- 682 Hoaglin, D. C. (2006). Summarizing Shape Numerically: The g-and-h Distributions, pp.
- 461–513. Hoboken, NJ, USA: John Wiley and Sons, Inc.
- 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.
- 686 Hwang, J., S. Lay, and A. Lippman (1994). Nonparametric multivariate density estimation:
- A comparative study. IEEE Trans. Signal Processing 42, 2795–2810.
- 688 Ilow, J. (2000). Forecasting network traffic using farima models with heavy tailed inno-
- vations. Acoustics, Speech, and Signal Processing, IEEE International Conference on 6,
- 690 3814–3817.
- 691 Jodrá, P. (2009). A closed-form expression for the quantile function of the Gompertz-
- Makeham distribution. Math. Comput. Simul. 79, 3069–3075.
- 693 Kim, T.-H. and H. White (2003). On More Robust Estimation of Skewness and Kurtosis:
- Simulation and Application to the S&P500 Index.

- Lawrance, A. J. (1987). A note on the variance of the Box-Cox regression transformation
- estimate. Journal of the Royal Statistical Society. Series C (Applied Statistics) 36(2),
- 221-223.
- Lehmann, E. L. and G. Casella (1998). Theory of Point Estimation (2 ed.). Springer Texts
- in Statistics.
- Liu, C. and D. B. Rubin (1995). ML Estimation of the t distribution using EM and its
   extensions, ECM and ECME. Statistica Sinica 5, 19–39.
- Liu, H., J. Lafferty, and L. Wasserman (2009). The Nonparanormal: Semiparametric Es-
- timation of High Dimensional Undirected Graphs. Journal of Machine Learning Re-
- search 10, 2295–2328.
- Maiboroda, R. and N. Markovich (2004). Estimation of heavy-tailed probability den-
- sity 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). 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
- 718 Physics 46, 323–351.
- 719 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). Lambert's W, infinite divisibility and Poisson mixtures. Journal of 721 Mathematical Analysis and Applications 378, 480492. 722
- Palma, W. and M. Zevallos (2011). Fitting non-gaussian persistent data. Applied Stochastic 723 Models in Business and Industry 27(1), 23–36. 724
- R Development Core Team (2010). R: A Language and Environment for Statistical Com-725 puting. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. 726
- Rathie, R. N. and P. Silva (2011). Applications of Lambert W Function. *International* Journal of Applied Mathematics & Statistics 23(D11), -. 728
- Rayner, G. D. and H. L. MacGillivray (2002). Numerical maximum likelihood estimation for 729 the g-and-k and generalized g-and-h distributions. Statistics and Computing 12, 57–75. 730
- Rosenlicht, M. (1969). On the explicit solvability of certain transcendental equations. Pub. 731 Math. Institut des Hautes Etudes Scientifiques 36, 15 – 22. 732
- Sakia, R. M. (1992). The Box-Cox transformation technique: A review. Journal of the 733 Royal Statistical Society. Series D (The Statistician) 41(2), 169–178. 734
- Smith, V. K. (1973). Least Squares Regression with Cauchy Errors. Oxford Bulletin of 735 Economics and Statistics 35(3), 223–31. 736
- Thode, Jr., H. C. (2002). Testing for Normality. CRC Press.

743

- Tsiotas, G. (2007). On the use of the Box-Cox transformation on conditional variance 738 models. Finance Research Letters 4(1), 28–32. 739
- Valluri, S. R., D. J. Jeffrey, and R. M. Corless (2000). Some Applications of the Lambert 740 W Function to Physics. Canadian Journal of Physics 78, 823 – 831. 741
- Vázquez, A., J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási (2006). 742 Modeling bursts and heavy tails in human dynamics. Phys. Rev. E 73(3), 036127.
- Wasserman, L. (2007). All of Nonparametric Statistics. Springer Texts in Statistics.

- 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. 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. ideas.repec.
  org/p/ces/ceswps/\_1505.html.

# <sup>755</sup> A Auxiliary Results and Properties

# 756 A.1 Inverse Transformation $W_{\delta}(z)$

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  $\delta = 0$ ,

$$W_{\delta}(z_i) \mid_{\delta=0} = z_i, \quad W'(\delta z_i^2) \mid_{\delta=0} = z_i^2, \quad and \ W(\delta z_i^2) \mid_{\delta=0} = 0.$$
 (38)

By definition  $\frac{W_{\delta}(z)}{z} = e^{-\frac{\delta}{2}W_{\delta}(z)^2}$  and therefore

$$\log \frac{W_{\delta}(z)}{z} = -\frac{\delta}{2}W_{\delta}(z)^2 = -\frac{W(\delta z^2)}{2}.$$
(39)

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

$$\frac{d}{dz}W_{\delta}(z) = -\frac{W_{\delta}(z)}{z\left(1 + \delta W_{\delta}(z)^{2}\right)} = e^{-\frac{1}{2}W(\delta z^{2})} \frac{1}{1 + W(\delta z^{2})}$$
(40)

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

$$W'(z) = \frac{W(z)}{z(1+W(z))} = \frac{1}{e^{W(z)}(1+W(z))}, \quad z \neq 0, -1/e.$$
(41)

759 Hence,

$$\frac{d}{dz}\frac{W\left(\delta z^{2}\right)}{\delta} = W'\left(\delta z^{2}\right) \cdot 2z = \frac{W\left(\delta z^{2}\right)}{\delta z^{2}\left(1 + W\left(\delta z^{2}\right)\right)} \cdot 2z = \frac{2W\left(\delta z^{2}\right)}{\delta z\left(1 + W\left(\delta z^{2}\right)\right)}$$
(42)

Therefore, 760

$$\frac{d}{dz}W_{\delta}(z) = \frac{1}{2} \left(\frac{1}{\delta}W\left(\delta z^{2}\right)\right)^{-1/2} \cdot \frac{d}{dz}\frac{W\left(\delta z^{2}\right)}{\delta}$$
(43)

$$= \frac{1}{2} \left( \frac{1}{\delta} W \left( \delta z^2 \right) \right)^{-1/2} \cdot \frac{2W \left( \delta z^2 \right)}{\delta z \left( 1 + W \left( \delta z^2 \right) \right)}$$
(44)

$$= \frac{1}{\delta^{1/2}} \left( W \left( \delta z^2 \right) \right)^{-1/2} \cdot \frac{W \left( \delta z^2 \right)}{z \left( 1 + W \left( \delta z^2 \right) \right)} \tag{45}$$

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

$$\frac{1}{\delta^{1/2}} \frac{1}{\delta^{1/2} u} \cdot \frac{\delta u^2}{z (1 + \delta u^2)} = \frac{u}{z (1 + \delta u^2)}.$$
 (46)

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

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

$$\frac{\partial}{\partial \delta} \left[ W_{\delta}(z) \right]^2 = -\frac{1}{1 + W(\delta z^2)} W_{\delta}(z)^4 \le 0. \tag{47}$$

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

$$\frac{\partial}{\partial \delta} \frac{W\left(\delta z^{2}\right)}{\delta} = \frac{\delta \frac{\partial}{\partial \delta} W\left(\delta z^{2}\right) - W\left(\delta z^{2}\right) \cdot 1}{\delta^{2}} \tag{48}$$

$$=\frac{\delta W'\left(\delta z^2\right)z^2-W\left(\delta z^2\right)}{\delta^2}\tag{49}$$

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

$$(50)$$

$$=\frac{\frac{W(\delta z^2)}{1+W(\delta z^2)} - W(\delta z^2)}{\delta^2}$$
(51)

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

$$= \frac{\frac{-W(\delta z^2)^2}{1+W(\delta z^2)}}{\delta^2}$$
(51)

$$= -\frac{1}{1 + W(\delta z^2)} [W_{\delta}(z)]^4.$$
 (53)

Since both terms are non-negative for all  $z \in \mathbb{R}$ , the result follows. 762

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 763 tails the more z gets shrinked (non-linearly) towards  $0 = \lim_{\delta \to \infty} W_{\delta}(z)$ . In particular, 764

 $[W_{\delta}(z)]^2 < z^2 \Leftrightarrow \frac{W_{\delta}(z)}{z} < 1 \text{ and } \frac{W_{\delta+\varepsilon}(z)}{z} < \frac{W_{\delta}(z)}{z} \text{ for } \delta \geq 0 \text{ and } \varepsilon > 0.$ 

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

$$\frac{\partial}{\partial \delta} W_{\delta}(z) = -\frac{1}{2} \frac{1}{1 + W(\delta z^2)} W_{\delta}(z)^3 \tag{54}$$

Proof.

$$\frac{\partial}{\partial \delta} W_{\delta}(z) = \operatorname{sgn}(z) \frac{\partial}{\partial \delta} \left( \frac{W(\delta z^2)}{\delta} \right)^{1/2}$$
(55)

$$= \operatorname{sgn}(z) \frac{1}{2} \left( \frac{W(\delta z^2)}{\delta} \right)^{-1/2} \frac{\partial}{\partial \delta} \frac{W(\delta z^2)}{\delta}$$
 (56)

$$= \frac{1}{2} \frac{1}{W_{\delta}(z)} \frac{\partial}{\partial \delta} \left[ W_{\delta}(z) \right]^2 \tag{57}$$

$$= -\frac{1}{2} \frac{1}{1 + W(\delta z^2)} W_{\delta}(z)^3, \qquad (58)$$

where the last line follows by Lemma A.3.

# 767 A.2 Penalty $\log R\left(\delta \mid z_i\right)$ for Standard Gaussian Input

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

$$R\left(\delta \mid z_{i}\right) = \frac{W_{\delta}\left(z_{i}\right)}{z_{i}\left[1 + \delta\left(W_{\delta}\left(z_{i}\right)\right)^{2}\right]} = \frac{W_{\delta}\left(z_{i}\right)}{z_{i}\left[1 + W\left(\delta z_{i}^{2}\right)\right]}$$
(59)

and thus

$$\log R\left(\delta \mid z_{i}\right) = \log \frac{W_{\delta}\left(z_{i}\right)}{z_{i}} - \log\left[1 + W\left(\delta z_{i}^{2}\right)\right]$$

$$(60)$$

$$= -\frac{W(\delta z_i^2)}{2} - \log\left[1 + W\left(\delta z_i^2\right)\right] \tag{61}$$

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

$$\frac{\partial \log R\left(\delta \mid z\right)}{\partial \delta} = -z^2 W'(\delta z^2) \left(\frac{1}{2} + \frac{1}{1 + W\left(\delta z^2\right)}\right) \le 0.$$
 (62)

Proof. We have

$$\frac{\partial \log R(\delta \mid z)}{\partial \delta} = \frac{1}{W_{\delta}(z)} \frac{\partial W_{\delta}(z)}{\partial \delta} - \frac{1}{1 + W(\delta z^2)} W'(\delta z^2) z^2$$
(63)

$$\stackrel{\text{Lemma A.4}}{=} \frac{1}{W_{\delta}(z)} \left( -\frac{1}{2} \frac{1}{1 + W(\delta z^2)} W_{\delta}(z)^3 \right) - \frac{1}{1 + W(\delta z^2)} W'(\delta z^2) z^2 \quad (64)$$

$$= -\frac{1}{1 + W(\delta z^2)} \left( \frac{1}{2} W_{\delta}(z)^2 + W'(\delta z^2) z^2 \right)$$
 (65)

Using  $W'(\delta z^2) = \frac{W(\delta z^2)}{\delta z^2 (1 + W(\delta z^2))}$  and re-factorizing gives (62).

769

# $_{70}$ A.3 Gaussian log-Likelihood at $W_{\delta}(z)$

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

$$\frac{\partial}{\partial \delta} \ell(\mu_x = 0, \sigma_x = 1 \mid W_{\delta}(z)) = \frac{1}{2} \frac{1}{1 + W(\delta z^2)} [W_{\delta}(z)]^4 \ge 0.$$
 (66)

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

$$\log \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[W_{\delta}(z)]^2} = \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2} [W_{\delta}(z)]^2.$$
 (67)

The rest follows by Lemma A.3.

Lemma A.6 shows that increasing  $\delta$  always increases the input log-likelihood  $\ell(\delta \mid \mathbf{u}_{\delta} = W_{\delta}(\mathbf{z}))$  - see also Fig. 6b. For  $\delta \to \infty$  the Gaussianized  $\mathbf{u}_{\delta}$  goes to  $\mathbf{0}$ , which clearly maximizes the Gaussian likelihood if  $\mu = 0$ .

## 775 B Proofs

### 776 B.1 Inverse transformation

Proof of Lemma 2.5. Without loss of generality assume that  $\mu_x = 0$  and  $\sigma_x = 1$ . Squaring (2) and multiplying by  $\delta$  yields

$$\delta Z^2 = \delta U^2 \exp\left(\delta U^2\right) \tag{68}$$

The inverse of (68) is by definition Lambert's W(z) function (Rosenlicht, 1969)

$$W(z) \exp W(z) = z, \quad z \in \mathbb{C}.$$
 (69)

W(z) is bijective for  $z \ge 0$ . Since  $\delta U^2 \ge 0$  for all  $\delta \ge 0$ , applying  $W(\cdot)$  to (68), dividing by  $\delta$ , and taking the square root gives

$$U = \pm \sqrt{\frac{W(\delta Z^2)}{\delta}} \tag{70}$$

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

### $^{783}$ B.2 Cdf and pdf

Proof of Theorem 2.8. By definition,

$$G_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}\left(\left\{U \exp\left(\frac{\delta}{2}U^2\right)\right\}\sigma_x + \mu_x \le y\right)$$
 (71)

$$= \mathbb{P}\left(U \exp\left(\frac{\delta}{2}U^2\right) \le z\right) = \mathbb{P}\left(U \le W_{\delta}(z)\right) \tag{72}$$

$$= F_U(U \le W_\delta(z)). \tag{73}$$

Taking the derivative with respect to y gives

$$\frac{d}{dy}G_Y(y \mid \boldsymbol{\beta}, \delta) = f_X(W_{\delta}(z)\sigma_x + \mu_x \mid \boldsymbol{\beta}) \cdot \sigma_x \frac{d}{dy}W_{\delta}\left(\frac{y - \mu_x}{\sigma_x}\right)$$
(74)

$$= f_U(W_{\delta}(z) \mid \boldsymbol{\beta}) \cdot \sigma_x \frac{1}{\sigma_x} \frac{d}{dz} W_{\delta} \left( \frac{y - \mu_x}{\sigma_x} \right)$$
 (75)

$$= f_U(W_{\delta}(z) \mid \boldsymbol{\beta}) \cdot \frac{d}{dz} W_{\delta}(z). \tag{76}$$

Using Lemma A.2 yields (15).

### 786 B.3 MLE for $\delta$

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

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

$$= \frac{1}{2} \sum_{i=1}^{N} \frac{W_{\delta}(z_i)^4}{1 + \delta W_{\delta}(z_i)^2} - \sum_{i=1}^{N} \frac{W_{\delta}(z_i)^2}{1 + \delta W_{\delta}(z_i)^2} \left( \frac{1}{2} + \frac{1}{1 + \delta W_{\delta}(z_i)^2} \right)$$
(78)

$$= \frac{1}{2} \sum_{i=1}^{N} \frac{W_{\delta}(z_{i})^{4}}{1 + W(\delta z_{i}^{2})} - \sum_{i=1}^{N} \frac{W_{\delta}(z_{i})^{2}}{1 + W(\delta z_{i}^{2})} \left(\frac{1}{2} + \frac{1}{1 + W(\delta z_{i}^{2})}\right). \quad (79)$$

787 Proof. Apply Lemmas A.5 and A.6 to  $\frac{\partial}{\partial \delta} \ell(\delta \mid \mathbf{z}) = \frac{\partial}{\partial \delta} \log R(\delta \mid z) + \frac{\partial}{\partial \delta} \ell(\mu_x = 0, \sigma_x = 1 \mid W_{\delta}(z)).$ 

Proof sketch of Theorem 4.1. a) If condition (35) holds, then  $D(\delta \mid \mathbf{z}) < 0$  at  $\delta = 0$  and stays negative for all  $\delta > 0$ . Hence the maximizer occurs at the boundary  $\delta = 0$ .

b) If (35) does not hold, then  $D(\delta = 0 \mid \mathbf{z}) > 0$ , decreases in  $\delta$  and crosses the zero line (one candidate for  $\widehat{\delta}_{MLE}$  occurs here).

793 c) As  $\delta$  gets larger,  $D(\delta \mid \mathbf{z})$  reaches a minimum (negative value) and starts increasing.

794 However, for  $\delta \to \infty$  the derivative approaches zero from below and never equals zero

795 again; thus  $\widehat{\delta}_{MLE}$  is unique.

796

Proof of Theorem 4.1. a) The log-likelihood is increasing at  $\delta = 0$  if and only if (set  $\delta = 0$  in (79) and use Property A.1)

$$\sum_{i=1}^{N} z_i^4 > 3 \sum_{i=1}^{N} z_i^2. \tag{80}$$

Eq. (80) means that transforming the data (choosing  $\hat{\delta} > 0$ ) increases the overall likelihood only if the data is heavy-tailed enough. Note that the sum of squares is not squared again. Hence condition (80) is not equivalent for the data having empirical kurtosis larger than 3.

- b) If (80) does not hold, then  $\widehat{\delta}_{MLE}$  must satisfy  $D(\delta \mid \mathbf{z}) \mid_{\delta = \widehat{\delta}_{MLE}} = 0$  from (77) 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  $z \in \mathbb{R}$ , (79) is also true in the limit; however, we can ignore this solution as we require  $\widehat{\delta}_{MLE} \in \mathbb{R}$ .
  - ii) By continuity and  $\lim_{\delta \to \infty} W_{\delta}(z) = 0$ , 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

$$\frac{1}{2} \sum_{i=1}^{N} \frac{W_{\delta}(z_{i})^{4}}{1 + \delta W_{\delta}(z_{i})^{2}} < \frac{1}{2} \sum_{i=1}^{N} \frac{W_{\delta}(z_{i})^{2}}{1 + \delta W_{\delta}(z_{i})^{2}}$$
(81)

$$<\sum_{i=1}^{N} \frac{W_{\delta}(z_{i})^{2}}{1+\delta W_{\delta}(z_{i})^{2}} \left(\frac{1}{2}+\frac{1}{1+\delta W_{\delta}(z_{i})^{2}}\right) \text{ for } \delta \geq \delta_{M}, \quad (82)$$

- showing that  $D(\delta \mid \mathbf{z}) \mid_{\delta \geq \delta_M} < 0$ . That is,  $D(\delta \mid \mathbf{z})$  approaches 0 from below for  $\delta \to \infty$ .
- 807 iii) By continuity and  $D(\delta \mid \mathbf{z}) \mid_{\delta=0} > 0$  (if (80) does not hold), it must cross the  $D(\delta \mid \mathbf{z}) = 0$  line at least once in the interval  $(0, \delta_M)$ , proving the existence of  $\widehat{\delta}_{MLE}$ .
  - c) The log-likelihood can be decomposed in

812

813

814

815

816

$$\ell\left(\delta \mid \mathbf{z}\right) \propto \underbrace{-\frac{1}{2} \sum_{i=1}^{N} \left[W_{\delta}(z_{i})\right]^{2} + \sum_{i=1}^{N} \log \frac{W_{\delta}\left(z_{i}\right)}{z_{i}} - \log\left[1 + W\left(\delta z_{i}^{2}\right)\right]}_{\mathcal{R}\left(\delta \mid \mathbf{z}\right)}.$$
(83)

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

Furthermore,  $\lim_{\delta\to\infty} \ell(\mu_x = 0, \sigma_x = 1 \mid W_{\delta}(\mathbf{z})) = 0$ , that is the input likelihood is monotonically increasing but bounded from above (by  $0 = \log 1$ ). On the other hand the penalty is decreasing without bounds,  $\lim_{\delta\to\infty} \mathcal{R}(\delta \mid \mathbf{z}) = -\infty$ . Thus their sum attains a global maximum either at the unique mode of  $\ell(\delta \mid \mathbf{z})$  or at the boundary  $\delta = 0$  - see also Fig. 6b.

817

# **Algorithm 1** Find optimal $\delta$ : function delta\_GMM(·) in the LambertW package.

**Input:** standardized data vector  $\mathbf{z}$ ; theoretical kurtosis  $\gamma_2(X)$ 

Output:  $\hat{\delta}_{GMM}$  as in (84)

- 1:  $\widehat{\delta}_{GMM} = \underset{\widehat{\alpha}}{\arg\min_{\delta}} ||\widehat{\gamma}_{2}(\mathbf{u}_{\delta}) \gamma_{2}(X)||$ , where  $\mathbf{u}_{\delta} = W_{\delta}(\mathbf{z})$  subject to  $\delta \geq 0$
- 2: **return**  $\delta_{GMM}$

### 818 C Details on IGMM

Here I present an iterative method to obtain  $\hat{\tau}$ , which builds on the input/output aspect 819 and theoretical properties of the input X. For example, if a random variable should be 820 exponentially distributed (e.g. waiting times), but the observed data shows heavier tails 821 then it is natural to estimate  $\sigma_x = \lambda^{-1}$  and  $\delta$  such that the back-transformed data has 822 skewness 2, as this is a particular property of exponential RVs - independent of the rate 823 parameter  $\lambda$ ; to remove heavy tails in y we should choose  $\tau$  such that the back-transformed 824 data  $\mathbf{x}_{\tau}$  has sample kurtosis 3; or for uniform input, we can try to find a  $\tau$  such that  $\mathbf{x}_{\tau}$ 825 has a flat density estimate. 826

Here I describe the estimator for  $\tau$  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.

830

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  $\mathbf{x}_{\tau}$  ( $\tau = (\mu_x^{(0)}, \sigma_x^{(0)}, \delta)$ ) with sample kurtosis  $\widehat{\gamma}_2(\mathbf{x}_{\tau})$  equal to the theoretical kurtosis  $\gamma_2(X)$ . Formally,

$$\widehat{\delta}_{GMM} = \arg\min_{\delta} ||\gamma_2(X) - \widehat{\gamma}_2(\mathbf{x}_{\tau})||, \qquad (84)$$

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  $\delta$  is continuous and monotonically increasing (see (22)). Also  $u = W_{\delta}(z)$  has a smaller slope than the identity

**Algorithm 2** Iterative Generalized Method of Moments (IGMM): function IGMM(·) in the LambertW package.

Input: data vector  $\mathbf{y}$ ; tolerance level tol; theoretical kurtosis  $\gamma_2(X)$ Output: IGMM parameter estimate  $\widehat{\tau}_{\text{IGMM}} = (\widehat{\mu}_x, \widehat{\sigma}_x, \widehat{\delta})$ 

- 1: Set  $\tau^{(-1)} = (0,0,0)$
- 2: Starting values:  $\tau^{(0)} = (\mu_x^{(0)}, \sigma_x^{(0)}, \delta^{(0)})$ , where  $\mu_x^{(0)} = \tilde{\mathbf{y}}$  and  $\sigma_x^{(0)} = \overline{\sigma}_y \cdot \left(\frac{1}{\sqrt{(1-2\delta^{(0)})^{3/2}}}\right)^{-1}$  are the sample median and standard deviation of  $\mathbf{y}$  divided by the standard deviation factor (see also (21)), respectively.  $\delta^{(0)} = \frac{1}{66} \left(\sqrt{66\hat{\gamma}_2(\mathbf{y}) 162} 6\right) \rightarrow \text{see}$  (??) for details.
- 3: k = 0
- 4: while  $||\tau^{(k)} \tau^{(k-1)}|| > tol \ do$
- 5:  $\mathbf{z}^{(k)} = (\mathbf{y} \mu_x^{(k)}) / \sigma_x^{(k)}$
- 6: Pass  $\mathbf{z}^{(k)}$  to Algorithm  $\mathbf{1} \longrightarrow \delta^{(k+1)}$
- 7: back-transform  $\mathbf{z}^{(k)}$  to  $\mathbf{u}^{(k+1)} = W_{\delta^{(k+1)}}(\mathbf{z}^{(k)})$ ; compute  $\mathbf{x}^{(k+1)} = \mathbf{u}^{(k+1)} \sigma_x^{(k)} + \mu_x^{(k)}$
- 8: Update parameters:  $\mu_x^{(k+1)} = \overline{\mathbf{x}}_{k+1}$  and  $\sigma_x^{(k+1)} = \widehat{\sigma}_{x_{k+1}}$
- 9:  $\tau^{(k+1)} = (\mu_x^{(k+1)}, \sigma_x^{(k+1)}, \delta^{(k+1)})$
- 10: k = k + 1
- 11: **return**  $\tau_{IGMM} = \tau^{(k)}$

u=z, and the slope is decreasing as  $\delta$  is increasing. Thus if the kurtosis of the original data is larger than the target kurtosis of the back-transformed data,  $\hat{\gamma}_2(\mathbf{y}) > \gamma_2(X)$ , then there always exists a  $\delta^{(*)}$  that achieves  $\hat{\gamma}_2(\mathbf{x}_{\tau^*}) \equiv \gamma_2(X)$ . By the re-parametrization  $\tilde{\delta} = \log \delta$  the bounded optimization problem can be solved by standard (unbounded) optimization algorithms.

In practice,  $\mu_x$  and  $\sigma_x$  are rarely known but also have to be estimated from the data.

As  ${f y}$  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

846 obtained iteratively.

842

Discussion of Algorithm 2: Algorithm 2 first computes  $\mathbf{z}^{(k)} = (\mathbf{y} - \mu_x^{(k)})/\sigma_x^{(k)}$  using  $\mu_x^{(k)}$  and  $\sigma_x^{(k)}$  from the previous step. This normalized output can then be passed to Algorithm

1 to obtain an updated  $\delta^{(k+1)} = \hat{\delta}_{GMM}$ . Using this new  $\delta^{(k+1)}$  one can back-transform  $\mathbf{z}^{(k)}$ to  $\mathbf{u}^{(k+1)} = W_{\delta^{(k+1)}}(\mathbf{z}^{(k)})$ , and consequently obtain a better approximation to the "true"

latent  $\mathbf{x}$  by  $\mathbf{x}^{(k+1)} = \mathbf{u}^{(k+1)}\sigma_x^{(k)} + \mu_x^{(k)}$ . However,  $\delta^{(k+1)}$  - and therefore  $\mathbf{x}^{(k+1)}$  - has been

obtained using  $\mu_x^{(k)}$  and  $\sigma_x^{(k)}$ , which are not necessarily the most accurate estimates in light of the updated approximation  $\hat{\mathbf{x}}_{(\mu_x^{(k)}, \sigma_x^{(k)}, \delta^{(k+1)})}$ . Thus Algorithm 2 computes new estimates  $\mu_x^{(k+1)}$  and  $\sigma_x^{(k+1)}$  by the sample mean and standard deviation of  $\hat{\mathbf{x}}_{(\mu_x^{(k)}, \sigma_x^{(k)}, \delta^{(k+1)})}$ , and starts another iteration by passing the updated normalized output  $\mathbf{z}^{(k+1)} = \frac{\mathbf{y} - \mu_x^{(k+1)}}{\sigma_x^{(k+1)}}$  to Algorithm 1 to obtain a new  $\delta^{(k+2)}$ .

It returns the optimal  $\hat{\tau}_{\text{IGMM}}$  once convergence has been reached, i.e., if  $||\tau^{(k)} - \tau^{(k+1)}|| < tolerapse tol.$ 

859

**Remark C.1** (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

$$\left(\widehat{\delta}_{\ell}, \widehat{\delta}_{r}\right)_{\text{GMM}} = \arg\min_{\delta_{\ell}, \delta_{r}} h\left(\gamma_{2}(X) - \widehat{\gamma}_{2}(\mathbf{x}_{(\mu_{x}^{*}, \sigma_{x}^{*}, \delta_{\ell}, \delta_{r})})\right). \tag{85}$$

60 Algorithm 2 remains unchanged.

Algorithm 3 Random sample generation: function rLambertW(·) in LambertW package.

Input: number of observations n; parameter vector  $\theta$ ; specification of the input distribution  $F_X(x)$ 

**Output:** random sample  $(y_1, \ldots, y_n)$  of a Lambert W  $\times$   $F_X$  RV.

- 1: Simulate n samples  $\mathbf{x} = (x_1, \dots, x_n) \sim F_X(x)$ .
- 2: Compute  $\mu_x = \mu_x(\boldsymbol{\beta})$  and  $\sigma_x = \sigma_x(\boldsymbol{\beta})$  (for scale family set  $\mu_x = 0$ , for non-central, non-scaled also set  $\sigma_x = 1$ )
- 3: Compute normalized  $\mathbf{u} = (\mathbf{x} \mu_x)/\sigma_x$ .
- 4:  $\mathbf{z} = \mathbf{u} \exp\left(\frac{\delta}{2}\mathbf{u}^2\right)$
- 5: **return**  $\mathbf{y} = \mathbf{z}\sigma_x + \mu_x$

# Ball D Simulation Details

Slightly heavy-tailed:  $\delta = 1/10$ . Here the RV Y has slight excess kurtosis (3 + 2.51) and  $\sigma_y(\delta, \sigma_x = 1) = 1.18$ . The Lambert W estimates of  $\hat{\tau}$  are unbiased, and have smaller empirical standard deviation for  $\hat{\mu}_x$  than the Gaussian MLE or the median. Also using Lambert W estimators does not give worse estimates for  $\sigma_y$ .

$\delta = 1/10$	median	Gaussia	n MLE	IGMM				Lambert W MLE				NA
N		$\mu_y$	$\sigma_y$	$\mu_x$	$\sigma_x$	δ	$\sigma_y$	$\mu_x$	$\sigma_x$	δ	$\sigma_y$	ratio
50	-0.02	-0.02	1.15	-0.02	1.02	0.08	1.18	-0.02	0.99	0.09	$\infty$	0
100	0.00	0.00	1.17	0.00	1.02	0.09	1.18	0.00	1.00	0.09	1.18	0
250	0.00	0.00	1.18	0.00	1.01	0.09	1.18	0.00	1.00	0.10	1.18	0
1000	0.00	0.00	1.18	0.00	1.00	0.10	1.18	0.00	1.00	0.10	1.18	0
50	0.56	0.53	0.61	0.54	0.48	0.64	0.55	0.53	0.55	0.58	0.56	0
100	0.50	0.49	0.57	0.50	0.45	0.61	0.54	0.49	0.51	0.56	0.54	0
250	0.50	0.48	0.56	0.47	0.46	0.56	0.53	0.48	0.51	0.54	0.53	0
1000	0.48	0.49	0.53	0.48	0.50	0.54	0.51	0.48	0.52	0.51	0.52	0
50	1.27	1.22	1.13	1.18	1.03	0.52	1.27	1.16	1.07	0.62	NA	0
100	1.28	1.19	1.21	1.15	1.07	0.60	1.26	1.12	1.09	0.64	1.28	0
250	1.26	1.19	1.20	1.12	1.09	0.63	1.22	1.09	1.09	0.65	1.23	0
1000	1.23	1.17	1.26	1.11	1.14	0.66	1.26	1.08	1.11	0.63	1.23	0

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

Table 5: Based on n = 1,000 replications. In each sub-table: (first rows) average, (middle rows) proportion of estimates below true value, (bottom rows) empirical standard deviation times  $\sqrt{N}$ .