MATH 399 Project

Kernel Estimation for Compositional Data

Student: Florian Stern

Supervisor: Dr Kamila Zychaluk

January - June 2013

Contents

1 Introduction **1**

2 Kernel density Estimation **2**
2.1 Naive estimator or Histogram 2
2.2 Kernel estimator 3
2.3 The bandwidth h 5
2.3.1 Measures of discrepancy: mean square error and mean integrated square error 6
2.3.2 Expectation, variance and bias 8
2.3.3 The ideal width and kernel 12
2.4 Examples of Kernel efficiency 13
2.5 Kernel Estimator for non-negative data 15

3 Estimation of Compositional Data **17**
3.1 Compositional Data 17
3.2 Beta Distribution 18
3.3 Parametric Estimation: Method of moments 21
3.4 Non-Parametric Estimation: Kernel Estimation 22
3.5 Simulations 24
3.5.1 Parametric and non parametric estimates comparison . . 24
3.5.2 Real data application 25

4 Conclusion **27**

Bibliography **28**

Appendix: codes **29**

1 Introduction

In Linear Models, a model linking the observations and the covariates is assumed. The goal is often the estimation of the parameters of the model, and clearly not the estimation of the density f, which is assumed known. A completely different kind of problem is to consider the case where the density f is unknown. In this case, the interest is in estimating the density f itself, and not anymore the parameters, as no model is assumed. Hence the aim is to compute an estimate $\hat{f}$ of the true density f.

To achieve this goal, there exist essentialy two kinds of methods. On one hand, these called *parametric* methods assume some previous knowledge about the density f. On the other hand, the methods dealing with no previous knowledge about the density f, which is unknown, are known as *non-parametric* methods. The main method studied in this report, called *Kernel estimation* from [2] and [3] is a non-parametric method to obtain the estimate $\hat{f}$.

The first section describes this estimator and some of its properties. The quality of the Kernel estimator is also discussed. Indeed, some statement is made about the particular parameter that determines the estimator, and how to choose this parameter optimally. Then, it is shown that some problems can occur using the non-parametric Kernel estimator for non-negative data, and thus a correction is proposed to improve the estimator to work better for this specific type of data. In a second section, another kind of data called compositional data from [7] are considered. Again, some problems that can occur using the non-parametric Kernel estimator for these data are stated, and a correction is proposed as for the non-negative data. Also, a parametric method using beta distribution, which is called *Moment Estimator* is considered for these compositional data. Finally, both non-parametric [Kernel] and parametric [Moment] estimators for simulated and real data are compared.

2 Kernel density Estimation

The goal of this part is to study a non-parametric method for density estimation. That means defining an estimator of a density without previous knowledge about the ditribution of the data. Then a criterion of optimality for this estimator is discussed. The last part shows how to improve the estimator for data which have bounded support, more precisely that are non-negative.

In the whole report, X is assumed a random variable with density f. For $n > 0$, the univariate observations X_1,X_2,...X_n are assumed independent and with same distribution as X.

2.1 Naive estimator or Histogram

It is the first non-parametric estimator presented. It is also one of the most well-known. Figure 1 shows an histogram of *Chiken weights*, based on 72 observations from [8]. Intuitively, it can be seen as a sum of boxes centred at the observations. That means, one observation at some point x will contribute to a box, another observation at the same point x will contribute to another box, that will be added to the first one and so on.

Mathematically, the *Naive estimator*, introduced in [1], is defined by

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

where the *weight function* w is defined by

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

Thus the contribution from $X_i, i = 1, ..., n$ at point x is $\frac{1}{nh} \times \frac{1}{2}$ if $X_i, i = 1, ..., n$ belongs to interval $[x - h, x + h]$, and 0 otherwise. Note that the term $\frac{1}{nh}$ is the averaging factor of the Kernel Estimator, and the term $\frac{1}{2}$ is given by the weight function w.

Indeed, the following equivalences hold, for $i = 1, ..., n$,

$$\begin{aligned} & X_i \in [x - h, x + h] \\ \Leftrightarrow \quad & x - h < X_i < x + h \\ \Leftrightarrow \quad & -h < X_i - x < h \\ \Leftrightarrow \quad & \mid X_i - x \mid < h \\ \Leftrightarrow \quad & \mid \frac{X_i - x}{h} \mid < 1 \\ \Leftrightarrow \quad & w\left(\frac{X_i - x}{h}\right) = \frac{1}{2}. \end{aligned}$$

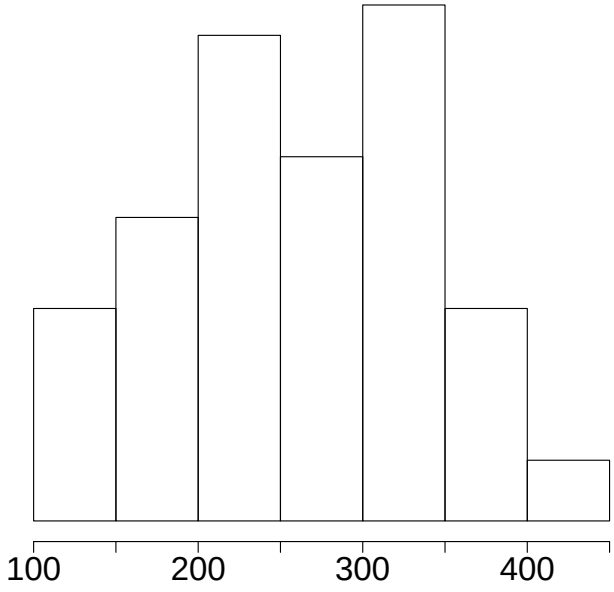

Figure 1: Histogram of chicken weights (in grammes).

Also, this shows that the parameters h has an important role because it directly determines the contribution of the obsverations and thus the shape of the estimator. Later, some further discussion is made about h.

2.2 Kernel estimator

The second non-parametric estimator presented is called *Kernel estimator* first introduced in [2] and [3]. Figure 2 shows the Kernel Estimator with $h = 0.25$ for *The duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA*, based on 272 observations from [9]. Whereas the Naive estimator which can be seen as a sum of boxes, this estimator is a smoother version in the sense that it can be seen as a sum of "bumps" placed at the observations.

The Kernel estimator may be seen as a generalization of the Naive estimator in the sense that it offers a wider choice for the weight function w. Indeed, the Kernel estimator is defined by

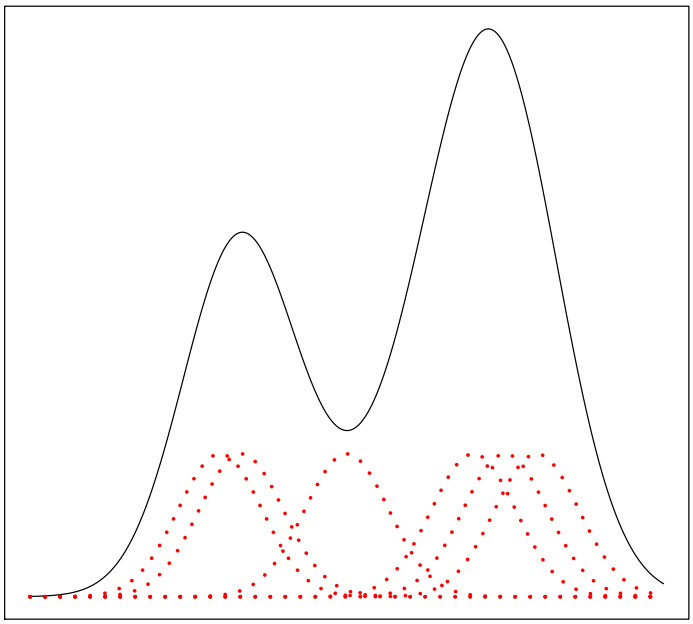

Figure 2: Kernel Estimator in black, and how it is defined by adding the red bumps.

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

where the *kernel* K is a function satisfying the following condition

$$\int_{-\infty}^{+\infty} K(x)dx = 1.$$

For example, all the probability distribution functions are integrating to unity, and thus are likely to be used as kernel K. Note also that the shape of the bumps is determined by the choice of K.

Figure 3 compares the Naive Estimator and the Kernel estimator with $h = 0.25$ for *The duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA*, based on 272 observations from [9]. Note that the

first one is erratic whereas the second is smooth. Also, the bimodal shape of the data is well estimated by both estimators.

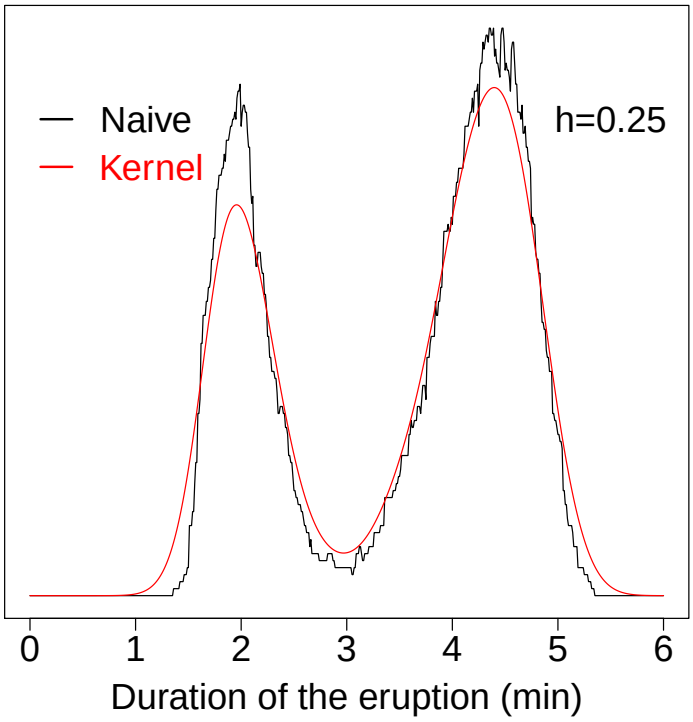

Figure 3: Naive Estimator in black versus Kernel Estimator in red for $h = 0.25$.

2.3 The bandwidth h

Note that both estimators depend on a parameter h, called the bandwidth. It determines the width of the boxes (*Naive estimator*) and the width of the bumps (*Kernel estimator*). Indeed, it is shown in Figure 4, considering again *The duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA*, based on 272 observations from [9] how h changes the shape of the estimators. Indeed, on the left panel for which $h = 0.5$, the bimodal shape of the data is well recognized by both estimators. However, on the right panel for which $h = 0.75$, the bimodal shape is still present, but less obvious at least for the Kernel estimator.

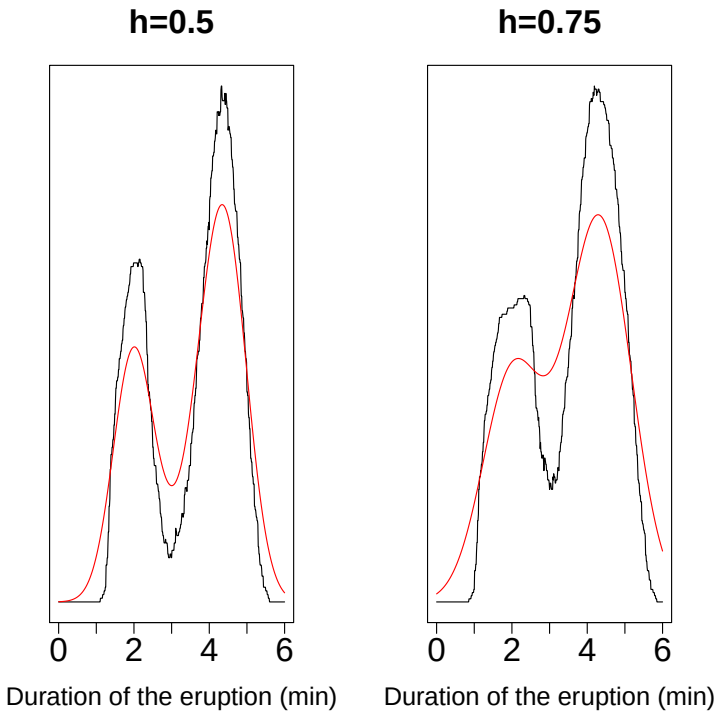

Figure 4: Naive estimator in black versus Kernel estimator in red. Both estimators are shown for $h = 0.5$ (left panel) and for $h = 0.75$ (right panel).

2.3.1 Measures of discrepancy: mean square error and mean integrated square error

In this part, two criterion of quality for the Kernel Estimator are introduced.

Definition

The *bias* of an estimator $\hat{f}$ is defined as

$$\text{bias}\hat{f}(x) = \mathbf{E}\hat{f}(x) - f(x).$$

An estimator $\hat{f}$ is unbiased if $\text{bias}\hat{f}(x) = 0$.

Definition

The *Mean Square Error*(MSE) is defined as

$$\text{MSE}_x(\hat{f}, f) = \mathbf{E}\left\{\hat{f}(x) - f(x)\right\}^2.$$

Intuitively, note that if $\hat{f}$ is a good estimator of f, then $\hat{f}$ is close to f. Hence the difference between the two of them is small, and so is the MSE. In this sense of proximity, having a small MSE is a criterion of quality. Another useful expression for the MSE is the following.

Proposition

$$\mathrm{MSE}_x(\hat{f}, f) = \left\{\mathrm{bias}\hat{f}(x)\right\}^2 + \mathbf{Var}\hat{f}(x).$$

Proof

The idea is to expand the square to make appear the bias and the variance of $\hat{f}$. The product term vanishes.

$$\begin{aligned}
\mathrm{MSE}_x(\hat{f}, f) &= \mathbf{E}\left\{\hat{f}(x) - f(x)\right\}^2 \\
&= \mathbf{E}\left\{\hat{f}(x) - \mathbf{E}\hat{f}(x) + \mathbf{E}\hat{f}(x) - f(x)\right\}^2 \\
&= \mathbf{E}\left\{\hat{f}(x) - \mathbf{E}\hat{f}(x)\right\}^2 + \left\{\mathbf{E}\hat{f}(x) - f(x)\right\}^2 + 2\mathbf{E}\left\{\left(\hat{f}(x) - \mathbf{E}\hat{f}(x)\right)\left(\mathbf{E}\hat{f}(x) - f(x)\right)\right\} \\
&= \mathbf{Var}\hat{f}(x) + \left\{\mathrm{bias}\hat{f}(x)\right\}^2 + 2\left(\mathbf{E}\hat{f}(x) - f(x)\right)\mathbf{E}\left\{\hat{f}(x) - \mathbf{E}\hat{f}(x)\right\} \\
&= \left\{\mathrm{bias}\hat{f}(x)\right\}^2 + \mathbf{Var}\hat{f}(x) + 2\left(\mathbf{E}\hat{f}(x) - f(x)\right)\left\{\mathbf{E}\hat{f}(x) - \mathbf{E}\hat{f}(x)\right\} \\
&= \left\{\mathrm{bias}\hat{f}(x)\right\}^2 + \mathbf{Var}\hat{f}(x).
\end{aligned}$$

Note that this quantity is a trade-off between the bias and the variance of the estimator $\hat{f}$. Ideally, an estimator is good if it has small bias and/or a small variance. Indeed, if the estimator of density has small bias, it means that in average it is very close to the true density. Also, if the estimator of density has small variance, it means that in average, the estimator is close to its mean. Thus a criterion of quality for $\hat{f}$ is to have small bias and/or a small variance, and equivalently a small MSE. This confirms the discussion above.

The *Mean Integrated Square Error*, MISE, is the integrated version of the MSE.

Definition

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

Similarly, a criterion of quality for $\hat{f}$ is to have a small MISE, and a similar proposition holds.

Proposition

$$\text{MISE}(\hat{f}, f) = \int \text{bias}_h(x)^2 + \int \mathbf{Var}\hat{f}(x). \tag{2}$$

Proof

The idea is to use the proposition proved below for MSE, that is it can be written as sum of squared bias and variance.

$$\begin{aligned}
\text{MISE}(\hat{f}, f) &= \mathbf{E} \int \left\{ \hat{f}(x) - f(x) \right\}^2 dx \\
&= \int \mathbf{E} \left\{ \hat{f}(x) - f(x) \right\}^2 dx \\
&= \int \text{MSE}_x(\hat{f}, f) dx \\
&= \int \text{bias}_h(x)^2 + \int \mathbf{Var}\hat{f}(x).
\end{aligned}$$

So the MISE is sum of integrated squared bias and integrated variance. Note that MSE looks at the estimator at a single point x whereas MISE is for the whole function, at all x-values.

2.3.2 Expectation, variance and bias

The aim of this part is to work out a simple version of MISE for the Kernel Estimator, or equivalentlty of its bias and its variance according to formula (2). It will turn out that the exact expressions for the bias and the variance of the Kernel Estimator are too complicated to compute, hence approximate formulas for these quantities based on Taylor expansions are stated. Finally, using these approximations and formula (2), the approximated MISE is derived.

Proposition

The expected value and the variance of the Kernel Estimate $\hat{f}$ are given by

$$\mathbf{E}\hat{f}(y) = \int \frac{1}{h} K\left(\frac{x-y}{h} \right) f(x) dx \tag{3}$$

$$\mathbf{Var}\hat{f}(y) = \frac{1}{n} \int \frac{1}{h^2} K\left(\frac{x-y}{h} \right)^2 f(x) dx - \frac{1}{n} \left(\frac{1}{h} \int K\left(\frac{x-y}{h} \right) f(x) dx \right)^2 \tag{4}$$

Proof

For expected value, the identical distribution of the X_i, $i = 1, ..., n$, is used. Indeed, this property implies that the sum over i vanishes and can be replaced by n.

$$\begin{aligned}
\mathbf{E}\hat{f}(y) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(\frac{1}{h}K\left(\frac{X_i - y}{h}\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int\frac{1}{h}K\left(\frac{x-y}{h}\right)f(x)dx \\
&= \frac{1}{n}\times n\int\frac{1}{h}K\left(\frac{x-y}{h}\right)f(x)dx \\
&= \int\frac{1}{h}K\left(\frac{x-y}{h}\right)f(x)dx
\end{aligned}$$

For variance, again, the identical distribution of the X_i is used. Also the first equality is true because the X_i, $i = 1, ..., n$ are independent, and thus the variance of a sum of random variables is equal to the sum of the variances of the random variables.

$$\begin{aligned}
\mathbf{Var}\hat{f}(y) &= \sum_{i=1}^{n}\mathbf{Var}\left(\frac{1}{n}\frac{1}{h}K\left(\frac{X_i - y}{h}\right)\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left(\mathbf{E}\left(\frac{1}{h}K\left(\frac{X_i - y}{h}\right)\right)^2 - \left(\mathbf{E}\frac{1}{h}K\left(\frac{X_i - y}{h}\right)\right)^2\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left(\int\frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(x)dx - \left(\int\frac{1}{h}K\left(\frac{x-y}{h}\right)f(x)dx\right)^2\right) \\
&= \frac{1}{n^2}\times n\left(\int\frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(x)dx - \left(\int\frac{1}{h}K\left(\frac{x-y}{h}\right)f(x)dx\right)^2\right) \\
&= \frac{1}{n}\int\frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(x)dx - \frac{1}{n}\left(\frac{1}{h}\int K\left(\frac{x-y}{h}\right)f(x)dx\right)^2.
\end{aligned}$$

From formulas (3) and (4), approximate formulas based on Taylor expansions for the bias and the variance of $\hat{f}$ can now be derived. Assume first that K is a symmetric function satisfying

$$\int K(t)dt = 1, \int tK(t)dt = 0 \,\text{and}\, \int t^2K(t)dt = k_2 \neq 0 \qquad (*)$$

Why does one set such assumptions? Recall that usually K will be a symmetric probability density, e.g. the normal density. This motivates that K integrates to unity. Also, the second hypothesis means that the random variable whose density is K has expectation zero, and the third hypothesis ensure that the

variance k_2 is non-zero.

Proposition

Assume that $(*)$ holds and the unknown density f has continuous derivatives of order 1 and 2. Thus

$$\int \text{bias}_h(x)^2 dx \approx \frac{1}{4} h^4 k_2^2 \int f''(x)^2 dx \tag{5}$$

Proof

By definition of bias and using formula (3)

$$\begin{aligned} \text{bias}_h(\hat{f}(x)) &= \mathbf{E}\hat{f}(x) - f(x) \\ &= \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y) dy - f(x) \end{aligned}$$

Make the change of variable $y = x - ht$ and use the assumption that K integrates to unity

$$\begin{aligned} \text{bias}_h(\hat{f}(x)) &= \int K(t) f(x-ht) dt - f(x) \\ &= \int K(t) \left\{ f(x-ht) - f(x) \right\} dt \end{aligned}$$

f has continuous derivatives of order 1 and 2, thus a Taylor expansion gives

$$f(x-ht) = f(x) - htf'(x) + \frac{1}{2} h^2 t^2 f''(x) + ...$$

So that, by the second assumption in $(*)$, the first term in the following equality vanishes

$$\begin{aligned} \text{bias}_h(\hat{f}(x)) &= -hf'(x) \int tK(t) dt + \frac{1}{2} h^2 f''(x) \int t^2 K(t) dt + ... \\ &= \frac{1}{2} h^2 f''(x) k_2 + \text{higher-order terms in } h \end{aligned}$$

Thus, for small h, the integrated squared bias is

$$\int \text{bias}_h(\hat{f}(x))^2 dx \approx \frac{1}{4} h^4 k_2^2 \int f''(x)^2 dx.$$

Using this approximation, the following formula for variance is derived.

Proposition

Assume that $(*)$ holds and f has continuous derivatives of order 1 and 2. Thus

$$\int \mathbf{Var}\hat{f}(x)dx \approx \frac{1}{nh}\int K(t)^2 dt \tag{6}$$

Proof

Recall that for two sequences of real numbers u_n and v_n, we say that $u_n = O(v_n)$ if an only if there exists a positive real number M and a integer n_0 such that $|u_n| \leq M|v_n|$ for $n \geq n_0$.

Using formula (4), the definition of bias, and from the previous proof that $\text{bias}_h(\hat{f}(x)) = O(h^2)$ yields

$$\begin{aligned}
\mathbf{Var}\hat{f}(x) &= \frac{1}{n}\int \frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(y)dy - \frac{1}{n}\left(\frac{1}{h}\int K\left(\frac{x-y}{h}\right)f(y)dy\right)^2 \\
&= \frac{1}{n}\int \frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(y)dy - \frac{1}{n}\left[\mathbf{E}\hat{f}(x)\right]^2 \\
&= \frac{1}{n}\int \frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(y)dy - \frac{1}{n}\left[\hat{f}(x) + \text{bias}_h(\hat{f}(x))\right]^2 \\
&\approx \frac{1}{hn}\int K(t)^2 f(x-ht)dt - \frac{1}{n}\left[\hat{f}(x) + O(h^2)\right]^2,
\end{aligned}$$

using the substitution $y = x - ht$ in the integral. Assume now that h is small and n is large. By using the third assumption in $(*)$ and expanding $f(x-ht)$ as a Taylor series, one gets

$$\begin{aligned}
\mathbf{Var}\hat{f}(x) &\approx \frac{1}{hn}\int \left[f(x) - htf'(x) + ...\right]K(t)^2 dt + O\left(\frac{1}{n}\right) \\
&= \frac{1}{hn}f(x)\int K(t)^2 dt + O(\frac{1}{n}) \\
&\approx \frac{1}{hn}f(x)\int K(t)^2 dt.
\end{aligned}$$

Recall that f is a probability density function, thus integrates to unity. Hence integrating the last formula over x yields

$$\int \mathbf{Var}\hat{f}(x)dx \approx \frac{1}{nh}\int K(t)^2 dt \tag{7}$$

2.3.3 The ideal width and kernel

Recall the general idea is to minimise the MISE. As the expression of MISE is too complicated, an easier approximate expression for MISE based on the previous paragraph is stated. Thus the minimisation can be done on this approximate MISE. It is possible to minimise it either over h or over K. Minimisation is done succesively over h, and then over K.
The formula (2) for MISE for which the approximation (5) and (6) are used, gives the following approximation of MISE, called *asymptotic MISE*

$$\text{MISE}(\hat{f}, f) \approx \frac{1}{4}h^4 k_2^2 \int f''(x)^2 dx + \frac{1}{nh}\int K(t)^2 dt \tag{8}$$

The value h_{opt}, which is the optimal value of h from the point of view of minimizing the *asymptotic MISE* satisfies

$$\frac{d\text{MISE}(\hat{f}, f)}{dh} = 0$$

which yields

$$h_{opt}^3 k_2^2 \int f''(x)^2 dx - \frac{1}{nh^2}\int K(t)^2 dt = 0$$

and finally

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

The problem with this choice of h is that it depends itself on the unknown density being estimated. Substituting the value of h_{opt} back into (8), the approximate corresponding value of the MISE is given by the following proposition.

Proposition

$$\text{MISE}(\hat{f}, f) = \frac{5}{4}C(K)\left\{\int f''(x)^2 dx\right\}^{1/5} n^{-4/5}$$

where the constant $C(K)$ is given by

$$C(K) = k_2^{2/5}\left\{\int K(t)^2 dt\right\}^{4/5} \tag{9}$$

Proof

After inserting the value of h_{opt} in (8), it consists in grouping similar terms.

$$
\begin{aligned}
\mathrm{MISE}(\hat{f}, f) &= \frac{1}{4}k_2^{-8/5}\left\{\int K(t)^2\right\}^{4/5}\left\{\int f''(x)^2dx\right\}^{-4/5}k_2^2\int f''(x)^2dxn^{-4/5} \\
&\quad +\frac{1}{n}k_2^{2/5}\left\{\int K(t)^2dt\right\}^{-1/5}\left\{\int f''(x)^2dx\right\}^{1/5}n^{1/5}\left\{\int K(t)^2dt\right\} \\
&= \frac{1}{4}k_2^{2/5}\left\{\int K(t)^2\right\}^{4/5}\left\{\int f''(x)^2dx\right\}^{1/5}n^{-4/5} \\
&\quad +n^{-4/5}k_2^{2/5}\left\{\int K(t)^2dt\right\}^{4/5}\left\{\int f''(x)^2dx\right\}^{1/5} \\
&= \frac{5}{4}C(K)\left\{\int f''(x)^2dx\right\}^{1/5}n^{-4/5}.
\end{aligned}
$$

The minimisation has been done over h, it can now be done over K. Note that the problem of minimising the asymptotic MISE then reduces to minimising $C(K)$ under the assumptions $(*)$. According to the definition of $C(K)$, this reduces again to the problem of minimizing $\int K(t)^2dt$ under the assumptions $(*)$. It is shown in [1] that this problem is solved by setting $K(t)$ to be

$$
\begin{cases}
\frac{3}{4}(1-\frac{1}{5}t^2)/\sqrt{5} & \text{if } \mid t \mid < \sqrt{5} \\
0 & \text{otherwise.}
\end{cases}
$$

which is called *Epanechnikov Kernel*, because it has been first suggested by him in density estimation.

2.4 Examples of Kernel efficiency

This part introduces a coefficient that compares different kernels according to a specific criterion based on Epanechnikov kernel.

Definition

The *efficiency* of a symmetric kernel K is defined as

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

where the constant $C(K)$ is defined in (9), and K_e denotes the Epanechnikov kernel, which can be seen as a reference-kernel.

Note that the efficiency belongs to the interval $[0, 1]$. Efficiency close to 1 is good in the sense of minimising the MISE. Table 1 shows definition of some commonly used kernels and their efficiency. Note that all values in the table are close to 1, so the choice of Kernel is not so that important.

Kernel	K(t)	Efficiency (exact and to 4 d.p)
Epanechnikov	$\begin{cases} \frac{3}{4}(1-\frac{1}{5}t^2)/\sqrt{5} & \text{if } \mid t \mid < \sqrt{5} \\ 0 & \text{otherwise.} \end{cases}$	1
Biweight	$\begin{cases} \frac{15}{16}(1-t^2)^2 & \text{if } \mid t \mid < 1 \\ 0 & \text{otherwise.} \end{cases}$	$(\frac{3087}{3125})^{1/2} \simeq 0.9939$
Triangular	$\begin{cases} 1- \mid t \mid & \text{if } \mid t \mid < 1 \\ 0 & \text{otherwise.} \end{cases}$	$(\frac{243}{250})^{1/2} \simeq 0.9859$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}t^2)$	$(\frac{36\pi}{125})^{1/2} \simeq 0.9512$
Rectangular	$\begin{cases} \frac{1}{2} & \text{if } \mid t \mid < 1 \\ 0 & \text{otherwise.} \end{cases}$	$(\frac{108}{125})^{1/2} \simeq 0.9295$

Table 1: Some Kernels and their efficiency, from [6].

2.5 Kernel Estimator for non-negative data

It is very often the case that the natural domain of definition of a density to be estimated is not the whole real line but an interval bounded on one side. For example, when data represent the ages of people which are positive quantities, estimates which give any weight to the negative number are likely to be unacceptable.

One possible way of ensuring that $\hat{f}(x)$ is zero for negative x is simply to calculate the estimate for positive x ignoring the boundary conditions, and then to set $\hat{f}(x)$ to zero for negative x. A drawback of this approach is that if we use a method, for example the kernel method, which usually produces estimates which are probability densities, the estimates obtained will no longer integrate to unity. To make matters worse, the contribution to $\int_0^\infty \hat{f}(x)$ of points near zero will be much less than that of points well away from the boundary, and so, even if the estimate is rescaled to make it a probability density, the weight of the distribution near zero will be underestimated.

Let's introduce the following weight function w

$$w(X_i, y) = K\left(\frac{x - X_i}{h}\right) - K\left(\frac{x + X_i}{h}\right)$$

The associated estimate, called *Kernel adapted* from [4] yields

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left[K\left(\frac{x - X_i}{h}\right) - K\left(\frac{x + X_i}{h}\right)\right] \tag{10}$$

Assume that the kernel K is symmetric and differentiable (so that $\hat{f}(x)$ also shares this property). The advantage of this estimate is that it deals well with the boundary 0 by having the following property $\hat{f}(0) = 0$.

Figure (5) shows the true density for Gamma(3,2) distribution and two estimators, the Kernel estimator and the Kernel estimator adapted for which $h = 0.25$ based on 100 observations from a simulated Gamma(3,2) distribution. Clearly, the estimator adapted is better than the classical Kernel estimator at the boundary 0, in the sense that it is zero for non-positive data and thus is closer to the true distribution, whereas this is not achieved by the Kernel estimator.

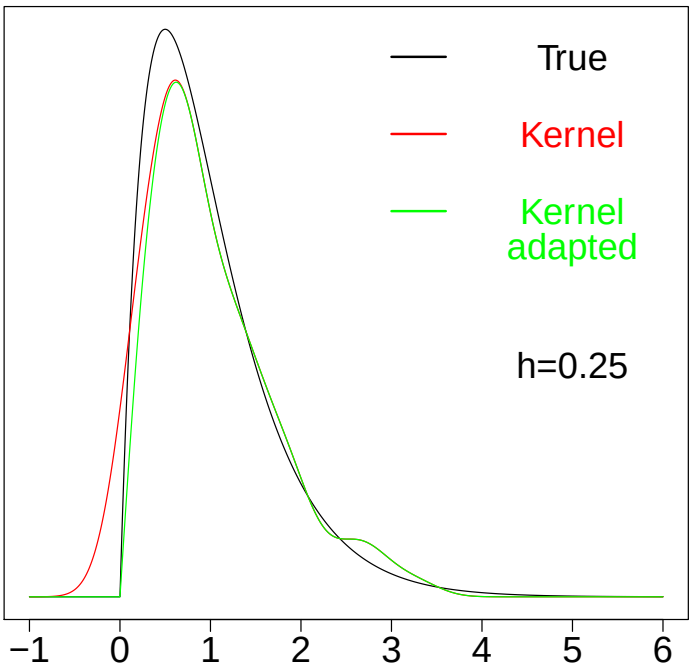

Figure 5: True density in black, Kernel estimator in red and Kernel estimator adapted to 0 boundary in green by formula (10). $h = 0.25$ for both estimators.

3 Estimation of Compositional Data

In this section, the concept of *Compositional Data* is introduced. Then the reader is reminded some points about Beta distribution, which is needed to introduce a parametric method called *Moment Estimator* for compositional data. Finally a non-parametric method based on Kernel Estimator for compositional data is also explained, and some simulations compare the two methods.

3.1 Compositional Data

Any vector $\mathbf{x}$ with non-negative elements x_1,...,x_D representing proportions of some whole is subject to the obvious constraint:

$$x_1 + ... + x_D = 1.$$

Compositional data, consisting of such vectors of proportions, play an important role in many disciplines. For example, in geology, it can represent the composition of a rock in terms of major-oxides. Such data naturally lead us to define a particular sample space.

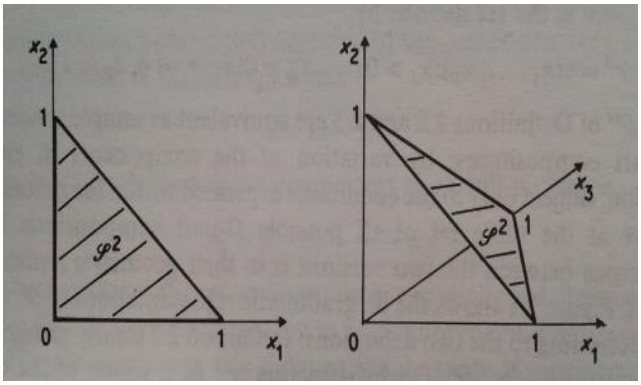

Figure 6: Simplex for $D = 2$ from [7].

Definition

The d-dimensional *simplex* is the set defined by

$$S^d = \{(x_1, ..., x_d) : x_1 > 0, ..., x_d > 0; x_1 + ... + x_d < 1\}.$$

where $d = D - 1$.

Equivalently, the *simplex* can be defined as

$$S^d = \{(x_1, ..., x_D) : x_1 > 0, ..., x_D > 0; x_1 + ... + x_D = 1\}.$$

Figure (6) shows two possible representations of the simplex for which $D = 2$ where the striped area represents the possible values of the components of x.

Thus compositional data "live" in such simplex. Note that the proportions considered are between 0 and 1, thus to model such data, a density with support [0,1] would be useful. This motivates the introduction in the next section of the beta-distribution, which is defined over the interval [0,1].

3.2 Beta Distribution

Definition

The *Beta function* with parameters p and q is the function defined by

$$B(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1}dt$$

for $p > 0$ and $q > 0$.

Definition

The *Gamma function* with parameter p is the function defined by

$$\Gamma(p) = \int_0^\infty t^{p-1}e^{-t}dt.$$

It can be shown that for $p > 0$, $\Gamma(p) = (p-1)!$ and $B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$.

Definition

A random variable Y is *beta-distributed* over the interval [0,1] with indices p and q, written Be(p,q), if it has density function

$$f_Y(y) = \frac{y^{p-1}(1-y)^{q-1}}{\mathbf{B}(p, q)} \quad (0 \leq y \leq 1).$$

Note that the quantity $B(p, q)$ is a normalizing constant, as it allows the density to integrate to unity. Figure (7) shows various true Beta distribution for different choice of parameters p and q. For identical p and q, the distribution is symmetrical, which can be seen for example on the top left panel for which

$p = q = 0.5$, the top right panel for which $p = q = 2$ or the bottom right panel for which $p = q = 1$. However, when p is different from q, as in the bottom left panel for which $p = 5$ and $q = 2$, the distribution is asymmetrical. The asymmetry property is as pronounced as big the difference between p and q is. Note that the right-bottom panel for $p = q = 1$ is the Uniform Distribution over [0,1].

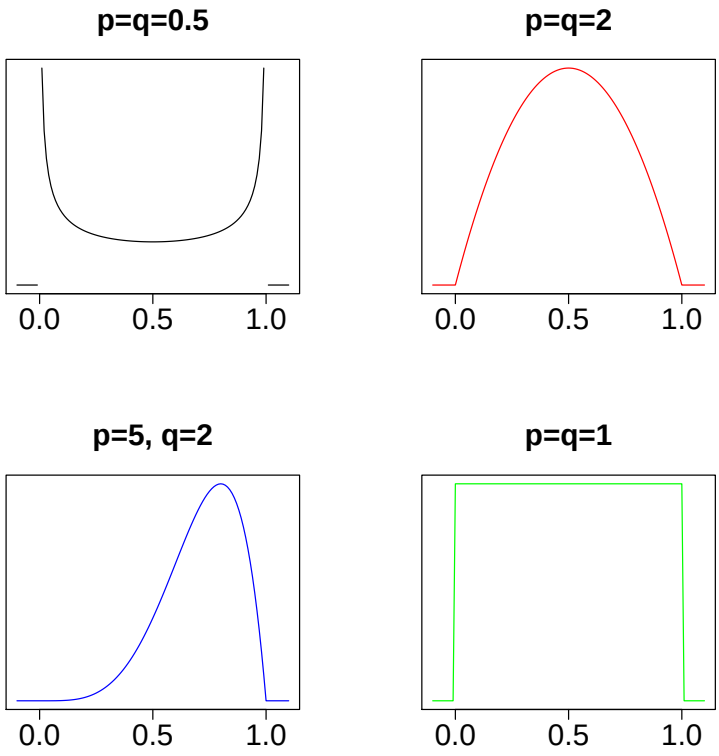

Figure 7: Various true Beta distribution for different parameters.

Proposition

Assume that Y has distribution $B(p, q)$ on interval $[0, 1]$. Thus

$$\mathbf{E}(Y) = \frac{p}{p+q} \tag{11}$$

and

$$\text{Var}(Y) = \frac{pq}{(p+q)^2(p+q+1)} \tag{12}$$

Proof

Using that for $p > 0$, $\Gamma(p) = (p-1)!$ and $B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$, the moment of order 1 and 2 are respectively given by

$$\begin{aligned}
\mathbf{E}(Y) &= \int_0^1 y f(y) dy \\
&= \int_0^1 y \frac{y^{p-1}(1-y)^{q-1}}{\mathbf{B}(p,q)} dy \\
&= \frac{1}{\mathbf{B}(p,q)} \int_0^1 y^p (1-y)^{q-1} dy \\
&= \frac{\mathbf{B}(p+1,q)}{\mathbf{B}(p,q)} \\
&= \frac{\mathbf{\Gamma}(p+1)\mathbf{\Gamma}(q)}{\mathbf{\Gamma}(p+1+q)} \frac{\mathbf{\Gamma}(p+q)}{\mathbf{\Gamma}(p)\mathbf{\Gamma}(q)} \\
&= \frac{p\mathbf{\Gamma}(p)}{(p+q)\mathbf{\Gamma}(p+q)} \frac{\mathbf{\Gamma}(p+q)}{\mathbf{\Gamma}(p)} \\
&= \frac{p}{p+q},
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{E}(Y^2) &= \int_0^1 y^2 f(y) dy \\
&= \int_0^1 y^2 \frac{y^{p-1}(1-y)^{q-1}}{\mathbf{B}(p,q)} dy \\
&= \frac{1}{\mathbf{B}(p,q)} \int_0^1 y^{p+1} (1-y)^{q-1} dy \\
&= \frac{\mathbf{B}(p+2,q)}{\mathbf{B}(p,q)} \\
&= \frac{\mathbf{\Gamma}(p+2)\mathbf{\Gamma}(q)}{\mathbf{\Gamma}(p+2+q)} \frac{\mathbf{\Gamma}(p+q)}{\mathbf{\Gamma}(p)\mathbf{\Gamma}(q)} \\
&= \frac{p(p+1)\mathbf{\Gamma}(p)}{(p+q)(p+q+1)\mathbf{\Gamma}(p+q)} \frac{\mathbf{\Gamma}(p+q)}{\mathbf{\Gamma}(p)} \\
&= \frac{p(p+1)}{(p+q)(p+q+1)}.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathbf{E}(Y^2) - \{\mathbf{E}(Y)\}^2 \\
&= \frac{p(p+1)}{(p+q)(p+q+1)} - \frac{p^2}{(p+q)^2} \\
&= \frac{p(p+1)(p+q) - p^2(p+q+1)}{(p+q)^2(p+q+1)} \\
&= \frac{pq}{(p+q)(p+q+1)}.
\end{aligned}$$

3.3 Parametric Estimation: Method of moments

Denote $\bar{x} = \widehat{E(X)} = \frac{1}{n}\sum_i^n x_i$ which is called *empirical mean* and $\bar{v} = \widehat{Var(X)} = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2$ which is called *empirical variance*. The method of moments is a method which gives estimates for $\hat{p}$ and $\hat{q}$. It is based on the moments of the variable X. More precisely, the idea is to estimate $\mathbf{E}(X)$ by $\bar{x}$ and $\mathrm{Var}(X)$ by $\bar{v}$, and then deduce the expression $\hat{p}$ and $\hat{q}$.

Proposition

Assume $\bar{v} < \bar{x}(1-\bar{x})$. For Beta distribution, the estimator of moment of $\hat{p}$ and $\hat{q}$ are given by

$$\hat{p} = \bar{x}\left[\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1\right]$$

and

$$\hat{q} = (1-\bar{x})\left[\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1\right].$$

Proof

From (11) and (12), and by replacing respectively $\mathbf{E}$(Y) and Var(Y) by their empirical version $\bar{x}$ and $\bar{v}$ yields

$$\bar{x} = \frac{\hat{p}}{\hat{p}+\hat{q}}$$

and

$$\bar{v} = \frac{\hat{p}\hat{q}}{(\hat{p}+\hat{q})^2(\hat{p}+\hat{q}+1)}.$$

Rearranging these expressions implies

$$\bar{x}(\hat{p}+\hat{q}) = \hat{p} \tag{13}$$

and

$$\bar{v}(\hat{p}+\hat{q})^2(\hat{p}+\hat{q}+1) = \hat{p}\hat{q} \tag{14}$$

Let $c = \frac{1}{\bar{x}}$, then from equation (13) we have $\hat{p}+\hat{q} = c\hat{p}$ and $\hat{q} = \hat{p}(c-1)$

Insert back in (14) yields

$$\bar{v}(c\hat{p})^2(c\hat{p}+1) = \hat{p}\hat{p}(c-1).$$

Finally

$$\hat{p} = \frac{1}{c}\left[\frac{(c-1)}{\bar{v}c^2} - 1\right] = \bar{x}\left[\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1\right]$$

and

$$\hat{q} = \hat{p}(c-1) = (1-\bar{x})\left[\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1\right].$$

3.4 Non-Parametric Estimation: Kernel Estimation

All the remarks of the section (2.5) can be extended to the case where the required support of the estimator is a finite interval [0,1]. Transformation methods from [5] can be based on transformations of the form

$$Y_i = H^{-1}(X_i) \tag{15}$$

for $i = 1, ..., n$, where H is any cumulative probability distribution function stricly increasing on $(-\infty, \infty)$.

Proposition

Assume X has density f_X, Y has density f_Y. Let F_X and F_Y be respectively the cumulative distribution function of X and Y. Thus

$$\hat{f}_X(x) = \hat{f}_Y(H^{-1}(x))\frac{1}{H'(H^{-1}(x))}$$

where $y = H^{-1}(x)$.

Proof

First

$$\begin{aligned} F_X(x) &= P(X_i \leq x) \\ &= P(H(Y_i) \leq x) \\ &= P(Y_i \leq H^{-1}(x)) \\ &= F_Y(H^{-1}(x)) \end{aligned}$$

Hence, by the derivative of the inverse of composed function,

$$\begin{aligned} f_X(x) &= F'_X(x) \\ &= F'_Y(H^{-1}(x)))(H^{-1}(x)))' \\ &= f_Y(H^{-1}(x)))(H^{-1}(x)))' \\ &= f_Y(H^{-1}(x)))\frac{1}{H'(H^{-1}(x))}. \end{aligned}$$

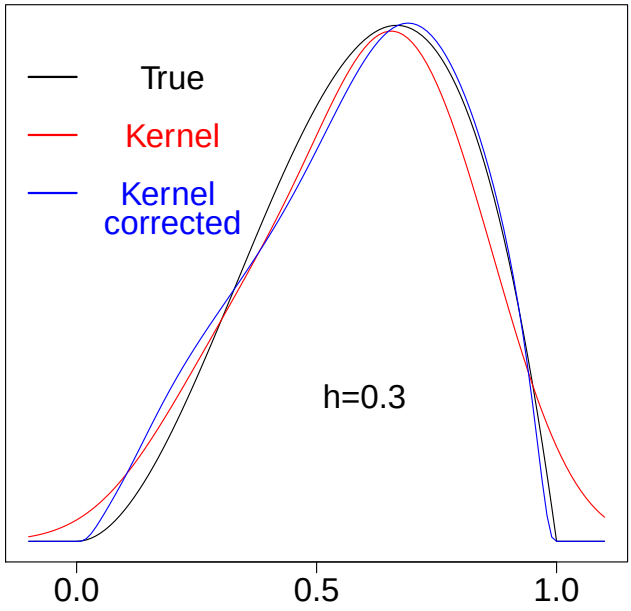

Figure 8: True density in black, Kernel estimator in red and Kernel estimator corrected by formula (16) in blue. $h = 0.3$ for both estimators.

Figure 8 shows the true density for Beta(3,2) distribution on interval [0,1] and two estimators, the Kernel estimator and the Kernel estimator corrected from formula (16) for which $h = 0.3$ based on 100 observations from a simulated Beta(2,2) distribution. Clearly, the Kernel estimator corrected is better than the classical Kernel estimator at the boundaries 0 and 1, in the sense that it is zero for values less than 0 or greater than 1 and thus is closer to the true distribution, whereas this is not achieved by the Kernel estimator.

Finally,

$$f_X(x) = f_Y(H^{-1}(x)) \frac{1}{H'(H^{-1}(x))}.$$

According to this proposition, the Kernel estimator based on the transformation (15), called *Kernel Etimator corrected* is given by

$$\hat{f}_X(x) = \hat{f}_Y(H^{-1}(x)) \frac{1}{H'(H^{-1}(x))} \tag{16}$$

3.5 Simulations

3.5.1 Parametric and non parametric estimates comparison

Example 1

Figure (9) shows the true density for Beta(2,2) density and two estimators, the Parametric [Moment] and the non-parametric [Kernel corrected] from formula (16) with $h = 0.3$ based on 100 observations from a simulated Beta(2,2) distribution. The parametric estimation is better in the sense that compared to the non-parametric, it is closer to the true density.

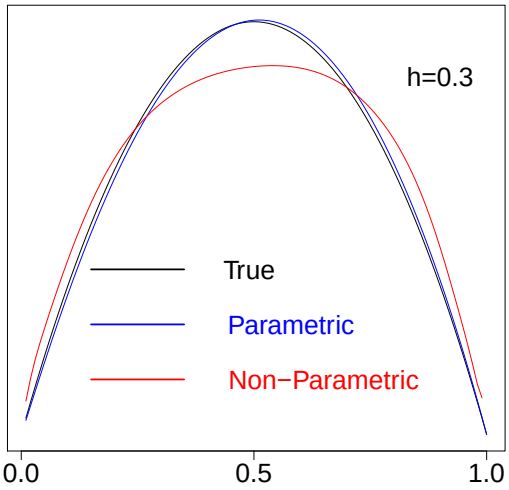

Figure 9: True density in black. Parametric Estimation in blue ($\hat{p} = 2.02, \hat{q} = 2.05$) versus Non-Parametric in red, for which the bandwidth $h = 0.3$.

Example 2

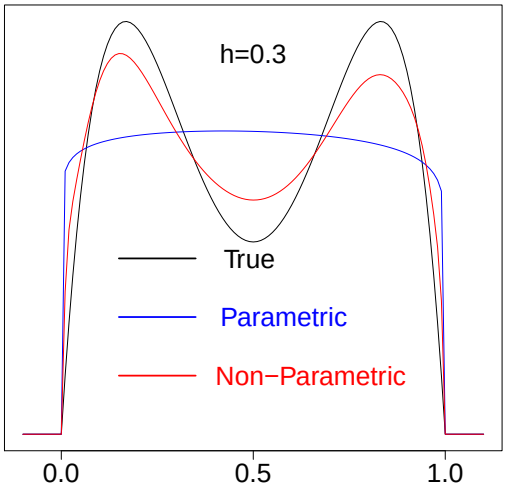

Figure 10: Parametric Estimation in blue ($\hat{p} = 1.05, \hat{q} = 1.06$) versus Non-Parametric in red, for which the bandwidth $h = 0.3$.

Figure (10) shows the true density for a mixture from Beta(6,2) and Beta(2,6) distribution with respective weights 1/2 and 1/2 and two estimators, the Parametric [Moment] and the non-parametric [Kernel corrected] from formula (16)with $h = 0.3$ based on 100 observations from a simulated mixture from Beta(6,2) and Beta(2,6) distribution. The non-parametric estimation is better in the sense that compared to the parametric one, it is able to detect the bimodal shape of the data whereas the parametric one cannot. Thus the non-parametric method is more flexible.

3.5.2 Real data application

Figure (11) shows the histogram of the *Percentage of attendance of students for the course of statistics during the first semester in Liverpool* and two estimators, the Parametric [Moment] and the non-parametric [Kernel corrected] from formula (16) with $h = 0.2$ based on 100 students attending the course. The data show a very asymetrical shape, most of the students attending the course whereas a few of them don't attend it. There are not big difference between the non-parametric estimation and the parametric one, except maybe at the right where the first one is a bit better. Note that the parametric estimator is a func-

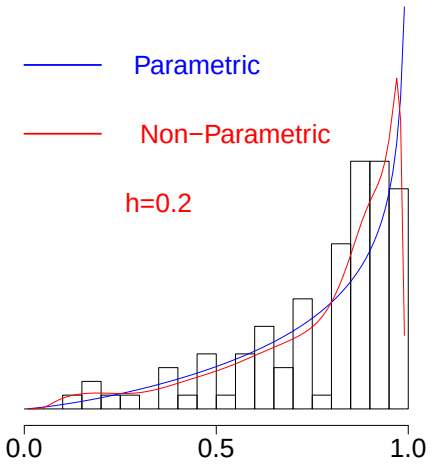

Figure 11: Histogram of the *Percentage of attendance of students for the course of statistics during the first semester in Liverpool.* Parametric Estimation in blue ($\hat{p} = 2.27, \hat{q} = 0.65$) versus Non-Parametric in red, for which the bandwidth $h = 0.2$.

tion which is not continuous at 1. If this is the case for the true distribution, then the assumptions for the non-parametric estimator are violated.

4 Conclusion

In this report, the main focus was the density estimation. More precisely, the non-parametric Kernel estimator defined in (1) is studied. Compared to the Naive estimator, it has the advantage to be smoother. One huge part is dedicated to both parameters governing the Kernel estimator: the bandwidth h and kernel K. One way to choose them optimally is to minimise the asymptotic MISE. The optimal h in this sense, called h_{opt}, has the problem of depending on the unknown density f, which makes it impossible to compute. The optimal K is called Epanechnikov Kernel, it can be computed explicitely. For non-negative data, an adapted Kernel estimator is defined in (10) to deal better with boundary 0.

Regarding compositional data which are vectors of proportions, Beta distribution is introduced because it suits well the kind of data in the sense that it is defined over [0,1]. Moment estimator has been derived for such a distribution, and also a corrected Kernel estimator defined in (16) to deal better with bounded data over [0,1]. Both parametric method [Moment] and non-parametric method [Kernel] have been experienced. In some cases, that is for example when the data are simulated from some known distribution, the first method works better. However, when the data are more complex, that is for example when the data are simulated from a mixture ditribution, then the non-parametric method is better. The non-parametric-method is hence more flexible, and more robust in the sense that it is not constrained by some previous knowledge about the data.

Due to limited time, I could not study some topics. Indeed, recall that the bandwidth h_{opt} is not computable as it depends on the second derivative of unknown density f. Thus, in the future, some work could be done to choose the bandwidth h by approximating it. Also, it could be interesting to investigate how the Kernel estimator works when dealing with multivariate data, that is when the observations X_1,X_2,...X_n are multidimensional.

References

[1] Fix, E. and Hodges, J.L. (1951). Discriminatory analysis, nonparametric estimation: consistencey properties. Report No.4, Project no.21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.

[2] Parzen,E. (1962). On estimation of a probability density function and mode. The Annals of Mathematical Statistics, **33**, 1065-1076.

[3] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics, **27**, 832-837.

[4] Boneva, L.I., Kendall,D.G. and Stefanov,I. (1971). Spline transformations: three new diagnostic aids for the statistical data-analyst (with Discussion). Journal of Royal Statistical Society. B, **33**, 1-70.

[5] Copas, J.B. and Fryer, M.J. (1980). Density Estimation and suicide risks in psychiatric treatment. Journal of Royal Statistical Society. A, **143**, 167-176.

[6] Silverman, B.W.(1986). Density Estimation for Statistics and Data Analysis.

[7] Aitchison, J.(1986). The statistical Analysis of Compositional Data.

[8] McNeil, D. R. (1977). Interactive Data Analysis. New York: Wiley.

[9] Azzalini, A. and Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. Applied Statistics **39**, 357365.

Appendix: codes

Naive Estimator

```
NaiveEstimator=function(x,y,h){
  p=length(x)
  n=length(y)
  diff=matrix(NA,p,n)
  vect=c(rep(NA,p))
  X=matrix(x,p,n)
  Y=matrix(y,n,p)
  Y=t(Y)
  diff=X-Y
  diff=diff/h
  diff=apply(diff,c(1,2),weight)
  vect=apply(diff,1,sum)/(n*h)
  vect
}
```

Kernel Estimator

```
KernelEstimator=function(x,y,h,K){
  p=length(x)
  n=length(y)
  diff=matrix(NA,p,n)
  vect=c(rep(NA,p))
  X=matrix(x,p,n)
  Y=matrix(y,n,p)
  Y=t(Y)
  diff=X-Y
  diff=diff/h
  diff=apply(diff,c(1,2),K)
  vect=apply(diff,1,sum)/(n*h)
}
```

Moment Estimator

```
MomentEstimator=function(y){
  ybar=mean(y)
  p=ybar*(ybar*(1-ybar)/var(y)-1)
  q=(1-ybar)*(ybar*(1-ybar)/var(y)-1)
  return(c(p,q))
}
```