amse
aix marseille
school of economics

Working Papers / Documents de travail

Log-Transform Kernel Density Estimation of Income Distribution

Arthur Charpentier
Emmanuel Flachaire

WP 2015 - Nr 06

Log-transform kernel density estimation of income distribution

by

Arthur Charpentier

Université du Québec à Montréal
CREM & GERAD
Département de Mathématiques
201, av. Président-Kennedy, PK-5151,
Montréal, Qc H2X 3Y7, Canada.
charpentier.arthur@uqam.ca

and

Emmanuel Flachaire

Aix-Marseille University
(Aix-Marseille School of Economics), CNRS & EHESS
Institut Universitaire de France
2 rue de la charité,
13002 Marseille, France,
emmanuel.flachaire@univ-amu.fr

November 2014

We are grateful to Karim Abadir and Taoufik Bouezmarni for helpful comments. Charpentier acknowledges the support of the Natural Sciences and Engineering Research Council of Canada. Flachaire acknowledges the support of the Institut Universitaire de France, the Aix-Marseille School of Economics and the A*MIDEX project (ANR-11-IDEX-0001-02) funded by the "Investissements d'Avenir" French Government program, managed by the French National Research Agency (ANR)

Abstract

Standard kernel density estimation methods are very often used in practice to estimate density function. It works well in numerous cases. However, it is known not to work so well with skewed, multimodal and heavy-tailed distributions. Such features are usual with income distributions, defined over the positive support. In this paper, we show that a preliminary logarithmic transformation of the data, combined with standard kernel density estimation methods, can provide a much better fit of the density estimation.

JEL: C15

Keywords: nonparametric density estimation, heavy-tail, income distribution, data transformation, lognormal kernel

1 Introduction

Heavy-tailed distributions defined over the positive support have a upper tail that decays more slowly than exponential distribution (as the Gaussian distribution).[1] The probability to observe large values in sample datasets is then higher, which may cause serious problems in finite samples. For instance, standard kernel density estimation are known to perform poorly and statistical inference for inequality measures may be seriouly misleading with heavy-tailed income distributions.[2]

In this paper, we study kernel density estimation applied to a preliminary logarithmic transformation of the sample. The density of the original sample is obtained by back-transformation. The use of a logarithmic transformation of the sample is rather common when dealing with positive observations and heavy tailed distributions.[3] A feature of the log-transformation is that it squashes the right tail of the distribution. When the distribution of X is lognormal or Pareto-type in the upper tail, the distribution of $\log X$ is no longer heavy-tail.[4] Since most income distributions are Pareto-type in the upper tail, if not lognormal,[5] the logarithmic transformation is appealing in such cases: kernel density estimation is applied to a distribution that is no longer heavy-tail. The quality of the fit is then expected to be improved in finite samples.

Nonparametric density estimation based on a transformation of the data is not a new idea. It has been suggested by Devroye and Györfi (1985), a rigorous study can be found in Marron and Ruppert (1994), and, it has increasingly been used in the recent years.[6] Even if the logarithmic transformation has been briefly discussed by Silverman (1986), with bounded domains and directional data, no great attention has been given to this transformation. In this paper, we study the logarithmic transformation combined with the

[1] Heavy-tailed distributions are probability distributions whose tails are heavier than the exponential distribution. The distribution F of a random variable X is heavy-tail to the right if $\lim_{x\to\infty} e^{\lambda x}(1-F(x)) = \infty, \forall \lambda > 0$.

[2] see Davidson and Flachaire (2007), Cowell and Flachaire (2007), Davidson (2012) and Cowell and Flachaire (2015).

[3] For instance, the Hill estimator of the tail index is expressed as a mean of logarithmic differences.

[4] Indeed, if a random variable X has a Pareto type distribution (in the upper tail), $\mathbb{P}(X > x) \sim x^{-\alpha}$, then $\log X$ has an exponential-type distribution (in the upper tail) since $\mathbb{P}(X > x) \sim e^{-\alpha x}$. Moreover, if X is lognormal, $\log X$ is Gaussian.

[5] see Kleiber and Kotz (2003)

[6] see for instance Buch-Larsen et al. (2005), Markovich (2007), Charpentier and Oulidi (2010), Abadir and Cornea (2014)

use of kernel density estimation, leading us to provide new insights on non-parametric density estimation with heavy-tailed distributions.

In section 2, we present kernel density estimation methods. In section 3, we derive the bias and variance for the log-transform kernel method. In section 4, simulation experiments are investigated to study the quality of the fit in finite samples with heavy-tailed distributions. Section 5 is concerned by an application and section 6 concludes.

2 Kernel density estimation

Let us assume that we have a sample of n positive *i.i.d.* observations, $X_1, \ldots, X_n$. We want to estimate the underlying density function f_X, without any *a priori* hypothesis on its shape, namely assuming that the distribution belongs to some parametric family.

2.1 Standard kernel density estimation

The kernel density estimator with kernel K is defined by

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1}$$

where n is the number of observations and h is the bandwidth. It is a sum of 'bumps' - with shape defined by the kernel function - placed at the observations.

Kernel density estimation is known to be sensitive to the choice of the bandwidth h, while it is not really affected by the choice of the kernel function when we use symmetric density functions as kernels. For a detailed treatment of kernel density estimation, see the book of Silverman (1986), as well as Härdle (1989), Scott (1992), Wand and Jones (1995), Simonoff (1996), Bowman and Azzalini (1997), Pagan and Ullah (1999), Li and Racine (2006) and Ahamada and Flachaire (2011).

A popular choice for the kernel is the standard Normal distribution, with expectation zero and standard deviation one, $K(t) = (2\pi)^{-1/2} \exp(-0.5\, t^2)$. The *standard Gaussian kernel density estimator* is equal to:

$$\begin{aligned} \hat{f}_X(x) &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - X_i}{h}\right)^2\right], & (2) \\ &= \frac{1}{n} \sum_{i=1}^{n} \phi(x; X_i, h). & (3) \end{aligned}$$

It is a sum of 'bumps' defined by Normal distributions, ϕ, with expectations X_i and a fixed standard deviation h.

The question of which value of h is the most appropriate is particularly a thorny one, even if automatic bandwidth selection procedures are often used in practice. Silverman's rule of thumb is mostly used, which is defined as follows:[7]

$$\hat{h}_{opt} = 0.9 \min\left(\hat{\sigma}\,;\, \frac{\hat{q}_3 - \hat{q}_1}{1.349}\right) n^{-\frac{1}{5}}, \tag{4}$$

where $\hat{\sigma}$ is the standard deviation of the data, and $\hat{q}_3$ and $\hat{q}_1$ are respectively the third and first quartiles calculated from the data. This rule boils down to using the minimum of two estimated measures of dispersion: the variance, which is sensitive to outliers, and the interquartile range. It is derived from the minimization of an approximation of the mean integrated squared error (MISE), a measure of discrepancy between the estimated and the true densities, where the Gaussian distribution is used as a reference distribution. This rule works well in numerous cases. Nonetheless, it tends to over-smooth the distribution when the true density is far from the Gaussian distribution, as multimodal and highly skewed.

Several other data-driven methods for selecting the bandwidth have been developed. Rather than using a Gaussian reference distribution in the approximation of the MISE, the *plug-in* approach consists of using a prior non-parametric estimate, and then choosing the h that minimizes this function. This choice of bandwidth does not then produce an empirical rule as simple as that proposed by Silverman, as it requires numerical calculation. For more details, see Sheather and Jones (1991).

Rather than minimizing the MISE, the underlying idea of *cross-validation by least squares* is to minimize the integrated squared error (ISE). Let $\hat{f}_{-i}$ be the estimator of the density based on the sample containing all of the observations except for y_i. The minimization of the ISE criterion requires us to minimize the following expression:

$$\mathrm{CV}(h) = \int \hat{f}^2(y)\,\mathrm{d}y - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(y_i).$$

This method is also called *unbiased cross-validation*, as $\mathrm{CV}(h) + \int f^2 \mathrm{d}y$ is an unbiased estimator of MISE. The value of h which minimizes this expression converges asymptotically to the value that minimizes the MISE (see Stone 1974, Rudemo 1982, Bowman 1984).

[7] See equation (3.31), page 48, in Silverman (1986).

2.2 Adaptive kernel density estimation

If the concentration of the data is markedly heterogeneous in the sample then the standard approach, with fixed bandwidth, is known to often oversmooth in parts of the distribution where the data are dense and undersmooth where the data are sparse. There would advantages to use narrower bandwidth in dense parts of the distribution (the middle) and wider ones in the more sparse parts (the tails). 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 (Portnoy and Koenker 1989). It is a sum of 'bumps' defined by Normal distributions, ϕ, with expectations X_i and varying standard deviations $h\lambda_i$.

A *pilot* estimate of the density at point y_i, denoted by $\tilde{f}(y_i)$, is used to measure the concentration of the data around this point: a higher value of $\tilde{f}(y_i)$ denotes a greater concentration of data, while smaller values indicate lighter concentrations. The parameter λ_i can thus be defined as being inversely proportional to this estimated value: $\lambda_i = [g/\tilde{f}(y_i)]^\theta$, where g is the geometric mean of $\tilde{f}(y_i)$ and θ is a parameter that takes on values between 0 and 1.[8] The parameter λ_i is smaller when the density is greater (notably towards the middle of the distribution), and larger when the density is lighter (in the tails of the distribution).

2.3 Log-transform kernel density estimation

It is also possible to estimate the underlying density of a sample, by using a preliminary transformation of the data, and obtaining the density estimate of the original sample by back-transformation. Let us consider a random variable X and define Y such that $Y = G(X)$, where G is a monotonically strictly increasing function. The underlying density functions are, respectively, f_X and f_Y. By the change of variable formula, we have

$$f_X(x) = f_Y[G(x)] \, . \, G'(x), \tag{5}$$

where $G'(x)$ is the first-derivative of G. An estimation of the density of the original sample is then obtained by back-transformation, replacing $f_Y(.)$ in (5) by a consistent estimator $\hat{f}_Y(.)$.

[8] In practice, an initial fixed-bandwidth kernel estimator can be employed as $\tilde{f}(y_i)$, with $\theta = 1/2$ and λ obtained with Silverman's rule of thumb.

In this section, since we have a density on $[0, +\infty)$ (and therefore positive observations), we consider the special case of the logarithmic transformation function, $Y = G(X) = \log X$. If the density of the transformed data f_Y is estimated with the Gaussian kernel density estimator, defined in (2), then the *log-transform kernel density estimation* is given by:

$$\hat{f}_X(x) = \hat{f}_Y(\log x)\ \frac{1}{x} = \frac{1}{nh}\sum_{i=1}^{n}\phi\left(\frac{\log x - \log X_i}{h}\right)\ \frac{1}{x} \tag{6}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{xh\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\log x - \log X_i}{h}\right)^2\right] \tag{7}$$

$$= \frac{1}{n}\sum_{i=1}^{n} Ln(x; \log X_i, h) \tag{8}$$

The bandwidth h can be selected with the Silverman's rule of thumb, the plug-in method of Sheather and Jones, or the cross-validation method, applied to the transformed sample $\log X_1, \ldots, \log X_n$.

It is worth noticing that (8) is a sum of 'bumps' defined by Lognormal distributions, Ln, with medians X_i and variances $(e^{h^2}-1)e^{h^2}X_i^2$. The kernel density estimation based on a log-transformation of the data is then similar to use a lognormal kernel density estimation on the original data. Note that the dispersion of the 'bumps' varies: it increases as X_i increases. In some way it can be viewed as an adaptive kernel method.

3 The bias and variance

With nonnegative data, the support of x is bounded to the left: $x \in [0, +\infty)$. A problem encountered by standard and adaptive kernel methods is that they may put positive mass to some values outside the support. Indeed, when smallest values are close to the lower bound zero, the 'bumps' placed at those observations can cross over the bound and, then, significant positive mass can be assigned to some negative values. A simple solution would be to ignore the boundary condition and to set $\hat{f}(x)$ to zero for negative x. However, the density estimate would no longer integrates to unity over the support $[0, +\infty)$ and this would cause a bias in the boundary region.[9]

The log-transform kernel density estimation does not encountered such problem, it integrates to unity. Observe that $\hat{f}_X$ is a density since it is

[9] And using a multiplicative factor to insure that the density integrates to one will increase the probability to have large values.

positive (as a sum of positive terms) and integrates to one, with a change of variable, $y = \log x$, so that $\mathrm{d}y = \mathrm{d}x/x$, from (6) we have:

$$\int_0^\infty \hat{f}_X(x)\,\mathrm{d}x = \int_0^\infty \hat{f}_Y(\log x)\frac{\mathrm{d}x}{x} = \int_{-\infty}^\infty \hat{f}_Y(y)\,\mathrm{d}y = 1. \tag{9}$$

Since the log-transform kernel is a sum of lognormal distributions, and lognormal distributions are defined over the the nonnegative support only, it is also clear from (8) that it integrates to unity.

The behavior of the log-transform kernel density estimation in the neighborhood of 0 will depend on the behavior of $f_X(x)$ (and its derivatives) in the neighborhood of 0. Since $\hat{f}_Y(\epsilon)$ is estimated with a standard Gaussian kernel estimator, from a Taylor expansion, we have

$$\mathrm{bias}\{\hat{f}_Y(\epsilon)\} = \mathbb{E}[\hat{f}_Y(\epsilon)] - f_Y(\epsilon) \sim \frac{h^2}{2} f_Y''(\epsilon) \tag{10}$$

see Silverman (1986, p.39). From (6) and (10), we have:[10]

$$\mathbb{E}[\hat{f}_X(\varepsilon)] = \frac{1}{\varepsilon}\,\mathbb{E}[\hat{f}_Y(\log\varepsilon)] \sim \frac{1}{\varepsilon}\left(f_Y(\log\varepsilon) + \frac{h^2}{2} f_Y''(\log\varepsilon)\right) \tag{11}$$

Since $f_Y(\log x) = x \cdot f_X(x)$, it follows that

$$\mathbb{E}[\hat{f}_X(\varepsilon)] \sim f_X(\varepsilon) + \frac{h^2}{2\varepsilon} f_Y''(\log\varepsilon) \tag{12}$$

By deriving twice $f_Y(y) = e^y \cdot f_X(e^y)$ with respect to y and replacing y by $\log x$, we obtain:[11]

$$f_Y''(\log\varepsilon) = \varepsilon \cdot f_X(\varepsilon) + 3\,\varepsilon^2 \cdot f_X'(\varepsilon) + \varepsilon^3 \cdot f_X''(\varepsilon) \tag{13}$$

Finally, replacing (13) in (12) gives

$$\mathrm{bias}\{\hat{f}_X(\varepsilon)\} \sim \frac{h^2}{2}\left(f_X(\varepsilon) + 3\varepsilon \cdot f_X'(\varepsilon) + \varepsilon^2 \cdot f_X''(\varepsilon)\right) \tag{14}$$

When the underlying density is zero at the boundary, $f_X(0) = 0$, putting $\varepsilon = 0$ in (14) shows clearly that the bias is zero. The log-transform kernel density estimation is then free of boundary bias. However, when $f_X(0) > 0$

[10] The relationship can be related to the Equation at the end of Section 2 in Marron and Ruppert (1994): $\mathbb{E}[\hat{f}_X(x) - f_X(x)] = \frac{1}{x}(\mathbb{E}[\hat{f}_Y(\log x) - f_Y(\log x)])$.

[11] Replace $y = \log x$ in $f_Y''(y) = e^y f_X(e^y) + 3e^{2y} f_X'(e^y) + e^{3y} f_X''(e^y)$

and is finite, there might be a significant bias depending on the behavior of the first and second derivatives of f_X in the neighborhood of 0.

We now turn to the variance. Since $\hat{f}_Y(\epsilon)$ is estimated with a standard Gaussian kernel estimator, from a Taylor expansion, we have

$$\mathrm{Var}[\hat{f}_Y(\epsilon)] = \frac{1}{nh} f_Y(\epsilon) \int K^2(u)\,\mathrm{d}u \tag{15}$$

see Silverman (1986, p.40). For the log-transform kernel density estimator, we then have

$$\mathrm{Var}[\hat{f}_X(\varepsilon)] = \frac{1}{\varepsilon^2}\mathrm{Var}[\hat{f}_Y(\log\varepsilon)] \sim \frac{1}{\varepsilon^2}\left(\frac{1}{nh} f_Y(\log\varepsilon) \int K^2(u)\,\mathrm{d}u\right) \tag{16}$$

Finally, using $f_Y(\log\varepsilon) = \varepsilon f_X(\varepsilon)$ in (16), we obtain

$$\mathrm{Var}[\hat{f}_X(\varepsilon)] \sim \frac{1}{\varepsilon nh} f_X(\varepsilon) \int K^2(u)\,\mathrm{d}u \tag{17}$$

so that the variance of $\hat{f}_X(\varepsilon)$ is classically of order $(nh)^{-1}$, and is divided by ε, which would be large in the neighborhood of 0. Compared to the variance of standard kernel estimator (15), equation (17) suggests that the variance of the log-transform kernel estimator would be larger for values of ε close to zero (in the bottom of the distribution) and smaller for values of ε far from zero (in the top of the distribution).

It is important to note that the log-transform kernel may perform poorly when the underlying distribution is not equal to zero at the boundary. Indeed, when $f_X(0) \neq 0$, putting $\varepsilon = 0$ in (14) and (17) shows clearly that the bias can be significant and the variance huge. As illustrated by Silverman (1986, Fig. 2.13), a large spike in zero may appear in the estimate. In such case, Gamma kernel density estimation may be more appropriate, as suggested by the application of Chen (2000) to the Silverman's data, and the application of Bouezmarni and Scaillet (2005) to the Brazilian income distribution, which exhibits an accumulation of observed points near the zero boundary.[12]

To the opposite, when the distribution is equal to zero at the boundary, the log-transform kernel may perform well. It should be more efficient than

[12]The use of other asymmetric kernels - beta, inverse and reciprocal inverse gaussian distributions - may also be used. They have been developped in the literature with non-negative data to remove boundary bias near zero, see Chen (1999), Abadir and Lawford (2004), Scaillet (2004), Hagmann and Scaillet (2007), Bouezmarni and Rombouts (2010) and Kuruwita et al. (2010).

the Gamma kernel. Indeed, the variance of $\hat{f}_X(\varepsilon)$ is of order $(nh)^{-1}$ with the log-transform kernel (see above), while it is of order $(nh^2)^{-1}$ in the boundary area with Gamma kernel.[13] In addition, the log-transform kernel allows us to use standard bandwidth selection methods on the transformed sample, with well-known properties. With Gamma and other asymmetric kernels, bandwidth selection is often more problematic, there is no general rule-of-thumb bandwidth and cross-validation can be burdensome for large samples.

4 Finite sample performance

We now turn to the performance in finite samples of the kernel density estimation methods presented in the previous section.

4.1 Model design

In our experiments, data are generated from two unimodal and one bimodal distributions:

- Lognormal : $Ln(x; 0, \sigma)$, with $\sigma = 0.5, 0.75, 1$
- Singh-Maddala : $\text{SM}(x; 2.8, 0.193, q)$, with $q = 1.45, 1.07, 0.75$
- Mixture : $\frac{2}{5}\,\text{SM}(x; 2.8, 0.193, 1.7) + \frac{3}{5}\,\text{SM}(x; 5.8, 0.593, q)$, with $q = 0.7, 0.5, 0.36$, plotted in Figure 1. It is a mixture of two Singh-Maddala distributions.

In the upper tail, the Singh-Maddala distribution $\text{SM}(x; a, b, q)$, also known as the Burr XII distribution, behaves like a Pareto distribution with tail-index $\alpha = aq$. We select parameters such that the upper tail behaves like a Pareto distribution with α, respectively, close to 4, 3, 2. As σ increases and q decreases the upper tail of the distribution decays more slowly: we denote the three successive cases as moderately, mediumly and strongly heavy-tailed. This design has been used in Cowell and Flachaire (2015).

We consider several distributional estimation methods. We first consider standard kernel density estimation based on the original sample, $X_1, \ldots, X_n$, with the Silverman rule-of-thumb bandwidth (Ksil), the plug-in bandwidth of Sheather and Jones (Ksj) and the cross-validation bandwidth (Kcv). We also consider the adaptive kernel density estimation with a pilot density

[13] Chen (2000) shows that the variance of the Gamma kernel density estimator is of order $(nh^2)^{-1}$ in the boundary area and of order $(nh)^{-1}$ elsewhere.

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

obtained by standard kernel density estimation, with the Silverman rule-of-thumb bandwidth (AKsil), the plug-in bandwidth of Sheather and Jones (AKsj) and the cross-validation bandwidth (AKcv). Then, we consider the log-transform kernel density estimation based on the transformed sample, $\log X_1, \ldots, \log X_n$, with respectively, the Silverman rule-of-thumb bandwidth (LKsil), the plug-in bandwidth of Sheather and Jones (LKsj) and the cross-validation bandwidth (LKcv) obtained from the transformed sample.

The sample size is $n = 500$ and the number of experiments $R = 1000$.

4.2 Overall estimation

To asses the quality of the overall density estimation, we need to use a distance measure between the density estimation and the true density. Here we use the mean integrated absolute errors (MIAE) measure,[14]

$$\text{MIAE} = E\left(\int_{-\infty}^{+\infty} \left|\hat{f}(x) - f(x)\right| \, \mathrm{d}x\right). \tag{18}$$

[14]Another appropriate measure is the MISE$= E\left(\int_{-\infty}^{+\infty}[\hat{f}(x) - f(x)]^2 \, \mathrm{d}x\right)$, but it puts smaller weights to differences in the tails.

Table 1 shows the quality of the fit obtained with standard, adaptive and log-transform kernel density estimation methods. The results show that:

- The popular standard kernel density estimation method with the Silverman's rule of thumb bandwidth performs very poorly with bimodal and heavy-tailed distributions (MIAE=0.225, 0.266, 0.300).
- 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 (last column).

The MIAE criteria gives one specific picture of the quality of the fit: it is the *mean* of the IAE values obtained from each sample.[15] Boxplots of IAE values are presented in Figure 2: they provide information on the *median*, *skewness*, *dispersion* and *outliers* of IAE values. The median is the band inside the box. The first and third quartiles (q_1, q_3) are the bottom and the top of the box. The outlier detection is based on the interval $[\underline{b}; \overline{b}]$, where $\underline{b} = q_1 - 1.5\,\text{IQR}$, $\overline{b} = q_3 + 1.5\,\text{IQR}$ and $\text{IQR}= q_3 - q_1$ is the interquartile range. Any values that fall outside the interval $[\underline{b}; \overline{b}]$ are detected as outliers, they are plotted as individuals circles. The horizontal lines at the top and bottom of each boxplot correspond to the highest and smallest values that fall within the interval $[\underline{b}; \overline{b}]$, see Pearson (2005).[16]

The top plot in Figure 2 shows boxplots of IAE values for the lognormal distribution with moderate heavy-tail, the most favorable case in Table 1 (first row). Boxplots of log-transform kernel density estimation are closer to zero, while they provide quite similar dispersion, compared to boxplots of standard and adaptive kernel density estimation. Moreover, the cross-validation bandwidth selection exhibits more outliers than the Silverman rule of thumb and Plug-in bandwidths. These results suggest that log-transform kernel density estimation, with the Silverman and plug-in bandwidth selection, perform slightly better.

The bottom plot in Figure 2 shows boxplots of IAE values for the mixture of two Singh-Maddala distributions with strong heavy-tail, the least

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

[16] If all observations fall within $[\underline{b}; \overline{b}]$, the horizontal lines at the top and bottom of the boxplots correspond to the sample maximum and sample minimum. It is the default boxplot command in R.

favorable case in Table 1 (last row). As suggested in Table 1, the standard kernel method with the Silverman bandwidth performs poorly, with a boxplot far from zero. In addition, we can see that the standard kernel method with cross-validation bandwidth exhibits many huge outliers. Overall, these results suggest that the log-transform kernel density estimation, with the plug-in bandwidth selection, outperforms other methods.

4.3 Pointwise estimation

The approximate expressions derived for the bias and variance of the log-transform kernel density estimation at point x, given in (14) and (17), suggest that the log-transform kernel method exhibits smaller bias in the neigborhood of zero, compared to the standard kernel estimator, and smaller (larger) variance at the top (bottom) of the distribution, see the discussion in section 3.

To illustrate the bias and variance of pointwise kernel density estimation, Figure 3 shows boxplots, biases and variances of standard, adaptive and log-transform kernel estimation at points $x = 0.01, 0.02, \ldots, 3$, for the worst case in Table 1, that is, with a mixture of Singh-Maddala with strong heavy-tail. The bandwidth is obtained with the plug-in method of Shether and Jones. By comparing the boxplots of standard (top left plot) and log-transform (top right plot) kernel methods, we can see that:

- The standard kernel method exhibits significant biases in the bottom of the distribution (boxes are far from the line of the true density), not the log-transform kernel method.
- Compared to the standard kernel method, the log-transform kernel method exhibits larger variances in the bottom of the distribution and smaller variances in the top.

These results are confirmed by the plots of biases (bottom left plot) and of variances (bottom right plot). It appears that the log-transform kernel method fits the upper tail much better. Figure 4 shows results for the more favourable case in Table 1, that is, with a lognormal distribution with moderate heavy-tail. The same features are observed, even if less pronounced.

5 Application

As an empirical study, we estimate the density of the income distribution in the UK in 1973. The data are from the family expenditure survey (FES),

a continuous survey of samples of the UK population living in households. The data are made available by the data archive at the University of Essex: Department of Employment, Statistics Division. We take disposable household income (i.e., post-tax and transfer income) before housing costs, divide household income by an adult-equivalence scale defined by McClements, and exclude the self-employed, as recommended by the methodological review produced by the Department of Social Security (1996). To restrict the study to relative effects, the data are normalized by the arithmetic mean of the year. For a description of the data and equivalent scale, see the annual report produced by the Department of Social Security (1998). The number of observations is large, $n = 6968$.

With these data, Marron and Schmitz (1992) showed that a nonparametric estimation of the income distribution in the United Kingdom produced a bi-modal distribution, which was not taken into account in preceding work which had used parametric techniques to estimate this same density.

Figure 5 presents the results from the estimation of the UK income distribution in 1973 with standard, adaptive and log-transform kernel method. As a benchmark, we plot a histogram with many bins, since we have a large number of observations.

The top plot shows results for the standard kernel estimation methods (see section 2.1). The value of the bandwidth obtained with the Silverman rule of thumb (Ksil) is equal to $h = 0.08559$: it allows us to reproduce the results in Marron and Schmitz (1992). We also plot the results with the plug-in bandwidth of Sheather and Jones (Ksj) and with the cross-validation bandwidth (Kcv). The comparison of the three estimators reveals that the results differ significantly. With the Silverman rule of thumb, the first mode is smaller than the second, while in the two other cases the reverse holds. Clearly, the kernel density estimation with the Silverman rule of thumb bandwidth fails to fit appropriately the underlying density function. The cross-validation or plug-in bandwidths give better results, as expected from our simulation study (see section 4.2).

The middle plot shows adaptive kernel density estimation methods, based on three different preliminary *pilot* density estimates: based on the Silverman rule of thumb bandwidth (AKsil), the cross-validation bandwidth (AKcv) and the plug-in bandwidth of Shether and jones (AKsj). The results are not very different from the standard kernel density estimation (top plot) except that the first mode is slightly higher and the estimation is more volatile on the second mode.

The bottom plot shows log-transform kernel density estimation methods, based on the three different bandwidth (LKsil, LKsj, LKcv). It is difficult

to distinguish by eyes the three lines, but the plug-in bandwidth provides a slightly higher first mode. Compared to standard and adaptive kernel methods, the estimation is smoothed everywhere and a small bumps is captured at the extreme bottom of the distribution. With respect to the histogram, used as benchmark, it appears that the log-transform kernel density estimation provides better results than the standard and adaptive kernel density estimation methods.

Finally, new features of the income distribution in the UK in 1973 are exhibited with the log-transform kernel density estimation. Compared to the initial estimation of Marron and Schmitz (1992), the main part of the income distribution is bimodal, but the first mode is higher than the second mode, and, a small group of very poor people appears in the bottom of the distribution.

We have estimated the income distribution in the UK for several years, from 1966 to 1999, with the FES dataset. We obtain similar results: the log-transform kernel density estimation provides better fit of the distribution, compared to standard and adaptive kernel methods, with the histogram used as benchmark.

6 Conclusion

With heavy-tailed distributions, kernel density estimation based on a preliminary logarithmic transformation of the data seems appealing, since kernel estimation may then be applied to a distribution which is no longer heavy-tailed.

We have seen that Gaussian kernel density estimation applied to the log-transformed sample is equivalent to use Lognormal kernel density estimation on the original data. We have then derived the bias and variance of the log-transform kernel density estimation at one point. It leads us to show that the method behaves correctly at the boundary if the underlying distribution is equal to zero at the boundary. Otherwise a significant bias and a huge variance may appear.

At first sight, our simulation study shows that using a preliminary logarithmic transformation of the data can greatly improve the quality of the density estimation, compared to standard and adaptive kernel methods applied to the original data. It is clear from our simulation study based on a measure of discrepancy between the estimated and the true densities over all the positive support. Studying the bias and variance at pointwise estimation, we show that the log-transform kernel density estimation exhibits

smaller bias in the bottom of the distribution, but the variance is larger. Our simulation results show that the top of the distribution is much better fitted with a preliminary log-transformation.

In our application, the use of a histogram as benchmark and a visual inspection help us to show that the log-transform kernel density estimation performs better than other kernel methods. It provides new features of the income distribution in the UK in 1973. In particular, the presence of a small group of very poor individuals is not captured by standard and adaptive kernel methods.

References

Abadir, K. M. and A. Cornea (2014). Link of moments before and after transformations, with an application to resampling from fat-tailed distributions. Paper presented at the 2nd Conference of the International Society for NonParametric Statistics, Cadiz, Spain.

Abadir, K. M. and S. Lawford (2004). Optimal asymmetric kernels. *Economics Letters 83*, 61–68.

Ahamada, I. and E. Flachaire (2011). *Non-Parametric Econometrics*. Oxford University Press.

Bouezmarni, T. and J. V. K. Rombouts (2010). Nonparametric density estimation for multivariate bounded data. *Journal of Statistical Planning and Inference 140*, 139–152.

Bouezmarni, T. and O. Scaillet (2005). Consistency of asymmetric kernel density estimators and smoothed histograms with application to income data. *Econometric Theory 21*, 390–412.

Bowman, A. (1984). An alternative method of cross-validation for the smoothing of kernel density estimates. *Biometrika 71*, 353–360.

Bowman, A. W. and A. Azzalini (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press.

Buch-Larsen, T., J. P. Nielsen, M. Guillén, and C. Bolancé (2005). Kernel density estimation for heavy-tailed distribution using the champernowne transformation. *Statistics 39*, 503–518.

Charpentier, A. and A. Oulidi (2010). Beta kernel quantile estimators of heavy-tailed loss distributions. *Statistics and Computing 20*, 35–55.

Chen, S. X. (1999). Beta kernel estimators for density functions. *Computational Statistics & Data Analysis 31*, 131–145.

Chen, S. X. (2000). Probability density function estimation using Gamma kernels. *Annals of the Institute of Statistical Mathematics 52*, 471–480.

Cowell, F. A. and E. Flachaire (2007). Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics 141*, 1044–1072.

Cowell, F. A. and E. Flachaire (2015). Statistical methods for distributional analysis. In A. B. Atkinson and F. Bourguignon (Eds.), *Handbook of Income Distribution*, Volume 2A, Chapter 6. New York: Elsevier Science B. V.

Davidson, R. (2012). Statistical inference in the presence of heavy tails. *The Econometrics Journal 15*, C31–C53.

Davidson, R. and E. Flachaire (2007). Asymptotic and bootstrap inference for inequality and poverty measures. *Journal of Econometrics 141*, 141 – 166.

Department of Social Security (1996). Households Below Average Income: Methodological Review Report of a Working Group. Corporate Document Services, London.

Department of Social Security (1998). Households Below Average Income 1979-1996/7. Corporate Document Services, London.

Devroye, L. and L. Györfi (1985). *Nonparametric Density Estimation: The L_1 View*. Wiley Series in Probability and Mathematical Statistics.

Hagmann, M. and O. Scaillet (2007). Local multiplicative bias correction for asymmetric kernel density estimators. *Journal of Econometrics 141*, 213–249.

Härdle, W. (1989). *Applied Nonparametric Regression*. Econometric Society Monograph, Cambridge.

Kleiber, C. and S. Kotz (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Hoboken. N.J.: John Wiley.

Kuruwita, C. N., K. B. Kulasekera, and W. J. Padgett (2010). Density estimation using asymmetric kernels and bayes bandwidths with censored data. *Journal of Statistical Planning and Inference 140*, 1765–1774.

Li, J. and J. S. Racine (2006). *Nonparametric Econometrics*. Princeton University Press.

Markovich, N. (2007). *Nonparametric Analysis of Univariate Heavy-Tailed Data*. Wiley Series in Probability and Statistics.

Marron, J. S. and D. Ruppert (1994). Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society. Series B 56*, 653–671.

Marron, J. S. and H. P. Schmitz (1992). Simultaneous density estimation of several income distributions. *Econometric Theory 8*, 476–488.

Pagan, A. and A. Ullah (1999). *Nonparametric Econometrics*. Cambridge: Cambridge University Press.

Pearson, R. K. (2005). *Mining Imperfect Data: Dealing with Contamination and Incomplete Records*. Society for Industrial and Applied Mathematics.

Portnoy, S. and R. Koenker (1989). Adaptive L estimation of linear models. *Annals of Statistics 17*, 362–381.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics 9*, 65–78.

Scaillet, O. (2004). Density estimation using inverse and reciprocal inverse Gaussian kernels. *Journal of Nonparametric Statistics 16*, 217–226.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New-York: John Wiley & Sons.

Sheather, S. J. and M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society 53*, 683–690.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. New-York: Springer Verlag.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society B 36*, 111–47.

Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing*. London: Chapman & Hall.

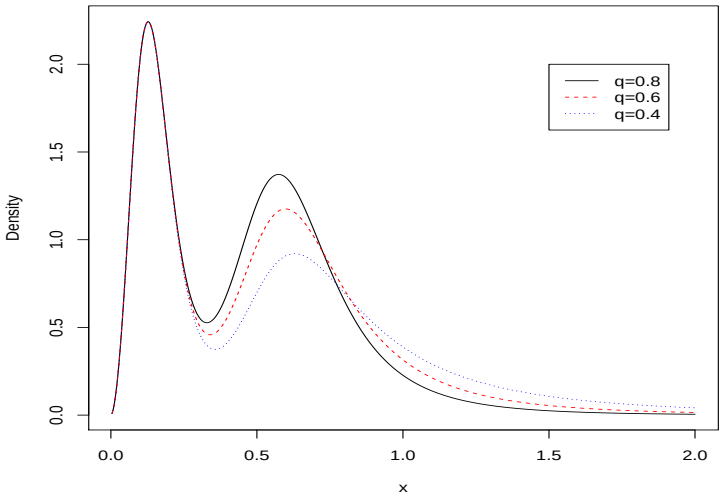

Figure 1: Mixture of two Singh-Maddala distributions

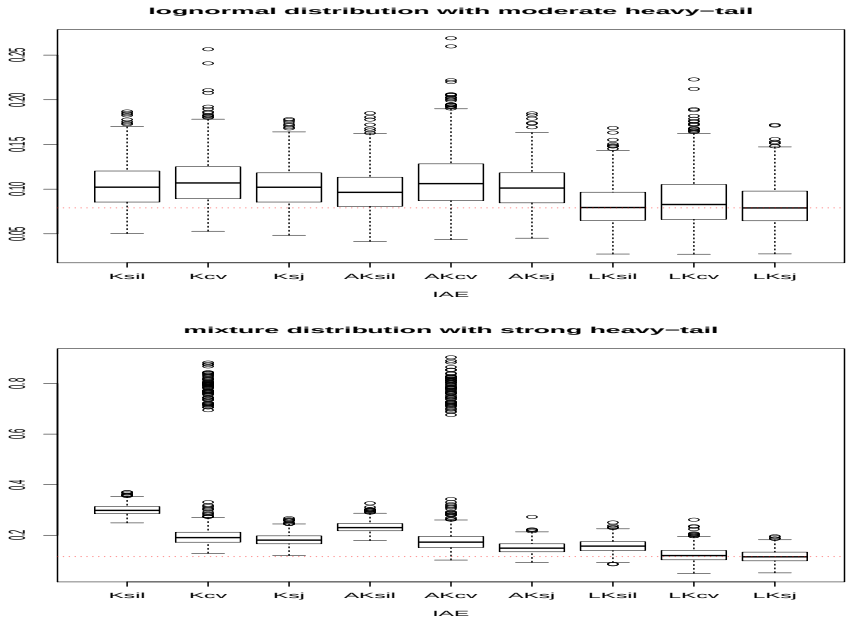

Figure 2: Boxplots of IAE values for the most and least favorable cases in Table 1 (first and last lines), that is, for the lognormal with moderate heavy-tail and for the mixture of two Singh-Maddala distributions with strong heavy-tail.

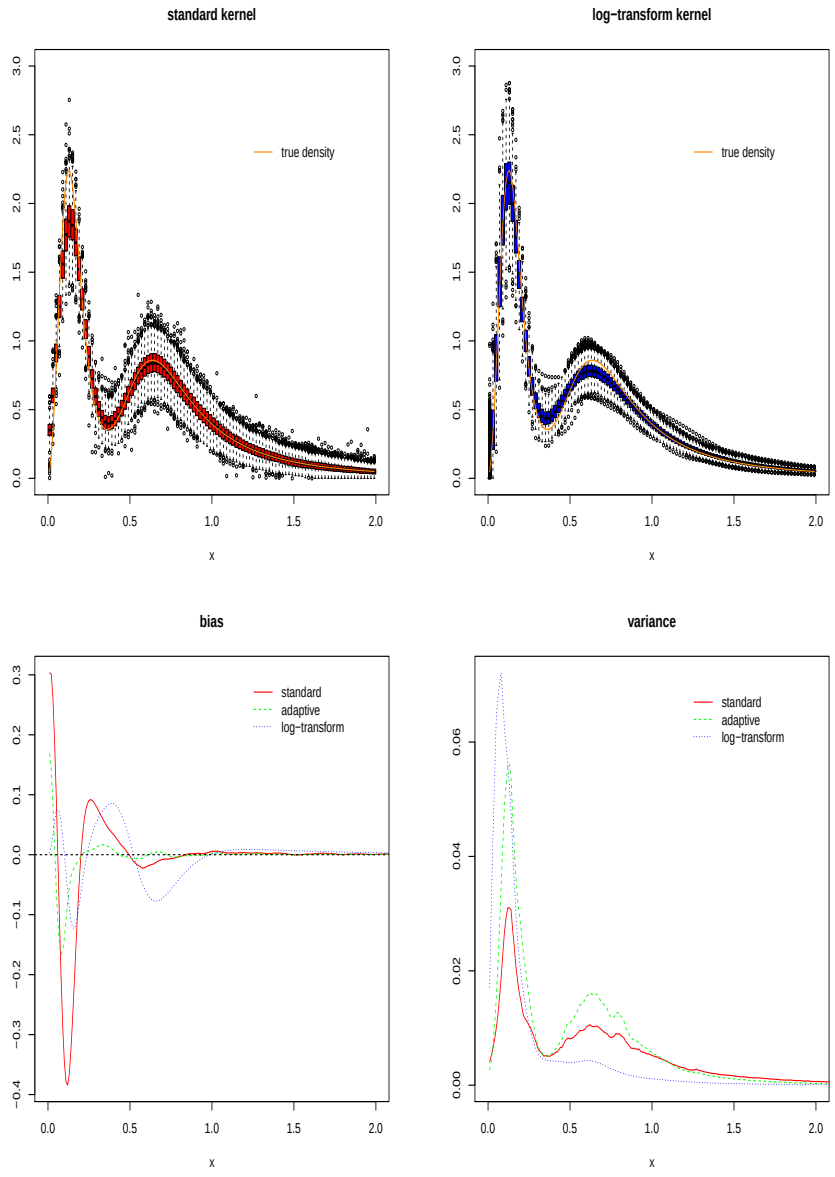

Figure 3: Pointwise estimation: boxplot, bias and variance of standard, adaptive and log-transform density estimation at point x for the less favourable case (mixture of Singh-Maddala with strong heavy-tail).

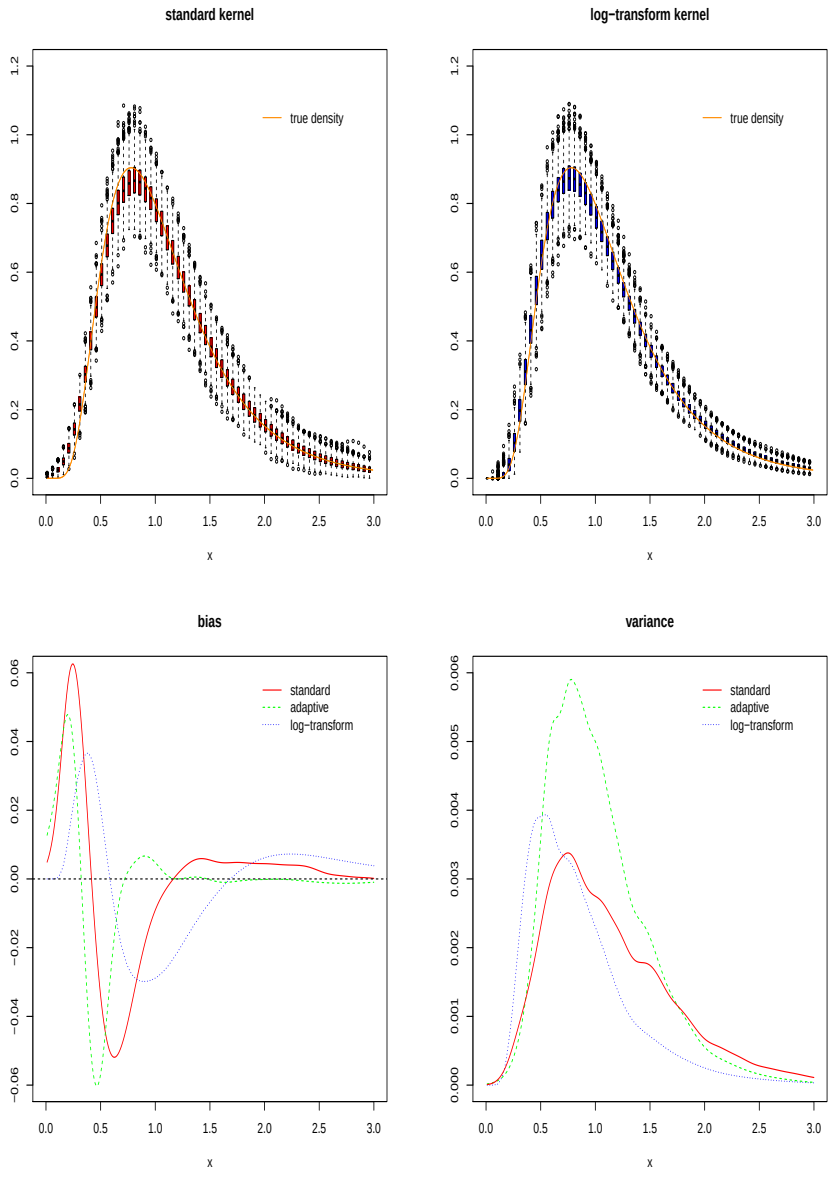

Figure 4: Pointwise estimation: boxplot, bias and variance of standard, adaptive and log-transform density estimation at point x for the most favourable case (lognormal with moderate heavy-tail).

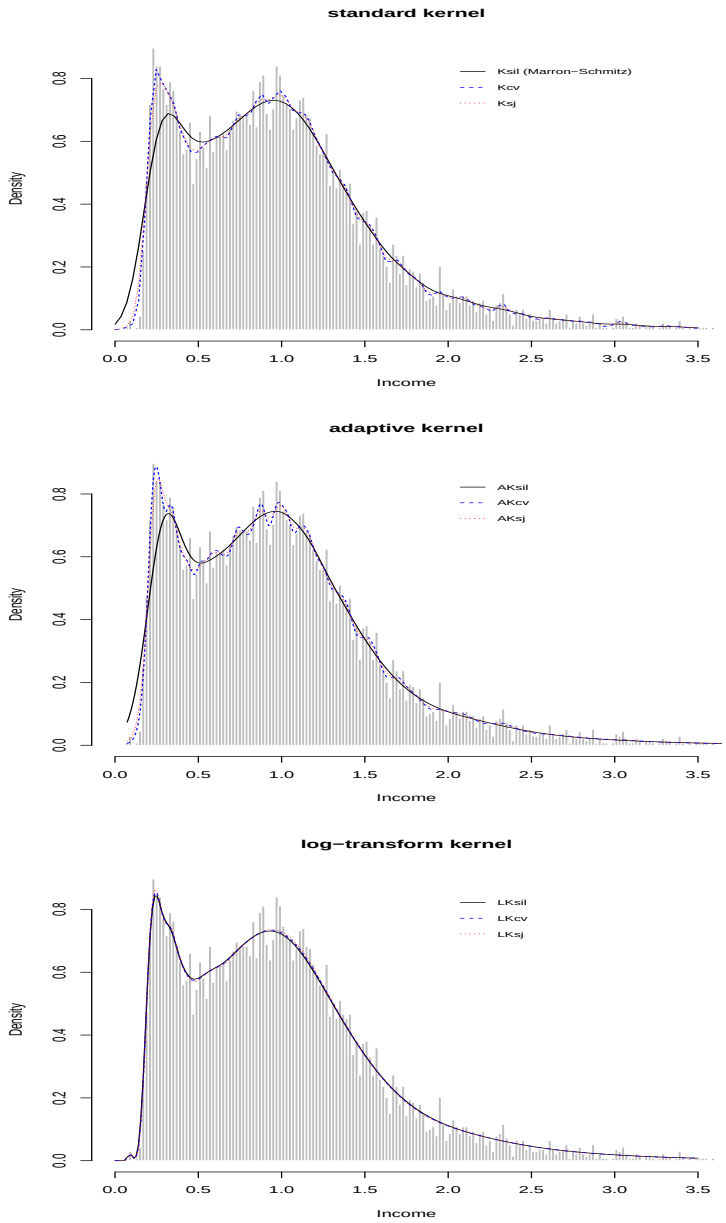

Figure 5: Standard, adaptive and log-transform kernel density estimation of income distribution in the United Kingdom (1973)