DOCUMENT DE TREBALL

XREAP2011-06

Estimation of Parametric and Nonparametric Models for Univariate Claim Severity Distributions - an approach using R

David Pitt
Montserrat Guillen (RFA-IREA)
Catalina Bolancé (RFA-IREA)

XARXA
DE REFERÈNCIA
EN ECONOMIA APLICADA

Estimation of Parametric and Nonparametric Models for Univariate Claim Severity Distributions - an approach using R

David Pitt[1], Montserrat Guillen[2] and Catalina Bolancé[2]

[1] Department of Applied Finance and Actuarial Studies, Macquarie University
Sydney, New South Wales, 2109 Australia
E-mail: david.pitt@mq.edu.au

[2] Dept. Econometrics, RFA-IREA, University of Barcelona
Diagonal, 690, 08034 Barcelona, Spain
E-mail: mguillen@ub.edu

May 19, 2011

Abstract

This paper presents an analysis of motor vehicle insurance claims relating to vehicle damage and to associated medical expenses. We use univariate severity distributions estimated with parametric and non-parametric methods. The methods are implemented using the statistical package R. Parametric analysis is limited to estimation of normal and lognormal distributions for each of the two claim types. The nonparametric analysis presented involves kernel density estimation. We illustrate the benefits of applying transformations to data prior to employing kernel based methods. We use a log-transformation and an optimal transformation amongst a class of transformations that produces symmetry in the data. The central aim of this paper is to provide educators with material that can be used in the classroom to teach statistical estimation methods, goodness of fit analysis and importantly statistical computing in the context of insurance and risk management. To this end, we have included in the Appendix of this paper all the R code that has been used in the analysis so that readers, both students and educators, can fully explore the techniques described.

1 Introduction

Kernel estimation is an easy nonparametric method to analyze the distribution of a random variable, that unlike parametric models requires little assumptions. When analysing the cost of individual claims in non-life insurance, we often encounter right skewness, because there are lots of small claims while only a few claims have a very large cost. When there is large skewness there is little awareness that classical kernel estimation is not a good method for approximating the probability density function (pdf) and the cumulated distribution function (cdf). In this work we show how nonparametric estimation of the pdf for right skewed random variables can be done in practice. We show an example with a cost insurance data base and provide the R code that is necessary to implement this approach.

The purpose of the analysis presented here is to illustrate univariate density estimation procedures using both parametric and non-parametric methods and to provide educators in insurance and risk analysis with a fully worked example of this form of data analysis using the statistical package R. We only consider the estimation of separate univariate models for two sets of positive insurance claims data. Bivariate analysis of these data, including estimation of correlations between claim cost types have been considered by [13] and [5] where bivariate skew-normal and bivariate normal distributions were fitted. Given that real claim severity data are usually positive and right-skewed, [5] also fitted the bivariate lognormal and log-skew-normal distributions along with a bivariate kernel density estimate.

Density estimation is necessary in insurance for many reasons including pricing and optimal capital allocation (see [9], [15], [10] and [27]). The book by [12] provides a comprehensive reference on the estimation of univariate and bivariate claims distribution models in insurance. In [14] an overview on risk measures for loss distributions is provided.

We study two positive cost claims data from a major Spanish motor insurer, namely property damage mainly resulting from third party liability and medical expenses that are not covered by the public health system. To obtain all the results in this paper we use the software R and the QRM library (see [14]). The claim amounts in the original data set were expressed in thousands of pesetas. To express these in thousands of Euros we used the standard conversion and divided by 166,386.

Next, in section 2 we describe the kernel density estimator and the transformed kernel density estimator. In section 3 we present different measures of goodness of fit, for parametric and nonparametric estimations. Then, in section 4 we describe the data set used in our application. Finally, we present the results and conclusions. The R programs used to obtain results are showed in the Appendix.

2 Kernel density estimation

2.1 Classical kernel density estimation

For a random sample of n independent and identically distributed observations $x_1, ..., x_n$ of a random variable X with pdf f_X, the kernel density estimator is

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^{n} K_b(x - x_i), \tag{1}$$

where $K_b(\cdot) = \frac{1}{b} K(\cdot / b)$, K is the kernel function and b is the bandwidth (see [25]). The bandwidth parameter is used to control the amount of smoothing in the estimation so that the greater b is, the smoother is the estimated density curve. The kernel function is usually a symmetric density with zero mean. In our work we use a Gaussian kernel, that is $K(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)$.

Many methods have been proposed for the selection of the bandwidth parameter in kernel density estimation. In program R the "rule of thumb" of Silverman (see [20], Chapter 3) is implemented while inter-quantile range as dispersion parameter. Unbiased and biased cross-validation methods and the plug-in method proposed by Sheather and Jones (see [19]) are also available in the library. We use all these methods and we select the one which represents better our pdf, if we compare it with the histogram. Note that when we use unbiased and biased cross-validation methods with skewed data, we can have problems to obtain a value for b. When data have right skewness these methods there may not be a global minimum and the value of b may be at the boundary of the grid.

2.2 Transformations and kernel density estimation

Classical kernel density estimation does not perform well when the true density is asymmetric. For instance, when one is interested in the density of the claim cost variable, the presence of many small claims produces a concentration of the mass near the low values of the domain and the presence of some very large claims causes positive skewness.

The lack of information in the right tail of the domain makes it difficult to obtain a reliable nonparametric estimate of the density in that area. Many authors have worked with heavy-tailed distributions and have adapted kernel estimation methods to this context. Different papers have proposed different transformed kernel estimation (TKE) methods of a pdf, based on parametric families (see [26], [8], [6], [7], [4], [2] and [1]).

Let $T(.)$ be an increasing and monotonic transformation function that has a first derivative $T'(.)$. If the true density is right-skewed, then the chosen transformation $T(.)$ must be a concave function. In this paper we propose a TKE of the pdf that consists of transforming the original data $y_i = T\left(x_i\right)$ so that the new transformed data can be assumed to have been generated from a symmetric random variable Y and hence the true density of the transformed variable can be reliably approximated using the classical kernel estimation method. Using a change of variable, once the kernel estimation is obtained for the transformed variable, estimation on the original scale is also obtained.

In [6] the authors proposed to select the transformation function from a transformation family proposed first in [26] named shifted power transformation family,

$$T_\lambda\left(x\right) = \begin{cases} \left(x+\lambda_1\right)^{\lambda_2} sign\left(\lambda_2\right) \\ \ln\left(x+\lambda_1\right) \end{cases}, \tag{2}$$

where $\lambda = (\lambda_1, \lambda_2)$, $\lambda_1 \geq -\min\left(x_i, i = 1, ..., n\right)$ and $\lambda_2 \leq 1$ for right-skewed data. This approach has a simple mathematical formulation and works particularly well for TKE of asymmetric distributions. In order to estimate the optimal parameters of the shifted power transformation function, the algorithm described in [6] can be used.

Let us assume a sample of n independent and identically distributed observations $x_1, ..., x_n$ is available. We also assume that a transformation function $T_\lambda(\cdot)$ has been selected so that the data can be transformed to give $y_i = T_\lambda(x_i)$, $i = 1, ..., n$. We denote the transformed sample by $y_1, ..., y_n$.

Having transformed the data, we then estimate the density of the transformed data set using the classical kernel density estimator

$$\widehat{f}_Y(y) = \frac{1}{n}\sum_{i=1}^{n} K_b\left(y - y_i\right),$$

where $K_b(\cdot) = \frac{1}{b}K(\cdot/b)$, K is the kernel function and b is the bandwidth. The estimator of the original density is obtained by back-transformation of $\widehat{f}_Y(y)$:

$$\widehat{f}_X(x,\lambda) = T'_\lambda\left(x\right)\widehat{f}_Y(y) = \frac{T'_\lambda\left(x\right)}{n}\sum_{i=1}^{N} K_b\left\{T_\lambda\left(x\right) - T_\lambda\left(x_i\right)\right\}, \tag{3}$$

where as we have mentioned we have assumed that the transformations are differentiable. The estimator defined in (3) is named transformed kernel density estimation.

In order to implement the transformation approach, a method to select the transformation parameters and the bandwidth is necessary.

2.3 Selecting the transformation parameters and the bandwidth

As in [6], we restrict the set of transformation parameters, $\lambda = (\lambda_1, \lambda_2)$, to those values that give approximately zero skewness for the transformed data $y_1, .., y_n$ (which have also been scaled to have the same variance as the original sample, see [26]).

We define skewness as:

$$\widehat{\gamma}_y = \left\{ n^{-1} \sum_{i=1}^{n} (y_i - \overline{y})^3 \right\} / \left\{ n^{-1} \sum_{i=1}^{n} (y_i - \overline{y})^2 \right\}^{\frac{3}{2}},$$

where $\overline{y}$ is the sample mean.

To select the λ parameter vector, we aim at minimizing the mean integrated square error (MISE) of $\widehat{f_Y}(y)$

$$MISE_Y\left(\widehat{f_Y}\right) = E\left[\int_{-\infty}^{+\infty} \left(\widehat{f_Y}(y) - f_Y(y)\right)^2 dy\right],$$

which, when b is asymptotically optimal, can be approximated by:

$$\frac{5}{4}\left[k_2 \alpha(K)^2\right]^{\frac{2}{5}} \beta\left(f_Y''\right)^{\frac{1}{5}} n^{-\frac{4}{5}}, \tag{4}$$

where $k_2 = \int t^2 K(t)\, dt$, $\alpha(K) = \int K(t)^2\, dt$ and $\beta(f_Y'') = \int_{-\infty}^{+\infty} [f_Y''(y)]^2\, dy$. Minimizing (4) with respect to the transformation parameters is equivalent to minimizing $\beta(f_Y'')$. The transformation parameters that minimize asymptotically $MISE_Y$ also minimize $MISE_X$ of $\widehat{f_X}(x, \lambda)$ in (3) (see [26]).

In [11] the following estimator for $\beta(f_Y'')$ is suggested:

$$\widehat{\beta}(f_Y'') = n^{-1}(n-1)^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c^{-5} K * K\left\{c^{-1}(y_i - y_j)\right\}, \tag{5}$$

where $K*K(t) = \int_{-\infty}^{+\infty} K(t-s)K(s)ds$ is the kernel convolution and c is the bandwidth used in the estimation of $\beta(f_Y'')$, which can be estimated by minimizing the mean square error (MSE) of $\widehat{\beta}(f_Y'')$. When it is assumed that f_Y is a normal density as in the "rule of thumb" approach, c can be estimated by $\widehat{c} = \widehat{\sigma}_y \left(\frac{21}{40\sqrt{2}n^2}\right)^{\frac{1}{13}}$, where $\widehat{\sigma}_y = \sqrt{n^{-1} \sum\limits_{i=1}^{n} (y_i - \overline{y})^2}$ (see [16] and [26]).

In our application we implement two strategies: we can obtain the transformation parameters by directly minimizing (5) and, alternatively, we can obtain a set of transformation parameter where skewness is zero and then search the transformation parameters that minimize (5) only within this set.

Finally, we need to make the selection of the bandwidth that is going to be used for the transformation. Here we simply use the "rule of thumb" developed by Silverman (see [20], Chapter 3, p. 45) for a standard normal density. Since our transformation aims at a transformed density with zero skewness, this approach seems very plausible. Following [20], the estimator of the bandwidth b is $\widehat{b} = 1.059\widehat{\sigma}_x n^{-\frac{1}{5}}$ and the corresponding transformation estimator will be denoted $\widehat{f_X}(x, \widehat{\lambda}; \widehat{b})$.

3 Measuring the goodness of fit

We are interested in evaluating the quality of our density estimates over the whole domain both for a parametric and a nonparametric setting. Let us begin with the log-likelihood function. However, this function is not a good criterion to evaluate the performance of non-parametric estimation. Log-likelihood depends on the values of the density exclusively in sample points. In kernel estimation the density shows bumps around isolated sample observations.

The parametric estimates that we will analyzed have been obtained using maximum likelihood estimation. We can compare the difference between the value of log-likelihood under several estimation alternatives, i.e. for any given pdf estimate, we can compute the sum of the logarithm of the estimated density in the sample points. This will provide with a straightforward measure of comparative goodness of fit.

Let us assume that we have $\widehat{f_X}(x)$, an estimate of the density for every point x in the domain. Let us also assume that a sample of n independent and identically distributed observations $x_1, ..., x_n$ is available. Then, we can estimate the log-likelihood function as:

$$\ln \hat{L}(\widehat{f_X}(\cdot); x_1, ..., x_n) = \sum_{i=1}^{n} \ln \widehat{f_X}(x_i).$$

If a transformation method were used as $\widehat{f_X}(x, \widehat{\lambda}; \widehat{b})$, in this case, the estimated log-likelihood function is

$$\ln \hat{L}(\widehat{f_X}(\cdot); T_{\widehat{\lambda}}(\cdot); x_1, ..., x_n) = \sum_{i=1}^{n} \ln \widehat{f_X}(x_i, \widehat{\lambda}; \widehat{b}).$$

A widely used measure for evaluating the quality of kernel density estimators over the whole domain is the integrated square error (ISE). Let $\widehat{f_X}(x)$ a kernel estimation of $f_X(x)$, then:

$$ISE_X\left(\widehat{f_X}\right) = \int_{-\infty}^{+\infty} \left(\widehat{f_X}(x) - f_X(x)\right)^2 dx.$$

The problem of ISE_X is that it depends on the true density f_X that is unknown. In [20] it is proved that minimizing ISE_X it is equal to minimizing the cross-validation function:

$$CV_X = \int_{-\infty}^{+\infty} \left[\widehat{f}_X(x)\right]^2 dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_i(x_i), \tag{6}$$

where $\widehat{f}_i$ is the "leave-one-out" estimation, that is the kernel estimation of f_X without observation x_i. We can obtain (6) for the transformed kernel density estimation, replacing $\widehat{f}_X(x)$ by $\widehat{f}_X(x, \widehat{\lambda}; \widehat{b})$ and $\widehat{f}_i(x_i)$ by $\widehat{f}_i(x_i, \widehat{\lambda}; \widehat{b})$.

We can generalize the definition of log-likelihood based earlier goodness of fit statistics by providing a statistic that gives more weight to the right tail of the distribution. This is important when we require our estimation to be more accurate in the upper right tail of the distribution. Also, we can generalize ISE_X and its approximation in (6) to a weighted ISE_X ($WISE_X$) that gives more weight to the right tail.

A weighted log-likelihood can be estimated if weights w_i, $i = 1, ..., n$ are included preceding each summation term as:

$$\ln_w \hat{L}(\widehat{f}_X(\cdot); x_1, ..., x_n) = \sum_{i=1}^{n} w_i \ln \widehat{f}_X(x_i).$$

If $w_i = 1$, $i = 1, ..., n$, then we would have the usual log-likelihood expression. We can also use some distance as a weight, so that observations that are located close to zero have much less importance than those located in the right tail.

We have tried two different expressions for weights. The first one is giving more weight to those observations that are distant from sample mean $\overline{x}$. Note that our data are always positive. The form of the weights is

$$w_i^{(1)} = \frac{n x_i}{\sum_{i=1}^{n} x_i}.$$

Using these weights in the estimated weighted log-likelihood implies that more importance is given to the fit in the right tail. Then, since for a given i, we have that $\ln \widehat{f}_X(x_i)$ is negative and it is smaller when $\widehat{f}_X(x_i)$ tends to zero (which is exactly what happens in the long tail region) then weighting those summation terms more, means that the $\ln_w \hat{L}(\widehat{f}_X(\cdot); x_1, ..., x_n) \leq \ln \hat{L}(\widehat{f}_X(\cdot); x_1, ..., x_n)$.

The second form for the weights considered is inspired by the same principle as the weighted integrated mean squared error that was proposed in [6], where contributions are weighted with a squared distance. In this case:

$$w_i^{(2)} = \frac{n x_i^2}{\sum_{i=1}^{n} x_i^2}.$$

When a transformation is used, the corresponding expression would be:

$$\ln_w \hat{L}(\widehat{f}_X(\cdot); T_{\widehat{\lambda}}(\cdot); x_1, ..., x_n) = \sum_{i=1}^{n} w_i \widehat{f}_X(x_i, \widehat{\lambda}; \widehat{b}).$$

Similarly, we can approximate a weighted ISE_X ($WISE_X$), weighting by x or by x^2:

$$WISE_X^1\left(\widehat{f}_X\right) = \int_{-\infty}^{+\infty} \left(\widehat{f}_X(x) - f_X(x)\right)^2 xdx$$

or

$$WISE_X^2\left(\widehat{f_X}\right) = \int_{-\infty}^{+\infty} \left(\widehat{f}_X(x) - f_X(x)\right)^2 x^2dx,$$

that cat be approximated with:

$$WCV_1 = \int_{-\infty}^{+\infty} \left[\widehat{f}_X(x)\right]^2 xdx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_i(x_i)x_i \tag{7}$$

or

$$WCV_2 = \int_{-\infty}^{+\infty} \left[\widehat{f}_X(x)\right]^2 x^2dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_i(x_i)x_i^2. \tag{8}$$

4 Data and results

The claims we considered relate to motor insurance policies of a major insurer in Spain for accidents that occurred in the year 2000. Data correspond to a cost of claims, expressed in thousands of Euros, in a random sample of all claims related to property damage expenses and to medical expenses.

Bodily injury is universally covered by the National Health System. This means that medical costs considered here are medical expenses that are not covered by the public system such as technical aids, drugs or chiropractic-related expenses. Those expenses have to be paid by the insurer. No compensation for pain and suffering or loss of income are included. Medical expenses contain medical costs related to all those who were injured in the accident. Property damage expenses include the insured's liability for damages he or she caused to vehicles, property or objects when the accident occurred.

The claims included in our sample are all claims that have already been settled. Although claims for compensation relating to bodily injury may take a long time to settle, these data were gathered in 2002, so that we can safely assume that there has been enough time for the claimant to include most costs, and we therefore consider that these are closed claims.

The sample size contains 518 claims, and for each claim we observe X_1 the cost of property damage and X_2 the cost of medical expenses, both expressed in thousands of Euros.

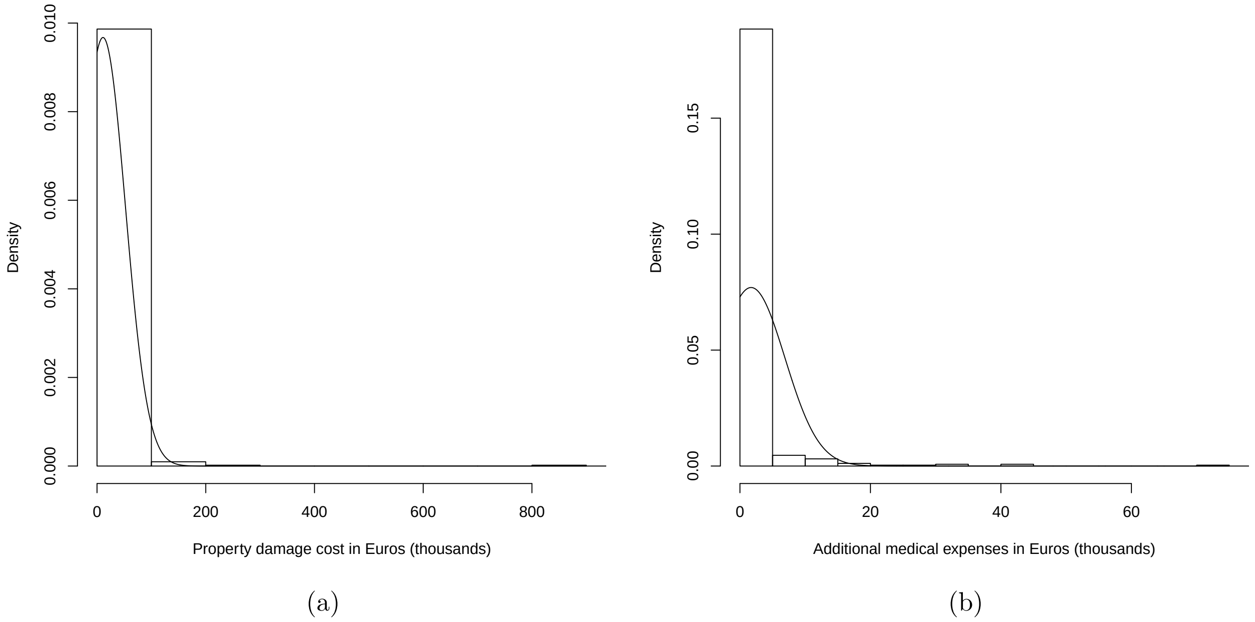
Table 1: Univariate descriptive statistics for the positive claims data set (in 1,000 Euros)

	Mean	Std. Dev.	Skewness	Kurtosis	Min	Max
X_1	10.984	41.276	15.652	297.142	0.078	829.012
X_2	1.706	5.188	8.037	82.019	0.006	71.250

4.1 Descriptive statistics and parametric fitting

In Table 1 we show the descriptives statistics obtained using the R commands in the ”Descriptive Statistics” subsection in the Appendix. The data set is called ”KEURcostes.txt”. In the R commands we assume that the observations of X_1 are located in keurcost[,1] and, correpondingly, X_2 is found in keurcost[,2].

We provide univariate histograms of the individual claim data for both components of the claim costs. These are shown in Figure 1. On each of these histograms we have overlaid a normal probability density function, estimated by maximum likelihood. The R commands to obtain Figure 1 are also shown in the ”Descriptive Statistics” subsection in the Appendix.



(a)

(b)

Figure 1: Histograms and Normal fit

From Figure 1 it becomes clear that a symmetric distribution, such as the normal, does not provide a good fit to these data. Much of the density under the fitted normal distribution relates to claim sizes smaller than the minimum observed claim value.

As the next step in the modeling, we investigate estimation using lognormal distribution. Equivalently, we investigate taking the log-transforms of each of the two components of our claim data set and fitting normal distributions to the resulting transformed data. Histograms of the log-transformed data with overlaid normal density functions are shown in Figure 2. The estimation is again conducted using R with the QRM library. The improvement in fit obtained using the lognormal distribution compared to the normal distribution is apparent in Figure 2.

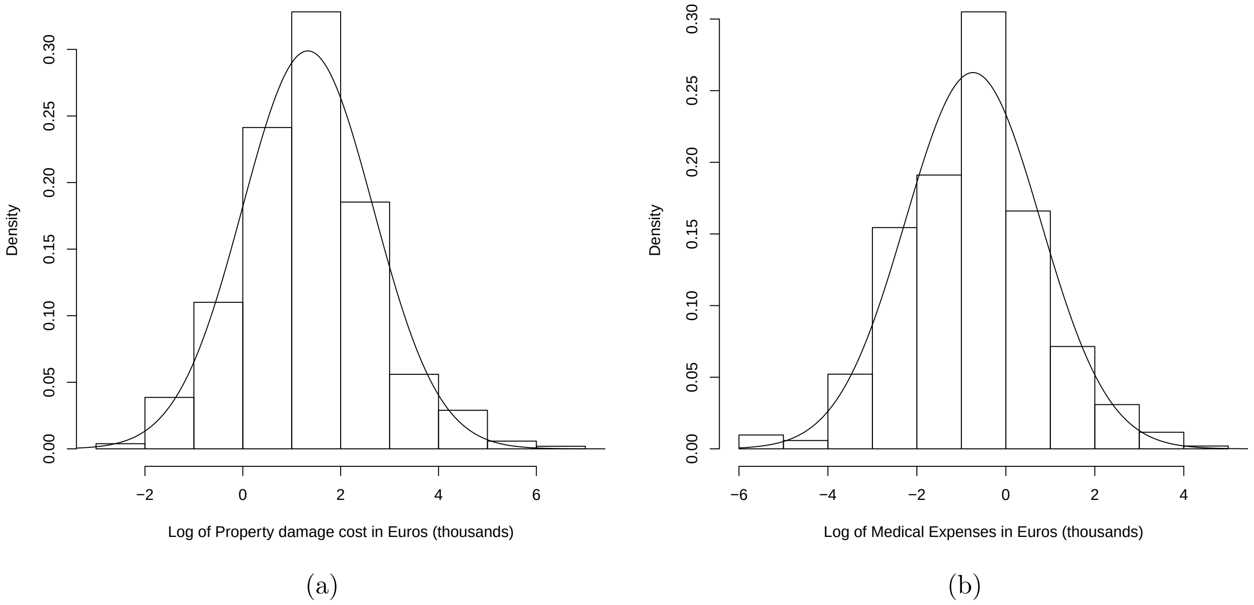


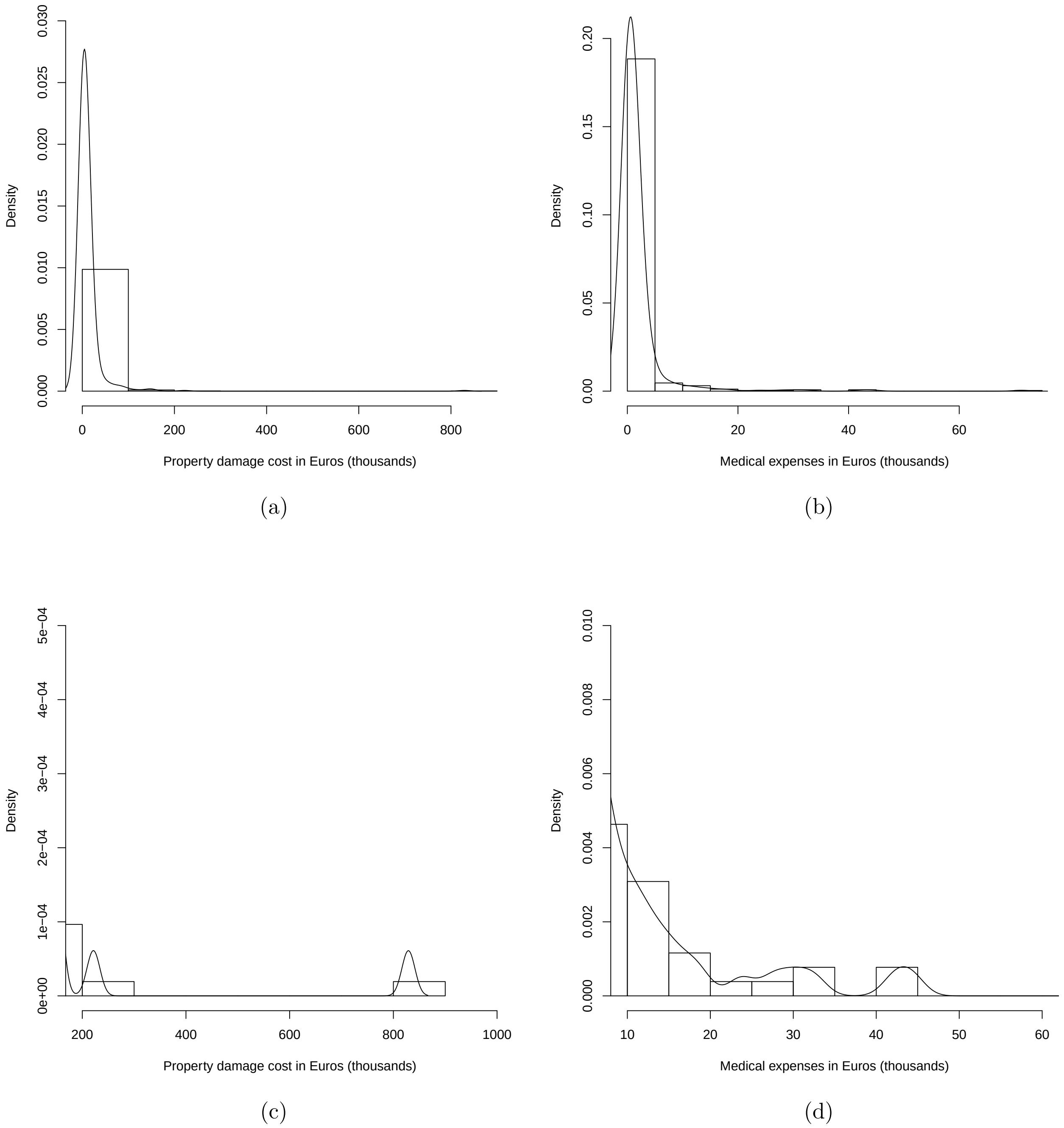
Figure 2: Histograms and lognormal fit

4.2 Results of nonparametric fitting

In this section we describe the results of kernel density estimation and different transformed kernel density estimation approaches to univariate claims data. Finally, we calculate the goodness of fit measure that we described in section 3 to compare the proposed estimations.

In Figures 3 and 4 we show the classical kernel density estimates and the log-transformation kernel estimates for both components of the univariate claims data: property damage and medical expenses. In Appendix we have provided the R commands used to obtain the density estimates shown in Figures 3 and 4. In program R the ”rule of thumb” proposed by [20] is implemented (bw=”nrd0” or bw=”nrd”), together with the unbiased and biased cross-validation methods (bw=”ucv” and bw=”bcv”) and the plug-in method proposed by [19] (bw=”sj”). A numeric value for the bandwidth b in classical kernel density estimation can be imposed in the R programme.

In Figure 3c and 3d we show the classical kernel density estimates in the right tail of the pdf. We note that the density does not have a smooth shape, as it has some bumps around the sample observations. In Figures 4a and 4b we can see that there is a considerable improvement in the kernel density estimate when it is applied to the log-transformed claim data compared to the fit obtained when applied to the untransformed data. Based on this fact, we further explore the optimal transformation to apply to our claim datasets.



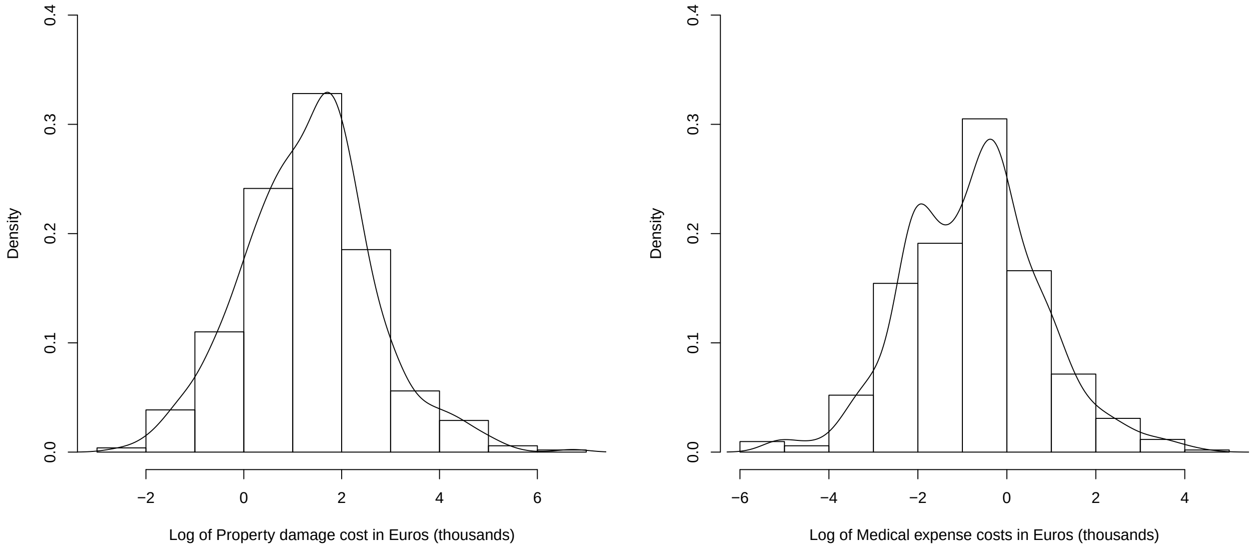
(a)

(b)

(c)

(d)

Figure 3: Histograms and classical kernel fit



(a) (b)

Figure 4: Histograms and classical kernel fit to log-transformed data

Optimal transformation parameters are obtained via expression (2). In section 2.3 we propose two forms to estimate the transformation parameters $\lambda = (\lambda_1, \lambda_2)$. The first is easier and only involves minimizing expression (5). We call this method "Method 1". The R code used to perform the minimization is shown in the Appendix. Note the use of the 'optim' function in R.

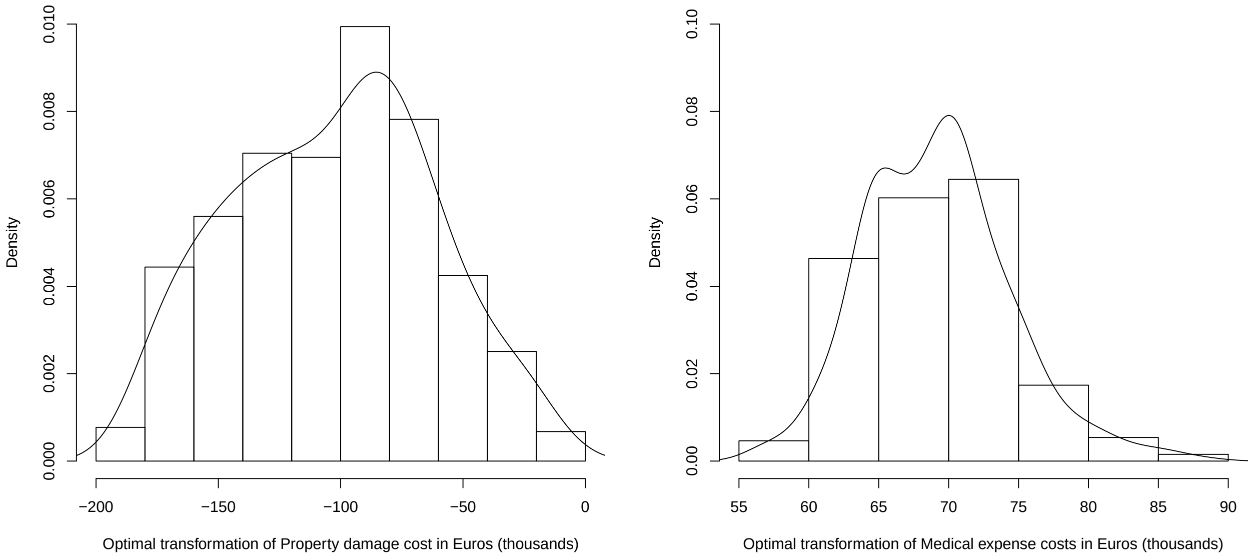
The second method which we propose to obtain the transformation parameters $\lambda = (\lambda_1, \lambda_2)$ needs to search within a set of transformation parameters that generate transformed variables with zero skewness, to look for the pair of parameters minimizing expression (5). We call this method "Method 2". The R code used to perform this algorithm is also given in the Appendix. Note the use of the 'optimize' function in R.

In Table 2 the transformation estimates of parameters $\lambda = (\lambda_1, \lambda_2)$ are shown. The results using the two methods are similar; in fact, asymptotically, the two methods converge, because the density that minimizes the functional $\beta\left(f''_Y\right)$ is symmetric (see [22] and [21]). The differences between Method 1 and Method 2 are caused because $\beta\left(f''_Y\right)$ is unknown and an estimation in expression (5) must be used. Note also that for X_2 the values of λ_1 and λ_2 are near zero, this indicates that the distribution associated to the cost of medical expenses is very similar to a lognormal distribution.

Table 2: Estimates of transformation parameters $\lambda = (\lambda_1, \lambda_2)$

	Method 1	Method 2
X_1	$(1.9931, -0.6201)$	$(1.870333, -0.5700)$
X_2	$(0.0219, -0.0054)$	$(0.0041, 0.0500)$

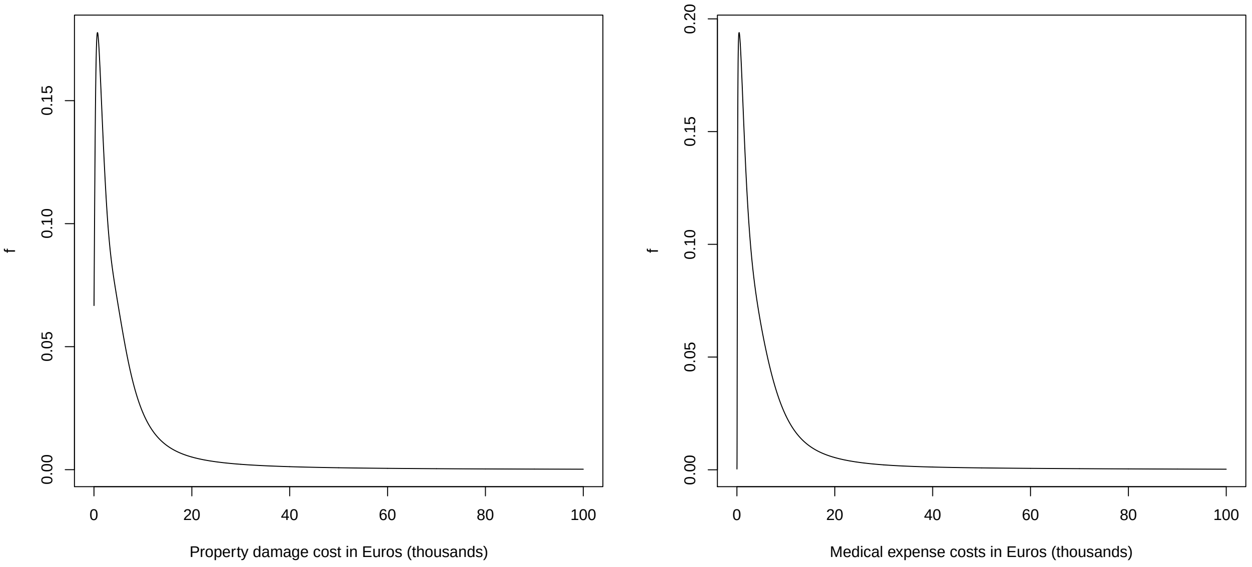
In Figure 5 we show the kernel density estimates of optimally transformed variable using Method 2 of both components of the univariate claims data. In Figure 6 we plot the TKE of pdf of property damage and medical expenses costs, we can see the smoothed shape of the pdf estimated in the right tail.



(a)

(b)

Figure 5: Histograms and classical kernel fit to optimally transformed data



(a)

(b)

Figure 6: TKE pdf estimate

4.3 Goodness of fit results

Following the discussion about goodness of fit in Section 3, we calculate log-likelihoods and two different weighted log-likelihoods for each of the estimated models. Then we approximate ISE_X using expression (6) and $WISE_X^1$ and $WISE_X^2$ using (7) and (8), respectively. In this way, we compare the nonparametric approaches. In order to store the results of these calculations in R, first we create three different R objects, namely lnL, w1lnL and w2lnL, and second we create ISE, w1ISE and w2ISE (see "Goodness of fit" subsection in the Appendix).

Note from the R code in Appendix, that lnL, w1lnL and w2lnL have six rows and two columns. The six rows correspond to the six different models considered: normal, lognormal, classical kernel, transformed kernel density estimation using log-transformation and finally the transformed kernel density estimation using optimally transformed, where the optimal transformation is found as discussed in Section 2.3, initially minimizing only expression (5) (Method 1), and second, searching the optimum within a set of transformation parameters, where the transformed variable has zero skewness (Method 2). In the Appendix, we only show the R code for property damage (keurcost[,1]), for medical expenses (keurcost[,2]) the programme is similar.

In Table 3 we show the results of log-likelihood and weighted log-likelihood for the six fitted densities for the damage cost (X_1) and below we show the same results for medical expenses cost (X_2). A higher value incates a better fit. We can see that normal distribution does not fit our data well; the two analyzed variables are right skewness and a symmetric distribution is not adequate. We can see that log-likelihood ($\ln \hat{L}$) results for lognormal and different TKE are similar. For classical kernel estimation, $\ln \hat{L}$ is clearly lower.

Results for weighted log-likelihood ($\ln_{w^{(1)}} \hat{L}$ and $\ln_{w^{(2)}} \hat{L}$) show that the classical kernel is the method that provides the best fit once the tail of the distribution gains importance with the use of weights. However this is a distorted result, because, as we can see in Figures 3c and 3d, classical kernel is not smooth in the tail, so the fitted density in this zone forms little bubbles around the observed data points. Because of lack of smoothness $\ln_{w^{(1)}} \hat{L}$ and $\ln_{w^{(2)}} \hat{L}$ do not provide a net goodness of fit measure for classical kernel.

The transformed kernel density estimation is smooth in the tail of the distribution. If we compare the values of $\ln_{w^{(1)}} \hat{L}$ and $\ln_{w^{(2)}} \hat{L}$ for TKE with the values for the lognormal fit, we observe that the three TKE work better. Moreover, when the true distribution departs from the lognormal distribution, as it occurs for damage costs X_1, the TKE with Method 2 has higher values for $\ln_{w^{(1)}} \hat{L}$ and $\ln_{w^{(2)}} \hat{L}$ and, therefore, the fit improves other alternatives.

Log-likelihood and weighted log-likelihood are not good measures for comparing nonparametric fits. In section 3 we proposed the use of CV, WCV_1 and WCV_2 to compare the fit of classical kernel estimation and TKE. The results for damage cost X_1 and medical expenses cost X_2 are found in Tables 4. The lower the value, the better the fit. Note that the values of CV, WCV_1 and WCV_2 can be negative. The minimum values of CV, WCV_1 and WCV_2 for damage cost X_1 are found for TKE with Method 1 and Method 2. For medical expenses cost X_2 these minimum values are found for TKE with log-transformation. So, we can conclude that when distribution is similar to lognormal, X_2 in our case, the TKE with log-transformation is sufficient.

5 Conclusions

In this paper we fitted univariate distributions to a real data set from motor insurance claims amounts.

The kernel estimation approach provides a smoothed version of the empirical distribution. We also provided details of goodness of fit criteria based on standard likelihood theory and also

Table 3: Log-likelihood and weighted log-likelihoods

Damage cost (X_1)

	$\ln \hat{L}$	$\ln_{w^{(1)}} \hat{L}$	$\ln_{w^{(2)}} \hat{L}$
Normal	-2661.62	-17924.79	-77309.75
Lognormal	-1573.23	-3837.87	-7303.22
Classical Kernel	-1998.40	-3249.17	-4770.71
Kernel Transformed (TKE with log transformation)	-1562.12	-3522.54	-5984.21
Kernel Transformed (TKE with Method 1)	-1574.85	-3556.95	-6002.81
Kernel Transformed (TKE with Method 2)	-1560.61	-3497.71	-5881.02

Medical expenses cost (X_2)

	$\ln \hat{L}$	$\ln_{w^{(1)}} \hat{L}$	$\ln_{w^{(2)}} \hat{L}$
Normal	-1587.29	-7927.78	-22537.08
Lognormal	-568.82	-2742.53	-4601.96
Classical Kernel	-980.70	-2401.07	-3604.91
Kernel Transformed (TKE with log transformation)	-551.17	-2590.77	-4155.92
Kernel Transformed (TKE with Method 1)	-560.40	-2587.88	-4149.64
Kernel Transformed (TKE with Method 2)	-552.53	-2586.62	-4152.70

Table 4: Cross-valitation

Damage cost (X_1)

	CV	WCV_1	WCV_2
Classical Kernel	0.0129	0.1301	2.0993
Kernel Transformed (TKE with log transformation)	-0.0747	-0.2201	-1.2030
Kernel Transformed (TKE with Method 1)	-0.0856	-0.2207	-1.2120
Kernel Transformed (TKE with Method 2)	-0.0856	-0.2207	-1.2114

Medical expenses cost (X_2)

	CV	WCV_1	WCV_2
Classical Kernel	0.0973	0.1283	0.2763
Kernel Transformed (TKE with log transformation)	-0.7428	-0.1974	-0.1488
Kernel Transformed (TKE with Method 1)	-0.7146	-0.1948	-0.1484
Kernel Transformed (TKE with Method 2)	-0.7114	-0.1943	-0.1478

using weighted likelihoods where greater weight is given to density estimation in the right tail of the distribution.

We can see that the value of the log-likelihood function is not a good method to compare nonparametric fits given that its value increase when the bandwidth b decrease; thus we proposed alternative criteria based on the minimization of Integrated Square Error (ISE) and Weighted Integrated Square Error (WISE). Finally, we conclude that transformed kernel density estimation with a Shifted Power Transformation Family is a good alternative to fit distributions with heavy tails.

References

[1] Bolancé, C. (2010) Optimal Inverse Beta(3,3) Transformation in kernel density estimation, SORT Statistics and Operations Research Transaction, 34, 223-238.

[2] Bolancé, C., Guillén, M. and Nielsen, J.P. (2009) Transformation kernel estimation of insurance claim cost distribution, in Corazza, M. and Pizzi, C. (Eds), Mathematical and Statistical Methods for Actuarial Sciences and Finance, Springer, Roma, 223-231.

[3] Bolancé, C., Guillén, M. and Nielsen, J.P. (2008) Inverse Beta transformation in kernel density estimation, Statistics and Probability Letters, 78, 1757-1764.

[4] Bolancé, C., Guillén, M. and Nielsen, J.P., 2008. Inverse Beta transformation in kernel density estimation. Statistics & Probability Letters, 78, 1757-1764.

[5] Bolancé, C., Guillén, M., Pelican, E. and Vernic, R. (2008) Skewed bivariate models and nonparametric estimation for the CTE risk measure. Insurance: Mathematics and Economics, 43, 386-393.

[6] Bolancé, C., Guillén, M. and Nielsen, J.P., (2003) Kernel density estimation of actuarial loss functions, Insurance: Mathematics and Economics, 32, 19-36.

[7] Buch-Larsen, T., Guillen, M., Nielsen, J.P. and Bolancé, C. (2005) Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. Statistics, 39, 503-518.

[8] Clements, A.E., Hurn, A.S. and Lindsay, K.A. (2003) Möbius-like mappings and their use in kernel density estimation, Journal of the American Statistical Association, 98, 993-1000.

[9] Denault, M. (2001) Coherent allocation of risk capital. Working paper. Ecole des H.E.C., Montreal.

[10] Dhaene, J., Goovaerts, M.J. and Kaas, R. (2003) Economic capital allocation derived from risk measures, North American Actuarial Journal 7, 44-59.

[11] Hall, P. and Marron J.S. (1987) Estimation of integrated squared density derivatives, Statistics & Probability Letters, 6, 100-115.

[12] Klugman, S.A., Panjer, H.H, and Willmot, G.E., 2008. Loss models: from data to decisions. 3rd Edition. John Wiley & Sons, New Jersey.

[13] Liu, Q., Pitt, D., Zhang, X. and Wu, X. (2011) A Bayesian approach to parameter estimation for kernel density estimation via transformation, Annals of Actuarial Science, in press.

[14] McNeil, A.J., Frey. R. and Embrechts, P. (2005) Quantitative risk management: concepts, techniques and tools. Princeton University Press., Princeton Series in Finance, London.

[15] Panjer, H.H. (2002) Measurement of risk, solvency requirements and allocation of capital within financial conglomerates. 27th International Congress of Actuaries, Cancun 2002. (see also http://www.actuaries.org/events/congresses/Cancun/afir_subject/afir_14_panjer.pdf)

[16] Park, U.B. and Marron, J.S. (1990) Comparison of data-driven bandwidth selectors, Journal of the American Statistical Association, 8, 66-72.

[17] Reiss, R.D., 1981. Nonparametric estimation of smooth distribution functions. Scandinavian Journal of Statistics 8, 116-119.

[18] Scott, D.W., 1992. Multivariate Density Estimation. Theory, Practice and Visualization. John Wiley & Sons, Inc.

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

[20] Silverman, B.W. (1986) Density Estimation for Statistics and Data Analysis, Chapman & Hall, London.

[21] Terrell, G.R. (1990) The maximal smoothing principle in density estimation, Journal of the American Statistical Association, 85, 270-277.

[22] Terrell, G.R. and Scott, D.W. (1985) Oversmoothed nonparametric density estimates, Journal of the American Statistical Association, 80, 209-214.

[23] Valdez, E.A. and Chernih, A., 2003. Wang's capital allocation formula for elliptically contoured distributions. Insurance: Mathematics and Economics 33, 517-532.

[24] Vernic, R. 2006. Multivariate Skew-Normal distributions with applications in insurance. Insurance: Mathematics and Economics 38, 413-426.

[25] Wand, M.P. and Jones, M.C. (1995) Kernel Smoothing. Chapman & Hall, London.

[26] Wand, P., Marron, J.S. and Ruppert, D. (1991) Transformations in density estimation. Journal of the American Statistical Association, 86, 343-361.

[27] Wang, S. (2002) A set of new methods and tools for enterprise risk capital management and portfolio optimization. Working paper. SCOR reinsurance company (http://www.casact.com/pubs/forum/02sforum/02sf043.pdf).

[28] Wu, T.-J., Chen, C.-F. and Chen, H.-Y., 2007. A variable bandwidth selector in multivariate kernel density estimation. Statistics & Probability Letters 77, 462-467.

[29] Zhang, X., King, M.L. and Hyndman, R.J., 2006. A Bayesian approach to bandwidth selection for multivariate kernel density estimation. Computational Statistics & Data Analysis 50, 3009-3031.

Appendix

Descriptive statistics

```
load(QRMlib)
cost <- read.table("KEURcostes.txt",header=TRUE)
keurcost <- cbind(cost[1],cost[2])
keurcost <- as.matrix(keurcost)
n <- nrow(keurcost)
colMeans(keurcost) #Means of each variable
var(keurcost) # Note that sample variance divides by n-1
sd(keurcost) # Standard deviation of each variable
min(keurcost[,1])
max(keurcost[,1])
#Here it is necessary to load QRM library
skewness(keurcost[,1])
kurtosis(keurcost[,1])

# Fitting a normal distribution
normfit1=fit.norm(keurcost[,1])
# Histogram of property damage cost data with overlaying normal fit
hist(keurcost[,1],
xlab="Property damage cost in Euros (thousands)",freq=FALSE)
x=seq(0,1000,length.out=10000)
lines(x,dnorm(x,normfit1$mu,sqrt(normfit1$Sigma)))
```

```
# Fitting of lognormal distribution
ln.keurcost=log(keurcost)
ln.normfit1=fit.norm(ln.keurcost[,1])
# Histogram of log of property damage cost data with overlaying normal fit
a=seq(-4,8,length.out=10000)
hist(ln.keurcost[,1],
xlab="Log of Property damage cost in Euros (thousands)",freq=FALSE)
lines(a,dnorm(a,ln.normfit1$mu, sqrt(ln.normfit1$Sigma)))
```

Kernel density estimation

```
#Kernel density estimates. Univariate, Gaussian kernel and
#rule-of-thumb bandwidth calculated using standard deviation
#as dispersion parameter.
bw1<-1.06*sx[1,]*n^(-1/5)
dens1=density(keurcost[,1],bw=bw1,kernel="gaussian")
#Histogram of property damage cost data with overlaying kernel density estimate
hist(keurcost[,1],
xlab="Property damage cost in Euros (thousands)",
freq=FALSE,ylim=c(0,0.03))
lines(dens1)
#Histogram of property damage cost data with overlaying kernel density estimate
# for right tail
hist(keurcost[,1],
xlab="Property damage cost in Euros (thousands)",
freq=FALSE, xlim=c(200,1000),ylim=c(0,0.0005))
lines(dens1)

#Kernel density estimates of log-transformation. Univariate, Gaussian kernel
#and plug-in bandwidth by Sheather and Jones.
slnx<-sd(ln.keurcost)
slnx<-as.matrix(slnx)
ln.dens1=density(ln.keurcost[,1],bw="sj",kernel="gaussian")
#Histogram of log of property damage cost data with overlaying kernel
#density estimate
hist(ln.keurcost[,1],
xlab="Log of Property damage
cost in Euros (thousands)",freq=FALSE,ylim=c(0,.4))
```

```
lines(ln.dens1)
```

Optimal transformation

```
#METHOD 1
# Gaussian kernel, we define function ker
ker<-function(x){return(dnorm(x,0,1))}

# We define the transformation function
transf<-function(ll1,ll2,n,x)
{
yd=(x+(ll1*rep(1,n)))
if (ll2!=0)
     y=sign(ll2)*(yd^ll2)
  else
     y=log(yd)
return(y)
}

# We define the derivative of the transformation function
dtransf<-function(ll1,ll2,n,x)
{
yc=(x+(ll1*rep(1,n)))
if (ll2!=0)
     gy=sign(ll2)*ll2*(yc^(ll2-1))
  else
     gy=1/yc
return(gy)
}

# We define expression (5) to be minimized
beta<-function(ll)
{
ll1=ll[1]
ll2=ll[2]
y<-transf(ll1,ll2,n,x)
sy=sd(y)
sx=sd(x)
scal=sx/sy
yscal=scal*y
 an=sx*((21/(8*sqrt(2)*5*n*n))^(1/13))
 ssgaus=0
for (i in 2:n)
         {
         y1=yscal[i-1]
         n1=n-(i-1)
         y1=y1*rep(1,n1)
         y2=yscal[i:n]
         t=(y1-y2)/an;
         bgaus=dnorm(0,0,1)*(3/(2*sqrt(2)*(n*(n-1))*(an^5)))*
               sum((rep(1,n1)-(t^2)+((1/12)*(t^4)))*
               exp(-(1/4)*(t^2)))
         ssgaus=ssgaus+bgaus
      }
      ssgaus=ssgaus^(1/5)
      return(ssgaus)
}

lambda=rep(0,4)
dim(lambda)=c(2,2)
ll1=0.5
ll2=0.5
# Search for optimal parameters in cost 1
x<-as.matrix(keurcost[,1])
n<-nrow(x)
betaopt=optim(c(ll1,ll2),beta)

lambda[1,1]=betaopt$par[1]
lambda[2,1]=betaopt$par[2]

#we print the optimal parameters for cost 1
lambda[,1]

#METHOD 2
# We define expression of squarred skewness to be minimized
sk3<-function(ll1)
```

```
{y<-transf(ll1,ll2,n,x)
sy=sd(y)
sx=sd(x)
scal=sx/sy
yscal=scal*y
      a=sum(yscal^3);
      bb=sum(yscal^2);
      c=sum(yscal);
      ff=(a+((2*(c^3))/n^2)-((3*bb*c)/n))^2;
      ff=ff/(sx^6)
      return(ff)
}

nm=4000
grid=rep(0,nm)
dim(grid)=c((nm/4),4)
for(i in 1:(nm/4)){
grid[i,2]=-3+0.01*(i-1)
}
# Search for optimal parameters in cost 1
x<-as.matrix(keurcost[,1])
n<-nrow(x)
for(i in 1:(nm/4)){
ll1=0
ll2=grid[i,2]
skopt=optimize(sk3,c(-min(x)+0.01,1000))
grid[i,1]=skopt$minimum
grid[i,3]=skopt$objective
}
for(i in 1:(nm/4)){
grid[i,4]=beta(c(grid[i,1],grid[i,2]))
}
ll1=grid[which.min(grid[,4]),1]
ll2=grid[which.min(grid[,4]),2]
ll1
ll2
```

Transformed kernel density estimation

```
# We define the transformation function
transf<-function(ll1,ll2,n,x)
{
yd=(x+(ll1*rep(1,n)))
if (ll2!=0)
     y=sign(ll2)*(yd^ll2)
  else
     y=log(yd)
return(y)
}

# We define the derivative of the transformation function
dtransf<-function(ll1,ll2,n,x)
{
yc=(x+(ll1*rep(1,n)))
if (ll2!=0)
     gy=sign(ll2)*ll2*(yc^(ll2-1))
  else
     gy=1/yc
return(gy)
}
#Calculate TKE of pdf for property damage cost
#Transformation parameters are required
l1=1.870333
l2=-0.5700
grid<-as.matrix((1:10000)/100)
ng<-nrow(grid)
fkt<-as.matrix(rep(0,ng))
x<-as.matrix(keurcost[,1])
y=transf(l1,l2,n,x)
tgrid=transf(l1,l2,ng,grid)
fkt<-as.matrix(rep(0,ng))
sx=sd(x)
hnt=1.059*sx*((1/n)^(1/5))
sy=sd(y)
yscal=(sx/sy)*y
```

```
dy=dtransf(l1,l2,n,x)
dyscal=(sx/sy)*dy
tgscal=(sx/sy)*tgrid
dtg=dtransf(l1,l2,ng,grid)
dtgscal=(sx/sy)*dtg
for (i in 1:ng) {dif=(tgscal[i,]-yscal)/hnt

                 fkt[i,]=(dtgscal[i,]/(n*hnt))*sum(ker(dif))}

# Plot of TKE of pdf for property damage cost
plot(grid,fkt, type="l",
xlab="Property damage cost in Euros (thousands)", ylab="f")
```

Goodness of fit

Consider now the two parametric models, namely the univariate normal and univariate log-normal distributions. The R code given below calculates the log-likelihood function and both weighted forms of the log-likelihood function of two parametric models, namely the univariate normal and univariate log-normal distributions.

```
lnL=rep(0,12)
w1lnL=rep(0,12)
w2lnL=rep(0,12)
dim(lnL)=c(6,2)
dim(w1lnL)=c(6,2)
dim(w2lnL)=c(6,2)

#Calculation of log-likelihood for univariate normal and univariate log-normal
fits for both components of claim
lnL[1,1]=sum(log(dnorm(keurcost[,1],normfit1$mu,sqrt(normfit1$Sigma))))
lnL[2,1]=sum(log(dlnorm(keurcost[,1],ln.normfit1$mu,sqrt(ln.normfit1$Sigma))))

#Calculation of weighted log-likelihood for univariate normal and univariate
#log-normal fits for both components of claim (weights=claim size)
w1lnL[1,1]=518*sum(keurcost[,1]*log(dnorm(keurcost[,1],normfit1$mu,
sqrt(normfit1$Sigma))))/sum(keurcost[,1])
w1lnL[2,1]=518*sum(keurcost[,1]*log(dlnorm(keurcost[,1],ln.normfit1$mu,
sqrt(ln.normfit1$Sigma))))/sum(keurcost[,1])
```

```
#Calculation of weighted log-likelihood for univariate normal and univariate
#log-normal fits for both components of claim (weights=claim size^2)
w2lnL[1,1]=518*sum((keurcost[,1])^2*log(dnorm(keurcost[,1],normfit1$mu,
sqrt(normfit1$Sigma))))/sum((keurcost[,1])^2)
w2lnL[2,1]=518*sum((keurcost[,1])^2*log(dlnorm(keurcost[,1],ln.normfit1$mu,
sqrt(ln.normfit1$Sigma))))/sum((keurcost[,1])^2)
```

Next we give the R code used to calculate the log-likelihood and both versions of the weighted log-likelihood for the density estimate obtained by applying the classical kernel to the non-transformed data.

```
#Calculation of log-likelihood for univariate kernel density estimate applied
#to non-transformed data for both components of claim
n <- nrow(keurcost)
dens1val=c(rep(0,n))
for(i in 1:n){dens1val[i]=1/(n*dens1$bw)
*sum(dnorm((keurcost[i,1]-keurcost[,1])/dens1$bw),0,1)}
lnL[3,1]=sum(log(dens1val))

#Calculation of weighted log-likelihood for univariate kernel density estimate
#applied to non-transformed data for both components of claim
#(weights=claim size)
w1dens1val=c(rep(0,n))
for(i in 1:n){w1dens1val[i]=n*keurcost[i,1]*log(1/(n*dens1$bw)
*sum(dnorm((keurcost[i,1]-keurcost[,1])/dens1$bw,0,1)))/sum(keurcost[,1])}
w1lnL[3,1]=sum(w1dens1val)

#Calculation of weighted log-likelihood for univariate kernel density estimate
#applied to non-transformed data for both components of claim
#(weights=claim size^2)
w2dens1val=c(rep(0,n))
for(i in 1:n){w2dens1val[i]=n*(keurcost[i,1])^2*log(1/(n*dens1$bw)
*sum(dnorm((keurcost[i,1]-keurcost[,1])/dens1$bw,0,1)))/sum((keurcost[,1])^2)}
w2lnL[3,1]=sum(w2dens1val)
```

Below, we provide the R code used to calculate the log-likelihood and both versions of the weighted log-likelihood for the density estimate obtained by applying the classical kernel to the log-transformed data.

```
#Calculation of log-likelihood for univariate kernel density estimate
#obtained from log-transformed data for both components of claim
ln.dens1val=c(rep(0,n))
for(i in 1:n){ln.dens1val[i]=1/(n*keurcost[i,1]*ln.dens1$bw)
*sum(dnorm((ln.keurcost[i,1]-ln.keurcost[,1])/ln.dens1$bw,0,1))}
lnL[4,1]=sum(log(ln.dens1val))

#Calculation of weighted log-likelihood for univariate kernel density estimate
#obtained from log-transformed for both components of claim
#(weights=claim size)
w1ln.dens1val=c(rep(0,n))
for(i in 1:n){w1ln.dens1val[i]=n*keurcost[i,1]
*log(1/(n*keurcost[i,1]*ln.dens1$bw)
*sum(dnorm((ln.keurcost[i,1]-ln.keurcost[,1])/ln.dens1$bw,0,1)))
/sum(keurcost[,1])}
w1lnL[4,1]=sum(w1ln.dens1val)

#Calculation of weighted log-likelihood for univariate kernel density estimate
#obtained from log-transformed for both components of claim
#(weights=claim size^2)
w2ln.dens1val=c(rep(0,n))
for(i in 1:n){w2ln.dens1val[i]=n*(keurcost[i,1])^2
*log(1/(n*keurcost[i,1]*ln.dens1$bw)
*sum(dnorm((ln.keurcost[i,1]-ln.keurcost[,1])/ln.dens1$bw,0,1)))
/sum((keurcost[,1])^2)}
w2lnL[4,1]=sum(w2ln.dens1val)
```

Now we give the R code used to obtain the log-likelihood and both forms of the weighted log-likelihood for the kernel density estimate obtained by applying the optimal transformation parameters, to our data. In this case, it is necessary to write the values of transformation parameters: ll1 (λ_1) and ll2 (λ_2).

```
#Calculate shifted power transformed kernel density estimation
#(given optimal lambdas) and log-likelihoods for cost 1
ll1=1.8703337
ll2=-0.57
x<-as.matrix(keurcost[,1])
  fkt=as.matrix(rep(0,n))
```

```
   sx=sd(x)
   hnt=1.059*sx*((1/n)^(1/5))
   y=transf(ll1,ll2,n,x)
   sy=sd(y)
   yscal=(sx/sy)*y
   gy=dtransf(ll1,ll2,n,x)
   gyscal=(sx/sy)*gy
   for (i in 1:(n-1))      {
      vecy<-as.matrix((yscal[1:i]-yscal[(n-i+1):n])/hnt)[1:i]
      newvecy=ker(vecy)
      auxy=c(as.matrix(newvecy),as.matrix(rep(0,(n-i))))
      aux2y=c(as.matrix(rep(0,(n-i))),as.matrix(newvecy))
      fkt=fkt+(auxy+aux2y)
      }
  k0=dnorm(0,0,1)
  fkt=fkt+rep(k0,n)
   fkt=(gyscal*fkt)/(n*hnt)
#Calculation of log-likelihood for transformation kernel density estimate
tkd.dens1val=c(rep(0,n))
for(i in 1:n){
tkd.dens1val[i]=log(fkt[i])}
lnL[6,1]=sum(tkd.dens1val)

#Calculation of weighted log-likelihood for transformation kernel density
#estimate (weights=claim size)
w1tkd.dens1val=c(rep(0,n))
for(i in 1:n){
w1tkd.dens1val[i]=n*x[i]*log(fkt[i])}
w1lnL[6,1]=sum(w1tkd.dens1val)/(sum(x))

#Calculation of weighted log-likelihood for transformation kernel density
#estimate (weights=claim size^2)
w2tkd.dens1val=c(rep(0,n))
for(i in 1:n){
w2tkd.dens1val[i]=n*x[i]*x[i]*log(fkt[i])}
w2lnL[6,1]=sum(w2tkd.dens1val)/(sum(x^2))
```

Newt we give the R code used to calculate the ISE and and both versions of the weighted ISE for the density estimate obtained by applying the classical kernel to the non-transformed data.

```
#Calculation ISE for classical kernels estimations cost1
x<-as.matrix(keurcost[,1])

ise1<-rep(0,12)
ise2<-rep(0,12)

dim(ise1)<-c(4,3)

dim(ise2)<-c(4,3)

grid<-as.matrix((1:500000)/100)
ng=nrow(grid)
ng
fk<-as.matrix(rep(0,ng))
bw1<-dens1$bw
for (i in 1:ng) {dif=(grid[i,]-x)/bw1

                 fk[i,]=(1/(n*bw1))*sum(ker(dif))}


first<-sum((fk^2)*0.01)
first
fk_i<-as.matrix(rep(0,n))
dim(fk_i)<-c(n,1)

fk_i[1,]<-(1/(n*bw1))*sum(ker((x[1,]-x[2:n,])/bw1))
fk_i[n,]<-(1/(n*bw1))*sum(ker((x[n,]-x[1:(n-1),])/bw1))
for (i in 2:(n-1)) {
    fk_i[i,]<-(1/(n*bw1))*(sum(ker((x[i,]-x[1:(i-1),])/bw1))
    +sum(ker((x[i,]-x[(i+1):n,])/bw1)))}
second<-sum(fk_i)/n
second
isek<-first-2*second
isek
```

```
ise1[1,1]<-isek

#Calculation WISE1 for classical kernels estimations cost1
first<-sum((fk^2)*grid*0.01)
first

second<-sum(fk_i*x)/n
second

w1isek<-first-2*second
w1isek
ise1[1,2]<-w1isek

#Calculation WISE2 for classical kernels estimations cost1
first<-sum((fk^2)*(grid^2)*0.01)
first

second<-sum(fk_i*(x^2))/n
second

w2isek<-first-2*second
w2isek
ise1[1,3]<-w2isek
```

Below, we provide the R code used to calculate the ISE and both versions of the weighted ISE for the density estimate obtained by applying the classical kernel to the log-transformed data.

```
#Calculation ISE for log-transformed kernels estimations cost1
x<-as.matrix(keurcost[,1])
y<-log(x)
lgrid=log(grid)
fkt<-as.matrix(rep(0,ng))
bw.ln1<-ln.dens1$bw
for (i in 1:ng) {dif=(lgrid[i,]-y)/bw.ln1
```

```
                 fkt[i,]=(1/(n*bw.ln1*grid[i,]))*sum(ker(dif))}


first<-sum((fk^2)*0.01)
first
fk_i<-as.matrix(rep(0,n))
dim(fk_i)<-c(n,1)
fk_i[1,]<-(1/(n*bw.ln1*x[1,]))*sum(ker((y[1,]-y[2:n,])/bw.ln1))
fk_i[n,]<-(1/(n*bw.ln1*x[n,]))*sum(ker((y[n,]-y[1:(n-1),])/bw.ln1))
for (i in 2:(n-1)) {
    fk_i[i,]<-(1/(n*bw.ln1*x[i,]))*(sum(ker((y[i,]-y[1:(i-1),])/bw.ln1))
    +sum(ker((y[i,]-y[(i+1):n,])/bw.ln1)))}
second<-sum(fk_i)/n
second
isek<-first-2*second
isek

ise1[2,1]<-isek
#Calculation WISE1 for log-transformed kernels estimations cost1
first<-sum((fkt^2)*grid*0.01)
first

second<-sum(fk_i*x)/n
second

w1isek<-first-2*second
w1isek
ise1[2,2]<-w1isek

#Calculation WISE2 for log-transformed kernels estimations cost1
first<-sum((fkt^2)*(grid^2)*0.01)
first

second<-sum(fk_i*(x^2))/n
second

w2isek<-first-2*second
w2isek
```

```
ise1[2,3]<-w2isek
```

Now we give the R code used to obtain the ISE and both forms of the weighted ISE for the kernel density estimate obtained by applying the optimal transformation parameters, to our data. In this case, it is necessary to write the values of transformation parameters: ll1 (λ_1) and ll2 (λ_2).

```
#Calculation ISE for shifted power transformed kernels estimations
#(given optimal lambdas) of cost1
x<-as.matrix(keurcost[,1])

ll1=1.993066
ll2=-0.620136
y=transf(ll1,ll2,n,x)
tgrid=transf(ll1,ll2,ng,grid)
fkt<-as.matrix(rep(0,ng))
sx=sd(x)
hnt=1.059*sx*((1/n)^(1/5))
sy=sd(y)
yscal=(sx/sy)*y
dy=dtransf(ll1,ll2,n,x)
dyscal=(sx/sy)*dy
tgscal=(sx/sy)*tgrid
dtg=dtransf(ll1,ll2,ng,grid)
dtgscal=(sx/sy)*dtg
for (i in 1:ng) {dif=(tgscal[i,]-yscal)/hnt

                 fkt[i,]=(dtgscal[i,]/(n*hnt))*sum(ker(dif))}
first<-sum((fkt^2)*0.01)
first
fk_i<-as.matrix(rep(0,n))
dim(fk_i)<-c(n,1)
fk_i[1,]<-(dyscal[1,]/(n*hnt))*sum(ker((yscal[1,]-yscal[2:n,])/hnt))
fk_i[n,]<-(dyscal[n,]/(n*hnt))*sum(ker((yscal[n,]-yscal[1:(n-1),])/hnt))
for (i in 2:(n-1)) {
   fk_i[i,]<-(dyscal[i,]/(n*hnt))*(sum(ker((yscal[i,]-yscal[1:(i-1),])/hnt))
     +sum(ker((yscal[i,]-yscal[(i+1):n,])/hnt)))}
```

```
second<-sum(fk_i)/n
second
isek<-first-2*second
isek

ise1[3,1]<-isek

#Calculation WISE1 for shifted power transformed kernels estimations
#(given optimal lambdas) of cost1
first<-sum((fkt^2)*grid*0.01)
first

second<-sum(fk_i*x)/n
second

w1isek<-first-2*second
w1isek
ise1[3,2]<-w1isek

#Calculation WISE2 for shifted power transformed kernels estimations
#(given optimal lambdas) of cost1
first<-sum((fkt^2)*(grid^2)*0.01)
first

second<-sum(fk_i*(x^2))/n
second

w2isek<-first-2*second
w2isek
ise1[3,3]<-w2isek
```

2006

CREAP2006-01
Matas, A. (GEAP)**; Raymond, J.Ll.** (GEAP)
"Economic development and changes in car ownership patterns"
(Juny 2006)

CREAP2006-02
Trillas, F. (IEB)**; Montolio, D.** (IEB)**; Duch, N.** (IEB)
"Productive efficiency and regulatory reform: The case of Vehicle Inspection Services"
(Setembre 2006)

CREAP2006-03
Bel, G. (PPRE-IREA)**; Fageda, X.** (PPRE-IREA)
"Factors explaining local privatization: A meta-regression analysis"
(Octubre 2006)

CREAP2006-04
Fernàndez-Villadangos, L. (PPRE-IREA)
"Are two-part tariffs efficient when consumers plan ahead?: An empirical study"
(Octubre 2006)

CREAP2006-05
Artís, M. (AQR-IREA)**; Ramos, R.** (AQR-IREA)**; Suriñach, J.** (AQR-IREA)
"Job losses, outsourcing and relocation: Empirical evidence using microdata"
(Octubre 2006)

CREAP2006-06
Alcañiz, M. (RISC-IREA)**; Costa, A.; Guillén, M.** (RISC-IREA)**; Luna, C.; Rovira, C.**
"Calculation of the variance in surveys of the economic climate”
(Novembre 2006)

CREAP2006-07
Albalate, D. (PPRE-IREA)
"Lowering blood alcohol content levels to save lives: The European Experience”
(Desembre 2006)

CREAP2006-08
Garrido, A. (IEB)**; Arqué, P.** (IEB)
“The choice of banking firm: Are the interest rate a significant criteria?”
(Desembre 2006)

CREAP2006-09
Segarra, A. (GRIT)**; Teruel-Carrizosa, M.** (GRIT)
"Productivity growth and competition in spanish manufacturing firms:
What has happened in recent years?"
(Desembre 2006)

CREAP2006-10
Andonova, V.; Díaz-Serrano, Luis. (CREB)
"Political institutions and the development of telecommunications"
(Desembre 2006)

CREAP2006-11
Raymond, J.L.(GEAP)**; Roig, J.L..** (GEAP)
"Capital humano: un análisis comparativo Catalunya-España"
(Desembre 2006)

CREAP2006-12
Rodríguez, M.(CREB)**; Stoyanova, A.** (CREB)
"Changes in the demand for private medical insurance following a shift in tax incentives"
(Desembre 2006)

CREAP2006-13
Royuela, V. (AQR-IREA)**; Lambiri, D.**; **Biagi**, **B.**
"Economía urbana y calidad de vida. Una revisión del estado del conocimiento en España"
(Desembre 2006)

CREAP2006-14
Camarero, M.; Carrion-i-Silvestre, J.LL. (AQR-IREA)**.**;**Tamarit**, **C.**
"New evidence of the real interest rate parity for OECD countries using panel unit root tests with breaks"
(Desembre 2006)

CREAP2006-15
Karanassou, M.; Sala, H. (GEAP).;**Snower** , **D. J.**
"The macroeconomics of the labor market: Three fundamental views"
(Desembre 2006)

2007

XREAP2007-01
Castany, L (AQR-IREA)**; López-Bazo, E.** (AQR-IREA)**.;Moreno** , **R.** (AQR-IREA)
"Decomposing differences in total factor productivity across firm size"
(Març 2007)

XREAP2007-02
Raymond, J. Ll. (GEAP)**; Roig, J. Ll.** (GEAP)
"Una propuesta de evaluación de las externalidades de capital humano en la empresa"
(Abril 2007)

XREAP2007-03
Durán, J. M. (IEB)**; Esteller, A.** (IEB)
"An empirical analysis of wealth taxation: Equity vs. Tax compliance"
(Juny 2007)

XREAP2007-04
Matas, A. (GEAP)**; Raymond, J.Ll.** (GEAP)
"Cross-section data, disequilibrium situations and estimated coefficients: evidence from car ownership demand"
(Juny 2007)

XREAP2007-05
Jofre-Montseny, J. (IEB)**; Solé-Ollé, A.** (IEB)
"Tax differentials and agglomeration economies in intraregional firm location"
(Juny 2007)

XREAP2007-06
Álvarez-Albelo, C. (CREB)**; Hernández-Martín, R.**
"Explaining high economic growth in small tourism countries with a dynamic general equilibrium model"
(Juliol 2007)

XREAP2007-07
Duch, N. (IEB)**; Montolio, D.** (IEB); **Mediavilla, M.**
"Evaluating the impact of public subsidies on a firm's performance: a quasi-experimental approach"
(Juliol 2007)

XREAP2007-08
Segarra-Blasco, A. (GRIT)
"Innovation sources and productivity: a quantile regression analysis"
(Octubre 2007)

XREAP2007-09
Albalate, D. (PPRE-IREA)
"Shifting death to their Alternatives: The case of Toll Motorways"
(Octubre 2007)

XREAP2007-10
Segarra-Blasco, A. (GRIT); **Garcia-Quevedo, J.** (IEB); **Teruel-Carrizosa, M.** (GRIT)
"Barriers to innovation and public policy in catalonia"
(Novembre 2007)

XREAP2007-11
Bel, G. (PPRE-IREA); **Foote, J.**
"Comparison of recent toll road concession transactions in the United States and France"
(Novembre 2007)

XREAP2007-12
Segarra-Blasco, A. (GRIT);
"Innovation, R&D spillovers and productivity: the role of knowledge-intensive services"
(Novembre 2007)

XREAP2007-13
Bermúdez Morata, Ll. (RFA-IREA); **Guillén Estany, M.** (RFA-IREA), **Solé Auró, A**. (RFA-IREA)
"Impacto de la inmigración sobre la esperanza de vida en salud y en discapacidad de la población española"
(Novembre 2007)

XREAP2007-14
Calaeys, P. (AQR-IREA); **Ramos, R.** (AQR-IREA), **Suriñach, J**. (AQR-IREA)
"Fiscal sustainability across government tiers"
(Desembre 2007)

XREAP2007-15
Sánchez Hugalbe, A. (IEB)
"Influencia de la inmigración en la elección escolar"
(Desembre 2007)

2008

XREAP2008-01
Durán Weitkamp, C. (GRIT)**; Martín Bofarull, M.** (GRIT) ; **Pablo Martí, F.**
"Economic effects of road accessibility in the Pyrenees: User perspective"
(Gener 2008)

XREAP2008-02
Díaz-Serrano, L.; Stoyanova, A. P. (CREB)
"The Causal Relationship between Individual's Choice Behavior and Self-Reported Satisfaction: the Case of Residential Mobility in the EU"
(Març 2008)

XREAP2008-03
Matas, A. (GEAP)**; Raymond, J. L.** (GEAP)**; Roig, J. L.** (GEAP)
"Car ownership and access to jobs in Spain"
(Abril 2008)

XREAP2008-04
Bel, G. (PPRE-IREA) **; Fageda, X.** (PPRE-IREA)
"Privatization and competition in the delivery of local services: An empirical examination of the dual market hypothesis"
(Abril 2008)

XREAP2008-05
Matas, A. (GEAP); **Raymond, J. L.** (GEAP); **Roig, J. L.** (GEAP)
"**Job accessibility and employment probability**"
(Maig 2008)

XREAP2008-06
Basher, S. A.; Carrión, J. Ll. (AQR-IREA)
Deconstructing Shocks and Persistence in OECD Real Exchange Rates
(Juny 2008)

XREAP2008-07
Sanromá, E. (IEB)**; Ramos, R.** (AQR-IREA); Simón, H.
Portabilidad del capital humano y asimilación de los inmigrantes. Evidencia para España
(Juliol 2008)

XREAP2008-08
Basher, S. A.; Carrión, J. Ll. (AQR-IREA)
Price level convergence, purchasing power parity and multiple structural breaks: An application to US cities
(Juliol 2008)

XREAP2008-09
Bermúdez, Ll. (RFA-IREA)
A priori ratemaking using bivariate poisson regression models
(Juliol 2008)

XREAP2008-10
Solé-Ollé, A. (IEB), **Hortas Rico, M.** (IEB)
Does urban sprawl increase the costs of providing local public services? Evidence from Spanish municipalities
(Novembre 2008)

XREAP2008-11
Teruel-Carrizosa, M. (GRIT), **Segarra-Blasco, A.** (GRIT)
Immigration and Firm Growth: Evidence from Spanish cities
(Novembre 2008)

XREAP2008-12
Duch-Brown, N. (IEB), **García-Quevedo, J.** (IEB), **Montolio, D.** (IEB)
Assessing the assignation of public subsidies: Do the experts choose the most efficient R&D projects?
(Novembre 2008)

XREAP2008-13
Bilotkach, V., **Fageda, X.** (PPRE-IREA), **Flores-Fillol, R.**
Scheduled service versus personal transportation: the role of distance
(Desembre 2008)

XREAP2008-14
Albalate, D. (PPRE-IREA), **Gel, G.** (PPRE-IREA)
Tourism and urban transport: Holding demand pressure under supply constraints
(Desembre 2008)

2009

XREAP2009-01
Calonge, S. (CREB)**; Tejada, O.**
“A theoretical and practical study on linear reforms of dual taxes”
(Febrer 2009)

XREAP2009-02
Albalate, D. (PPRE-IREA)**; Fernández-Villadangos, L.** (PPRE-IREA)
“Exploring Determinants of Urban Motorcycle Accident Severity: The Case of Barcelona”
(Març 2009)

XREAP2009-03
Borrell, J. R. (PPRE-IREA)**; Fernández-Villadangos, L.** (PPRE-IREA)
“Assessing excess profits from different entry regulations”
(Abril 2009)

XREAP2009-04
Sanromá, E. (IEB)**; Ramos, R.** (AQR-IREA), **Simon, H.**
“Los salarios de los inmigrantes en el mercado de trabajo español. ¿Importa el origen del capital humano?”
(Abril 2009)

XREAP2009-05
Jiménez, J. L.; Perdiguero, J. (PPRE-IREA)
“(No)competition in the Spanish retailing gasoline market: a variance filter approach”
(Maig 2009)

XREAP2009-06
Álvarez-Albelo,C. D. (CREB)**, Manresa, A.** (CREB)**, Pigem-Vigo, M.** (CREB)
“International trade as the sole engine of growth for an economy”
(Juny 2009)

XREAP2009-07
Callejón, M. (PPRE-IREA)**, Ortún V, M.**
“The Black Box of Business Dynamics”
(Setembre 2009)

XREAP2009-08
Lucena, A. (CREB)
“The antecedents and innovation consequences of organizational search: empirical evidence for Spain”
(Octubre 2009)

XREAP2009-09
Domènech Campmajó, L. (PPRE-IREA)
“Competition between TV Platforms”
(Octubre 2009)

XREAP2009-10
Solé-Auró, A. (RFA-IREA),**Guillén, M.** (RFA-IREA)**, Crimmins, E. M.**
"Health care utilization among immigrants and native-born populations in 11 European countries. Results from the Survey of Health, Ageing and Retirement in Europe"
(Octubre 2009)

XREAP2009-11
Segarra, A. (GRIT)**, Teruel, M.** (GRIT)
"Small firms, growth and financial constraints"
(Octubre 2009)

XREAP2009-12
Matas, A. (GEAP)**, Raymond, J.Ll.** (GEAP), **Ruiz, A.** (GEAP)
"Traffic forecasts under uncertainty and capacity constraints"
(Novembre 2009)

XREAP2009-13
Sole-Ollé, A. (IEB)
"Inter-regional redistribution through infrastructure investment: tactical or programmatic?"
(Novembre 2009)

XREAP2009-14
Del Barrio-Castro, T., **García-Quevedo, J.** (IEB)
"The determinants of university patenting: Do incentives matter?"
(Novembre 2009)

XREAP2009-15
Ramos, R. (AQR-IREA), **Suriñach, J.** (AQR-IREA)**, Artís, M.** (AQR-IREA)
"Human capital spillovers, productivity and regional convergence in Spain"
(Novembre 2009)

XREAP2009-16
Álvarez-Albelo, C. D. (CREB), **Hernández-Martín, R.**
"The commons and anti-commons problems in the tourism economy"
(Desembre 2009)

2010

XREAP2010-01
García-López, M. A. (GEAP)
"The Accessibility City. When Transport Infrastructure Matters in Urban Spatial Structure"
(Febrer 2010)

XREAP2010-02
García-Quevedo, J. (IEB), **Mas-Verdú, F.** (IEB), **Polo-Otero, J.** (IEB)
"Which firms want PhDs? The effect of the university-industry relationship on the PhD labour market"
(Març 2010)

XREAP2010-03
Pitt, D., **Guillén, M.** (RFA-IREA)
"An introduction to parametric and non-parametric models for bivariate positive insurance claim severity distributions"
(Març 2010)

XREAP2010-04
Bermúdez, Ll. (RFA-IREA), **Karlis, D.**
"Modelling dependence in a ratemaking procedure with multivariate Poisson regression models"
(Abril 2010)

XREAP2010-05
Di Paolo, A. (IEB)
"Parental education and family characteristics: educational opportunities across cohorts in Italy and Spain"
(Maig 2010)

XREAP2010-06
Simón, H. (IEB), **Ramos, R.** (AQR-IREA), **Sanromá, E.** (IEB)
"Movilidad ocupacional de los inmigrantes en una economía de bajas cualificaciones. El caso de España"
(Juny 2010)

XREAP2010-07
Di Paolo, A. (GEAP & IEB), **Raymond, J. Ll.** (GEAP & IEB)
"Language knowledge and earnings in Catalonia"
(Juliol 2010)

XREAP2010-08
Bolancé, C. (RFA-IREA), **Alemany, R.** (RFA-IREA), **Guillén, M.** (RFA-IREA)
"Prediction of the economic cost of individual long-term care in the Spanish population"
(Setembre 2010)

XREAP2010-09
Di Paolo, A. (GEAP & IEB)
"Knowledge of catalan, public/private sector choice and earnings: Evidence from a double sample selection model"
(Setembre 2010)

XREAP2010-10
Coad, A., **Segarra, A.** (GRIT), **Teruel, M.** (GRIT)
"Like milk or wine: Does firm performance improve with age?"
(Setembre 2010)

XREAP2010-11
Di Paolo, A. (GEAP & IEB), **Raymond, J. Ll.** (GEAP & IEB), **Calero, J.** (IEB)
"Exploring educational mobility in Europe"
(Octubre 2010)

XREAP2010-12
Borrell, A. (GiM-IREA), **Fernández-Villadangos, L.** (GiM-IREA)
"Clustering or scattering: the underlying reason for regulating distance among retail outlets"
(Desembre 2010)

XREAP2010-13
Di Paolo, A. (GEAP & IEB)
"School composition effects in Spain"
(Desembre 2010)

XREAP2010-14
Fageda, X. (GiM-IREA), **Flores-Fillol, R.**
"Technology, Business Models and Network Structure in the Airline Industry"
(Desembre 2010)

XREAP2010-15
Albalate, D. (GiM-IREA), **Bel, G.** (GiM-IREA), **Fageda, X.** (GiM-IREA)
"Is it Redistribution or Centralization? On the Determinants of Government Investment in Infrastructure"
(Desembre 2010)

XREAP2010-16
Oppedisano, V., **Turati, G.**
"What are the causes of educational inequalities and of their evolution over time in Europe? Evidence from PISA"
(Desembre 2010)

XREAP2010-17
Canova, L., **Vaglio, A.**
"Why do educated mothers matter? A model of parental help"
(Desembre 2010)

2011

XREAP2011-01
Fageda, X. (GiM-IREA), **Perdiguero, J.** (GiM-IREA)
“An empirical analysis of a merger between a network and low-cost airlines”
(Maig 2011)

XREAP2011-02
Moreno-Torres, I. (ACCO, CRES & GiM-IREA)
“What if there was a stronger pharmaceutical price competition in Spain? When regulation has a similar effect to collusion”
(Maig 2011)

XREAP2011-03
Miguélez, E. (AQR-IREA); **Gómez-Miguélez, I.**
“Singling out individual inventors from patent data”
(Maig 2011)

XREAP2011-04
Moreno-Torres, I. (ACCO, CRES & GiM-IREA)
“Generic drugs in Spain: price competition vs. moral hazard”
(Maig 2011)

XREAP2011-05
Nieto, S. (AQR-IREA)**, Ramos, R.** (AQR-IREA)
“¿Afecta la sobreeducación de los padres al rendimiento académico de sus hijos?”
(Maig 2011)

XREAP2011-06
Pitt, D., **Guillén, M.** (RFA-IREA), **Bolancé, C.** (RFA-IREA)
“Estimation of Parametric and Nonparametric Models for Univariate Claim Severity Distributions - an approach using R”
(Juny 2011)

XARXA
DE REFERÈNCIA
EN ECONOMIA APLICADA

xreap@pcb.ub.es