Quantile, risk measure and inequality index with heavy-tailed distribution

Arthur Charpentier
Université du Québec à Montréal

Emmanuel Flachaire
Aix-Marseille University

Motivation

In the presence of heavy-tails:

- Density: kernel estimation can be very poor
- Indices: sample estimates can be very far from true values

Our contribution:

- Density: more efficient estimation
- Indices: more efficient estimators

Kernel density estimation

Distribution	Upper-tail	Kernel methods		
		Ksil	Kcv	Kad
Lognormal	moderate	0.1031	0.1087	0.1090
	medium	0.1323	0.1319	0.1284
	strong	0.1630	0.1670	0.1572
Singh-Maddala	moderate	0.0985	0.1055	0.0986
	medium	0.1089	0.1153	0.1041
	strong	0.1290	0.1409	0.1237
Mixture of S-M.	moderate	0.2262	0.1456	0.1348
	medium	0.2665	0.1678	0.1455
	strong	0.3005	0.2310	0.1684

Table : Quality of density estimation: MIAE criteria, $n = 500$.

Sample estimates

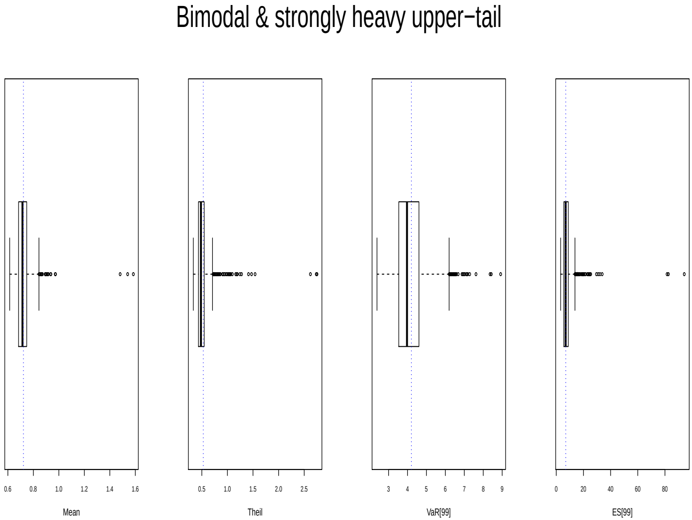


Figure : Boxplots of sample estimates (EDF-based estimates), $n = 500$

The transformed kernel estimate

Devroye and Györfi (1985), Devroye and Lugosi (2001):[1]

The *transformed kernel estimate* (Devroye et al., 1983) is based upon a transformation $T\colon R^1 \to [0,1]$ which is strictly monotonically increasing, continuously differentiable, one-to-one and onto, and which has a continuously differentiable inverse. The transformed data sequence is $Y_1, \ldots, Y_n$, where $Y_i = T(X_i)$. Note that Y_1 has density

$$g(x) = f(T^{-1}(x))T^{-1\prime}(x).$$

Now, g is estimated by g_n from $Y_1, \ldots, Y_n$, and f is estimated by

$$f_n(x) = g_n(T(x))T'(x). \tag{2}$$

Estimate the density via a transformation into the unit interval

$$\mathbb{R} \longrightarrow [0,1] \longrightarrow \mathbb{R}$$

[1] see Ruppert and Cline (1994), Hössjer and Ruppert (1995), Swanepoel and Van Graan (2005), Buch-Larsen et al. (2005), Charpentier and Oulidi (2010)

Method

Nonparametric density estimation via a transformation into $[0, 1]$:

1. Transform the sample with a function $T : \mathbb{R} \to [0, 1]$
2. Estimate the density of the transformed sample $Y = T(X)$
3. Obtain the density of the original sample as,

$$\hat{f}_X(x) = \hat{f}_Y[T(x)] \, . \, T'(x),$$

Optimal transformation function:

- If T is the CDF of X, the distribution of Y is $\mathcal{U}(0, 1)$

Method

Nonparametric density estimation via a transformation into $[0, 1]$:

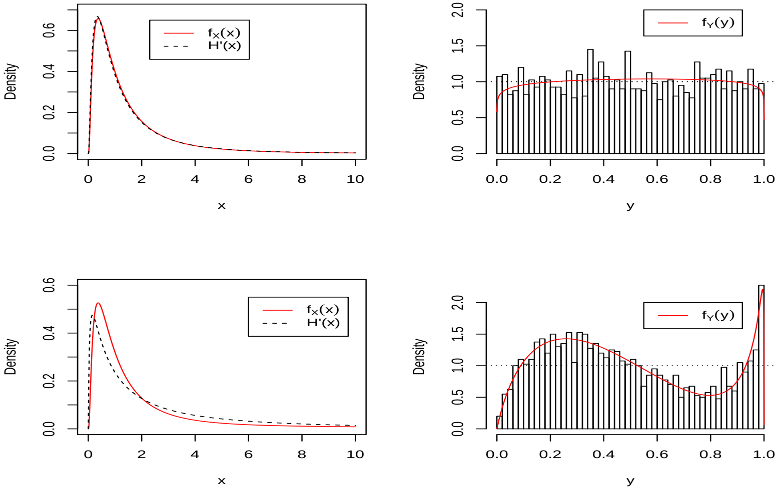
1. Transform data with a parametric CDF estimate $H(X, \hat{\theta})$
2. Estimate the density of the transformed sample $Y = H(X, \hat{\theta})$
3. Obtain the density of the original sample as,

$$\hat{f}_X(x) = \hat{f}_Y[H(x, \hat{\theta})] \,.\, H'(x, \hat{\theta}),$$

Optimal transformation function:

- If $H(X, \hat{\theta})$ is a correctly specified dist. of X, $Y \underset{as}{\sim} \mathcal{U}(0, 1)$
- If $H(X, \hat{\theta})$ is a misspecified distribution of X, a nonparametric estimation will capture any deviation from $\mathcal{U}(0, 1)$

Method



Main idea

Questions:

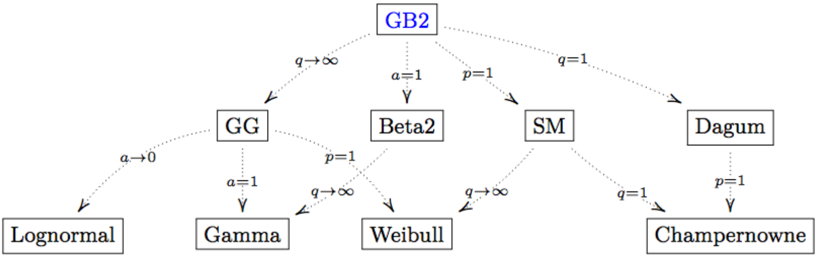
- which transformation function?
- what should we do on $[0, 1]$?

With heavy-tailed distribution:

- use a transformation function to fit appropriately the tails $\rightarrow$ parametric CDF
- use a nonparametric estimation on $[0, 1]$ to capture any deviation from $\mathcal{U}(0, 1)$

The transformation function (positive data)

$$GB2: \; t(y; a, b, p, q) = \frac{|a| y^{ap-1}}{b^{ap} \mathrm{B}(p,q)[1 + (y/b)^a]^{p+q}}, \qquad \text{for } y > 0,$$



GB2 permits to capture upper tail of many different types, including heavy-tails and also light-tails

What should we do on $[0, 1]$?

Nonparametric density estimation with bounded support:[2]

- Beta kernel
- mixture of beta distributions
- (Bernstein)

[2] Chen (1999), Bouezmarni & Rolin (2003), Bouezmarni & Rombouts (2010)

Beta distribution

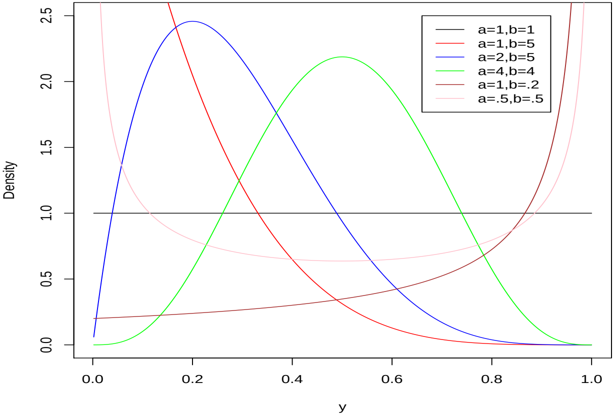
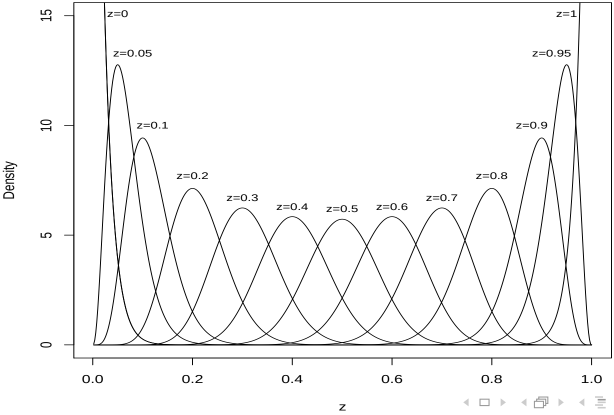


Figure : Beta density functions, beta$(y; a, b)$

Beta kernel estimation

$$\hat{f}(u) = \sum_{i=1}^{n} \frac{1}{n} \text{ beta}\left(u; \frac{U_i}{h}, \frac{1-U_i}{h}\right)$$

with some possible boundary corrections (Chen 1999)



Beta kernel estimation: bandwidth sensitivity

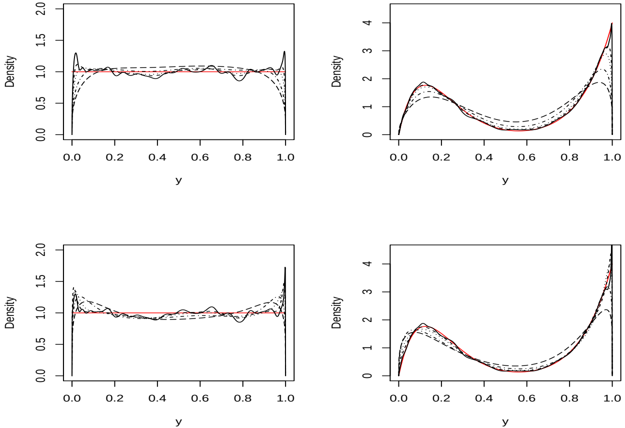


Figure : Beta kernel density estimation and bandwidth sensitivity. The red line is the true density function, other lines are beta kernel estimators with $h = .002, .01, .02, .05, .1$. The top figures are computed with the standard beta kernels and the bottom figures with the modified beta kernels

Beta kernel estimation: bandwidth selection

We capture deviations from the uniform distribution!

- Rule-of-thumb? (reference distribution uniform: $h = \infty$)
- Cross-validation? (regularity conditions)

Automatic bandwidth selection is problematic.

Mixture of Beta distributions

The beta mixture estimator is

$$\hat{f}(u) = \sum_{k=1}^{K} \hat{\pi}_k \text{ beta}\left(u; \hat{a}_k, \hat{b}_k\right)$$

with $0 \leq \hat{\pi}_k \leq 1$ and $\sum_{k=1}^{K} \hat{\pi}_k = 1$

- It allows us to bring out the link between parmetric ($K = 1$) and nonparametric ($K = n$) estimation.
- K is selected by minimizing a criteria, as the BIC.[3]

[3] $\text{BIC} = -2\hat{\ell} + \#\text{param} \log n$

Simulations

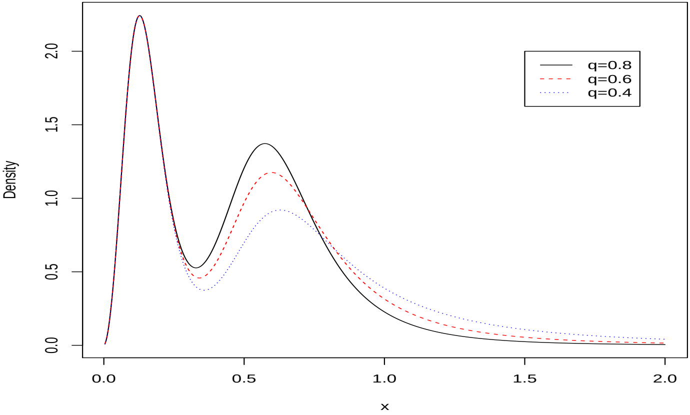


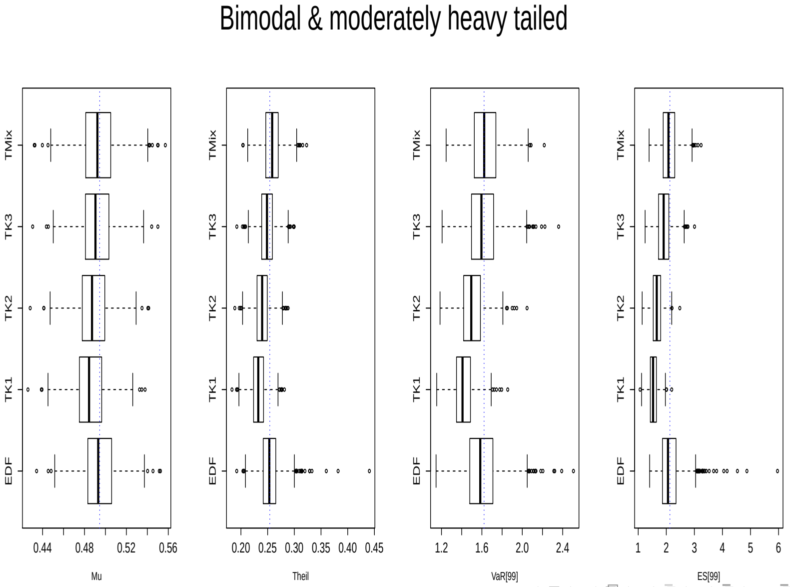
Figure : Mixture of two Singh-Maddala distributions

Results: density estimation

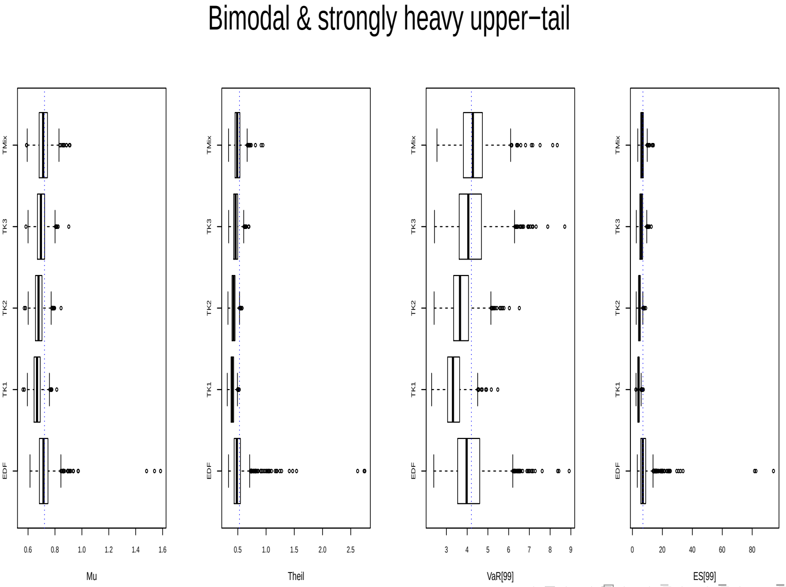
	Kernel methods			Transformed methods			
Upper-tail	Ksil	Kcv	Kad	TK1	TK2	TK3	TMix
Lognormal							
moderate	0.1031	0.1087	0.1090	0.1023	0.1178	0.1773	0.0899
medium	0.1323	0.1319	0.1284	0.1023	0.1176	0.1766	0.0904
strong	0.1630	0.1670	0.1572	0.1021	0.1175	0.1767	0.0890
Singh-Maddala							
moderate	0.0985	0.1055	0.0986	0.1058	0.1200	0.1774	0.0927
medium	0.1089	0.1153	0.1041	0.1061	0.1201	0.1774	0.0931
strong	0.1290	0.1409	0.1237	0.1042	0.1187	0.1765	0.0905
Mixture of two Singh-Maddala,							
moderate	0.2262	0.1456	0.1348	0.1096	0.1172	0.1714	0.1022
medium	0.2665	0.1678	0.1455	0.1093	0.1178	0.1722	0.1030
strong	0.3005	0.2310	0.1684	0.1072	0.1171	0.1730	0.1078

Table : Quality of density estimation: MIAE criteria, $n = 500$.

Results: estimation of indices (moderately heavy-tailed)



Results: estimation of indices (strongly heavy-tailed)



Conclusion

With transformed kernel/mixture density estimation:

- Density: the quality of the fit is better, compared to standard kernel estimation, it does not deteriorate as the tail is heavier
- Indices: more reliable estimation, compared to sample indices

We should also improve inference (work in progress) ...