ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

METHODOLOGY AND APPLICATIONS OF KERNEL TECHNIQUES IN MORTALITY DATA

By

Paraskevi M. Peristera

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
2001

ISBN : 960-8287-06-5

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

METHODOLOGY AND APPLICATIONS OF KERNEL TECHNIQUES IN MORTALITY DATA

By

Paraskevi M. Peristera

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
2001

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΤΩΝ ΤΕΧΝΙΚΩΝ KERNEL ΣΕ ΔΗΜΟΓΡΑΦΙΚΑ ΔΕΔΟΜΕΝΑ

Παρασκευή Μ. Περιστερά

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής

του Οικονομικού Πανεπιστημίου Αθηνών

ως μέρος των απαιτήσεων για την απόκτηση

Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα

2001

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

A Thesis submitted in partial fulfilment of

the requirements for the degree of

Master of Science

Methodology and applications of kernel techniques in demographic and economic data

Paraskevi M. Peristera

Supervisor : Anastasia Kostaki
Lecturer

External examiner :JJ McCutcheon
Professor, Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh

Approved by the Graduate Committee

Professor J. Panaretos
Director of the Graduate Program
May 2001

DEDICATION

To my parents

ACKNOWLEDGEMENTS

I would like to express my deep gratefulness to my supervisor Kostaki Anastasia for giving me the opportunity of becoming familiar with an extremely interesting field of Statistics, this of Demography and Actuarial Statistics. Her help and guidance were valuable for the fulfilment of this thesis. Furthermore the advice of Dr. Christos Papatheodorou concerning the applications in economic data were crucial for a complete understanding and presentation of them in this thesis. I really thank both of them for their time and their patience. Finally I would like to thank my parents and especially my sister for their support and their encouragement for all the years of my studies.

VITA

I was born in 1976 in Athens and graduated from the high school of Lavrion in 1994. I started my studies in mathematics at the university of Thessaloniki. I got my degree in mathematics in 1998 from the university of Patras. On September of the same year I was admitted at Athens University of Economics and Business in the postgraduate program in Statistics. The second semester was held at the Katholieke Universiteit of Leuven in Belgium. At the second year of the master I was dealt with the completion of my thesis.

ABSTRACT

Paraskevi Peristera

Methodology and Applications of Kernel Techniques in mortality data

May 2001

Graduation techniques are extensively used in actuarial field and in analysis of demographic data. Both parametric and non-parametric methods are used for achieving smoothness of the mortality data. Recently, particular emphasis is given at kernel techniques. A presentation of the methodology of these techniques is provided here. A review of methods for the choice of the bandwidth parameter is also provided. In order to illustrate the applicability of kernel graduation method, we use it for graduating different mortality data sets. The classical kernel approach is also compared to the local linear approach. In order to evaluate the efficiency and accuracy of the kernel method as a graduation method, we compare its results with those of a parametric model.

Furthermore we present the applicability of kernel techniques in the case of income data in order to reveal the shape of income distribution of different countries. A short review of the most appropriate method for the choice of the bandwidth parameter is also provided. Finally we use this method for comparing the income distribution of several European countries as well as of the USA.

ΠΕΡΙΛΗΨΗ

Παρασκευή Περιστερά

Μεθοδολογία και Εφαρμογές των τεχνικών kernel σε Δημογραφικά δεδομένα

Μάιος 2001

Οι τεχνικές εξομάλυνσης χρησιμοποιούνται ευρέως στο πεδίο της Αναλογιστικής Επιστήμης και στην ανάλυση δημογραφικών δεδομένων. Τόσο παραμετρικές όσο και μη παραμετρικές μέθοδοι χρησιμοποιούνται για την εξομάλυνση των δεδομένων θνησιμότητας. Τελευταία ιδιαίτερη έμφαση δίνεται στις τεχνικές kernel. Στην εργασία αυτή παρέχεται μία παρουσίαση της μεθοδολογίας αυτών των τεχνικών. Με σκοπό την αξιολόγηση της αποτελεσματικότητας και της ακρίβειας των τεχνικών kernel σαν μέθοδο εξομάλυνσης χρησιμοποιούμε την τεχνική αυτή για την εξομάλυνση εμπειρικών δεδομένων θνησιμότητας. Ακόμα για να αξιολογήσουμε τη μέθοδο αυτή, συγκρίνουμε τα αποτελέσματά της με αυτά που προκύπτουν από την χρήση ενός παραμετρικού μοντέλου.

Στη συνέχεια παρουσιαζούμε την εφαρμογή των τεχνικών kernel σε δεδομένα εισοδημάτων διαφόρων χωρών με σκοπό να περιγράψουμε την μορφή της κατανομής τους.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF PLOTS

Chapter 1

INTRODUCTION

Mortality measurement is a subject of interest both in Demography and Actuarial science. Sets of mortality rates are widely used by actuaries to calculate life insurance premiums, annuities, industrial assurance premiums and so on. On the other hand demographers will use national mortality tables in order to project the population of a country.

The purpose of measuring mortality is to enable inferences to be drawn about the likelihood of death occurring within a specific population during a specific period of time. In estimating mortality, actuaries and demographers employ life tables as a model. However fluctuations will be inherent in the observations since the actual observations from which the life table has been derived are a sample of total experience in terms of time i.e. covering a short period of years. So, the construction of a new table is the adjustment of the observed rates to produce smooth decrement rates, which are accurate estimates of the underlying mortality. The adjustment procedure that reduces the random errors in the observed rates as well as smoothing them is known as graduation.

Graduation techniques should be distinguished between parametric and non-parametric. Both parametric and non-parametric techniques can produce close adherence to the data. The most widely used graduation techniques are the graphic method, the spline graduation technique and mortality laws. Another method is the use of summation and adjusted-average formulae especially used by actuaries. An alternative approach for graduating the age specific mortality pattern is a non-parametric technique known as kernel graduation technique.

The kernel graduation technique allows to show in an effective way the structure of a data set without the imposition of a parametric model. Generally, kernel estimates approximate the density f(x) from observations x. These estimates depend on the data, the kernel function K(.) and on the bandwidth parameter h(.). In effect, a kernel density estimator is formed by placing a kernel function at each data point and then by

summing these functions to form the density estimate. So, the kernel estimator can be considered as sum of "bumps" placed at the observations (Silverman, 1986). The kernel function K determines the shape of the bumps, while the bandwidth h determines their width. The choice of the bandwidth is of great importance since it controls the degree of smoothing. If a small bandwidth is used then nearby points are more influential. On the other hand if a large bandwidth is used then information is averaged over a larger region and as a consequence individual points have less influence on the estimate. In the case of mortality data, two significant kernel estimators used for graduation are the Copas-Haberman estimator and the Nadaraya-Watson estimator (Gavin, Haberman and Verall, 1994). The Copas-Haberman estimator has been used for estimating mortality data by Copas and Haberman (1983) as well as by Bloomflied and Haberman (1987). The Nadaraya-Watson estimator is closely related to Moving Weighed Average (Gavin, Haberman and Verall, 1992). For the choice of the optimal bandwidth several methods exist in the literature. Bloomfield and Haberman (1987) in order to find the optimal bandwidth first fit a curve to the data and then test the graduated rates for smoothness using standard actuarial tests of fit. Gavin, Haberman and Verall (1994) use the cross-validation technique for the choice of the optimal bandwidth, since this method allows a balance between variance and bias. A small bandwidth reflects exactly the crude death rates while for larger bandwidths more smoothing occurs at the expense of fit between graduated rates and the actual data. Generally kernel estimates fail to deal satisfactorily with the tails of the distributions without oversmoothing the main part of the distribution.

Although kernel graduation techniques are mainly used for smoothing demographic data sets they can provide satisfactory results in the case of economic data. This method has proved to be a very useful tool particularly for graphical illustration of the shape of income distributions. In particular kernel techniques result in smooth density estimates that make easier the comparison between different states (such as differences in time, differences between population groups, countries etc).

In this study we present the kernel method as a graduation technique and its application to a wide range of fields. At the outset, some methodological issues concerning this method are presented. In addition in order to illustrate the kernel technique as a graduation method we use it for graduating different mortality data sets.

Since this method provides a simple way of finding structure of data we also use for displaying the shape of income distribution of different populations. More specifically, Chapter 2 provides a review of existing density estimation techniques since kernel methods were firstly developed in order to estimate a probability density from a sample of observations. Chapter 3 is devoted to the presentation of the kernel method as an estimation technique. Here are presented the properties of the kernel estimates. In addition a review of the existing methods for the choice of the bandwidth parameter is provided. Chapter 4 focuses on graduation techniques. A review of the existing graduation methods is presented. Particular emphasis is given on kernel estimators used for graduation of mortality rates. Chapter 5 illustrates the results of the kernel graduation technique applied to mortality data sets. A discussion about the existing methods for the choice of the bandwidth parameter is provided. In addition we compare the classical kernel approach to the local linear approach. In order to evaluate the efficiency and accuracy of the kernel method with respect to graduation we compare with the Helligman-Pollard model with height parameters. In Chapter 6 we present the application of the kernel technique to income data. We refer to the ability of kernel estimates to provide a picture of income distributions in an informative way. Finally in Chapter 7 some concluding remarks are provided.

Chapter 2

DENSITY ESTIMATION

2.1 AN OVERVIEW OF METHODS FOR DENSITY ESTIMATION

The estimation of an unknown density function has generated a considerable literature over the past few years.

This happens because the density function is a fundamental statistical concept underlying the empirical frequency distribution. In fact, it can be thought of as the theoretical analogue of the frequency distribution, which captures the essential characteristics of the shape of the distribution.

The density function can be estimated by adopting a parametric or non-parametric approach.

The parametric approach requires that a particular functional form is specified which appears to be a good way of capturing the essential characteristics of the shape of the distribution.

The non-parametric approach can be thought of as an approach to sketch the curve that is traced out by the empirical distributions in order to fit an empirical frequency distribution.

The kernel method of density estimation has proved to be one of the most popular non-parametric approaches over the past few years.

This method of density estimation can be considered as an improvement of the already existing methods of non-parametric density estimation.

Before the analytic study of the kernel method of density estimation let us provide a presentation some of the methods of density estimation that are widely used for the purpose of estimating of a density function and have been developed before the kernel density estimation.

2.1.1 HISTOGRAMS

The method most widely used to represent the shape of a probability density function is the histogram. The histogram is calculated by partitioning the real line into equal sized line segments, called bins. The fraction of observations of the variable of

interest that fall within a given bin is taken as an estimate of the probability of observing future realizations in that bin. The histogram is drawn as a sequence of bars, representing the simplifying assumption that the probability density is constant within each bin.

This approach to probability has its beginnings with John Grant, (Tapia and Tompson (1970a)).

He developed a near histogram (it was not normalized for its cell size) out of the need to represent birth and death data of early 17th century in London.

The histogram maybe defined in the following manner:

First partition the range of the sample. If y_1 is the first order statistic, then we define

$$y_1=t_0<t_1<\ldots<t_{k-1}<t_k=y_N$$

Then calculate

$$\hat{f}_N(x)=\frac{q_i}{N(t_i-t_{i-1})},\ for\ x\in(t_i-t_{i-1})$$

where q_i is the number of sample points in the interval (t_i,t_{i-1}) and

$$\hat{f}_N(x)=\frac{q_2}{N(t_2-t_1)},$$

while $$\hat{f}_N(x)=0\ for\ x\notin[y_1,y_N]$$

If F_N is the sample distribution function, that is

$$F_N(x)=\frac{\#of\ observations\le x}{N}$$

then the histogram may also be written as:

$$\hat{f}_N(x)=\frac{F_N(t_i)-F_N(t_{i-1})}{t_i-t_{i-1}}$$

$for\ x \in (t_i - t_{i-1}), i = 1,2\ldots,k$

Rosenblatt (1956) proposed an extension of the classical histogram method. The estimator he proposed is given by:

$$\hat{f}_N(x) = \frac{F_N(x + h_N) - F_N(x - h_N)}{2h_N}$$

where $h_N \sim N^{-a}, 0 < a < 1$

and as before

$$F_N(x) = \frac{\#of\ observations \leq x}{N}$$

Silverman (1986) define the histogram density estimator as follows:
Let x be the variable of interest, m the number of bins, h the half width of each bin and f(x) the density function.
If x_o is the origin of the histogram, that is, the lower endpoint of the left most bin, the end points are given by the sequence

$$x_0,\ x_0 + 2h,\ x_0 + 4h, \ldots, x_0\ m(2h).$$

then a particular observation x_i falls in bin j, if:

$$x_0 + (j-1)2h \leq x_i < x_0 + j(2h).$$

The histogram estimate is given by:
for x located in bin j.

$$\hat{f}_N(x) = \frac{\#\left[x_o + (j-1)2h \leq x_i < x_o + j(2h)\right]}{n(2h)}$$

Advantages

- The most attractive feature of the histogram as a density estimation method is its simplicity in comparison with other methods.
- Furthermore it is an excellent tool for data exploration and presentation and in the univariate case they are a useful class of density estimates.

Disadvantages

- The histogram in many cases may be an inappropriate method for density estimation since the exploration of the data may be severely influenced by the choice of the origin.
- Another disadvantage is that although most densities are not step functions, the histogram has the unattractive feature of estimating all densities by a step function.
- Furthermore the discontinuity of histograms causes extreme difficulty if derivatives of the estimates are required.
- Because of its mathematical inaccuracy the histogram makes inefficient use of the data if it is used as a density estimator in procedures like cluster analysis and non-parametric discriminant analysis.
- A further problem is the extension of the histogram to the multivariate setting, especially the graphical display of a multivariate data set.
- In addition the choice of the amount of smoothing is required which may be considered as another inconvenient of the method.

So, more sophisticated methods than histogram are necessary to be used in practice for density estimation.

2.1.2 THE NAIVE ESTIMATOR

Rosenblatt (1956) generalized his initial estimate of histogram to a class of estimates, which turn out to be one of the most important non-parametric density estimates. This class of estimates is defined as follows:

Let $W_N(u)$ be a weighting function such that

$$\int W_N(u)du=1$$

where the integral is taken over a set which shrinks with N at an appropriate rate.

So, he proposes the following estimate of the pdf f:

$$\hat{f}_N(x) = \frac{\sum_{i=1}^{N} W_N(x - x_i)}{N}$$

Whittle (1958) proposed estimates similar to these suggested by Rosenblatt but he also considered the assumption that the sample size is a Poisson random variable with mean M provided the observations are independent and the 'stopping rule' for sampling is independent of the sample. Then he estimated the function $\varphi(x)=Mf(x)$, where the estimate of φ is :

$$\hat{\varphi}(x) = \int W_x(y) dN(y)$$

where N(y) is the number of observations ≤ y

Silverman (1986) suggested a density estimate with a rectangular weight function that provides a natural introduction to kernel density estimates. So, the density at a point x can be thought of as the limit of the height of a histogram bar centered at x as the half-width h of the bar goes to zero:

$$f(x) = \lim_{h \to \infty} \frac{1}{2h} P(x - h < X < x + h)$$

Thus, a simple density estimator is one, which replaces the probability in a small region (window) around x with the sample proportion, scaling the estimate so the total area under f(x) integrates to unity:

$$\int f(x) dx = 1$$

So, the density estimator is similar to a histogram with bin width equal to 2h,but with every point as the centre of a sampling interval, that is, with no fixed origin and therefore freeing the histogram from a particular choice of bin positions. Then this is called the naïve estimator.

$$\hat{f}_s(x) = \frac{1}{2h}\frac{\#[x-h < X_i < x+h]}{n}$$

If we consider the weight function

$$w(x) = \begin{cases} \frac{1}{2} & if\ |x| < 1 \\ 0 & otherwise \end{cases}$$

then the naïve estimator can be written in the form:

$$\hat{f}_s(x) = \frac{1}{nh}\sum_{i=1}^{n} S\left(\frac{x - X_i}{h}\right)$$

Advantages

- This method of density estimation has the advantage over the histogram that the number of bin positions is not fixed and consequently it reduces the effect of disturbing results because of the inappropriate choice of the bin edge.

Disadvantages

- The most important drawback of this method is the choice of the binwidth. This is governed by the parameter h, which controls the amount by which the data are smoothed to produce the data.
- Furthermore the density estimator $\int f''(x)^2$ is not a continuous function but has jumps at points $X_i \pm h$ and zero derivatives everywhere else. As a result we may have a misleading impression of the estimate since they usually are ragged.

2.1.3 THE KERNEL DENSITY ESTIMATOR

It should be noticed that the first published paper on kernel density estimation is by Rosenblatt (1956) who describes the naïve estimator and by Cacoullos (1966) for the multivariate case. However Fix and Hodges were those who first introduced the idea

of kernel density estimation in an unpublished paper in 1951 (This paper is however reprinted by B.W.Silverman and M.C. Jones).

As already mentioned, the kernel methods have been developed in order to estimate a probability density from a sample of observations.

Suppose we have a sample of n observations $y_1,\ldots,y_n$ from a density g(y) which it is required to estimate.

The general form of the kernel estimator of g(y) is given by:

$$\hat{g}(y)=\frac{1}{nh}\sum_{j=1}^{n}\psi\left(\frac{y-y_j}{h}\right)$$

for some function ψ, where h=h(n) is positive and $\to 0$ as $n\to\infty$.

Advantages

- Kernel density estimation is a very effective way of showing the structure of a set of data.
- Furthermore it has an easy to understand explicit definition, which enhance its popularity over other methods of density estimation.

Disadvantages

- The only practical drawback of the kernel method of density estimation is its inability to deal satisfactory with the tails of distributions without oversmoothing the main part of the density. In fact when applied to data from long-tailed distributions, because the window width is fixed across the entire sample, spurious noise appears in the tails of the estimates. If the estimates are smoothed sufficiently to deal with this the essential detail in the main part of the distribution is masked.

Chapter 3

KERNEL DENSITY ESTIMATION

3.1 THE KERNEL ESTIMATOR IN THE UNIVARIATE CASE

Parzen (1962) introduced the kernel estimate as a weighted average over the sample distribution function. So, he proposes the following estimate:

$$\hat{f}_M(x) = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x-y}{h}\right) dF_M(y)$$

where $$F_n(x) = \frac{\#\, of\ observations \le x}{n},$$

or the estimator can be written as:

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

He considers kernels, which satisfy the following conditions:

i. $sup|K(u)| < \infty$

ii. $\int_{-\infty}^{\infty} |K(u)| du < \infty$

iii. $\lim_{|u| \to \infty} |uK(u)| = 0$

iv. $\int_{-\infty}^{\infty} K(u) du = 1$

3.1.1 PROPERTIES OF KERNEL ESTIMATES

Some elementary properties of kernel estimates follow at once from the definition.

- The kernel estimator can be considered as a sum of 'bumps' placed at the observations. The kernel function K determines the shape of the bumps while the window width h determines their width.
- Provided the kernel K is everywhere non-negative and satisfies the condition (iv)- in other words is a probability density function- it follows that $\hat{f}_x$ will itself be a probability density.
- $\hat{f}_x$ will inherit all the continuity and differentiability properties of the kernel K. So, if K is the normal density function the $\hat{f}_x$ will be a smooth curve having derivatives of all orders.
- Sometimes we can use kernels, which take negative as well positive values and then the estimate may itself be negative in places. However for most practical purposes non-negative kernels are used.

3.1.2 SAMPLE PROPERTIES OF KERNEL ESTIMATES

The sample properties of the kernel estimator are studied. The expressions for the mean value and the variance are derived directly from the definition of the kernel estimator.

So, the mean value of the kernel estimator is defined as:

$$E\,\hat{f}(x) = \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)dy$$

while the variance of the kernel estimator (Silverman, 1986) is given by the following expression:

$$\operatorname{var}\hat{f}(x) = \frac{1}{n}\int \frac{1}{h^2} K\left(\frac{x-y}{h}\right)^2 f(y)dy - \left\{\frac{1}{h}\int K\left(\frac{x-y}{h}\right) f(y)dy\right\}^2$$

Finally the bias of the kernel estimator is given by the formula:

$$bias_h(x) = E\hat{f}(x) - f(x)$$

$$= \int \frac{1}{h} K\left\{\frac{(x-y)}{h}\right\} f(y)dy - f(x)$$

The expressions of the mean and variance can be written in an alternative way if the convolution notation is used (Wand and Jones, 1995).

The convolution of two functions, f and g is defined as:

$$(f * g)(x) = \int f(x-y)g(y)dy$$

Using this notation the expression of the variance is written as:

$$\text{var}\,\hat{f}(x) = n^{-1}\left\{(K_h^2 * f)(x) - (K_h * f)^2(x)\right\}$$

while the bias is now expressed as:

$$bias\hat{f}(x) = E\hat{f}(x) - f(x) = (K_h * f)(x) - f(x)$$

Although these expressions are important because they are used in order to obtain the exact expressions for measures of discrepancy such as the MSE and MISE criteria, but except in very special cases their calculations become intractable and therefore it is more instructive to obtain approximations of them.

3.1.3 APPROXIMATE PROPERTIES OF KERNEL ESTIMATES

Silverman (1986), using Taylor series expansion gives the approximate expressions for the bias and the variance of the kernel density estimator.

For simplicity it is assumed that the kernel K is a symmetric function satisfying the conditions: i) $\int K(t)dt = 0$ ii) $\int tK(t) = 0$ and iii) $\int t^2 K(t)dt = k_2 \neq 0$

where k_2 is a constant. In addition it is assumed that the unknown density f has continuous derivatives of all orders required.

So, under these assumptions the approximate expression for the bias is:

$$bias\ _n(x) = -hf'(x)\int tK(t)dt + \frac{1}{2}h^2 f''(x)\int t^2 K(t)dt + \ldots$$
$$= \frac{1}{2}h^2 f''(x)k_2 + higher\ \ order\ \ terms\ \ in\ h$$

while the variance is approximately given by the expression :

$$\text{var}\,\hat{f}(x) \approx n^{-1}h^{-1}f(x)\int K(t)^2\,dt$$

These approximate expressions of the bias and the variance rather than true ones are used in order to investigate how the mean square error and the mean integrated square error behave in the case of the kernel density estimator.

It is important to notice that the bias in the estimation of f(x) does not depend directly on the sample size but it does depend on the window width h.

Furthermore the choice of the smoothing parameter implies a trade-off between random and systematic error. Since if in an attempt to eliminate the bias, a very small value of h is chosen then the integrated variance will become large. On the other hand when a large value of h is used, the random variation as quantified by the variance is reduced at the expense of introducing systematic error or bias into the estimation.

The trade-off between the bias and variance terms is considered as one of the most fundamental problems in density estimation. Consequently the choice of the smoothing parameter h is of great importance and attention should be given in methods that are used for the choice of h.

3.1.4 ASYMPTOTIC PROPERTIES OF KERNEL ESTIMATES

Unbiasdness and consistency are important properties of any density estimator. Therefore attention should be paid to conditions under which the kernel estimate is consistent and if it is an unbiased estimator of the true density.

Parzen (1962) studied when a kernel estimator is asymptotically unbiased as well as the conditions required in order to be consistent.

Firstly he assumes that h should be chosen as a function of n which tends to 0 as n tends to ∞.

i) Unbiasdness of the kernel estimator

An estimator is asymptotically unbiased in the sense that if h=h(n) is chosen as a function of n such that

$$\lim_{n\to\infty} h(n) = 0 \qquad (2.1.1)$$

then

$$\lim_{n\to\infty} E[f_n(x)] = f(x)$$

A kernel estimator is asymptotically unbiased at all points x at which the probability density function is continuous if the constants h satisfy (2.1.1) and if the kernel K is a bounded Borel function satisfying the conditions

$$\sup_{-\infty<y<\infty}|K(y)| \qquad (2.1.2)$$

$$\int_{-\infty}^{\infty}|K(y)dy| = 0 \qquad (2.1.3)$$

$$\lim_{y\to\infty}|yK(y)dy = 0| \qquad (2.1.4)$$

and in addition

$$\int_{-\infty}^{\infty} K(y)dy = 1 \qquad (2.1.5)$$

ii) Consistency of the kernel estimator

A kernel estimator $f_n(x)$ is a consistent estimate in quadratic mean of f(x), in the sense that

$$E|f_n(x) - f(x)|^2 \to 0 \qquad as \quad n\to\infty$$

if the conditions (2.1.1)-(2.1.5) are satisfied and in addition the constants h=h(n) satisfy the condition

$$\lim_{n\to\infty} nh(n) = \infty$$

The conditions under which consistency is achieved, imply that while the window width must get smaller as the sample size increases it must not converge to zero as rapidly as n^{-1}. This means that the expected number of points in the sample falling in the interval $x \pm h_n$ must tend to infinity as n tends to infinity but with a slow rate (Silverman, 1986).

3.2 MEASURES OF DISCREPANCY

The theoretical treatment when using various methods of density estimation is to study the closeness of the density estimator $\hat{f}$ to the true density f.

Various measures of the discrepancy of the density estimator $\hat{f}$ from the density f have been studied. Firstly it is specified an error criterion for measuring the error when estimating the density at a single point and then when estimating the density over the whole real line.

So, the mean square error (MSE) is a natural measure in the case of a single point estimation, which is defined as:

$$MSE(f) = E\left\{\hat{f}(x) - f(x)\right\}^2 = \left\{E\hat{f}(x) - f(x)\right\}^2 + Var\hat{f}(x)$$

i.e. it is expressed as the sum of the squared bias and the variance at x.

This error criterion is often preferred to other criteria such as mean absolute error, which is defined as:

$$MAE(\hat{f}) = E\left|\hat{f} - f\right|$$

since it is mathematically simpler to work with.

Furthermore the variance-bias decomposition allows easier analysis and interpretation of the performance of the kernel density estimator.

Another measure of the global accuracy of $\hat{f}$ as an estimator of f is the mean integrated squared error (MISE) (Rosenblatt, 1956). This is defined by:

$$MISE\left(\hat{f}\right) = E\int\left\{\hat{f}(x) - f(x)\right\}^2 dx$$

$$= \int E\left\{\hat{f}(x) - f(x)\right\}^2 dx$$

$$= \int MSE_x\left(\hat{f}\right)dx$$

$$= \int \left\{E\hat{f}(x) - f(x)\right\}^2 dx + \int Var\hat{f}(x)dx$$

The MISE is preferred for measuring the global performance of the kernel density estimator because of its mathematical simplicity.

Using the convolution notation the expressions for the MSE and MISE become:

$$MSE\left(\hat{f}(x)\right) = n^{-1}\left\{\left(K_h^2 * f\right)(x) - \left(K_h * f\right)^2(x)\right\} + \left\{\left(K_h * f\right)(x) - f(x)\right\}^2$$

while,

$$MISE\left(\hat{f}(x)\right) = n^{-1}\int\left\{\left(K_h^2 * f\right)(x) - \left(K_h * f\right)^2(x)\right\}dx + \int\left\{\left(K_h * f\right)(x) - f(x)\right\}^2 dx$$

3.2.1 APPROXIMATE EXPRESSIONS OF MSE AND MISE

The expressions of MSE and MISE are rather complicated and it is difficult to investigate how these behave. Also it is difficult to interpret the influence of the bandwidth on the performance of the kernel density estimator. So, in order to overcome this problem, approximate expressions for MSE and MISE can be used, which are simpler and allow deeper understanding of how the techniques behave without having to grasp complicated formulae.

These approximate expressions are obtained if we use the approximate expressions for the bias and the variance (Wand and Jones, 1995). So, the asymptotic MSE approximation is:

$$MSE\left(\hat{f}\right) = (nh)^{-1} f(x)\int K(t)^2 dt + \frac{1}{4}h^4 k_2 \int f''(x)^2 dx$$

while the asymptotic MISE approximate expression is:

$$MISE\left(\hat{f}\right) \approx \frac{1}{4}h^4 k_2{}^2 \int f''(x)^2 dx + n^{-1}h^{-1}\int K(t)^2 dt\,.$$

3.3 CHOICE OF THE BANDWIDTH PARAMETER

In kernel density estimation methods, as already mentioned, it is required the specification of the bandwidth h.

There are cases where the choice of the bandwidth could be done subjectively by eye or other where it is more beneficial to select it automatically from the data.

The first approach is preferred when there is knowledge about the structure of the data while the second is used when there is no prior knowledge of the structure of them or any suspicion about the bandwidth that could give an estimate close to the true density.

Generally the bandwidth selectors can be divided in two classes (Wand and Jones, 1995). These are "quick and simple" and data-driven bandwidth selectors.

The first class consists of bandwidth selectors that are easily computed but they are not always approximate the optimal bandwidth. Usually they are used in order to provide starting points for subjective choice of the bandwidth.

The second class consists of bandwidth selectors that are consistent with respect to MISE i.e. they minimize the mean integrated squared error. These bandwidth selectors require more computational effort.

Many data based methods have been proposed for selecting the bandwidth but none of these proposals has gained wide acceptance.

3.3.1 OPTIMAL SMOOTHING PARAMETER

Before these methods are presented more analytically, it is given the ideal value of the window width since the aim of the use of most of the methods is to approximate this value.

The ideal value of the smoothing parameter h (Parzen, 1962, Lemma 4A) is:

$$h_{opt} = k_2^{-\frac{2}{5}} \left\{ \int K(t)^2 dt \right\}^{\frac{1}{5}} \left\{ \int f''(x)^2 dx \right\}^{-\frac{1}{5}} n^{-\frac{1}{5}}$$

This value can be obtained if we minimize the approximate mean integrated square error:

$$\frac{1}{4}h^4 k_2^2 \int f''(x)^2 dx + n^{-1}h^{-1}\int K(t)^2 dt$$

Then according to Silverman (1986), the following conclusions can be drawn:

i. h_{opt} itself depends on the unknown density being estimated

ii. the ideal window width will converge to zero as the sample size increases but at a very slow rate

iii. since the term $\int f''^2$ measures in a sense the rapidity of fluctuations in the density, smaller values of h will be appropriate for more rapidly fluctuating densities.

At the sequel we present some methods existed in the literature which are used for the choice of the bandwidth parameter.

3.3.2 BANDWIDTH SELECTORS

Quick and simple bandwidth selectors

1) Rule of Thumb

The simplest proposals are various versions of the 'rule of thumb' (Silverman, 1986; Hardle, 1991).

The idea of this method is that we could estimate the unknown term $\int f''(x)^2$ in the expression of the optimal bandwidth assuming that it belongs to a prespecified class of density functions.

Thus, if the unknown distribution is normal with parameters μ and σ then:

$$\begin{aligned}\|f''\|_2^2 &= \int f''(x)^2 dx \\ &= \sigma^{-5}\int \phi''(x)^2 dx \\ &= \sigma^{-5}\frac{3}{8\sqrt{5}} \\ &\approx 0.2.12\sigma^{-5}\end{aligned}$$

where ϕ is the standard normal density.

Then if the gaussian kernel is being used the rule of thumb suggest estimating $\|f''\|_2^2$ through an estimator $\hat{\sigma}$ for σ. So, in this case the expression for the optimal bandwidth is:

$$h_{opt} = \left(\frac{\|\phi\|_2^2}{\|\hat{f}''\|_2^2 \mu_2^2(\phi) n} \right)^{\frac{1}{5}} = \left(\frac{4\hat{\sigma}^5}{3n} \right)^{\frac{1}{5}} \approx 1.06\,\hat{\sigma} n^{-\frac{1}{5}}$$

Instead of $\hat{\sigma}^2$ more robust estimate for the scale parameter of the distribution can be used such that the the interquantile range $\hat{R}$ which is defined as $\hat{R} = X_{[0.75n]} - X_{[0.25n]}$.

In this case the rule of thumb is modified into: $h_{opt} = 0.79\hat{R}n^{-\frac{1}{5}}$. Furthermore if instead of σ is used the adaptive estimate A=min(standard deviation, interquantile range/1.34) then the rule of thumb is modified into:

$$h_{opt} = 1.06 \min\left(\hat{\sigma}, \frac{\hat{R}}{1.34} \right) n^{-\frac{1}{5}}$$

Advantages

The advantage of this method is that it provides a quick "first guess" bandwidth (Wand and Jones, 1995).

- Also it gives reasonable results when the data follow a normal distribution.

Disadvantages

Although this method is satisfactory when the underlying density is close to Gaussian, in every other case it oversmooths and mask important features in the data (Hall, Sheather, Jones and Marron, 1991).

- Furthermore the bandwidth does not converge at all to the optimum although its probability mass does have a relative rate of convergence to its mean of $n^{-1/2}$ (Silverman, 1986).

These bandwidths lead to smooth estimates but fail to detect the bimodality of the density (Hardle, 1991)

2) *Maximal smoothing principle*

A widely applicable method for choosing the smoothing parameters for non-parametric density estimators is this which makes use of "the maximal smoothing principle" (Terrel, 1990).

The maximal smoothing principle suggests that we should choose the largest degree of smoothing that is compatible with the estimated scale of the density.

In order to choose the best degree of smoothing, we must have a criterion for optimality of density estimates. This will be to make the expected L^2 metric $E\left(\left(\int \hat{f}(y) - f(y)\right)^2\right)$ or mean integrated squared error (MISE) as small as possible.

So, the maximal smoothing principle underlies the idea that measures of scale tend to place upper-bounds on the smoothing parameters that minimise the asymptotic mean integrated squared error of density estimates.

In the case of fixed kernel estimates the MISE is asymptotically minimised by choosing the smoothing parameter h to be (Parzen, 1962):

$$h = \left[\frac{\int K^2(x)dx}{n\sigma_k^4 \int f''(x)^2 dx}\right]^{1/5}$$

for sufficiently smooth f.

The problem then can be thought of as the minimisation of $\int (f'')^2 \left[or \int (f')^2\right]$ with the constraint that $T(F) = T\left(\hat{F}_n\right)$, where T is a measure of scale, such as the standard deviation and f is a density.

So, knowing some measure of the scale of the underlying density we can obtain tight upper bounds on the asymptotically kernel window parameter.

Terrel and Scott (1985) suggested computing the bounds by using the range as a scale statistic.

Advantages

The main advantage of this method is that it tends to eliminate accidental features such as asymmetries and multiple modes that could have come about by chance.

Furthermore it avoids the extreme sampling variability of cross-validation by using ordinary scale estimators such as the standard deviation and interquantile range. These

scale estimators have order n-1 variability, while cross-validated parameters have orders of variability such as n-1/5.

Disadvantages

Maximal smoothing parameters are conservative rather than asymptotically optimal. They tend to retain information so they should be used in conjuction with other data displays that retain more of the features of the original sample.

Data-driven bandwidth selectors

1) **Cross-Validation**

Cross-validation is an automatic and simple method for selecting a bandwidth that reflects the data but also considers smoothness. There are two forms of cross-validation: maximum likelihood CV and least-squares CV.

Maximum Likelihood Cross-Validation

The method of likelihood cross-validation is based on the idea of using likelihood to judge the adequacy of fit of a statistical model.

The score function $CV_{KL}(h)$ was suggested and by Habbema, Herbmans and Van Der Broek (1974). It is defined as:

$$CV_{KL}(h) = n^{-1}\sum_{i=1}^{n}\log \hat{f}_{-i}(X_i)$$

$$= n^{-1}\sum_{i=1}^{n}\log\left[\sum_{j\neq i}K\left(\frac{X_i - X_j}{h}\right)\right] - \log[(n-1)h]$$

The value of h is obtained from the maximization of the function CV(h) for the given data. Thus,

$$h_{opt} = \max CV_{KL}(h)$$
$$= \max n^{-1}\log \hat{f}_{-i}(X_i)$$

It must be noticed that maximizing $CV_{KL}(h)$ is similar to optimizing the Kullback-Leibber information: $d_{KL}(f,\hat{f}) = \int \log\left(\frac{f}{\hat{f}}\right)(x)f(x)dx$

Advantages

- This bandwidth selector is of general applicability not just in density estimation (Stone, 1974; Geisser, 1975).
- Also the use of this score function does not present severe computational difficulties (Silverman, 1986).

Disadvantages

- A disadvantage of this method is that in the case of identical observations in one point, $CV_{KL}(h)$ may have an infinite value and hence cannot be defined an optimal bandwidth
- Also, $CV_{KL}(h)$ is very sensitive to outliers (Scott and Factor, 1981). To overcome this problem a large bandwidth is used and this leads to a slight oversmoothing for the other observations.
- Schuster and Gregory (1981) showed that the use of likelihood cross-validation may lead to inconsistent estimates of the true density if the true probability density function has a sufficiently long tail.
- Finally, in order to obtain good estimates of the quantity of interest, they must be used large samples.

Least-Squares Cross-Validation

Least-squares cross-validation was firstly suggested by Rudemo (1982) and Bowman (1984). Silverman (1986) considers least-squares cross-validation as the most widely studied data-based bandwidth selector.

Rudemo (1982) proposes that the smoothness of an estimate should be dictated by minimizing an estimate of a quadratic risk function.

So, in least-squares cross-validation, we consider an alternative measure of distance between $\hat{f}$ and f, the Integrated Squared Error (ISE), which is defined as:

$$d_1(h) = \int \left(\hat{f} - f\right)^2 (x)dx \ .$$

$$= \int \hat{f}^2(x)dx - 2\int \left(\hat{f}f\right)(x)dx + \int f^2(x)dx$$

Since the last term is independent of $\hat{f}$, the ideal choice of the window width (in the sense of minimizing the integrated squared error) would be the same as if minimizing the quantity

$$R(\hat{f}) = \int \hat{f}^2(x)dx - 2\int (\hat{f}f)(x)dx .$$

The basic principle of least-squares cross-validation is to construct an estimate of $R(\hat{f})$ from the data themselves and then to minimize this estimate over h to give the choice of window width.

If $\hat{f}_{-i}$ is the density estimate constructed from all the data points except X_i i.e.

$$\hat{f}_{-i}(x) = (n-1)^{-1} h^{-1} \sum_{i \neq j} K\{h^{-1}(x - x_j)\}$$

and

$$CV(h) = \int \hat{f}^2(x)dx - 2n^{-1} \sum_i \hat{f}_{-i}(x_i)$$

then cross-validation implies the minimization of CV(h) over h.

Thus,

$$h_{opt} = \min CV(h)$$

$$= \min(\int \hat{f}^2(x)dx - 2n^{-1} \sum_{i=1}^{n} \hat{f}_{-i}(X_i))$$

Advantages

- Stones (1984) proved that asymptotically, that least-squares cross-validation achieves the best possible choice of smoothing parameter in the sense of minimizing the integrated square error.

Disadvantages

- A disadvantage of this method is that it suffers from too much sample variability (Park & Marron, 1990) even though it avoids the oversmoothing problem of the rule of thumb selector. This means that different data sets from the same

distribution will all too often give results which are very different and this make this method unacceptable in practice.

- Furthermore the theoretical and practical performance of this bandwidth selector is rather disappointing (Hall and Marron, 1987a; Park and Marron, 1990).
- Also, it does converge to the optimum with the slow relative rate of $n^{-1/10}$ (Hall and Marron, 1987 b).

2) Biased Cross-Validation method

The biased cross-validation method (Scott & Terrel, 1987) can be considered as an 'improvement' of cross-validation. This bandwidth selector is based on the minimisation of the asymptotic mean integrated squared error i.e of:

$$AMISE\left(\hat{f}\right) = (nh)^{-1}\int K(x)^2\,dx + \frac{1}{4}h^4 k_2{}^2 \int f''(x)^2\,dx$$

if the quantity

$$R(f'') = \int f''(x)^2\,dx$$

is estimated by :

$$\tilde{R}(f'') = R\left(\hat{f}\right) - \left(nh^5\right)R(K'')$$

$$= n^{-2}\sum\sum \left(K_h'' * K_h''\right)\left(X_i - X_j\right)$$

Advantages

- The biased cross-validation method can be considered as an 'improvement' of cross-validation in the sense that although it has an $n^{-1/10}$ rate of convergence it has an improved small sample performance in some cases.

Disadvantages

- The main disadvantage is that the reduction in variance results in an increase in bias.

3) Plug-in selectors

Another possible bandwidth selector is the plug-in selector. In fact this was the first proposed method for using data to choose the bandwidth of a kernel density estimator (Woodroofe, 1970).

Plug-in bandwidth selectors are based on the idea of substituting estimates into the asymptotic representation of the optimal bandwidth.

So, the plug-in bandwidth selector requires the minimization of the AMISE optimal bandwidth i.e. of the quantity:

$$h_{AMISE} = \left\{ \frac{\int K(x)^2 dx}{\sigma_K^4 \int f''(x)^2 dx} \right\}^{1/5}$$

where the unknown quantity

$$R(f'') = \int f''(x)^2 dx$$

is estimated by $R(\hat{f}_a'')$ where $\hat{f}_a$ is a kernel density estimate with a different bandwidth α. Then α should be expressed in terms of h and then solve the resulting expression of h_{AMISE} for h.

Advantages

- This bandwidth selector is very efficient when the underlying density is smooth (Park and Marron, 1990).
- Also Park and Marron (1990) consider this method as the most practical existing method.

Disadvantages

- This bandwidth selector requires the appropriate choice of the estimates that will be plugged-in.
- When there is not enough smoothness present it may give not robust results.

3.3.3 COMPARISON OF DATA-DRIVEN BANDWIDTH SELECTORS

Various data –driven methods for choosing the bandwidth have been proposed and studied. There is interest in finding which is the most promising of them since all have cases where they perform best but also have several drawbacks.

A mean of comparing the bandwidth selectors is by asymptotic rate of convergence to the optimum (Park and Marron, 1990; Wand and Jones, 1995). While computer simulation is also an important tool for the comparison of bandwidth selectors.

The bandwidth $\hat{h}$ has a relative rate of convergence to h_0 of order n^{-v} with asymptotic variance σ^2 if it satisfies the following:

$$n^{v}\left(\frac{\hat{h}}{h_0}-1\right)\rightarrow_D N\left(\mu,\sigma^2\right)$$

where μ and $\sigma^2>0$ depend only on f and K and h_0 is the optimal bandwidth.

The rate of convergence of the least-squares cross-validated bandwidth as well as of the biased cross-validated bandwidth is $n^{-1/10}$ (Hall and Marron, 1987a; Scott and Terrel, 1987).

3.4 CHOICE OF THE KERNEL FUNCTION

From the definition of the kernel estimator it follows that apart from the choice of the bandwidth that plays an important role in kernel density estimation, the effect of the shape of the kernel function should be investigated.

Usually the kernel function is a unimodal probability density function that is symmetric about zero and in addition it satisfies the condition:

$$\int K(x)dx=1$$

The above situation ensures that $\hat{f}$ will also be a density.

There are of course kernels that do not satisfy the above requirements, these are not preferred not only for reasons of simplicity in interpretation but also because they are in a sense inadmissible.

Since there are many kernel functions that satisfy these requirements there is interest in comparing them in order to find whether certain kernel shapes perform better than others.

The optimal value of the kernel is obtained by the formula of AMISE when minimizing it with respect to K.

In the case where the bandwidth is chosen optimally, the approximate value of the mean integrated square error will be

$$\frac{5}{4}C(K)\left\{f''(x)^2\,dx\right\}^{\frac{1}{5}} n^{-\frac{4}{5}}$$

where the constant C(K) is given (Silverman, 1986) by:

$$C(K) = k_2^{2/5}\left\{\int K(t)^2\,dt\right\}^{4/5}$$

Then a small value of the mean integrated square error is obtained if the kernel K is chosen such that C(K) will be small.

So, the problem of determining the optimal kernel is equivalent of minimizing C(K) subject to the constraints:

$$\int K(x)dx = 1, \qquad \int xK(x)dx = 0, \qquad \int x^2K(x)dx = a^2$$

as well as

$$K(x) \geq 0 \quad \text{for all } x.$$

This problem is solved by setting K(x) to be the Epanechnikov kernel (firstly described by Epanechnikov in 1969, Hodges and Lehman, 1956) i.e. if:

$$K(x) = K_e(x) = \begin{cases} \dfrac{3}{4\sqrt{5}}\left(1 - \dfrac{1}{5}x^2\right) & -\sqrt{5} \leq x \leq \sqrt{5} \\ 0 & \textit{otherwise} \end{cases}$$

3.4.1 EFFICIENCY OF THE KERNEL FUNCTIONS

The efficiency of any symmetric kernel is defined to be (Silverman, 1986):

$$eff(K) = \left\{\frac{C(K_e)}{C(K)}\right\}^{\frac{5}{4}}$$

$$= \frac{3}{5\sqrt{5}}\left\{\int x^2 K(x)dx\right\}^{-\frac{1}{2}}\left\{\int K(x)^2 dx\right\}^{-1}$$

Thus the efficiency of any symmetric kernel is obtained by comparing it with the Epanechnikov kernel and represents the ratio of sample sizes necessary to obtain the same minimum AMISE when using the kernel K_e as well as K (Wand and Jones, 1995).

Below are presented efficiencies of several kernels compared to the optimal kernel.

Some kernels and their efficiencies

Kernel	K(x)	Efficiency
Epanechnikov	$\frac{3}{4}\left(1-\frac{1}{5}x^2\right)/\sqrt{5}, for\|x\|<\sqrt{5}$ 0 ,otherwise	1
Biweight	$\frac{15}{16}(1-x^2)^2, for\|x\|<1$ 0 otherwise	0.994
Triweight	$\{2^7 B(4,4)\}^{-1}(1-x^2)^3 \quad for\|x\|<1$ 0 otherwise	0.987
Normal	$\frac{1}{\sqrt{2\pi}}e^{-\left(\frac{1}{2}\right)x^2}$	0.951
Triangular	$1-\|x\|, \quad for\|x\|<1$	0.986

	0 otherwise	
Rectangular	$\frac{1}{2}\ for\|x\|<1$ 0 otherwise	0.9295

From this table the following conclusions can be drawn:

The efficiencies of all these kernels are very close to 1 and they perform almost the same, so in the sense of minimizing the mean integrated error there is no difference no matter which kernel is being used.

Furthermore the choice of the kernel should be done in terms of other considerations such as computational efficiency or the degree of differentiability required.

3.4.2 HIGHER ORDER KERNELS

A kernel K is a kth –order kernel if:

$$\mu_0(K)=1,\quad \mu_j(K)=0\ for\ j=1,\ldots,k-1\quad and\quad \mu_k(K)\neq 0$$

where $$\mu_j(K)=\int x^j K(x)$$

is the jth moment of the kernel K.

Furthermore K is supposed to be symmetric.

The kernels of higher order are used in order to achieve a best rate of convergence or in other words reducing the order of the approximate bias. This happened because the kernel is not anymore constrained to be a probability density function.

Although the use of higher order kernels may lead in a better rate of convergence it also results in an increase of the error of the sample sizes except for the case where the sample size is very large,

Furthermore higher order kernels take negative values, which makes more difficult the interpretation of the resulting estimators.

Also the resulting estimate will have similar behavior to these kernels and it will not be a density itself.

3.5 TRANSFORMED KERNEL DENSITY ESTIMATORS

Although kernel density estimators provide good results in many cases, they do not perform well in others and therefore it is necessary the use of modified kernel density estimators.

Improvement of univariate or multivariate kernel density estimates may be achieved when letting the bandwidth vary by the point of observation and by the point of the sample observation.

Below are given some kernel density estimators that differ from the fixed kernel estimator but none of them has been widely accepted.

3.5.1 LOCAL KERNEL DENSITY ESTIMATORS

The local kernel density estimator is given by:

$$\hat{f}_L(x;h(x)) = \{nh(x)\}^{-1}\sum_{i=1}^{n} K\left\{(x - X_i)/h(x)\right\}$$

This estimator is also called the balloon estimator due to Tukey and and Tukey (1981).

The difference from the basic kernel density estimator is that the local kernel density estimator has a different bandwidth h(x) for each point x at which f(x) is estimated, i.e. the bandwidth depends on the density of observations near the point it is estimated.

In the multivariate case the local kernel density estimator is:

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

These estimators belong to the general class of nearest neighbour estimators.

The k-th nearest neighbour estimator in d dimensions of Loftsgaarden and Quesenberry (1965) is defined as:

$$\hat{f}(x) = \frac{k}{nV_d h_k(x)^d}$$
$$= \frac{1}{nh_k(x)^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_k(x)}\right)$$

where $h_k(x)$ is the Euclidean distance from x to the k-th nearest data point and V_d is the volume of the unit sphere S_d in R^d.

So, the kernel as well as the nearest neighbor methods will give identical estimates for every value of k in the second case that corresponds to a kernel estimate with a certain value of h.

Advantages

- The most important advantage of this estimator is that it has a straightforward asymptotic analysis (Mack and Rosenblatt, 1979).
- In addition these estimates have the advantage of reducing the variance in the tails but with an increase in bias.

Disadvantages

- A disadvantage of this estimate is that it will not be a density even when K is. Furthermore is not everywhere differentiable.
- Also these estimators are prone to local noise and have very heavy tails. Furthermore as all the nearest-neighbor estimators have a poor performance in low dimensions.

3.5.2 VARIABLE KERNEL DENSITY ESTIMATORS

These are defined as follows:

$$\hat{f}_V(x;a) = n^{-1} \sum \{\alpha(X_i)\}^{-1} K\left\{\frac{x - X_i}{\alpha(X_i)}\right\}$$

In this class of estimators the bandwidth parameter is replaced by n values $a(X_i)$

This estimator is called a "sample smoothing estimator" (Scott and Terrel, 1992) and can be considered as a mixture of identical but individually scaled kernels centered at each observation.

Furthermore variable kernel method of density estimation is a method in which the amount of smoothing is adapted to the local density of the data.

These estimators can be considered as a special case of a general class of density estimators the "adaptive kernel estimates".

In fact adaptive kernel estimators are based on a two-stage procedure. Firstly, it is obtained an estimate of the data in order to have an idea of the density but also this enables us to have an idea about the possible pattern of the bandwidths corresponding to different observations, which are then used for the construction of the adaptive estimator.

Silverman (1986) describes the procedure for obtaining an adaptive estimator as follows:

i) It is found a pilot estimate $\tilde{f}(t)$ that satisfies $\tilde{f}(X_i) > 0$ for all i.

ii) Local bandwidth factors are defined by

$$\lambda_i = \left\{\tilde{f}(X_i)/g\right\}^{-a}$$

where g is the geometric mean of the $\tilde{f}(X_i)$:

$$\log g = n^{-1}\sum \tilde{f}(X_i)$$

and α is the sensitivity parameter, which satisfies the inequality:

$$0 \le \alpha \le 1.$$

iii) The adaptive kernel estimate is defined by:

$$\hat{f}(t) = n^{-1}\sum_{i=1}^{n} h^{-d}\lambda_i^{-d} K\left\{h^{-1}\lambda_i^{-1}(t - X_i)\right\}$$

where K is the kernel function and h is the bandwidth.

So according to this procedure, pilot estimation is necessary in order to obtain $\alpha(X_i)$. But the construction of the pilot estimator will be based on another density estimation method such as kernel density estimation or nearest neighbor method.
The Abramson estimator (1982) is of this form where $\alpha(X_i)$ equals

$$\alpha(X_i) = hf^{-\frac{1}{2}}(X_i)$$

and is considered as a good choice because then a bias of order h^4 is achieved.
Breiman, Meisel and Purcell (1977) suggested an estimate, which is a special case of the adaptive kernel estimates and has the form of a variable kernel estimator. They considered as a pilot estimator a nearest neighbor estimate with a large value of the smoothing parameter k and the sensitivity parameter α to be equal with 1/d, where d is the dimensionality of the space in which the density is being estimated.
Nevertheless the general view in the literature (Breiman, Meisel and Purcell, 1977; Abramson, 1982; Silverman, 1986) is that any convenient estimator can be used as a pilot and a possible choice would be a fixed kernel estimate with a bandwidth chosen with a reference to a life table.

Advantages

- The variable kernel estimate will itself be a probability density function in the case where the kernel K is and consequently it will have all the local smoothness properties of the kernel.

Disadvantages

- The most important drawback of this estimator is that it may be severely influenced by observations that are very far away and not only by points nearby.

3.5.3 TRANSFORMATIONS OF KERNEL DENSITY ESTIMATORS

In many cases it is necessary to transform the data in order to be able to estimate their density. This may be the case when the underlying density has features that require different amount of smoothing at different locations and therefore the fixed kernel density estimator does not perform well. When changing the data, the estimate of the

density of the transformed data can be back-transformed to obtain an estimate of the density of the original data.

The use of transformations in kernel density estimation is studied by Devroye and Guorfi (1985), Silverman (1986) and Wand, Marron and Ruppert (1991), Rudemo(1991), as well as by Ruppert and Cline(1994).

The theory for transformed density estimators can be described as follows (Wand, Marron and Ruppert, 1991):

Let $X_1,\ldots,X_n$ be a random sample with density f_x and support $S(f_x)$. Also consider a family of monotonic increasing transformations $\{\tilde{g}_\lambda : \lambda \in \Lambda\}$ that map $S(f_x)$ into the real line. If X is a random variable corresponding to the untransformed data and $\tilde{Y} = \tilde{g}(X)$ then the transformed variable Y is given by:

$$Y = \left(\sigma_X / \sigma_Y\right)\tilde{Y} = g_\lambda(X)$$

Then the density is given by:

$$f_Y(y;\lambda) = f_X\left(g_\lambda^{-1}(y)\right)\left\{\frac{d}{dy}g_\lambda^{-1}(y)\right\}$$

The kernel density estimate will be:

$$\hat{f}_Y(y;h,\lambda) = n^{-1}\sum_{i=1}^{n} K_h\left(y - Y_i\right)$$

while the density estimate of f_X is the back-transform:

$$\hat{f}_X(x;h,\lambda) = n^{-1}\sum g'_\lambda(x)K_h\{g_\lambda(x) - g_\lambda(X_i)\}.$$

This is called a transformation-kernel density estimator (TKDE) (Ruppert and Cline, 1994).

The choice of the transformation depends on the shape of the density of the data. In the case of a symmetric density f with a large amount of kurtosis an appropriate transformation would be a concave to the left and convex to the right of the centre of the symmetry of f (Ruppert and Wand, 1992).

In the case of skewed data then the transformation could be a convex one that belongs to the shifted power transformation family with the aim of reducing the skewness of f (Wand, Marron and Ruppert, 1991).

Such transformations may be the two-parameter sifted power transformation:

$$\tilde{g}_{\lambda_1,\lambda_2}(x) = (x+\lambda_1)^{\lambda_2} sign(\lambda_2), \qquad \lambda_2 \neq 0$$
$$= \ln(x+\lambda_1), \qquad \lambda_2 = 0$$

where $\lambda_1 > -\min(X)$ and min(X) denotes the lower endpoint of the support of f, or the Box-Cox transformation.

Furthermore it must be noticed that transformation kernel density estimators can be used in order to reduce the bias and appear to be more effective at small sample sizes and for densities with multiple modes than higher-order kernels (Ruppert and Cline, 1994).

In fact this approach to bias reduction was firstly proposed by Abramson (1982) and is also described by Silverman (1986).

Advantages

- These estimators can achieve the same rate of convergence as high-order kernels but with a reduced bias.
- Also non-parametric transformation kernel estimators suggested by Ruppert and Cline (1994) have a very good performance for densities with sharp peaks.

3.6 THE KERNEL METHOD FOR MULTIVARIATE DATA

Since most of the important applications in practice involve the analysis of multivariate data it is interesting to study the kernel method in the case of multivariate data.

So, the multivariate kernel density estimator with kernel K and bandwidth h is defined by (Cacoullos, 1966)

$$\tilde{q}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left\{\frac{1}{h}(x - x_i)\right\}$$

supposing that the given data set is defined in the d-dimensional space where underlying density is to be estimated.

The kernel function K defined for d-dimensional x satisfies the following condition: $\int_{R^d} K(x)dx = 1$ and will usually be a radially symmetric unimodal probability density function such as the standard multivariate normal density function.

i.e.

$$K(x) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}X^T X\right)$$

or the multivariate Epanenchnikov kernel

$$K_e(x) = \begin{cases} \frac{1}{2}cd^{-1}(d+2)\left(1-X^T X\right) & \text{if } X^T X < 1 \\ 0 & \text{otherwise} \end{cases}$$

where c_d is the volume of the unit d-dimensional sphere $c_1=2$, $c_2=\pi$, ...

For the case d=2 kernels of great importance are the following:

$$K_2(x) = \begin{cases} 3\pi^{-1}\left(1-X^T X\right)^2 & \text{if } X^T X < 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$K(x) = \begin{cases} 4\pi^{-1}\left(1-X^T X\right)^3 & \text{if } X^T X < 1 \\ 0 & \text{otherwise} \end{cases}$$

These kernels in comparison with the Epanechnikov kernel have the advantage of possessing derivatives of higher order and in addition they can be calculated more quickly then the normal kernel.

The choice of the bandwidth as well as the choice of the kernel is a 'problem' in the sense, with which we should deal with also in the case of multivariate data.

Silverman (1986) shows that the approximately optimal window width in the sense of minimizing the mean integrated squared error i.e.

$$\frac{1}{4}h^4 a^2 \int \left\{\nabla^2 f(x)\right\}^2 dx + n^{-1}h^{-d}\beta$$

where a and β are constants

and

$$\nabla^2 f(x) = \sum_{i=1}^{d} \left(\partial^2 \Big/ \partial x_i^2 \right) f(x)$$

is :

$$h_{opt} = d\beta a^{-2} \left\{ \int \left(\nabla^2 f\right)^2 \right\}^{-1} n^{-1}.$$

It must be noticed that h_{opt} converges to zero but with the slow rate of $n^{-1/(d+4)}$.

Silverman (1986) considers the Epanechnikov kernel to be optimum among non-negative kernels in the sense of minimizing the smallest mean integrated squared error achievable.

Furthermore the choice of the bandwidth is of great importance. As in the case of univariate data, the methods of cross-validation and likelihood cross-validation can be used for this purpose. As Silverman (1986) mentioned the method of least-squares cross-validation is probably preferred since the likelihood cross-validation method requires the outlying of observations, which is rather difficult in the case of multivariate data but also because there is more space in which outliers can occur.

3.7 APPLICATIONS OF KERNEL DENSITY ESTIMATORS

Kernel smoothing techniques are mostly used for density estimation and non-parametric regression. Although in these cases these methods are of great importance and provide effective solutions they can also be very useful when tackling with more complicated problems. Therefore kernel density estimators can be used in discriminant or cluster analysis. In fact they were firstly used for this reason and not for data presentation (Fix and Hodges, 1951). In addition they are broadly used for estimating other functions such as the hazard rates or spectral densities and intensities functions. Furthermore when data are of irregular type i.e. have dependencies or are observed with error, kernel density estimators and more generally non-parametric density estimators gain attention and this emphasize their applicability to a variety of topics.

3.7.1 NON-PARAMETRIC DISCRIMINANT ANALYSIS

Fix and Hodges (1951) were the first that considered the topic of non-parametric discriminant analysis. They also established the consistency of the approach consisting in non-parametrically estimating the likelihood ratio.

The discrimination problem may be defined as follows:

A random variable Z with observed value z, is distributed over some space either according to distribution F with density f or according to distribution G with density g. Then the problem is to decide which distribution does the variable Z has.

The classical approach is to assume that F and G are completely known and the solution is given explicitly by Neyman and Pearson (1936) but also by Welch (1939) Thus the allocation of Z to F or G depends only on the likelihood ratio $f(z)/g(z)$.

In fact if $f(z)/g(z) > c$ where c is an appropriate constant then Z is allocated to distribution F while if $f(z)/g(z) < c$, Z has the distribution G. Finally in the case where $f(z)/g(z) = c$ the decision may be made in an arbitrary manner. This is called the likelihood procedure and is denoted by L(c).

In most practical problems the densities f and g are not assume to be known and in order to overcome this problem it is assumed that f and g belong to a parametric family which allows us to estimate them more easily.

However Fix and Hodges (1951) considered the case where the densities f and g are not known apart from some assumptions about their existence. Then they suggest

estimating f and g using kernel estimates of the form $\hat{f}(z) = m^{-1} \sum_{i=1}^{m} K_m(X_i - z)$ and apply the procedure L(c) with these estimates in order to obtain a non-parametric discriminant rule.

Also, Silverman (1986) suggests using the kernel method for estimating the densities f and g. Non-parametric discriminant analysis was studied when using separate kernel estimates for each population in the training set and chose smoothing parameters by likelihood cross-validation. Their results enhance the use of kernel estimates in order to obtain a non-parametric discrimination rule, since in most cases the kernel method performed very well.

Furthermore it should be noticed that non-parametric discriminant analysis can be applied even in the case of discrete or mixed data (Silverman, 1986).

3.7.2 ESTIMATION OF HAZARD FUNCTIONS USING KERNEL DENSITY ESTIMATES

Kernel density estimators can be used in the case that the density itself is not of interest, but some functionals of the density are the quantities for which there is need to use an appropriate method of estimating them. Such quantities may be the hazard rate as well as the intensity function.

The hazard rate is defined as follows:

$$\lambda(x) = \frac{f(x)}{\int_x^{\infty} f(u)du} = \frac{f(x)}{1 - F(x)}, \qquad F(x) < 1$$

where F(x) is the cumulative distribution function of a random variable X, which usually represents the lifetime of a subject.

The hazard function is of great importance in the context of reliability theory and survival analysis. It can be interpreted as the approximate probability of failure in the time interval [x, x+dx] given that the subject has survived to time x (McCune and McCune, 1987).

A non-parametric estimator of the hazard rate, obtained by the kernel method is the following:

$$\lambda(x;h) = \hat{f}(x)/1 - \hat{F}(x)$$

$$= n^{-1}\sum_{i=1}^{n} K_h(x - X_i) \Big/ n^{-1}\sum_{i=1}^{n} \tilde{K}\left\{(x - X_i)/h\right\}$$

where $\tilde{K}(u)du$ is the cumulative distribution function of the kernel

$$\tilde{K}(u) = \int_{-\infty}^{u} K(v)dv\,.$$

From the definition of the estimate of the hazard rate follows that in order to get the best possible estimate of the hazard, errors that appear usually in the tails should be minimized. Therefore in this case it is preferred to use the adaptive kernel estimator rather than the fixed one (Silverman, 1986). In addition an alternative form of the hazard estimator is this that uses the empirical distribution function in the denominator.

3.7.3 KERNEL SPECTRAL DENSITY ESTIMATION

The problem of estimating spectral density functions is closely related to that of estimating probability density functions. So, the kernel method can be applied in order to get an estimate of the spectral density function.

Suppose that X(t), t=0, ±1, ... be a stationary time series process and we are interested in estimating the spectral density function:

$$f(\alpha) = (2\pi)^{-1}\int_{-\pi}^{\pi} e^{-i\alpha t} r(t)dt$$

where r(t) = Cov(X(0), X(t)) is the covariance function of the process and α represents frequency.

The basic estimate of the spectral density function is the periodogram of X(0).....X(n-1) defined as:

$$I_n(\alpha) = (2\pi n)^{-1} \left| \sum_{t=0}^{n-1} \exp\{-iat\} X(t) \right|^2$$

However, the periodogram is a very wiggly estimator. Therefore the kernel spectral estimator can be used since it overcomes some of the deficiencies of the periodogram. The kernel spectral density estimate is given by:

$$\hat{f}(a;h) = \int_{-\pi}^{\pi} K_h(a-u) I_n(u) du$$

The choice of the bandwidth parameter still is of great importance. Beltrao and Bloomfield (1987) suggest using a cross-validated likelihood function in kernel spectrum estimation for the choice of the bandwidth. While the method discussed in Park and Marron (1990) as well as Hall et al. (1991), which suggest plugging estimates of integrated squared derivatives into asymptotic representation for the optimal bandwidth may also be used for selecting the bandwidth in kernel spectrum estimation.

Furthermore kernel spectral density estimators can be used for the estimation of integrals of squared derivatives of a spectral density (Park and Cho, 1991).

3.7.4 KERNEL DENSITY ESTIMATION FOR INVESTIGATING MULTIMODALITY

Several authors have studied methods to investigate the number of modes in a density or its derivatives. The existence of modes is of great importance not only from statistical point of view but also because they may indicate other phenomenon such as clusters in the data from which the density estimate was constructed. Furthermore the existence of bumps in a density is of interest since it indicative of a mixture (Cox, 1966). The definition of a mode in a density f will be a local maximum. While a bump is an interval [a, b] over which the density is concave but this does not hold for any larger interval.

Most of the methods which investigate the number of modes depend on some arbitrary implicit or explicit choice of the scale of the effects being studied (Silverman, 1980).

However Silverman 's (1981 b) approach is based on a technique that makes use of kernel density estimates in order to investigate the number of modes in a population. In fact he considers a test statistic for examining the number of modes in the density by constructing kernel density estimates of the data.

So, in order to test the null hypothesis that the density f has k modes against the alternative that f has more than k modes he defines the k-critical window width h_{crit} by

$$h_{crit} = \inf\left\{h; \hat{f}(.,h) \text{ has at most } k \text{ modes}\right\}$$

and rejects the null hypothesis for large values of h_{crit}.

From the definition of the critical value h_{crit} it follows that the $\hat{f}(.,h)$ has more than k modes if and only if $h < h_{crit}$ (Silverman, 1981 b). Furthermore it follows that a simple binary search procedure can be used to find h_{crit} in practice.

3.7.5 KERNEL DENSITY ESTIMATION FOR IRREGULAR TYPE OF DATA

The type of the data influences the performance of kernel density estimators. Apart from the usual case where the data are assumed to be independent kernel density estimates can be used when data are not independent, they are length biased, censored or even data measured with some error.

i) Length-biased data

Such data arise when the sampling mechanism includes observations to the final sample according to the following rule: the probability of including an observation X from a density f to the sample is proportional to its value.

The problem of density estimation in this case can be defined as follows: given a sample $X_1, \ldots, X_n$ of positive -valued random variables with density

$$g(x) = xf(x)/\mu, \qquad x > 0$$

where $\mu = \int zf(z)dz < \infty$ and the kernel method is used for the estimation of f.

The simplest kernel estimate that could be obtained is the following:

$$\tilde{f}_L(x;h) = \hat{\mu}\,\hat{g}(x;h)/x$$

where $\hat{\mu}$ is an appropriate estimate of μ and $\hat{g}(x;h)$ a kernel estimate of g(x). However an improved estimator can be used. This is based on the idea of convolving a kernel weight with the estimate of the distribution function and is given by:

$$\hat{f}_L(x;h) = \int K_h(x-y)d\hat{F}_L(y)$$

$$= n^{-1}\hat{\mu}\sum_{i=1}^{n} X_i^{-1}K_h(x-X_i)$$

where $\hat{F}_L(y)$ is a distribution function that takes into account the sampling mechanism for length biased data and is defined as :

$$\hat{F}_L(y) = n^{-1}\hat{\mu}\sum_{i=1}^{n} X_i^{-1}1_{\{X_i \le x\}}$$

and $$\hat{\mu} = \left(n^{-1}\sum_{i=1}^{n} X_i^{-1} \right)^{-1}$$

This estimator has the same bias properties as the fixed kernel density estimate but there is some loss of information due to the length biased mechansim (Wand and Jones, 1995).

ii) censored data

Right censoring arise frequently in practice for life data. In such studies the variable of interest is lifetime, which is subject to right censoring during the follow-up period after the start of the study. Thus, right-censoring expresses the possibility that the observed variable is removed before the end of the study.

The problem of density estimation can then be expressed as follows: let $X_1, \ldots, X_n$ be the uncensored lifetimes with distribution function F_X and $Z_1, \ldots, Z_n$ the censoring variables with distribution function F_Z. The observed variable is

$$Y_i = \min(X_i, Z_i)$$

while

$$I_i = 1_{\{x_i \le z_i\}}$$

is an indicator function which allows us to know whether the observed lifetime is a censored one or not.

The kernel method is used for the estimation of the density F_X of the X_i's and the kernel density estimator is given by

$$\hat{f}_X(x;h) = \int K_h(x-y)\,d\hat{F}_X^{KM}(y)$$

$$= \sum_{i=1}^{n} s_i K_h\left(x - Y_{(i)}\right)$$

where $\hat{F}_X^{KM}(x)$ is the Kaplan-Meier estimate of F_X (Kaplan and Meier, 1958) i.e. a generalization of the empirical distribution function for right censored data and is given by:

$$\hat{F}_X^{KM}(x) = \begin{cases} 0, & 0 \le x \le Y_{(1)} \\ 1 - \prod_{i=1}^{j-1}\left(\dfrac{n-i}{n-i+1}\right)^{\delta_{(i)}}, & Y_{(j-1)} < x < Y_{(j)}, \quad j = 2,\ldots,n \\ 1, & x > Y_{(n)} \end{cases}$$

where $\left(Y_{(i)}, I_{[i]}\right)$ is the (Y_i, I_i) ordered with respect to the Y_i's and s_i is the size of the jump of $\hat{F}_X^{KM}(x)$ at $Y_{(i)}$.

3.7.6 CLUSTER ANALYSIS

Another possible field where density estimation methods can be used is cluster analysis. Cluster analysis is used to divide a population into a number of clusters or classes. The relationship between clusters and the density estimator can be interpreted

as follows: the points $X_1,\ldots, X_n$ in the set to be clustered are represented by modes or peaks in the density estimate constructed from these points. There are several methods of cluster analysis but here we consider only hierarchical clustering and the way kernel estimates can be used to define such a hierarchical structure for a set of points. More precisely hierarchical clustering of the data is obtained by the family trees that consist of "parent-child" relationships between the data.

The following algorithm for hierarchical clustering:

Let f be the density estimate of $X_1,\ldots, X_n$ and d_{ij} the Euclidean distance between X_i and X_j. Also consider a threshold t_i for each object X_i. Then, for objects within distance d_{ij}, X_j will be a parent of X_i if j is chosen so that to maximize

$$\frac{\hat{f}(X_j)-\hat{f}(X_i)}{d_{il}}$$

over objects X_l for which

$$d_{il} \le t_i \quad and \quad \hat{f}(X_l) > \hat{f}(X_i)$$

If there are no points X_l satisfying the above conditions, then X_i will not have a parent and therefore will be the root of the family tree.

They also suggest using the kernel method to estimate $\hat{f}$ and choose all the thresholds to be equal to the bandwidth h.

3.7.7 A KERNEL APPROACH TO A SCREENING PROCEDURE

The kernel method can be applied to a screening problem (Boys, 1992).

Screening methods can be applied to several fields of science such as medicine, quality control and education (Eddy, 1980; Madsen, 1982).

Screening procedures retain individuals, so that of those retained, the probability of possessing an attribute has some prespecified value δ. A continuous variable is used to screen future individuals for the presence (or absence) of an attribute.

So, the screening procedure may be described by a specification region C_τ for the feature variable X where an individual passes the screen only if $X \in C_\tau$.

Thus, the screening problem may be described as finding a region C_x that retains individuals possessing the attribute while also satisfying an appropriate probability statement. Usually two types of screening procedure may be used in such situations. These are:

- A 'local' procedure which considers a probability at the level of individual. This means that the region C_x is chosen so that $\Pr(T = 1/X = x)$ is greater or equal to δ if $x \in C_x$ or otherwise less than δ i.e.

$$\Pr(T = 1/X = x) = \begin{cases} \geq \delta & if\ x \in C_x \\ \prec \delta & if\ x \notin C_x \end{cases}.$$

- A 'global' procedure that uses the conditional probability over all individuals passing the procedure, i.e. C_x satisfies

$$\Pr(T = 1/X \in C_x) = \delta$$

where T is a binary variable denoting the presence (T=1) or the absence(T=0) of the attribute.

The use of kernel density estimators in screening procedures

Usually screening procedures are developed using parametric models that in many cases are invalid. Therefore there is interest in developing screening procedures that retain individuals possessing an attribute (T=1) within a classical framework but without assuming parametric models.

In fact a specification region for the screening variable can be derived using a kernel smoothing approach (Boys, 1992).

In the case of local screening procedure it is easy to obtain a non-parametric solution using Copas's (1983a) kernel estimate for the binary function Pr(T=1/X=x).

In the case of global screening procedures a non-parametric solution can be obtained by using kernel smoothing techniques to estimate $\Pr(T=1, X \in C_x)$ and $\Pr(X \in C_x)$.

Boys (1992) suggests that a simple solution to the global screening problem is found by obtaining the value w satisfying

$$\hat{\Delta}(w;h)=\delta,$$

where $$\hat{\Delta}(w;h)=\frac{\sum_{i=1}^{n} t_i K\left\{\frac{(x_i-w)}{h}\right\}}{\sum_{i=1}^{n} K\left\{\frac{(x_i-w)}{h}\right\}}$$

is a kernel estimate of the true success probability in the screened population i.e. of

$$\Delta(w)=\Pr(T=1/X\geq w).$$

Before a screening solution is found it is necessary to obtain a value for the smoothing parameter that does not depend on the unknown model.

Estimates of the smoothing parameter h are given by Copas (1983a) and Silverman (1986). Finally the value of w satisfying the condition is given by

$$\hat{w}=\min\left[x:\hat{\Delta}\left\{x;\hat{h}(x)\right\}=\delta\right]$$

Furthermore the proportion of individuals possessing the attribute may be estimated by $\hat{\gamma}=\#(t_i=1)/n$, while the proportion retained by the procedure by

$$\hat{\beta}(\hat{w})=\frac{1}{n}\sum_{i=1}^{n}\Phi\left\{(x_i-\hat{w})/h_f\right\}$$

where Φ is the standard normal kernel density.

In addition, the proportion of individuals possessing the attribute in the screened-out population by

$$\hat{\varepsilon}(\hat{w}) = \frac{\{\hat{\gamma} - \delta\hat{\beta}(\hat{w})\}}{\{1 - \hat{\beta}(\hat{w})\}}.$$

Also it must be noticed that when using the kernel methods the screening procedure will on average attain the correct success rate in the retained population δ=0.95.
Kernel methods produce smaller absolute mean bias and mean deviation than other methods such as empirical and estimative methods, while these are more sensitive with respect to skewness in comparison with the kernel method (Boys 1990).

Chapter 4

GRADUATION

4.1 INTRODUCTION

Smoothing techniques are of great importance since they have applications in many fields of science such as biometry, econometrics, engineering, mathematics and economics. The idea of smoothing is widely used when analyzing demographic and economic data. Therefore graduation techniques are applied for smoothing data sets arising in different fields. Especially graduation by the kernel method is used for data arising in criminology and medicine (Copas, 1982) as well as in actuarial field (Copas and Haberman, 1983; Bloomfield and Haberman, 1987) but also for analysing income distributions (Cowell, Jenkins and Litchfield; Schluter, 1996).

4.2 DESCRIPTION AND USES OF GRADUATION

More formally, as described by Bloomfield and Haberman (1987) graduation may be regarded as the principles and methods by which a set of observed (or crude) probabilities are adjusted in order to provide a suitable basis for inferences to be drawn and further practical computation to be made.

One of the principal applications of graduation in actuarial field is building a survival model, which is usually presented under the form of a mortality table. Then graduation may be regarded as the process of smoothing the separate empirical mortality rates to obtain the best possible estimates of the underlying unknown mortality pattern of the population.

So, let E be an event whose probability of occurrence depends on some continuous variable x : $P(E/X) = q_x$.

In the case of mortality data, E may be death and x age. Then given observations on n individuals with characteristic x and incidence of E it is required to estimate q_x.

The fact that the observed data may be regarded as a sample from a large population implies that the observed probabilities (or crude rates), derived therefore are subject to sampling errors. Providing these errors are random in nature and taking into account the fact that set of probabilities progress smoothly with x we can use graduation to remove the random errors and produce smooth estimates of the true rates.

So, the most important role of graduation is the smoothing of the data since it allows handling them in a more efficient way. However Bloomfield and Haberman (1987) refer to other situations where graduation techniques are important such as the case of incomplete data and the use of estimation methods based on graduation are necessary or even in the case of forecasting and projection.

4.3 METHODS OF GRADUATION

The methods of graduation can be described broadly as (a) graphical methods, (b) parametric methods and (c) non-parametric methods. Graduation by reference to a standard table and spline graduation as well as their advantages and disadvantages are described in the context of parametric graduation. Also "laws of mortality" that describe the mortality pattern are presented. Furthermore, non-parametric methods of graduation are of special interest since the kernel method is used for the purpose of non-parametric graduation.

4.3.1 THE GRAPHIC METHOD

The graphic method (Benjamin and Pollard, 1980) is one of the most widely used graduation techniques.

In the case of mortality data this method can be used to derive the graduated mortality rates as follows:

Firstly the crude mortality rates $\overset{\circ}{q}_x$ are calculated from the available data and are represented graphically by the data.

Then a smooth curve is drawn as close as possible to these points providing a reasonable progression of the mortality rates so that the graduated mortality rates can be read from this curve.

Finally the first few orders of differences $\left(\Delta, \Delta^2, \Delta^3\right)$ for the graduated rates are calculated since they reveal if and where the smoothness is unsatisfactory but also where points of inflexion occur. These differences are adjusted in order to improve the progression of the rates and obtain graduated rates, which satisfy the prescribed requirements of smoothness and adherence to data. This process is referred to as 'hand-polishing'.

In addition it is helpful to record on the graph the approximate 95 per cent confidence limits for the observed mortality rates at each age and then draw one line connecting consecutive pairs of the upper limit points and another such line connecting the lower limit points. These two lines could be used as a guide when drawing the smooth curve since this curve should not pass outside the limit lines more often than about once for every twenty observations.

It is noticed that in the case where the actual deaths at any given age are not less than about ten, the 95 per cent confidence limits may be taken as:

$$\overset{\circ}{q}_x \pm \left(2\sqrt{\theta_x}\right) / E_x$$

Advantages

- The graphic method gives good results even in cases where the data are very scanty.
- Furthermore it allows to draw conclusions based on previous experience and knowledge obtained from other tables of a similar type especially for data at the ends of the table.

Disadvantages

- Different results may be obtained from the same data due to the fact that individual judgement is allowed. Also this may lead to results that are prone to individual bias and prejudice.
- The graphic method turns out to be unsuitable for large data sets since it is practically non-efficient to achieve a very high degree of smoothness for such data. This happens because it is impossible to obtain sufficient places of decimals in the graduated rates, which is caused of the difficulty of reading more than three figures from a graph.

4.3.2 GRADUATION BY REFERENCE TO A STANDARD LIFE TABLE

Graduation by reference to a standard life table (Benjamin and Pollard, 1980) is a graduation technique that can be used for scanty data for which there is already

experience related to a known standard graduated table, which can be considered as a 'base curve' for graduating the new data.

Several ways of applying this graduation technique are suggested in the literature.

The simplest method is to calculate the ratios of the crude mortality rates q_x to the corresponding q_x^s of the standard life table and graduate these ratios $\overset{\circ}{q}_x / q_x^S$.

Lidstone improved on this method by graduating graphically the quantity

$$\log(p_x^S / \overset{\circ}{p}_x).$$

Apart from the graphic method of graduation the Lidstone's transformation can be used in conjunction with some mathematical formulae of graduation with reference to a standard life table. Some of them are given below:

$$q_x = aq_x^s + b \quad (1)$$

$$\mu_x = a\mu_x^s + b \quad (2)$$

$$q_x = q_x^s(ax + b) \quad (3)$$

$$\mu_x = \mu_{x+n}^s + K \quad (4)$$

$$q_x = aq_x^{(1)} + bq_x^{(2)} \quad (5)$$

where a, b, K and n are constants, while $q_x^{(1)}$ refers to one standard life table and $q_x^{(2)}$ to a second.

Each of the above formulae is likely to produce a satisfactory graduation depending on the situation to which will be applied.

Thus, in the case of linear relationship between the observed mortality rates and the standard rates, formula (1) may be appropriate. While when the ratio of the observed rate to the rate from the standard table, follow a linear trend with age then formula (4) is probably the correct one.

Advantages

- The method can be used in the case of extremely scanty data, when all other methods are out of question even the graphic method of graduation.

Disadvantages

- The choice of an appropriate standard table is not always possible. Thus the adherence of the graduated rates to the data can be unsatisfactory even if the parameters are correctly estimated, since any special feature in the graduation of the standard table will be reproduced even exaggerated in the graduation of the new data.

4.3.3 SPLINE GRADUATION

Another popular method for graduation incorporates the use of spline functions. A spline function is called a polynomial for which the maximum possible number of derivatives exist. In fact a spline is a polynomial chosen so that derivatives up to and including the order one less than the degree of the polynomial used, are continuous everywhere. Thus, a spline s of degree k, defined on the interval [a, b] with interval knots $x_1,\ldots,x_n$ $\{a = x_0 < x_1 \ldots < x_n < x_{n+1} = b\}$, is a function such that if $0 \le i \le n$ and $x_i \le x \le x_{i+1}$ then $s(x)=p_i(x)$, where $p_i(x)$ is a polynomial in x of degree not greater than k. The polynomials $p_0(x),\ldots, p_x(x)$ fit together such that s is differential (k-1) times in the interval (a, b).

Furthermore the spline function offers continuity of the greater possible number of derivatives consistent with the use of polynomials of lower degree that would be needed to fit all data by a single polynomial.

According to Benjamin and Pollard (1980) the natural cubic spline is very useful for graduation purposes, especially in the case of mortality data.

Given a small number of knots $x_1,\ldots,x_n$, the natural cubic spline passing through the n data points is defined as:

$$s(x) = \alpha_0 + \alpha_1 x + \sum_{j=1}^{n-2} b_j \Phi_j(x)$$

where

$$\Phi_j(x) = \phi_j(x) - \left\{\frac{x_n - x_j}{x_n - x_{n-1}}\right\}\phi_{n-1}(x) + \left\{\frac{x_{n-1} - x_j}{x_n - x_{n-1}}\right\}\phi_n(x)$$

and

$$\phi_j(x) = \begin{cases} 0, & x \prec x_j \\ (x - x_j)^3, & x \ge x_j \end{cases}$$

The best n-knot spline $s(x)=q_x$, given a set of n knots is one which minimizes

$$X^2 = \sum_x \frac{(\theta_x - E_x q_x)^2}{E_x q_x (1-q_x)}$$

$$= \sum_x \frac{E_x}{q_x(1-q_x)} \left[\frac{\theta_x}{E_x} - q_x \right]^2$$

$$= \sum_x w_x \left(\overset{\diamond}{q_x} - q_x \right)^2$$

where $$w_x = \frac{E_x}{q_x(1-q_x)}$$

The choice of the number as well as the position of knots is of great importance. It is not possible to obtain an acceptable graduation if there are few knots, on the contrary if the number of knots is excessive there would be little graduation as the spline will adhere too closely the crude rates. McCutcheon (1981) in order to determine the positions of a fixed number of knots, suggests minimizing X^2 subject to some constraints that arise from the fact that the knots must lie in the interval [a,b].

Advantages

- The method provides a smooth graduation for small munber of ages when these are well chosen.

Disadvantages

- A disadvantage of the spline graduation method is that for the choice of knot points is required considerable skill.
- In addition this method is not readily applicable to very sparse experiences.

4.3.4 LAWS OF MORTALITY

In order to graduate the mortality rates several "laws of mortality" i.e. of mathematical expressions have been developed.

The first model that gave very close fits to empirical mortality at all ages was that suggested by Heligman and Pollard (1980):

$$\frac{q_x}{p_x} = A^{(x+B)^C} + De^{-(E(\ln x - \ln F)^2} + GH^x$$

where q_x is the probability of dying within a year, $p_x=(1-q_x)$ and A to H are parameters to be estimated.

All the parameters in the model have demographic interpretation. The mathematical expression given above contains three terms, the first of them reflects the early childhood years, the second refers to the middle age mortality (accident hump) while the third, known as Compertz exponential represents senescent mortality. Thus, A measures the level of mortality, which is nearly equal to q_1, while B is an age displacement to account for infant mortality. C measures the rate of mortality decline in childhood. The parameters D and E represent severity and spread of the accident term while F indicates the location of the accident term. Finally the parameter G represents the base level of senescent mortality while H reflects the rate of increase of that mortality. Heligman and Pollard (1980) estimated the parameters of the model using a least-squares approach in order to minimise the function:

$$S^2 = \sum_x \left[\frac{\hat{q}_x}{q_x} - 1\right]^2$$

where $\hat{q}_x$ is the fitted value at age x and q_x is the observed mortality rate.

Advantages

- This "law of mortality" has the advantage of being continuous and applicable over the entire age range.
- Another advantage is the relatively few parameters that have to be estimated as well as the fact that all these parameters have demographic interpretation and they fully describe the age pattern of mortality.

Disadvantages

- The only disadvantage is that in cases where the accident hump is too intense it provides systematic deviations from the adult ages since the accident hump of the estimated set of q_x values is located at a higher age than the empirical accident hump.

4.4 NON-PARAMETRIC METHODS OF GRADUATION

The non-parametric (or distribution free) approach to graduation does not involve functional forms or parameters of such forms. In general non-parametric methods apply to very wide families of distribution rather than only to families specified by a particular functional form. Two approaches to graduation by non-parametric methods are described: the summation and adjusted average graduation formulae as well as the kernel graduation, firstly used by Copas and Haberman (1983) and by Ramlau-Hansen (1983).

4.4.1 SUMMATION AND ADJUSTED-AVERAGE GRADUATION FORMULAE

Benjamin and Pollard (1980) consider the graduation method of summation and adjusted -average graduation formulae.

Let represent any ungraduated value by v_x, which consists of two parts the true value u_x and a superimposed error e_x, so that $v_x = u_x + e_x$. The $\{v_x\}$ are independent unbiased estimators of $\{u_x\}$, while the $\{e_x\}$ are independent random variables with zero expectations since only sampling errors are taken into account. The $\{e_x\}$ will contain in practice apart from the sampling errors inaccuracies, which may be either random or systematic. In the case of random inaccuracies both sources of contribution to $\{e_x\}$ (i.e. inaccuracies and sampling errors) will be both redistributed by the graduation formulae. If these inaccuracies are systematic then this method of graduation is not anymore applicable since the $\{e_x\}$ cease to be independent random variables with zero expectations. An ideal graduation would eliminate all the $\{e_x\}$ but in practice what can only be attained is a reduction of the error as well as a smooth progression of the graduated rates.

Moving or running averages can be used for smoothing observations with irregularities of the form of riples or undulations. However they distort values that

already follow a smooth distortion and do not require adjustment. This distortion will not affect the smoothness of the values but may introduce a downward or upward bias to them and may be corrected by adjustment of the moving average operation. In addition moving averages do not produce values at the beginning and end of the table.

Summation formulae

This graduation method is based on the smoothing properties of moving averages. These formulae involve three summation operators [l], [m], and [n] as well as a fourth one which is a linear combination of two or more summation operators. The operator [n] or 'summation n' is defined to be:

$$[n]v_0 = v_{\frac{n-1}{2}} + v_{\frac{n-3}{2}} + \ldots + v_{\frac{n-3}{2}} + v_{\frac{n-1}{2}}$$

The moving average of n terms is proved to be (Kendall and Stuart, 1968):

$$\frac{[n]}{n}v_x = \left\{1 + \frac{(n^2-1^2)}{2^2 3!}\delta^2 + \frac{(n^2-1^2)(n^2-3^2)}{2^4 5!}\delta^4 + \ldots\right\}v_x$$

where the second and fourth differences of v_x are defined as follows:

$$\delta^2 v_x = v_{x+1} - 2v_x + v_{x-1}$$

$$\delta^4 v_x = v_{x+2} - 4v_{x+1} + 6v_x - 4v_{x-1} + v_{x-2}$$

This formula shows the distortion inherent in the use of a simple moving average while the summation formulae are designed to be free of second-difference distortion. The most widely used by actuaries summation formulae is the Spencer's 21-term formula given by:

$$V_x = \frac{[5][5][7]}{350}\{[1]+[3]+[5]-[7]\}v_x$$

However a complete analysis of graduation formulae includes an investigation of the following features: the range, the error-reducing power of the formula, its smoothing power as well as its 'wave-cutting' properties.

Range

It is useful to calculate the range of a formula, which is defined as the span of the number of ungraduated v's involved in the calculation of a single graduated value. This may be greater than the number of v's if some of the coefficients are zero. The rule for determining the range of a summation formula is the following: firstly take the number of terms of the widest summation operator in the linear compound operator and then for each individual summation operator [n] outside the linear compound operator add n-1 to the range. Generally it is preferable to use a formula with the shorter range since then it is easier to apply, as well as the assumption that fourth and higher differences are negligible over the range is more likely to be accurate. Furthermore a smaller number of terms at the ends of the series of graduated values remain to be filled by other methods.

Most summation graduation formulae do not produce first, second or third difference distortion.

Adjusted-average graduation formulae

Any summation formulae of range 2r+1 can be written in the explicit form

$$K_0 v_x + K_1(v_{x+1} + v_{x-1}) + K_2(v_{x+2} + v_{x-2}) + \ldots + K_r(v_{x+r} + v_{x-r})$$

Although every summation formulae can be expressed in this way, there are an infinite number of formulae of this expanded type, which cannot be derived from summation formulae. Those are referred in the literature as 'adjusted-average' graduation formulae and can be used for graduation purposes.

Error-reducing power

When a summation graduation formula with the expanded form is applied to the observed data, the ungraduated errors $\{e_x\}$ are smoothed to yield graduated errors $\{$

where $e'_x = K_0 e_x + K_1(e_{x+1} + e_{x-1}) + \ldots + K_r(e_{x+r} + e_{x-r})$.

In the case where the $\{v_x\}$ are unbiased observations then the expectations of $\{e_x\}$ as well as of e'_x will be zero. So, the summation graduation formulae will achieve its goal as an error reducer more successfully, if the e'_x are closer to zero than the $\{e_x\}$. A necessary condition in order to be this true is the e'_x to have a considerably smaller variance than e_x. Finally a measure of an error-reducing power is given by the ratio of the standard deviation of e'_x to the standard deviation of e_x:

$$\phi_E = \sqrt{\left(K_0^{\,2} 2K_1^{\,2} + 2K_2^{\,2} + \ldots + 2K_r^{\,2}\right)}$$

In fact the error-reducing power of the formula will be better if this error-reducing index is as small as possible.

Smoothing power

The smoothing power of a summation graduation formula depends on the size of the coefficients as well as on their order.

A summation graduation formula has good smoothing properties if the third differences of the graduated errors $\{e'_x\}$ tend to concentrate closer to zero in comparison with the $\{e_x\}$. This happens if the variance of $\Delta^3 e'_x$ is considerably smaller than the variance of $\Delta^3 e_x$. A measure of the smoothing power of a formula is given by the smoothing index ϕ_s i.e. the ratio of standard deviation of $\Delta^3 e'_x$ to the standard deviation of $\Delta^3 e_x$ under the assumption that the $\{e_x\}$ are independent with common variance σ^2.

Wave cutting

A formula typified by a curve that spreads the effect of an ungraduated error over a wider field is said to have good wave-cutting properties. The wave-cutting index ϕ_w of a graduation formula is defined as the sum of the five central coefficients, while for even number of terms it is defined as the sum of the four middle coefficients and the next one at either end. The smaller its values the better the wave cutting of the

formula. The wave-cutting properties of a summation formula are of little importance in the case of demographic and actuarial statistics since systematic inaccuracies usually do not produce distortions in the form of waves.

Adjusted-average graduation formulae with optimal error-reducing and smoothing power

The error-reducing, smoothing and wave-cutting indices allow to compare alternative graduation formulae. However, without defining absolute scales for these indices we cannot decide which of the formulae are absolutely satisfactory. Thus, the error - reducing efficiency of a summation or adjusted-average formula can be calculated by dividing the optimal error-reducing index for a formula by the error-reducing index of the formula. While the smoothing efficiency of a summation or adjusted-average formula can be calculated by dividing the optimal smoothing index for a formula of that range by the smoothing index of the formula.

Advantages

- The calculations required are very simple since British actuaries developed these graduation methods at a time when available calculating equipment was very limited. Although nowadays, electronic calculators have eliminated computational difficulties this method is convenient for hand computation.

Disadvantages

- The most important disadvantage is that the ends of the table need to be completed by other methods. In the case of a graduation formula of range 2r+1, r graduated values are calculated at either end of the table. These values can be obtained for the first r ages by fitting an unweighted lest-squares cubic near the first 2r+1 unadjusted values. While the graduated values for the last r ages can be obtained in a similar way using the final 2r+1 unadjusted values. An alternative approach may be the use of extrapolation methods. Furthermore Compertz and Makeham - type curves are used to graduate mortality rates at high ages.
- It is impossible to take into account the weight of the exposed to risk at each age since it is assumed that the variances of the random errors $\{e_x\}$ are constant.
- It is required a lot of experience in order to have crude rates that progress fairly smoothly which gives satisfactory results.

4.4.2 KERNEL GRADUATION

An alternative non-parametric *approach* for the graduation of the mortality probabilities q_x is the use of kernel methods (Copas and Haberman, 1983). Kernel methods were firstly developed for estimating density functions.

Since graduation requires the estimation of two such density functions kernel methods can be used for the purpose of graduation.

Consider E the event of death, denoted by E=d, whose probability of occurrence depends on the continuous variable x (age)

$$P(E=d/X=x) = q_x$$

The crude estimate of the mortality rate q_x is given by

$$\overset{\circ}{q}_i = \frac{d_i}{e_i}, i = 1,\ldots,n$$

where d_i denotes the number of deaths, while e_i the exposed to risk for age x_i.

A kernel estimator of q_x (Copas and Haberman, 1983) is given by

$$\hat{q}_x = \frac{\sum_{i,E\,only} \psi\left(\frac{x - x_i}{h}\right)}{\sum_{i,all\,cases} \psi\left(\frac{x - x_i}{h}\right)}$$

where the sum is taken over all cases in the denominator but in the nominator over only these cases where E has occurred.

In this expression all cases are counted separately even if there are ties among x's.

In this estimate h is a constant and the kernel function is assumed to be positive, symmetric about zero and with a single at the origin.

Some properties of $\hat{q}_x$ are the following:

a) it always lies in [0,1]

b) h controls the degree of smoothing – if h is very small, $\hat{q}_x$ is essentially the proportion of E cases at x i.e. the crude rate at age x - if h is large observations at

other ages have greater influence on $\hat{q}_x$ and more smoothing occurs at the expense of fit between graduated rates ant the actual data. As h→∞, $\hat{q}_x$ tends to the overall probability of death. Thus h also measures the information contributing to the estimate $\hat{q}_x$.

Advantages

- The estimate of the crude mortality rate is given by a single formula and it is not required any statistical fitting of the parameters.
- Furthermore in contrast with parametric methods, in this case the degree to which the estimated curve responds to features of the data can be controlled in a continuous way.

Disadvantages

- It is required the choice of the smoothing parameter, which governs the degree of smoothness in the method. In practice the choice of the smoothing parameter is a subjective one, although several methods have been proposed in the literature, with the aim of reflecting important features of the data but without over reacting to spurious chance fluctuations.

4.5 KERNEL ESTIMATORS FOR GRADUATION

In the literature two estimators are used for estimating q_x by kernel methods, the Nadaraya-Watson and Copas-Haberman estimators (Gavin, Haberman and Verall, 1994). These are defined as follows:

Copas-Haberman estimator

$$\hat{q}_x^{CH} = \frac{\left[\sum_{i=1}^{n} d_i K\left(\frac{x - x_i}{h}\right)\right]}{\left[\sum_{i=1}^{n} e_i K\left(\frac{x - x_i}{h}\right)\right]}$$

Nadaraya-Watson estimator

$$\hat{q}_x^{NW} = \frac{\left[\sum_{i=1}^{n} q_i K\left(\frac{x - x_i}{h}\right)\right]}{\left[\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)\right]}$$

The main difference between the two estimators is that the $\hat{q}_x^{CH}$ estimator makes explicit use of the number of deaths and the amount of exposure at each age while the $\hat{q}_x^{NW}$ is based on evenly spaced data since all the data at each age are represented by a single observation d_i / e_i.

For mortality data the $\hat{q}_x^{NW}$ estimator is considered more succesful than the $\hat{q}_x^{CH}$, which usually causes a bias due to the fact that the interval (x-h, x) contains more data than the interval (x, x+h).

DERIVATION OF $\hat{q}_x^{CH}$ AND $\hat{q}_x^{NW}$:

It has already been mentioned that the mortality rate may be expressed as

$$q_x = P(E=d/X=x)$$

The application of Bayes theorem results in three probability functions to be estimated namely:

$$P(E = d / X = x) = \frac{P(X = x / E = d)}{P(X = x)} P(E = d) \qquad (4.5.1)$$

The Copas-Haberman estimator is obtained if P(X=x/E=d) and P(X=x) are estimated by kernel functions and a simple estimate of P(E=d) is taken (Copas and Haberman, 1983).

Let the total exposure be denoted by $t = \sum_{i=1}^{n} e_i$ and the total number of deaths by u,

where $u = \sum_{i=1}^{n} d_i$.

A kernel estimator of P(X=x) is:

$$P(X = x) = \frac{1}{th} \sum_{j=1}^{t} K\left(\frac{x - x_j}{h}\right)$$

$$= \frac{1}{th} \sum_{i=1}^{n} \sum_{j=1}^{e_i} K\left(\frac{x - x_{i,j}}{h}\right) \qquad (4.5.2)$$

where $x_{i,j}$ denotes the age of the jth life out of the group of lives aged x_i .

However, $K((1/h)(x-x_{i,j}))$ is constant for $j=1,2,\ldots e_i$ and so

$$\sum_{j=1}^{e_i} K\left(\frac{x - x_{i,j}}{h}\right) = e_i K\left(\frac{x - x_i}{h}\right) \qquad (4.5.3)$$

Then (3.5.2) becomes:

$$P(X = x) = \frac{1}{th} \sum_{i=1}^{n} e_i K\left(\frac{x - x_i}{h}\right) \qquad (4.5.4)$$

Similarly, a kernel estimator for P(X=x/E=d) is:

$$P(X = x / E = d) = \frac{1}{th} \sum_{i=1}^{n} d_i K\left(\frac{x - x_i}{h}\right) \qquad (4.5.5)$$

while an estimator for P(E=d) is $\sum_{i=1}^{n} d_i \Big/ \sum_{i=1}^{n} e_i$

If we substitute the equations (4.5.4) and (4.5.5) into (4.5.1) we obtain the kernel graduation estimator $\hat{q}_x^{CH}$.

Finally, we have:

$$\hat{q}_x^{CH} = \left[\frac{1}{ih}\sum_{i=1}^{n} d_i K\left(\frac{x-x_i}{h}\right)\left(\frac{u}{t}\right)\right] \Big/ \left[\frac{1}{ih}\sum_{i=1}^{n} e_i K\left(\frac{x-x_i}{h}\right)\right] \quad (4.5.6)$$

$$= \left[\sum_{i=1}^{n} d_i K\left(\frac{x-x_i}{h}\right)\right] \Big/ \left[\sum_{i=1}^{n} e_i K\left(\frac{x-x_i}{h}\right)\right] \quad (4.5.7)$$

Based on the $\hat{q}_x^{CH}$ estimator, Gavin, Haberman and Verall (1992) derived a different kernel estimator that is closely related to MWA graduation. This estimator includes the case where the data are condensed to n equally spaced observations.

This is obtained from the $\hat{q}_x^{CH}$ estimator as follows:

$$\hat{q}_x^{CH} = \left[\sum_{i=1}^{n}\left(\frac{d_i}{e_i}\right) e_i K\left(\frac{x-x_i}{h}\right)\right] \Big/ \left[\sum_{i=1}^{n} e_i K\left(\frac{x-x_i}{h}\right)\right]$$

$$\Rightarrow \hat{q}_x^{CH} = \left[\sum_{i=1}^{n}\sum_{j=1}^{e_i}\left(\frac{d_i}{e_i}\right) K\left(\frac{x-x_{i,j}}{h}\right)\right] \Big/ \left[\sum_{i=1}^{n}\sum_{j=1}^{e_i} K\left(\frac{x-x_{i,j}}{h}\right)\right]$$

In this estimator there is a contribution in the numerator and denominator which is the same for each life aged x_i .So, if instead of contributing this for each life we count it just one then we obtain the Nadaraya-Watson estimator.

$$\hat{q}_x^{NW} = \left[\sum_{i=1}^{n}\left(\frac{d_i}{e_i}\right) K\left(\frac{x-x_i}{h}\right)\right] \Big/ \left[\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)\right]$$

$$\Rightarrow \hat{q}_x^{(1)} = \left[\sum_{i=1}^{n} \dot{q}_i K\left(\frac{x-x_i}{h}\right)\right] \Big/ \left[\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)\right]$$

4.5.1 PROPERTIES OF THE NON-PARAMETRIC ESTIMATE

Below are presented some properties of the kernel estimate of the mortality rate, studied by Copas and Haberman (1983).

i) Let Z_i be an indicator random variable which takes the value 1 if death occurs at age x_i and 0 otherwise. Then the estimate in the general case can be written as follows:

$$\hat{q}_x = \frac{\sum_i Z_i \psi_{x,i}^{(h)}}{\sum_i \psi_{x,i}^{(h)}} \qquad (4.5.8)$$

where $$\psi_{x,i}^{(h)} = \psi\left(\frac{x - x_i}{h}\right)$$

ii) If q_{xi} is the true probability for dying at age x_i then the Z_i 's are independent and have a binomial distribution with mean q_i and variance $q_i(1-q_i)$.

iii) The expected value of the estimate $\hat{q}_x$ will be :

$$E(\hat{q}_x) = q_x + \frac{\sum (q_i - q_x)\psi_{x,i}^{(h)}}{\sum \psi_{x,i}^{(h)}}$$

while the variance will be :

$$Var(\hat{q}_x) = \frac{\sum q_i(1-q_i)\psi_{x,i}^{2(h)}}{\left(\sum \psi_{x,i}^{(h)}\right)^2}$$

$$\approx q_x(1-q_x)\frac{\sum \psi_{x,i}^{2(h)}}{\left(\sum \psi_{x,i}^{(h)}\right)^2} \qquad (4.5.9)$$

iv) bias of the estimate:

Any graduation method will involve bias, since in order to make the crude mortality rates more smooth it is inevitably introduced some bias or distortion by the process of graduation. So the resulting estimates are usually biased.

From equation (4.5.9) we can obtain the bias of the estimate. In fact, using a Taylor series expansion the bias takes the form:

$$BIAS = q_x' \frac{\sum (x_i - x)\psi_{x,i}^{(h)}}{\sum \psi_{x,i}^{(h)}} + \frac{1}{2} q_x'' \frac{\sum (x_i - x)^2 \psi_{x,i}^{(h)}}{\sum \psi_{x,i}^{(h)}} + \ldots \qquad (4.5.10)$$

If the values of x_i are symmetrically located about x the coefficient of q'_x will be zero. In practice this coefficient will be small except near the ends of the age range. In fact if x is close to the lower limit then all the values $x_i > x$, produce a positive bias at a point where the curve tends to be convex and therefore $\hat{q}_x$ will tend to overestimate. While when x is close to the upper limit, the curve of q_x is concave and $\hat{q}_x$ will tend to underestimate.

The second term in (4.5.10) will be small if q_x'' is small i.e. the curve of the true population values q_x is approximately linear in the neighborhood of x, or if h is small so that the effective size of $(x_i-x)^2$ in the appropriate sum is limited.

Also it must be noticed that the coefficient of q_x'' is always positive. This means that if the curve of the true population values is convex, the second term in (4.5.10) will be positive and $\hat{q}_x$ will tend to overestimate q_x. The converse is true if the q_x curve is concave.

4.5.2 BIAS FOR THE NADARAYA-WATSON AND COPAS-HABERMAN KERNEL ESTIMATORS

Haberman , Gavin and Verall (1994) examine the bias in the cases of the Nadaraya-Watson estimator and the Copas-Haberman estimator.

The bias for $\hat{q}_x^{NW}$ has the form:

$$\frac{\sum_{i=1}^{n}(x_i - x)K(x - x_i)}{\sum_{i=1}^{n}K(x_i - x)}q'_x + \frac{1}{2}\frac{\sum_{i=1}^{n}(x_i - x)^2 K(x - x_i)}{\sum_{i=1}^{n}K(x - x_i)}q''_x + R \quad (4.5.11)$$

The same conclusions as in the general case are drawn if the data are symmetrically placed around x. Furthermore between the ages 25 and 80 the estimates are not heavily influenced by boundary effects and the mortality curve is approximately exponential in shape. So if the data were transformed by taking logs then the mortality curve would approximately be a straight line and then the Nadaraya-Watson estimator is expected to give an unbiased estimator of the true mortality rate for this region.

The bias for the $\hat{q}_x^{CH}$ has the formula:

$$\frac{\sum_{i=1}^{n}(x_i - x)e_i K_b(x - x_i)}{\sum_{i=1}^{n} e_i K_b(x_i - x)} q'_x + \frac{1}{2}\frac{\sum_{i=1}^{n}(x_i - x)^2 e_i K_b(x - x_i)}{\sum_{i=1}^{n} e_i K_b(x - x_i)} q''_x + R \quad (4.5.12)$$

Usually the data are asymmetrically placed since the number of lives exposed to risk of dying tends to decrease with increasing age. So the coefficient of the $\hat{q}'_x$ in (4.5.12) is negative which means that $\hat{q}'^{CH}_x$ gives a negative bias for most ages.

Several methods have been proposed in order to reduce bias near the ends of the table. Hall and Wehrly (JASA 86(1991): 665-72) suggest reflecting the data, which means generate pseudo-data that effectively extend the boundaries so that the original data lie in the interior of an enlarged data set. In this way the original data are less influenced by boundary effects.

Rice (Communications in Statistics. Theory and Methods 13 (1984): 893-900) uses an extrapolation method which combines two different kernels with different bandwidths to eliminate the first -order bias.

Jones (Statistics and Computing (1993): 135-46) considers a kernel function that is defined as a linear combination of K(x) and xK(x). Then we have a kernel function for the right-hand boundary

Also Bloomfield and Haberman (1987) suggest to exclude ages with few recorded deaths in order to eliminate the waves at each end of the curve.

Another approach is to group scanty data. The effect of this grouping depends on the size of h depends on the size of h in relation to the width of the group.

4.6 CHOICE OF THE BANDWIDTH PARAMETER

In non-parametric graduation methods the amount of smoothing can be varied over a continuous range, by the choice of bandwidth.

This is often cited as an advantage over parametric techniques in which the amount of smoothing can only be varied over a discrete range. This may happen for example by changing the number of parameters, by increasing the degree of a polynomial, by increasing the number of knots in a cubic spline or changing from one family of curves to another.

Both kernel estimators (i.e. $\hat{q}_x^{CH}$ and $\hat{q}_x^{NW}$) contain a bandwidth (or smoothing parameter) which governs the amount of smoothing that is applied to the graduation process.

Generally when the bandwidth h is large there is a lot of smoothing while if h is small the estimate will just have points of density at each observation.

In the case of mortality data, the relationship between the smoothing parameter and the kernel estimator $\hat{q}_x$ can be interpreted as follows: if h is very small, $\hat{q}_x$ is virtually the crude death rate at age x. Whereas when h becomes larger, observations at other ages have greater influence on $\hat{q}_x$ and more smoothing occurs at the expense of the fit between graduated rates and the actual data.

In fact as $h \rightarrow \infty$, $\hat{q}_x$ tends to the overall probability of death. Thus h measures in some way the information contributing to an estimate $\hat{q}_x$.

So, in the context of kernel graduation the choice of bandwidth is dominant and attention should focus on ways of choosing this parameter.

4.6.1 METHODS FOR CHOOSING THE BANDWIDTH PARAMETER

While it is sometimes the case that the amount of smoothing that is appropriate can be decided by studying the resulting graduations (Bloomfield and Haberman, 1987) it is desirable to have an objective, data dependent technique for choosing the bandwidth.

Gavin, Haberman and Verrall (1994) consider cross-validation as an objective and risk-based method for selecting the smoothing parameter in a non-parametric graduation, which also achieves a balance between variance and bias.

i) Use of actuarial tests of fit

So, in the first approach in order to obtain the bandwidth parameter a curve is fitted to the data and then the graduated rates are tested for smoothness using actuarial tests of fit.

The x^2 test.

For testing the fidelity of the kernel graduations to the original data the following tests are applied: firstly the standardized deviation between actual and expected deaths at each age is calculated by the formula:

$$z_x = \frac{d_x - e_x \hat{q}_x}{\sqrt{e_x \hat{q}_x (1 - \hat{q}_x)}},$$

To test where the z's are normally distributed, $x^2 = \sum_x z_x^2$ is calculated. When the number of degrees of freedom exceeds 50, the statistic $t(x^2) = \sqrt{2x^2} - \sqrt{2(n-1)}$ is approximately normally distributed with zero mean and unit variance. Furthermore a runs test as well as one for serial correlation is applied to the deviations z_x.

Run Test

The run test is a non-parametric test that checks the randomness of the deviations z_x. If the number of deaths at each age were distributed according to the normal model, the deviations at successive ages would be independent and the signs of the deviations would be randomly distributed, with neither too many nor too few runs of successive deviations with the same sign. So, if the number of positive signs is n_1 while the number of negative signs is n_2 and both of them are larger than about 20 then the number of runs formed by the signs of deviations is approximately Normally distributed with mean μ_r and variance σ_r^2, where

$$\mu_r = \frac{2n_1 n_2}{n_1 + n_2} + 1$$

$$\sigma_r^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$$

A small number of runs would indicate a graduation that is too straight compared with the observed values, cutting across waves or bends in the observed rates (Forfar, McCutchen and Wilkie, 1988). Too large a number of runs would indicate that the graduation follows the observed experience too closely and is an indication of over-fitting. However Bloomfield and Haberman (1987) notice that a large number of runs is an indication of peculiar data and not of an unsatisfactory fit.

Serial Correlation Test

In addition a serial correlation test was applied on the values of the z_x's. If the z's are randomly distributed then the correlation coefficient between successive successive values of z is approximately Normally distributed with zero mean and variance 1/n, where n is the number of ages. The correlation coefficient is defined as:

$$\rho = \frac{\sum_{y=x_0}^{x_0+n-2}(z_y - \bar{z})(z_{y+1} - \bar{z})}{\sum_{y=x_0}^{x_0+n-1}(z_y - \bar{z})^2}$$

for the age range (x_0, x_0+n-1) and $\bar{z} = \frac{\sum_{y=x_0}^{x_0+n-1} z_y}{n}$.

Too high a positive value indicates of ρ indicates an unsatisfactory fit and the graduation is not satisfactory. While, a high negative value shows that the z's were alternative positive and negative too a great extent.

It should be noted that graduations generally fail or pass the serial correlation test and the runs test together (Forfar, McCuthceon and Wilkie, 1988).

Finally it is obtained an interval with values of the bandwidth h which produce graduations satisfying the tests mentioned above. From these graduations this with the smallest sum of absolute values for the test statistics is this that gives the best fit.

Smoothness

For testing smoothness several criteria exist in the literature, Benjamin and Pollard (1980), suggest that the third differences of the graduated curve $\Delta^3 \hat{q}_x$ should be smooth and small. Barnett (1985) suggests that a series of the graduated values are smooth to the k'th order if k'th differences are insignificant, while second differences should not pass through zero no more often than 1+2n where n is the number of acceptable inherent inflections or occurences of roughness. Bloomfield and Haberman (1987) suggest using a relative measure of smoothness, defined as $D^k = \left(\hat{q}_i / \left|\Delta^k \hat{q}_i\right|\right)^{1/k}$, which expresses the k'th difference of the graduates rates relative to the graduated rates.

ii) Cross-Validation

The cross-validation method in the case of mortality data requires the minimisation of the score function CV (b) in order to choose the bandwidth, where CV (b) is defined as follows:

$$CV(b) = \frac{\sum_{j=1}^{n} \left(\overset{\circ}{q}_j - \hat{q}_j^{(-j)} \right)^2}{n}$$

where $\hat{q}_j^{(-j)} = \begin{cases} \sum_{\substack{i= \\ i \neq j}}^{n} d_i K_b\left(x_j - x_i \right) \Big/ \sum_{\substack{i= \\ i \neq j}}^{n} e_i K_b\left(x_j - x_i \right), & for\ \hat{q}_x^{CH} \\ \sum_{\substack{i=1 \\ i \neq j}}^{n} \overset{\circ}{q}_i K_b\left(x_j - x_i \right) \Big/ \sum_{\substack{i=1 \\ i \neq j}}^{n} K_b\left(x_j - x_i \right), & for\ \hat{q}_x^{NW} \end{cases}$

depending on which estimator is being used.

In fact $\hat{q}_j^{(-j)}$ is the estimate of the rate of mortality using all the crude rates except the one for which i=j.

Theoretically minimising CV(b) is equivalent to minimising the mean integrated squared error that is defined as:

$$MISE(\hat{q}) = \int (\hat{q}_x - q_x) dx^2 = \int (E\hat{q}_x - q_x)^2 dx + \int V(\hat{q}_x) dx$$

$$= \text{int}\, egrated\ squared\ bias + \text{int}\, egrated\ \text{var}\, iance$$

Thus we can have a balance between bias and variance.

4.7 CHOICE OF THE KERNEL FUNCTION

The current literature indicates that the choice of the kernel function is not as influential as the value of bandwidth. However as already mentioned there are several kernel functions that can be used. Some kernel functions such as the Epanechnikov kernel with the property of minimizing asymptotically the mean square error or the

normal kernel are usually used for graduating mortality data (Gavin, Haberman and Verrall, 1994,1995). Another approach to define a kernel function is to choose them in such a way, so that to minimise the variance of the k'th differences of the graduated rates relative to the variance of the k'th differences of the crude rates subject to the constraints: $\int_{-\infty}^{\infty} K(x)dx = 1$ *and* $\int_{-\infty}^{\infty} x^2 K(x)dx = 0$, where K is a bounded function. This approach was firstly used for the purpose of choosing the best theoretical weights in moving weighted averages (London, 1985;Benjamin and Pollard, 1980; Ramsay, 1993). Gavin, Haberman and Verrall (1994) applied this approach in the context of kernel estimation and obtained the following kernel functions, for k=0 and k=1:

$$K(x) = \begin{cases} 3(3h^2 - 5x^2)/8h^3, & |x| \le h \\ 0, & |x| \succ h \end{cases}$$

and

$$K(x) = \begin{cases} 15(h^2 - x^2)(3h^2 - 7x^2)/32h^5, & |x| \le h \\ 0, & |x| \succ h \end{cases}$$

The kernel function for k=0 is discontinuous at the points $\pm h$ and therefore is rejected, while this for k=1 is continuous with support on the interval (-h, h).

Finally the use of higher order kernel functions is suggested as a bias reduction technique (Hastie and Loader, 1993). However Gavin, Haberman and Verrall (1994) using the kernel functions described above conclude to the fact that they do not perform well as the normal kernel and in addition they do not have better results in the sense of reducing bias over simpler positive kernel functions.

4.8 ADAPTIVE KERNEL ESTIMATOR

The Nadaraya–Watson kernel estimator of the true mortality rate has the disadvantage of the increase of the bias near the ends of the mortality table.

So the boundary problem may force cross-validation to select a smaller bandwidth at the boundary to reduce the bias, but this may lead to undersmoothing in the middle of the table.

In order to overcome this problem, Gavin Haberman and Verrall (1995) suggest using an adaptive kernel estimator.

The adaptive kernel estimator allows the bandwidth to vary according to the variability of the data. This means that in the case where the amount of exposure is large, a low value for the bandwidth results in an estimate that more closely reflects the crude rates. While when the amount of exposure is small such as at older ages then the estimate of the true rates of mortality, progress more smoothly if the bandwidth has a higher value. So at older ages the adaptive kernel estimator calculates local averages over a greater number of observations, which results in reducing the variance of the graduated rates with a possible increase of the bias.

Some adaptive models are proposed by Gavin, Haberman and Verrall (1995). Below are given two of these models:

a) $\hat{q}_i^t = \sum_{j=1}^{n} S_{ij}\dot{q}_j^t$ where $S_{ij} = \dfrac{K_{bi}(x_i - x)}{\sum_{j=1}^{n} K_{bi}(x_i - x_j)}$ $for\ i = 1,\ldots,n$

and $\hat{q}_i^t$ denotes the transformed crude rates.

In this model a different bandwidth is calculated for each age at which the curve is to be estimated. Then, using this bandwidth it is measured the distance from the age at which the curve is to be estimated to each of the observed ages. So, if the age to be estimated is x_i then it is measured the distance from x_i to x_j using b_i, for j =1,.....n.

b) $\hat{q}_i^t = \sum_{j=1}^{n} S_{ij}\dot{q}_j^t$ where $S_{ij} = \dfrac{K_{bj}(x_i - x_j)}{\sum_{j=1}^{n} K_{bj}(x_i - x_j)}$ $for\ i = 1,\ldots,n$

In this model it is calculated a different bandwidth, b_j, for each observed age x_j, for j =1, ..., n. Then using the bandwidth that corresponds to each observed age it is possible to measure the distance from the observed age to the age at which the curve is to be estimated. For example if the age to be estimated is x_i we measure the distance from x_i to x_j using b_j, for j = 1, ..., n.

4.9 KERNEL ESTIMATES OF INCOME DISTRIBUTIONS

In the literature there is great interest in analysing income distributions. In fact changes in the shape of income distributions are studied as well as how different groups are affected by them. The following features describe the shape of the income distribution: i) income levels and changes in the location of the distribution of the distribution as a whole ii) in income inequality and changes in the spread of the distribution and iii) clumping and polarisation and changes in patterns of clustering at various points along the income scale.

Generally increases in incomes shift the density concentration along to the right. Furthermore changes in income clumping and polarisation are revealed by shifts in the 'bumps' of income concentration at different points along the income scale. In addition changes in distributional location and clumping can be examined using the density function. Therefore a non-parametric approach can be adopted such as the technique of kernel density estimation.

Cowell, Jenkins and Litchfield (1994) suggest the use of kernel density estimates in order to reveal the features of the shape of the UK income distribution because it provides a succinct and informative summary of the details of the changes in ways that are easily understood. So they prefer this method over other methods such as indices of inequality or Lorenz curves and Pen's parade. Furthermore Schluter (1998) suggests the use of kernel estimates for studying the mobility of several countries since as it claims standard approaches based on mobility indices and transition matrices lead to misleading conclusions.

In addition Martin Biewen (2000) in order to get the shape of the income distribution of Germany considers kernel density estimates of the equivalent income.

However the results of the analysis depend on various underlying assumptions such as the choice of the smoothing parameter as well as the choice of a particular equivalence scale. Schluter (1996) considers the informal method of inspection also used by Deaton (1989). According to this method they firstly use a small bandwidth which produces erratic density estimates and they gradually use increased bandwidths until they get a smooth estimate.

Finally Marron and Schmitz (1992) consider the technique of kernel density estimation for simultaneously estimating several income distributions. However since the estimates depend on the bandwidth in order to compare several density estimates the same amount of smoothing must be applied to each curve. To overcome this

problem they consider taking the average of cross-validated bandwidths for individua samples. In fact they use a weighted average of the bandwidth coefficients. So they propose the following bandwidth:

$$\hat{h}_{p,j} = \hat{C}_p n_j^{-\frac{1}{5}}, for j = 1,\ldots,m$$

where $$\hat{C}_p = \sum_{j=1}^{n} w_j \hat{C}_{CV,j} \text{ and } w_j = \frac{n_j^{\frac{1}{5}}}{\left(\sum_{j=1}^{n} n_j^{\frac{1}{5}}\right)}.$$

Chapter 5

APPLICATIONS TO DEMOGRAPHIC DATA

In order to illustrate the applicability of the kernel technique as a graduation method we use it for graduating mortality data.

5.1 THE DATA

We use six sets of age-specific empirical death frequencies of both sexes of Finland for the year 1983, of New-Zealand for the year 1982 and finally mortality rates for the male and female population of Germany for the year 1988. The life tables are taken from the Central Statistical Office of each country.

For each age a measure of exposure (E_x) and the corresponding number of deaths (d_x) are given. We thus form the age specific death frequencies q_x, where $q_x = \frac{d_x}{E_x}$.

Scrutiny of these values for both the female and the male populations indicates a roughly exponential increase. The exceptions are the first years of life where the probability of dying drops sharply as well as for people between the ages 20 to 30, where the probabilities of dying draw a hump known in demographic literature as "the accident hump".

5.2 GRADUATION

The purpose of graduation in actuarial field is to provide a smooth sequence of graduated rates in order to, more closely reflect the variation due to age in the unknown true probabilities of dying compared to the observed death frequencies. The kernel technique as remarked in Chapter 3 provides a satisfactory smoothing of mortality data, although, the estimates may be influenced by boundary effects, at older ages.

As it is pointed out in Chapter 3, when using kernel graduation techniques the choice of the kernel function as well as the choice of the bandwidth parameter should be taken into account.

CHOICE OF THE BANDWDITH

Bloomfield and Haberman (1987) in order to choose the bandwidth they first fitted a curve to the data and then separately tested the graduated rates for smoothness using standard actuarial tests of fit. We followed this approach for the choice of the bandwidth for our data sets.

Firstly, we applied kernel graduation over the entire age range. Graduations were compared from the view points of goodness of fit and smoothness. In order to test the fidelity of kernel graduations to the original data we used the tests suggested by Bloomfield and Haberman (1987), described in Section 3.6.1. Then the graduation having the smallest sum of absolute test statistic was considered to have the best fit. The smoothness is then checked using the criterion suggested by Benjamin and Pollard i.e. to accept as the better graduation that which gives rise to the smaller total of the sum of the absolute values of the third differences. In addition the criterion suggested by Barnett and Haberman and Bloomfield (1987) is used, where the second differences are checked for changes of sign while third and fourth differences should satisfy the inequality $\left|\frac{\Delta^k \hat{q}_x}{\hat{q}_x}\right| \leq \frac{1}{A^k}$. Choosing a target value for A in order to check for smoothness is a matter of judgement, Barnett (1985) suggests a target value of A that equals 7, while Bloomfield and Haberman (1987) considered the case where A equals 4. When graduating the full age range for all the populations the goodness-of-fit tests do not always give statistics within the 5% limits which means that the null hypothesis is rejected and consequently the graduation is not acceptable. In fact the values of the deviations z_x are extremely large for the first four ages, which means that the kernel estimates for these ages are not close to the crude rates. So, as a result the values of $t(\chi^2)$ are not acceptable. In Tables 1 and 2 in Appendix A are illustrated the values of the test statistic $t(\chi^2)$ produced by the kernel graduation of the full age range. Generally the values of the test statistic $t(\chi^2)$ increase as the value of bandwidth h increase.

Therefore a second kernel graduation was carried out where the first four ages were not taken into account. Although the values of the test statistic $t(\chi^2)$ were dramatically reduced there were also cases where the null hypothesis was still rejected.

More analytically Table 1 in Appendix A displays the values of the test statistic $t(\chi^2)$ for the male and female populations of Finland, New-Zealand and Germany for the

full age range when using the normal kernel function. As it is obvious the test statistic is not accepted for any value of the bandwidth which means that all these graduations give unsatisfactory results. This is probably a phenomenon that is caused by the unsatisfactory kernel estimates produced for the first ages. Table 2, presents the values of the test statistics when the four first ages are excluded for the graduation, for the male and the female population of Finland. In the second column is given the value of the test statistic $t(\chi^2)$ with which we examine the fidelity of the kernel graduations to the original data, while the third and fourth columns give the values of the run test t(r) and of the serial correlation test $t(\rho)$. The last column is the absolute sum of the test statistic.

As we observe the only accepted graduations, for the male population of Finland are those for h=2 and h=2.25. According to the rule of thumb we use, the best fitted curve is produced for h=2, for the restricted age range. Investigating Barnett's criterion either with A = 7, or A = 4 the curve is very poorly smooth. When increasing the bandwidth to 5.5 or even to 7 with A=7 the curve is found to be poorly smooth while for A=4 the third differences satisfy the criterion up to ages 20 and 18 for h=5.5 and h=7 respectively. However most of the fourth differences failed the test. There also were 12 and 9 second sign changes for h=5.5 and h=7. In order to test smoothness we also used Benjamin and Pollard's criterion. Table 5 in Appendix A display the absolute sum of third differences when h=2, 5.5 or 7. We observe that bigger bandwidths produce smaller sums of absolute third differences. Although graduations for the entire age range were not accepted we checked the smoothness of the graduated rates. The same conclusions can be drawn as in the case of the restricted age range. A solely difference is that using Barnett's criterion with A=4 the third differences are smooth up to age 25.

Concerning the female population of Finland for the year 1983 and after applying the rule of thumb we get that the best fitted curve is produced for h=2. It is remarkable that the goodness-of-fit tests yield statistics within the 5% limits except for the $t(\chi^2)$ values for cases where h is greater than 2.25. Barnett 's criterion gives a very poorly smooth curve if A=7 or A=4 and bandwidth h=2 in both the restricted and the entire age range. Also in the case of the restricted age range, the smoothing is poor for h=5.5. In the case of the full age range if h=5.5 and A=4 third differences do satisfy the criterion for ages up to 33 while fourth differences fail the test but the criterion fails when A=7.

Figures 1 and 2 in Appendix-B present graphically the kernel normal estimates against the empirical ones for the male population of Finland for the year 1982. These figures refer to bandwidth equal to 2, 5.5 or 7, for both the restricted as well as the full age range. The corresponding results when the bandwidth equals 2 or 5.5 are given for the female population of Finland in Figures 4 and 5 in Appendix-B. It is obvious in these figures that for small bandwidths the graduated rates adhere too closely to the data. Furthermore it should be noticed that even for greater values of bandwidth than those displayed in the graphs since the data are too scanty no better smoothing is achieved.

The same analysis was performed to the male and female population of New-Zealand for the year 1982. Table 1 in Appendix-A shows the values of the test statistic $t(\chi^2)$ that are again rejected. A second kernel graduation was carried out without including the first four ages. The values of the test statistic are presented in Table 3 for both sexes. The best fitting curve for males is produced for h=4.75. However it should be noticed that in all graduations the number of positive durations n_1 was always smaller than 20. If we ignore the runs test the best fitting curve is produced for h=5.25. Using Barnett's criterion with A=7 the curve is found to be poorly smooth for the biggest part of the age range, for both the restricted as well as the full age range. Setting A = 4 the curve exhibit a smooth progression up to ages 18 for the restricted data set and up to age 22 for the entire data set. For h=5.25 third differences satisfy the criterion for the second half of the age range if A=7 while fourth differences fail the test for their majority. In the contrary if A=4 third differences up to age 21 satisfy this criterion and fourth differences for the second half of the full age range. The same conclusions are drawn for the restricted data set where the Barnett' s test is satisfied up to age 13. Table 6, shows the sum of absolute third differences due to Benjamin-Pollard criterion for testing smoothness.

For the restricted data set of females the best fitting curve is produced for h=4.75. However the curve is poorly smooth until age 29.

Figures 5 to 8 in Appendix-B display graphically the kernel normal estimates for different values of the bandwidth against the empirical ones for the males and females for both the restricted and the full data sets.

Finally goodness of fit tests were applied to the male and female population of Germany for the year 1988. Table 4, displays the values of test statistic for both males and females. Concerning the male population the best fitting curve is obtained for

h=4.5. Barnett's criterion for smoothness shows that if we set A=4 for both the restricted and the full age range, few third and fourth differences satisfy the inequality. However for A=4, the majority of the third differences satisfy the criterion while fourth differences do satisfy Barnett's criterion for the second half of the age range. Furthermore smoothness is checked when the bandwidth equals 6.5. If A=7, for the entire age range Barnett's criterion is not satisfied. However if A =4 all third differences up to age 23 as well as fourth differences for the second half of the age range satisfy the criterion. In the case of the restricted age range the curve is found to be smooth with the exception of the early ages. In Table 5 in Appendix A, the sum of the absolute third differences is given in order to check smoothness using Benjamin-Pollard's criterion.

Concerning the female population of Germany, the goodness of fit test $t(\chi^2)$ is rejected for any value of the bandwidth we tried. This may happen because of the unsatisfactory kernel estimates at higher ages and consequently the large deviations for these ages that contribute to a large value of the test statistic $t(\chi^2)$. Although these estimates are rejected when we test for their fidelity to the original data, we also applied the tests for checking their smoothness. When applying Barnett's test for bandwidth equal to 4.5 the curve is found to be very poorly smooth in the case A equals 7 for both the full and the restricted age range. To the contrary if A =4 most of the third differences satisfy Barnett's criterion while fourth differences do not. In addition we checked smoothness when the bandwidth equals 6. For the full age range if A=7 few third and fourth differences satisfy the criterion. However if A =4 the curve is found to be smooth for the second half of the age range. The same situation describes smoothness for the restricted age range.

Figures 9 to 12 in Appendix B present graphically the fitted q_x-values using kernel graduation technique with bandwidth equal to 4.5 or 6.5 for males and for females the bandwidth takes the values of 4.5 and 6, against the empirical ones. It also becomes obvious from the graphs that at early ages as well as at the end of the age range there is distortion from the true mortality rates.

CROSS-VALIDATION

Apart from the method used above where we first choose a model that best fits the data and then test for it smoothness, Gavin, Haberman and Verrall (1994) consider cross-validation as an objective method for selecting the smoothing parameter. As

mentioned in section 3.6.1, minimizing the cross-validation function CV(h) is equivalent to minimizing the mean integrated squared error which allows to have a balance between variance and bias. The only disadvantage of the method is that there are cases where it undersmooths the data.

So, in order to find the optimal bandwidth we applied the cross-validation method. In order to obtain the cross-validation score we used the C routine suggested by Hardle (1994) which is implemented in S-plus. However the program failed to give results for many cases where the generalized cross-validation function took negative values. An additional problem was that the method did not give cross-validation scores that corresponded to a clear minimum.

Although the choice of the kernel function is not as important as the choice of the bandwidth parameter it may influence the resulted kernel estimates. So in order to check how a different kernel function may influence the graduation of mortality data we used the Parzen kernel function. Generally the kernel estimates were not influenced by using this kernel function. The results were very similar to those produced when using the normal kernel function.

In order to evaluate the efficiency and accuracy of the kernel method with respect to graduation we compare it with the Heligman-Pollard model with 8 parameters (hereafter HP8).

This model was selected because it provides a satisfactory representation of the age pattern of mortality over the entire age range. The only disadvantage is that it provides systematic deviations from the adult ages since the accident hump of the estimated set of q_x values is located at a higher age than the empirical accident hump, in cases where the accident hump is too intense.

The parameters of the model have been estimated by a non-linear least-squares procedure. The function minimized was:

$$S^2 = \sum_x \left[\frac{\hat{q}_x}{q_x} - 1 \right]^2$$

and the algorithm E04FDF, part of the NAG library was used in order to calculate this sum of squares.

Table 6 in Appendix-A displays the values of the sums S^2 on the exit of the iterative procedure for both male and female population.

The estimated parameters of HP8 model for male and female populations of all the data sets considered, are shown in Table 7.

In order to compare the efficiency of the kernel method and the parametric model the sum $S^2 = \sum_x \left[\frac{\hat{q}_x}{q_x} - 1\right]^2$ was also calculated for the kernel estimates. Furthermore we computed the sum $S_1^2 = \sum_x (\hat{q}_x - q_x)^2$ for both the kernel estimates and the HP8 model. Finally we also calculated the $t(\chi^2)$ test for the HP8 model. These values are presented in the following tables:

MALES

	$S^2 = \sum_x \left[\frac{\hat{q}_x}{q_x} - 1\right]^2$	$S_1^2 = \sum_x (\hat{q}_x - q_x)^2$	$t(\chi^2)$
HP8	3.6845	0.00014	0.000838
Kernel Method-h=2.25	5.827	0.000108	11.972
Kernel Method-h=5.5	14.98	0.000436	34.43

FEMALES

HP8	7.8156	0.00028	0.12224
Kernel Method-h=5.5	43.2935	0.000273	35.82

Table 5.1.1: Values of the sum S^2, S_1^2 and $t(\chi^2)$ for male and female population of Finland 1983.

MALES

	$S^2 = \sum_x \left[\frac{\hat{q}_x}{q_x} - 1\right]^2$	$S_1^2 = \sum_x (\hat{q}_x - q_x)^2$	$t(\chi^2)$
HP8	0.316728	0.000475	7.52078
Kernel Method-h=4.75	11.61979	0.000253	41.45

Kernel Method-h=5.25	12.52767	0.000316	45.51

FEMALES

HP8	0.38189	0.0002075	8.2151
Kernel Method-h=5.5	10.552	0.005539	39.81

Table 5.1.2: Values of the sum S^2 , S_1^2 and $t(\chi^2)$ for male and female population of New- Zealand 1982.

MALES

	$S^2 = \sum_x \left[\frac{\hat{q}_x}{q_x} - 1\right]^2$	$S_1^2 = \sum_x (\hat{q}_x - q_x)^2$	$t(\chi^2)$
HP8	0.92786	0.000226	5.187
Kernel Method-h=4.5	9.6067	0.000295	32.93
Kernel Method-h=6.5	11.873	0.00077	43.92

FEMALES

HP8	0.84148	0.0018	37.916
Kernel Method-h=4.5	6.72831	0.000247	32.084
Kernel Method-h=6	9.23399	0.000527	40.75

Table 5.1.3: Values of the sum S^2 , S_1^2 and $t(\chi^2)$ for male and female population of Germany 1988.

We observe that the values of S^2 of the Heligman-Pollard are small in comparison with those of the kernel method. Thus the fit of the HP8 model at first ages is closer to the empirical mortality data. This verifies that at early ages the kernel method is influenced by boundary effects. The test S_1^2 produces similar values for both the parametric model and the kernel methods. This means that at older ages both methods produce similar results. Finally the $t(\chi^2)$ values reveal the fit of the graduated rates to the empirical ones. The results are also presented graphically in Figures 13 to 18 in Appendix B. From Figure 15 we observe that the kernel method is more "robust" in comparison with the HP8 model in the sense that it is not influenced by observations

that differ too much from the general pattern of the observed death frequencies. In fact the HP8 model is severely affected by only one observation that is too small (this that corresponds at age 11), while the kernel method does not seem to be influenced by it. Also it is remarkable that in comparison with the HP8 model, the smoothing of the kernel method at the edges of the age range and especially at ages between 0 and 20 is rather disappointing.

Finally a further comparison of the kernel regression approach was performed considering the case where the least squares approach is used in order to fit a curve to the local rates. The idea of using a local linear approach was introduced by Cleveland (1979). This method has the advantage over the classical approach of the kernel regression technique that it has a better behavior near the edges of the data range.

Therefore in order to find the optimal bandwidth we use the cross-validation technique. In order to obtain the cross-validation score we run the S-plus code that makes use of the hcv function which exist in the sm library (the sm library consists of a set of tools and functions within the S-plus environment that refer to the operations of the smoothing procedures), created by Bowman and Azzalini (1997). Furthermore the sm.density function was used in order to obtain the local linear estimators with normal kernel functions and constant bandwidth.

Generally the cross-validation technique did not result to a local minimum and the bandwidths suggested by the method undersmoothed the data. Figure 20 in Appendix B shows graphically the cross-validation scores and the corresponding bandwidths. No local minimum is apparent in the graph.

Figure 19 presents the estimates produced by the local linear approach when using the normal kernel function and different values of bandwidth. As it is obvious for small values of bandwidth the smoothing is more satisfactory at early ages but there is the problem that it may reproduces closely the observed values. Furthermore as it is presented in Figures 21 to 26 for large bandwidths the kernel estimates behave better than the local kernel estimates at early ages while the opposite happens for the end of the age range. Finally it is remarkable that the HP8 model achieves a satisfactory smoothing at early ages in comparison to both kernel and local kernel estimates.

Chapter 6

APPLICATIONS TO ECONOMIC DATA: THE CASE OF INCOME DATA

The kernel method has primarily been developed for density estimation problems. This method has proved to be a very useful tool particularly for graphical illustration of the shape of income distributions. In particular kernel techniques result in smooth density estimates that make easier the comparison between different states (such as differences in time, differences between population groups, countries etc). In this chapter we evaluate this method for representing differences in time and between countries.

6.1 THE DATA

In our analysis we use the data from Panel Comparability project (PACO)[1].

PACO Database contains comparable micro-data based on national and regional data for seven European countries: Germany, Lorraine/ France, Luxembourg, United kingdom, Poland, Hungary and United States (USA). It has been created by CEPS/INSTEAD (Centre d' Etude de Population, de Pauvrete et de Politique Socio-Economiques/ International Networks for Studies in Technology, Emvironment, Alternatives, Development) in partnership with DIW (Deutches Institut fur Wirtschoftsforschung). Table 5.1contains PACO data for different countries.

Our analysis we based on microdata and meta-data from the national household panels of the following countries: Germany, France, Luxembourg, United Kingdom (UK), Poland, Hungary and United States (USA).

[1] This project was funded by the European Commission, under the Human Capital and Mobility Program (1993-1996).

COUNTRY	YEAR	COUNTRY	YEAR
France (Lorraine)	1985-1990	Germany	1984-1996
Hungary	1992-1994	Luxembourg	1985-1992
Poland	1987-1990 & 1994-1996	UK	1991-1993
USA	1983-1987		

Table 6.1: PACO data for different countries.

The variables contained in the PACO database refer to a large number of social and economic characteristics of the household and its members such as: income, demographic characteristics, labour force and work history, to education and family background and housing elements.

In our analysis the annual disposable household income is used as a measure of economic status. In order to compare households with different size and composition an equivalence scale is applied to the income distributions. According to this equivalence scale, each household's disposable income is weighted by the number of persons in that household. In particular a weight of 1 is assigned to the first member of the household and 0.5 for each additional member of the household. Furthermore we consider distributions of the relative income i.e. the annual disposable income divided by the corresponding mean income in order to be comparable between countries and have a better view of inequalities changes. Thus the relative disposable income is considered for all the European countries mentioned above as well as for the USA.

6.2 KERNEL DENSITY ESTIMATION

The non-parametric approach of kernel density estimation is proposed for describing the income distributions since it allows a useful graphical representation of the data but also it overcomes problems that arise when using parametric approaches. So, we used the kernel estimation method in order to reveal the shape of the income distribution of different European countries as well as of the USA. However when using the kernel estimation method the choice of the bandwidth parameter is of great importance since it controls the degree of smoothing.

As remarked in section 3.2 several methods have been proposed in the literature but none of them has gained great acceptance. Among them the cross-validation technique is widely used in the case of income distributions (Marron and Schmitz, 1992). In order to find the appropriate bandwidth in our analysis we followed two different approaches. Firstly we calculated the optimal bandwidth by using the expression suggested by Silverman (1986) (also used by Cowell et al. (1994)) for the pilot bandwidth. This bandwidth is defined as:

$$h_{opt} = 0.9An^{-1/5} \qquad (5.1)$$

where A= min(sample standard deviation, interquantile range/1.34).

Furthermore we used the informal method of inspection for the choice of the bandwidth parameter. When using this method the following algorithm is applied: the starting point is a small width producing an erratic density estimate then the bandwidth is gradually increased until a smooth estimate is arrived at.

In order to examine more analytically the behavior of the kernel estimates we calculated them for representing the shape of income distributions of different countries. All the estimates were derived using STATA kernel density estimation programs written by Salgado-Ugarte *et al.*(1993). We calculated the optimal bandwidth defined in equation 5.1. We also calculated the kernel estimates using a smaller bandwidth than the optimal that equals 0.01. In Appendix C in Figures 1 to 4 we present the kernel estimates of the following countries: Germany (1990), UK (1991), Luxembourg (1985) and Poland (1987) using both the optimal as well as a bandwidth equal to 0.01. From these figures we observe the importance of choosing an appropriate bandwidth parameter since a very small bandwidth undersmooths the data while the optimal bandwidth implies a smooth curve. In order to get estimates that allow us to have a representative view of the income distributions we use the informal method of inspection for the choice of the bandwidth parameter. Thus, we proceeded in our analysis by using bandwidths that are greater than the optimal. In Figures 5 to 8 in Appendix C are shown the distributions of disposable incomes for the following countries: Germany (1990), UK (1991), Luxembourg (1985) and Poland (1987) using different bandwidths. From these graphs we can observe that for the purposes of this chapter the most representative shape of income distributions is obtained for bandwidth equal to 0.15. In the case of larger bandwidths (e.g. for

bandwidth equal to 0.2 or 0.5) we get a more smooth picture of income distributions at a loss of some important characteristics of their shape. This phenomenon is apparent for the income distribution of UK for the year 1991, shown in Figure 6. So, for a large bandwidth (i.e. for a bandwidth equal to 0.5) the important feature of the existence of a double mode that characterises the UK income distribution is lost. For a bandwidth equal to 0.15 we get a realistic picture of the UK income distribution. When using the optimal bandwidth we get estimates that are very close to those obtained for a bandwidth equal to 0.15. However the optimal bandwidth is influenced by individual observations, especially around the mode of the distribution. From Figures 5, 7 and 8 we observe that we have a more representative picture of the income distributions of Germany (1990), Luxembourg (1985) and Poland (1987) for a bandwidth equal to 0.15. It is obvious that small values of bandwidth produce density estimates that are more close to the observed data i.e. they display the variation associated with individual observations rather than the underlying structure of the whole sample. In the contrary when the bandwidth is large the structure of the data is obscured, by oversmoothing them.

It is known that for larger values of the bandwidth more smoothing is achieved however we should choose the bandwidth and consequently the degree of smoothing depending on what is the aim of our analysis. Although the informal method of inspection allows examining how the density estimate changes its characteristics when different bandwidth parameters are used, a more objective method should be considered in order to find the optimal bandwidth.

From Figures 1 to 8 we observe that the kernel functions have transferred positive weight to the negative axis. Bowman and Azzalini (1995) mentioned that in the case where only positive values can be recorded the kernel functions centred on observations that are very close to zero transfer positive weight to the negative axis. They also suggest reducing this effect by using a smaller bandwidth but then we get estimates that undersmooth the data. An alternative approach in order to overcome this problem is to transform the data by taking logarithms (Bowman & Azzalini, 1995 ; Silverman,1986). So, we applied the kernel method to the logarithms of the data points and we performed the appropriate inverse transformation. Thus, if the density estimates of the logarithms of the data are given by $\hat{g}(\log x)$, the estimates in the original scale are given by the expression: $\hat{f}(x) = \frac{1}{x}\hat{g}(\log x)$.

In Figures 8 to 12 are presented the kernel estimates using the log transformation of the following countries: Germany (1990), UK (1991), Luxembourg (1985) and Poland (1987). We observe that the estimates lie on the positive section of the axis. A bandwidth equal to 0.15 has proved to be more appropriate for the log transformed kernel estimates.

Since kernel methods allow us to compare several income distributions we used this method to detect inequality changes. In particular changes in the income distributions are examined for the periods 1985 to 1990 for the five following countries: France, Germany, Luxembourg, Poland and USA. The results obtained using kernel estimates are shown in Figures 13 and 14 in Appendix C. In the case that no data were available for the years 1985 and 1990 we considered available data for years close to the desirable ones.

From Figure 13 we observe that for the year 1985 the proportion of population at higher income levels is greater for USA, France and Germany. In addition the concentration of population at very low incomes is higher for Poland as well as for USA.

For the disposable incomes of the year 1990, shown in Figure 14 we observe that the distribution with the higher modes around the mean is that of Luxembourg (1990). The proportion of the population with high incomes is greater for the following countries: Germany and USA. In addition for Poland, France and Luxembourg the proportion of the population with high incomes is concentrated at the same levels.

Furthermore in Figures 15 to 19 in Appendix C is displayed the relative disposable income of Poland for the years 1987 and 1990, as well as the relative disposable income of Germany, France, Luxembourg and of the USA for the years 1985 and 1990. From Figure 16 is obvious that there was a shift of the upper tail of the distribution to the right. In addition concentration of the population at very high incomes is greater for the year 1990. Furthermore the distribution for the year 1985 has a higher mode around the mean in comparison with the distribution of the year 1990.

From Figure 17 we observe that the distribution for the year 1990 has a higher mode around the mean in comparison with the distribution of the year 1985. Furthermore there is a shift in density towards low relative incomes combined with a shift towards higher relative incomes. From Figure 18 it is obvious that there is a shift of the low tail of the distribution to low incomes. Furthermore the distribution for the year 1990

has a higher mode around the mean. Finally from Figure 19 where the income distribution of the USA is presented we observe that there are no great differences between the income distributions for the years 1985 and 1990.

Chapter 7

CONCLUSIONS

The subject of graduating the age pattern of mortality is of great interest in demographic analysis. In addition the graduation of economic data has great practical value since it allows an easier interpretation of income densities and consequently a comparison of how income inequality changes between different countries or different years. We focused on a non-parametric graduation using kernel methods. A presentation of kernel methods was provided in the context of a density estimation technique as well as a regression technique. Furthermore its applicability and accuracy was examined by graduating mortality as well as income data. A comparison of the kernel method with the eight-parameter Heligman-Pollard model was also performed for the mortality data sets. Finally local least-squares estimates were applied to the same data sets since these are considered to produce better estimates than the classical kernel estimates at the edges of the age range.

Kernel graduation technique was applied to six mortality data sets. The kernel method provides an easily applied method of graduation, which in comparison with parametric techniques does not require the estimation of large number of parameters. In addition a balance between a high level of smoothing or a close fit to the data can be achieved through the choice of the optimal bandwidth. In order to choose the bandwidth we first used the most traditional actuarial approach, according to which we determine graduations that provide a good fit and then test for its smoothness. In addition in order to find the optimal bandwidth we used the cross-validation technique. In the first method used for the choice of the bandwidth and due to standard graduation criterion, graduations for the full age range were not accepted for any one of the data sets. When we excluded the four first ages from graduation the results became more satisfactory however there were still case for which the graduations were rejected. Especially for the data set of Finland 1983 for both male and female population, where data are very scanty, the bandwidth that provided the best fit undersmoothed the data. The smoothness criterion with a value of A=7 provided a unsatisfactory smoothing for all the data sets for both the full and the restricted age range. While the same criterion for a value of A=4 proved more satisfactory since it provides a smooth curve for the second half of the age range.

However even in this case the curve is poorly smoothed for the early ages. Generally even for larger values of the bandwidth parameter was not achieved a better smoothing. So, the kernel method failed to provide a satisfactory smoothing over the entire age range. Furthermore the cross-validation technique failed to give bandwidths that provide a satisfactory smoothing of the curve. As a consequence none of the methods mentioned above proved able to provide a bandwidth that adequately smooths the data. Similar results were produced using either the normal or the parzen kernel function.

The kernel graduation technique was also compared with the Heligman-Pollard model with eight parameters (HP8) that generally provides a complete and appropriate representation for the mortality pattern for the entire age range. For all the data sets the HP8 model provides a better smoothing of the data at earlier ages as well as at the ends of the age ranger in comparison with the kernel method. However the kernel method is more robust than the HP8 model since it is not influenced by isolated observations.

Finally, Gavin, Haberman and Verrall (1994) suggested fitting higher order functions locally since then the bias is eliminated without a great increase in variance. Therefore we used a local linear estimates by a least-squares approach. These estimates provided a better smoothing than the kernel estimates for small values of bandwidth, especially at the beginning of the age range. However for larger values of bandwidth the method provided estimates at early ages that differ too much from the observed values. At the end of the age range local linear estimates provide a more satisfactory smoothing than the kernel estimates. However the HP8 model provides a superior smoothing at early ages in comparison with the local linear estimates.

So, the HP8 model proves more accurate for the graduation of the mortality pattern of the overall age range while the non-parametric estimates are influenced by boundary effects.

Furthermore the kernel density estimation method was applied to income data in order to reveal the income distribution of four European countries as well as of the USA. Kernel density estimation methods provide an easy and informative summary of the details of the changes in ways that are easily understood. Also the fact that it provides smooth estimates of the income densities allows an easier comparison between different states (such as differences in time, between countries etc.). Generally the choice of the bandwidth parameter may influence the resulting estimates and

consequently the conclusions about the shape of the income densities revealing some special characteristics of them. We followed two different approaches for the choice of the appropriate bandwidth for our analysis. The optimal bandwidth obtained using the expression suggested by Silverman (1986) resulted in smooth estimates. However these estimates were influenced by individual observations especially around the mode of the distribution. So the informal method of inspection was considered more appropriate for the choice of the bandwidth parameter. Small bandwidths produced erratic estimates that were very close to the individual observations, while larger values of bandwidths oversmoothed the data. A bandwidth equal to 0.15 was considered appropriate for our analysis. Although kernel methods provide smooth estimates they have the disadvantage of transferring positive weight to the negative axis. In order to overcome this problem we transformed the data by taking logarithms. Furthermore we presented the income distributions of Germany, France, Poland, Luxembourg and of the USA for the periods 1985 and 1990 in order to detect inequality changes of their income distributions.

APPENDIX A

MALES					
FINLAND		**NEW-ZEALAND**		**GERMANY**	
h	$t(\chi^2)$	***h***	$t(\chi^2)$	***h***	$t(\chi^2)$
2	21.74535	1.5		2	9.206036
2.5	14.42553	2.5	18.849	2.5	14.72895
3	18.76371	3	24.777	3	19.30029
4	25.66436	3.5	30.009	3.5	26.50741
4.5	28.70522	4	34.914	4	29.79535
5	44.50805	4.25	37.103	4.5	32.93405
5.5	34.43451	4.5	39.354	5	35.82427
6	38.76288	4.75	41.459	5.5	38.76288
		5	43.528	6	40.754
				6.5	43.927
FEMALES					
FINLAND		**NEW-ZEALAND**		**GERMANY**	
h	$t(\chi^2)$	***H***	$t(\chi^2)$	***h***	$t(\chi^2)$
2	9.2	1.5	1.992	2	18.035
2.5	14.728	2.5	16.25	2.5	21.086
3	19.3	3	21.518	3	23.85
4	26.5	3.5	26.293	3.5	26.589
4.5	29.7	4	30.75	4	29.333
5	32.934	4.25	32.888	4.5	32.084
5.5	35.824	4.5	34.914	5	35.086
6	38.762	4.75	36.87	5.5	37.798
		5	38.839	6	40.75
				6.5	43.49

Table A.1: Test Statistic $t(\chi^2)$ for the full are-range of the male and female population of Finland, New-Zealand and Germany respectively, using the normal kernel function.

BANDWDITH	$t(\chi^2)$	t(r)	t(ρ)	ABSOLUTE SUM*
		FINLAND-MALES		
1.5	-4.3466	3.577	-0.59924	
2	0.690184	3.384	-0.551	
2.25	2.334261	2.606	-0.51354	4.625184
2.5	3.569003	1.624	-0.47278	5.453801
2.75	4.64863	1.624	0.43196	
3	5.522127	0.832	0.950664	
3.25	6.26352	0.832	-0.34406	
3.5	7.01522	0.7299	-0.29768	
3.75	7.700204	0.7299	0.25894	
4	8.261344	0.318	-0.21304	
4.5	9.470811	0.05226	-0.01591	
5	10.657	0.05226	-0.00669	
5.5	11.84177	0.0185	-0.30522	
		FINLAND-FEMALES		
2	0.092658	4.07	0.011486	4.174144
2.25	1.702776	3.892	-0.46529	6.060066
2.5	3.023922	3.892	-0.42516	
2.75	4.6115687	3.892	-0.00326	
3	5.168033	1.997	-0.00414	
3.88	7.923828	1.967	-0.152	
4	-12.3062	1.5	-0.13315	
4.5	9.818287	1.199	-0.03747	
5	11.5209	0.69	0.05335	

Table A.2: Test Statistics of the restricted age range, for the female and male population of Finland of the year 1983, using the normal kernel function.

* Absolute sum of the test statistics, calculated when the value of $t(X^2)$ is accepted. The bandwidth for the smaller absolute sum gives the best fit of the curve

BANDWDITH	$t(\chi^2)$	t(r)	t(ρ)	ABSOLUTE SUM*
		NEW-ZEALAND MALES		
3	-7.49027	-4.96	0.412048	
3.5	-6.13914	-5.397	0.486704	
4	-4.62127	-4.0112	0.542268	
4.25	-3.81065	-4.0112	0.567982	
4.5	-2.99237	-4.0112	0.592169	7.595739
4.75	-2.13042	-3.9028	0.610473	6.643693
5	-1.27408	-7.478	0.630984	9.383064
5.25	-0.40084	-7.478	0.647245	8.526085
5.5	0.572899	-7.478	0.662632	8.713531
5.75	1.413185	-7.478	0.675541	9.566726
6	2.377808	-7.564	0.691466	9.941808
6.25	3.356285	-7.564	0.701318	
6.5	4.350665	-7.564	0.713093	
		NEW-ZEALAND FEMALES		
3	-6.65225	-5.17	0.019873	
3.5	-5.00486	-5.17	0.458265	
4	-3.17008	-5.04	0.523858	
4.25	-2.20675	-5.109	0.54445	7.8602
4.5	-1.24104	-5.604	0.570238	7.415278
4.75	-0.21416	-5.544	0.592758	6.350918
5	0.83536	-5.544	0.6071	6.98646
5.25	1.875192	-5.544	0.623714	8.042906
5.5	2.95702	-5.544	0.63814	9.13916
5.75	4.089742	-5.544	0.648017	
6	5.196759	-5.99	0.663723	

Table A.3: Test Statistic for the restricted are-range of the male and female population of New Zealand for the year 1982, using the normal kernel function.

* Absolute sum of the test statistics. The bandwidth for the smaller absolute sum gives the best fit of the curve

BANDWDITH	$t(\chi^2)$	t(r)	t(ρ)	ABSOLUTE SUM*
		GERMANY-MALES		
3	-.48648	-3.903	0.70664	
3.5	-1.912264	-4.363	0.69392	6.969184
4	-1.1018	-4.86	0.685641	6.647441
4.25	-0.44222	-4.86	0.67887	5.98109
4.5	0.160195	-4.86	0.680653	5.700848
4.75	0.853203	-4.86	0.682009	6.395212
5	1.574874	-4.86	0.93746	7.372334
5.5	3.223032	-5.204	0.694964	
6	4.887244	-5.66	0.707725	
		GERMANY-FEMALES		
1.5	13.34721	-3.306	-0.03144	
2	13.17072	-3.784	-0.03161	
2.5	13.48705	-3.223	-0.03446	
3	13.58727	-2.524	-0.03017	
3.5	13.86522	-3.3	-0.01509	
4	14.32667	-4.11	0.010353	
4.5	15.02144	-4.11	0.047902	
5	16.26845	-4.24	0.084023	

Table A.4: Test Statistic for the restricted are-range of the male and female population of Germany, using the normal kernel function.

* Absolute sum of the test statistics. The bandwidth for the smaller absolute sum gives the best fit of the curve

BANDWIDTH	$\Sigma \mid 10^5 * \Delta^3 q_x \mid$	
	RESTRICTED AGE RANGE	FULL AGE RANGE
FINLAND-MALES		
2	8187	8356
5.5	550	570
7	352	361
FINLAND-FEMALES		
2	3353	3513
5.5	351	333
NEW-ZEALAND MALES		
4.75	443	377
5.25	400	347
NEW-ZEALAND FEMALES		
4.75	443	377
GERMANY-MALES		
4.5	564	613
6.5	352	368
GERMANY-FEMALES		
4.5	505	538
6	347	364

Table A.5: Benjamin-Pollard criterion for checking smoothness for the male and female population of Finland.

HP8-FORMULA			
	FINLAND	***NEW-ZEALAND***	**GERMANY**
MALES	7.81	0.31	0.75
FEMALES	7.81	0.38	0.84

Table A.6: Sums of squares of the relative deviations between the empirical and the fitted qx-values using HP8 formula.

	$A*10^4$	**$B*10^3$**	**$C*10^3$**	**$D*10^4$**	**E**	**F**	**$G*10^6$**	**$H*10^2$**
			FINLAND					
MALES	4.43	42.29	128.11	2.76	4.84	25.83	9.57	111.8
FEMALES	4.42	42.14	128.01	2.77	4.94	25.77	9.59	111.8
			NEW-ZEALAND					
MALES	13.49	29.6	121.11	15.08	11.33	20.94	42.58	110.51
FEMALES	11.06	40.85	127.77	4.75	12.02	18.95	34.08	109.98
			GERMANY					
MALES	9.38	53.26	132.29	6.17	10.45	21.21	65.04	110.00
FEMALES	9.03	66.95	126.46	2.10	11.25	19.68	22.26	110.77

Table A.7: Estimated values for parameters A, B, C, D, E, F, G and H for males and females of Finland, New-Zealand and Germany respectively, using the HP8 formula.

APPENDIX B

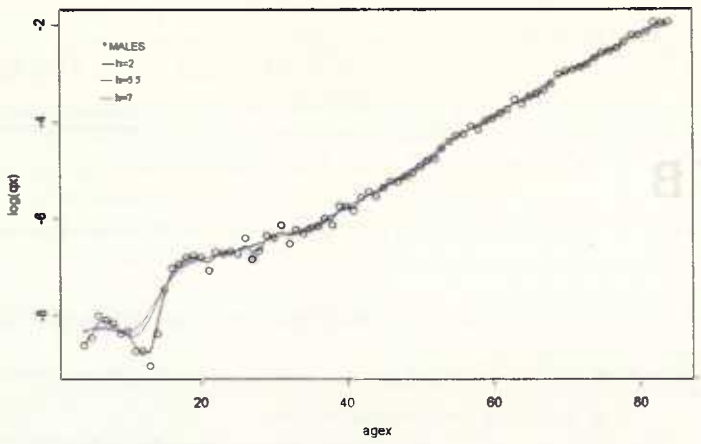


Figure B.1: Empirical q_x-values of the restricted age range for the male population of Finland 1983 (circles) and kernel normal estimates using different values of the bandwidth parameter.

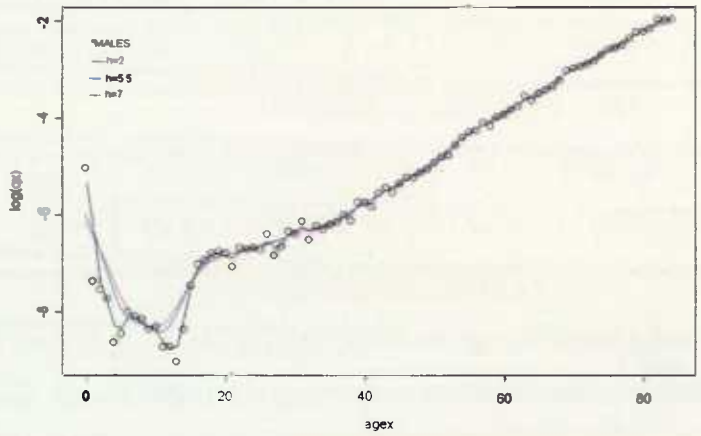


Figure B.2: Empirical q_x-values of the full age range for the male population of Finland 1983 (circles) and kernel normal estimates using different values of the bandwidth parameter.

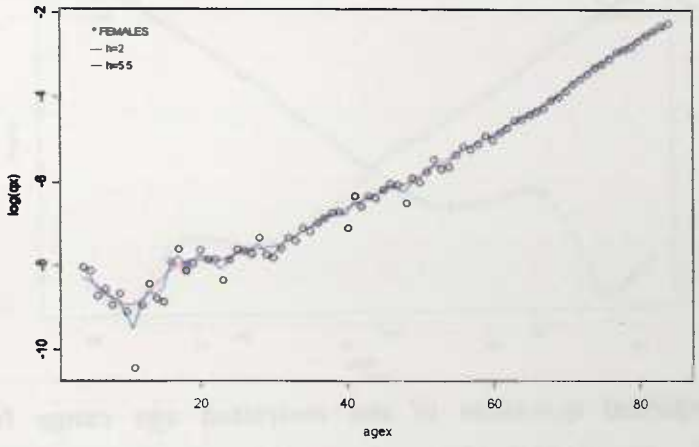


Figure B.3: Empirical q_x-values of the restricted age range for the female population of Finland 1983 (circles) and kernel normal estimates using different values of the bandwidth parameter.

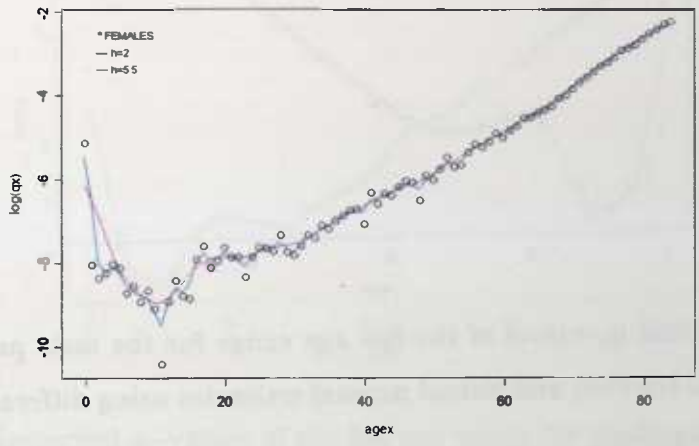


Figure B.4: Empirical q_x-values of the full age range for the female population of Finland 1983 (circles) and kernel normal estimates using different values of the bandwidth parameter.

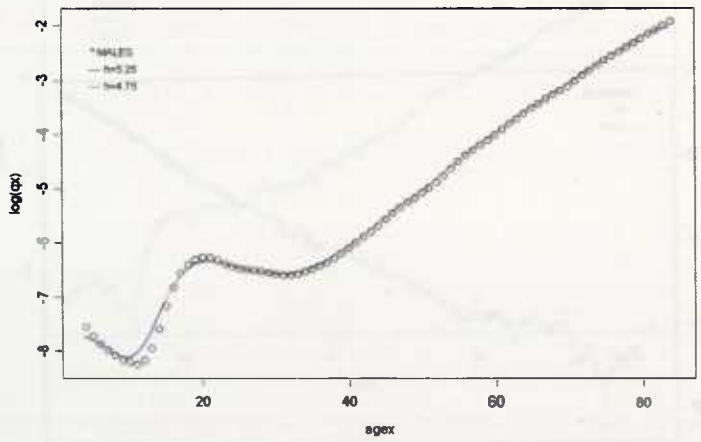


Figure B.5: Empirical q_x-values of the restricted age range for the male population of New-Zealand 1982 (circles) and kernel normal estimates using different values of the bandwidth parameter.

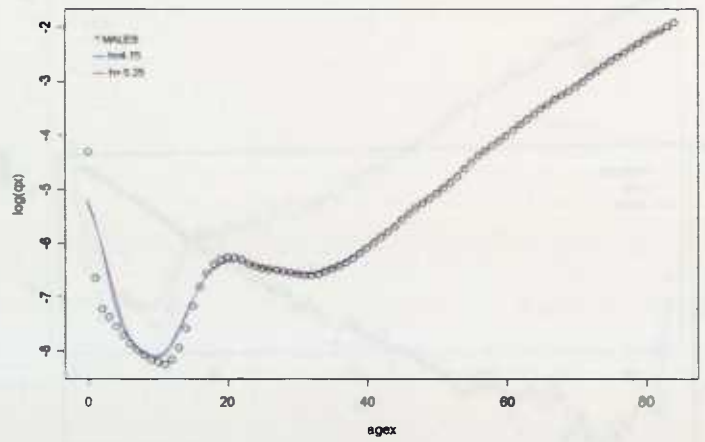


Figure B.6: Empirical q_x-values of the full age range for the male population of New-Zealand 1982 (circles) and kernel normal estimates using different values of the bandwidth parameter.

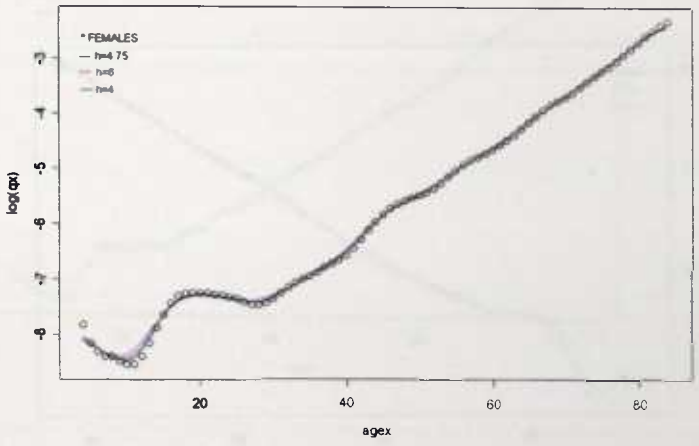


Figure B.7: Empirical q_x-values of the restricted age range for the female population of New-Zealand 1982 (circles) and kernel normal estimates using different values of the bandwidth parameter.

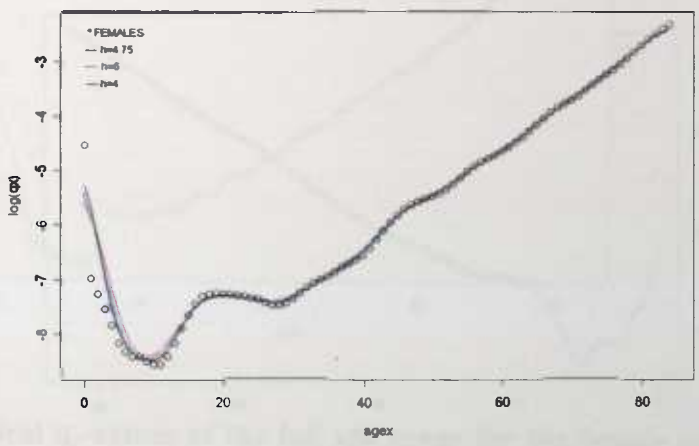


Figure B.8: Empirical q_x-values of the full age range for the female population of New-Zealand 1982 (circles) and kernel normal estimates using different values of the bandwidth parameter.

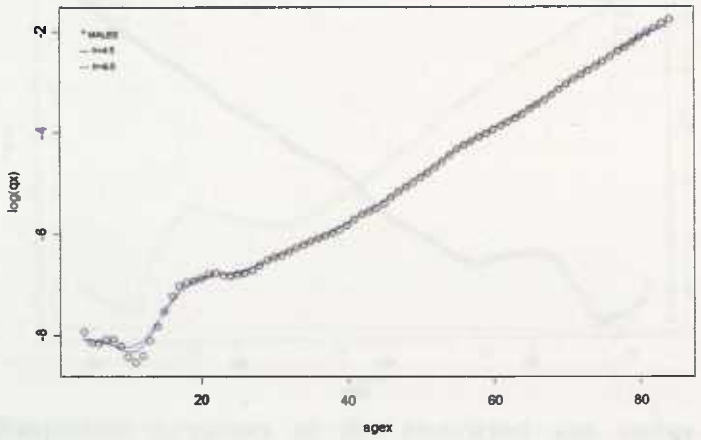


Figure B.9: Empirical q_x-values of the restricted age range for the male population of Germany 1988 (circles) and kernel normal estimates using different values of the bandwidth parameter.

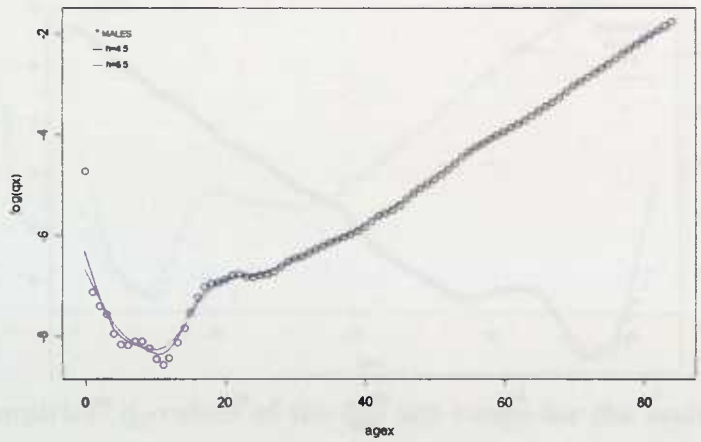


Figure B.10: Empirical q_x-values of the full age range for the male population of Germany 1988 (circles) and kernel normal estimates using different values of the bandwidth parameter.

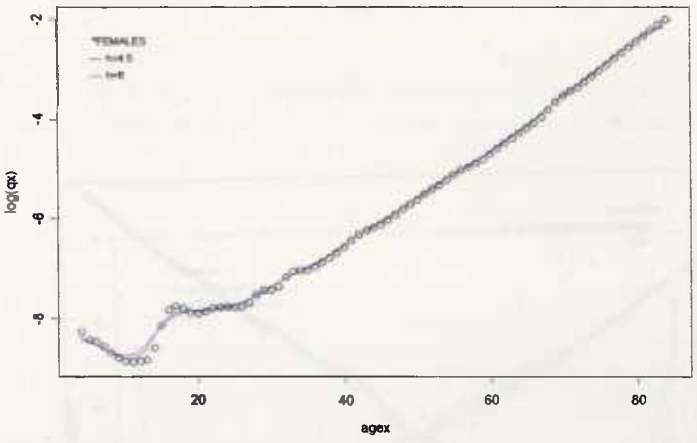


Figure B.11: Empirical q_x-values of the restricted age range for the female population of Germany 1988 (circles) and kernel normal estimates using different values of the bandwidth parameter.

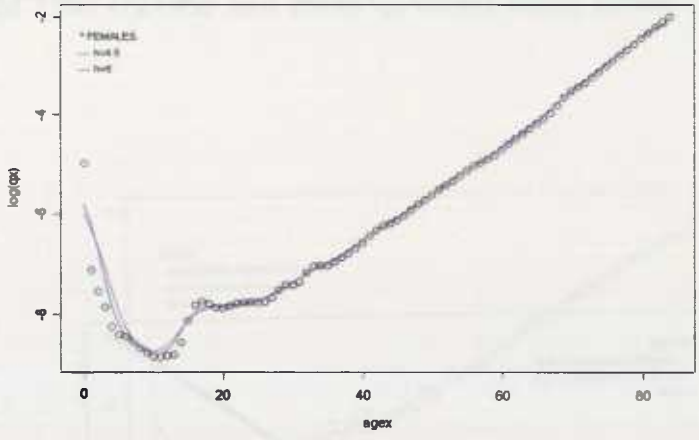


Figure B.12: Empirical q_x-values of the full age range for the female population of Germany 1988 (circles) and kernel normal estimates using different values of the bandwidth parameter.

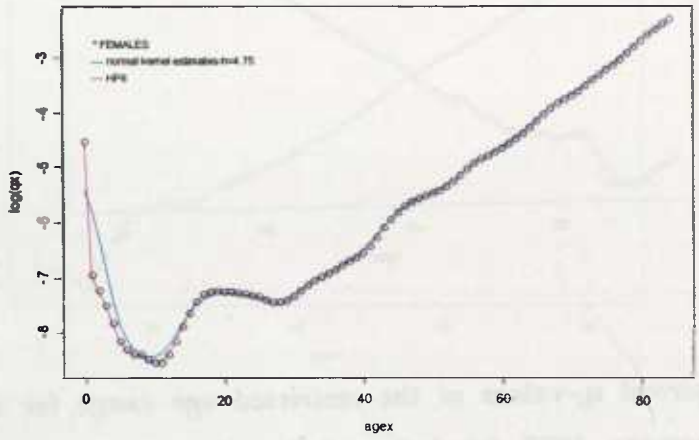


Figure B.13: Empirical q_x-values of the full age range for the female population of New-Zealand 1982 (circles) and fitted q_x-values using kernel graduation and HP8 formula.

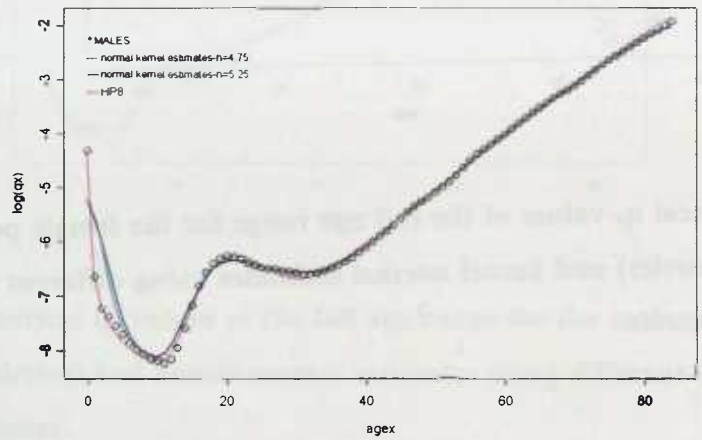


Figure B.14: Empirical q_x-values of the full age range for the male population of New-Zealand 1982 (circles) and fitted q_x-values using kernel graduation and HP8 formula.

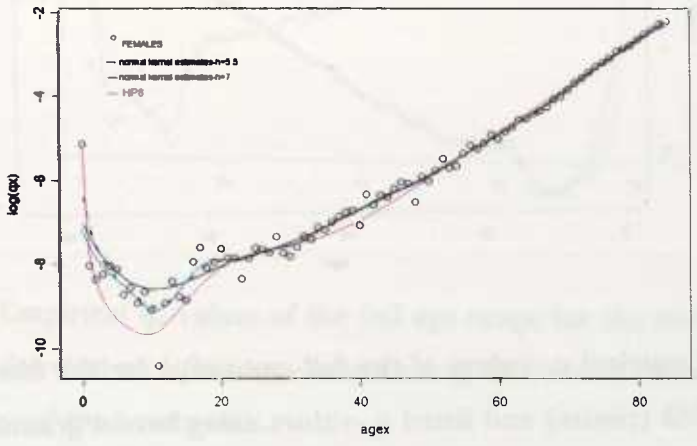


Figure B.15: Empirical q_x-values of the full age range for the female population of Finland 1983 (circles) and fitted q_x-values using kernel graduation and HP8 formula.

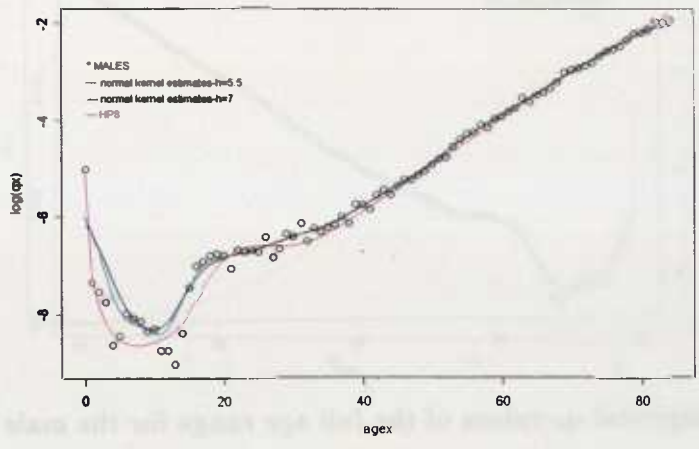


Figure B.16: Empirical q_x-values of the full age range for the male population of Finland 1983 (circles) and fitted q_x-values using kernel graduation and HP8 formula.

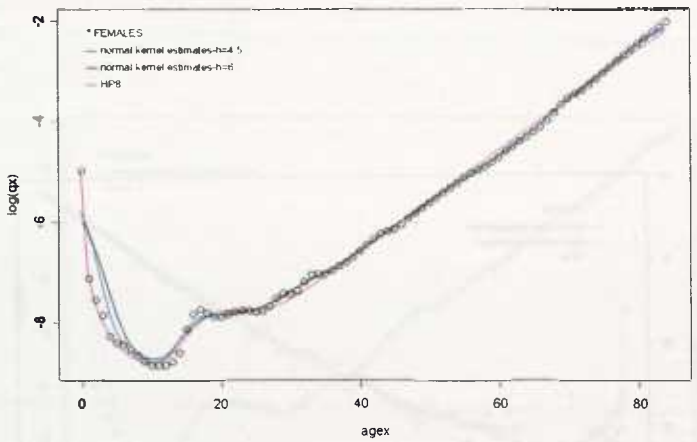


Figure B.17: Empirical q_x-values of the full age range for the female population of Germany 1988 (circles) and fitted q_x-values using kernel graduation and HP8 formula.

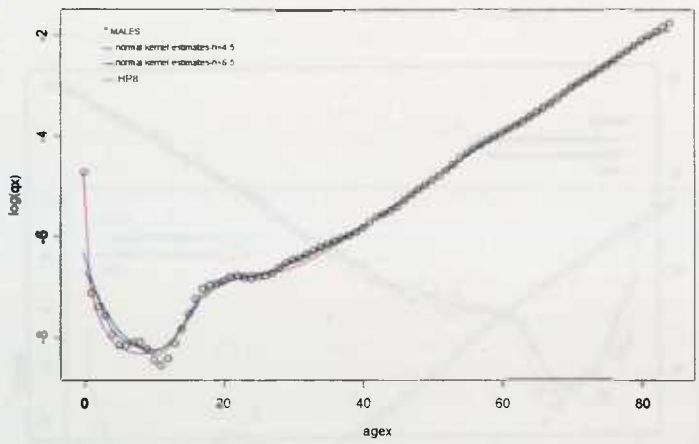


Figure B.18: Empirical q_x-values of the full age range for the male population of Germany 1988 (circles) and fitted q_x-values using kernel graduation and HP8 formula.

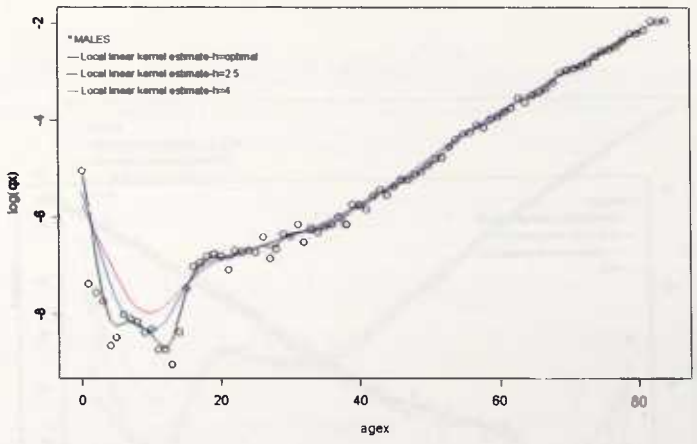


Figure B.19: Empirical q_x-values of the full age range for the male population of Finland 1983 (circles) and fitted q_x-values using local linear kernel estimates for different values of the bandwidth.

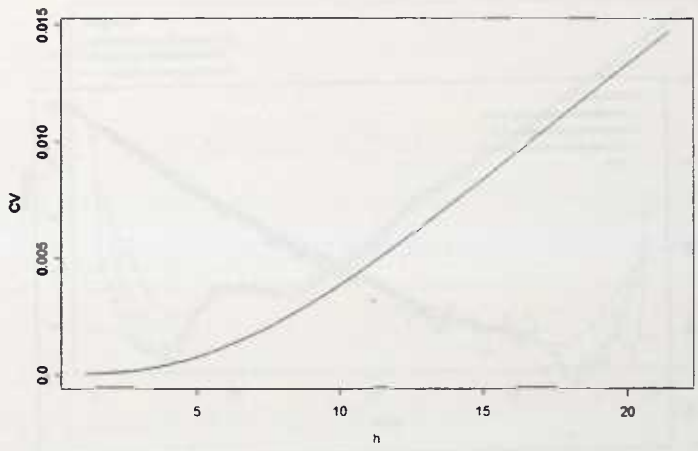


Figure B.20: CV scores for the female population of Germany 1988.

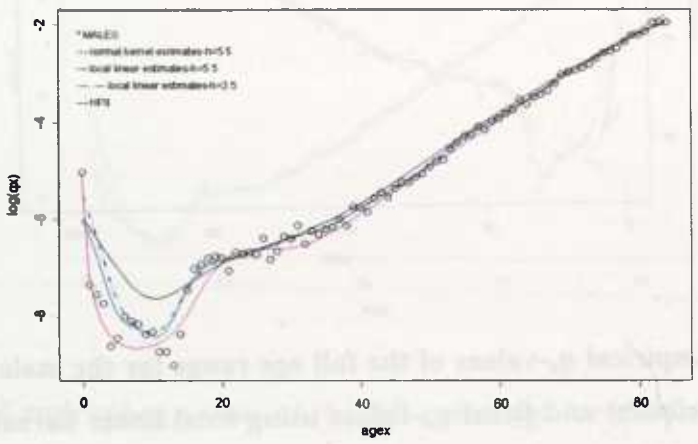


Figure B.21: Empirical q_x-values of the full age range for the male population of Finland 1983 (circles) and fitted q_x-values using normal kernel estimates, local linear kernel estimates and HP8 model.

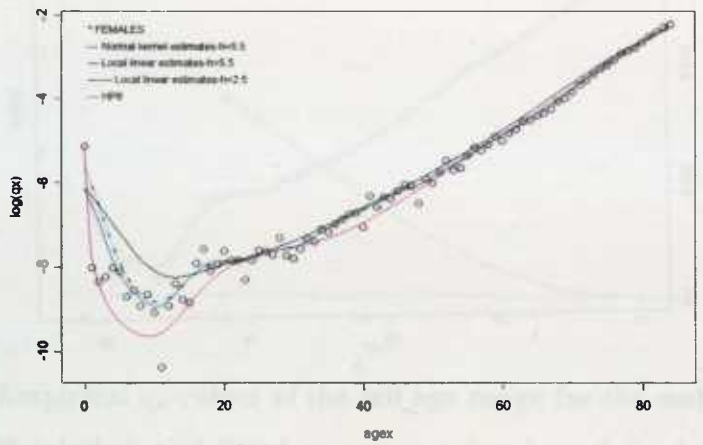


Figure B.22: Empirical q_x-values of the full age range for the female population of Finland 1983 (circles) and fitted q_x-values using normal kernel estimates, local linear kernel estimates and HP8 model.

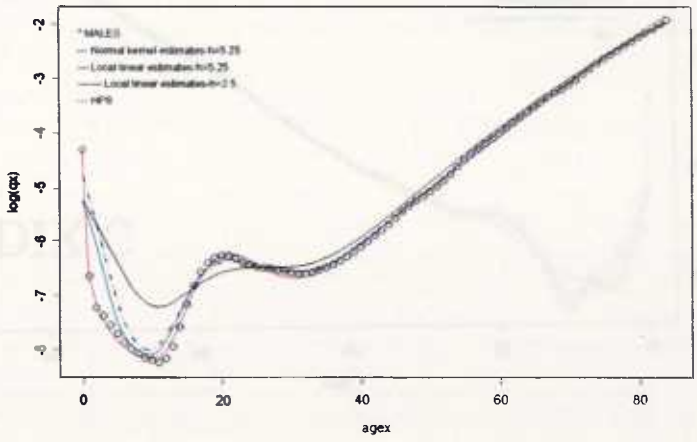


Figure B.23: Empirical q_x-values of the full age range for the male population of New-Zealand 1982 (circles) and fitted q_x-values using normal kernel estimates, local linear kernel estimates and HP8 model.

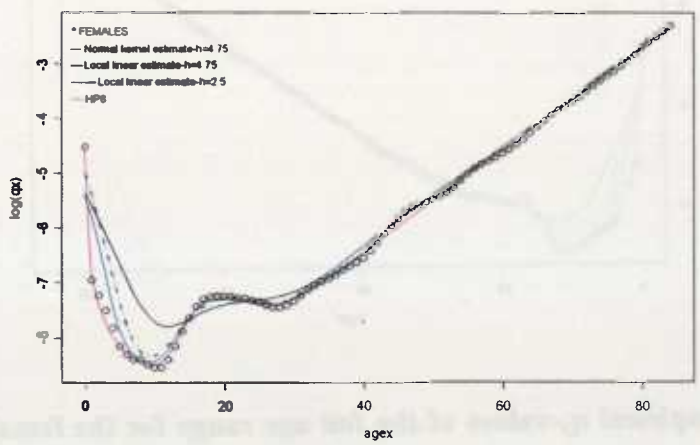


Figure B.24: Empirical q_x-values of the full age range for the female population of New-Zealand 1982 (circles) and fitted q_x-values using normal kernel estimates, local linear kernel estimates and HP8 model.

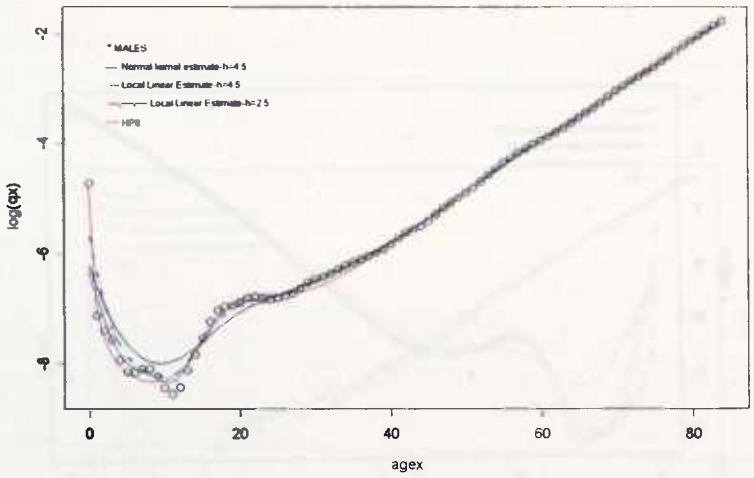


Figure B.25: Empirical q_x-values of the full age range for the male population of Germany 1988 (circles) and fitted q_x-values using normal kernel estimates, local linear kernel estimates and HP8 model.

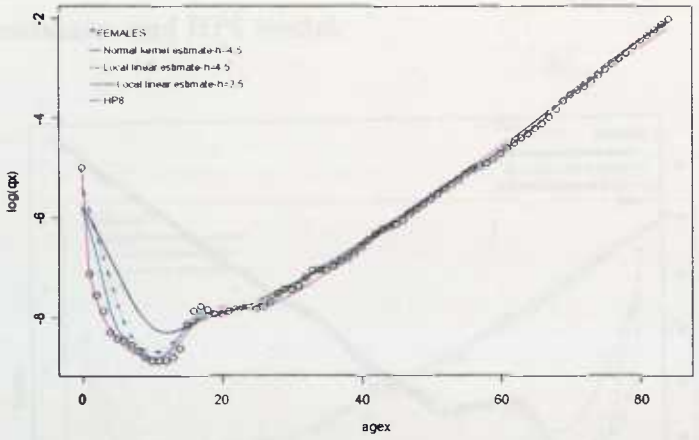
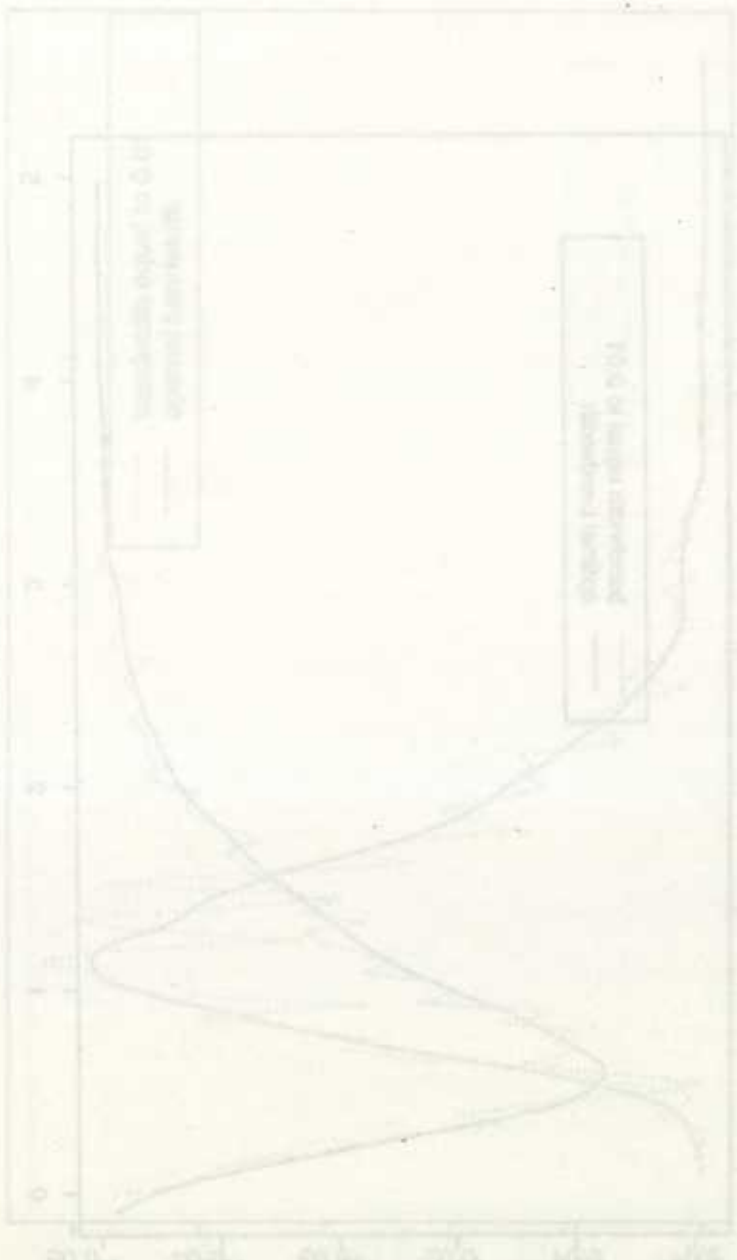


Figure B.26: Empirical q_x-values of the full age range for the female population of Germany 1988 (circles) and fitted q_x-values using normal kernel estimates, local linear kernel estimates and HP8 model.

APPENDIX C



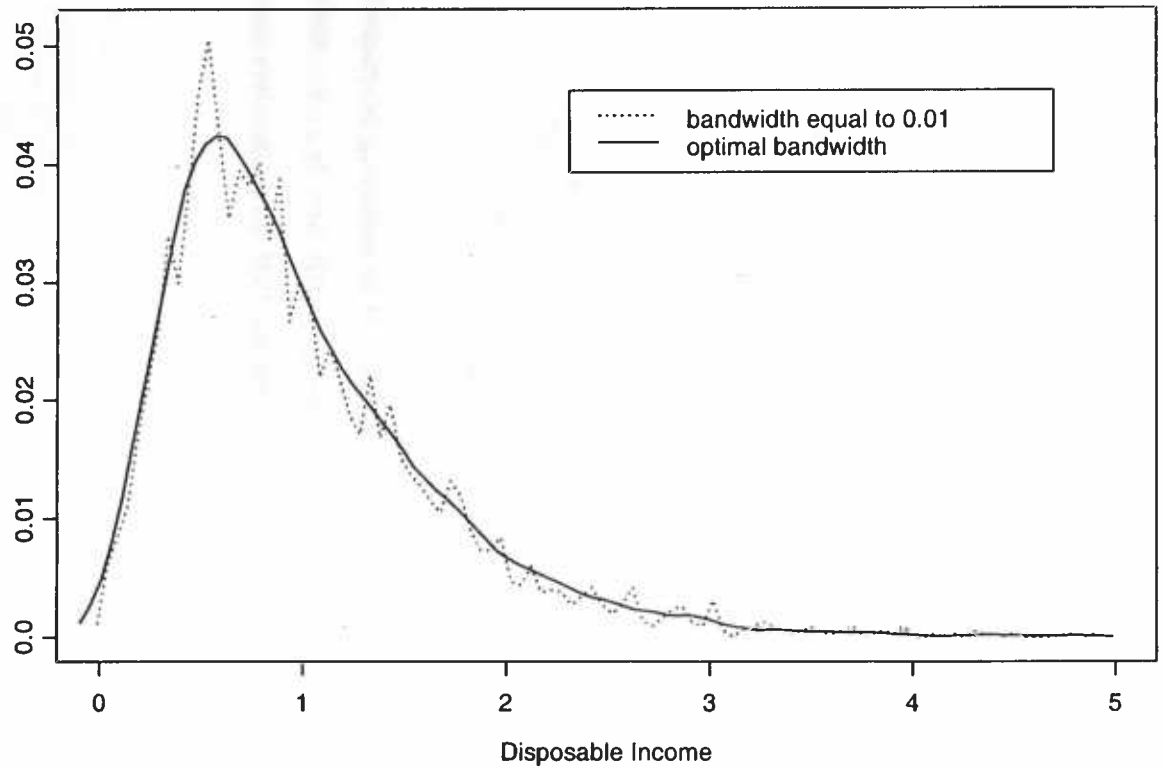


Figure C1: Disposable Income of Germany (1990) using a bandwidth equal to 0.01 and the optimal bandwidth.

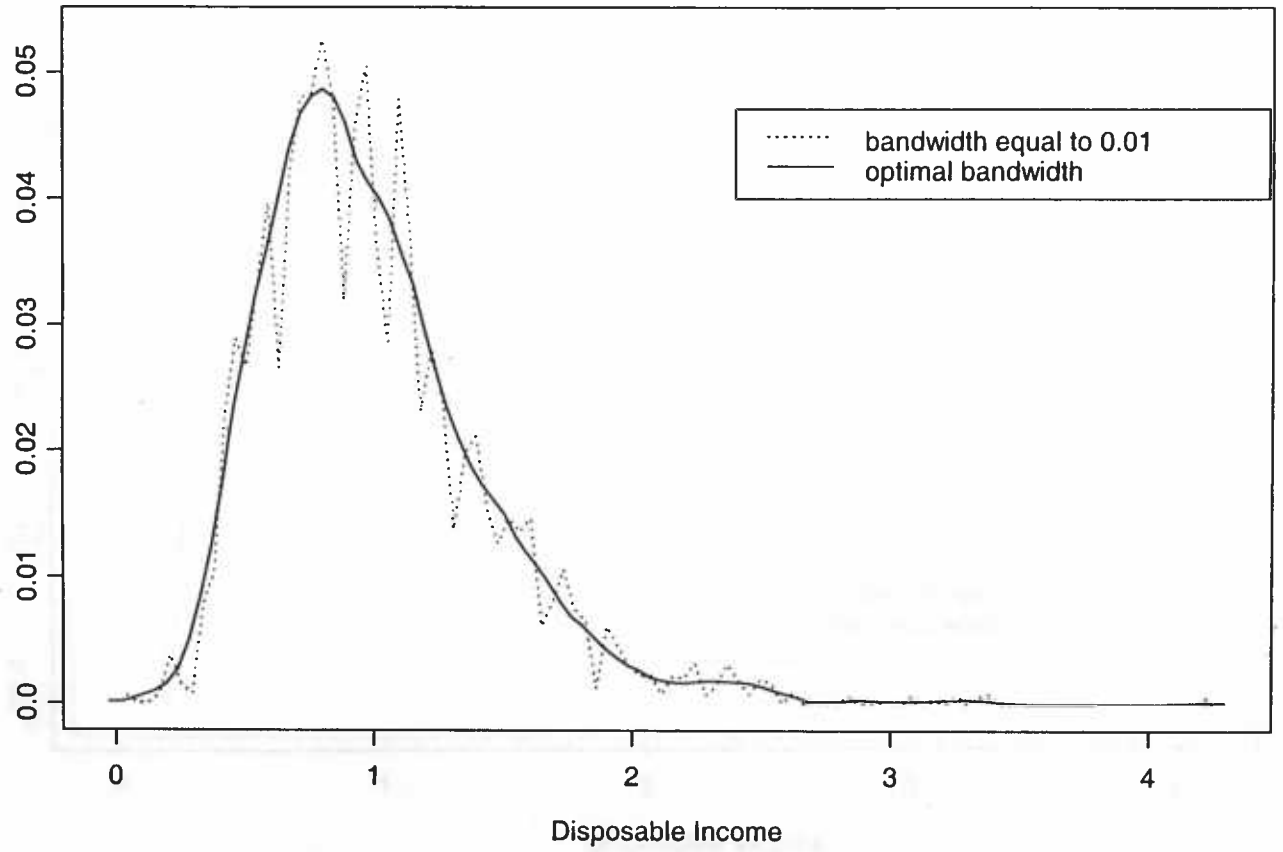


Figure C3: Disposable Income of Luxembourg (1990) using a bandwidth equal to 0.01 and the optimal bandwidth.

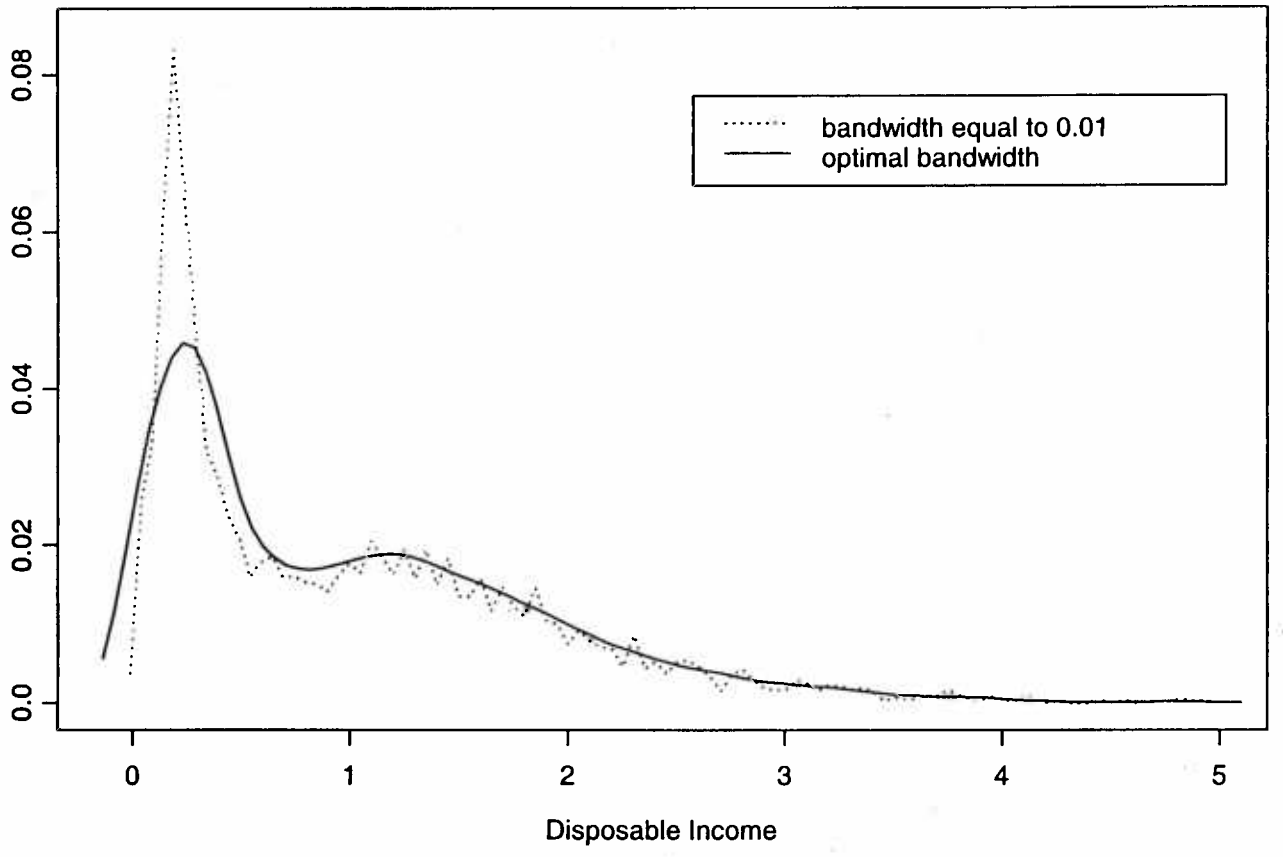


Figure C2: Disposable Income of UK (1991) using a bandwidth equal to 0.01 and the optimal bandwidth.

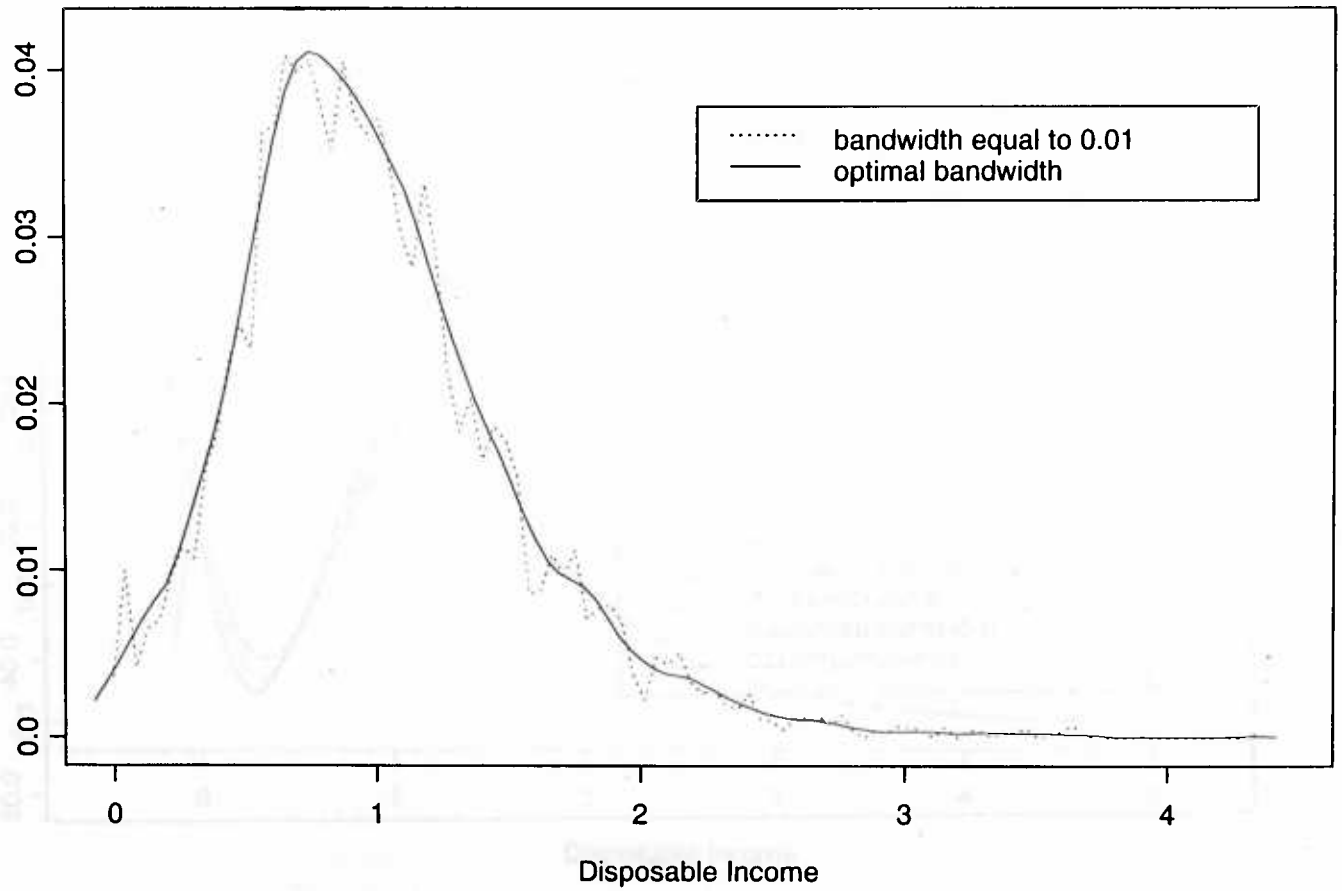


Figure C4: Disposable Income of Poland (1987) using a bandwidth equal to 0.01 and the optimal bandwidth.

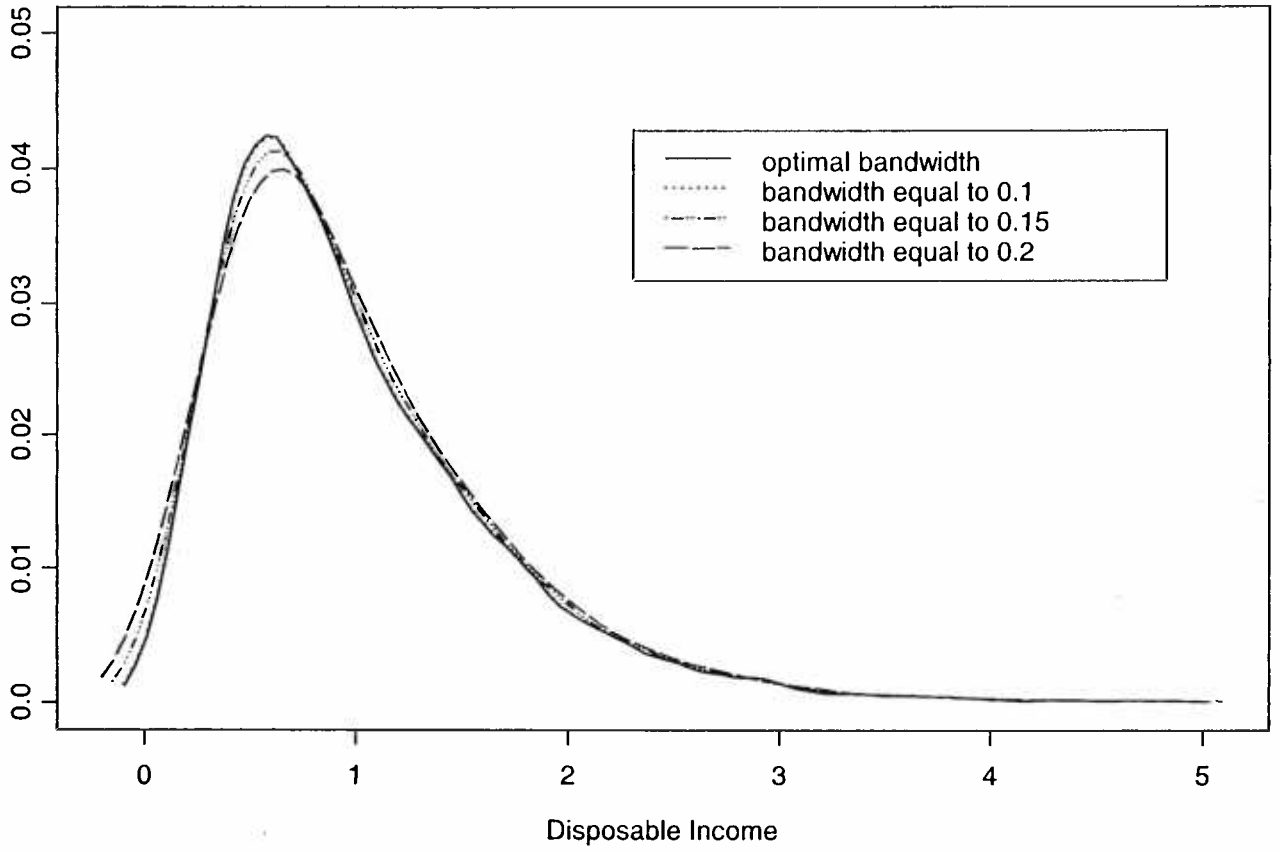


Figure C5: Disposable Income of Germany (1990) using different bandwidths.

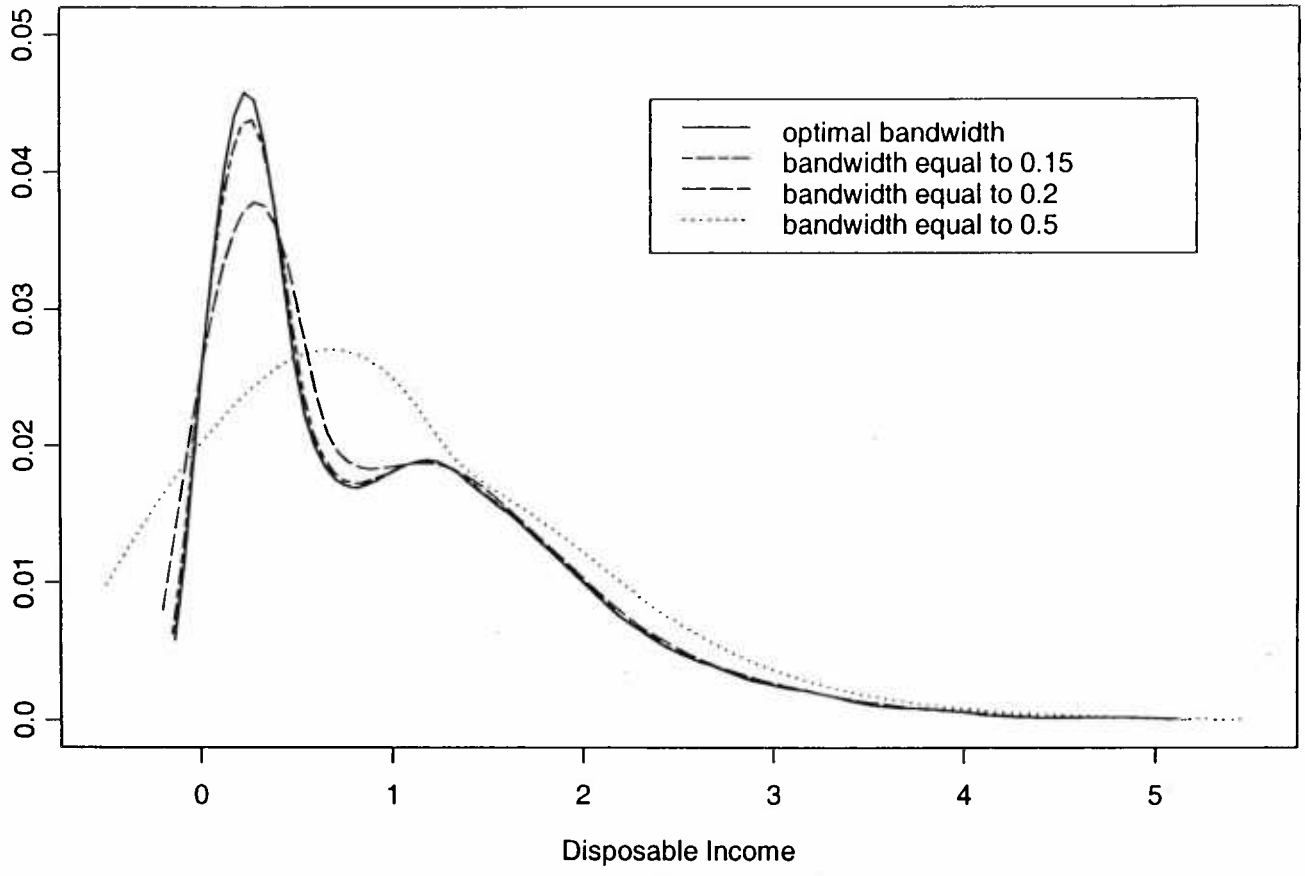


Figure C6: Disposable Income of UK (1991) using different bandwidths.

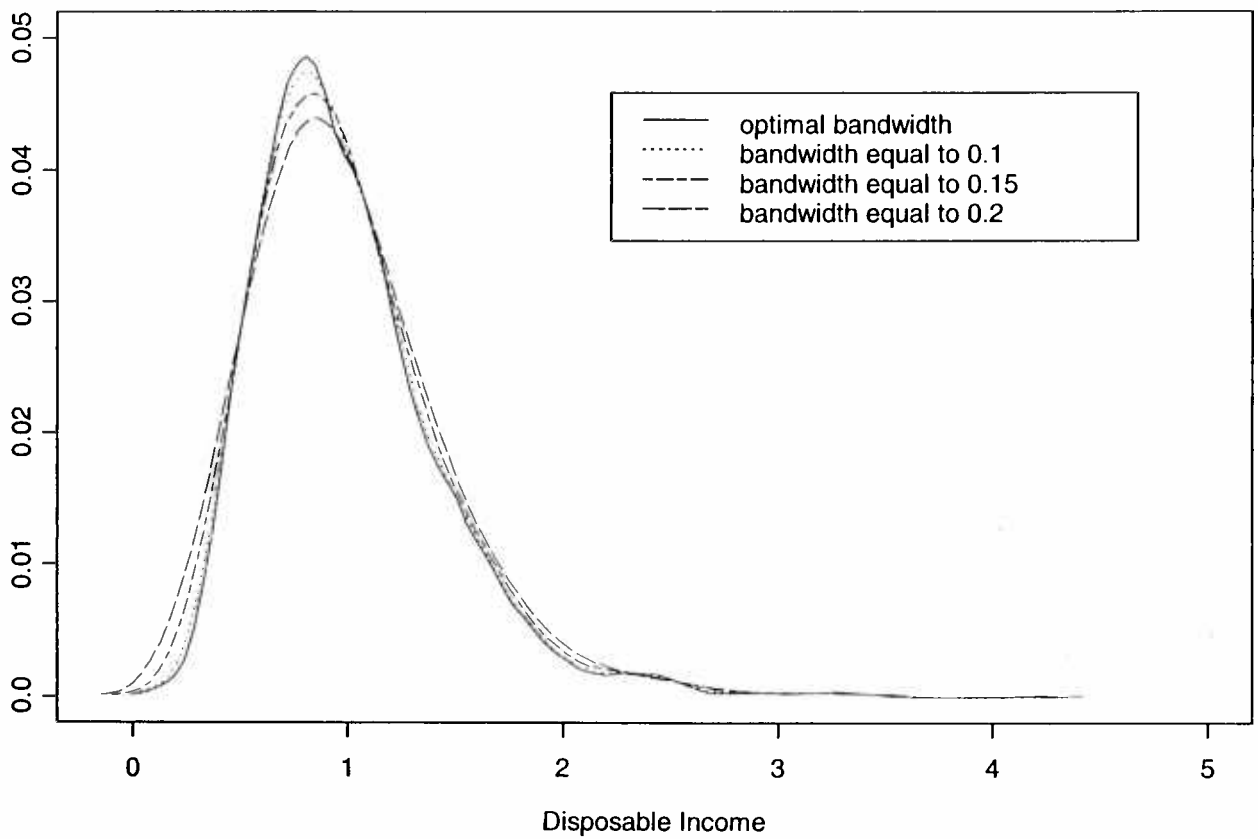


Figure C7: Disposable Income of Luxembourg (1985) using different bandwidths.

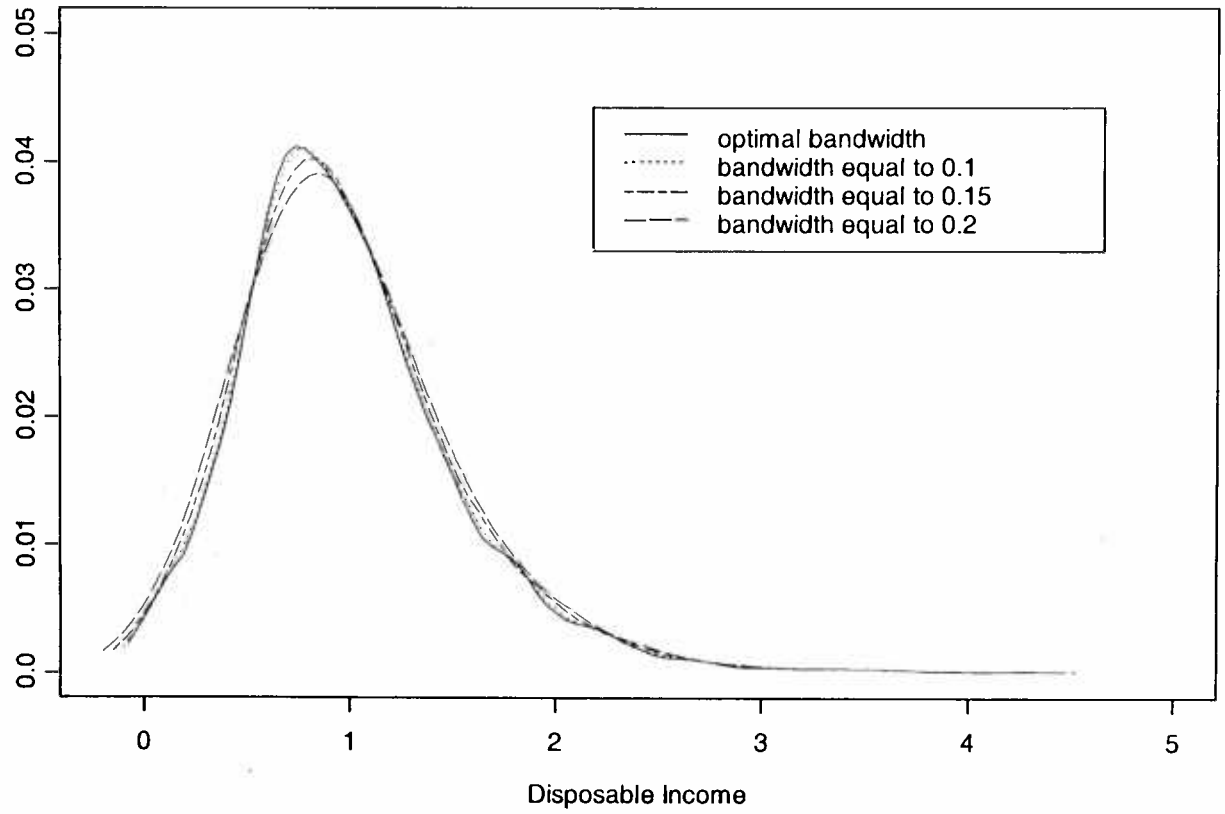


Figure C8: Disposable Income of Poland (1987) using different bandwidths.

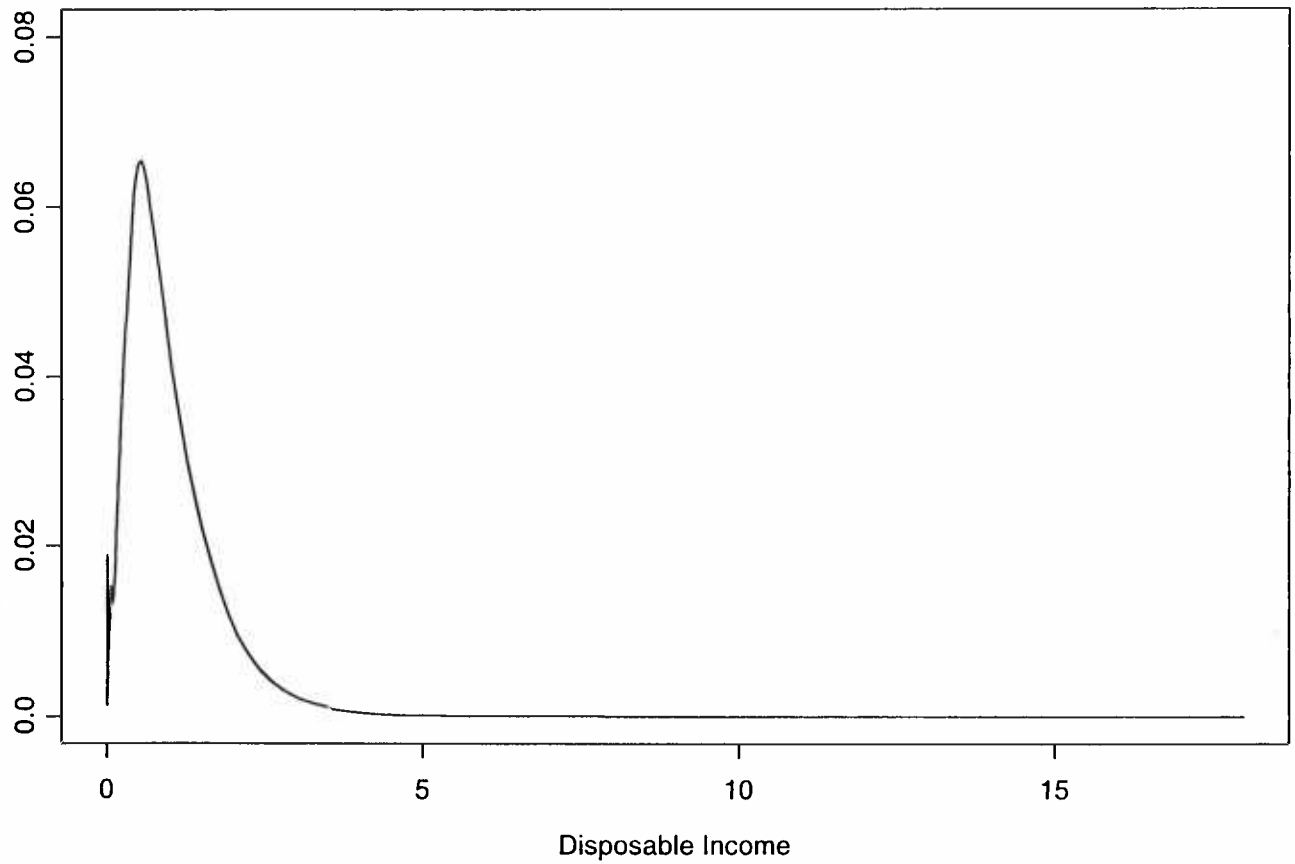


Figure C9: Disposable Income of Germany (1990), using the log transformation and a bandwidth equal to 0.15

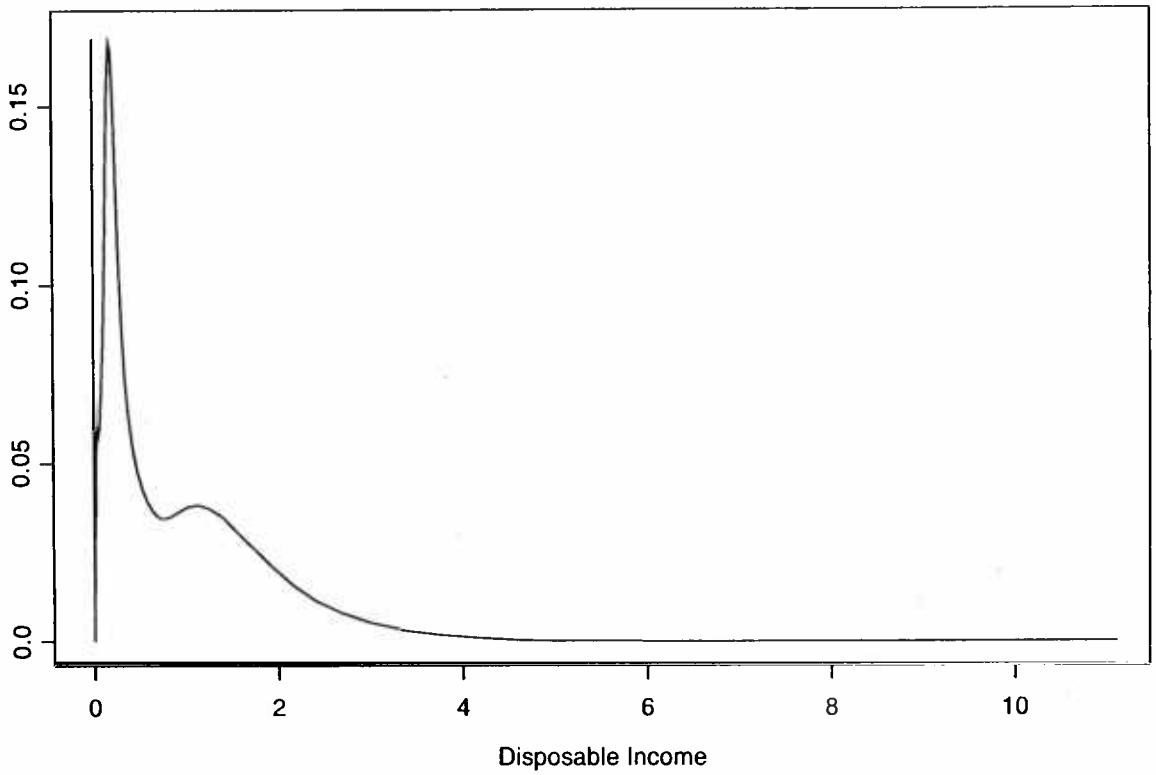


Figure C10: Disposable Income of UK (1991), using the log transformation and a bandwidth equal to 0.15.

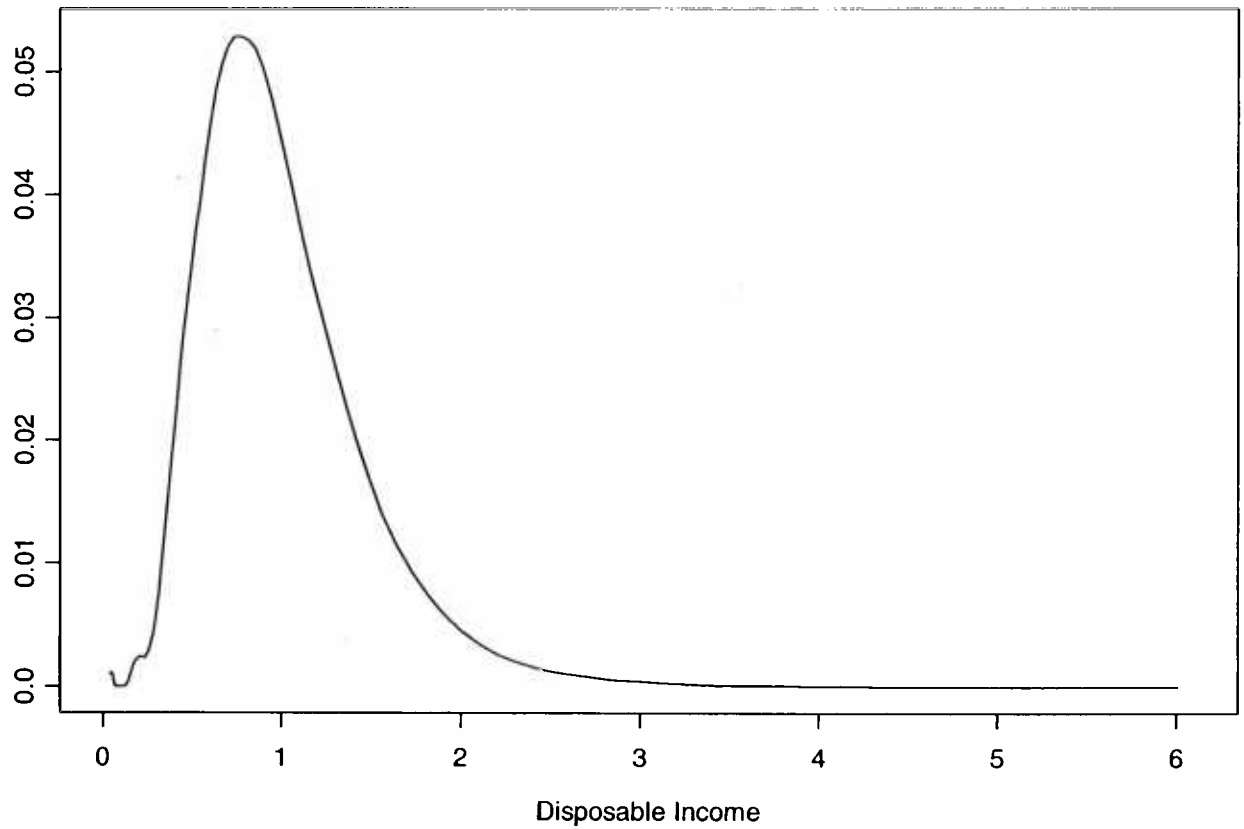


Figure C11: Disposable Income of Luxembourg (1985), using the log transformation and a bandwidth equal to 0.15.

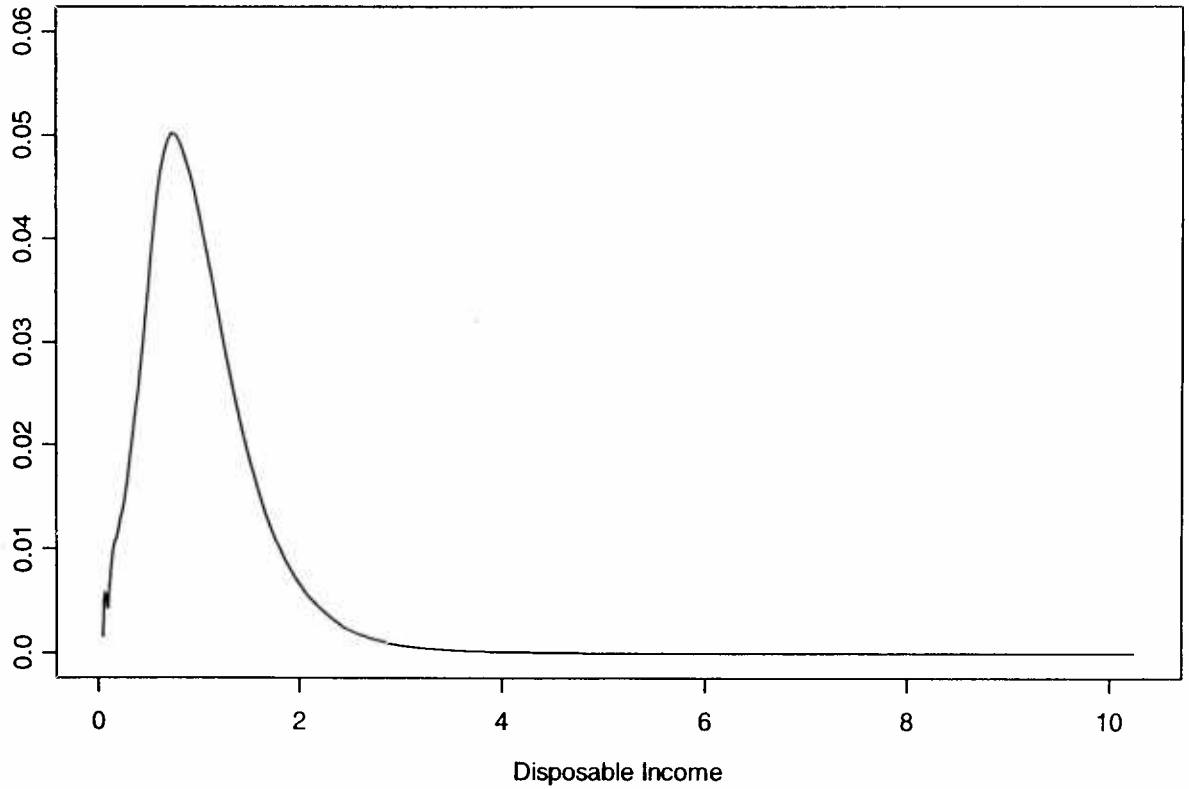


Figure C12: Disposable income of Poland (1987), using the log transformation and a bandwidth equal to 0.15.

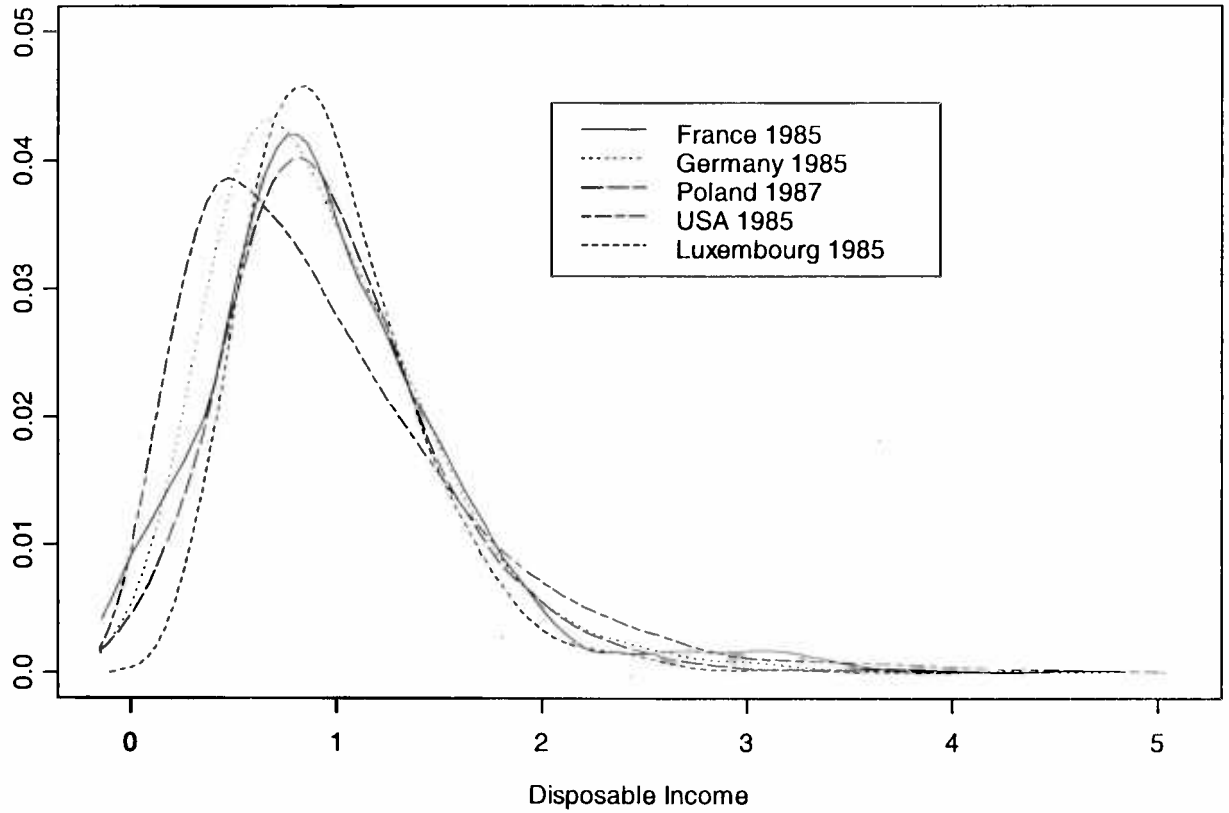


Figure C13: Disposable Income of the following countries: France (1985), Germany (1985), Poland (1987), USA (1985) and Luxembourg (1985), using a bandwidth equal to 0.15.

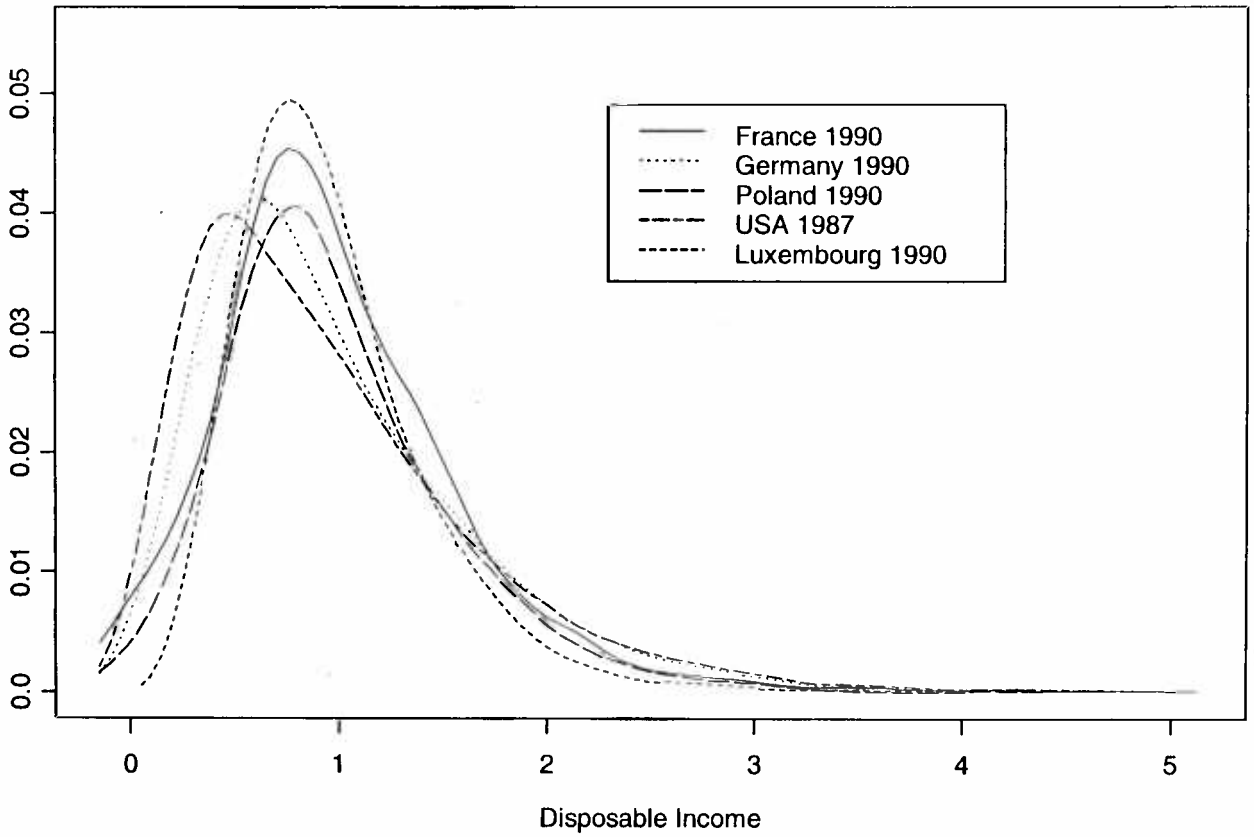


Figure C14: Disposable Income of the following countries: France (1990), Germany (1990), Poland (1990), USA (1987), Luxembourg (1990), using a bandwidth equal to 0.15.

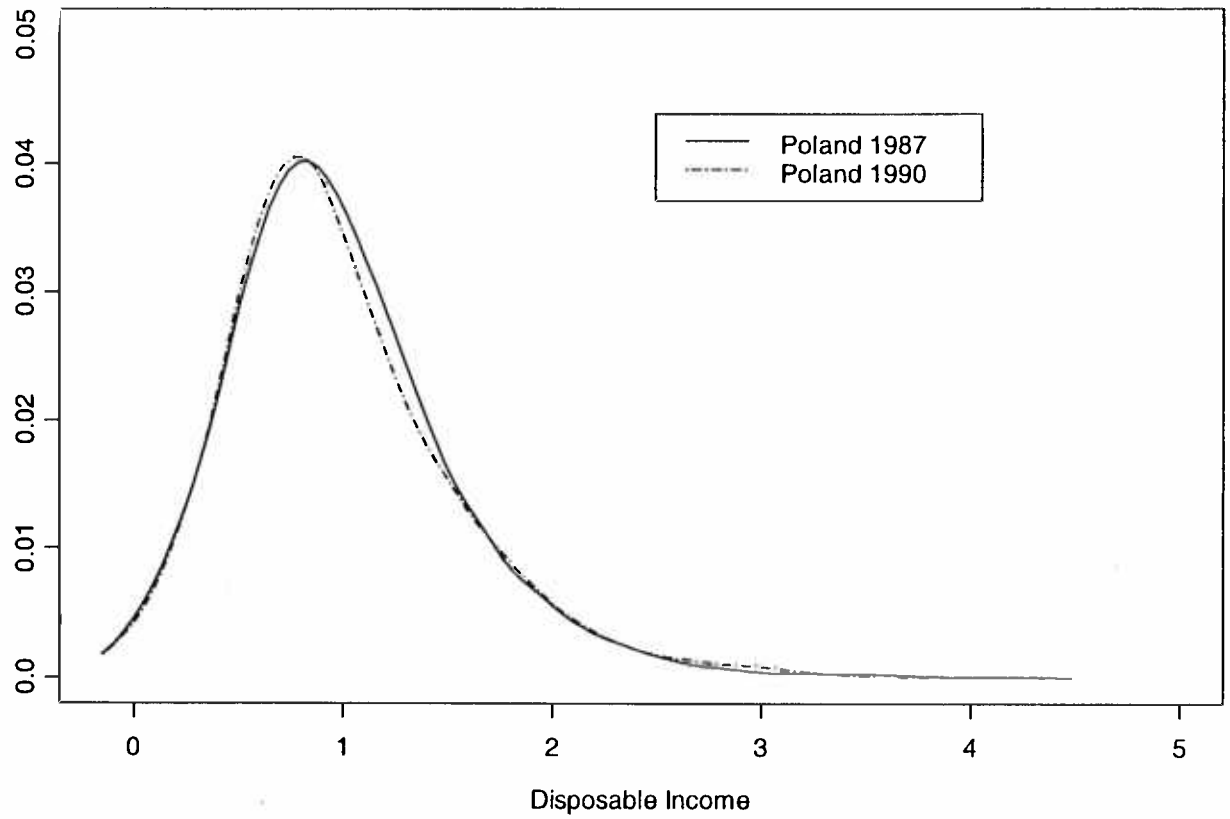


Figure C15: Disposable Income of Poland for the years 1987 and 1990, using a bandwidth equal to 0.15.

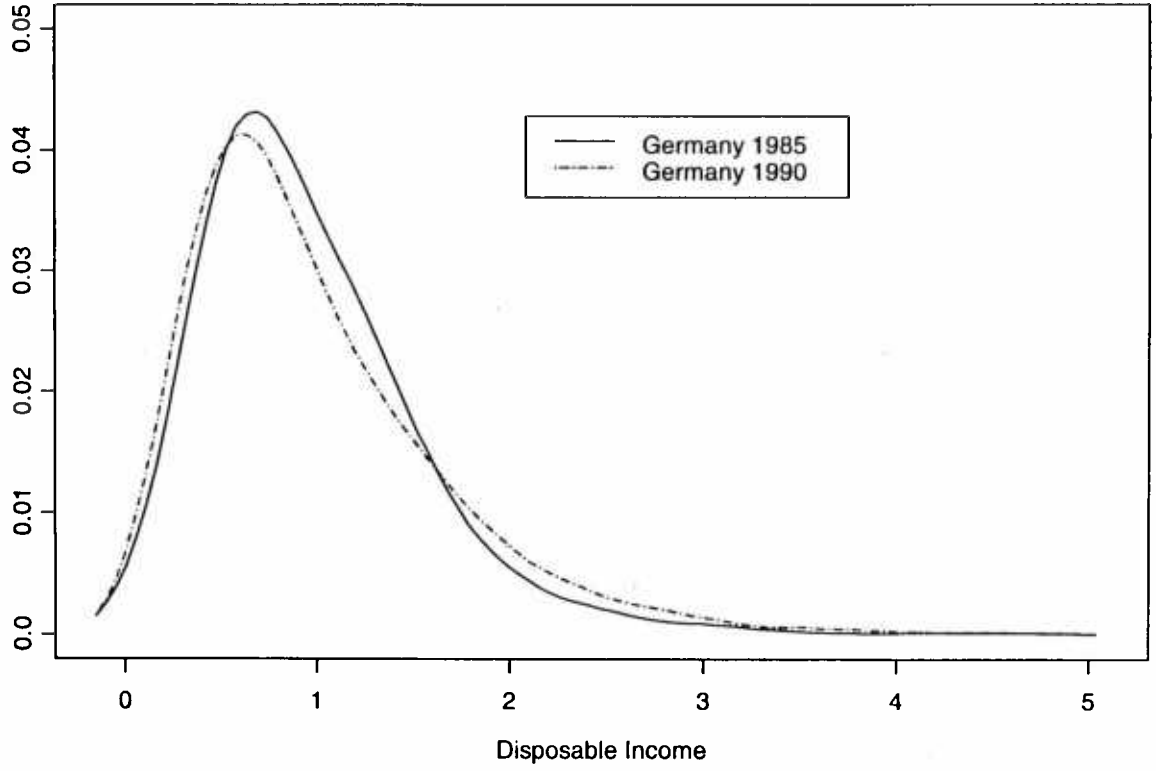


Figure C16: Disposable Income of Germany for the years 1985 and 1990, using a bandwidth equal to 0.15.

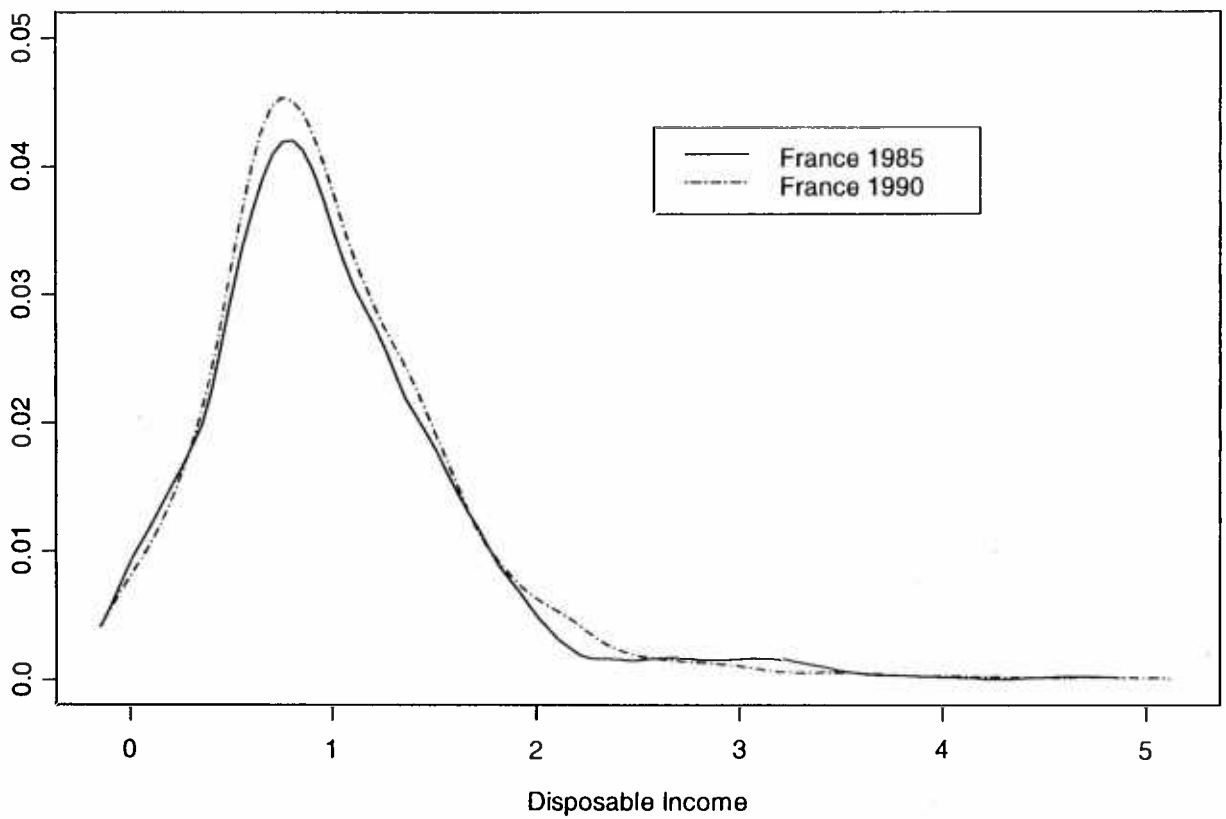


Figure C17: Disposable Income of France for the years 1985 and 1990, using a bandwidth equal to 0.15.

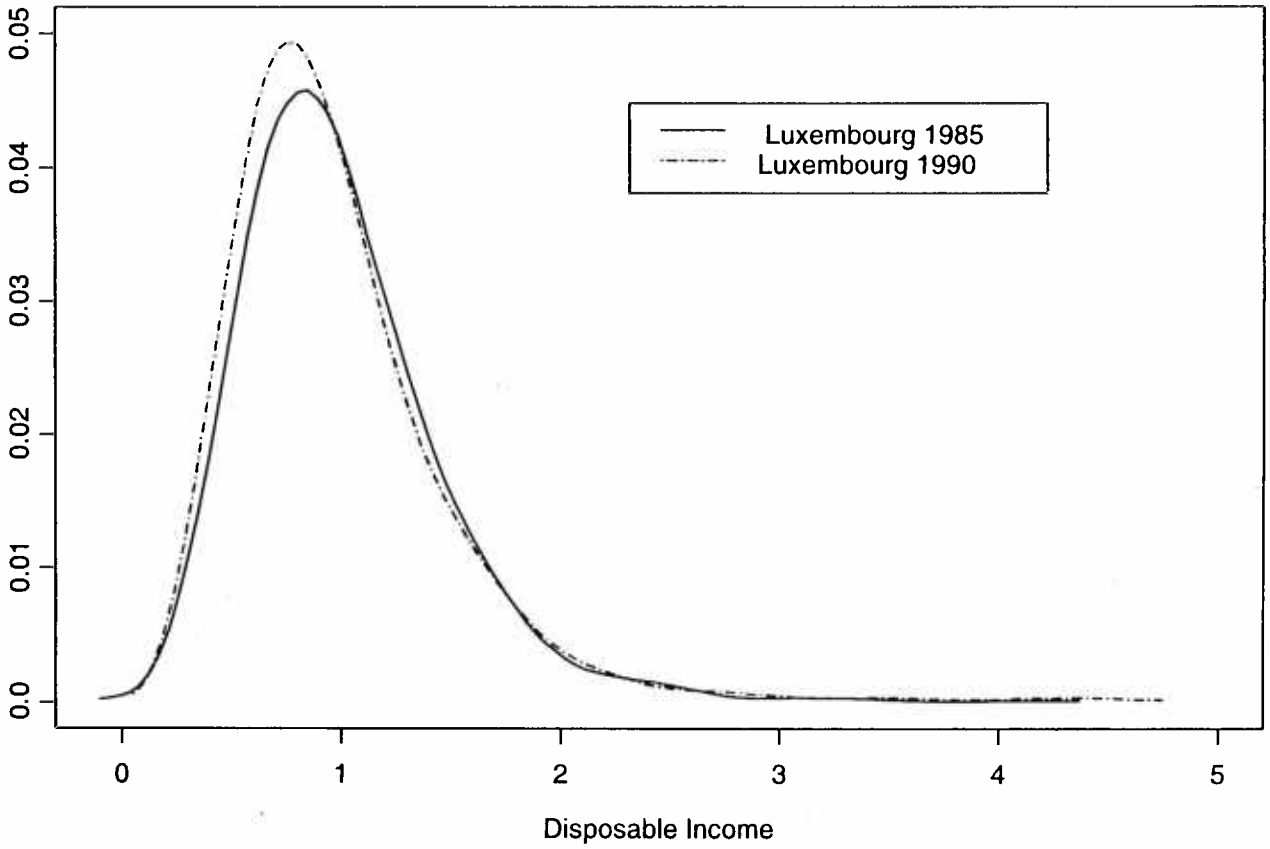


Figure C18: Disposable Income of Luxembourg for the years 1985 and 1990, using a bandwidth equal to 0.15.

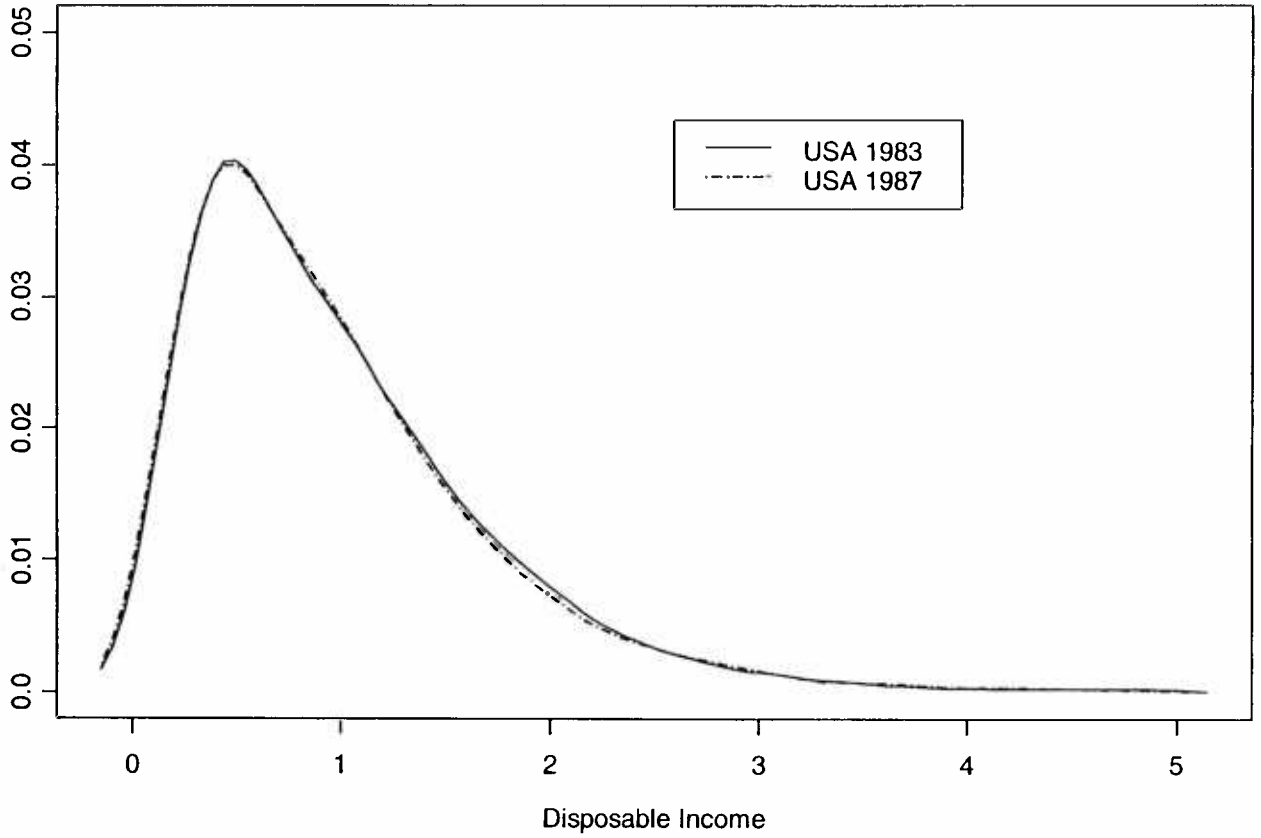


Figure C19: Disposable Income of USA for the years 1983 and 1987, using a bandwidth equal to 0.15.

REFERENCES:

Abramson, I.S. (1982). On bandwidth variation in kernel estimates-a square root law. **The Annals of Statistics**, 10, 1217-1223.

Akaike, H. (1970). Statistical predictor information. **Annals of the Institute of Statistical Mathematics**, 22, 203-17.

Akaike, H. (1974). A new look at the statistical model identification. **IEEE Transactions of Automatic Control AC**, 19, 716-23.

Barnett, H.A.R. (1985). Criteria of Smoothness. **Journal of Institute of Actuaries,** 112, 331-352.

Beltrao, K.I. and Bloomfield, P. (1987). Determining the bandwidth of a kernel spectrum estimate. **Journal of Time Series Analysis,** 8, 21-38.

Benjamin, B. and Pollard, J.H. (1980). The analysis of mortality and other actuarial statistics. London, Heinemann.

Biewen, M. (2000). Income Inequality in Germany during the 1980s and 1990s. **Review of Income and Wealth**, 46, 1, 1-19.

Bloomfield, D.S.F. and Haberman, S. (1987). Graduation: Some experiments with kernel methods. **The Journal of the Institute of Actuaries**, 114, 339-369.

Boys, R. (1992). On a kernel approach to a screening problem. **Journal of Royal Statistical Society, B,** 54, 1, 157-169.

Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates. **Biometrika,** 71, 2, 353-360.

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

Breiman, L., Meisel, W. and Purcell, E. (1977). Variable kernel estimates of multivariate densities. **Technometrics,** 19, 135-144.

Cacoulos, T. (1966). Estimation of a multivariate density. **Annals of the Institute of Statistical Mathematics,** 18, 178-189.

Copas, J. B. (1983a). Plotting p against x. **Applied Statistics,** 32, 25-31.

Copas, J. B. (1982). Regression, prediction and shrinkage. **Journal of Royal Statistical Society,** 3, 311-354.

Copas, J. B. and Haberman, S. (1983). Non-parametric graduation using kernel methods. **Journal of Institute of Actuaries,** 110, 135-156.

Cowell, F.A. Jenkins, S.P. and Litchfield, J.A. (1994). The changing shape of the UK income distribution: kernel density estimates. LSE mimeo.

Craven, P. and Wabba, G. (1979). Smoothing noisy data with spline functions. **Numerical Mathematics,** 31, 377-403

Deaton, A. (1989). Rice distribution in Thailand. **The economic journal,** supplement, 1-137.

Devroye, L. and Gyorfi, L. (1985). Nonparametric Density Estimation: The L1 view. New York, Wiley.

Eddy, D. M. (1980). Screening for Cancer: Theory, Analysis and Design. Englewood Cliffs, Prentic-Hall.

Epanechnikov, V. A. (1969). Nonparametric Estimators of a Multivariate Probability Density. **Theory of Probability and its Applications,** 14, 153-158.

Fix, E. and Hodges, J.L. (1951). Discriminatory Analysis, nonparametric estimation: consistency properties. Technical Report, Report Number 4, Project Number 21-49-004, **USAF School of Aviation Medicine**, Randolph Field, Texas.

Forfar, D.O., McCutcheon, M.A. and Wilkie, M.A. (1988). On graduation by mathematical formula. **Institute of Actuaries,** 1-149.

Gavin, J., Haberman, S. and Verrall, R. (1993). Moving Weighted average graduation using kernel estimation. **Insurance, Mathematics and Economics,** 12, 113-126.

Geisser, S. (1975). The predictive sample reuse method with applications. **Journal of American Statistical Association,** 70, 320-328.

Habbema, J.D.F., Herbmans, J. and Van der Broek, K. (1974). A stepwise discrimination program using density estimation. **In Bruckman, G. (ed.), Compstat 1974.** Vienna, Physica Verlag, 100-110.

Heligman, M.A. and Pollard, J.H (1980). The age pattern of mortality. **Institute of Actuaries**, 49-80.

Hall, P. and Marron, J.S. (1987). On the amount of noise inherent in bandwidth selection for a kernel density estimator. **The Annals of Statistics,** 15, 1, 163-181.

Hall, P. and Marron, J.S. (1987a). Extend to which Least-Squares Cross-Validation minimises Integrated Squared Error in non-parametric density estimation. **Probability Theory and Related fields**, 74, 567-581.

Hall, P. and Marron, J.S. (1987b). Estimation of Integrated Squared Density Derivatives. **Statistics and Probability Letters,** 6, 109-115.

Hall, P., Seather, S.J., Jones, M.C. and Marron J.S. (1991). On optimal data based bandwidth selection in kernel density estimation. **Biometrika,** 78, 2, 263-269.

Hall, P. and Wehrly, T.E. (1991). A geometrical method for removing edge effects from kernel type non-parametric regression estimates. **Journal of the American Statistical Association** 86, 665-72.

Hardle, W. (1989). Applied non-parametric regression. Cambridge University Press.

Hardle, W. (1991). Smoothing Techniques With Implementation in S. Springer Series In Statistics.

Hastie and Loader. (1993). Local Regression: automatic kernel carpentry. **Statistical Science**, 8(2), 120-143.

Jones, M.C. (1993). Simple Boundary Correction for Kernel Density Estimation. **Statistics and Computing**, 135-46.

Kaplan, E. and Meier, P. (1958). Non-parametric estimation from incomplete observations. **Journal of American Statistical Association**, 53, 457-81.

Loftsgaarden, D.O. and Quesenberry, C.P. (1965). A non-parametric estimate of a multivariate density function. **Annals of Mathematical Statistics,** 36, 1049-1051.

London, D. (1985). Graduation -The revision of Estimates. Winsted and Abington, Connecticut, USA:ACTEX Publications.

Mack, Y.P. and Rosenblatt, M. (1979). Multivariate K-nearest neighbor density estimates. **Journal of Multivariate Analysis,** 9, 1-15.

Madsen, R.W. (1982). A selection procedure using a screening variate. **Technometrics**, 24, 301-306.

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

McCune, S.K. and McCune, E.D. (1987). On improving convergence rates for non negative kernel failure-rate function estimators. **Statistics and Probability Letters** 6, 71-6.

McCutcheon, J.J. (1981). Some remarks on splines. **Transactions of the Faculty of Actuaries,** 37, 4, 421-438.

Park, B. U. and Marron, J. S. (1990). Comparison of Data -Driven Bandwidth Selectors. **Journal of the American Statistical Association, Theory and Methods,** 85, 409, 66-72.

Parzen, E. (1962). On the estimation of a probability density function and the mode. **Annals of Mathematical Statistics**, 40, 1056-1076.

Ramlau-Hansen, H. (1983). The choice of a kernel function in the graduation of counting process intensities. **Scandinavian Actuarial Journal**, 3, 165-182.

Ramsay, C.M. (1993). Minimum variance moving-weighted-average graduation. **Transaction of Society of Actuaries**, 43., 305-333.

Rice, J.A. (1984). Boundary modification for kernel regression. **Communications in Statistics, Theory and Methods,** 13, 893-900.

Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function. **Annals of Mathematical Statistics**, 27, 832-837.

Rudemo, M. (1982). Empirical Choice of Histogram and Kernel Density Estimators. **Scandinavian Journal of Statistics,** 9, 65-78.

Rudemo, M. (1991). Comment on "Transformations in density estimation", by M.P. Wand, J.S. Marron and D. Ruppert. **Journal of American Statistical Association**.86, 353-354.

Ruppert, D. and Cline. (1994). Bias reduction in kernel density estimation by smoothed empirical transformations. **The Annals of Statistics**, 22, 1, 185-210.

Ruppert, D. and Wand, M.P. (1992). Correcting for kurtosis in density estimation. **Australian Journal of Statistics** 34, 19-29.

Salgado-Ugarte, I.It., Sihimizu, M. and Taniuchi, T. (1993). Exploring the shape of Univariate Density using kernel density estimation, **STATA Technical Bulletin IG,** 8-19.

Schuster, E.F. and Gregory, C.G. (1981). On the nonconsistency of maximum likelihood nonparametric density estimators. In Eddy, W.F. (ed.), **Computer Science and Statistics,** Proceedings of the 13 th Symposium on the Interface. New York, Springer-Verlag, 295-298.

Schluter, C. (1996). Income distribution and inequality in Germany: evidence from panel data. Discussion paper, No. DARP 16, **London School of Economics.**

Schluter, C. (1998). Income Dynamics in Germany, the USA and the UK: Evidence from Panel data, CASE-Paper 8, **London School of Economics,** London.

Scott, D.W. and Factor, L.E. (1981). Monte Carlo study of three data-based nonparametric density estimators. **Journal of American Statistical Association,** 76, 9-15.

Scott, D. W. and Terrel, G. R. (1987). Biased and Unbiased Cross-Validation in Density Estimation. **Journal of the American Statistical Association,** 82, 1132-1146.

Scott, D. W. and Terrel, G. R. (1992). Variable kernel density estimation. **The Annals of Statistics,** 20, 3, 1236-1265.

Shibata, R. (1981). An optimal selection of regression variables. **Biometrika** 68, 45-54.

Silverman, B.W. (1980). Comment on Good and Gaskins (1980): "density estimation and bump-hunting by the penalized likelihood method, exemplified by scattering and meteorite data". **Journal of American Statistical Association**, 75, 67-68.

Silverman, B.W. (1981b). Using kernel density estimates to investigate multimodality. **Journal of Royal Statistical Society, B,** 43, 97-99.

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

Silverman, B.W. and Jones, M.C. (1989). Commentary on Fix and Hodges. **International Statistical Review,** 57, 3, 233-247.

Stone, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates. **The Annals of Statistics,** 12, 1285-1297.

Stone, M. (1974). Cross-validatory choice and assesment of statistical predictions (with discussion). **Journal of Royal Statistical Society B,** 36, 111-147.

Tapia, R.A. and Thompson, J.R. (1978). Non-parametric Probability Density Estimation. Johns Hopkins University Press, Baltimore, Maryland.

Terrel, G. (1990). The maximal smoothing principle in density estimation. **Journal of the American Statistical Association**, 85, 410, 470-477.

Terrel, G. R. and Scott, D. W. (1985). Oversmoothed Non-parametric Density Estimates. **Journal of the American Statistical Association,** 80, 209-214.

Tukey, J.W. (1977). Exploratory data analysis. Reading, MA, Addison-Welsley.

Tukey, P.A. and Tukey, J.W. (1981). Graphical display of data sets in 3 or more dimensions. In Barnett, V.(ed.), Interpreting Multivariate Data. Chilsester: Wiley, 189-275.

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

Wand, M.P., Marron, J.S. and Ruppert, D. (1991). Transformations in Density Estimation. **Journal of the American Statistical Association, Theory and methods,** 86, 414.

Watson, G.S. (1964). Smooth regression analysis. **Sankya, Series A,** 26, 359-72.

Welch, B.L. (1939). Note on discriminant functions. **Biometrika,** 31, 218-220.

Whittle, P. (1958). On smoothing of probability density functions. **Journal of the Royal Statistical Society, B,** 20, 334-343.