Log-transform Kernel Density Estimation of Income Distributions

by

Arthur Charpentier
Université du Québec à Montréal

and

Emmanuel Flachaire
Aix-Marseille Université

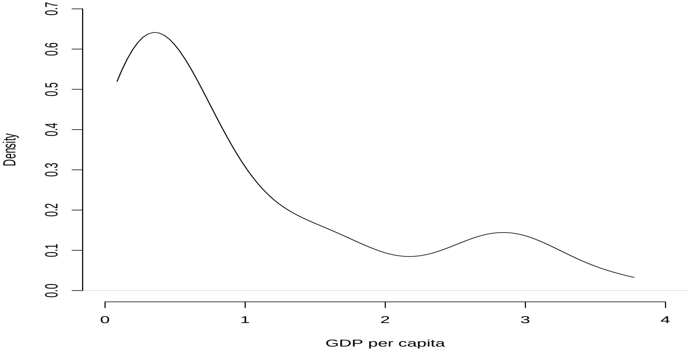
What can we say from sample values?

```
> GDP
  [1] 0.590 0.156 0.212 0.514 0.111 0.119 0.284 0.218 0.125 0.085 0.130 0.499 0.404
 [14] 0.700 0.171 0.173 0.165 0.139 0.303 0.192 0.226 0.146 0.105 0.110 0.168 1.127
 [27] 0.454 0.162 0.594 0.115 0.200 0.710 0.161 0.250 0.718 0.169 0.717 0.515 0.114
 [40] 0.136 0.575 0.112 0.098 0.161 0.232 1.584 3.680 0.713 0.492 0.393 0.453 0.181
 [53] 0.302 0.521 1.141 0.305 0.599 1.786 1.724 3.776 1.141 0.356 0.897 0.851 0.689
 [66] 0.603 0.316 0.429 0.581 0.651 0.999 1.445 0.277 0.283 2.979 0.257 0.369 0.686
 [79] 1.934 2.805 0.749 1.196 0.909 0.098 0.292 0.357 1.509 2.200 0.432 0.890 1.529
 [92] 0.634 2.539 2.627 1.571 0.876 2.894 2.852 2.827 2.869 1.377 2.887 1.684 2.541
[105] 3.206 1.305 2.577 3.129 1.282 0.445 1.868 3.072 3.382 0.729 2.765 1.624 1.054
[118] 3.135 0.723 2.452 0.345
> mean(GDP)
[1] 0.9999587
> median(GDP)
[1] 0.594
> sd(GDP)
[1] 1.000607
```

Standard descriptive statistics provide some limited information.

What can we say from density function?

```
> plot(density(GDP))
```



Density function provides much more information.

Motivation

Density estimation can be very useful:

- Standard descriptive statistics outline some features of density function (location, scale, skewness, . . .)
- Much more information is given by the density function

How to estimate density function:

- parametric estimation
- histogram
- kernel method

With income distributions: standard methods often perform poorly!

Income distributions

Specific features of income distributions?

- they are defined over the positive support, $[0, +\infty)$
- they are often very skewed, with heavy upper-tail
- they can be multimodal

What is the problem with heavy-tailed distributions?

- The upper tail decays slowly (more than exponential dist.)
- Higher probability to have large observations in a sample
- Heavy-tail is known to cause statistical problems[1]

[1] Davidson and Flachaire (2007, JoE), Cowell and Flachaire (2007, JoE; 2015, Handbook), Schluter and van Garderen (2009, JoE), Davidson (2012, EJ)

Our contribution

Log-transformation!

- Transform the data with a logarithmic transformation[2]
- Apply kernel methods to $\log X_1, \ldots, \log X_n$
- Use a back-transformation

Why?

- log-transformation squashes the right tail of the distribution.
- if $f(x)$ is lognormal or Pareto-type in the upper tail, $f(logX)$ is no longer heavy- tail.
- Income distributions are Pareto-type in the upper tail, if not lognormal

$\rightarrow$ Kernel methods applied to a distribution no longer heavy-tailed!

[2] Devroye and Györfi (1985), Marron and Ruppert (1985), Silverman (1986)

Our results

1. We first show that a preliminary logarithmic transformation of the data, combined with standard kernel density estimation methods, can provide a much better fit of the overall density estimation.
2. We calculate the bias and variance of the log-transform kernel density estimation method.
3. Then, we show that the fit of the bottom of the distribution may not be satisfactory, even if a better fit of the upper tail can be obtained in general.

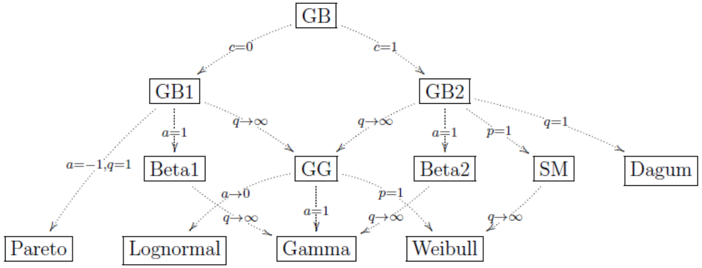
Outline

1. Standard density estimation
2. Log-transform kernel density estimation
3. Finite sample performance: overall estimation
4. Finite sample performance: pointwise estimation
5. Application

Standard density estimation

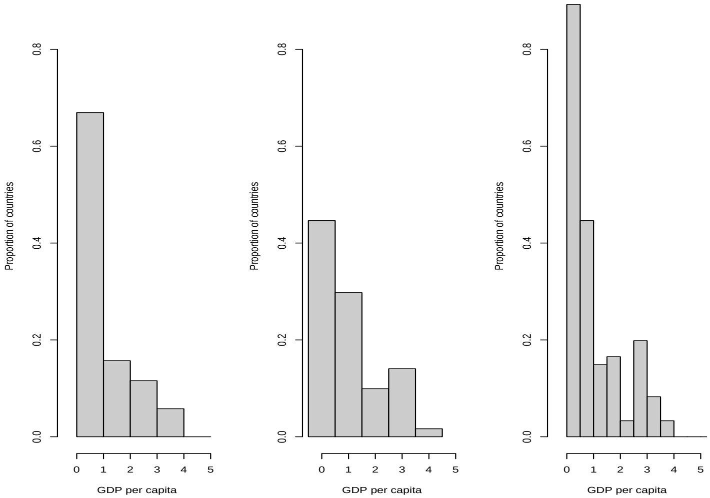
Parametric density estimation

Most of the standard parametric income distributions are special cases of the Generalized Beta distribution:



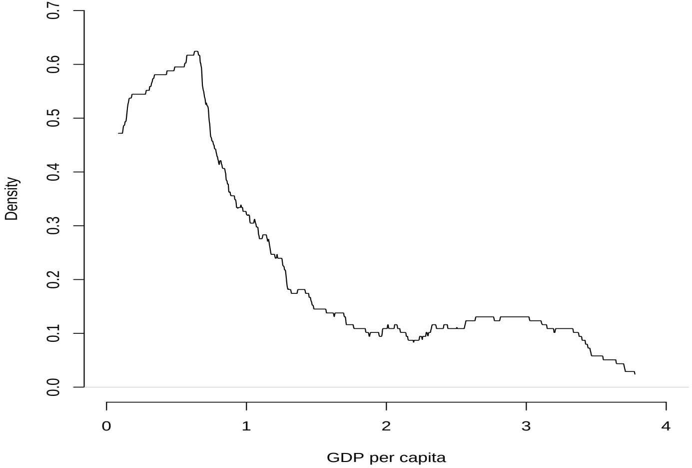
However, they are unimodal. Nonparametric density estimation maybe more appropriate (Marron and Schmitz 1992, ET).

Histograms



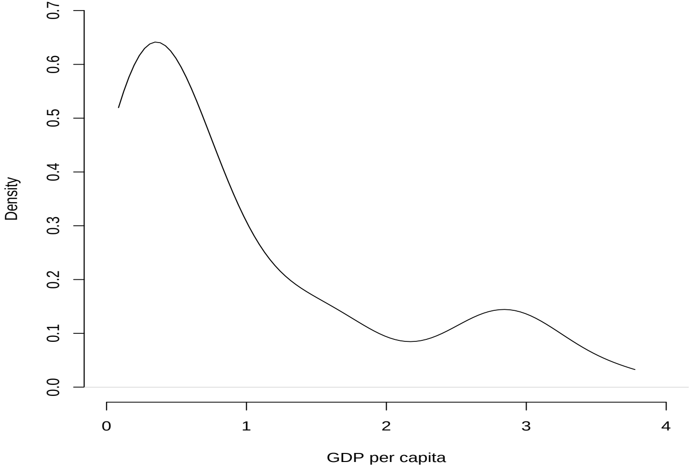
Histogram's sensitivity to the position and the number of bins

Naive estimator



Naive estimator of GDP per capita

Kernel estimator



Kernel estimator of GDP per capita

Bandwidth selection

The kernel density estimator

$$\hat{f}(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right)$$

is not really affected by the choice of the kernel $K()$, but it is sensitive to the choice of the bandwidth h

Bandwidth selection:

- Silverman's rule-of-thumb, $\hat{h}_{opt} = 0.9 \min\left(\hat{\sigma}\,;\, \frac{\hat{q}_3 - \hat{q}_1}{1.349}\right)\, n^{-\frac{1}{5}}$.
- Plug-in method of Sheather and Jones (1991, JRRS)
- Cross-validation

Adaptive kernel

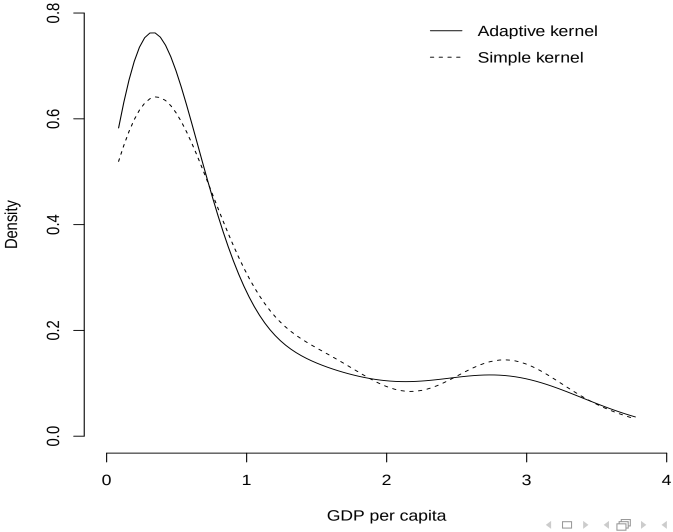
- When the concentration of the data is markedly heterogeneous, a fixed bandwidth may be quite restrictive.

- The adaptive kernel estimator is defined as follows:

$$\hat{f}(y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h\lambda_i}K\left(\frac{y-y_i}{h\lambda_i}\right),$$

where λ_i is a parameter that varies with the local concentration of the data, $\lambda_i = [g/\tilde{f}(y_i)]^{\alpha}$.

Adaptive kernel



Log-transform kernel density estimation

Method

Let us consider $Y = G(X)$, by the change of variable formula,[3]

$$f_X(x) = f_Y[G(x)].G'(x),$$

Let us consider $Y = \log X$, the *log-transform kernel* method is

$$\begin{aligned}
\hat{f}_X(x) &= \hat{f}_Y(\log x)\,\frac{1}{x} = \frac{1}{nh}\sum_{i=1}^{n}\phi\left(\frac{\log x - \log X_i}{h}\right)\,\frac{1}{x} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{xh\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\log x - \log X_i}{h}\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathit{lognormal}(x; \log X_i, h)
\end{aligned}$$

h: Silverman, plug-in, or CV methods applied to $\log X_1, \ldots, \log X_n$.

[3] where G is a monotonically strictly increasing function

Kernel method: a sum of 'bumps'

- Standard (Gaussian) kernel method:

$$\hat{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n} normal(x; X_i, h)$$

It is a sum of 'bumps' defined by Normal distributions with expectations X_i and a fixed variance h^2.

- Log-transform kernel method:

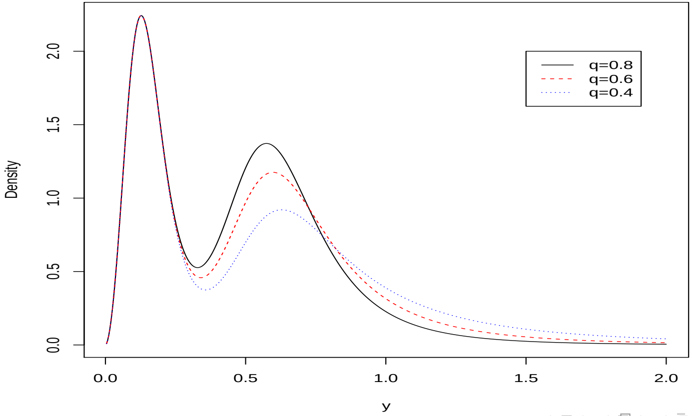
$$\hat{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n} lognormal(x; \log X_i, h)$$

It is a sum of 'bumps' defined by Lognormal distributions with medians X_i and varying variances $(e^{h^2} - 1)e^{h^2} X_i^2$.

Finite sample performance: overall estimation

Model design

- Quality of the fit: MIAE $= E\left(\int_0^\infty \left|\hat{f}(y) - f(y)\right| \mathrm{d}y\right)$.
- Data are drawn from lognormals, Singh-Maddala and mixtures of two SM distributions.



Overall estimation: results

	standard			adaptive			log-transform		
Tail	Ksil	Kcv	Ksj	AKsil	AKcv	AKsj	LKsil	LKcv	LKsj
Lognormal									
moderate	0.104	0.109	0.103	0.098	0.110	0.103	0.082	0.087	0.082
medium	0.133	0.133	0.125	0.110	0.128	0.118	0.082	0.087	0.082
strong	0.164	0.172	0.152	0.126	0.161	0.136	0.082	0.087	0.082
Singh-Maddala									
moderate	0.098	0.105	0.099	0.093	0.102	0.096	0.087	0.094	0.087
medium	0.108	0.115	0.109	0.096	0.109	0.102	0.088	0.094	0.088
strong	0.129	0.138	0.128	0.103	0.126	0.114	0.090	0.096	0.090
Mixture of two Singh-Maddala,									
moderate	0.225	0.145	0.139	0.172	0.140	0.125	0.163	0.120	0.115
medium	0.266	0.164	0.157	0.206	0.154	0.135	0.158	0.121	0.115
strong	0.300	0.229	0.182	0.232	0.212	0.150	0.157	0.122	0.117

Table: Quality of density estimation: standard, adaptive, log-transform kernel methods, MIAE criteria (worst in red, best in blue), $n = 500$.

Overall estimation: summary

- The popular standard kernel density estimation method with the Silverman's rule of thumb bandwidth performs very poorly with bimodal and heavy-tailed distributions.
- Standard and adaptive kernel methods deteriorate as the upper tail becomes heavier (from moderate to strong).
- Log-transform kernel methods do not deteriorate as the upper tail becomes heavier.
- The log-transform kernel estimation method with the plug-in bandwidth of Sheather and Jones outperforms other methods .

$\rightarrow$ Log-transform kernel estimator outperforms other methods

The MIAE criteria

- The MIAE criteria gives one specific picture of the quality of the fit

$$\mathrm{MIAE} = \mathbb{E}\left(\int_{-\infty}^{+\infty} \left|\hat{f}(x) - f(x)\right| \mathrm{d}x\right)$$

- It is the *mean* of the IAE values obtained from each sample

$$\mathrm{IAE} = \int_{-\infty}^{+\infty} \left|\hat{f}(x) - f(x)\right| \mathrm{d}x$$

- It is a global measure and nothing can be said on the quality of the fit at different parts of the distribution.

Finite sample performance: pointwise estimation

Density property

The support of x is bounded to the left: $x \in [0, +\infty)$.

Standard and adaptive kernel methods:

- They may put positive mass outside the support
- We can set $\hat{f}(x) = 0, \forall x < 0$, but, then $\int_0^{+\infty} \hat{f}(x) \neq 1$!!

Log-transform kernel method

- It does not put positive mass outside the support
- We have $\int_0^{+\infty} \hat{f}(x) = 1$. [4]

---

[4] with a change of variable, $y = \log x$, so that $dy = dx/x$, we have:

$$\int_0^{\infty} \hat{f}_X(x)dx = \int_0^{\infty} \hat{f}_Y(\log x)\frac{dx}{x} = \int_{-\infty}^{\infty} \hat{f}_Y(y)dy = 1. \qquad (1)$$

The bias

The bias of the standard (Gaussian) kernel estimator is obtained from a Taylor expansion:

$$\text{bias}\{\tilde{f}_X(\epsilon)\} = \mathbb{E}[\tilde{f}_X(\epsilon)] - f_X(\epsilon) \sim \frac{h^2}{2} f''_X(\epsilon) \tag{2}$$

We derive the bias of the log-transform kernel estimator:

$$\text{bias}\{\hat{f}_X(\varepsilon)\} \sim \frac{h^2}{2}\left[f_X(\varepsilon) + 3\varepsilon \cdot f'_X(\varepsilon) + \varepsilon^2 \cdot f''_X(\varepsilon)\right] \tag{3}$$

With distributions such that $f_X(0) = 0$, in the neigborhood of the boundary, the bias of the log-transform kernel estimator is small:

$$\varepsilon \sim 0 \quad \Rightarrow \quad \text{bias}\{\hat{f}_X(\varepsilon)\} \sim 0 \tag{4}$$

$\rightarrow$ The bias of the log-transform kernel estimator should be smaller in the bottom of the distribution.

The variance

The variance of the standard (Gaussian) kernel estimator is :

$$\mathrm{Var}[\tilde{f}_X(\epsilon)] = \frac{1}{nh} f_X(\epsilon) \int K^2(u)du \qquad (5)$$

We derive the variance of the log-transform kernel estimator:

$$\mathrm{Var}[\hat{f}_X(\varepsilon)] \sim \frac{1}{\varepsilon nh} f_X(\varepsilon) \int K^2(u)du \qquad (6)$$

It is divided by ε, which would be large in the neighborhood of 0.

→ The variance of the log-transform kernel estimator should be larger for values of ε close to zero (in the bottom of the distrib.) and smaller for values of ε far from zero (in the top of the distrib.).

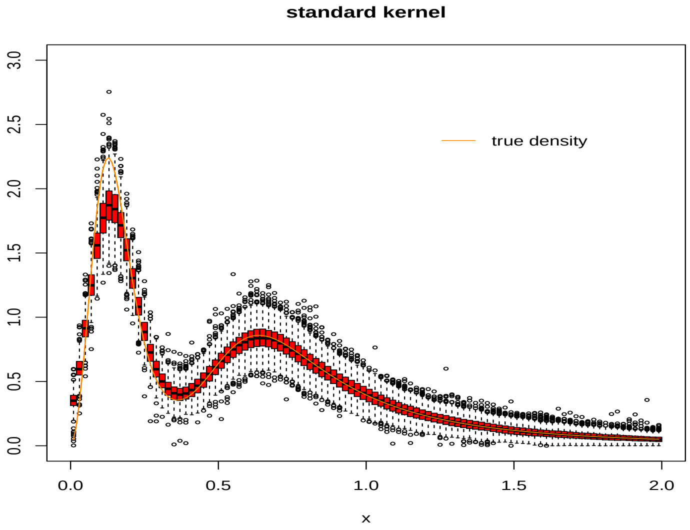


Figure: Pointwise estimation: boxplot of standard density estimation at point x for the less favourable case (bimodal with strong heavy-tail).

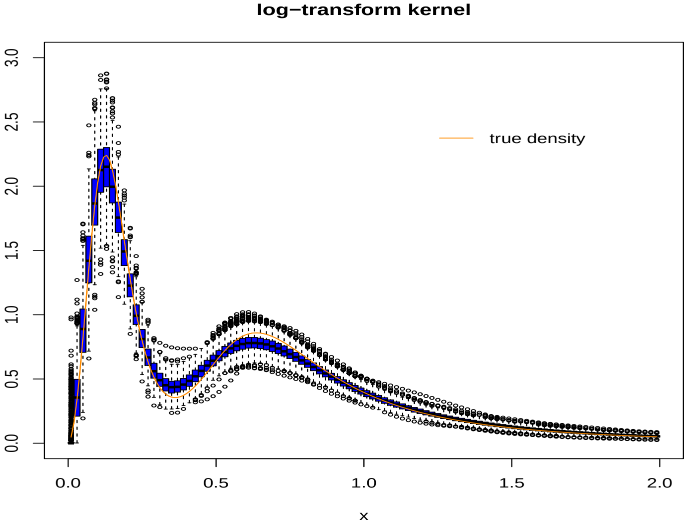
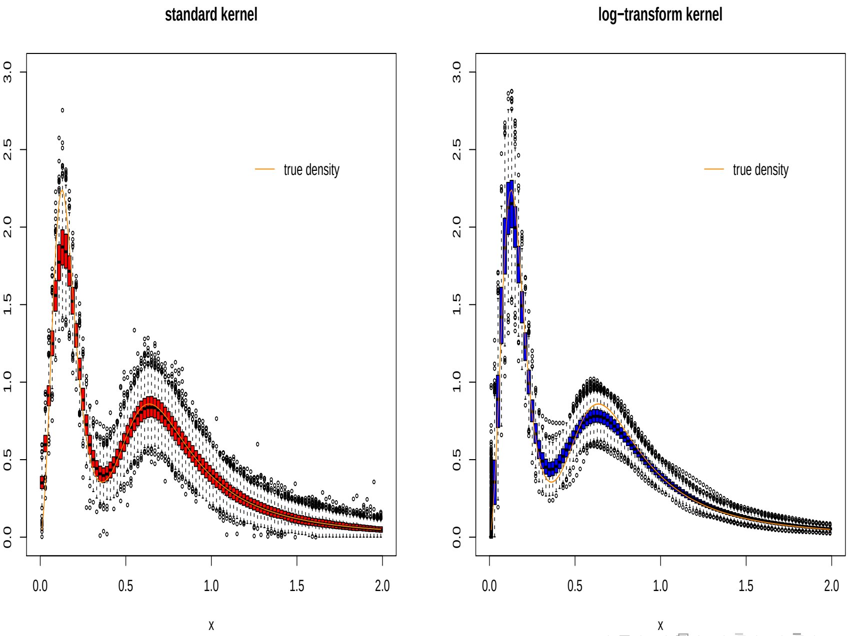
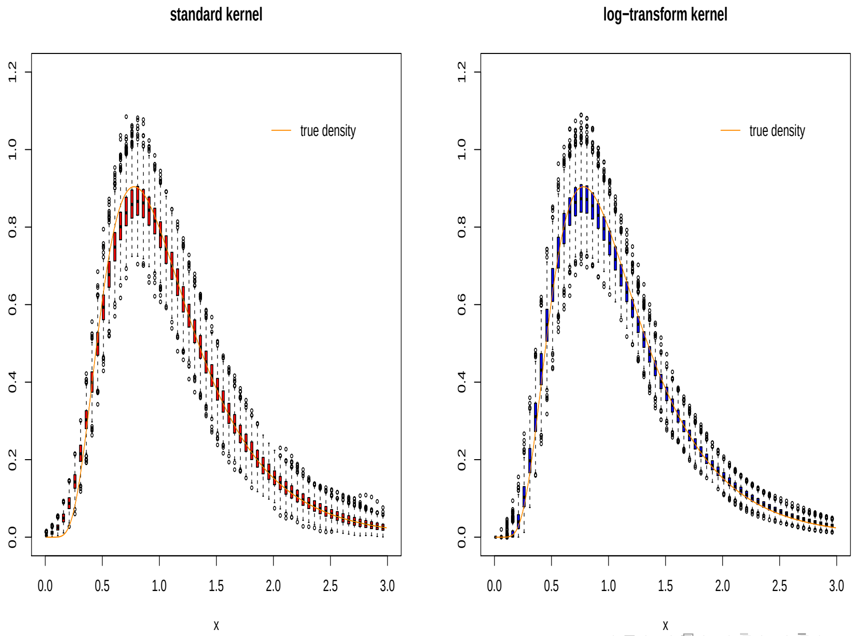


Figure: Pointwise estimation: boxplot of log-transform density estimation at point x for the less favourable case (bimodal with strong heavy-tail).


standard kernel
log−transform kernel
true density
true density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
x
x


standard kernel
log−transform kernel
true density
true density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
x
x

Poitwise estimation: summary

Compared to the standard kernel method, the log-transform kernel density estimator exhibits:

- smaller bias in the bottom of the distribution
- larger variance in the bottom of the distribution
- smaller variances in the top of the distribution

→ The top of income distributions is much better fitted, but the bottom of the distribution may not be satisfactory.

Silverman example

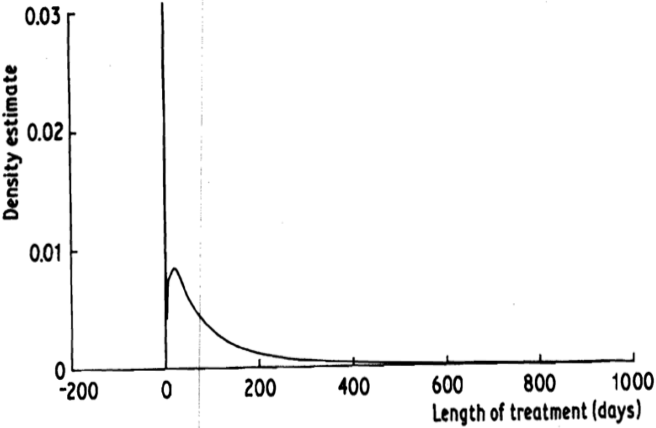
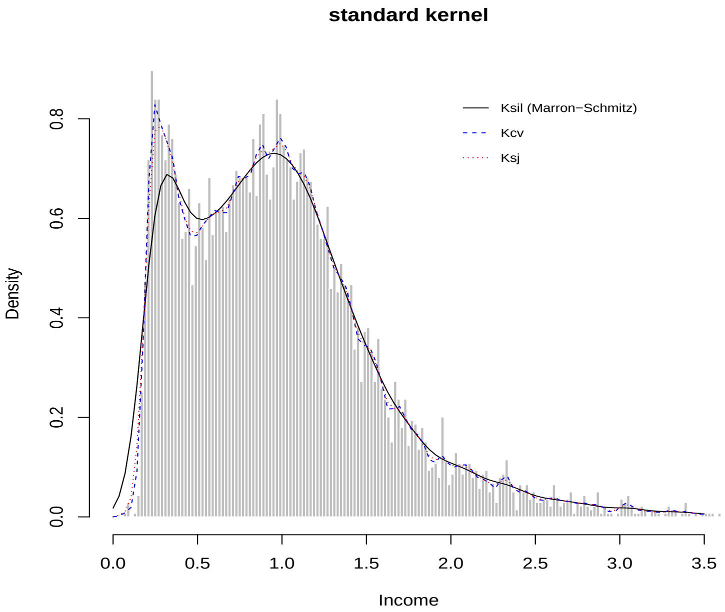


Fig. 2.13 *Log-normal weight function estimate for suicide study data, obtained by transformation of Fig. 2.12. Note that the vertical scale differs from that used in previous figures for this data set.*

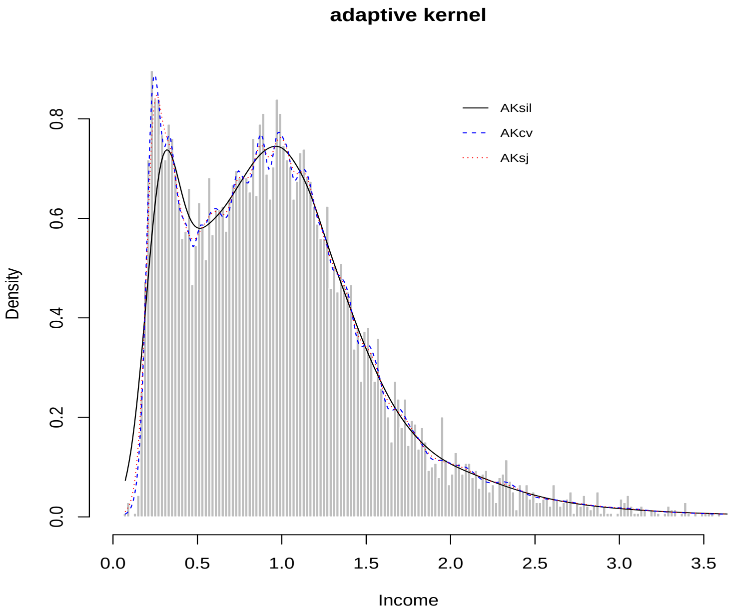
Application

Application

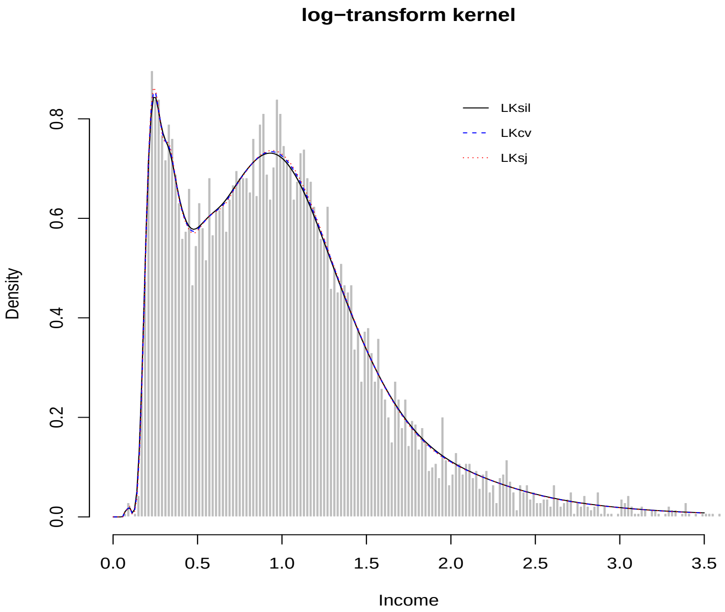
- Density estimation of income distribution in the UK in 1973
- Family expenditure survey (FES), a continuous survey of samples of the UK population living in households
- To restrict the study to relative effects, the data are normalized by the arithmetic mean of the year.
- The number of observations is large, $n = 6968$.
- Similar data as in Marron and Schmitz (1992).



standard kernel
Ksil (Marron−Schmitz)
Kcv
Ksj
Density
0.0
0.2
0.4
0.6
0.8
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Income



adaptive kernel
AKsil
AKcv
AKsj
Density
0.0
0.2
0.4
0.6
0.8
Income
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5


log−transform kernel
LKsil
LKcv
LKsj
Density
0.0
0.2
0.4
0.6
0.8
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Income

Conclusion

Log-transform kernel density estimation:

- appealing: kernel method applied to distrib. not heavy-tailed
- similar to lognormal kernel density estimation on original data
- greatly improve the quality of the *overall* density estimation
- derive the bias and variance
- greatly improve the quality of the *top* distribution estimation
- the *bottom* distribution estimation may not be satisfactory

Application: when n is large, a visual inspection and the use of a histogram as benchmark can help to see if the log-transform kernel density estimation is appropriate in the bottom of the distribution.