CDFquantile Supplementary Materials

Michael Smithson and Yiyun Shou

11 August 2016

Contents

Introduction

The materials in this document supplement the paper titled "CDF-Quantile Distributions for Modeling Random Variables on the Unit Interval". The supplementary materials include sections on skew-kurtosis maps for distributions in the CDF-Quantile family, simulation studies investigating the sampling distributions of the location and dispersion parameters, the sensitivity of selected quantile estimates to outliers, and the R code from the package cdfquantreg for two examples involving real data-sets.

Skew-Kurtosis Maps

We investigated two kinds of skewness and kurtosis measures, both of which take advantage of an explicit expression for the quantile. First, we employed the classical quantile-based measures of skewness (Bowley, 1920) and kurtosis (Moors, 1988). Second, we also used the L-moment measures, which may be computed via linear combinations of probability-weighted moments (Greenwood, 1979) and provide a check on the findings from the classical measures.

The CDF-Quantile family is well-behaved with regard to skew and kurtosis as functions of its parameters. The upper part of Figure 1 shows typical graphs of Bowley skew as a function of μ and σ, demonstrating the two shapes produced by this subfamily. For distributions producing Shape 1, skew is entirely influenced by μ unless it is fairly far from 0 or when σ is large, in which case it quickly becomes entirely influenced by σ. In Shape 2, on the other hand, when μ shifts sufficiently far from 0 the influence of σ increases but does not take over. The lower part of Figure 1 shows typical graphs of Moors kurtosis as a function of μ and σ. These show a consistent joint influence of μ and σ on kurtosis.

Figure 2 shows examples of skew-kurtosis (S-K) maps, using Bowley skew and Moors kurtosis (L-moment measures gave qualitatively similar results). The characteristic shapes correspond to the four categories of

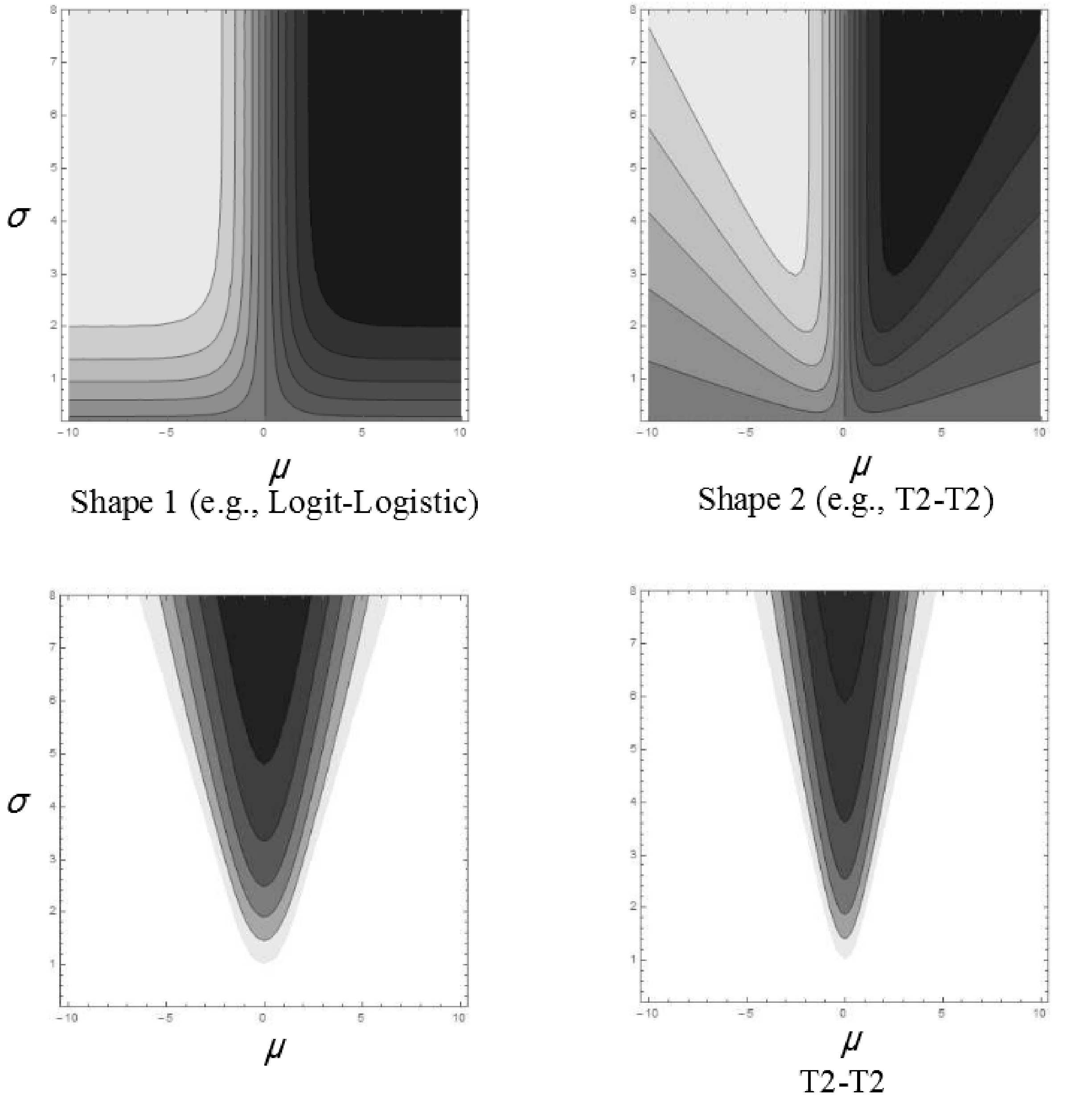

Figure 1: Figure 1. Skew (upper panels) and Kurtosis (lower panels) as functions of μ and σ, for the Logit-Logistic and T2-T2 distributions, with $\mu \in [-10, 10]$ and $\sigma \in [0.2, 8.0]$. Darker shades in the skew plots indicate greater positive skew, and in the kurtosis plots they indicate more leptokurtic distributions.

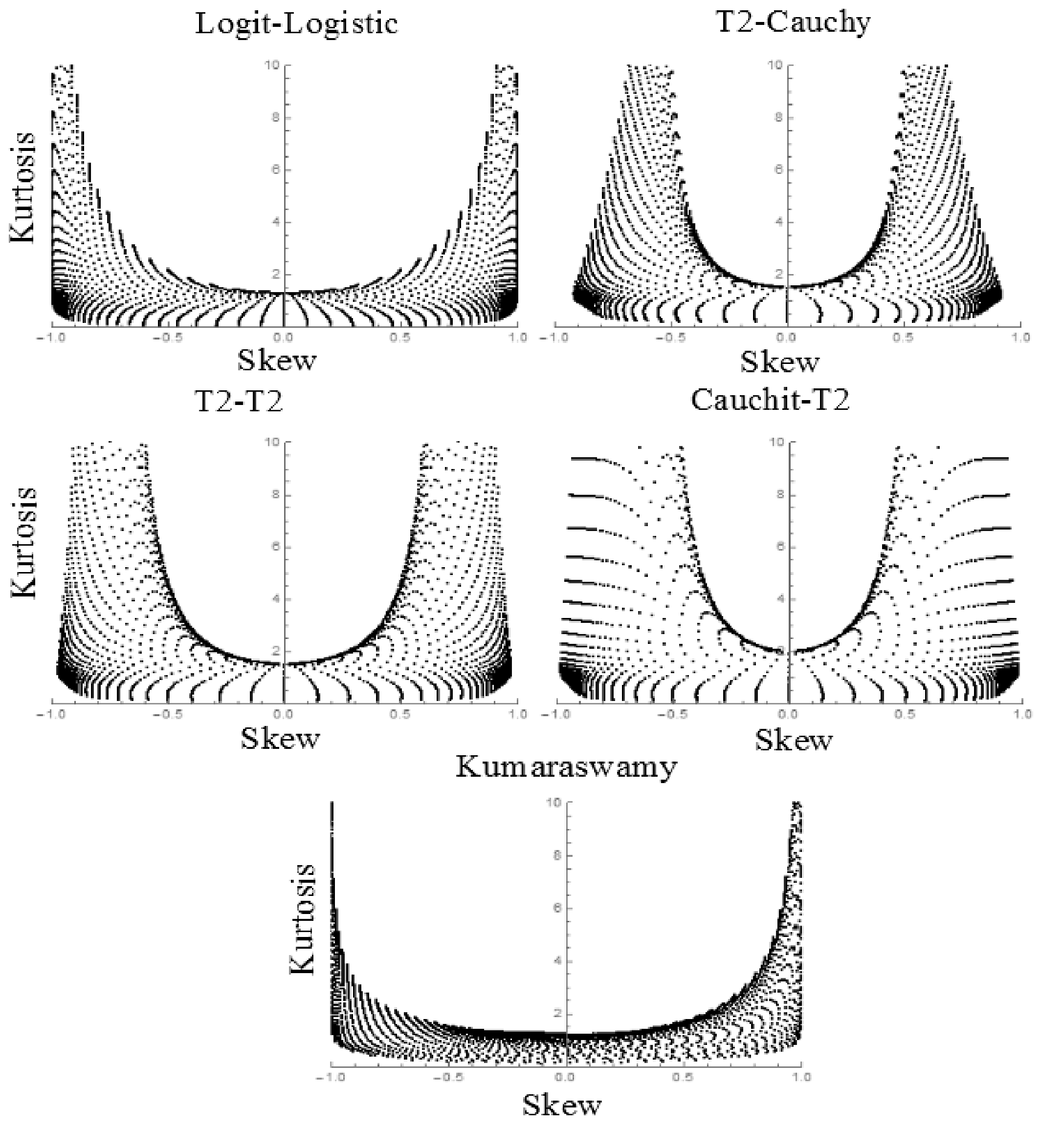

Figure 2: Figure 2. Skew-kurtosis maps for the Logit-Logistic, T2-Cauchy, T2-T2, Cauchit-T2, and Kumaraswamy distributions, with $\mu \in [-10, 10]$ and $\sigma \in [0.2, 8.0]$.

distribution shape and tail behavior described in the paper. Thus, the LL-shaped distributions produce the shape of skew-kurtosis map illustrated by the logit-logistic distribution in Figure 2; the bimodal (BM) and finite-tailed (FT) distributions exhibit the horn-shaped skew-kurtosis maps illustrated by the T2-Cauchy and T2-T2 distributions; and the trimodal (TM) distributions produce the skew-kurtosis map shown for the Cauchit-T2 distribution.

All Burr VII-X S-K maps are covered by corresponding Logit-X maps, thereby favouring the Logit over the Burr VII transform in this regard. Likewise, all Cauchit-X maps are covered by corresponding ArcSinh-X maps, although the difference between them is small. On the other hand, owing to the relationship between quantile-dual distributions described by Assertion 4, the X-Logistic maps are covered by corresponding X-Burr VII maps, and the X-ArcSinh maps are covered by corresponding X-Cauchy maps.

Maximum Likelihood Estimator Computation and Distributions

The cdfquantreg package produces MLEs in a two-stage estimation process. Users may provide their own starting values. The default starting values in the package are the sample values of μ and σ derived from the marginal distribution, and 0 for all predictor coefficients in the model. The first stage utilizes the Nelder-Mead estimator as implemented in a current version of R. The parameter estimates from that stage are then used in the second stage as starting-values for estimation by the BFGS method, also as implemented in the current version of R. The BFGS method is appropriate because the likelihood functions for the distributions in our package all are twice-differentiable, and the explicit formulas for all gradients are supplied by our code. The standard errors are estimated using the sample Fisher's information matrix, which in turn is derived from the Hessian.

As mentioned in the paper, under mild regularity conditions the joint distribution of the MLEs is multivariate normal. However, this behavior is not guaranteed for small samples. In a simulation study, we investigated the behavior of the sampling distributions for the location and dispersion parameter MLEs for finite samples. The beta distribution was utilized for comparisons with members of the CFD-Quantile family.

There were three sample size conditions: 20, 100 and 200. For each sample size condition, 10000 simulation samples were generated from the 16 distributions (including 15 CDF-Quantile distributions and the beta distribution) with $\mu = 0$ and $\sigma = 1.5$. The 16 distributions were then used to fit to the simulation samples that were generated from their own distribuions (e.g., the beta distribution was fitted to the data generated from the beta distribution).

The results for $N = 20$ reveal some differences among the CDF-Quantile distributions. Figures 3 and 4 display quantile-quantile plots for these simulations. The Arcsinh-X distributions exhibit a slight tendency to be platykurtic for the location parameter and several distributions show a mild tendency to be leptokurtic for the dispersion parameter. For sample sizes of 100 and 200, all of the distributions' estimators' sampling distributions are well-behaved and closely approximate normal distributions.

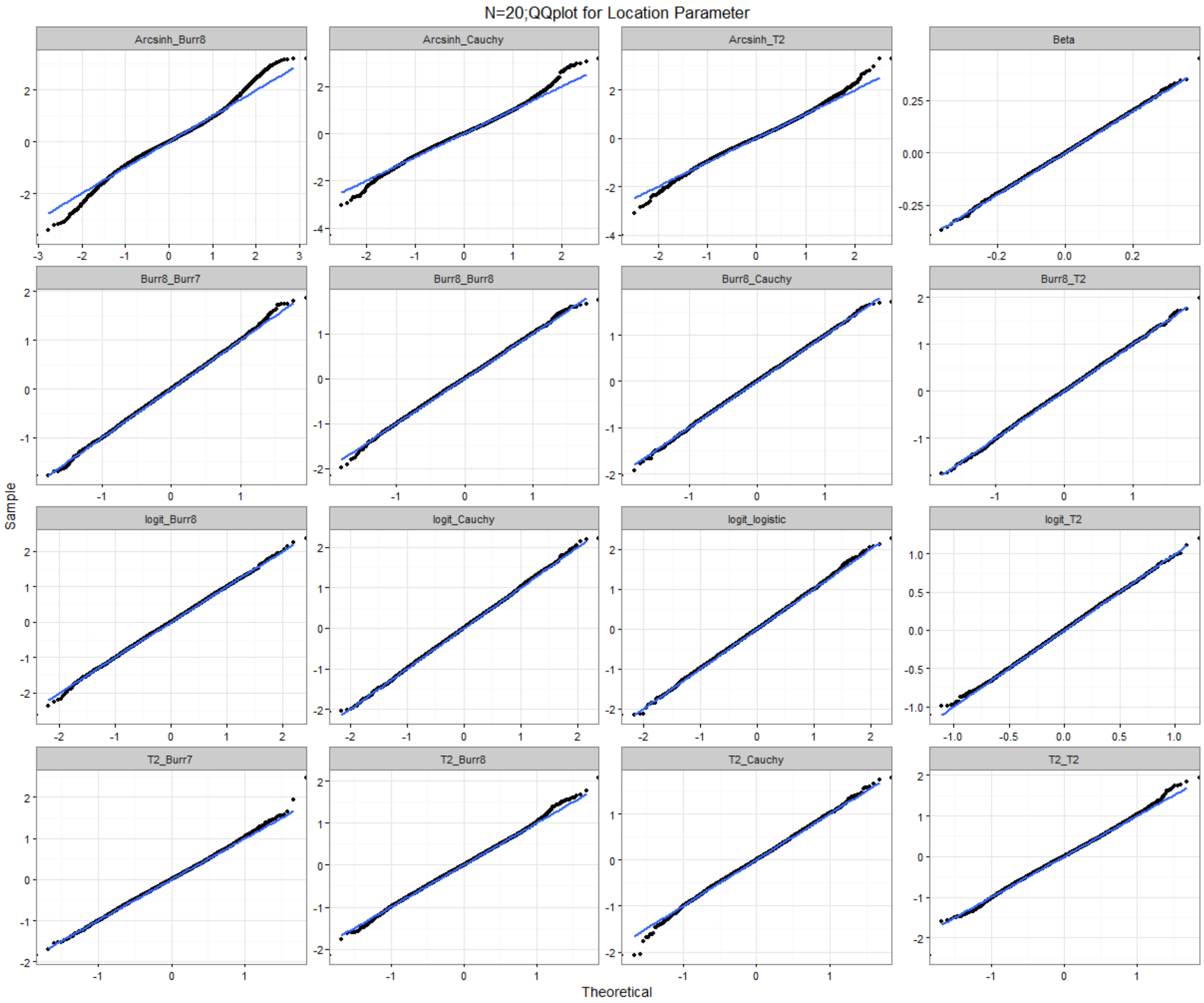

Figure 3. Quantile-quantile plots for location parameter estimations for sample size = 20

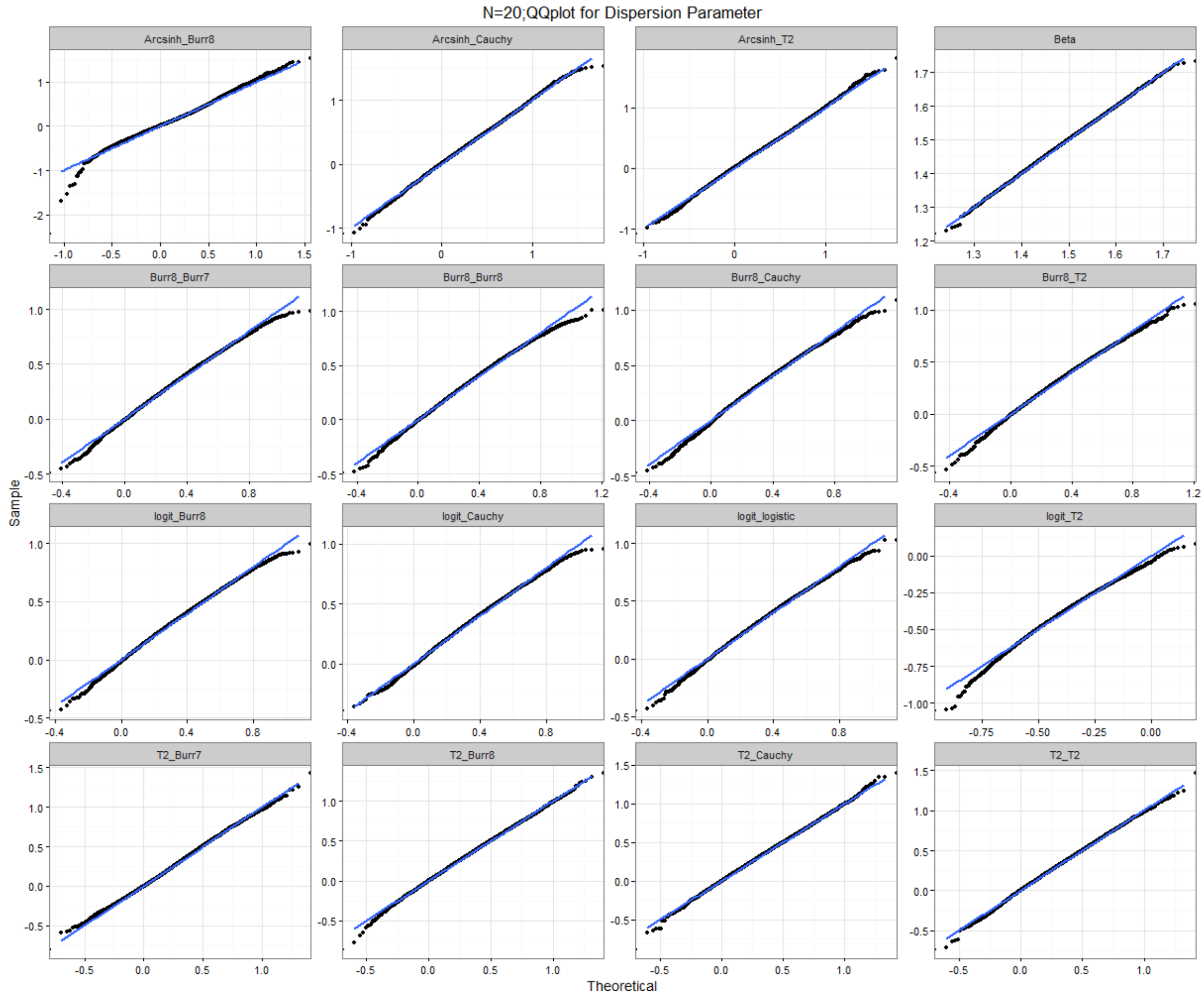

Figure 4. Quantile-quantile plots for dispersion parameter estimations for sample size = 20

Outlier Influence on Estimation

Outliers (near 0 or 1) can present difficulties for estimation when working with random variables on the (0,1) interval. We investigated outlier influence for representative members of the CDF-Quantile family in comparison with the beta distribution. The CDF-Quantile distributions consisted of one from each of the four shape categories: Logit-Logistic (LL), T2-T2 (FT), T2-Cauchy (BM), and T2-Burr-VIII (TM). The data for the simulations were random samples of a logit-transformed random variable distributed $N(0, 1.5)$. The sample sizes employed were 40 and 100. Each simulation had 10,000 samples.

Three outlier values were examined: .01, .001, and .0001 (.99, .999, and .9999). For every sample in a simulation, the appropriate outliers were substituted for the lowest (highest) cases in the sample. The goal was to examine the impact of these outliers on the sample 25th, 50th, and 75th percentiles. The 50th percentile was chosen because it is solely a function of the location parameter μ and therefore would provide an indication of outlier influence on the location parameter estimate. The other two percentiles were chosen because they are functions of both the location and dispersion parameters (μ and σ), thereby providing indications of outlier influence on the dispersion parameter estimate.

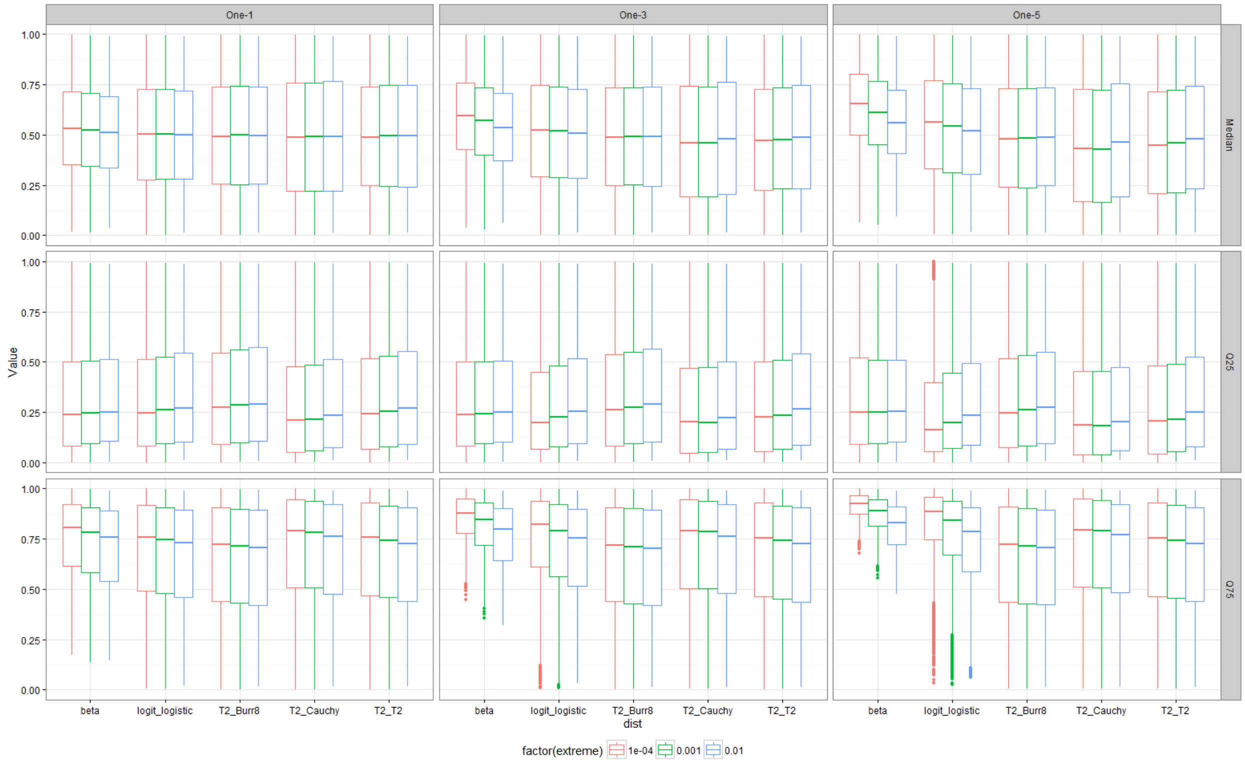

Figure 5. Boxplots of quantiles (median, 25%, and 75%; row panels) given different levels extremity (boxplot colours) at one, and different numbers (1, 3, 5) of extreme values (column panels) for sample size = 40.

As Figure 5 illustrates, the quantile estimates by the CDF-Quantile distribution models are less strongly biased by outliers than the beta distribution. The logit-logistic model is moderately biased by the outliers, but the other CDF-Quantile family members' quantile estimates are relatively stable. These differences in outlier influence are smaller but still detectable for samples of 100. The diferences also are borne out in simulation results for estimates of the location and dispersion parameters themselves (not shown here). That said, it also is the case that the sampling variability of the quantile and parameter estimates is greater for the CDF-Quantile distributions than for the beta. Nevertheless, the evidence from the simulations indicates that the CDF-Quantile distributions are better-suited than the beta for modeling quantiles.

Examples

Example 1: Lay Interpretations of IPCC Report Probability Phrases

We provide here an example of the procedures used to fit CDF-Quantile distributions to the IPCC data, namely the T2-T2 distribution fitted to the data for the questions which differed from one another by the use of "likely" versus "unlikely" in the sentence presented to participants. Thus, we will reproduce the results for the T2-T2 distribution displayed in Table 3 in the paper.

```
data <- subset(IPCC, mid == 1 & high == 0)
fit <- cdfquantreg(probm ~ valence | valence, fd = "t2", sd = "t2", data = data)
summary(fit)
```

```
## Family:  t2 t2
## Call:  cdfquantreg(formula = probm ~ valence | valence, data = data,
##     fd = "t2", sd = "t2")
```

```
##
## Mu coefficients (Location submodel)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.79843    0.03436  23.240  < 2e-16 ***
## valence     -0.18600    0.04120  -4.514 6.35e-06 ***
##
## Sigma coefficients (Dispersion submodel)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.36789    0.04500  -8.176 2.22e-16 ***
## valence     -0.42067    0.06228  -6.755 1.43e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Converge:  0
## Log-Likelihood:  435.2941
##
## Gradient:  -0.0114 -0.01 0.0294 0.0203
```

The output shows that valence (“likely” vs “unlikely”) has had an effect on both μ and σ. To find the relevant quantile estimates from these parameter estimates, we need the T2-T2 quantile function:

$$Q\left(\gamma,\mu,\sigma\right)=\frac{\sigma\left(\frac{\mu}{\sigma}+w(\gamma)\right)}{2\sqrt{\left(\sigma\left(\frac{\mu}{\sigma}+w(\gamma)\right)\right)^2+2}}+\frac{1}{2},$$

where

$$w\left(x\right)=-\frac{\sqrt{(1-2x)^2}}{\sqrt{2}\sqrt{(1-x)x}},0\leq x<\frac{1}{2};$$

and

$$w\left(x\right)=\frac{\sqrt{(1-2x)^2}}{\sqrt{2}\sqrt{(1-x)x}},\frac{1}{2}\leq x\leq 1.$$

The median, 25^{th}, and 75^{th} quantiles are as follows:

$$Q\left(.5,\mu,\sigma\right)=\frac{1}{2}\left(\frac{\mu}{\sqrt{\mu^2+2}}+1\right)$$

$$Q\left(.25,\mu,\sigma\right)=\frac{3\mu-\sqrt{6}\sigma}{2\sqrt{\left(3\mu-\sqrt{6}\sigma\right)^2+18}}+\frac{1}{2}$$

$$Q\left(.75,\mu,\sigma\right)=\frac{3\mu+\sqrt{6}\sigma}{2\sqrt{\left(3\mu+\sqrt{6}\sigma\right)^2+18}}+\frac{1}{2}$$

We now obtain the quantile estimates and compare them with the empirical quantiles.

```r
mu1 <- coef(fit)[1]
sigma1 <- exp(coef(fit)[3])
q125 <- 0.5 * (1 + (3 * mu1 - sqrt(6) * sigma1)/(sqrt((3 * mu1 - sqrt(6) * sigma1)^2 +
    18)))
q15 <- 0.5 * (1 + mu1/sqrt(2 + mu1^2))
q175 <- 0.5 * (1 + (3 * mu1 + sqrt(6) * sigma1)/(sqrt((3 * mu1 + sqrt(6) * sigma1)^2 +
    18)))
# Here are the empirical quantiles and quantile estimates for 'unlikely':
matrix(c(quantile(data$probm[data$valence == 0], c(0.25, 0.5, 0.75)), q125,
    q15, q175), nrow = 2, byrow = TRUE, dimnames = list(c("empirical", "estimated"),
    c("Q.25", "Q.5", "Q.75")))
```

```
##                 Q.25       Q.5      Q.75
## empirical 0.4701345 0.7488789 0.8285202
## estimated 0.5813702 0.7458176 0.8470549
```

```r
mu2 <- coef(fit)[1] + coef(fit)[2]
sigma2 <- exp(coef(fit)[3] + coef(fit)[4])
q225 <- 0.5 * (1 + (3 * mu2 - sqrt(6) * sigma2)/(sqrt((3 * mu2 - sqrt(6) * sigma2)^2 +
    18)))
q25 <- 0.5 * (1 + mu2/sqrt(2 + mu2^2))
q275 <- 0.5 * (1 + (3 * mu2 + sqrt(6) * sigma2)/(sqrt((3 * mu2 + sqrt(6) * sigma2)^2 +
    18)))
# Here are the empirical quantiles and quantile estimates for 'likely':
matrix(c(quantile(data$probm[data$valence == 1], c(0.25, 0.5, 0.75)), q225,
    q25, q275), nrow = 2, byrow = TRUE, dimnames = list(c("empirical", "estimated"),
    c("Q.25", "Q.5", "Q.75")))
```

```
##                Q.25       Q.5      Q.75
## empirical 0.5895964 0.6991031 0.7886996
## estimated 0.5841082 0.6986953 0.7854783
```

The estimates fit the data better for the "likely" sentences than for the "unlikely" sentences, with the poorest fit being evident in the 25th quantile for the "unlikely" sentences.

Example 2: CDF analysis for Ambiguity-Conflict data

Smithson and Shou (2016) examined lay people's probability judgments when receiving ambiguous or conflicting information. The main hypotheses were that participants' probability judgments for the ambiguous source should be (1) higher, and (2) with greater consensus than those for the conflicting source.

The experimental task required participants to estimate the number of days on which there would be a storm or the probability of a storm, given pairs of expert forecasts (i.e., sources of information). There were 1570 participants (782 males) recruited from the United States, aged from 18 to 88 with a mean of 51. Half of the participants received the expert forecasts of the number of days with storms, while half received the forecasts of the probability of a storm.

Each participant viewed four scenarios, each of which presented two pairs of expert forecasts. One pair of experts provided ambiguous but agreeing forecasts, while another pair of experts provided conflicting forecasts. For example, the two ambiguous but agreeing experts in Scenario 2 predicted that 4 to 6 days out of the next 7 days (or 40% to 60% probability) would have a storm. The conflicting experts in the same scenario had one expert predicting 4 out of 7 days would have a storm while the other predicted 6 (or 40% versus 60% probability). Based on a pair of forecasts (i.e., [4, 6] and [4, 6] for the ambiguous pair; and 4 and 6 for the conflicting pair), participants were requested to provide their own estimates of how many days out of the 7 days would have a storm.

Distribution Selection

In order to choose an appropriate distribution, we selected two finite-tailed distributions and four trimodal distributions and examined how well these distributions fit the data. We fitted participants' responses in each scenario using the six CDF-Quantile distributions and the beta distribution. The type of uncertainty (Ambiguous vs. Conflicting) was used as the predictor of the mean and dispersion. The analyses for fitting the CDF-Quantile distributions were performed in R by using the `cdfquantreg` package. The beta distribution was fitted by the `gamlss` package that also utilizes a location (μ) submodel and a dispersion (σ) submodel.

Table 1 shows the fit of the distributions to the entire data-set, and for each of the four scenarios. The best fit distribution overall is Arcsinh-T2 distribution, followed by Arcsinh-Cauchy and Arcsinh-Burr VIII. The beta distribution's fit is substantially worse than both the finite-tailed and trimodal CDF-Quantile distributions.

Table 1: Log likelihood for selected distribtuions

	Overall	S1	S2	S3	S4	Scenario Average	Shape
Arcsinh-Cauchy	6345.16	1611.60	1331.95	1642.59	1862.65	1612.20	Finite-tailed
T2-T2	5337.27	1551.34	1044.15	1367.19	1530.67	1373.34	Finite-tailed
Arcsinh-T2	6520.37	1579.31	1445.30	1715.67	1889.72	1657.50	Trimodal
Arcsinh-Burr VIII	6293.93	1530.35	1378.53	1661.57	1836.68	1601.78	Trimodal
T2-Burr VII	5690.94	1576.14	1198.56	1481.02	1604.74	1465.11	Trimodal
T2-Burr VIII	5606.84	1570.92	1161.98	1454.66	1584.98	1443.13	Trimodal
Beta	1898.83	883.87	390.47	586.20	639.98	625.13	

Estimation Results

We then compared the estimation results of the Arcsinh-Cauchy (finite-tailed) and Arcsinh-T2 (trimodal) distributions, with those from the beta distribution. Table 2 shows the estimation results for these distributions for the four scenarios. The three distributions yielded similar estimates for Scenarios 1 and 3. Participants' probability judgments for the conflicting pair in Scenario 1 were significantly higher and more dispersed than for the ambiguous pair (the ambiguous pair is the reference category in the Table 2 output). On the other hand, participants' judgments for the conflicting pair in Scenario 3 were significantly lower and less dispersed than for the ambiguous pair. The three distributions yielded the same qualitative findings for Scenario 4 with no effect for the location and greater dispersion for the conflicting pairs, although the beta distribution returned a nearly significant negative coefficient in the location submodel.

Table 2: Fitting results for the 4 scenarios

	Arcsinh-Cauchy		Arcsinh-T2		Beta	
	coefficient(SE)	p	coefficient(SE)	p	coefficient(SE)	p
	Scenario 1					
β_0	-0.02(0.01)	.008	-0.01(0.01)	.074	0.01(0.02)	.581
β_1	0.03(0.01)	.003	0.03(0.01)	.002	0.06(0.03)	.023
δ_0	-1.85(0.04)	<.001	-1.73(0.03)	<.001	-0.60(0.02)	<.001
δ_1	0.13(0.05)	.012	0.14(0.04)	.001	0.16(0.03)	<.001
	Scenario 2					
β_0	0.01(0.01)	.170	0.04(0.01)	<.001	0.34(0.02)	<.001
β_1	-0.09(0.01)	<.001	-0.13(0.01)	<.001	-0.67(0.03)	<.001
δ_0	-1.83(0.04)	<.001	-1.51(0.03)	<.001	-0.10(0.02)	<.001
δ_1	0.16(0.06)	.006	0.08(0.05)	.096	-0.08(0.03)	.021
	Scenario 3					
β_0	0.02(0.01)	.001	0.05(0.01)	<.001	0.34(0.02)	<.001
β_1	-0.03(0.01)	<.001	-0.06(0.01)	<.001	-0.39(0.03)	<.001
δ_0	-1.82(0.04)	<.001	-1.49(0.03)	<.001	-0.10(0.02)	<.001
δ_1	-0.25(0.06)	<.001	-0.27(0.05)	<.001	-0.39(0.03)	<.001
	Scenario 4					
β_0	-0.01(0.00)	.047	-0.01(0.00)	.011	-0.03(0.02)	.126
β_1	0.00(0.01)	.793	0.00(0.01)	.957	-0.05(0.03)	.119
δ_0	-2.15(0.04)	<.001	-1.84(0.03)	<.001	-0.56(0.02)	<.001
δ_1	0.12(0.06)	.044	0.20(0.05)	<.001	0.40(0.03)	<.001

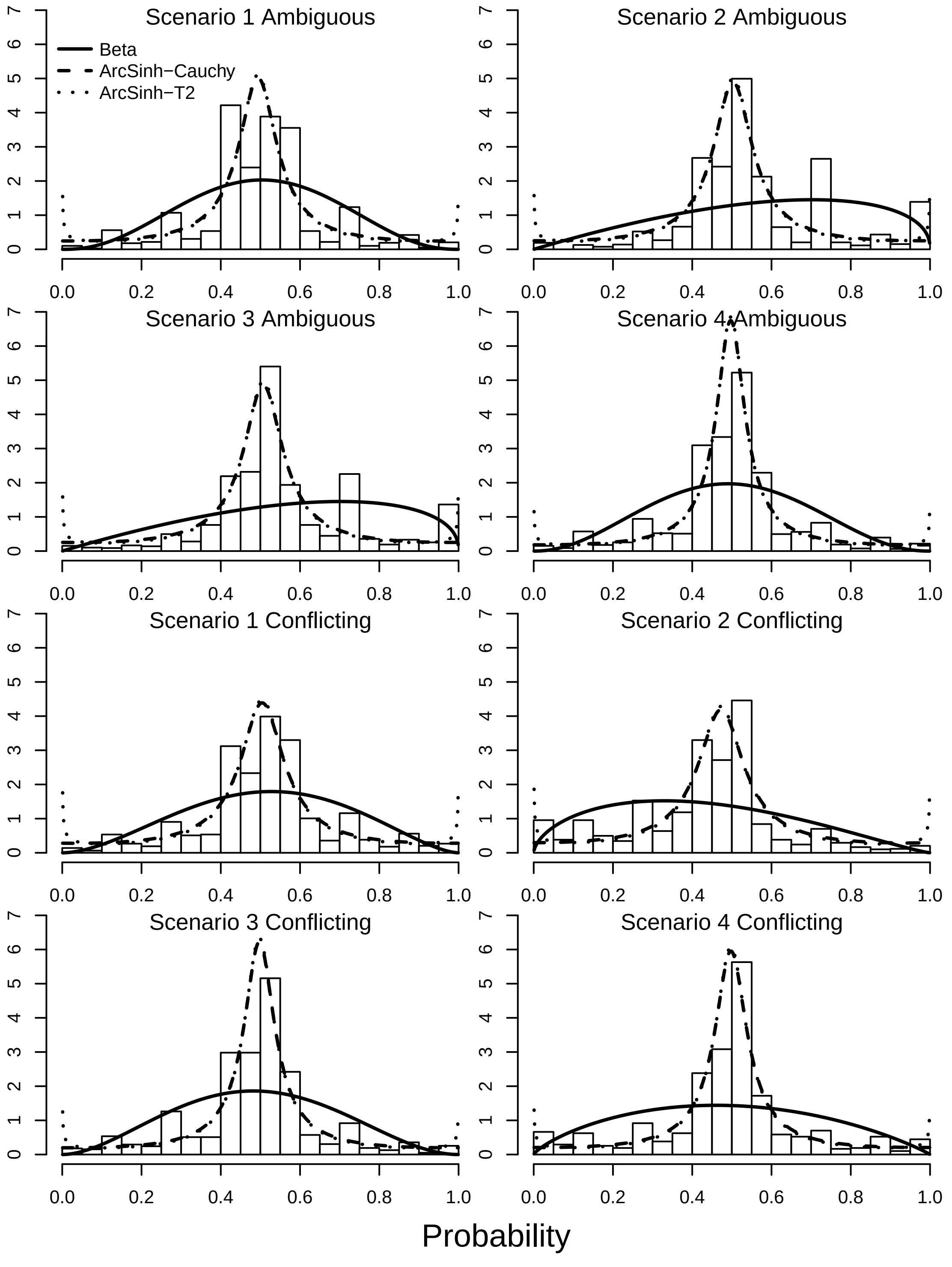

Figure 6. Fitting three distributions to the ambiguity-conflict data.

The three distributions agreed in their location submodels but differed in their dispersion submodels for Scenario 2. The Arcsinh-Cauchy model indicated that estimates for the conflicting forecasts were more dispersed than for the ambiguous forecasts, whereas the beta model found the opposite, and the ArcSinh-T2 model did not find a significant difference in dispersion (although its estimate is closer to the Arcsinh-Cauchy than to the beta). Figure 6 displays histograms of the data overlaid with the fitted distributions. It can be seen that in Scenario 2 the beta distribution was substantially influenced by the additional peak at the tails of the distributions. The overall shape of the beta distribution was biased towards the tail that had a second mode. In contrast, the overall shapes of the ArchSinh-Cauchy and ArcSinh-T2 were robust to the tails and captured the central modes more accurately than the beta, which accounts for their higher log-likelihoods in Table 1, most notably for Scenario 2.

Example 1 Continued: Modeling Best Estimates with Random Intercept

Given that every subject in our example provided ratings for the six estimation tasks, ratings within subjects are not likely to be independent, and a reasonable treatment of the data would be to add a subject-effect parameter in the form of a random intercept to our models. Thus, the original model is extended as follows:

$$\begin{aligned} \hat{\mu}_i &= \hat{\beta}_{0i} + \hat{\beta}_1 x_i, \\ \hat{\beta}_{0i} &= \hat{\gamma}_0 + u_{0i}, \\ \hat{\sigma}_i &= \hat{\delta}_0 + \hat{\delta}_1 x_i \end{aligned} \tag{1}$$

where $u_{0i} \sim N\left(0, \sigma_0^2\right)$.

We estimated such models (in SAS 9.3) for the beta, Logit-Logistic, and T2-T2 examples (SAS code is available at
https://dl.dropboxusercontent.com/u/1857674/cdfquantile/SAS_files.zip). The addition of the random intercept improved model fit in all three cases (LL values increased from 265.0 to 290.0 for the beta, 345.2 to 387.4 for the Logit-Logistic, and 435.3 to 545.5 for the T2-T2). Nonetheless, these models altered their respective fixed-effects coefficients only slightly. Thus, the qualitative pattern of findings remained the same as in the fixed-effects models (e.g., opposite-signed location submodel coefficients for the T2-T2 and beta models, and a non-significant location submodel effect for the Logit-Logistic).

References

Bowley, A L (1920). *Elements of statistics.* London : P. S. King and Son.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C. & Wallis, J. R. (1979). Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, **5**, 1049–1054. doi: 10.1029/WR015i005p01049.

Moors, J J A (1988). A Quantile Alternative for Kurtosis. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **1**, 25–32. doi: 10.2307/2348376.

SAS Institute Inc. (2011). SAS Software, Version 9.3. url: http://www.sas.com/.

Smithson M & Shou Y (2016). *Judgments under Ambiguity and Conflict.* The Australian National University, Canberra, Australia.