Finite mixture models: visualisation, localised regression, and prediction

NAJLA MOHAMMED A QARMALAH

Finite mixture models: visualisation, localised regression, and prediction

Najla Mohammed Ashiq Qarmalah

A thesis presented for the degree of
Doctor of Philosophy

Department of Mathematical Sciences
Durham University
United Kingdom

February 2018

Dedicated to

my parents
whose affection, love,
encouragement and prayers of day
and night make me able to get such
success and honour

Finite mixture models: visualisation, localised regression, and prediction

Najla Mohammed Ashiq Qarmalah

Submitted for the degree of Doctor of Philosophy

February 2018

Abstract: Initially, this thesis introduces a new graphical tool, that can be used to summarise data possessing a mixture structure. Computation of the required summary statistics makes use of posterior probabilities of class membership obtained from a fitted mixture model. In this context, both real and simulated data are used to highlight the usefulness of the tool for the visualisation of mixture data in comparison to the use of a traditional boxplot.

This thesis uses localised mixture models to produce predictions from time series data. Estimation method used in these models is achieved using a kernel-weighted version of an EM–algorithm: exponential kernels with different bandwidths are used as weight functions. By modelling a mixture of local regressions at a target time point, but using different bandwidths, an informative estimated mixture probabilities can be gained relating to the amount of information available in the data set. This information is given a scale of resolution, that corresponds to each bandwidth. Nadaraya-Watson and local linear estimators are used to carry out localised estimation. For prediction at a future time point, a new methodology of bandwidth selection and adequate methods are proposed for each local method, and then compared to competing forecasting routines. A simulation study is executed to assess the performance of this model for prediction.

Finally, double-localised mixture models are presented, that can be used to improve predictions for a variable time series using additional information provided by other time series. Estimation for these models is achieved using a double-kernel-weighted

version of the EM–algorithm, employing exponential kernels with different horizontal bandwidths and normal kernels with different vertical bandwidths, that are focused around a target observation at a given time point. Nadaraya-Watson and local linear estimators are used to carry out the double-localised estimation. For prediction at a future time point, different approaches are considered for each local method, and are compared to competing forecasting routines. Real data is used to investigate the performance of the localised and double-localised mixture models for prediction. The data used predominately in this thesis is taken from the International Energy Agency (IEA).

Declaration

The work in this thesis is based on research carried out in the Department of Mathematical Sciences, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

I would like to express my sincere gratitude to Allah for the countless blessings he has bestowed on me, both in general and particularly during my work on this thesis.

My appreciation and special thanks go to my supervisors, Prof. Frank Coolen and Dr. Jochen Einbeck, for their unlimited support, expert advice and guidance in all the time of research and writing of this thesis. I appreciate their high standards, ability, and knowledge of research methodology. It is really hard to find the words to express my gratitude and appreciation for them.

I am immensely grateful to my parents for their faith, support, and constant encouragement. I thank my parents for teaching me to believe in Allah, in myself, and in my dream. I hope my parents are proud of this thesis.

I want to acknowledge my closest supporters, my brothers and sisters, for their confidence and constant encouragement, and my family and friends, who have been a great source of motivation through all these years.

I am also grateful to my country, Saudi Arabia, represented by King Salman, Saudi Arabian Cultural Bureau in London, Ministry of Education and Princess Norah bint Abdul Rahman University in Riyadh for granting me the scholarship and providing me with this great opportunity for completing my studies in the UK and so enabling me to fulfil my ambition.

Many thanks to Durham University for offering such an enjoyable academic atmosphere and for the facilities, that have enabled me to study smoothly.

My final thanks go to everyone who has assisted me, stood by me or contributed to my educational progress in any way.

Contents

Chapter 1

Introduction

1.1 Overview

The use of finite mixture models is a source of much debate mainly, because of the flexibility of the models across a wide variety of random phenomena, and the increase in available computing power. Finite mixture models have been successfully applied in many fields. For example, according to Mclachlan and Peel [62], applications of finite mixture models have been used in astronomy, biology, medicine, psychiatry, genetics, economics, engineering, and marketing, among many other fields in the biological, physical, and social sciences. In addition, finite mixture models have applications including cluster and latent class analyses, discriminant analysis, image analysis and survival analysis [62].

The use of finite mixture models has increased considerably over the past decade and the use of these models has continued to receive increasing attention in the years, both from a practical and theoretical point of view. In the 1990s, finite mixture models were extended by mixing standard linear regression models as well as generalised linear models [90]. Lindsay [57] discusses the non-parametric and semi-parametric maximum likelihood estimation used in mixture models, while Mclachlan and Peel [62] discuss the major problems relating to mixture models. Some issues discussed by both include identifiability problems, the EM–algorithm, the properties of the maximum likelihood estimators so obtained, and the assessment of the number of components used in the mixture.

One of the most popular mixture models is the mixture of Gaussian distributions due to its applications in various fields. This model is classed as the first mixture model as used by Karl Pearson [65]. Pearson fits a mixture of two Gaussian probability density functions with different means and variances. In fact, Gaussian mixture models are used in the investigation of the performances of certain estimators as departures from normality [62]. Consequently, Gaussian mixture models have been used in the development of robust estimators [62]. For example, under the contaminated normal family as outlined by Tukey [81], the density of an observation is taken to be a mixture of two univariate Gaussian densities with the same means but where the second component has a greater variance than the first.

The initial focus of this thesis is to develop a new graphical tool, that can be used to visualise Gaussian mixture data. This new graphical tool can provide additional information to the data analyst, where traditional plots cannot. Notably, this new plot can be applied to data, which belongs to any mixture of density distributions. This idea will be discussed in more detail in Chapter 2. In addition, a mixture of local regression models is developed for prediction from time series using two approaches. In the first approach, a mixture of local regression model is developed for prediction using past information from a target time series. This will be discussed in more detail in Chapter 3. Moreover, in the second approach, additional information is used from other time series, that is relevant to the target time series, in order to stabilise the prediction of the target time series in comparison with other time series models. This problem will be investigated in Chapter 4.

This first chapter will review local modelling as represented in local polynomial regression, and local likelihood methods. Previous research about mixture models used unknown distributional shapes of data will be presented and one of the most important classes of mixture models, a mixture of non-parametric regression models, will be discussed. In addition, popular estimation methods used in mixture models, such as the EM–algorithm and its application on mixtures of non-parametric regression models will be explained in more detail. This thesis examines the EM–algorithm in more detail in Chapter 2 and more advanced versions of the EM–algorithm in Chapters 3 and 4. In addition, several methodologies used for prediction are reviewed such as the

ARIMA and Holt models, which can be compared with the newly proposed methods for prediction as outlined in Chapters 3 and 4. This context is used to develop the new graphical tool for visualising mixture models and prediction from time series using new methodologies based on mixtures and local regression models.

1.2 Local modelling

Regression is one of the most commonly used statistical methods, simply because it can be applied across many research fields, including: econometrics, social science, medicine, and psychology. Linear regression is a classical and widely technique used, in order to study the relationship between variables and to fit a line through data. A simple linear regression model for given pairs of data such as $(x_i, y_i), i = 1, \ldots, n$ takes the form as follows:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1.2.1}$$

where β_0 and β_1 are the parameters, that represent the intercept and slope of the model, respectively. If the observed data has a linear trend, then the model (1.2.1) can be said to fit the data well, and parameters can be estimated using the least squares method. However, when the observed data takes a more complex shape, and cannot be converted into a linear relationship using transformation, then linear regression may not be an appropriate method to use. Consequently, a non-parametric regression model is considered to be a more useful technique for relaxing linearity assumption, and to avoid the restrictive assumptions relating to the functional form of the regression function [23]. Indeed, non-parametric regression belongs to a class of regression techniques whereby, according to Wand and Jones [85], the model is shaped completely based on the data. The non-parametric method is particularly useful to use when a parametric model becomes too restrictive.

There are several different non-parametric regression techniques, that can be used, and these can, generally, be split into the categories of spline-based and local methods [23]. Thus, instead of solving a parametric problem as demonstrated in the model (1.2.1), we can solve many linear regression problems using only two parameters $\beta_0(x)$ and $\beta_1(x)$, and using local linear regression. For example, if h is the size of the local

neighbourhood, known as a bandwidth or a smoothing parameter, then the local linear regression model can be defined as follows:

$$y_i = \beta_0(x) + \beta_1(x)x_i + \epsilon_i,\ x - h \leq x_i \leq x + h$$

where $\beta_0(x)$ and $\beta_1(x)$ are the parameters that depend on x.

A general non-parametric regression model can be written as:

$$y_i = m(x_{i1}, x_{i2}, \ldots, x_{iq}) + \varepsilon_i$$

where $m(\cdot)$ is known as the regression function, while $x_{i1}, x_{i2}, \ldots, x_{iq}$ are q predictors for the i-th of n observations. The errors ε_i are assumed to be independently distributed, with a mean of 0 and a constant variance of σ^2 [26]. However, most methods of non-parametric regression implicitly assume, that $m(\cdot)$ is a smooth continuous function [26]. One important instance of non-parametric regression is known as non-parametric simple regression, where there is only one predictor [23]:

$$y_i = m(x_i) + \varepsilon_i \tag{1.2.2}$$

This type of non-parametric regression is often called 'scatterplot smoothing' , because it traces a smooth curve through a scatterplot of y against x [26]

An important issue, that must be settled before using a local fitting technique is determining the 'local neighbourhood' or the window, which is commonly described using the kernel function W and the bandwidth parameter h [23]. The kernel function W is a weight function, that it weighs observations close to the target point more heavily and assigns a weight of 0 to far away observations [39]. According to Härdle et al. [39], for all the other kernel methods, the bandwidth h determines the degree of smoothness of the regression function estimation $\hat{m}(\cdot)$ by controlling the weights of the observation points used in the local neighbourhood. The larger the local neighbourhood, then the smoother the estimated regression function is. The bandwidth h can be chosen to be constant or to depend on the location [21]. The choice of the bandwidth h is more crucial than the kernel function, because it determines the complexity of a model. For example, when a bandwidth $h = 0$, this results in interpolating the data, which leads to the most complex model. On the other hand, when h tends to ∞, the data is fitted

globally, which is the simplest model. As a result, a bandwidth governs the complexity of the model [21]. The choice of kernel function does not mainly affect the performance of the resulting estimation in non-parametric regression, both theoretically and empirically [21]. Wand and Jones [85] discuss the effect of the kernel function on non-parametric smoothing estimation based on asymptotic mean integrated squared error (AMISE) criterion. The result suggests that most unimodel kernel functions perform about the same as each other, see for more detail [59].

Local fitting is indeed a particularly useful technique to use in non-parametric estimation [79]. This local modelling approach aims to relax the global linearity assumption through the local linear model, which results in a new objective function called the local maximum likelihood function [23]. In order to fit the regression function $m(x)$ in the model (1.2.2) at a particular point of x_0 locally, there are many ways, that can be used to evaluate the estimator of $m(x)$, where the data set can fit the smoother $\hat{y} = \hat{m}(x)$.

1.2.1 Nadaraya-Watson estimator

Nadaraya [64] and Watson [88] proposed a kernel regression estimator, usually referred to as the Nadaraya-Watson estimator, or local constant regression estimator. It belongs to a class of kernel regression estimators, that correspond to a local constant least squares fit. The Nadaraya-Watson estimator weighs the local average of the response variables y_i. In the Nadaraya-Watson estimator, when $h \to 0$, then $\hat{m}(x_i)$ converges to y_i at an observation x_i. As discussed in Section 1.2, the behaviour is different for $h \to \infty$, where an infinitely large h makes all weights equal, and local modelling becomes global modelling [39]. The main difference between parametric and non-parametric modelling is that for the former the bandwidth h is always infinite, but different parametric families of models are used. In non-parametric modelling, as in local modelling, several different bandwidths used need to be considered, so that the resulting curve articulates a given data set [23].

In definition, W is the real-value kernel function for assigning weights, and h is the bandwidth, a non-negative number, that controls the size of the local neighbourhood.

The Nadaraya-Watson kernel regression estimator can be represented as:

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} W_h(x_i - x) y_i}{\sum_{i=1}^{n} W_h(x_i - x)}$$

where $W_h(\cdot) = W(\cdot/h)/h$ [23]. The function W is usually taken to be a symmetric probability density, because it yields smaller mean integrated squared error (MISE). The MISE can be presented as follows [73]

$$\text{MISE}(x) = \int \text{MSE}(u) W(u) du$$

where $W(\cdot)$ is a weight function where $W(\cdot) \geq 0$ and $\text{MSE}(\cdot)$ is a mean squared error which is defined as follows

$$\text{MSE}(x) = E\left[\{\hat{m}(x) - m(x)\}^2\right]$$

See for a detailed discussion of symmetric kernel function [14]. For example, the Gaussian kernel is widely used for non-parametric smoothing. It is defined by the Gaussian probability density function as follows:

$$W(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right) \tag{1.2.3}$$

The kernel function based on a Gaussian probability density function, as defined in Equation (1.2.3) will be used later in Chapter 4.

1.2.2 Local polynomial estimator

In local polynomial regression, we first apply a Taylor expansion to $m(x)$ in a neighbourhood of x_0 as follows:

$$m(x) \approx m^{(0)}(x_0)/0! + m^{(1)}(x_0)/1!(x - x_0) + \ldots + m^{(p)}(x_0)/p!(x - x_0)^p = \boldsymbol{x}^T \boldsymbol{\beta} \tag{1.2.4}$$

where $\boldsymbol{x} = \{1, x - x_0, \ldots, (x - x_0)^p\}^T$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ such that $\beta_j = m^{(j)}(x_0)/j!, j = 0, \ldots, p$ and p is the degree of the polynomial. The points close to x_0 will have more influence on the estimate of $m(x_0)$, while the points furthest from x_0 will have the least influence [23]. A kernel function W_h puts more weight on the points near x_0, and less weight on the points furthest from x_0. To estimate the mean function $m(x)$, a weighted

Method	Bias	Variance
Nadaraya-Watson	$\left(\frac{d_2}{dx^2}m(x) + \frac{2\frac{d}{dx}m(x)\frac{d}{dx}f(x)}{f(x)}\right) b_n$	V_n
Local linear	$\frac{d_2}{dx^2}m(x)b_n$	V_n

$b_n = \frac{1}{2}h^2 \int_{-\infty}^{\infty} u^2 W(u)du, V_n = \frac{\sigma^2(x)}{f(x)nh} \int_{-\infty}^{\infty} W^2(u)du$

Table 1.1: Asymptotic biases and variances

polynomial regression can be minimised with respect to $\beta_0, \beta_1 \ldots$, and β_p as follows:

$$\sum_{i=1}^{n} \{y_i - \beta_0 - \beta_1(x_i - x_0) \ldots - \beta_p(x_i - x_0)^p\}^2 W_h(x_i - x_0) \tag{1.2.5}$$

The kernel function W_h controls the weights of the points at different locations. The resulting estimator is called the local polynomial regression estimator. For convenience, this can be denoted as follows:

$$\begin{aligned} \boldsymbol{W} &= \text{diag}\left\{W_h(x_1 - x_0), \ldots, W_h(x_n - x_0)\right\}, \\ \boldsymbol{X} &= \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} = \begin{pmatrix} 1, & x_1 - x_0 & \ldots, & (x_1 - x_0)^p \\ \vdots & \vdots & \ldots, & \vdots \\ 1, & x_n - x_0 & \ldots, & (x_n - x_0)^p \end{pmatrix} \end{aligned}$$

Then, the solution to the locally weighted least squares problem as presented in Equation (1.2.5) as follows:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y} \\ \hat{m}(x_0) &= \boldsymbol{e}_1^T \times \hat{\boldsymbol{\beta}} \end{aligned}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, and $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$ is a $1 \times (p+1)$ vector with the first entry being 1 and the others 0. Furthermore, we can obtain an estimate of the q-th $(q < p)$ derivative of $m(x)$ is as follows:

$$\hat{m}^{(q)}(x_0) = q!\boldsymbol{e}_{q+1}^T\hat{\boldsymbol{\beta}}$$

where $\mathbf{e}_{q+1}$ is a $1 \times (p+1)$ vector with $(q+1)$-th entry one and others 0 [23].

The Nadaraya-Watson estimator or local constant estimator is a special case of polynomial regression estimator when $p = 0$. When $p = 1$, the local polynomial regression estimator is known as a local linear estimator [23]. The asymptotic bias and variance

properties for a random design of the two estimators are summarised in Table 1.1 [23]. If we look at Table 1.1, we can see that a local linear estimator produces a more concise form of asymptotic bias than the Nadaraya-Watson estimator, but the asymptotic variances are the same. In addition, the local linear estimator offers several useful properties, such as automatic correction of boundary effects [13, 22, 59], design adaptivity, and best asymptotic efficiency using the minimax criteria [20]. Fan and Gijbels [23] offer a comprehensive account of local polynomial regression. In general, asymptotic and boundary bias correction advantages correspond to the local linear $p = 1$ and the local cubic $p = 3$ estimators are obtained. In addition, higher values of the degree of polynomial estimators, such as p, for example 2 or 3, enjoy the advantage of producing a greater smoothness of $m(\cdot)$. For example, higher values of the degree of polynomial estimators, such as p can yield a faster convergence rate to 0 of the mean squared error (MSE) [48]. The Nadaraya-Watson estimator and the local linear estimator will be used later in Chapters 3 and 4.

Model selection criteria can still be used to select variables for the local model. This determines whether an estimate $\hat{m}(x)$ is satisfactory, or whether alternative local regression estimates, for example, with different bandwidths, can produce better results. A good bandwidth plays an important role in local modelling. The most popular methods of selecting bandwidth typically minimise the mean squared error of the fit, or employ a formula, that approximates MSE [26]. For example, an optimal bandwidth can be obtained by minimising MISE, or an asymptotic leading term of MISE. In practice, data driven methods can be used for bandwidth selection, including the cross-validation (CV) criterion [77], for example. The CV criterion is computationally intensive method of bandwidth selection using the data. It has useful feature allowed by the generality of its definition, and it can be applied in a wide variety of settings. The CV can be defined as:

$$\mathrm{CV}(h) = \sum_{i=1}^{n} \{y_i - \hat{m}_i(x_i)\}^2$$

where $\hat{m}_i(x_i)$ is the estimate of the smooth curve at x_i, and is constructed from the reminder of the data, excluding x_i. A more developed version of the CV is a generalized

cross-validation (GCV), which has an efficient computational form as follows:

$$\text{GCV}(h) = \frac{n\text{RSS}}{\text{tr}\{I - S\}^2}$$

where RSS=$\sum_{i=1}^{n} \{y_i - \hat{m}(x_i)\}^2$ is the residual sum of squares, while S is a smoothing matrix, which can be considered as the analogue to the hat matrix, that is $\hat{m} = Sy$. The value of h, which minimises the formulas of MSE, MISE , CV or GCV, should provide a suitable level of smoothing. There are many methods that can be used for bandwidth selection, and these are described in more detail for example in [23, 84].

1.3 Local likelihood estimation

Local likelihood estimation is a useful technique, that avoids parametric form assumption for the unknown target function based on the idea of local fitting [79]. It has been discussed by various researchers across different domains of application. For example, local likelihood techniques have been developed for generalized linear models [25], hazard regression models [24] and estimating equations [10]. It was Tibshirani and Hastie [79] who first extended the idea of non-parametric regression to likelihood based regression models, for more details, see Fan et al. [24], whose research examines developments in this area. It is important to make the right choice about the size of the neighbourhood in local likelihood estimation. When each window contains 100% of the data with equal weight, the local likelihood procedure exactly resembles the global likelihood method, for more details, see Tibshirani [79]. Indeed, Fan et al. [21] show the connection between local polynomial regression and local likelihood estimation.

To illustrate the local likelihood concept, the model (1.2.2) with a normal and independently distributed error $\varepsilon_i \sim N(0, \sigma^2)$ is considered. It is assumed, that the observed data $\{(x_i, y_i), i = 1, \ldots, n\}$ comprised independent random samples from a population (X, Y), and therefore, (x_i, y_i) follows a normal regression model. Conditioning on $X = x$, the density function of Y can be written as follows

$$\phi(y|m(x), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\{y - m(x)\}^2\right]$$

Suppose we are interested in estimating $m(x)$ at x_0, which has a $(p+1)$-th continuous

derivative at the point x_0 as follows:

$$\begin{aligned} m(x_i) &\approx m(x_0) + m'(x_0)(x_i - x_0) + \ldots + \frac{m^{(p)}(x_0)}{p!}(x_i - x_0)^p \\ &= \boldsymbol{x}_i^T \boldsymbol{\beta}^0 \end{aligned} \tag{1.3.1}$$

where $\boldsymbol{x}_i = \{1, x_i - x_0, \ldots, (x_i - x_0)^p\}^T$, $\boldsymbol{\beta}^0 = (\beta_0^0, \ldots, \beta_p^0)^T$ with $\beta_\nu^0 = \frac{m^{(\nu)}(x_0)}{\nu!}$, $\nu = 0, \ldots, p$. If a data points x_i in a neighbourhood around x_0, $m(x_i)$ is approximated using the Taylor expansion, then a kernel-weighted log-likelihood is considered, which puts more weight on the points in the neighbourhood of x_0 and less weight on the points furthest from x_0. This kernel-weighted log-likelihood is known as a log local likelihood. Therefore, the log local likelihood function for a Gaussian regression model is written as follows:

$$\ell(\boldsymbol{\beta}) = -\log(\sqrt{2\pi\sigma^2}) \sum_{i=1}^{n} W_h(x_i - x_0) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left\{ y_i - \sum_{j=0}^{p} \beta_j^0 (x_i - x_0)^j \right\}^2 W_h(x_i - x_0)$$

Maximising the above local likelihood function is equivalent to minimising the following, which yields the local polynomial regression estimator:

$$\sum_{i=1}^{n} \left\{ y_i - \sum_{j=0}^{p} \beta_j^0 (x_i - x_0)^j \right\}^2 W_h(x_i - x_0)$$

In general, local likelihood estimation can be defined as follows: suppose we have independent observed data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ from population (X, Y), and (x_i, y_i) has a log-likelihood $\ell\{m(x_i), y_i\}$, whereas $m(x)$ is an unknown mean function. If we approximate $m(x_i)$ in a neighbourhood of x_0 using the Taylor expansion as in the Equation (1.3.1). Then, a log local likelihood function is as follows:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell\left\{ \mathbf{x}_i^T \boldsymbol{\beta}, y_i \right\} W_h(x_i - x_0) \tag{1.3.2}$$

By maximising the Equation (1.3.2) in regards to $\boldsymbol{\beta}$, the estimator of the $m(x)$ at point x_0 is $\hat{m}(x_0) = \hat{\beta}_0$ where $\hat{\boldsymbol{\beta}}$ is the solution.

Fan et al. [21] detail the applications of using the local likelihood method in non-parametric logistic regression. Furthermore, the asymptotic normality of local likelihood estimates has been studied in other research, that explores different models, for example: the generalized linear model [25], the hazard model setting [24], and for local

estimating equations [10]. Tibshirani and Hastie [79] also show that using the local likelihood procedures for local linear regression estimation produces favourable advantages. For example, a local linear estimator works well to reduce bias at the end-points in comparison to a local constant estimator in non-parametric regression. The local likelihood technique will be used later in Chapters 3 and 4.

1.4 Mixture models

A mixture model is a mixture of density functions, which has the density form as follows:

$$f(x \mid \mathbf{\Phi}) = \sum_{k=1}^{K} \pi_k f_k(x \mid \boldsymbol{\beta}_k) \tag{1.4.1}$$

where $\pi_k \geq 0$ with $\sum_{k=1}^{K} \pi_k = 1$ is the mixing proportion of the k-th component, $f_k(x \mid \boldsymbol{\beta}_k)$ is the k-th component density function, $\mathbf{\Phi} = \{\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\}$ is a vector containing all the parameters in the mixture model, and $\boldsymbol{\beta}_k$ is a vector containing all the unknown parameters for the k-th component [62].

Mixture models play an important role in the statistical analysis of data due to their flexibility for modelling a wide variety of random phenomena. In addition, using mixture models could be viewed as taking a model-based clustering approach towards data obtained from several homogeneous sub-groups with missing grouping identities [27, 62, 72]. As a result, mixture models are being increasingly studied in the literature across different fields and applications. Lindsay studies the theory and applications of mixture models in detail [57].

One mixture model, that is particularly useful, is a mixture of regression models. Goldfeld and Quandt [35] introduced a mixture of regression model, that is especially known as a switching regression model in the field of econometrics. Mixtures of regression models are appropriate to use when observations come from several sub-groups with missing grouping identities, and when in each sub-group, the response has a linear relationship with one or more other recorded variables. Another useful mixture model is the finite mixture of linear regression model, which has received increasing attention in research recently [35]: it has applications in econometrics and marketing [30, 70, 89], in epidemiology [37], and in biology [86]. The model setting can be stated as shown

below. Let $\mathcal{K}$ be a latent class variable with $P(\mathcal{K} = k) = \pi_k$ for $k = 1, 2, \ldots, K$, and supposing that given $\mathcal{K} = k$, the response y depends on $\boldsymbol{x}$ in a linear way where $\boldsymbol{x}$ is a p-dimensional vector:

$$y = \boldsymbol{x}^T \boldsymbol{\beta}_k + \epsilon_k = \beta_{0k} + \beta_{1k} x + \epsilon_k, \quad \epsilon_k \sim N(0, \sigma_k^2)$$

The conditional distribution of Y given $\boldsymbol{x}$ can be written as follows:

$$Y|\boldsymbol{x} \sim \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}^T \boldsymbol{\beta}_k, \sigma_k^2) \tag{1.4.2}$$

where $\{(\boldsymbol{\beta}_k, \sigma_k^2), k = 1, \ldots, K\}$ are the parameters of each component density, and $\{\pi_k, k = 1, \ldots, K\}$ are the mixing proportions for each component. The conditional likelihood function of a mixture of regression model can be written as follows:

$$f(y|x) = \sum_{k=1}^{K} \pi_k \phi(y|\boldsymbol{x}^T \boldsymbol{\beta}_k, \sigma_k^2)$$

McLachlan and Peel [62] have studied and summarised model (1.4.2), while the Bayesian approaches used in model (1.4.2) and the selection of the number of components K have been studied by Frühwirth-Schnatter [31] and Hurn, Justel, and Robert [47]. Jordan and Jacobs [53] note that the proportions depend on the covariates present in a hierarchical mixtures of experts model in machine learning. Mixture models continue to be subject to intense research activity, with special issues being tackled in close succession [8, 42]. A large proportion of articles about special issues discuss the variants of mixture regression models, such as Poisson regression, spline regression, or regression under censoring.

Recently, mixtures of non-parametric regression models, which relax the linearity assumption on the regression functions, have received particular attention. For example, Young and Hunter [91] use kernel regression to model covariate-dependent proportions for mixture of linear regression models. This idea is further developed by Huang and Yao [45] to develop a semi-parametric approach . Furthermore, Huang et al. [44] have proposed a non-parametric finite regression mixture model, where the mixing proportions, the mean functions, and the variance functions are all non-parametric, and this model has been applied to U.S. house price index (HPI) data. This model will be discussed in more detail in Chapter 3.

1.4.1 Mixture models estimation

There has been much discussion and debate regarding methods of estimation for mixture distributions. Over the years, a variety of approaches have been used to estimate mixture distributions. These approaches include graphical methods, method of moments, minimum-distance methods, maximum likelihood, and Bayesian approaches. The main reason for the huge literature on estimation methodology for mixtures is the fact that explicit formulas for parameter estimates are typically not available. For example, the maximum likelihood estimates (MLE) for the mixing proportions and the component means, variances and covariances are not available in closed form for normal mixtures [62]. Markov chain Monte Carlo (MCMC) sampling within a Bayesian framework can be used to estimate the parameters of finite mixture models [17].

There are two major classes of estimation methods for mixture models, and these are the EM–algorithm and the Bayesian methods, especially Markov Chain Monte Carlo estimation [62]. In addition, other methods have been developed based on the EM–algorithm and the Bayesian methods, in order to fit mixture models. For example, Stephens [76] presented the birth-and-death algorithm to be used as an estimation method for a mixture model. Smith and Roberts [74] proposed a Gibbs sampling procedure for mixture models and Frühwirth-Schnatter [31] gave a comprehensive summary of the Bayesian analysis for mixture models and Markov switching models. Although, using Bayesian methods provides more information about unknown parameters, they are very expensive in terms of computational cost.

The EM–algorithm was proposed in Dempster et al. [16], and systematically studied by McLachlan and Krishnan [61]. It is a technique that provides iterative steps to maximise the likelihood function, when some of the data is missing, in order to estimate the parameters of interest. Dempster et al. [16] called this method the EM–algorithm, where E stands for “expectation”and M stands for “maximisation”. McLachlan and Peel [62] provide a comprehensive review of the formulation of the mixture problem in the EM framework as an incomplete data problem, which is summarised as follows: suppose that the complete data is $\{(x_i, G_i), i = 1, \ldots, n\}$, the data comprising independent samples from population (X, G), where $\{x_i, i = 1, \ldots, n\}$ is the observed data, and $\{G_i, i = 1, \ldots, n\}$ is a K–dimensional vector with $G_{ik} = (G_i)_k = 0$ or 1, $k = 1, \ldots, K$,

according to whether x_i does or does not arise from the k-th component of the mixture. Let $\mathcal{L}(\mathbf{\Phi})$ be the complete likelihood function if the missing data G is given, and then, the complete log likelihood from Equation (1.4.1), is given by the following:

$$\ell(\mathbf{\Phi}) = \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\{\log \pi_k + \log f_k(x|\boldsymbol{\beta}_k)\}$$

The EM–algorithm consists of two steps: the E–step and the M–step. In the E–step, we compute the expectation of the complete log-likelihood function $\ell(\mathbf{\Phi})$ over the missing data conditioned on the observed data with the given parameters. For the E–step of the l-th iteration, we compute as follow:

$$Q(\mathbf{\Phi}|\mathbf{\Phi}^{(l)}) = E(\ell(\mathbf{\Phi})|\mathbf{\Phi}^{(l)}, x)$$

Let $\mathbf{\Phi}^{(0)}$ be the value specified initially for $\mathbf{\Phi}$. Then, in the first iteration of the EM–algorithm, the E–step requires the computation of the conditional expectation of $\ell(\mathbf{\Phi})$ given x, using $\mathbf{\Phi}^{(0)}$ for $\mathbf{\Phi}$, which can be written as

$$Q(\mathbf{\Phi}|\mathbf{\Phi}^{(0)}) = E(\ell(\mathbf{\Phi})|\mathbf{\Phi}^{(0)}, x)$$

This expectation operator is effected by using $\mathbf{\Phi}^{(0)}$ for $\mathbf{\Phi}$. It follows that on the $(l+1)$-th iteration, the E–step requires the calculation of $Q(\mathbf{\Phi}|\mathbf{\Phi}^{(l)})$, where $\mathbf{\Phi}^{(l)}$ is the value of $\mathbf{\Phi}$ after the l-th EM iteration. Therefore, we get the following

$$Q(\mathbf{\Phi}|\mathbf{\Phi}^{(l)}) = \sum_{k=1}^{K}\sum_{i=1}^{n} r_k(x_i; \mathbf{\Phi}^{(l)})\{\log \pi_k + \log f_k(x_i|\boldsymbol{\beta}_k)\} \tag{1.4.3}$$

where we can see the following:

$$r_k(x_i; \mathbf{\Phi}^{(l)}) = \pi_k^{(l)} f_k(x_i|\boldsymbol{\beta}_k) / \sum_{g=1}^{K} \pi_g^{(l)} f_g(x_i|\boldsymbol{\beta}_g)$$

The quantity $r_k(x_i; \mathbf{\Phi}^{(l)})$ is the posterior probability, that the i-th member of the sample with an observed value x_i belongs to the k-th component of the mixture. The M–step on the $(l+1)$-th iteration requires the global maximisation of $Q(\mathbf{\Phi}|\mathbf{\Phi}^{(l)})$, with respect to $\mathbf{\Phi}$ over the parameter space, to give an updated estimate $\mathbf{\Phi}^{(l+1)}$. For the finite mixture model, the updated estimates $\pi_k^{(l+1)}$ of the mixing proportions π_k are calculated independently of the updated estimate $\boldsymbol{\beta}_k^{(l+1)}$ of the parameter vector $\boldsymbol{\beta}_k$ containing the unknown parameters in the component densities. The updated estimate of π_k is

given as follows:

$$\pi_k^{(l+1)} = \frac{1}{n}\sum_{i=1}^{n} r_k(x_i; \mathbf{\Phi}^{(l)}), \quad k = 1, \ldots, K$$

Concerning the updating of $\boldsymbol{\beta}_k$ in the M–step of the $(l+1)$-th iteration, it can be seen in the Equation (1.4.3) that $\boldsymbol{\beta}_k^{(l+1)}$ is obtained as an appropriate root of the following:

$$\sum_{k=1}^{K}\sum_{i=1}^{n} r_k(x_i; \mathbf{\Phi}^{(l)}) \frac{\partial}{\partial \boldsymbol{\beta}_k} \log f_k(x_i|\beta_k) = 0 \tag{1.4.4}$$

The EM–algorithm gives the solution of Equation (1.4.4) in a closed form [62]. In general, the EM–algorithm leads to closed form for the estimators of parameters, which give an advantage in programming.

The EM–algorithm is one of the most used algorithms in statistics [16]. It is applied for missing data structures, which makes the maximum likelihood inference based on such data possible. In addition, mixture models are certainly a favourite domain for the application of the EM–algorithm [62]. McLachlan and Krishnan [61] study the advantages and disadvantages of the EM–algorithm. For example, the likelihood function $\mathcal{L}(\mathbf{\Phi})$ is increasing at each EM iteration, that is $\mathcal{L}(\mathbf{\Phi}^{(l+1)}) \geq \mathcal{L}(\mathbf{\Phi}^{(l)})$ for $l = 0, 1, \ldots$ [16]. Hence, a convergence must be obtained with a sequence of likelihood values, that are shown above. In practice, the E and M–steps are alternated repeatedly until the difference $\ell(\mathbf{\Phi}^{(l+1)}) - \ell(\mathbf{\Phi}^{(l)})$ is sufficiently small in the case of convergence of the sequence of log likelihood values $\left\{\ell(\mathbf{\Phi}^{(l)})\right\}$ [62]. Mclachlan and Peel [62] discuss the stopping criterion of the EM–algorithm which adopted in term of either the size of the relative change in the parameter estimates or the log likelihood $\ell(\cdot)$.

Wu [52] and Mclachlan and Krishnan [61] note that the convergence behaviours of the EM–algorithm, and state that the EM–algorithm can provide global maximum likelihood estimators under fairly general conditions [16, 61]. However, the convergence of the EM–algorithm is relatively slow, and its solutions may be highly dependent on its initial position $\mathbf{\Phi}^{(0)}$. Baudry and Celeux [5] studied how EM–algorithm initialisation affects the estimation and the selection of a mixture model, especially in Gaussian mixture models. They presented strategies for choosing the initial values $\mathbf{\Phi}^{(0)}$ for the EM–algorithm. In conclusion, Baudry and Celeux state that no method can effectively be used to address the dependence of the EM–algorithm on its initial position in all situations. However, others have suggested solutions, in order to overcome this drawback

by using the penalized log-likelihood of Gaussian mixture models in a Bayesian regularization perspective and then choosing the best among several relevant initialisation strategies [5]. The EM–algorithm will be used in Chapter 2, and developed versions of the EM–algorithm will be used in Chapters 3 and 4.

1.4.2 Choosing the number of components

Choosing the number of components is a crucial issue in mixture modelling. In genetic analysis and other applications, a question arises as to whether observed data are a sample from a single population or whether the data have come from several separate populations. In the literature, two major approaches have been examined, in order to select the number of components K for a mixture model for unknown distributional shapes of data, and these are the classic and the Bayesian approaches. One approach for testing the number of components is to boot-strap a likelihood ratio test. The bootstrap test procedure was proposed by Hope [43], an illustration of its use in mixture models was given in Aitkin et al. [2]. It is a re-sampling approach used to assess the p-value of the likelihood ratio test statistic (LRTS) [62]. This is

$$H_0 : K = K_0 \qquad \text{versus} \qquad H_1 : K = K_1$$

with $K_1 = K_0 + 1$ in practice [62]. To illustrate this method, we suppose that $\hat{\mathbf{\Phi}}_K$ is the estimate of $\mathbf{\Phi}_K$ when K mixture components are used. Then, the likelihood ratio test can be defined as follows:

$$D = -2 \log \frac{\mathcal{L}(\hat{\mathbf{\Phi}}_{K_0})}{\mathcal{L}(\hat{\mathbf{\Phi}}_{K_1})} \tag{1.4.5}$$

To test the hypothesis above, the value D is computed from Equation (1.4.5), which is denoted as D_0. Then, N bootstrap samples of size n are generated from the mixture model fitted under the null hypothesis of the K_0 components. For each of N data sets, the process is repeated by recalculating $\hat{\mathbf{\Phi}}_{K_0}$ and $\hat{\mathbf{\Phi}}_{K_1}$ and by computing the corresponding value of D. Next, the position d of D_0 within all other values of D is determined. Finally, the test rejects the null hypothesis H_0 if D_0 is greater than the statistics χ^2_α where $\alpha = \frac{1-d}{N+1}$.

Polymenis and Titterington [66] have proposed modified version of Windham and Cut-

ler's method for determining the number of components in a mixture and compared the modified method with the bootstrap likelihood ratio method. The result of comparison is, that the bootstrap likelihood ratio method has some obvious advantages over the modified method, that it takes into account the single component densities and it performs better for small sample size. However, the modified method performs as well as the bootstrap likelihood ratio when the sample size is not small and the 'true' mixture, which the data come from is known [66].

There are also two popular model selection criteria, that can be used for choosing the number of components. The first one is the Akaike information criterion (AIC) [3], which is given by

$$\mathrm{AIC} = -2\ell(\hat{\mathbf{\Phi}}) + 2d$$

where d is the number of parameters in a K component mixture model. The second one is the Bayesian information criterion (BIC) [71], which is defined as

$$\mathrm{BIC} = -2\ell(\hat{\mathbf{\Phi}}) + d\log n$$

Leroux [56] established, under mild conditions, that certain penalized loglikelihood criteria, including AIC and BIC, do not underestimate the true number of components, asymptotically. Other satisfactory conclusions for the use of AIC or BIC in this situation are discussed by Biernacki, Celeux, and Govaert [62], Cwik and Koronacki [15], and Solka et al. [75]. Other nonparametric methods that have been used for this problem include the work of Henna [40] and a number of graphical displays, for example the normal scores plot [11, 38]. Previously, Lindsay and Roeder [58] had proposed the use of residual diagnostic for determining the number of components. Miloslavsky and van der Laan [63] have investigated minimisation of the distances between the fitted mixture model and the true density as a method for estimating the number of components using cross validation. They present simulation studies to compare the cross validated distance method with AIC, BIC, Minimum description length principle (MDL) and Information Complexity (ICOMP) on univariate normal mixtures [63]. For further information, see Chen and Kalbeisch [12], Lindsay [57], and McLachlan and Peel [62]. Bayesian methods provide estimates of K as well as their posterior distributions by assuming some prior distributions. There are many Bayesian methods that can be

used. For example, the reversible jump Metropolis-Hasting algorithm [36], and the birth-death processes [76].

1.5 Prediction

There have been significant discussion and debate in many fields of applied sciences and in the statistical literature about prediction of future values from time series. Statistical literature about prediction is abundant [4]. There are two major approaches taken when dealing with prediction: parametric and non-parametric. The most popular approaches used for prediction are the ARIMA model, which corresponds to a parametric approach, and the exponential smoothing, which is a non-parametric approach. These two methods are considered as automatic forecasting algorithms, which determine an appropriate time series model, estimate the parameters, and compute the forecasts [49]. In addition, there are robust versions of the ARIMA, exponential, and Holt-Winters smoothing method, which are suitable for forecasting univariate time series in presence of outliers, see for example [33, 83].

Although, the ARIMA and exponential smoothing models are different methodologies, and are correspond to different classes, they overlap [49]. For example, Hyndman et al. [50] claim that linear exponential smoothing models are all special cases of ARIMA models. However, the non-linear exponential smoothing models differ from ARIMA models. On the other hand, there are many ARIMA models that have no equivalent exponential smoothing models. As a result, there is overlap between these classes and they compliment each other. In addition, each class has its advantages and drawbacks [49]. For example, the exponential smoothing models can be used for modelling non-linear time series data. In addition, for seasonal data, the exponential smoothing models perform better than the ARIMA models for the seasonal M3-competition data, which are available in R package **Mcomp**. Although there are more the ARIMA models than the exponential smoothing models for seasonal data, the smaller exponential smoothing class can capture the dynamics of almost all real business and economic time series, see Hyndman and Khandakar [49] for more detail about the features of the ARIMA and the exponential smoothing models.

Exponential smoothing is an elementary non-parametric method of forecasting a future realisation from a time series. It is an algorithm for producing point forecasts only. Gardner [32] reviews earlier papers about the context of exponential smoothing since the 1950s. All exponential smoothing methods have been shown to produce optimal forecasts from innovation state space models (see for example, [51]). Taylor [78] extends the discussion about the exponential smoothing models by listing a total of fifteen methods. These models are summarised by Hyndman [49]. Some of these models are popularly approaches in forecasting, for example, the simple exponential smoothing (SES) method and Holt's linear method, respectively. In addition, the additive Holt-Winters' method and the multiplicative Holt-Winters' method are also more commonly used methods. To explain how to calculate the point forecast using such methods, we suppose that the observed time series is given by $y_1, y_2, \ldots, y_n$ and a forecast of m step ahead y_{T+m} based on all of the data up to time T, is denoted by $\hat{y}_{T+m|T}$. The point forecasts and updating equations for the Holt-Winters' additive method are as follows

$$\begin{aligned}
\text{Level} \quad &: \quad \ell_T = \alpha(y_T - s_{T-c}) + (1-\alpha)(\ell_{T-1} + b_{T-1}) \\
\text{Growth} \quad &: \quad b_T = \beta^*(\ell_T - \ell_{T-1}) + (1-\beta^*)b_{T-1} \\
\text{Seasonal} \quad &: \quad s_T = \gamma(y_T - \ell_{T-1} - b_{T-1}) + (1-\gamma)s_{T-c} \\
\text{Forecast} \quad &: \quad \hat{y}_{T+m|T} = \ell_T + b_T m + s_{T-c+m_c^+}
\end{aligned} \tag{1.5.1}$$

where c is the length of seasonality, for example, the number of months or quarters in a year, ℓ_T is the level of the series, b_T is the growth, s_T is the seasonal component, β^* and γ are the smoothing parameters. For Holt-Winters' additive method, the values for the initial states ℓ_0, b_0, $s_{1-c}, \ldots, s_0$, α, β^* and γ should be set. All of these initial values will be estimated from the observed data. The formula of s_T in Holt-Winters in Equation (1.5.1) is not unique. It has been modified in some literature to make it simpler. Hyndman [49] gives more details about the different forms of s_T, which can be found in previous literature. In addition, Hyndman [49] summarises formulae for computing point forecasts m periods ahead for all of the exponential smoothing methods.

The most basic exponential smoother is the exponentially weighted moving average (EWMA). The EWMA is a technique used to estimate the underlying trend in a

scatterplot without the use of restrictive models, and it is similar in concept to the non-parametric regression technique. In fact, the EWMA is virtually identical to the Nadayara-Watson kernel estimator with a half kernel function, that gives 0 in its positive arguments [34]. The EWMA can be defined as follows: Let $y_1, \ldots, y_T$ be a time series observed at equally spaced time points $t_1, \ldots, t_T$. The EWMA forecasts y_{T+1} by using a weighted average of past observations with geometrically declining weights is given by [23]

$$\hat{y}_{T+1} = \frac{\sum_{i=1}^{T} \exp\left(\frac{t_i - t_{T+1}}{h}\right) y_i}{\sum_{i=1}^{T} \exp\left(\frac{t_i - t_{T+1}}{h}\right)}$$

The ARIMA and Holt models are used and compared with models proposed for prediction in Chapters 3 and 4.

1.6 Outline of thesis

The rest of this thesis is organized as follows: Chapter 2 introduces a new graphical tool designed to summarise data, which possesses a mixture structure. This includes computational elements of the plot, real data examples and a simulation study. A paper presenting the results of Chapter 2 has already been published in *Statistical Papers* [67]. This chapter has also been presented at several conferences, including: the *Northern Postgraduate Mini-Conference in Statistics* in Durham (June 2015), and the *Saudi Student Conference* in Birmingham (February 2016).

Chapter 3 presents localised mixture models, that can be used for prediction. In this context, the estimation procedure and the identifiability of these models are also explained. A new methodology of bandwidth selection for prediction is also proposed. In addition, several approaches used for prediction based on bandwidth selection using these models are suggested. Furthermore, a simulation study is conducted, in order to assess the performance of these models in terms of prediction, and to compare them with other common time series models. At the end of this chapter, real data examples are given. A paper presenting some parts of Chapter 3 has already been published in the *Archives of Data Science Series A* [68]. Chapters 2 and 3 have also been jointly presented at several seminars and conferences, including: the *37th Annual Research Students' Conference in Probability and Statistics* in Nottingham (April 2014), the

Northern Postgraduate Mini-Conference in Statistics in Newcastle (June 2014), the *Durham Risk Day* in Durham (November 2014), the *Saudi Student Conference* in London (January 2015), the *European Conference on Data Analysis* (ECDA) in Essex (September 2015), the *22nd International Conference on Computational Statistics* in Oviedo (Spain) (August 2016), at a research seminar at the Department of Mathematical Sciences at Durham University (March 2017), and at the *40th Annual Research Students' Conference in Probability and Statistics* in Durham (April 2017).

In Chapter 4, double-localized mixture models are presented, and in the context, the chapter discusses the estimation procedure. Several approaches for prediction based on bandwidth selection using these models are suggested. At the end of this chapter, real data examples are given, that assess the performance of these models for prediction use.

Chapter 5 summarises the key results of this thesis and discusses ideas for future research. There are many interesting opportunities to develop and extend the research presented in this thesis. Some of these are mentioned in the final sections of Chapters 2 to 4 and in Chapter 5.

At the end of this thesis, three appendices are provided. Appendix A illustrates the key notations used in this thesis. In Appendix B, auxiliary results are presented related to Chapter 3. Finally, Appendix C shows auxiliary results related to Chapter 4.

Chapter 2

Visualisation of mixture data

This chapter introduces a new graphical tool, that can be used to visualise data, which possesses a mixture structure. Computation of the required summary statistics makes use of posterior probabilities of class membership, which can be obtained from a fitted mixture model. Real and simulated data are used to highlight the usefulness of this tool for the visualisation of mixture data, in comparison to using a traditional boxplot.

2.1 Introduction

Visualisation tools play an essential role in analysing, investigating, understanding, and communicating various forms of data, and the development of novel graphical tools continues to be a topic of interest in the statistics literature. For example, Wang and Bellhouse [87] recently introduced a new graphical tool, known as the shift function plot, in order to evaluate the goodness-of-fit of a parametric regression model. A boxplot is one of the most popular graphical techniques used in statistics. It was first proposed for use as a unimodal data display by Tukey [82], who referred to it as a "schematic plot" or a "box-and-whisker plot", but it is now commonly known as the boxplot. A boxplot, in its simplest form, aims at summarising a univariate data set by displaying five main statistical features as follows: the median, the first quartile, the third quartile, the minimum value and the maximum value.

The boxplot has become one of the most frequently used graphical tools for analysing data, because it provides information about the location, spread, skewness, and long-

tailedness of a data set at a quick glance. The median in a boxplot serves as a measure of location. The dispersion of a data set can be assessed by observing the length of a box or by examining the distance between the ends of the whiskers. The skewness can be observed by looking at the deviation of the median line from the center of the box, or by examining the length of the upper whisker as relative to the length of the lower one. In addition, the distance between the ends of the whiskers in comparison to the length of the box displays longtailedness [6]. Alternative specifications for the ends of the whiskers can be used with a particular view for outlier detection. Specifically, the boundaries $Q_1 - 1.5IQR$ to $Q_3 + 1.5IQR$ can be computed where Q_1, Q_3 and IQR represent the first quartile, third quartile and interquartile range, respectively. Then, any observations smaller than $Q_1 - 1.5IQR$, or greater than $Q_3 + 1.5IQR$ are labelled as "outliers", for more details, see for example [29]. Finally, whiskers are drawn from the box to the furthest non-outlying observations. Additionally, notches can be added, which approximate a 95% confidence interval for the median [55].

Further variants of the boxplot have been developed, in oreder to analyse special kinds of data. For example, Abuzaid et al. [1] proposed a boxplot for circular data. Additionally, Hubert and Vandervieren [46] presented an adjustment of the boxplot to tackle outliers present in skewed data by modifying the whiskers. Recently, Bruffaerts et al. [9] have developed a generalized boxplot, that is more appropriate for skewed distributions and distributions with heavy tails.

As observed by McGill et al. [60], the traditional boxplot is not able to adequately display data, which is divided into certain groups or classes. Therefore, they developed a version of the boxplot for grouped data, which sets the widths of each group-wise boxplot as proportional to the square root of the group sizes. However, this technique requires the groups to be defined a priori, and for the group membership of each observation to be known. In practice, it is common to deal with data sampled from heterogeneous sub-populations, for which the group membership is a latent variable. To our knowledge, there is no appropriate plot that can represent such mixture data properly. Consequently, this research introduces a new plot tailored to mixture data to which we refer as a K–boxplot, where K is the number of mixture components. Compared to a boxplot, the K–boxplot is able to display important additional infor-

mation regarding the structure of the data set. Both K–boxplots and boxplots have a similar constructions: they contain boxes and they display extreme values. However, the K–boxplot visualises the K components of mixture models by using K different boxes. This can be compared to a boxplot, which uses only one box. A boxplot is a special case of a K–boxplot with $K = 1$.

Figure 2.1 provides a schematic display of, what we will refer to as a 'full', using a K–boxplot in the special case of $K = 3$, which describes the main features of K–boxplots in general. The K–boxplot displays K rectangles oriented with the axes of a co-ordinate system, in which one of the axes has the scale of a data set. The key features that appear in a K–boxplot are the weighted median ($M(w)$), the first weighted quartile ($Q_1(w)$) and the third weighted quartile ($Q_3(w)$) in each box, where w is a set of corresponding non–negative weights. These are displayed as respective weighted quantiles using the posterior probabilities of group membership as weights, as will be explained in more detail later. The bottom and top of the boxes show the weighted first and third quartiles of the data in each group, respectively. Weighted medians are displayed as horizontal lines and drawn inside the boxes. Additional information is provided along the widths of the boxes, and these depend on the mixing proportions of the mixture.

Just as for usual boxplots, any data points outside the boxes can be displayed in several ways. Here, any points that appear fully outside of the boxes are displayed individually using horizontal lines, and can therefore, be used to identify outliers. The lengths of these lines correspond to the posterior probabilities of group membership, which will be explained in more detail using real data examples later. Furthermore, variants of the K–boxplot that display points outside the boxes in different ways will be introduced in Section 2.3.1.

K–boxplots can be used to show a mixture data, the location, spread and skewness for each component in a mixture, and this information is displayed transparently to viewers. Each of the component-wise boxplots can be interpreted in the same way as traditional boxplots with respect to these measures, allowing for a detailed appraisal of the data. The required information needed in order to draw a K–boxplot can be estimated using different methods, for example using the EM–algorithm. However, it

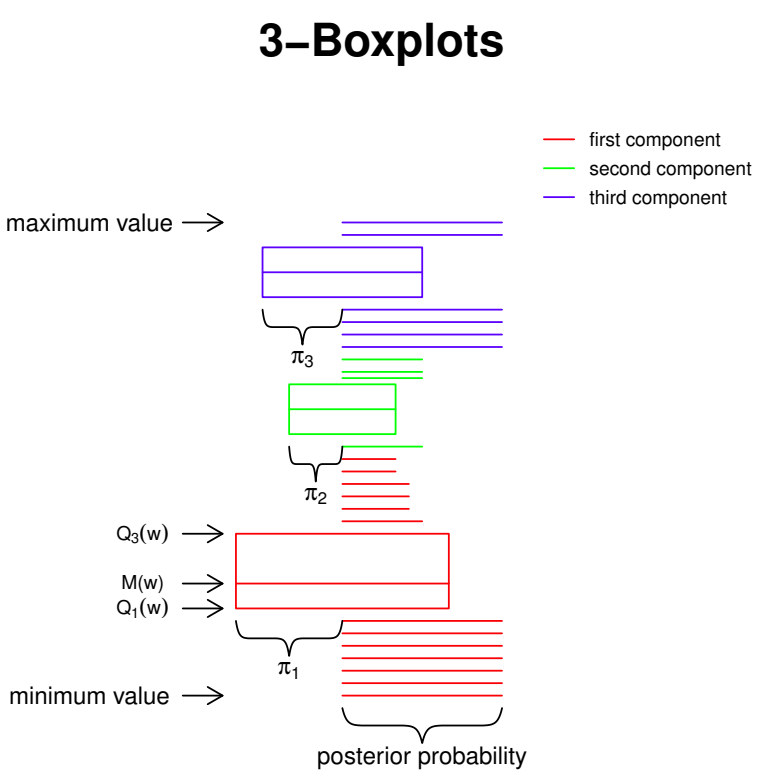

Figure 2.1: Summary of information provided by a 3–boxplot in its 'full' form. Here $M(w)$ denotes the weighted median, and $Q_j(w)$ the j-th weighted quartile, using the notation formally introduced in Section 2.2.2

should be noted that K–boxplots are not an *inferential* tool, and the K–boxplots will not make any automated decision about the choice of the mixture distributions, or the number of components, but they visualise the result of such inferential decisions made by the data analyst. Since the data analyst will be able to identify the impact of their model choices at a glance, K–boxplots will support them in making such choices in an informed manner.

The structure of the remainder of this chapter can be outlined as follows: Section 2.2 describes the computational elements of a K–boxplot, including the posterior probabilities derived from mixture models, as well as weighted quartiles. Section 2.3 discusses two real data examples, and Section 2.4 offers conclusions. Code used to execute K–boxplots is provided in the statistical programming language R [69] in the form of function `kboxplot` using the package **UEM** [18].

2.2 Computational elements of K–boxplots

2.2.1 Posterior probabilities

If we assume a random variable Y with density $f(y)$, which is a finite mixture of K probability density functions $f_k(y), k = 1, \ldots, K$, then it can be seen that

$$f(y) = \sum_{k=1}^{K} \pi_k f_k(y) \tag{2.2.1}$$

with the masses, or mixing proportions, $\pi_1, \ldots, \pi_K$ with $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. We can refer to $f_k(\cdot)$, which depend on the parameter vector θ_j, as the j-th component of the mixture of probability density functions. Just to clarify the terms, when speaking of 'mixture data' we mean data $y_i, i = 1, \ldots, n$, and it is plausible to assume that the data has been independently generated from, or at least can be represented by, a model of the type shown in Equation (2.2.1).

Now, let G be the random vector, which draws a class $k \in \{1, \ldots, K\}$, where the following applies:

$$G_{ik} = \begin{cases} 1, & \text{if observation } i \text{ belongs to component } k \\ 0, & \text{otherwise} \end{cases} \tag{2.2.2}$$

We assume that for an observation y_i, the value G is known. This means that we know to which of the K components the i-th observation belongs. If we interpret the π_k as 'prior' probability of class membership, then posterior probabilities of class membership can be produced using Bayes' theorem, that is, for the i-th observation y_i, $i = 1, \ldots, n$ can be represented as follows:

$$r_{ik} = P(G_{ik} = 1 | y_i) = \frac{\pi_k f_k(y_i)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(y_i)} \tag{2.2.3}$$

These posterior probabilities are combined into a weight matrix $R = (r_{ik})_{1 \leq i \leq n, 1 \leq k \leq K}$ form, which is the key ingredient of a K–boxplot. They will be used to compute the component-wise medians and quartiles, and, furthermore, it enables an immediate computation of the estimate as follows:

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} r_{ik} \tag{2.2.4}$$

This will be used to determine the width of the k-th K–boxplot. It should also be noted that assigning each data point y_i to the component k, which maximises r_{ik} for the fixed i, and posterior probabilities can be used as a classification tool. This is known as the maximum a posteriori (MAP) rule [62].

The estimates of θ_k are *not* needed for the construction of the K–boxplot itself. However, the computation of (2.2.3) involves the densities f_k and hence θ_k. Therefore, the θ_k need to be computed along the way as well. Most commonly, mixture models can be estimated using the EM–algorithm. In this case, the values θ_k are updated in the M–step, and the Equation (2.2.3) corresponds exactly to the E–step as discussed in Section 1.4.1 in Chapter 1, using the current estimates of π_k and θ_k. In practice, the r_{ik} can be conveniently extracted from the output of the final EM iteration.

The application of K–boxplots is not restricted to a certain choice of component densities. In principle, K–boxplots can be used to visualise the results of fitting a mixture of any combination of densities f_k, provided, that one is able to compute the parameters θ_k in the M–step. The choice of f_k is taken to the data analyst. In the absence of any strong motives to use a different distribution, a normal distribution will often be a convenient choice for the component densities. In this case, this is demonstrated as follows:

$$f_k(y) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$

where μ_k represent the component means and σ_k represent the component standard deviations. Maximising the complete log likelihood in the M–step gives the estimates as follows:

$$\begin{aligned} \hat{\mu}_k &= \frac{\sum_{i=1}^n r_{ik} y_i}{\sum_{i=1}^n r_{ik}} \\ \hat{\sigma}_k^2 &= \frac{\sum_{i=1}^n r_{ik}(y_i-\mu_k)^2}{\sum_{i=1}^n r_{ik}} \end{aligned} \tag{2.2.5}$$

The EM–algorithm consists of iterating the Equations (2.2.3) and (2.2.5) until convergence occurs [16]. The initial values $\theta_k^{(0)}, \pi_k^{(0)}$, $k = 1, \ldots, K$, are required for the first E–step. It is well known that different starting points can lead to different solutions, that correspond to the different local maxima of the log-likelihood, see [62] for a detailed discussion of this problem. Possible strategies for choosing the starting points include using: random initialisation, quantile-based initialisation, scaled Gaussian quadrature

points, or short EM runs [7]. These strategies continue to be under discussion and are the subject of research. A recent contribution on the topic is provided by Baudry and Celeux [5].

2.2.2 Weighted quartiles

If we suppose that $y_1 \leq \ldots \leq y_n$ indicate the ordered observations and $w = \{w_1, \ldots, w_n\}$ are a set of corresponding non-negative weights. Then, the equation below

$$m(w) = \max\{\ell : \frac{\sum_{i=\ell}^{n} w_i}{\sum_{i=1}^{n} w_i} \geq \frac{1}{2}\}$$

provides a maximal index ℓ, so that the total weight of observations larger, or equal than y_ℓ is at least 50%. Hence, the weighted median of $y_1, \ldots, y_n$ is defined by Fried et al. [28]:

$$M(w) = y_{m(w)}$$

There is no unique definition for quartiles, but in analogy to the above, one can define the first weighted quartile of $y_1, \ldots, y_n$ as $Q_1(w) = y_{q_1(w)}$, where the following applies:

$$q_1(w) = \max\{\ell : \frac{\sum_{i=\ell}^{n} w_i}{\sum_{i=1}^{n} w_i} \geq \frac{3}{4}\}$$

And the third weighted quartile of $y_1, \ldots, y_n$ as $Q_3(w) = y_{q_3(w)}$ can be represented as follows:

$$q_3(w) = \max\{\ell : \frac{\sum_{i=\ell}^{n} w_i}{\sum_{i=1}^{n} w_i} \geq \frac{1}{4}\}$$

In general, the weighted quartile of $y_1, \ldots, y_n$ as $Q_i(w) = y_{q_i(w)}, i = 1, 2, 3$ can be defined as follows:

$$q_i(w) = \max\{\ell : \frac{\sum_{i=\ell}^{n} w_i}{\sum_{i=1}^{n} w_i} \geq \alpha\}$$

where $\alpha = \frac{1}{2}, \frac{3}{4}$ and $\frac{1}{4}$ for $i = 2, 1$ and 3 respectively.

For example, the weighted median of 1, 3, 4, 7 and 9 with weights 0.2, 0.25, 0.3, 0.05, and 0.2 is $y_3 = 4$, because 0.2+0.05+0.3 $\geq$ 0.5. In addition, the first and third weighted quartile of the data are $y_2 = 3$ and $y_4 = 7$ respectively because 0.2+0.05+0.3+0.25 $\geq$ 0.75 and 0.2+0.05=0.25. An illustration of this process is provided in Table 2.1.

In the case of a K–boxplot, the box corresponding to the k-th component is fully

ℓ	1	2	3	4	5
y_ℓ	1	3	4	7	9
w_ℓ	0.2	0.25	0.3	0.05	0.2
$\sum_{i=\ell}^{n} w_i$	1	0.80	0.55	0.25	0.2

Table 2.1: Illustration of computation of weighted quantiles

determined by the observations y_i and the weights $w_i = r_{ik}$, $i = 1, \ldots, n$ which are the posterior probabilities of class membership. It should be noted, that these weights, for a fixed k, generally do not sum to 1. In addition, the weights $w_i, i = 1, \ldots, n$ can be determined basically using different methods rather than the E-step of the EM–algorithm in a Bayesian framework. For example, the Gibbs sampler could be used as a method to find the posterior probabilities. However, the EM–algorithm has good features to produce the posterior probabilities in comparison to the Gibbs sampler. For example, it gives a form of the posterior distribution at a lower cost than the Gibbs sampler, see McLachlan and Peel [62] for more details about the difference between the EM–algorithm and the Gibbs sampler.

2.3 Examples

In this section, two examples are presented to illustrate the usefulness of the K–boxplots for mixture data, in comparison to traditional boxplots. Moreover, an additional example is provided by Qarmalah et al. [67].

2.3.1 Example 1: energy use data

The data discussed in this example is taken from the International Energy Agency (IEA)[1]. The data used gives the annual energy use (in kg oil equivalent per capita) for 134 countries around the world between 1971 and 2011. The nature of the data, which is restricted to the positive range, and it features several countries with extremely large energy use. Therefore, a log-transformation will be applied in all further analyses. This example will consider only the year 2011 initially, for which Figure 2.2 presents four

[1]International Energy Agency, available at: http://www.iea.org/

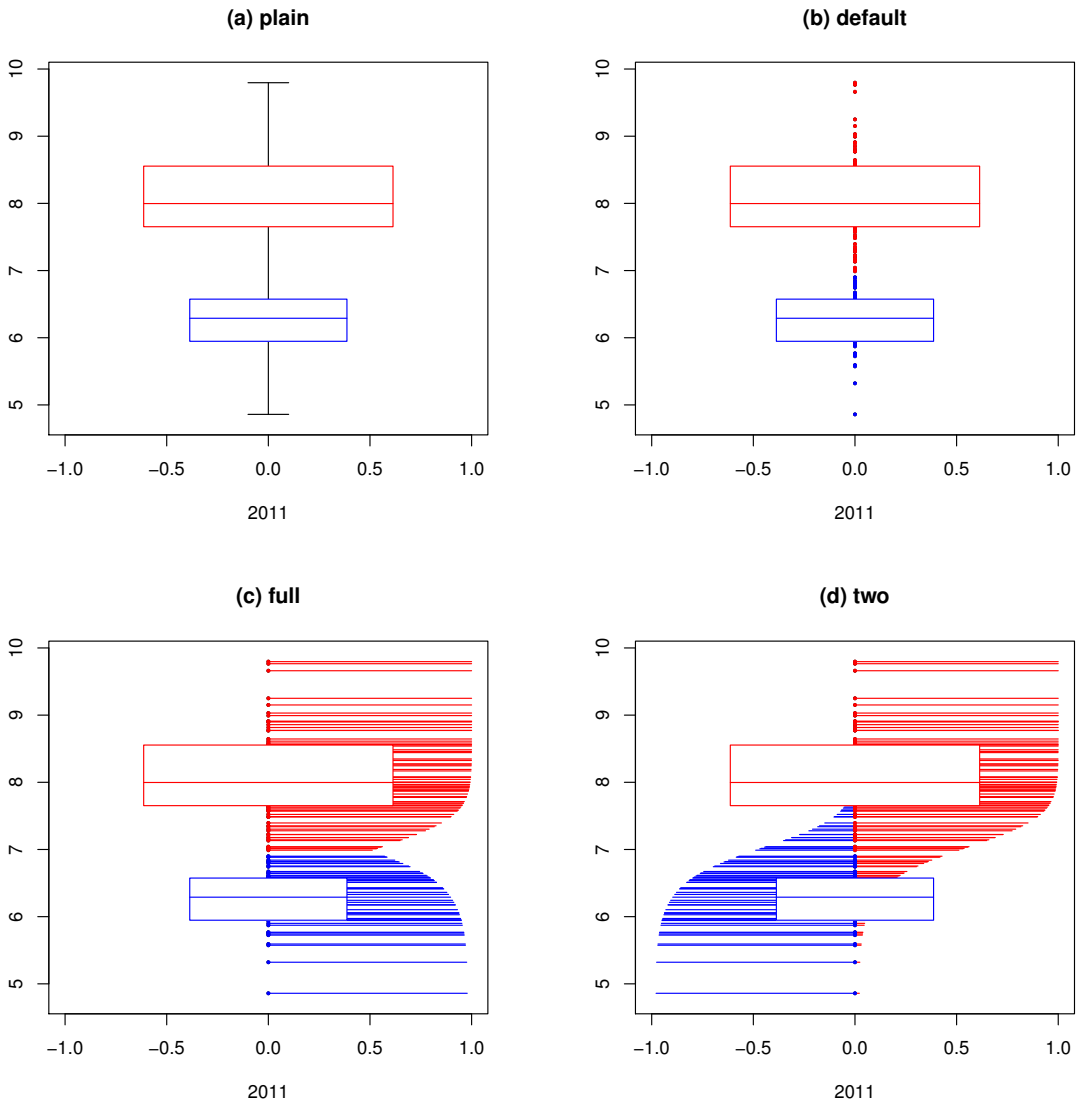

Figure 2.2: Four variants of 2–Boxplots of a log of energy use data in 2011.

different types of 2–boxplots for the log of energy use, the bimodal character of the country-wise log of energy data has already been reported in [19]. Figure 2.3 presents a histogram of log of energy use data in 2011 which visualises the distribution of mixture of two components. In fact, the likelihood ratio test in Section 1.4.2 of Chapter 1 was used to test the null hypothesis $H_0 : K = 1$ versus $H_1 : K = 2$. As a result, The p-value is 0.03 which suggests that the number of components K would be chosen to be equal to 2 at the 5% level of significance. In Figure 2.2, the 2–boxplots are labelled in the title area by the corresponding option, which needs to be specified as `type` argument in R function `kboxplot`. All four versions carry the main feature of a 2–boxplot, for example, the two boxes, which indicate the location, spread and size of

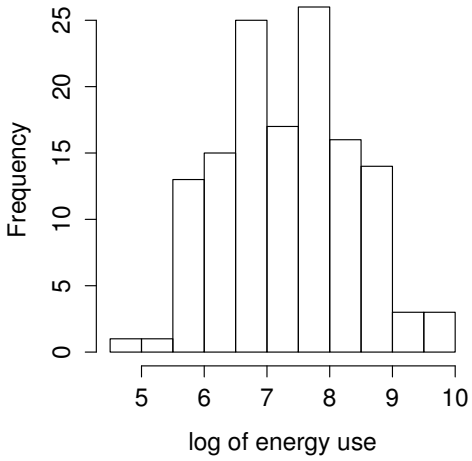

Figure 2.3: Histogram of a log of energy use data in 2011.

the two components. We can see that the blue box represents a group of low energy use countries and the red box visualises high energy use countries. One can observe from these figures, that the number of high energy use countries is higher than the number of low energy use countries, according to the widths of the boxes as determined by the fitted mixing proportions π_k, we can use the convention that the π_k correspond exactly to the half-width. Further, it is possible to obtain information on the spread and location of groups by observing the bottom, top and cut lines of the boxes, which represent the weighted first and third quartiles and the weighted medians, respectively.

The four types of K–boxplots differ in how individual observations are presented. The 'plain' version of the 2–boxplot shown in Figure 2.2 (a), most closely resembles a traditional boxplot in its simplest form: there are two boxes that represent the mixture components, with whiskers drawn-up to the overall maximum and minimum values. For K–boxplots, it is not considered a sensible option to draw-up the whiskers to a certain multiple of the interquartile range. The reason for this is, that this range would have to be calculated with respect to the corresponding top or bottom box, which would be little informative results, especially if the range of this box is small.

The 'default' option, see Figure 2.2 (b), provides slightly more information, that data points falling outside the boxes are plotted explicitly. Hence, making this representation

is particularly suitable to identify outlying cases. Furthermore, the points are coloured using the MAP classification rule. For example, for country i, it is necessary to identify the component k for, which the posterior probability r_{ik} is maximal, and then the colours of the points are the same colour as the box for that component. The 'full' version as shown in Figure 2.2 (c) provides another layer of detail, by giving explicitly the posterior probabilities of belonging to their component, to which they were assigned according to the MAP rule. The lines have a maximum length of 1, in which case a country is classified as having 100% posterior probability to one of the two groups. Finally, in Figure 2.2 (d), another variant is offered, which gives a full picture of all posterior probabilities, represented by lines of the length 1, which, are split-coloured around the ordinate axis, according to the values of r_{ik}, $k = 1, 2$. This variant is only supported for $K = 2$ as this produces presentational difficulties otherwise. Figure 2.4 [top] presents the boxplots of the log of energy use data of the countries in selected years between 1971 to 2011. The five main features of the boxplot are obvious for each year. The median of log of energy use data increased to the early 1990's. However, it should be noted that until 1989 only data for 112 countries were available, and that the sharp increase in 1991, and the subsequent decrease, can be explained by the inclusion of many new countries from 1991 onwards, after the fall of the iron curtain, and the subsequent political and economical developments of those countries previously belonging to the Soviet Union.

Overall, if we put the 1990 effect to one side, the boxplots show, that there has been a relatively steady increase of energy use throughout all countries over time. However, the sequence of the 2–boxplots of log of energy use data as shown in Figure 2.4 [bottom] reveals, that this interpretation would not be accurate. It can be seen that the data forms two groups, where one group corresponds to high energy use (supposedly so-called 'developed') countries, and one group corresponds to low energy use countries. The median as a measure of location almost changes slightly in either of the two groups, and this result appears to conflict with the information transmitted by the boxplots. However, what did change over time was that the low-energy-use group got much smaller, and the high-energy-use group became larger, this is represented by the boxes getting slimmer and wider, respectively. This can be interpreted as that, over the years, more and more countries have managed to make the transition from being a low to a

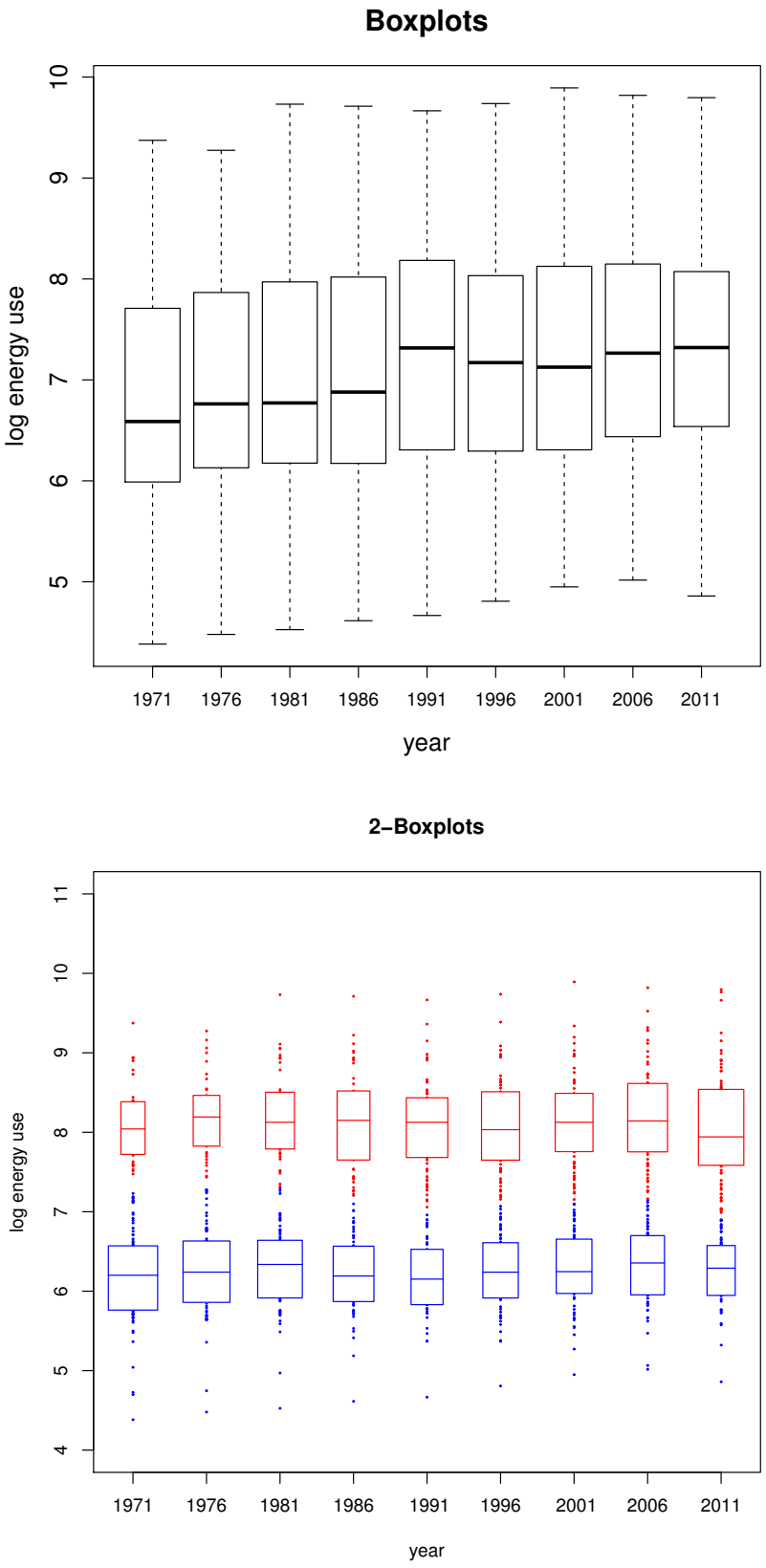

Figure 2.4: Boxplots [top] and 2–boxplots [bottom] of a log of energy use data between 1971 to 2011

high energy use country. This example demonstrates how misinterpretations based on the results of traditional boxplots can be avoided if the new graphical representation tool is used. This is because the tool takes in the account the mixture character of the data. However for completeness, it should be noted that, due to the non-linearity of the logarithm, the preceding analysis is not equivalent to fitting a mixture of log-normal distributions to the original data.

2.3.2 Example 2: internet users data

This example considers a data set of size $n = 100$, which was originally given in the form of a time series of the numbers of users who are connected to the internet through a server every minute. The data are available in the R package **datasets** under the name of `WWWusage` and visualised using a boxplot and a histogram, as shown in Figure 2.5. The histogram suggests, that distributions where either $K = 3$ or $K = 4$ may be adequate. If we first consider $K = 3$, it can be seen, that the 3–boxplots of the log(`WWWusage`) data uses a mixture of three normal distributions, where two different cases have been considered. In the first case, we allowed the components of the normal mixture to have unequal variances σ_k^2. In the second case, we assumed equal variances $\sigma_k^2 = \sigma^2, \forall k, k = 1, 2, 3$, in which case the second of the estimators as shown in (2.2.5) was to be adapted to become the following:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik}(y_i - \mu_k)^2$$

In Figure 2.6(a), the 3–boxplot of log(`WWWusage`) for the unequal variance case is presented. There are three boxes that represent three categories of the number of the internet users during different periods. It can be observed that the majority of the data falls into the central box, which represents the large majority of time points for which a medium number of internet users was observed. Additionally, two smaller clusters are observed that correspond to low and high internet usage, respectively. The 3–boxplots in the equal variance case is shown as in Figure 2.6(b). It can be seen that there is not much difference between the plots in this instance, though, expectedly, the spread of the smaller boxes for the equal variance case is slightly larger than for the

unequal variance case. All this information about the size and structure of clusters cannot be observed using a traditional boxplot. If we now proceed to the case $K = 4$, as shown in Figures 2.7 (a) and (b), it can be seen that 4–boxplots of log(`WWWusage`) is shown in the unequal and equal variance case, respectively. In comparison to the 3–boxplots, the boxes have been split differently as follows: in the unequal variance case (a), the low-usage box has been split, while in the equal variance case (b) the medium-usage box has been split. Furthermore, these 4–boxplots in their 'full' form, which allow insights into the MAP classification of data points to clusters, as well as the posterior probability of belonging to that cluster, as symbolized by the length of the horizontal line drawn to the right. As appreciable number of observations has been allocated to each cluster. If classification is the main purpose of the study, then this graphical information may be very useful.

In summary, the most suitable working assumption, in terms of the choice of K and the choice of equal or unequal component variances, will depend on the particular application. The point being made here is, that the impact of this choice on the fitted model may be quite large, and that the K–boxplots allow the data analyst to visualise the consequence of their choice at a glance, which will be helpful to support their decision process, on which model to choose. Therefore, a K–boxplot is a tool that can be used to visualise the different clusters in mixture data, however it is not an inference

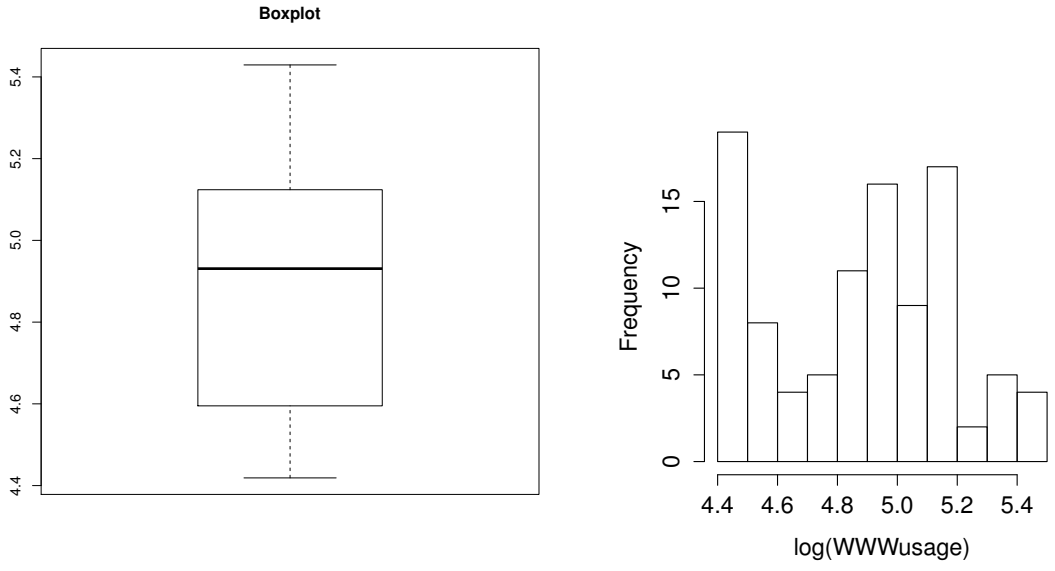

Figure 2.5: Boxplot and histogram of a log of the numbers of internet users

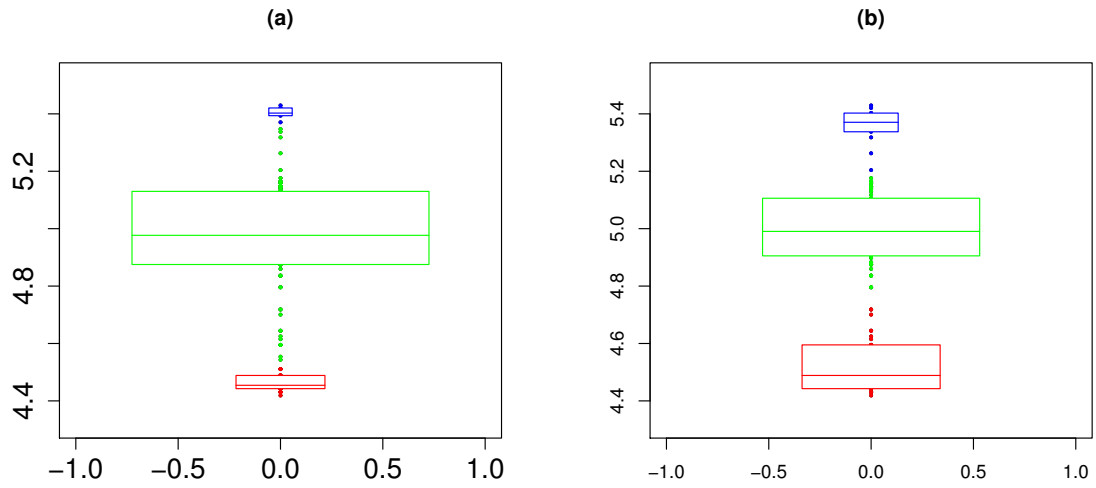

Figure 2.6: 3–Boxplots of a log of the numbers of internet users; (a) with unequal variances, (b) with equal variances

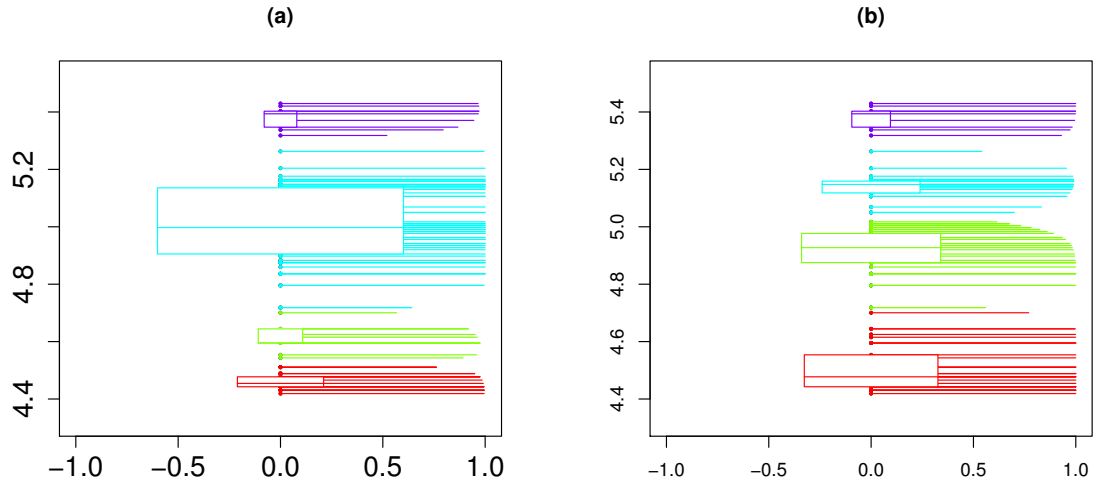

Figure 2.7: 4–Boxplots of a log of the numbers of internet users; (a) with unequal variances, (b) with equal variances

method in itself. Consequently, as for any other graphical tool, the data analyst should not solely rely on a K–boxplot to determine the distribution of data.

2.3.3 Simulation

In order to obtain insight into the behaviour of the K–boxplots under the use of component distributions other than Gaussian, and, in particular, under component

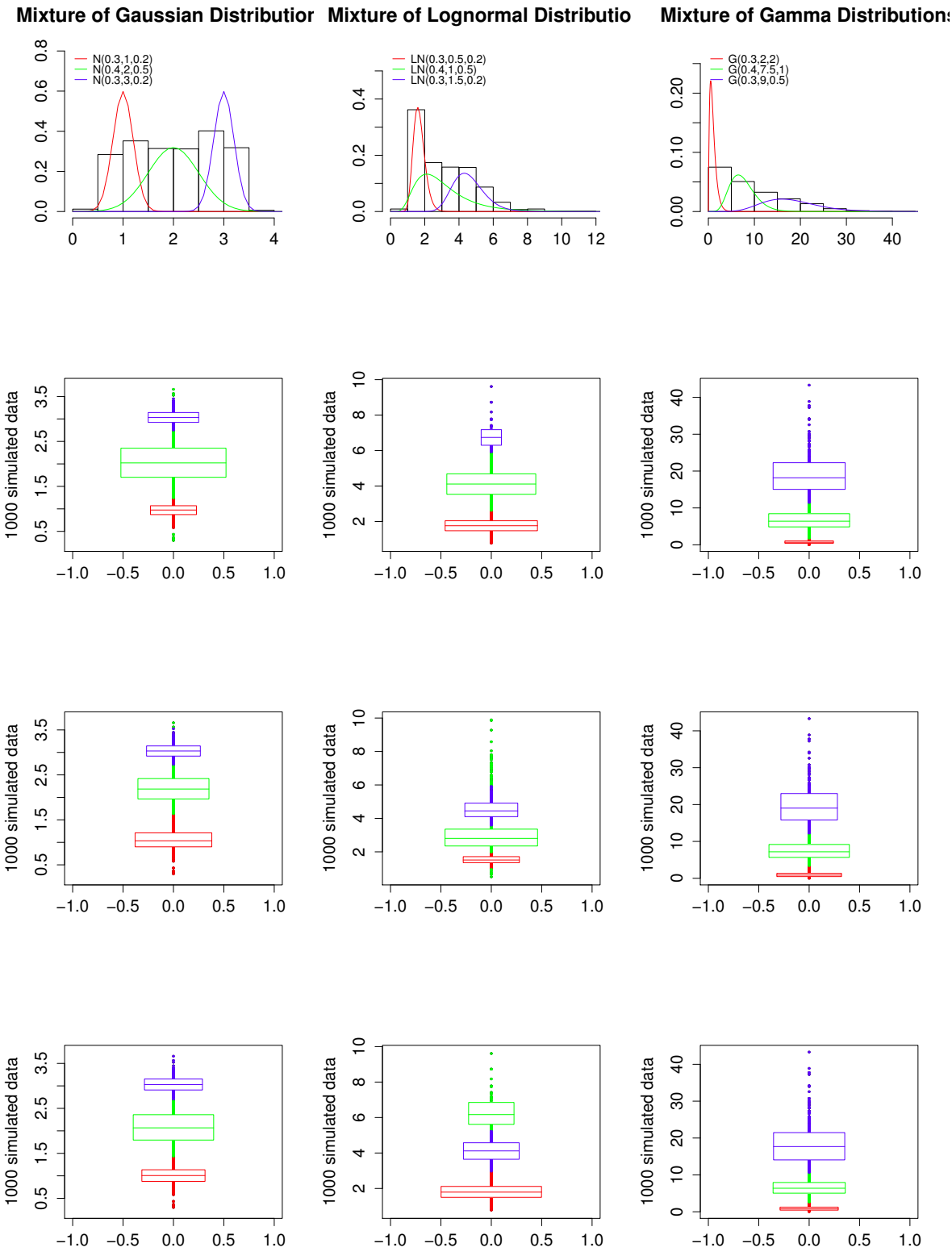

Figure 2.8: 3–Boxplots simulated from scenarios (a), (b), (c) [from left to right] and fitted using Gaussian, log-normal, and Gamma component densities [from top to bottom]

misspecification, we have carried out a small-scale simulation whereby data sets were simulated in three scenarios. Under all three simulated scenarios, we used $K = 3$, $\pi_1 = 0.3$, $\pi_2 = 0.4$ and $\pi_3 = 0.3$, but the component densities differed as follows:
(a) a mixture of three Gaussian component densities with $\mu_k = k$, $\sigma_1 = \sigma_3 = 0.2$ and $\sigma_2 = 0.5$;
(b) a mixture of three log-normal densities with $\mu_k = k/2$, and $\sigma_k, k = 1, 2, 3$ as in (a);
(c) a mixture of three Gamma densities with shape parameters 2,7.5,9 and scale parameters 2,1,0.5, respectively.
The true underlying densities can be seen in the top row of Figure 2.8 along with histograms of the simulated data sets. The panels below the top row show 3–boxplots fitted to the simulated data using a mixture of three Gaussian distributions, log-normal distributions and Gamma distributions, respectively. That is, the component distributions are correctly specified along the diagonal of the 3×3 panel of the 3–boxplots, but they are mis-specified off the diagonal.

The main conclusions drawn from Figure 2.8 are, that: (i) the mixture proportions were in the most cases approximately correctly captured; (ii) if the data is simulated from Gaussian components, shown in the first column, then the 3–boxplots are quite robust to component misspecification; (iii) in the bottom right 2×2 panel, we see that the skewness of the original distribution was correctly represented by the fitted distribution; (iv) if a Gaussian mixture is fitted to the 'true' log–normal or Gamma components, then the tail component tends to carry too much weight.

2.4 Conclusions

This chapter has presented a new powerful graphical tool, that can be used to visualise and analyse data stemming from a mixture of K distributions, and this has been called the K–boxplot. This plot can be used to visualise the different K groups of a mixture data, that a boxplot is not able to do. This tool is a useful extension of the traditional boxplot, and can especially be used to find out additional information about the location and spread of individual groups in mixture data, that is not visualised when using a traditional boxplot. However, in a similar way to a traditional boxplot,

a K–boxplot can visualise outliers in the data. It must be noted that the K–boxplot cannot be classed as an inference method, that can make automated decisions about the distribution or the number of components in mixture data. However, it is a useful tool to support the data analyst in this respect. For instance, overlapping or very small boxes may be a sign, that the number of components should be reduced, or long one-sided tails seen outside the boxes may be a sign, that the Gaussian component densities are not adequate.

K–boxplots can be implemented using the function `kboxplot`, that is available as part of the R package **UEM** [18]. The implemented R subroutine provides several graphical options for the data analyst, including a black and white option. There are two ways, in which this function can be used. The first option is to apply `kboxplot` directly onto the data itself, in which case the model will be fitted implicitly. The alternative option, which is the recommended option as it gives better control over the process, is to apply `kboxplot` onto a previously fitted model, for which the subroutines provided within the R package **UEM** could be used. However, also functions from alternative R packages, or even alternative software, may be considered for this purpose, as long as they provide access to the weight matrix R.

Taking into account the matrix R, the computational complexity of producing a K–boxplot is of the order $O(nK)$ as compared to $O(n)$ for a traditional boxplot. For all data sets, choices of K, and graphical variants considered in this chapter, the computational time to produce a K–boxplot, taking R into account, was less than 0.02 seconds on an Intel® Core(TM) i7–4790 CPU @ 3.60GHz machine. The computations required for the underlying inferential mechanism will usually contribute to a larger computational burden. For example, for Example 2, which has been computed using the EM routines built into R package **UEM**, this computation required 0.11 seconds for the (unequal variance) 3–component model (28 EM iterations), and 0.26 seconds for the 4–component model (52 EM iterations), using Gaussian quadrature points as starting points in each case. It should be noted that the R code used to reproduce the examples presented in this thesis is shown in the R Documentation files of R package **UEM**.

Finally, one issue that was given only marginal attention was the selection of the

number of components, K. This problem was inherent to the mixture fitting technique, and while there does exist a rich literature on suggested methods how to select this number K, this question is eventually still down to the subjective judgement by the data analyst, see Section 1.4.2 of Chapter 1 for more detail. In order to arrive at this judgement, the data analyst will undoubtedly benefit from using the simple graphical tool, as the proposed one, which can visualise the structure of the mixture model, obtained under the hypothesised number K at a glance.

Chapter 3

Localised mixture models for prediction

3.1 Introduction

This chapter explores how localised mixture of regression models can be used to produce predictions from time series data. For this purpose, estimation for these models is achieved using a kernel-weighted version of the EM–algorithm, using exponential kernels with different bandwidths as weight functions. Nadaraya-Watson and local linear estimators are used to carry out the localised estimation step, see Section 1.2 of Chapter 1 for more details about these estimators.

Using the first proposed model, forecasts can be calculated directly using historical data comprising the locally average of observed past values: the size of the local neighbourhood and the specific weights of the values are defined by an exponential kernel. Using the second model, forecasts are based on using a fitted intercept and slope for the local neighbourhood preceding the forecast point. These two models will be referred to as a mixture model using local constant estimators (MLC) and as a mixture model using local linear estimators (MLL), respectively. By modelling MLC and MLL at a target time point t_T, but with different bandwidths h_k, $k = 1, \ldots, K$, where K is the number of components in a mixture, it is possible to estimate a mixture of probabilities, that are informative using the amount of information available in the data set at

the scale of resolution corresponding to each bandwidth. For prediction at the time point t_{T+m} where m is the number of forward lag, adequate approaches are provided for each local method, and then compared to competing forecasting routines. Some of these approaches are discussed and applied using real data in [68].

Further consideration is given to optimal bandwidth choices used for forecasting. A new approach for bandwidth selection for forecasting is proposed in Section 3.7. This approach was applied using real and simulated data, in order to produce accurate predictions in Sections 3.8 and 3.9. At the end of this chapter, a simulation study is presented, that assesses the accuracy of the forecasting using MLC and MLL models. In addition, a comparison is presented between the new methods based on using MLC and MLL for prediction and traditional methods, that employ the ARIMA and Holt models, which are popular approaches used for time series forecasting.

This chapter is organised as follows: firstly, it presents the MLC model, and estimation relating to this model can be found in Section 3.2. Section 3.3, presents the MLL model and a guide about how to estimate the parameters for this model is included in the description. Sections 3.4 and 3.5 discuss identifiability property and model selection strategies for the MLC and MLL models. In addition, Section 3.6 will outline how these models can be used for predictions under consideration of bandwidth selection. Section 3.7 outlines the methodology used for a simulation study, while Section 3.8 presents the simulation study used to assess the performance of the MLC and MLL models for prediction in comparison to other traditional models, such as the ARIMA and Holt models. Section 3.9, uses data for energy use for Bolivia, Lebanon and Greece from 1971 to 2011, and compares results to point forecasts obtained using the ARIMA and Holt exponential smoothing models. Finally, conclusions are presented in Section 3.10.

3.2 Mixture models using local constant kernel estimators (MLC)

For a time series of the form $\{(t_i, y_i) : i = 1, \ldots, T\}$, a localised mixture of K non-parametric regressions $m_k(t_i)$, $k = 1, \ldots, K$ was considered. At a time point t_T, it is possible to define a locally constant model $m_k(t_i) \approx m_k(t_T)$ in a neighbourhood of t_T

by using Taylor's expansion as discussed in Section 1.2 of Chapter 1, where the $m_k(t_T)$ play the role of parameters and are denoted as $\beta_k(t_T)$. Effectively, the $m_k(t_T)$ were estimated using component-wise Nadaraya-Watson estimators. Then, the model can be locally defined:

$$y_i = \begin{cases} \beta_1(t_T) + \epsilon_{i1}, & \text{with probability } \pi_1(t_T) \\ \vdots & \\ \beta_K(t_T) + \epsilon_{iK}, & \text{with probability } \pi_K(t_T) \end{cases} \tag{3.2.1}$$

where $\beta_1(t_T), \ldots, \beta_K(t_T)$ are unknown constants, $\pi_k(t_T)$ is the proportion of the k-th component, such that $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$, and the errors $\epsilon_{ik} \sim N(0, \sigma^2)$ are independently distributed. Note that both $\beta_k(t_T)$ and $\pi_k(t_T), k = 1, \ldots, K$ are functions of the time point t_T. When $K = 1$, then model (3.2.1) is a simple non-parametric regression model, that can be defined as Equation (1.2.2) as outlined in Chapter 1. In this case, the model is non-parametric local constant regression model and is denoted as NLC. For ease of notation, we will often suppress dependence of the parameters on t_T.

For the component k, it is necessary to obtain estimators of π_k, β_k and σ at time t_T. In the estimation step, the EM–algorithm was used, see Section 1.4.1 of Chapter 1. Therefore, let G be a random vector, which is defined as in Equation (2.2.2) of Chapter 2. Then, we have $P(G = k) = \pi_k$ and we denote

$$f_{ik} = P(y_i|G = k) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_k)^2}{2\sigma^2}\right)$$

then

$$\begin{aligned} P(y_i, G = k) &= P(y_i|G = k)P(G = k) \\ &= \pi_k f_{ik} \end{aligned}$$

One-sided component-wise weight functions W_k are anchored at t_T and can be introduced as follows:

$$W_k(t_i, t_T) = \begin{cases} \frac{\exp\left(\frac{t_i - t_T}{h_k}\right)}{h_k} & t_i - t_T \leq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.2.2}$$

The exponential kernel W_k is a popular kernel for prediction in local regression models

[34]. For example, it is used in the EWMA forecasts, see Section 1.5 of Chapter 1. Therefore, we assume now that, for an observation y_i, the value of G is known. For example, we know to which of the K components the i-th observation belongs. This gives the "complete" data $(y_i, G_{i1}, \ldots, G_{iK})$, $i = 1, \ldots, n$, with local probability as follows:

$$P(y_i, G_{i1}, \ldots, G_{iK}) = \prod_{k=1}^{K} (f_{ik}\pi_k)^{G_{ik}W_k(t_i,t_T)}$$

Then, the corresponding local likelihood function $\mathcal{L}^*$, which is called complete local likelihood [62], is as follows:

$$\mathcal{L}^*(\mathbf{\Phi}_c|y_1, \ldots, y_T, G_{i1}, \ldots, G_{iK}) = \prod_{i=1}^{T}\prod_{k=1}^{K} (f_{ik}\pi_k)^{G_{ik}W_k(t_i,t_T)}$$

where $\mathbf{\Phi}_c = (\pi_1, \ldots, \pi_{K-1}, \beta_1, \ldots, \beta_K, \sigma)$ is a vector containing all the parameters in the mixture model MLC. Therefore, the log local likelihood function ℓ^* is as follows:

$$\begin{aligned}\ell^*(\mathbf{\Phi}_c|y_1, \ldots, y_T, G_{i1}, \ldots, G_{iK}) &= \log \mathcal{L}^*(\mathbf{\Phi}_c|y_1, \ldots, y_T, G_{i1}, \ldots, G_{iK}) \\ &= \sum_{i=1}^{T}\sum_{k=1}^{K} G_{ik}W_k(t_i,t_T)\log\pi_k + G_{ik}W_k(t_i,t_T)\log f_{ik}\end{aligned}$$

If we interpret the π_k as a 'prior' probability of class membership, then the posterior probabilities of class membership can be produced using Bayes' theorem. Because the G_{ik} are, in fact unknown, we can replace them by their conditional expectations, as follows:

$$r_{ik} = E(G_{ik}|y_i) = P(G_{ik} = 1|y_i) = P(G = k|y_i)$$

If we use Bayes' theorem, we can see the following:

$$r_{ik} = P(G = k|y_i) = \frac{P(G = k)P(y_i|G = k)}{\sum_\ell P(G = \ell)P(y_i|G = \ell)} = \frac{\pi_k f_{ik}}{\sum_\ell \pi_\ell f_{i\ell}}$$

This equates as follows:

$$r_{ik} = P(G_{ik} = 1|y_i) = \frac{\pi_k f_k(y_i)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(y_i)} \tag{3.2.3}$$

Equation (3.2.3) is identical to the E–step of the EM–algorithm. In the l-th cycle of the EM–algorithm iteration, we have the estimates $\pi_k^{(l)}$, $\beta_k^{(l)}$ and $\sigma^{(l)}$. Then, in the $(l+1)$-th cycle, using the estimates $\pi_k^{(l)}$, $\beta_k^{(l)}$ and $\sigma^{(l)}$, the posterior probabilities $r_{ik}^{(l+1)}$

can then be given as follows:

$$r_{ik}^{(l+1)} = \frac{\pi_k^{(l)} \exp\left(-\frac{1}{2}(\frac{y_i-\beta_k^{(l)}}{\sigma^{(l)}})^2\right)}{\sum_{\ell=1}^{K} \pi_\ell^{(l)} \exp\left(-\frac{1}{2}(\frac{y_i-\beta_\ell^{(l)}}{\sigma^{(l)}})^2\right)}$$

In the M–step, for the $\pi_k^{(l+1)}$, one needs to apply a Lagrange multiplier, since $\sum_{k=1}^{K} \pi_k^{(l+1)} = 1$ by setting

$$\partial\left(Q(\mathbf{\Phi}_c|\mathbf{\Phi}_c^{(l)}) - \lambda(\sum_{k=1}^{K} \pi_k^{(l+1)} - 1)\right)/\partial\pi_k^{(l+1)} = 0, \quad k = 1, \ldots, K$$

Thus it is possible to obtain the following:

$$\pi_k^{(l+1)} = \frac{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)}{\sum_{i=1}^{T}\sum_{k=1}^{K} r_{ik}^{(l+1)} W_k(t_i, t_T)} \tag{3.2.4}$$

In addition, by setting $\partial\ell^*/\partial\beta_k^{(l+1)} = 0$ and $\partial\ell^*/\partial\sigma^{(l+1)} = 0$, as the estimates, we can see the following:

$$\beta_k^{(l+1)} = \frac{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T) y_i}{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)} \tag{3.2.5}$$

and

$$\sigma^{2(l+1)} = \frac{\sum_{i=1}^{T}\sum_{k=1}^{K} r_{ik}^{(l+1)} W_k(t_i, t_T)(y_i - \beta_k^{(l+1)})^2}{\sum_{i=1}^{T}\sum_{k=1}^{K} r_{ik}^{(l+1)} W_k(t_i, t_T)} \tag{3.2.6}$$

For more details about the calculations of Equations (3.2.5) and (3.2.6), see Appendix C.

This iteratively updates the E–step and M–step using different initial values at each grid point locally, until the algorithm converges. In this study, the stopping criterion of the EM–algorithm depends on the size of the relative change in the parameter estimates which is approximately 0. In other wards, the iteratively updates the E–step and M–step locally until $\hat{\mathbf{\Phi}}_c^{(l)} \approx \hat{\mathbf{\Phi}}_c^{(l+i)}, i = 1, 2, \ldots$. According to this stopping criterion, the EM–algorithm stopped at $l = 200$ in this study. As a result, the kernel-weighted version of the EM–algorithm estimates π_k, β_k and σ for each component k and for given time point t_T. Once the estimates of π_k, β_k and σ are obtained using Equations (3.2.4)–(3.2.6), then different approaches for forecasting were considered to find the m-step-ahead forecasts at the given time point t_T in this chapter. The size of the local neighbourhood plays a role in both prediction and estimation. The optimal bandwidth

for prediction will be discussed in more detail later in Section 3.7 of this chapter.

3.3 Mixture models using local linear kernel estimators (MLL)

In this section, the model MLC was generalised using local linear kernel estimators rather than local constant estimators, in order to carry out the localised estimation step, this model named MLL. The motivation for using local linear estimators was to improve prediction from a time series points that have a linear trend. Local linear estimators possess favourable asymptotic bias properties and a high ability to control boundary effects in comparison to local constant estimators [23]. This feature of local linear estimator is important, in order to obtain accurate predictions, since the model MLL is, by definition, applied at a boundary target time point t_T. For more details about the properties of local linear estimators, see Section 1.2.2 in Chapter 1. The k-th non-parametric regression function around the time point t_T can be approximated as $m_k(t_i) \approx m_k(t_T) + m_k^{(1)}(t_T)(t_i - t_T)$. This motivates the localised model as follows:

$$y_i = \begin{cases} \beta_{01}(t_T) + \beta_{11}(t_T)(t_i - t_T) + \epsilon_{i1}, & \text{with probability } \pi_1(t_T) \\ \vdots & \\ \beta_{0K}(t_T) + \beta_{1K}(t_T)(t_i - t_T) + \epsilon_{iK}, & \text{with probability } \pi_K(t_T) \end{cases} \tag{3.3.1}$$

where the intercepts β_{0k} and the slopes β_{1k} are fixed unknown coefficients, which depend implicitly on a fixed time t_T. The errors $\epsilon_{ik} \sim N(0, \sigma^2)$ are independently distributed. When $K = 1$, the model (3.3.1) is a non-parametric linear regression model, and this is explained in more detail in Section 1.2.2 of Chapter 1. In this case, we can refer to this model as non-parametric local linear regression model (NLL).

For the given t_T, the data is weighted by exponential kernels W_k for each component, which is defined as in Equation (3.2.2). In the estimation step, the EM–algorithm is used to estimate $\mathbf{\Phi}_l = (\pi_1, \ldots, \pi_{K-1}, \beta_{01}, \ldots, \beta_{0K}, \beta_{11}, \ldots, \beta_{1K}, \sigma)$, which is a vector containing all the parameters in the mixture model MLL. Let G be a random vector, which is defined as in Equation (2.2.2) of Chapter 2. Then, we have $P(G = k) = \pi_k$

and we denote

$$f_{ik} \equiv P(y_i|G=k) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_{0k} - \beta_{1k}(t_i - t_T))^2}{2\sigma^2}\right)$$

Then

$$\begin{aligned} P(y_i, G=k) &= P(y_i|G=k)P(G=k) \\ &= \pi_k f_{ik} \end{aligned}$$

Therefore, we assume now that, for an observation y_i, the value of G is known. This gives the "complete" data $(y_i, G_{i1}, \ldots, G_{iK})$, $i = 1, \ldots, n$, with local probability as follows:

$$P(y_i, G_{i1}, \ldots, G_{iK}) = \prod_{k=1}^{K} (f_{ik}\pi_k)^{G_{ik}W_k(t_i,t_T)}$$

Then, the corresponding local likelihood function $\mathcal{L}^*$, which is called complete local likelihood [62], is as follows:

$$\mathcal{L}^*(\mathbf{\Phi}_l|y_1, \ldots, y_T, G_{i1}, \ldots, G_{iK}) = \prod_{i=1}^{T}\prod_{k=1}^{K} (f_{ik}\pi_k)^{G_{ik}W_k(t_i,t_T)}$$

Therefore, the log local likelihood function ℓ^* is as follows:

$$\begin{aligned} \ell^*(\mathbf{\Phi}_l|y_1, \ldots, y_T, G_{i1}, \ldots, G_{iK}) &= \log \mathcal{L}^*(\mathbf{\Phi}_l|y_1, \ldots, y_T, G_{i1}, \ldots, G_{iK}) \\ &= \sum_{i=1}^{T}\sum_{k=1}^{K} G_{ik}W_k(t_i,t_T)\log\pi_k + G_{ik}W_k(t_i,t_T)\log f_{ik} \end{aligned}$$

As the G_{ik} are in fact unknown, we replace them by their conditional expectations as follows

$$r_{ik} \equiv E(G_{ik}|y_i) = P(G_{ik}=1|y_i) = P(G=k|y_i)$$

Using Bayes' theorem, one has

$$r_{ik} = P(G=k|y_i) = \frac{P(G=k)P(y_i|G=k)}{\sum_\ell P(G=\ell)P(y_i|G=\ell)} = \frac{\pi_k f_i}{\sum_\ell \pi_\ell f_{i\ell}}$$

which is equivalent to the posterior probabilities in the E–step. Then, in the $(l+1)$-th cycle of the EM–algorithm, using the estimates $\pi_k^{(l)}$, $\beta_{0k}^{(l)}$, $\beta_{1k}^{(l)}$ and $\sigma^{(l)}$, the posterior

probabilities $r_{ik}^{(l+1)}$ can then be given by undertaking the following:

$$r_{ik}^{(l+1)} = \frac{\pi_k^{(l)} \exp\left(-\frac{1}{2}\left(\frac{y_i-\beta_{0k}^{(l)}-\beta_{1k}^{(l)}(t_i-t_T)}{\sigma^{(l)}}\right)^2\right)}{\sum_{\ell=1}^{K} \pi_\ell^{(l)} \exp\left(-\frac{1}{2}\left(\frac{y_i-\beta_{0\ell}^{(l)}-\beta_{1\ell}^{(l)}(t_i-t_T)}{\sigma^{(l)}}\right)^2\right)} \tag{3.3.2}$$

For the M–step, the estimates of $\pi_k^{(l+1)}$, $\beta_{0k}^{(l+1)}$, $\beta_{1k}^{(l+1)}$ and $\sigma^{(l+1)}$ are found as follows: One needs to apply a Lagrange multiplier for $\pi_k^{(l+1)}$ since $\sum_{k=1}^{K} \pi_k^{(l+1)} = 1$. Setting

$$\partial\left(Q(\mathbf{\Phi}_l|\mathbf{\Phi}_l^{(l)}) - \lambda(\sum_{k=1}^{K} \pi_k^{(l+1)} - 1)\right)/\partial\pi_k = 0, \quad k = 1, \ldots, K$$

One obtains

$$\pi_k^{(l+1)} = \frac{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)}{\sum_{i=1}^{T}\sum_{k=1}^{K} r_{ik}^{(l+1)} W_k(t_i, t_T)} \tag{3.3.3}$$

In addition, the estimates of $\beta_{0k}^{(l+1)}$, $\beta_{1k}^{(l+1)}$ and $\sigma^{(l+1)}$ are as follows:

$$\beta_{0k}^{(l+1)} = \frac{S_{k,T,2}S_{k,T,0}^{*} - S_{k,T,1}S_{k,T,1}^{*}}{S_{k,T,2}S_{k,T,0} - S_{k,T,1}^2}, \qquad \beta_{1k}^{(l+1)} = \frac{S_{k,T,0}S_{k,T,1}^{*} - S_{k,T,1}S_{k,T,0}^{*}}{S_{k,T,2}S_{k,T,0} - S_{k,T,1}^2}, \tag{3.3.4}$$

where $S_{k,T,j} = \sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T)^j$ and $S_{k,T,j}^{*} = \sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T)^j y_i$, and

$$\sigma^{2(l+1)} = \frac{\sum_{i=1}^{T}\sum_{k=1}^{K} r_{ik}^{(l+1)} W_k(t_i, t_T)(y_i - \beta_{0k}^{(l+1)} - \beta_{1k}^{(l+1)}(t_i - t_T))^2}{\sum_{i=1}^{T}\sum_{k=1}^{K} r_{ik}^{(l+1)} W_k(t_i, t_T)} \tag{3.3.5}$$

For more details about the calculations of Equations (3.3.4) and (3.3.5), see Appendix C.

The kernel-weighted version of the EM–algorithm uses the iteration of Equations (3.3.2)–(3.3.5) until convergence occurs. In addition, the stopping criterion of the EM–algorithm used is the same as the stopping criterion used for the MLC model in Section 3.2. Once the bandwidths h_k are set, and the estimates in Equations (3.3.3)–(3.3.5) are found, the localised model for prediction can be used to predict a future observation. This approach will be presented later in Section 3.6 of this chapter.

3.4 Identifiability

Identifiability is an important issue when considering mixture models. This property of mixture models must be investigated before exploring the specific problems

relating to estimation, testing of hypotheses, classification of random variables, and so forth. The concept of identifiability for different classes of finite mixture models is discussed by Titterington and others [80]. In addition, the identifiability of finite mixtures of regression models is studied by Hennig [41] and by Frühwirth-Schnatter [31]. The identifiability for mixture models is defined as follows: if we suppose that $f(y \mid \mathbf{\Phi}) = \sum_{k=1}^{K} \pi_k f_k(y \mid \boldsymbol{\beta}_k)$ and $f(y \mid \mathbf{\Phi}^*) = \sum_{k=1}^{K^*} \pi_k^* f_k(y \mid \boldsymbol{\beta}_k^*)$ are any two members of parametric family of mixture densities. This class of finite mixtures is identifiable for $\mathbf{\Phi} \in \mathbf{\Omega}$ where $\mathbf{\Omega}$ is the specified parameter space if

$$f(y \mid \mathbf{\Phi}) \equiv f(y \mid \mathbf{\Phi}^*)$$

if and only if $K = K^*$ and we can permute the component labels. Then $\pi_k = \pi_k^*$ and $f_k(y \mid \boldsymbol{\beta}_k) = f_k(y \mid \boldsymbol{\beta}_k^*)$, $k = 1, \ldots, K$ [62].

More recently, Huang et al. [44] proposed a class of non-parametric mixture of regression models, where the mixing proportions $\pi_k(t)$, and the mean functions $m_k(t)$, and the variance functions $\sigma_k^2(t)$ are all non-parametric, as defined by:

$$Y|T = t \sim \sum_{k=1}^{K} \pi_k(t) N\{m_k(t), \sigma_k^2(t)\} \tag{3.4.1}$$

Huang et al. [44] show the identifiability of the model (3.4.1) under certain conditions. One of these conditions is transversality of any two smooth curves $a(t) = (m_i(t), \sigma_i^2(t))$ and $b(t) = (m_j(t), \sigma_j^2(t)), i \neq j$. The transversality of any two smooth curves implies, that if $a(t) = b(t)$ then $a'(t) \neq b'(t)$. Then, the mean and variance functions of any two components cannot be tangent to each other. Huang et al. [44] present a theorem as follows:

Theorem [44]

If it is supposed that:

(i) $\pi_k(t) > 0$ are continuous functions, and $m_k(t)$ and $\sigma_k^2(t)$ are differentiable functions, $k = 1, \ldots, K$;

(ii) any two curves $(m_i(t), \sigma_i^2(t))$ and $(m_j(t), \sigma_j^2(t))$, $i \neq j$, are transversal;

(iii) the range $\mathcal{T}$ of t is an interval in $\mathbb{R}$.

then model (3.4.1) is identifiable [44].

It is possible to argue that the MLL model is identifiable. The different bandwidths used automatically imply different degrees of smoothness and different slopes. In other words, if $a(t) = (m_1(t), \sigma^2(t))$ and $b(t) = (m_2(t), \sigma^2(t))$, are two solutions at a time point t, such that $a(t) = b(t)$, this means $m_1(t) = m_2(t)$. However, since $h_1 \neq h_2$, this implies $m_1'(t) \neq m_2'(t)$. Hence, we have established the transversality, and assuming with the above concerns, that the above theorem holds in our context. It is possible to conclude that the MLC model is identifiable based on the theorem because it is a version of model (3.4.1), and when using different bandwidths for each component it is possible to make the local components of these models recognisable at each data point t_T. However, for the MLC model, further research should be considered in relation to the identifiability for more complex structures as in model (3.4.1).

Huang et al. [44] noted that for more complex structures of the model (3.4.1), the identifiability problem needs further consideration. For example, if we suppose that the mean functions $m_1(t)$ and $m_2(t)$ of a two-components structure are crossed at a point t. If the variance functions of the two components are the same, then there are two solutions of mean functions, such that $m_1(t)$ and $m_2(t)$ are tangent to each other. Now, if we assume that $m_1(t)$ is a monotone decreasing mean and the $m_2(t)$ is a monotone increasing mean. Then, the problem in this case is, that it is not clear, which paths the mean functions will follow without knowing the second derivatives of the mean functions at the "cross". Then, condition (ii) of the theorem does not hold in this case. As a result, the investigation of condition (ii) of the theorem should be subject to further research in this case, especially for the MLC and model (3.4.1), which are based on using local constant kernels. From the above discussion and previous theorem, we conclude the following:

Corollary

The MLC and MLL models are identifiable unless the bandwidths $h_i, i = 1, \ldots, K$, are identical.

It can be seen that MLC and MLL models are closely related to model (3.4.1) except, for the conditions on some parameters and the estimation method. Huang et al. [44] assumed that the parameters $m_k(t)$, $\pi_k(t)$ and $\sigma_k^2(t)$ in the model (3.4.1) are smooth functions. While this is a reasonable assumption in the MLC and MLL models too, it cannot be strictly guaranteed, since the allocations to components, using r_{ik}, can change abruptly and can produce discontinuities. The variances in models (3.2.1) and (3.3.1) can be classed as equal, namely $\sigma_k^2 = \sigma^2, \forall\ k = 1, \ldots, K$, which contributes to reducing the complexity of statistical analysis of the MLC and MLL models. The standard deviation is a noise term, which reflects the variability in the data unexplained by the mixture.

Moreover, a difference between the MLC and MLL models and model (3.4.1) is the equality of the bandwidths h_k in the model (3.4.1), but this is not the case in models (3.2.1) and (3.3.1). The use of different bandwidths for the MLC and MLL models plays an important role in localised estimation, by weighting the data inside different local neighbourhoods around a target point t_T. This technique allows us to control the amount of information used from past observations, in order to fit short and long-term linear trends of data. In addition, the use of different bandwidths is useful to avoid the label switching problem, which is crucial problem in mixture models.

The second main difference between the MLC and MLL models and model (3.4.1) is an estimation method used. The initial values used for the EM–algorthim are produced globally in model (3.4.1). For initialisation, Huang et al. [44] conducted a mixture of polynomial regressions with the constant proportions of π_k and variances of σ_k^2. Then, the estimates of the mean functions $m_k(t)$, and parameters $\pi_k(t)$ and σ_k^2 are obtained globally. These estimates are classed as initial values for the first iteration of the EM–algorithm. However, in the MLC and MLL models, we set different initial values locally at each grid points. In addition, Huang et al. [44] used local constant estimators in kernel regression to estimate the parameters.

Moreover, in model (3.4.1), the posterior probabilities in the E–step are calculated globally by estimating component's label curves for each of the observations. This means that the E–step does not depend on the location of observations. After that, for the M–step, the estimators are calculated locally at each grid points for the same

probabilistic label obtained in the E–step. Then, the estimators at each grid points are used to find the global estimators using linear interpolation. However, in the MLC and MLL models, the E–step and M–step for the EM–algorithm are conducted locally at each grid point without doing any further linear interpolation. As a result, the EM–algorithm method for the MLC and MLL models depends on the location of a time point t_T.

3.5 Model selection

Model selection for the MLC and MLL models includes the selection of the number of components K and the bandwidths h_k. Choosing the number of components is a very important issue when using mixture models. In the literature, many approaches have been suggested for choosing an adequate number of components, see Section 1.4.2 of Chapter 1. However, the selection of the number of components K remains a controversial issue and is becoming increasingly difficult when using complicated mixture models [44]. In this chapter, the number of components was generally fixed at 2. In addition, one of the components was a fixed bandwidth, $h_1 = 1$, and the second bandwidth h_2 of the second component was optimised in some cases and fixed in others. The reasons for these restrictions on the number of components and on the bandwidth selection are to reduce the expense of computational cost. For example, the computational time to predict a future observation in simulation, taking these restrictions into account, was 48 hours for MLC model on an Intel® Core(TM) i7–4790 CPU @ 3.60GHz machine. In addition, for the MLL model which has more parameters than the MLC model, the computational time was approximately a week. See Section 3.8 for more detail.

Bandwidth selection has been considered in the literature as shown in Section 1.2.2 Chapter 1. However, to the best of our knowledge, there is no statistical method for finding the best bandwidth for forecasting using localised mixture models, even for $K = 1$. As a result, in this chapter, a new methodology is proposed to select the optimal bandwidth for prediction. More details about this new criterion of bandwidth selection will be explained later in Section 3.7.

3.6 Forecasting

Once the bandwidths (h_1, h_2) are determined, different approaches for prediction based on the selected bandwidths can be suggested for localised mixture regression models MLC and MLL. These approaches produce m-step-ahead forecasts for a target time point t_T as follows:

Once the MLC model is fitted, two approaches can be proposed to forecast future observations of a time series at a time point t_T for a given bandwidth $h = (h_1, h_2)$. In the first approach, the m-step-ahead forecast equation is obtained by solving the minimisation problem as follows:

$$\hat{y}_{T+m}^{\mathrm{MLC}^{(1)}} = \min_{a} \sum_{i=1}^{T} \sum_{k=1}^{K} r_{ik} W_k(t_i, t_{T+m})(y_i - a)^2$$

Then, we apply the following m-step-ahead forecast equation

$$\hat{y}_{T+m}^{\mathrm{MLC}^{(1)}} = \frac{\sum_{i=1}^{T} \sum_{k=1}^{K} r_{ik} W_k(t_i, t_{T+m}) y_i}{\sum_{i=1}^{T} \sum_{k=1}^{K} r_{ik} W_k(t_i, t_{T+m})} \tag{3.6.1}$$

In the second approach, the fitted MLC is used for prediction, which gives the following forecast equation:

$$\hat{y}_{T+m}^{\mathrm{MLC}^{(2)}} = \sum_{k=1}^{K} \hat{\pi}_k(t_T) \hat{\beta}_k(t_T) \tag{3.6.2}$$

where $\hat{\pi}_k$ and $\hat{\beta}_k$ are the fitted parameters of MLC.

Moreover, a new approach is presented for prediction based on fitted MLL. The m-step-ahead forecast equation can be articulated as follows:

$$\hat{y}_{T+m}^{\mathrm{MLL}} = \sum_{k=1}^{K} \hat{\pi}_k \left[\hat{\beta}_{0k}(t_T) + \hat{\beta}_{1k}(t_T)(t_{T+m} - t_T) \right] \tag{3.6.3}$$

where $\hat{\pi}_k$, $\hat{\beta}_{0k}$ and $\hat{\beta}_{0k}$ are the fitted parameters at a time point t_T. If we examine Equation (3.6.1), it can be seen that the first approach used for forecasting for the MLC model does not depend on the fitted parameters $\hat{\pi}_k$ and $\hat{\beta}_k$ and $\hat{\sigma}$. However, the approach used in Equations (3.6.2) and (3.6.3) for MLC and MLL models, respectively are mainly based on the fitted parameters. As a result, we can consider the forecast approach as shown in Equation (3.6.3) to be a developed version of the forecast approach used in Equation (3.6.2). The performance of prediction for the MLC and MLL models are compared using Equations (3.6.2) and (3.6.3) in a simulation study. For the first

approach used for forecast in MLC model is compared with the second approach used in Equation (3.6.2) in the simulation study using fixed bandwidths. The methodology of the simulation will be discussed in the following section of this chapter.

3.7 Simulation methodology

A simulation study was conducted to assess the performance of the MLC and MLL models for forecasting based on a new methodology of bandwidth selection. In this study, the MLC and MLL models were compared to the ARIMA and Holt models, which are the most popular approaches used for prediction, see Section 1.5 of Chapter 1. In the reminder of this chapter, a collection of notations is used for ease of explanation. Specifically, $\text{MLC}^{(i)}(h), i = 1, 2$ ($\text{MLL}(h)$) refers to as forecasting based on a vector bandwidth $h = (h_1, h_2)$, where $i = 1$ and $i = 2$ indicate the forecasting approaches used for MLC model in Equations (3.6.1) and (3.6.2), respectively. In addition, $\text{NLC}^{(i)}(h), i = 1, 2$ ($\text{NLL}(h)$) denotes forecasting based on a bandwidth h for MLC and MLL models for one component, respectively. The simulation was executed according to the following steps, where the second and third steps are not applied for the ARIMA and Holt models:

1. 1000 data sets with size 100 for each data set were generated from a given model.

2. A new approach towards bandwidth selection for prediction was applied to find the optimal bandwidth h for $\text{MLC}^{(i)}(h), i = 1, 2$ and $\text{MLL}(h)$ models. The optimal bandwidth $\hat{h}$ was obtained by solving the minimisation problem as follows:

$$\hat{h} = \underset{h}{\text{argmin}} \frac{\sum_{i=a}^{b} \left(\hat{y}_i(h) - y_i\right)^2}{\sum_{i=a}^{b} y_i^2} \tag{3.7.1}$$

   where $\hat{y}_i$ is the forecast based on Equations (3.6.1)–(3.6.3), and a is the 76-th time point, and b is the 96-th time point in this analysis. Hence, we obtained 21 forecasts for each given data set and forward lag. To solve the optimisation problem as found in Equation (3.7.1), the Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS) is used. This is a popular numerical optimisation method used for solving unconstrained non-linear optimisation problems [54].

3. Once the problem of Equation (3.7.1) was solved for each data set in Step 1, the median of the findings was calculated and considered as an optimal bandwidth for all data sets. Since the optimal bandwidths had a right-skewed distribution in most cases in the study, using the median was favourable for obtaining an accurate insight into the optimal bandwidth for all data sets, see Figures 3.1 and 3.2.

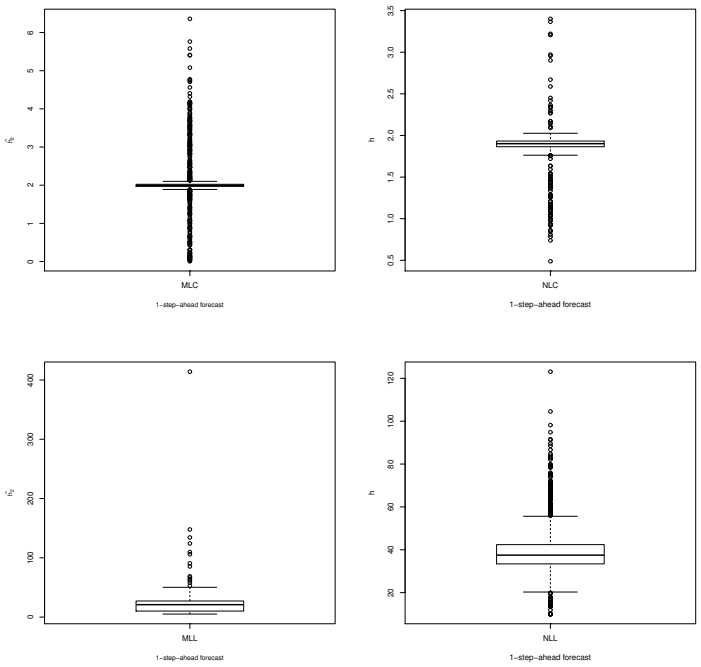

Figure 3.1: Summary of results provided by boxplots of optimal bandwidths of the $\text{NLC}^{(2)}$, $\text{MLC}^{(2)}$, NLL, MLL for data from model 3.8.1

4. Once the optimal bandwidth of $\text{MLC}^{(2)}(h)$, $\text{MLL}(h)$, $\text{NLC}^{(2)}(h)$, and $\text{NLL}(h)$ was found from $a = 76$ to $b = 96$, the m-step-ahead forecasts from $c = 77$ to $d = 97$ were found based on the optimal bandwidth in Step 3 because we assumed that the observations from $c = 77$ to $d = 97$ are unknown. The sum of the square relative error (SSRE) of forecasts is considered as an accuracy criterion for m-step-ahead forecasting for all considered models, which is denoted as $\text{SSRE}(m)$. It can be defined as follows:

$$\text{SSRE}(m) = \frac{\sum_{T=c}^{d} \left(\hat{y}_{T+m} - y_{T+m}\right)^2}{\sum_{T=c}^{d} y_{T+m}^2} \tag{3.7.2}$$

where c is the 77-th time point and d is the 97-th time point.

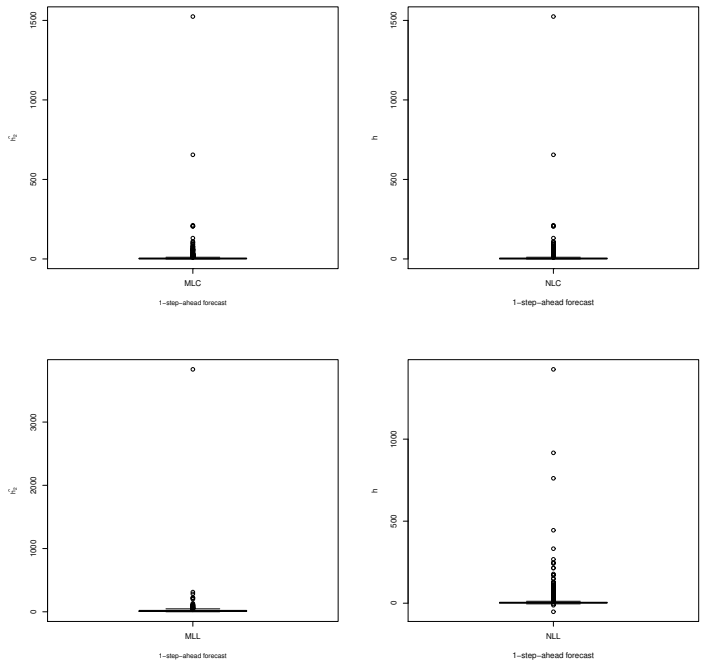

Figure 3.2: Summary of results provided by boxplots of optimal bandwidths of the $\text{NLC}^{(2)}$, $\text{MLC}^{(2)}$, NLL, MLL for data from model 3.8.2

In Step 2, the optimal bandwidth $\hat{h}_2$ for $\text{MLC}^{(2)}(h)$ and $\text{MLL}(h)$ is expected to be larger than 1, especially for the MLL, in order to capture the long-term trends of the past data. In addition, for the simulation study, the median of the 1000 optimal bandwidths was used for all data sets used for prediction rather than the different optimal bandwidths for each data set. In practice, it was found that there is no difference between using the median of all optimised bandwidths, and using different optimal bandwidths for each data set. As a result, this gives an advantage for the median to be an optimal bandwidth for all data sets under study.

In the simulation study, two cases were considered: simulation using optimised bandwidths and fixed non-optimised bandwidths. In the first case, $\text{MLC}^{(2)}(h)$ was considered only because the methodology used in this approach of forecasting does not include the bandwidth h in itself from Equation (3.6.2). In addition, it is noted that the forecast using Equation (3.6.1) depends on h itself, then optimisation over this bandwidth will give the $\text{MLC}^{(1)}(h)$ an advantage over the other techniques. As a result, the optimisation problem for bandwidth selection using Equation (3.6.1), this would be rather misleading in comparison to using the MLL approach, where the bandwidth only plays a role for the estimation but not in prediction. Therefore, the behaviour

of the MLC$^{(1)}(h)$ was considered in a separate experiment, else we aimed to compare between the performance of prediction between MLC$^{(1)}(h)$ and MLC$^{(2)}(h)$ using the same different choices of bandwidths, see Case study 2 of Section 3.8. In addition, we investigated the performance of NLC$^{(1)}(h)$ and NLC$^{(2)}(h)$ and compared this with the MLC$^{(1)}(h)$ and MLC$^{(2)}(h)$.

3.8 Simulation study

In this section, two examples are presented to investigate the performance of MLC$^{(i)}(h)$, $i = 1, 2$, MLL(h) models and their special cases for prediction based on bandwidth selection by simulation using different models. The bandwidth of one of the components was fixed to 1 namely $h_1 = 1$, because it is the smallest obvious bandwidth needed to capture the short-term trend of the past data. However, the second bandwidth h_2 was classed as unknown unless noted differently. In addition, the number of components of MLC$^{(i)}, i = 1, 2$ and MLL models was fixed at 2. These restrictions on the number of components and the bandwidth selection are considered in order to reduce the complexity of the models and the expense of computational cost as mentioned in Section 3.5. For NLC$^{(i)}(h), i = 1, 2$ and NLL(h), the bandwidth h was mainly classed as unknown. However, it was fixed in some simulation scenarios, which will be discussed later in the examples given in this chapter. For each example, two cases were discussed according to their bandwidth selection scenarios. In the first case, the simulation results are shown based on optimised bandwidths h_2 for MLC$^{(2)}$ and MLL, and the optimised bandwidth h for NLC$^{(2)}$ and NLL. In the second case, the simulation results for prediction are based on the fixed non-optimised bandwidth (h_1, h_2) for the MLC$^{(i)}, i = 1, 2$ and NLC$^{(i)}, i = 1, 2$.

3.8.1 Example 1

In the first example, the data was generated using the following model:

$$y_t = 0.1 + 0.1t + \sin(2\pi\frac{t}{12}) + 0.2\sin(2\pi\frac{2t}{12}) + 0.1\sin(2\pi\frac{4t}{12}) + 0.1\cos(2\pi\frac{4t}{12}) + e_t \quad (3.8.1)$$

the errors $e_t \sim N(0, 0.5^2)$ are independently distributed. This model has seasonal

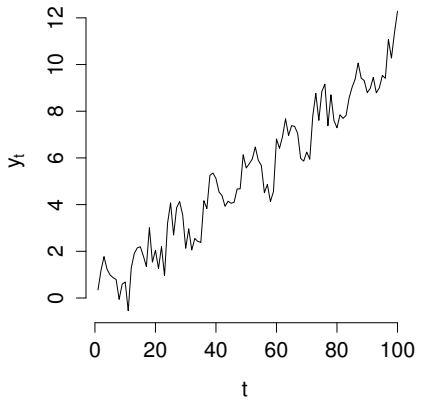

Figure 3.3: Time series of a data set from model (3.8.1).

harmonic components and contains an underlying linear trend. The data originating from model (3.8.1) has a strong variability and linear trend as shown in Figure 3.3. The performance of $\text{MLC}^{(i)}, i = 1, 2$ and MLL for prediction this data is investigated and compared with other common time series models, such as the SARIMA, ARIMA and Holt models. The SARIMA denotes seasonal autoregressive integrated moving average, which is a class of seasonal ARIMA model.

Case study 1: prediction based on optimised bandwidths

In this section, the performance of $\text{MLC}^{(2)}$, MLL, $\text{NLC}^{(2)}$, NLL, SARIMA, ARIMA and Holt models are discussed in relation to the simulation study. By applying steps 1, 2 and 3 of the simulation in section 3.7, $\hat{h}_2 = 2$ and $\hat{h}_2 = 20.93$ produced optimal bandwidths of $\text{MLC}^{(2)}$ and MLL, respectively. In addition, $\hat{h} = 1.90$ and $\hat{h} = 37.50$ produced optimal bandwidths for $\text{NLC}^{(2)}$ and NLL, respectively. Figure 3.1 shows boxplots of the optimal bandwidths for $\text{MLC}^{(2)}$ (top left), the MLL (bottom left), the $\text{NLC}^{(2)}$ (top right), and the NLL (bottom right). It is clear that the optimal bandwidths produced skewed distributions, which suggests that the median can be used as a favourable statistic to present the optimal bandwidths for all data sets in models $\text{MLC}^{(2)}$, MLL, $\text{NLC}^{(2)}$, and NLL.

Figure 3.4[1] shows boxplots of log(SSRE) of m-step-ahead forecasts based on the optimal

[1]A log-transformation is applied on SSRE for clear presentation and ease of data analysis.

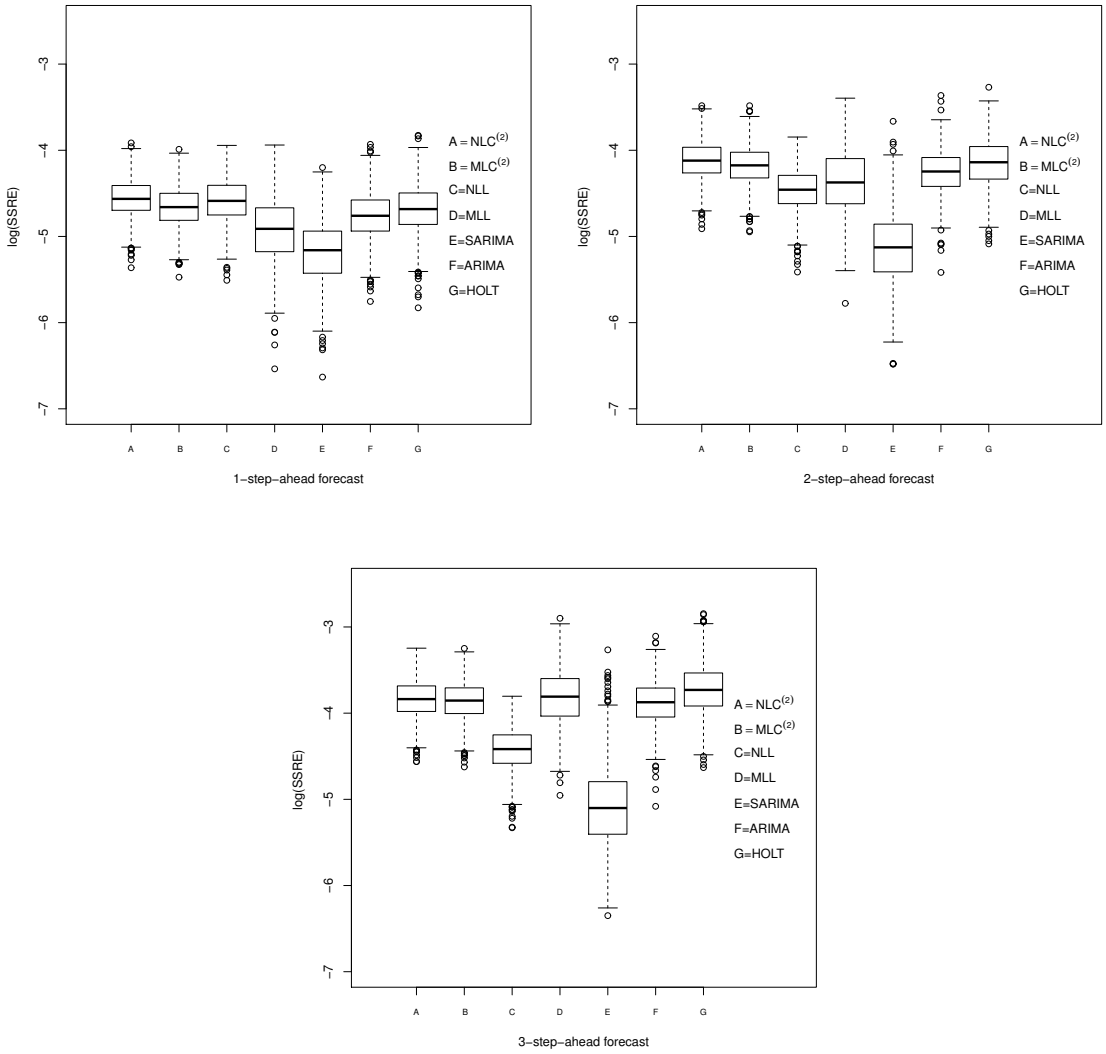

Figure 3.4: Summary of results provided by boxplots of log(SSRE) of the m-step-ahead forecasts, $m = 1, 2, 3$, of data from model (3.8.1) for NLC$^{(2)}$, MLC$^{(2)}$, NLL, MLL, SARIMA, ARIMA and Holt models.

bandwidths for the models under the study and log(SSRE) of m-step-ahead forecasts for traditional models, corresponding to the SARIMA, ARIMA and Holt models. From Figure 3.4 (left), we can see that MLL has performed well and has produced smaller errors for one forward lag than all other models except the SARIMA model. This is due to its ability to model the long-term linear trend. In fact, the SARIMA model is superior to the other models for prediction because the order of seasonality, which is 12, is given in the simulation and considered in the SARIMA model only. With the exception of the SARIMA results, the NLL is better than all other models for $m = 2$ and 3 from Figure 3.4 (right and bottom). In addition, the performance of MLC$^{(2)}$ is

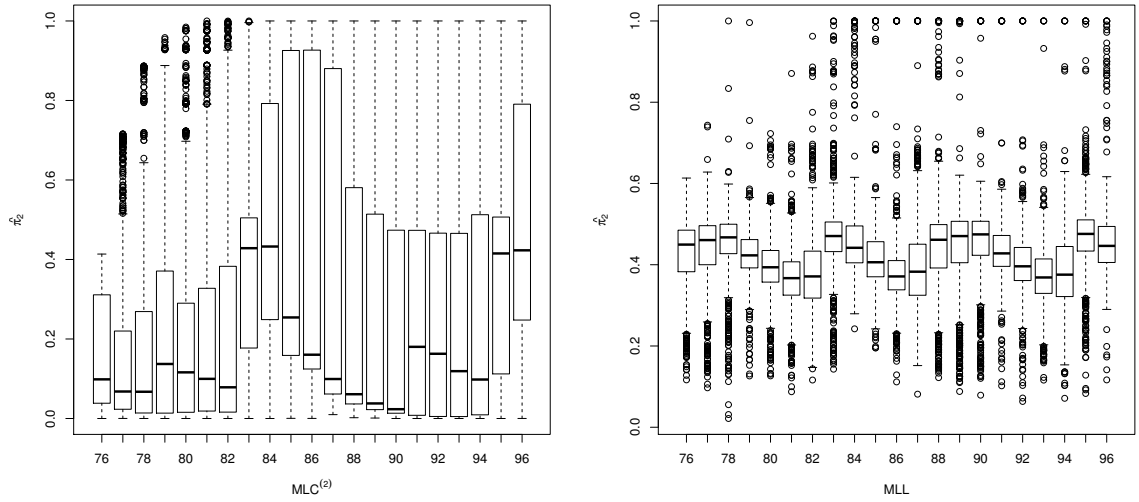

Figure 3.5: Summary of results provided by boxplots of the fitted probabilities of the second components of data from model (3.8.1) for MLC$^{(2)}$ (left) and MLL (right)

better than NLC$^{(2)}$ for all forward lags and becomes competitive with the ARIMA and Holt models only.

Further information is provided in Figure 3.5, which visualises the fitted $\hat{\pi}_2$ of the second components, which are defined in Equations (3.2.4) and (3.3.3) for MLC$^{(2)}$ (left) and MLL (right), respectively. Figure 3.5 shows the importance of the second components of MLC$^{(2)}$ and MLL for prediction. One can observe from Figure 3.5 (left) that the long-term component seems to become close to about 40% at the 83-th, 84-th, 95-th and 96-th grid points for MLC$^{(2)}$. In addition, the smallest percentage of the long-term components is 1% at the 90-th grid point. In fact, the medians of the proportions of the long-term components in the boxplots seem to produce a harmonic pattern through all grid points, which means that the MLC$^{(2)}$ technique has the ability to follow and pick the pattern of a given data set. As a result, the importance of the second component for MLC$^{(2)}$ is limited and restricted on certain grid points in comparison to MLL, as shown in Figure 3.5 (right). The medians of the proportions of the long-term components of MLL fluctuate between 30% and 40%, which make these components useful for prediction compared to the MLC$^{(2)}$ model. As a result, recent information corresponding to the short-term component is considered more relevant for MLC$^{(2)}$. However, the importance of the long-term component for prediction for the MLL is more than that for MLC$^{(2)}$ thanks to its ability to fit a long-term linear

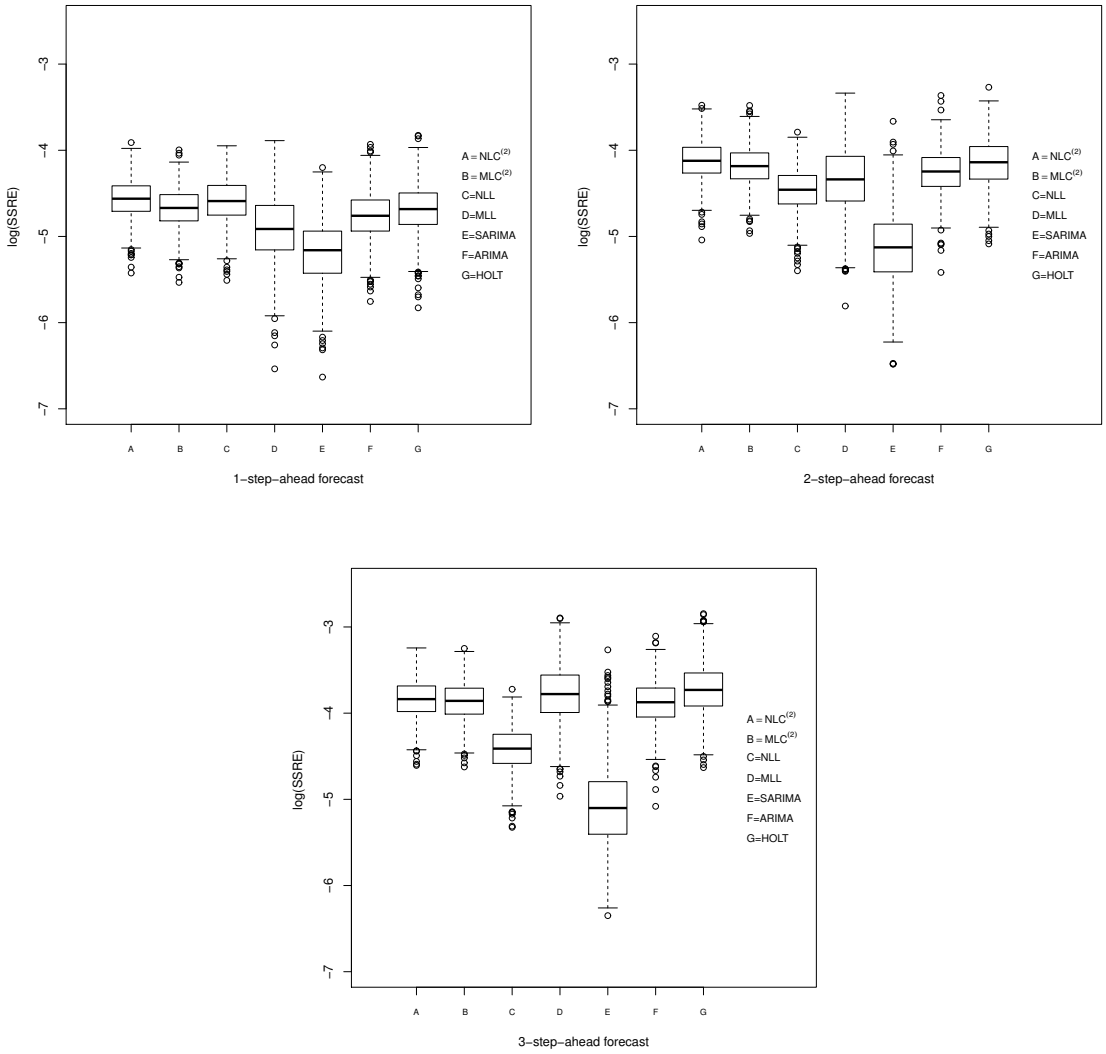

Figure 3.6: Summary of results provided by boxplots of log(SSRE) of m-step-ahead forecasts, $m = 1, 2, 3$, of data from model (3.8.1) for the NLC$^{(2)}$, MLC$^{(2)}$, NLL, MLL, SARIMA, ARIMA and Holt models.

trends in comparison with MLC$^{(i)}$, $i = 1, 2$ models.

Figures 3.6 and 3.7 show the different simulation results when using different optimised bandwidths for each individual data set. It is clear that the results are similar to the results shown in Figures 3.4 and 3.5. As a result, Figures 3.6 and 3.7 support the concept of using the median of all optimal bandwidths in the study, which led to Figures 3.4 and 3.5. This provides clear an indication that there was no need to use different optimal bandwidths for each data set in the simulation study.

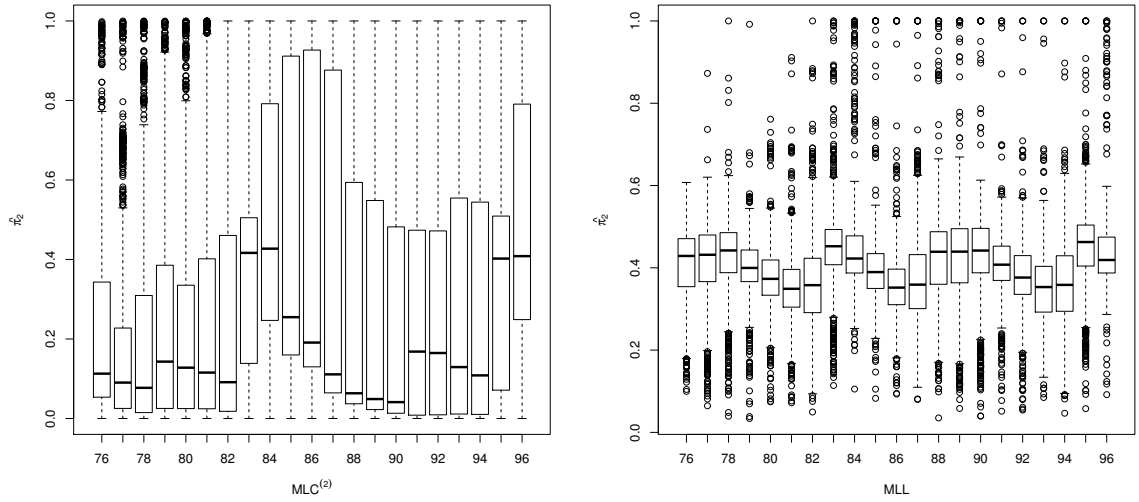

Figure 3.7: Summary of results provided by boxplots of the fitted of probabilities of the second components of data from model (3.8.1) for the MLC$^{(2)}$ (left) and MLL (right)

Case study 2: prediction based on fixed non-optimised bandwidths

In this section, performance of predictions for selected bandwidths (h_1, h_2), in relation to MLC$^{(1)}$ and MLC$^{(2)}$ is investigated and compared with those produced using the SARIMA, ARIMA and Holt models. Figure 3.8 shows boxplots of log(SSRE) of the m-step-ahead forecasts for NLC$^{(i)}$, $i = 1, 2(1)$, MLC$^{(i)}$, $i = 1, 2(1,4)$, SARIMA, ARIMA and Holt models. It is clear that NLC$^{(1)}$(1) has the best performance in prediction, with a minimum median for all forward lags of m. The MLC$^{(2)}$(1,4) produced a larger margin error than NLC$^{(2)}$(1) , but only for $m = 1$ as shown in Figure 3.8 (left) according to the log(SSRE). However, the MLC$^{(1)}$(1,4) model produces the worst performance for prediction in comparison to the NLC$^{(1)}$(1) for all forward lags. Figure 3.9 shows that MLC$^{(1)}$(0.5,1) is better than NLC$^{(1)}$(1) for prediction according to log(SSRE). In addition, for the MLC$^{(2)}$(0.5,1), the performance of prediction is slightly improved in comparison with that of NLC$^{(2)}$(1) for all forward lags of m. This result is plausible because the bandwidth $h_2 = 4$ is far from the real optimal bandwidth h_2 for MLC$^{(2)}$ in this example.

It is worth mentioning that using MLC and MLL methodologies for prediction can pick the right optimal bandwidths, which means that the bandwidth selection is implicit to these approaches. For example, in Figures 3.10 and 3.11, each grid point presents 1000

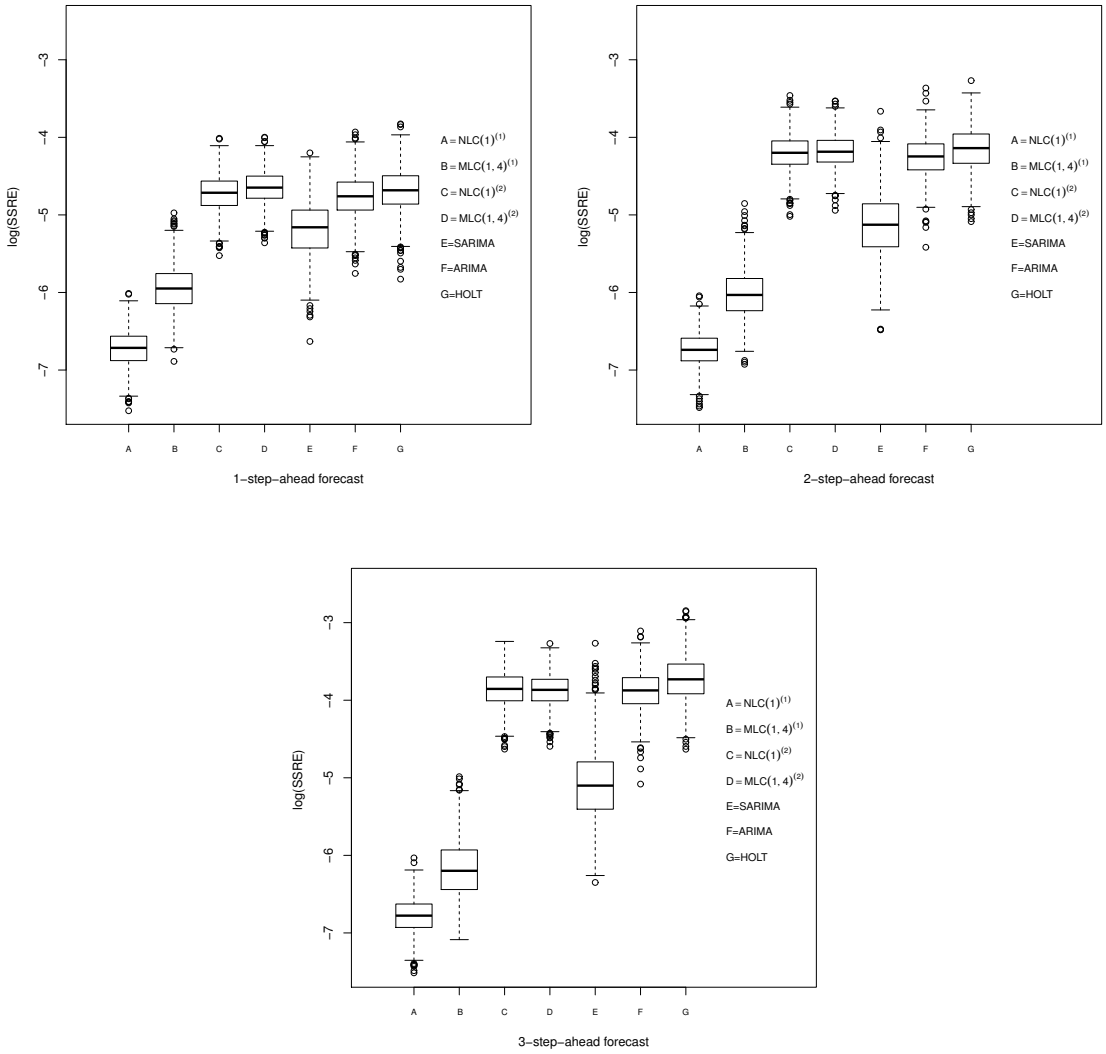

Figure 3.8: Summary of results provided by boxplots of log(SSRE) of the m-step-ahead forecasts, $m = 1, 2, 3$, of data from model (3.8.1) for NLC$^{(i)}$, MLC$^{(i)}$, SARIMA, ARIMA and Holt models when $h = (1, 4)$, $i = 1, 2$.

results for each boxplot of the fitted proportions $\hat{\pi}_i, i = 1, 2$ and 3 of MLC$^{(2)}$(1, 2, 5) and MLL(1, 20.93, 50), which are defined in Equations (3.2.4) and (3.3.3) respectively. Figure 3.10 shows the fitted proportions $\hat{\pi}_1$, $\hat{\pi}_2$, $\hat{\pi}_3$ from the 76-th to the 96-th grid points when using the model MLC$^{(2)}$(1, 2, 5) to fit the data sets in Example 1. In this case, the optimal bandwidth $h_2 = 2$ of MLC$^{(2)}$ in this example is included to investigate whether the MLC methodology for prediction can select the right bandwidth. It can be observed, that the bandwidth $h_3 = 5$ has the lowest proportions $\hat{\pi}_3$, and in general compared to the proportions $\hat{\pi}_1$'s and $\hat{\pi}_2$'s of the other components. This means that the third component, that is related to the selected bandwidth $h_3 = 5$, is less important

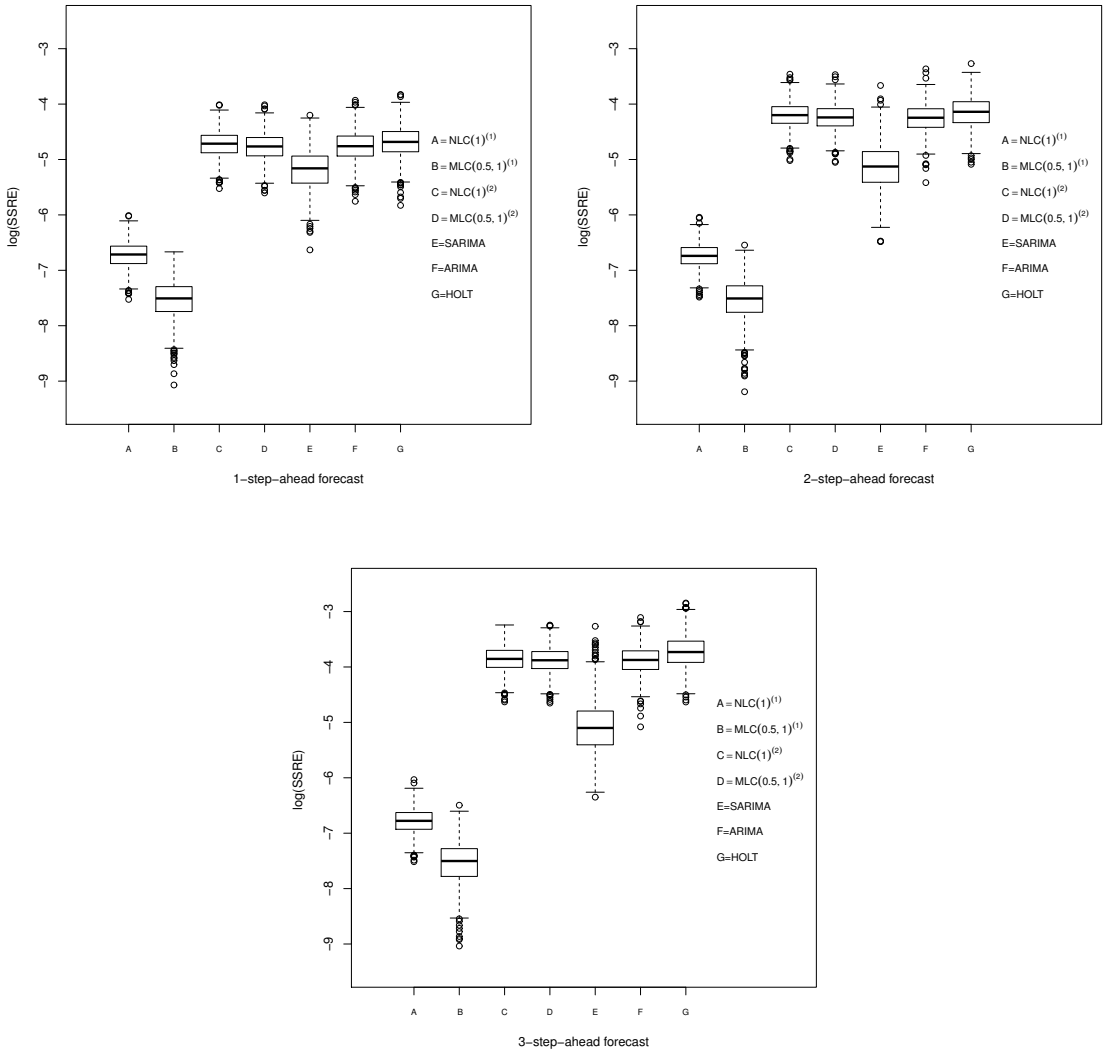

Figure 3.9: Summary of results provided by boxplots of log(SSRE) of the m-step-ahead forecasts, $m = 1, 2, 3$, of data from model (3.8.1) for the NLC$^{(i)}$, MLC$^{(i)}$, SARIMA, ARIMA and Holt models when $h = (0.5, 1)$, $i = 1, 2$.

than the other components. As a result, the MLC methodology overwhelmingly selects the most important components that correspond to bandwidths, which have previously been shown to produce good forecasts. Moreover, this result supports the results of Case study 1, which gives our new methodology of bandwidth selection an advantage.

In Figure 3.11, the optimal bandwidth obtained for the second component of the MLL model $h_2 = 20.93$ is used in the second component of the MLL(1, 20.93, 50) model to fit the data sets shown in Example 1. It seems that the first components for each grid point with the bandwidth $h_1 = 1$, have the highest proportions. In addition, the

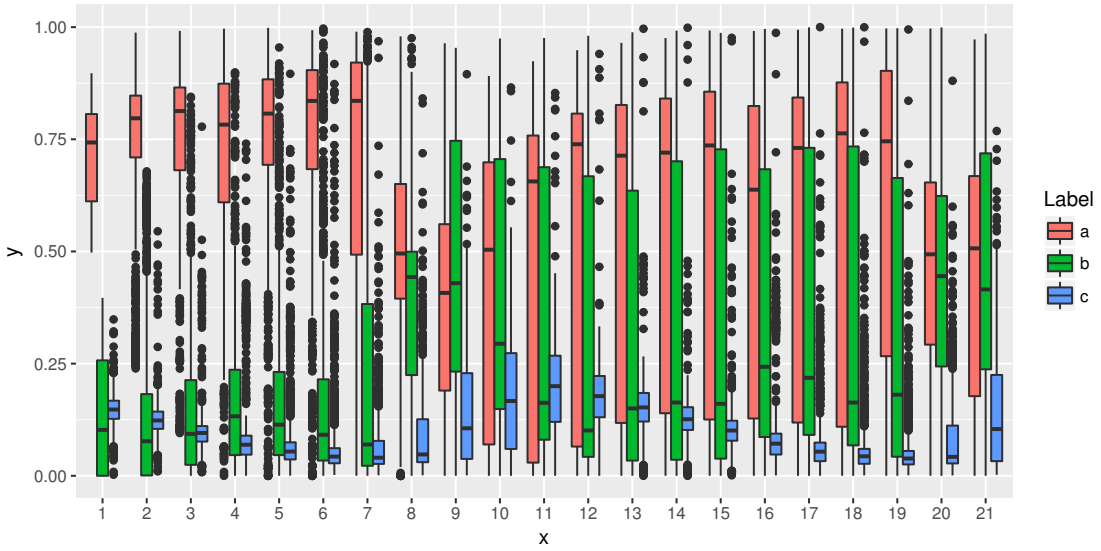

Figure 3.10: Summary of results provided by boxplots of the fitted probabilities $\hat{\pi}_1$, $\hat{\pi}_2$ and $\hat{\pi}_3$ of data from model (3.8.1) for MLC$^{(i)}$(1,2,5), $i = 1, 2$, where $a = \hat{\pi}_1$, $b = \hat{\pi}_2$, and $c = \hat{\pi}_3$. The horizontal axis denotes the grid points (from 77 to 97) and the vertical axis gives the proportions.

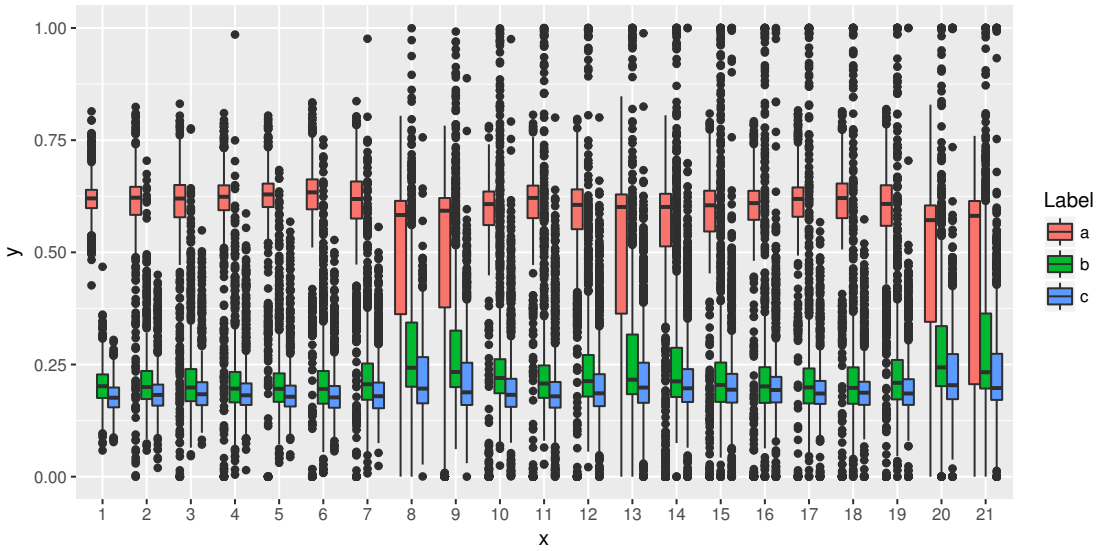

Figure 3.11: Summary of results provided by boxplots of the fitted probabilities $\hat{\pi}_1$, $\hat{\pi}_2$ and $\hat{\pi}_3$ of data from model (3.8.1) for MLL(1,20.93,50), where $a = \hat{\pi}_1$, $b = \hat{\pi}_2$, and $c = \hat{\pi}_3$. The horizontal axis denotes the grid points (from 77 to 97) and the vertical axis gives the proportions.

medians of the fitted proportions for the second components are larger than those of the third components. In this case, the MLL model shows the ability to capture the same bandwidths as those obtained in Case study 1 of the simulation. Figure 3.11 supports these results in Figure 3.5. For the MLL model, the fitted proportions $\hat{\pi}_2$ of the second

components with a bandwidth $h_2 = 20.93$ for each grid point in Figure 3.5 (right) are larger than the fitted proportions $\hat{\pi}_2$, which are found and presented in Figure 3.11. These differences between MLL(1, 20.93) and MLL(1, 20.93, 50) relating to the fitted proportions $\hat{\pi}_2$ is reasonable due to the increase in the number of components in MLL(1, 20.93, 50), which contributes to reducing the proportions for the components with the bandwidth $h_2 = 20.93$. This suggests that the MLL methodology in this example does not need more than two components. As a result, Figures 3.5 and 3.11 provide an indication that $K = 2$ is adequate to fit the short and long-term trend in this example.

3.8.2 Example 2

In this example, the data was generated from a model, which has no trends. This model can be defined as follows:

$$y_t = \sin(2\pi\frac{t}{12}) + 0.2\sin(2\pi\frac{2t}{12}) + 0.1\sin(2\pi\frac{4t}{12}) + 0.1\cos(2\pi\frac{4t}{12}) + e_t \qquad (3.8.2)$$

The errors are assumed firstly as $z_t \sim N(0, 0.5^2), t = 1, \ldots, 100$ and then the considered errors are $e_t = 0.7e_{t-1} + z_t, e_1 = z_1, t = 2, \ldots, 100$. This model includes seasonal harmonic components without linear trend. Figure 3.12 shows a data set from model 3.8.2 which is very variable time series. The main difference between the model shown here and the model used in the first example is, that the errors are strongly correlated,

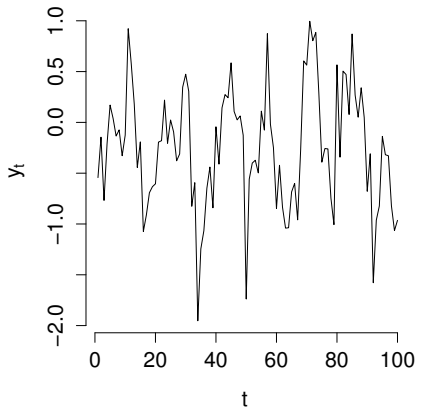

Figure 3.12: Time series of a data set from model (3.8.2).

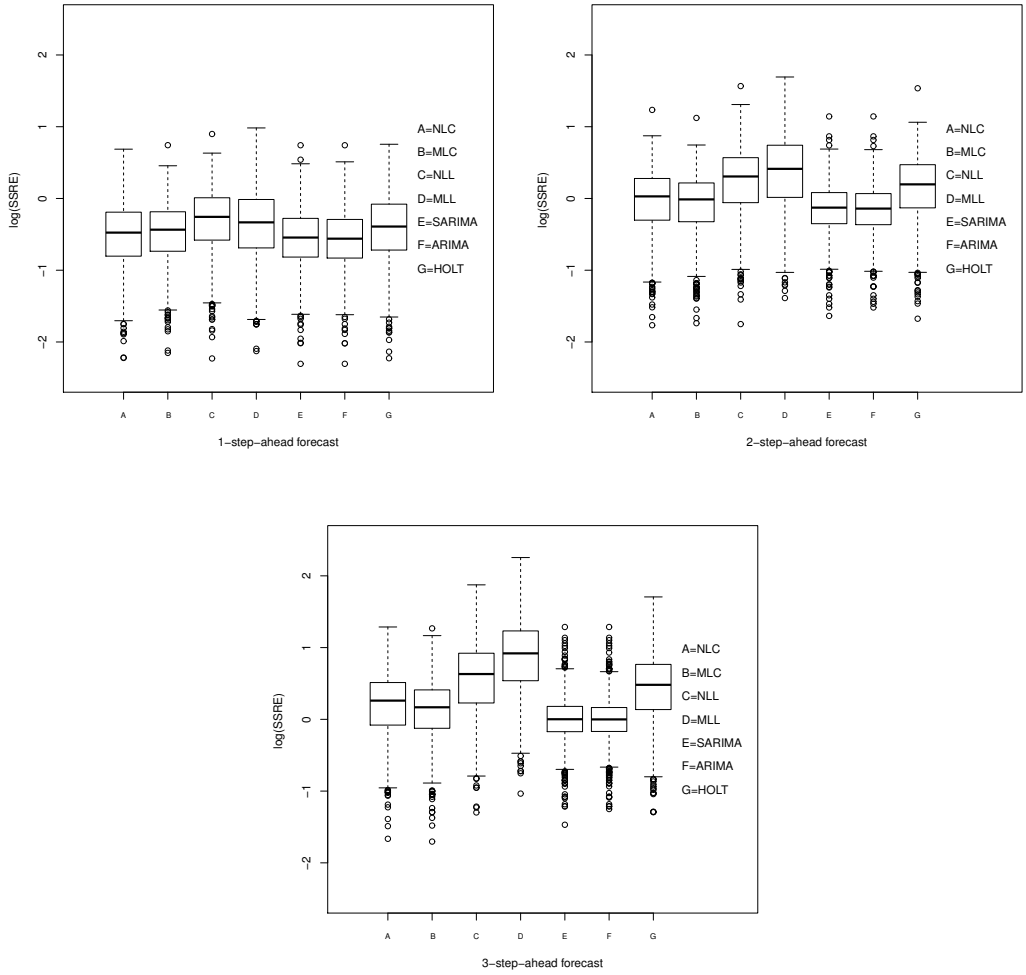

Figure 3.13: Summary of results provided by boxplots of log(SSRE) of m-step-ahead forecasts, $m = 1, 2, 3$, of data from model 3.8.2 for the NLC$^{(2)}$, MLC$^{(2)}$, NLL, MLL, SARIMA, ARIMA and Holt models

and the data from the former one do not have linear trend. As a result, there is a strong suggestion that the MLC$^{(2)}$ will perform well for prediction compared to the MLL. From Step 3 of the simulation set up in Section 3.7, $\hat{h}_2 = 2.15$ and $\hat{h}_2 = 10.11$ are the optimal bandwidths used for MLC$^{(2)}$ and MLL, respectively. For NLC$^{(2)}$ and NLL, the optimal bandwidths used are $\hat{h} = 0.80$ and $\hat{h} = 2.50$, respectively.

Case study 1: prediction based on optimised bandwidths

Figure 3.13 shows boxplots of log(SSRE) of the m-step-ahead forecasts for the MLC$^{(2)}$, MLL, NLC$^{(2)}$, NLL, SARIMA, ARIMA and Holt models. It is obvious as shown in Figure 3.13 (left), that the MLL model is competitive with the MLC$^{(2)}$ and Holt

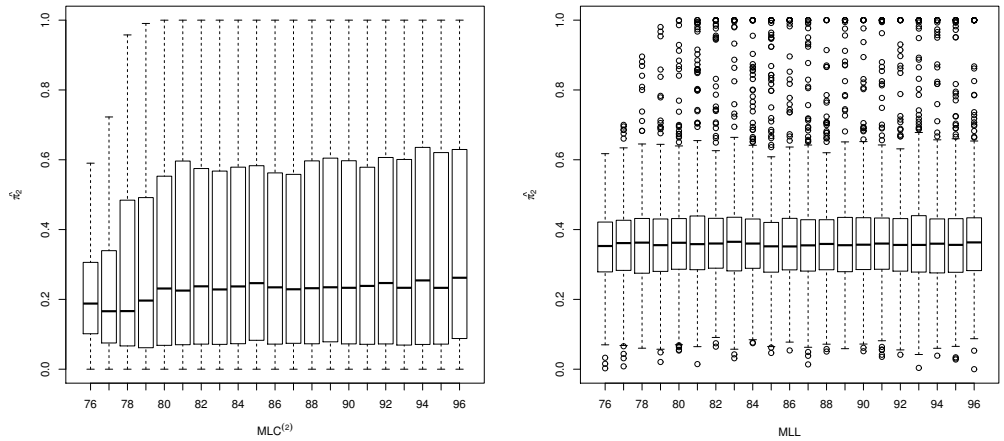

Figure 3.14: Summary of results provided by boxplots of the fitted probabilities of the second components for the MLC$^{(2)}$ (left) and the MLL (right)

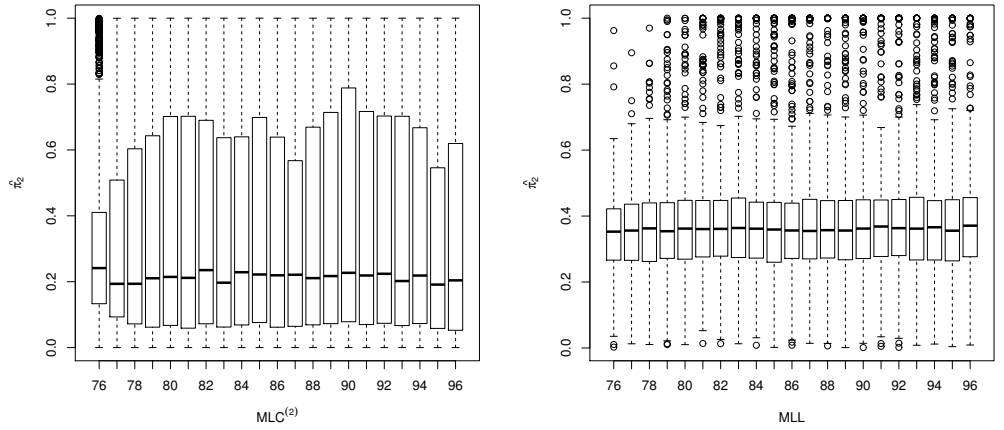

Figure 3.15: Summary of results provided by boxplots of the fitted probabilities of the second components of data from model (3.8.2) for the MLC$^{(2)}$ (left) and MLL (right)

models, and performs better than the NLL model for $m = 1$ only. In addition, there is no difference in performance for prediction between the SARIMA and ARIMA models, and they perform better than other models for all forward lags. In addition, the MLC$^{(2)}$ model performs better than the NLC$^{(2)}$ and Holt models for $m = 2$ and 3, as shown in Figure 3.13 (right and bottom). Figure 3.14 (left) shows that the proportions of the long-term components are around 0.22% for almost all grid points. This gives positive evidence regarding the ability of the MLC model to fit short-term trend components, rather than long-term trend components, especially for non-linear data. However, we can see from Figure 3.14 (right), that on all grid points, the proportions of the long-term

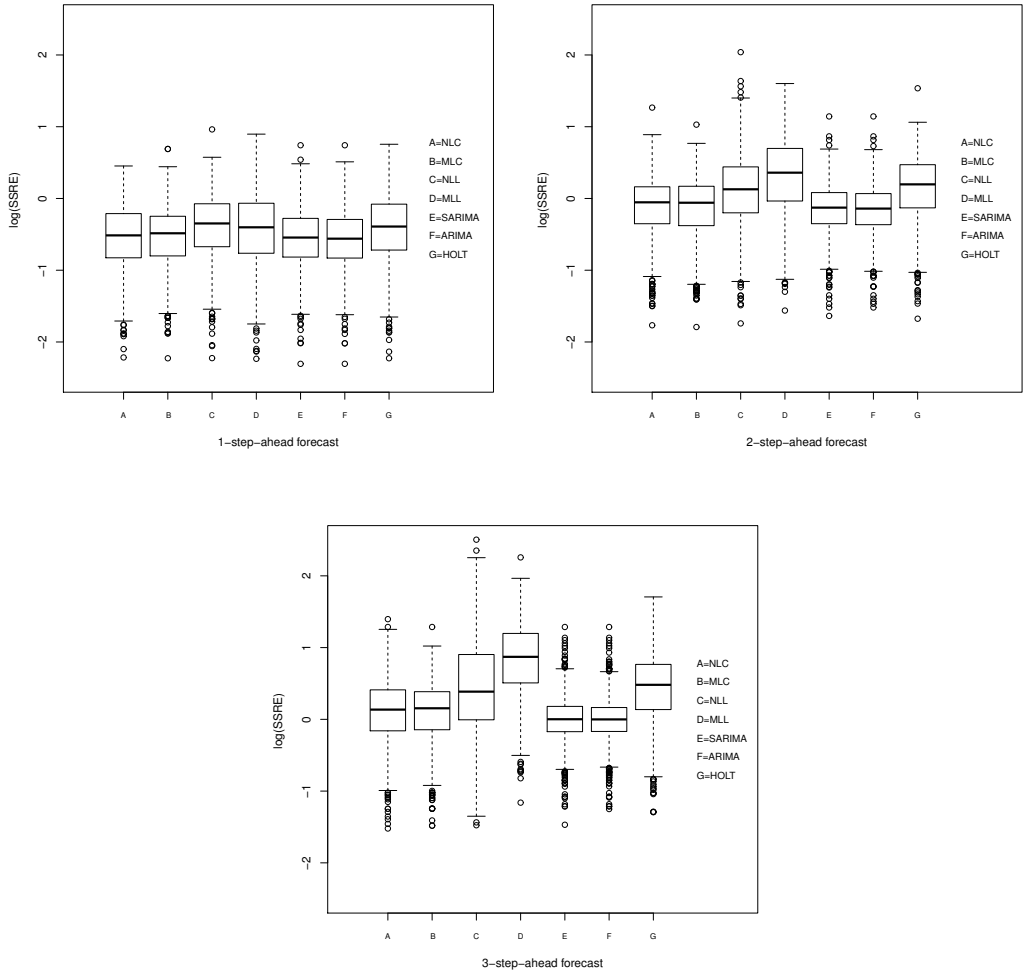

Figure 3.16: Summary of results provided by boxplots of log(SSRE) of m-step-ahead forecasts, $m = 1, 2, 3$ of data from model 3.8.2 for the NLC$^{(2)}$, MLC$^{(2)}$, NLL, MLL, SARIMA, ARIMA and Holt models

components settle at about 35% for the MLL model. This result is sensible, since the data has no trends, which makes these components in the mixture less important than the short-term components. Figures 3.15 and 3.16 as shown present the simulation results when using different optimised bandwidths for each data set shown in this example. As in Example 1, it is clear that the results are similar to the results shown in Figures 3.13 and 3.14.

Case study 2: prediction based on fixed non-optimised bandwidths

Figure 3.17 presents boxplots of log(SSRE) for NLC$^{(i)}(1)$, MLC$^{(i)}(1, 4)$, SARIMA, ARIMA and Holt models, for $i = 1, 2$. It is clear that NLC$^{(1)}(1)$ model is the best

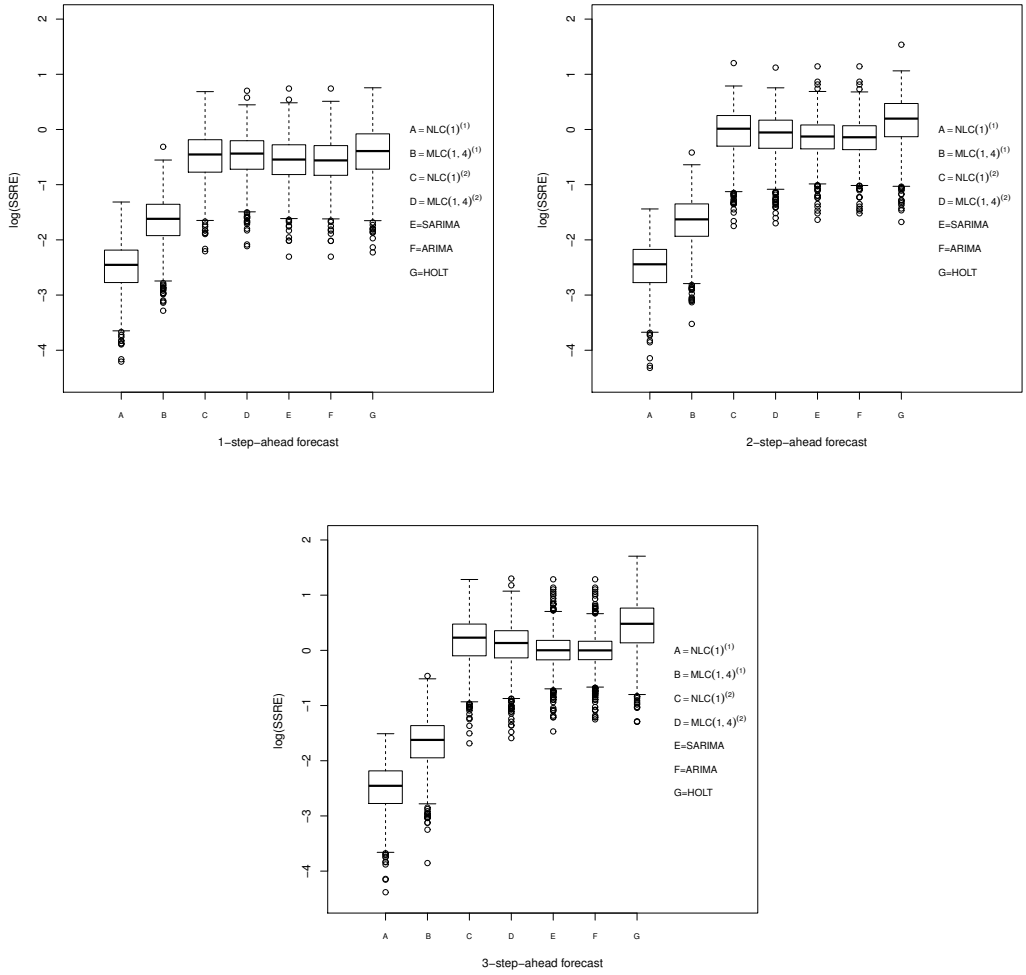

Figure 3.17: Summary of results provided by boxplots of log(SSRE) of m step-ahead forecasts, $m = 1, 2, 3$, of data from model 3.8.2 for the NLC$^{(1)}$, MLC$^{(1)}$, NLC$^{(2)}$, MLC$^{(2)}$, SARIMA, ARIMA and Holt when $h = (1, 4)$

model for prediction based on the log(SSRE) for all forward lags of m. In addition, the MLC$^{(1)}$(1,4) model performs well for all forward lags compared with all other models, except the NLC$^{(1)}$(1) model. For $m = 1$, it can be seen in Figure 3.17 (left), that MLC$^{(2)}$(1,4) is competitive with the NLC$^{(2)}$(1) and with the Holt model. However, for $m = 2$ and $m = 3$, MLC$^{(2)}$(1,4) shows better performance for prediction compared to the NLC$^{(2)}$(1) and the Holt models. In addition, it becomes competitive with the SARIMA and ARIMA models. From examining Figure 3.17, it can be seen that when the bandwidth h_2 for MLC$^{(1)}$ is larger (less) than the bandwidth h for NLC$^{(1)}$, and so performance of MLC$^{(1)}$ for prediction decreases (increases).

Figure 3.18 shows the boxplots of log(SSRE) for NLC$^{(i)}$(1), MLC$^{(i)}$(0.5,1), SARIMA,

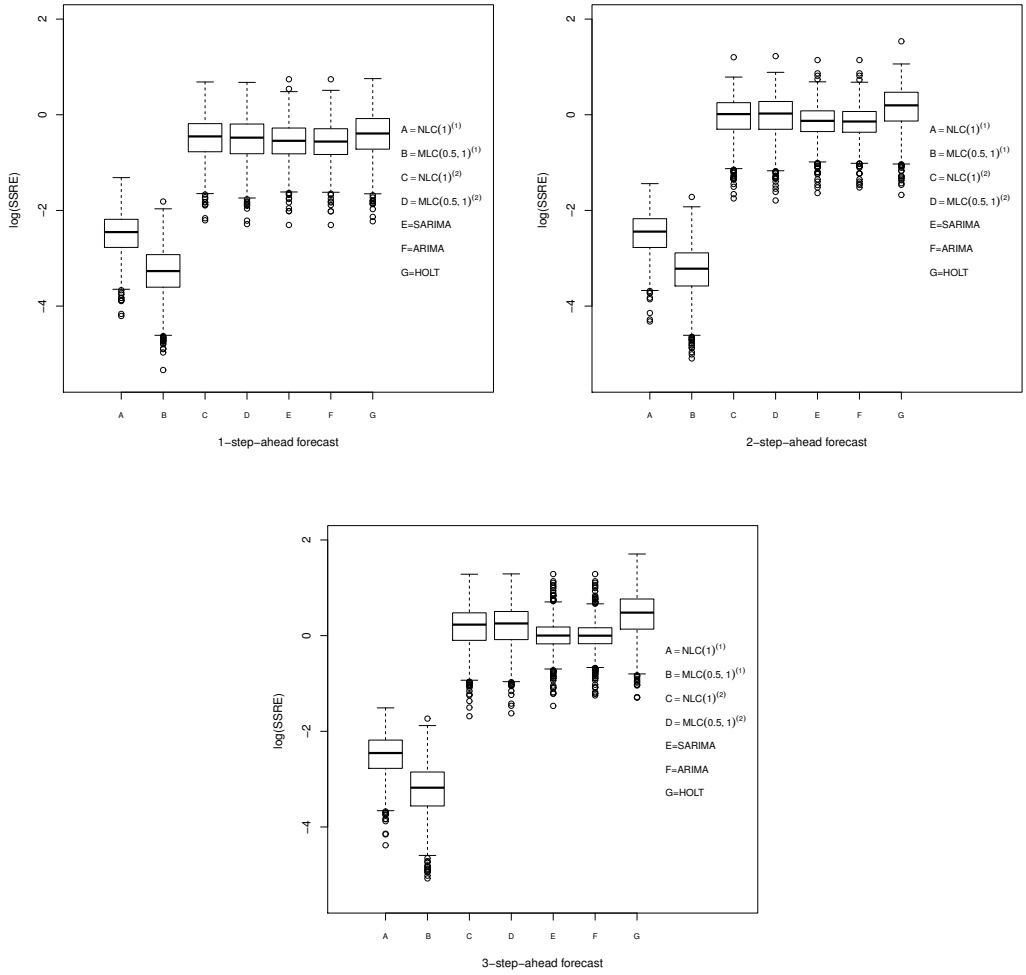

Figure 3.18: Summary of results provided by boxplots of log(SSRE) of m step-ahead forecasts, $m = 1, 2, 3$, of data from model 3.8.2 for the NLC$^{(1)}$, MLC$^{(1)}$, NLC$^{(2)}$, MLC$^{(2)}$, SARIMA, ARIMA and Holt when $h = (1, 0.5)$

ARIMA and Holt models, for $i = 1, 2$. It is clear that the MLC$^{(1)}$(0.5,1) is the best performing prediction model in comparison with the other models. However, MLC$^{(2)}$(0.5,1) shows marginal better performance for prediction compared with the NLC$^{(2)}$(1) for $m = 1$ as shown in Figure 3.18 (left).

In conclusion, from the above simulation study, we can see that the MLC$^{(i)}, i = 1, 2$ and MLL models reveal powerful methodologies, that could contribute to improving predictions based on bandwidth selection, if we compare them to traditional models, such as the ARIMA and Holt models. The MLL methodology was superior to other models for variable and linear trend data and for small forward lags, as shown in Example 1. The MLC$^{(2)}$ model performs well for prediction for short-term trends and

for large forward lags as shown in Example 2. The $\text{MLC}^{(1)}$ model in both Examples 1 and 2 shows better results for prediction compared to the $\text{MLC}^{(2)}$, ARIMA and Holt models. Moreover, the $\text{MLC}^{(i)}, i = 1, 2$ models are superior to the $\text{NLC}^{(i)}, i = 1, 2$ for very small bandwidths of the second components, that is $h_2 < 1$.

3.9 Applications

In this section, real data examples are presented, in order to investigate the performance of the $\text{MLC}^{(i)}, i = 1, 2$ and MLL models for forecasting compared to the ARIMA and Holt models. The data under study is the same data as discussed in Section 2.3 of Chapter 2, that is the annual energy use data. While the full data set contains more than 130 countries, we choose three countries with different patterns for purposes of this presentation.

Figure 3.19 displays the time series of log of energy use of Bolivia, Lebanon and Greece. It can be seen that the time series for Bolivia (left) shows two main features, which are shared by the large majority of countries in this data base: it shows an overall increasing linear trend, but considerable variability. The other two time series illustrate extreme cases where one of the features is more pronounced: in the case of Lebanon (see in the middle) it shows very strong variability, and in the case of Greece (right) it shows a very consistent linear trend with little variability.

The log of energy use data of these countries is fitted at the target points $t_T =$

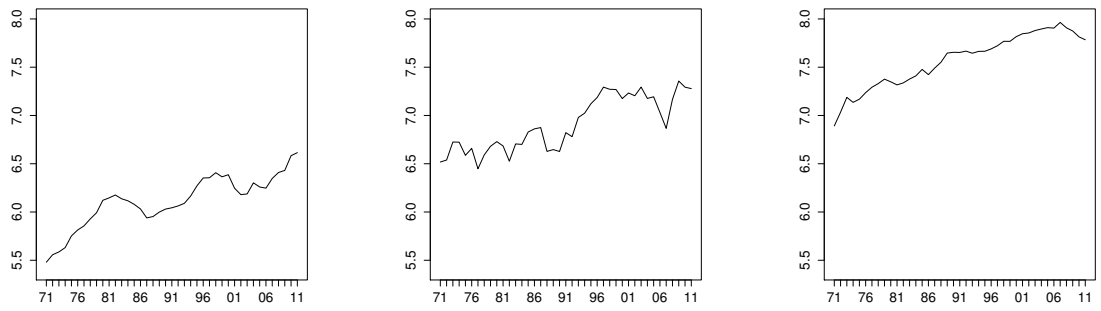

Figure 3.19: Time series of energy use data for Bolivia, Lebanon and Greece (from left to right). The horizontal axis denotes the calendar year (from 1971 to 2011), and the vertical axis gives the annual energy use (natural log of kg oil equivalent per capita).

Model	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(4)	SARE(1)	SARE(2)	SARE(3)	SARE(4)
				$(h_1, h_2) = (1, 5)$				
$MLC^{(1)}$	0.09	0.08	0.09	0.09	7.11	7.34	7.78	7.87
$MLC^{(2)}$	0.23	0.41	0.67	0.99	12.63	17.77	22.97	26.91
MLL	0.12	0.24	0.48	0.87	8.64	11.56	18.38	25.49
				$(h_1, h_2) = (1, 20)$				
$MLC^{(1)}$	0.15	0.17	0.20	0.22	10.07	10.97	12.19	13.13
$MLC^{(2)}$	0.29	0.47	0.77	1.11	14.13	19.09	24.43	28.24
MLL	0.14	0.31	0.59	1.06	9.50	14.84	21.04	27.09
Holt	0.14	0.44	0.90	1.50	8.85	17.60	26.33	34.20
ARIMA	0.13	0.37	0.67	0.97	8.74	14.62	21.38	25.86

Table 3.1: The SSRE and SARE of forecasting for Bolivia from 1991 to 2008 using fixed bandwidths.

$1990, \ldots, 2007$, in order to obtain the m-step-ahead forecasts ($m = 1, \ldots, 4$) for each time point t_T for different models. Hence, we have 18 forecasts for each model and forward lags. For the MLC and MLL models, $K = 2$ components are used to fit the data. To assess the performance of the forecasts using these models, we will consider the SSRE of forecasts and the sum of absolute relative error (SARE) of the m-step-ahead forecasts, which is denoted as SARE(m) and can be defined as follows:

$$\text{SARE}(m) = \frac{\sum_{T=c}^{d} |\hat{y}_{T+m} - y_{T+m}|}{\sum_{T=c}^{d} |y_{T+m}|} \tag{3.9.1}$$

where c is the first time point and d is the last time point, which for our analysis takes the values $c = 1990$ and $d = 2007$, respectively.

The new approach of bandwidth selection is applied for the energy use by the selected countries using the Equation (3.7.1) where $a = 1989$ and $b = 2006$ for this analysis. Tables 3.3, 3.6 and 3.9 summarise the results[2] of the m-step-ahead forecasts based on the optimised bandwidths according to the SSRE of the 1-step-ahead forecasts criteria for MLL and NLL models relating to the energy use of Bolivia, Lebanon and Greece, respectively. For $MLC^{(1)}$ and $NLC^{(1)}$ models, the optimal bandwidths were very small and convergent to 0. Consequently, three different settings of bandwidths, $(h_1, h_2) = (0.5, 1)$, $(h_1, h_2) = (1, 5)$ and $(h_1, h_2) = (1, 20)$, are considered randomly, in order to capture different sizes in term of trends, short, medium and long-term trend, prevailing in these data sets, as shown in Tables 3.1, 3.4 and 3.7. Table 3.2, 3.5 and 3.8 show the results of the SSRE and SARE of the m-step-ahead forecasts for the $MLC^{(i)}$ and $NLC^{(i)}, i = 1, 2$.

[2]All values of SSRE and SARE in tables are multiplied by 1000.

Model	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(4)	SARE(1)	SARE(2)	SARE(3)	SARE(4)
				$h = 1$				
NLC$^{(1)}$	0.02	0.02	0.03	0.03	3.96	4.04	4.48	4.65
				$h = (0.5, 1)$				
MLC$^{(1)}$	0.007	0.006	0.007	0.005	2.06	1.85	1.98	1.75
				$h = 1$				
NLC$^{(2)}$	0.17	0 .36	0.65	0.99	10.76	16.43	22.13	27.86
				$h = (0.5, 1)$				
MLC$^{(2)}$	0.14	0.34	0.62	0.95	9.94	15.87	21.71	27.33

Table 3.2: The SSRE and SARE of forecasting for Bolivia from 1991 to 2008 using fixed bandwidths.

From Table 3.1, we see that the MLC$^{(1)}$ model performs well for all forward lags, and produces smaller errors than all other methods, except in the case when $h = (1, 20)$ and $m = 1$. In this case, the MLL model shows slightly better performance than the MLC$^{(1)}$ model, due to its ability to model the long-term linear trend. This result is equivalent to the result provided by Table 3.3, which suggests that $h = (1, 5.2)$ is the optimal bandwidth for prediction for the MLL model. This finding supports the conclusion in Section 3.8, that the MLL model can pick a good bandwidth in itself. In addition, the optimal bandwidth of NLL model is $h = 0.9$ with SSRE= 0.13×10^{-3} as shown from Table 3.3. It is clear that the MLL model is superior to the NLL model for all forward lags for prediction. In this example, the MLC$^{(2)}$ and NLC$^{(2)}$ models do not predominantly depend on the bandwidth selection. It is found that a range of bandwidths, have approximately the same SSRE. As a result, the new methodology of bandwidth selection is just applied on the MLL and NLL models as it is shown in Table 3.3. As shown in Table 3.1, the MLL model performs better than the MLC$^{(2)}$ model for all forward lags, and for small and large-term trends components. It is clear that the ARIMA and Holt models become competitive with the MLC$^{(1)}$ and MLL models for short forward lags and large-term trend components as shown in Table 3.1. The MLC$^{(2)}$ model produces a larger margin of errors than the MLC$^{(1)}$ model, which makes the MLC$^{(1)}$ model a more favourable model to use for prediction in this example. It can be seen in Table 3.2, that the MLC$^{(1)}$ model performs better than all models and

Model	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(4)
		$h = 0.90$		
NLL	0.13	0.39	0.77	1.34
		$(h_1, h_2) = (1, 5.2)$		
MLL	0.12	0.24	0.48	0.86

Table 3.3: The SSRE of forecasting for Bolivia from 1991 to 2008 using optimised bandwidths.

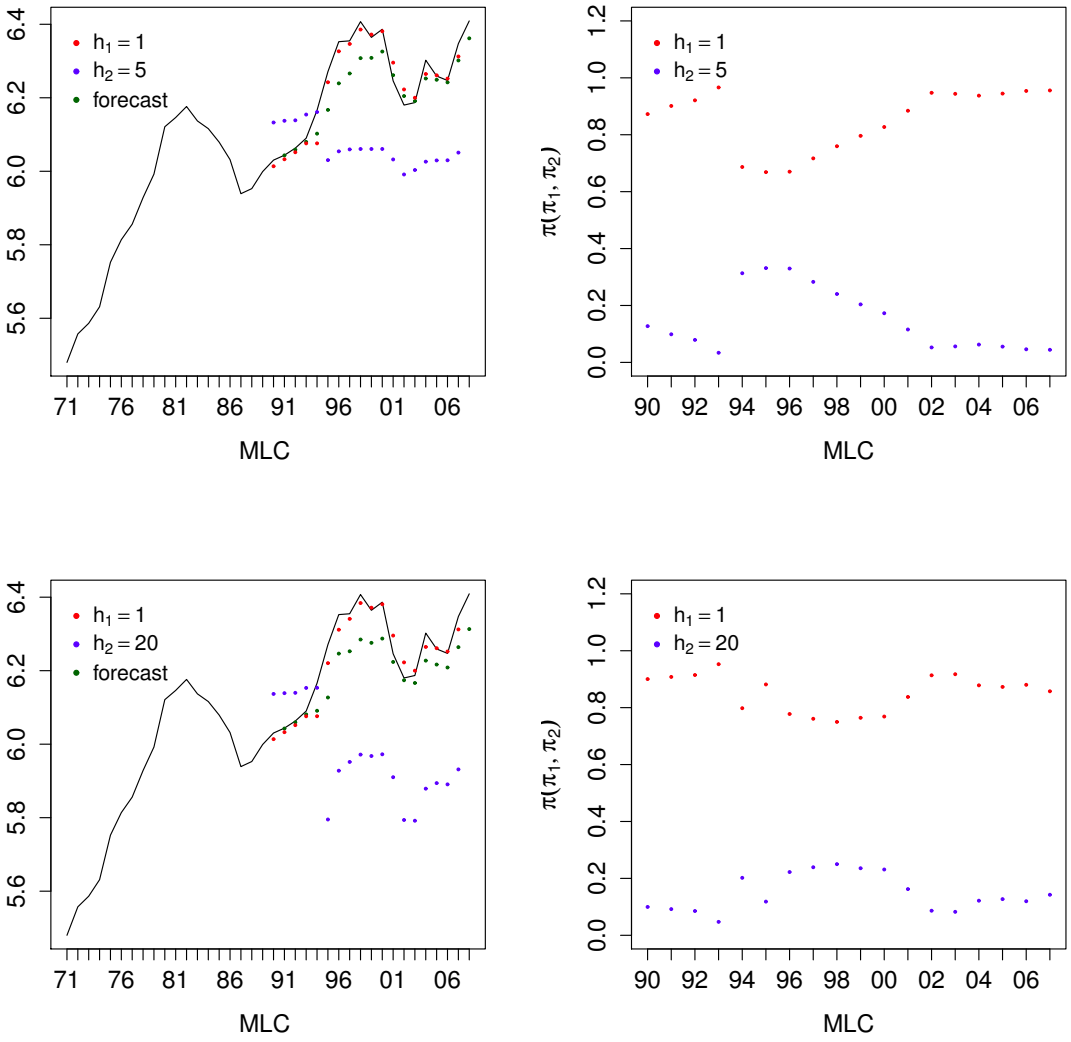

Figure 3.20: Data for Bolivia showing the parameters $\hat{\beta}_k(t_T)$ fitted using MLC$^{(1)}$ and resulting forecasts at $\hat{y}_{T+1}^{\text{MLC}^{(1)}}$ (left); and fitted parameters $\hat{\pi}_k(t_T)$ (right).

for all forward lags.

Further insight is provided in Figure 3.20, which shows the time series for Bolivia, as well as the fitted parameters and predictions (top and bottom left), and the fitted mixture probabilities (top and bottom right) for $t_T, T = 1990, \ldots, 2007$ for the one-step ahead forecasts from the MLC$^{(1)}$ model. One can observe that the long-term components seem to become close to irrelevant for the MLC$^{(1)}$ model from around $t_T = 2002$ onwards, but this effect is not observed for the MLL model, see Figure 3.21 (right). In most cases, the proportion of the short-term components settle at about 80%, which is plausible since the most recent information is considered more relevant.

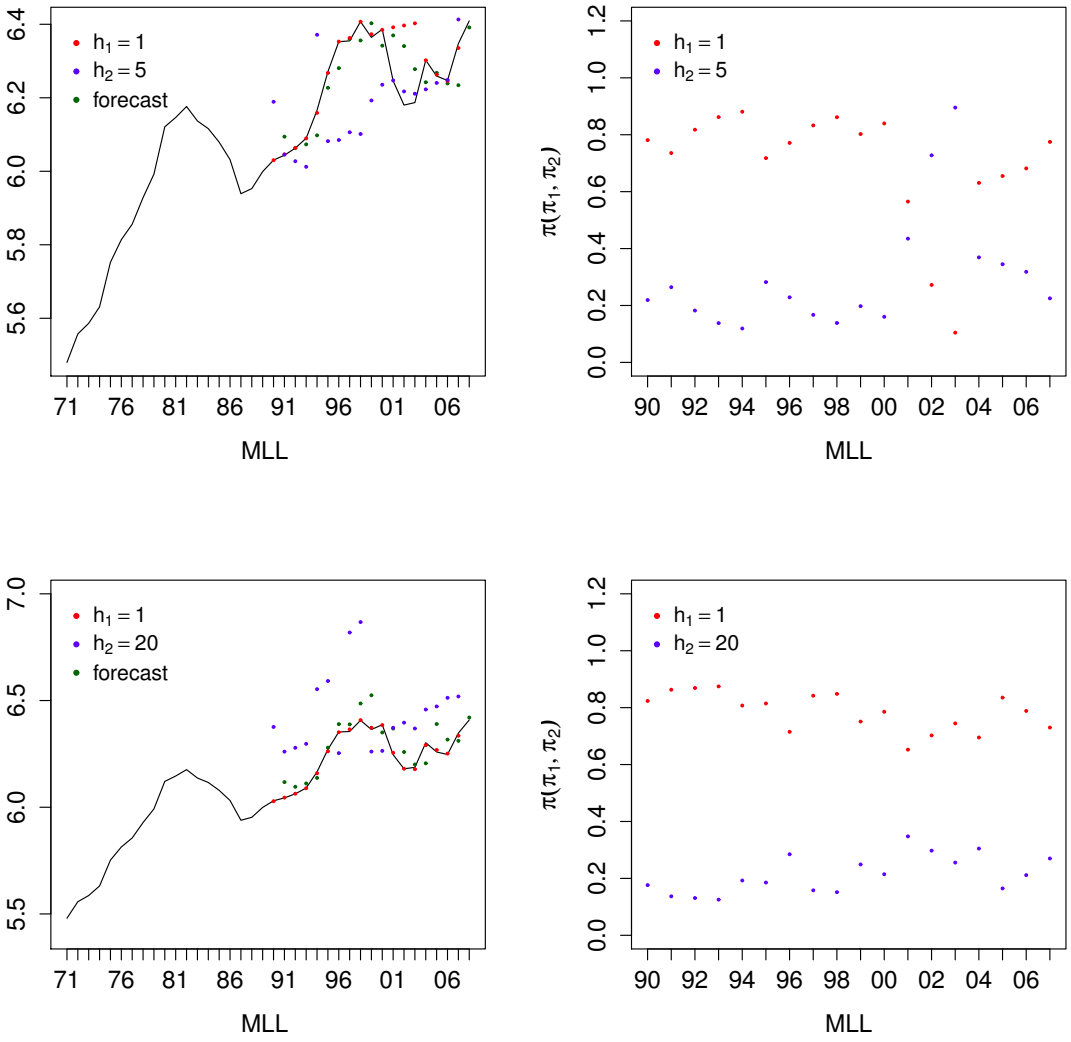

Figure 3.21: Data for Bolivia showing the parameters $\hat{\beta}_{0k}(t_T)$ fitted using MLL and resulting forecasts at $\hat{y}_{T+1}^{\text{MLL}}$ (left); and fitted parameters $\hat{\pi}_k(t_T)$ (right).

The additional information provided by the long-term component in the MLL model is useful for short-term predictions, but this advantage vanishes for $m > 1$ due to the increased variance.

For the Lebanon data, the errors shown in Table 3.4 are overall of a larger magnitude than those for Bolivia, due to the larger variability of the data itself, but otherwise the picture obtained previously is confirmed: using the MLC$^{(1)}$ model leads generally to favourable results, with the MLL model becoming superior only for $m = 1$ and a large long-term bandwidth. Both the ARIMA and Holt models can compete with the MLC$^{(1)}$ model only for $m = 1$ and $h = (1, 20)$. Table 3.6 presents the optimal

Model	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(4)	SARE(1)	SARE(2)	SARE(3)	SARE(4)
				$(h_1, h_2) = (1, 5)$				
$MLC^{(1)}$	0.16	0.18	0.18	0.17	10.95	11.51	11.48	10.75
$MLC^{(2)}$	0.45	0.83	1.14	1.32	18.46	23.75	28.04	30.19
MLL	0.24	0.70	1.21	1.67	11.70	19.02	27.07	31.55
				$(h_1, h_2) = (1, 20)$				
$MLC^{(1)}$	0.30	0.32	0.35	0.35	15.95	16.62	17.13	16.83
$MLC^{(2)}$	0.54	0.93	1.26	1.47	21.33	26.71	31.24	33.38
MLL	0.24	0.60	1.05	1.40	12.24	17.76	24.38	29.05
Holt	0.34	0.69	0.95	1.07	15.26	21.25	26.88	28.19
ARIMA	0.31	0.71	1.05	1.26	14.05	20.94	26.92	27.66

Table 3.4: The SSRE and SARE of forecasting for Lebanon from 1991 to 2008 using fixed bandwidths.

Model	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(4)	SARE(1)	SARE(2)	SARE(3)	SARE(4)
				$h = 1$				
$NLC^{(1)}$	0.04	0.05	0.05	0.04	5.50	5.74	5.74	5.15
				$h = (0.5, 1)$				
$MLC^{(1)}$	0.02	0.02	0.02	0.02	3.17	3.17	3.39	3.14
				$h = 1$				
$NLC^{(2)}$	0.33	0.68	0.99	1.18	14.95	21.06	26.28	27.28
				$h = (0.5, 1)$				
$MLC^{(2)}$	0.31	0.65	0.98	1.17	14.06	20.42	26.13	26.87

Table 3.5: The SSRE and SARE of forecasting for Lebanon from 1991 to 2008 using fixed bandwidths.

bandwidths for prediction using the NLL and MLL models, which are $h = 1.05$ and $h = (1, 5.89)$, respectively and it shows the corresponding SSRE for m-step ahead forecasting. It seems that the MLL model is better than the NLL model with an error SSRE=0.23×10^{-3}. In respect of the $MLC^{(2)}$ and $NLC^{(2)}$ models, the new methodology used for bandwidth selection has not been applied for the same reason as stated in the Bolivia data study case.

For the data for Greece, the situation is different, due to the specific nature of this time series, which shows an increase that is close to the linear. Here the ability to model a local linear trend plays a strong role in enhancing prediction, and due to the stability of this trend, this continues to hold for the forecast lags $m > 1$. However, it became clear that the $MLC^{(2)}$ model provided poor performance in comparison with the other models, especially for the long-term trend component. The optimal bandwidths of the NLL and MLL models were $h = 1.4$ and $h = (1, 4.2)$, respectively, as shown in Table

Model	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(4)
		$h = 1.05$		
NLL	0.38	0.95	1.69	2.31
		$(h_1, h_2) = (1, 5.89)$		
MLL	0.23	0.64	1.12	1.52

Table 3.6: The SSRE of forecasting for Lebanon from 1991 to 2008 using optimised bandwidths.

Model	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(4)	SARE(1)	SARE(2)	SARE(3)	SARE(4)
				$(h_1, h_2) = (1, 5)$				
$MLC^{(1)}$	0.03	0.02	0.02	0.02	5.11	4.43	3.92	3.64
$MLC^{(2)}$	0.07	0.10	0.15	0.22	7.68	9.71	11.62	13.85
MLL	0.01	0.03	0.07	0.14	2.71	3.53	5.68	8.32
				$(h_1, h_2) = (1, 20)$				
$MLC^{(1)}$	0.12	0.11	0.11	0.10	10.71	10.42	9.98	9.67
$MLC^{(2)}$	0.17	0.23	0.29	0.36	12.80	14.36	16.08	17.67
MLL	0.02	0.04	0.10	0.20	2.82	4.13	6.29	9.04
Holt	0.02	0.04	0.09	0.17	3.14	4.44	7.11	10.20
ARIMA	0.02	0.05	0.12	0.22	3.27	5.42	8.20	11.28

Table 3.7: The SSRE and SARE of forecasting for Greece from 1991 to 2008 using fixed bandwidths.

Model	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(4)	SARE(1)	SARE(2)	SARE(3)	SARE(4)
				$h = 1$				
$NLC^{(1)}$	0.003	0.003	0.004	0.004	1.43	1.51	1.65	1.77
				$h = (0.5, 1)$				
$MLC^{(1)}$	0.004	0.004	0.009	0.009	0.55	0.55	0.63	0.70
				$h = 1$				
$NLC^{(2)}$	0.02	0.05	0.09	0.14	3.91	5.77	8.06	10.70
				$h = (0.5, 1)$				
$MLC^{(2)}$	0.02	0.04	0.08	0.13	3.33	5.16	7.43	10.42

Table 3.8: The SSRE and SARE of forecasting for Greece from 1991 to 2008 using fixed bandwidths.

3.9, and are used to find m-step-ahead forecasts. From the data shown in Table 3.9, it is clear that the MLL model is better than the NLL model according to SSRE for m-step-ahead forecasts. Moreover, this result is supported by Table 3.7, which means that the MLL model has the ability to capture the optimal bandwidth from selected bandwidth. The optimal bandwidths for both the $MLC^{(2)}$ and $NLC^{(2)}$ models tend to be 0. In order to investigate the performance of prediction for different settings of bandwidths, the bandwidths were fixed as in Tables 3.7 and 3.8. Table 3.8 shows that the $MLC^{(1)}$ model produces a smaller error margin than the $MLC^{(2)}$ model, according the SSRE and SARE criteria.

In summary, the examples provide evidence for the superiority of the $NLC^{(1)}$ and $MLC^{(1)}$ methods, especially for greater forward lags and smaller bandwidths. Remarkably, the performance of the $MLC^{(1)}$ method almost does not depend on the forward lags. Here an apparent 'weakness' of the $MLC^{(1)}$ method, namely the non-adaptability

Model	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(4)
		$h = 1.4$		
NLL	0.02	0.04	0.10	0.17
		$(h_1, h_2) = (1, 4.2)$		
MLL	0.01	0.03	0.07	0.13

Table 3.9: The SSRE of forecasting for Greece from 1991 to 2008 using optimised bandwidths.

to linear trends, seems to turn into an advantage, as the technique does not 'learn' the direction of these local trends, and so avoids overshooting once the data take a turn. For the MLC$^{(i)}, i = 1, 2$, methods in general, the bandwidth choice of $h_2 = 0.5$ produces better results than $h_2 = 5$ and $h_2 = 20$, see Tables 3.2, 3.5 and 3.8. For the MLL model, this interpretation is less clear-cut, but it is right to say that the results, using the MLL model with $h_2 = 20$, were generally comparable to those obtained using the ARIMA and Holt models. It appears that the MLL method can only be recommended when $m = 1$ and h_2, is large, and here it is better than the MLC$^{(2)}$ model, because of its ability to fit the linear trend well. The MLC$^{(2)}$ model performed badly for prediction when compared to the other models.

It can be seen from Tables 3.3, 3.6 and 3.9, the MLL model had smaller error margin than the NLL model, and it performed well in prediction for all forward lags, which suggests that mixture models cam improve prediction compared with standard local linear model and traditional models such as the ARIMA and Holt models. In addition, the MLC$^{(i)}, i = 1, 2$, models show favourable performance for prediction compared to the NLC$^{(i)}, i = 1, 2$, models for small bandwidths $h_2 = 0.5$, and this can be seen in Tables 3.2, 3.5 and 3.8.

3.10 Conclusions

In conclusion, this chapter has presented a novel approach to forecasting based on localised mixtures of non-parametric regressions. Non-parametric regression allows a forecast to be calculated directly using historical data, as a local average of observed past values. In the first model, named the MLC model, local constant estimators were used to carry out the localised estimation step. In the second model, referred to as the MLL model, the MLC model was generalized using local linear estimators.

Estimation for these models was achieved using a kernel-weighted version of the EM-algorithm, and using exponential kernels with different bandwidths as weight functions. In addition, these localised mixture models show favourable property relating to identifiability, but this needs further consideration in some cases. In order to forecast, several approaches for prediction at the time t_{T+m}, these models were investigated as

shown above using a simulation study. The simulation was conducted on two different examples of data, in order to investigate in which cases the MLL and MLC models showed good performance for prediction in comparison with traditional models, such as the ARIMA and Holt models. In the first example, the data included seasonal and linear trends with uncorrelated errors. However, in the second example, the linear trend was eliminated and the errors were correlated.

A new approach for bandwidth selection for prediction was presented in this chapter. In each example given, two simulation scenarios were considered depending on the type of bandwidth selection. In the first case, the optimal bandwidth was optimised for prediction however in the second case, two settings of bandwidths were considered: $h = (0.5, 1)$ and $h = (1, 4)$. The results suggest that the MLL model can only improve predictions from time series data, as compared with the ARIMA and Holt models, that use long-term components and short forward lags, as in Example 1 in the simulation study and in the real data applications. However, further forecasting methods should be investigated to enhance and explore this comparison. In addition, the $\text{MLC}^{(1)}$ model shows good performance in prediction for short-term components and non-linear trend data as seen in Example 2 of the simulation and in the real data applications.

Although the simulation study was restricted the choice of one of the bandwidths to be 1, which means that one of the bandwidths was not optimised, the results show competitive and challenging arguments in favour of using the $\text{MLC}^{(2)}$ and MLL models in comparison to using other models used for prediction. This suggests that further study on the performance of $\text{MLC}^{(2)}$ and MLL models for prediction when all bandwidths are optimized should be undertaken. In addition, it was found that the new methodology for bandwidth selection in real data applications was more useful for MLL and NLL models than for MLC and NLC models. As a result, it is strongly recommended that this methodology of bandwidth selection should be used in the MLL and NLL models. In addition, there is strong evidence as presented above, that the $\text{MLC}^{(2)}$ and MLL models can select the good bandwidth which means that bandwidth selection is recommended to these approaches for prediction. This feature gives additional advantage to the $\text{MLC}^{(2)}$ and MLL models for use in prediction.

Chapter 4

Double-localised mixture models for prediction

4.1 Introduction

In this chapter, a developed model based on localised mixture regression models is presented, in order to improve the prediction from time series using information from other time series. The data under study is the same data as discussed in Section 2.3.1 of Chapter 2, that is the annual energy use data. Before 1995, there was no information available about energy consumption for some countries. As a result, the study in this chapter is restricted to the period between 1995 and 2011. This is to avoid issues relating to the missing data in our statistical analysis, which needs further consideration in future. In Figure 4.1, we have extended the analysis for the data sets, that is limited in Section 2.3.1 of Chapter 2 by selected years. It is clear that there are two clusters of data, that appear over time, as shown in Section 2.3.1 in Chapter 2. If we apply the bootstrap likelihood ratio test, see Section 1.4.2 of Chapter 1 for more detail, a mixture of two Gaussian distributions fit the data sets over time. The two components are visualised simultaneously over time by a sequence of 2–boxplots of log of energy use data as shown in Figure 4.1. As in Section 2.3.1 in Chapter 2, one group corresponds to high energy use countries, and one group corresponds to low energy use countries. In addition, the median changes slightly in either of the two groups, except in 2009, which appears to considerable decrease in both groups. The low-energy-use

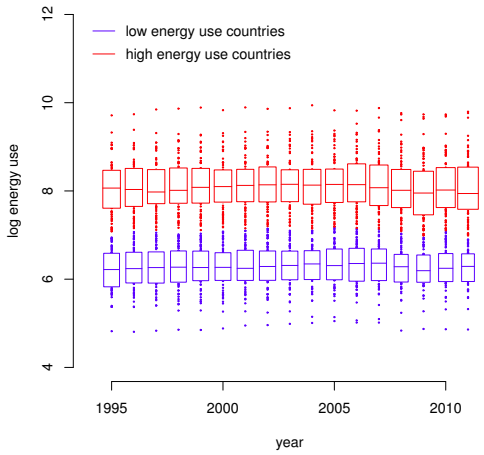

Figure 4.1: 2–Boxplots of log energy use data between 1995 to 2011

group and the high-energy-use group are of equivalent sizes up until 2007, when the high-energy-use group grows larger, and this is represented by the boxes getting wider. This trend can be interpreted as, in recent years, more countries have become developed and so energy consumption has increased.

Mixture of local regression models in a similar way to Chapter 3 are stated firstly. The models are the $\text{MLC}^{(i)}, i = 1, 2$ and MLL models, but with equal bandwidths for all components. These models are used to fit the data of all countries at a given time point t_T, in order to forecast energy use data for a given time series. In this case, the $\text{MLC}^{(i)}, i = 1, 2$ and the MLL models are closely related to model (3.4.1) by Huang et al. [44]. For these models, we can interpolate the estimates of the fitted energy use data to find the estimate of energy use data between any two years as in model (3.4.1). Figure 4.2 shows a sequence of β's fitted using Equation (3.2.5) of the $\text{MLC}^{(i)}, i = 1, 2$ models (left) and a sequence of β_0's fitted from Equation (3.3.4) of the MLL model (right) as outlined in Chapter 3 with 2–boxplots over time. We can class this methodology as a regression technique to fit a multi-valued non-parametric local regression. It is clear from Figure 4.2, that the $\text{MLC}^{(i)}, i = 1, 2$ and MLL models show favourable features that can separate the data into two groups. The weighted medians of a mixture in each year is approximating to the fitted β of the $\text{MLC}^{(i)}, i = 1, 2$ models as shown in Figure 4.2 (left), and the fitted β_0 of the MLL model as shown in Figure 4.2 (right).

Suppose for a certain country at a given time point t_T, we would like to forecast the

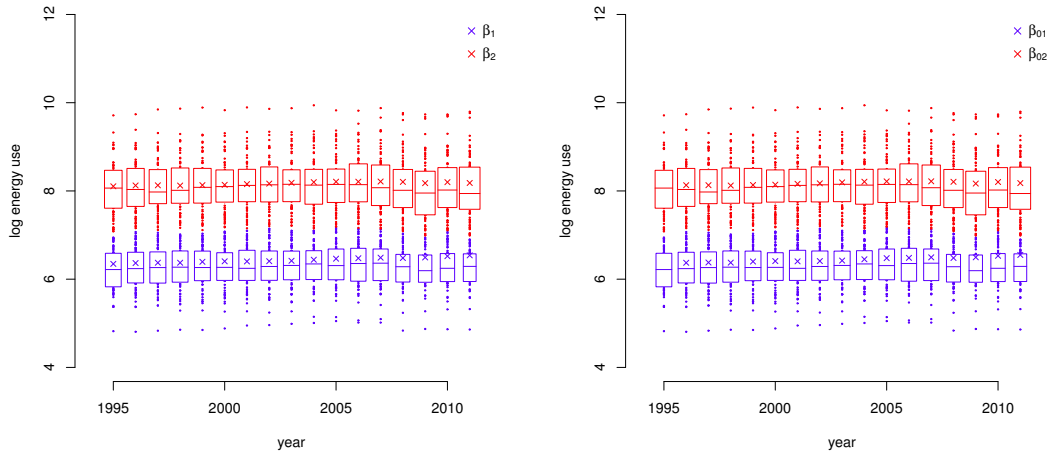

Figure 4.2: Data from all countries, 2–boxblots with parameters $\hat{\beta}_k(t_T)$ fitted using MLC$^{(1)}$ (left) and with parameters $\hat{\beta}_{0k}(t_T)$ fitted using MLL$^{(1)}$ (right) simultaneously over time.

energy use of this country based on the MLC$^{(i)}, i = 1, 2$ and MLL models, and using the information from other countries. The obvious technique for prediction in this case would be to use the respective posterior probabilities $r_{ijk}, i = 1, \ldots, T, j = 1, \ldots, J, k = 1, \ldots, K$, of the country of interest and the fitted parameters at the time point t_T to find the m-step-ahead forecasts y_{Tj+m} locally, where K is the number of mixture components and J is the number of time series. The technique for prediction relies on the historical information from all countries. So far, the effect of countries, which have similar data pattern to the target country, on the target country is not considered in prediction. This effect could play a role in providing additional information, which could contribute in improving the prediction for the target country.

4.1.1 Motivation

The motivation in this chapter was to improving the prediction for a very noisy time series using information from anther time series at a certain time point t_T and using past observations. In other words, for a given country, the aim was to use information from all countries, especially countries, that show similar energy use patterns over time, but without placing any consideration on the geographical aspects of these countries.

For this purpose two bandwidths were used: the horizontal h_k and the vertical v_k

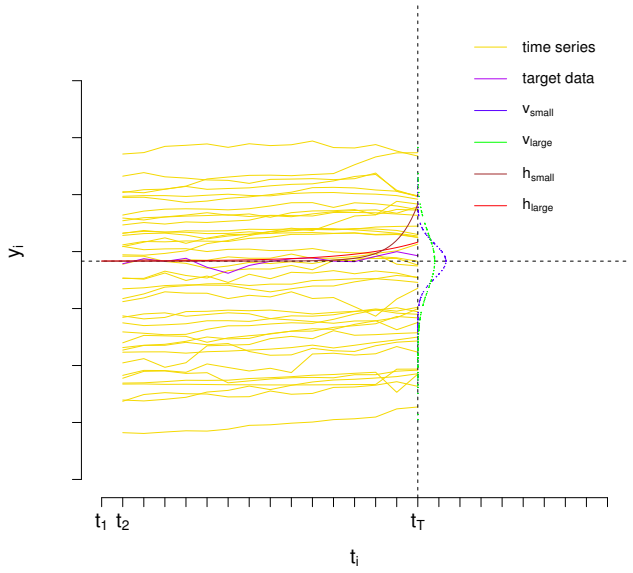

Figure 4.3: J Series of the time series, j-th time series, vertical kernels using small and large bandwidths v_{small} and v_{large}, and horizontal kernels using small and large bandwidths h_{small} and h_{large}

bandwidths for each component k. The horizontal bandwidths h_k control the amount of historical information gained from past data in a local neighbourhood. The vertical bandwidths determine how many countries should be taken into account in order to improve the prediction for the given country. Figure 4.3 illustrates the idea of using horizontal and vertical kernels for each component. The blue and green curves represent vertical kernels with small and large vertical bandwidths, respectively. However, the brown and red curves indicate horizontal kernels with small and large horizontal bandwidths, respectively. In addition, the purple time series is the time series of the country of interest, and the gold time series is time series for all other countries. As shown in Figure 4.3, in order to investigate, for example, the influence of political events on the energy consumption of a target country under consideration of the other countries, it is convenient to use a short-term trend bandwidth (h_{small}) and a large-term bandwidth v_{large}. However, if the aim is to investigate the trends of the target country, via other countries, which over a certain time span develop better economic policies and so forth, then using a long-term trend bandwidth h_{large} and a short-term bandwidth v_{small} appears adequate, that affects similar countries to the target country in data pattern. In this context, we will also see that the K–boxplot is a useful graphical

tool that can be used to visualise the similarity of countries in the data pattern [67].

This chapter aims to show how to use the localised mixture models, as presented in Chapter 3 in order to improve predictions for very variable time series using additional information provided by other time series. For this purpose, two models are proposed that can improve the prediction of a target time series using anther time series at a given time point t_T. The MLC and MLL models were developed using the additional bandwidths $v_k, k = 1, \ldots, K$, which determine the size of the local neighbourhoods around a given observation point y_{ij} for the target time series, and at a given time point t_T. These two developed models will be referred to as: a mixture model using local constant estimators and vertical kernels (MLCV) and a mixture model using local linear estimators and vertical kernels (MLLV).

Estimation for these models was achieved using a double-kernel-weighted version of the EM–algorithm, and using exponential kernels with different bandwidths h_k and normal kernels with different bandwidths v_k around a target observation y_{Tj} and at a given time point t_T. Nadaraya-Watson and local linear estimators were used to carry out the localised estimation step.

For prediction at time point t_{T+m}, adequate approaches were provided for each local method, and were compared to competing forecasting routines. By modelling MLCV and MLLV models with the bandwidths h_k and v_k, we can obtain an estimated mixture probabilities, that are informative for the amount of information available in the data sets, at the scale of resolution corresponding to each bandwidth.

At the end of this chapter, three examples will be presented to assess the accuracy of the forecasting undertaken using the MLCV and MLLV models. In addition, a comparison will be presented between the double-localised mixture models used for prediction and more traditional ones, such as the ARIMA and Holt models, which are popular used approaches for time series forecasting. It is worth to mention that there are no comparative methods to double-localised mixture models for prediction in the literature.

The rest of chapter is organised as follows: An overview of the MLCV model and estimation for this model can be seen in Section 4.2. Section 4.3 presents the MLLV model and how to estimate the parameters of this model. Section 4.4 discusses model

selection for the MLCV and MLLV models. Forecast approaches are proposed and discussed for models under study in Section 4.5. Section 4.6 presents real data examples, giving energy use for the Ivory Coast, Albania and Lithuania from 1995 to 2011. The results are compared to point forecasts obtained using the ARIMA and Holt exponential smoothing models. Finally, Section 4.7 presents the conclusions of this chapter of the thesis.

4.2 Mixture models using local constant kernel estimators and vertical kernels (MLCV)

If we assume J series of time series in the form $\{(t_i, y_{ij}) : i = 1, \ldots, T, j = 1, \ldots, J\}$, where J is the number of time series, it is possible to consider a double-localised mixture of K non-parametric regressions $m_k(t_i)$, $k = 1, \ldots, K$, where K is a fixed number of components, such that $K < J$ and $m_k(t_i)$ is a non-parametric regression function at the k-th component. At the time point t_T and for the j-th time series, it is possible to define a locally constant model $m_k(t_i) \approx m_k(t_T)$ in a neighbourhood of t_T by using Taylor's expansion, which can be denoted as $\beta_{kj}(t_T)$. Thus, the model can be written as follows

$$y_{ij} = \begin{cases} \beta_{1j}(t_T) + \epsilon_{ij1}, & \text{with proportion } \pi_{1j}(t_T) \\ \vdots & \\ \beta_{Kj}(t_T) + \epsilon_{ijK}, & \text{with proportion } \pi_{Kj}(t_T) \end{cases} \tag{4.2.1}$$

where $\beta_{1j}(t_T), \ldots, \beta_{Kj}(t_T)$ are unknown fixed parameters, that depend on the target point t_T, $\pi_{kj}(t_T)$ is the proportion of the k-th component, such that $0 \leq \pi_{kj} \leq 1$ and $\sum_{k=1}^{K} \pi_{kj} = 1$, and the errors $\epsilon_{ijk} \sim N(0, \sigma_j^2)$ are independently distributed. For ease of notation, the dependence of the parameters on t_T was regularly suppressed.
For the given component k, the aim was to obtain estimators of π_{kj}, β_{kj} and σ_j at a time point t_T and for the j-th time series using past historical information and information provided from other time series. In the estimation step, the EM–algorithm was used to carry out the localised estimation.

Therefore, it is proposed that G_j is the random vector, which indicates a class $k \in$

$1, \ldots, K$, where the following applies:

$$G_{ijk} = \begin{cases} 1, & \text{if observation } i \text{ of } j\text{-th time series belongs to component } k \\ 0, & \text{otherwise} \end{cases} \tag{4.2.2}$$

and $P(G_j = k) = \pi_{kj}$. Denote the following:

$$f_{ijk} = P(y_{ij}|G_j = k) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(y_{ij} - \beta_{kj})^2}{2\sigma_j^2}\right)$$

Then the following applies:

$$P(y_{ij}, G_j = k) = P(y_{ij}|G_j = k)P(G_j = k) = \pi_{kj} f_{ijk}$$

The one-sided component-wise weight functions W_k and V_k were anchored at t_T and were used, where W_k was the exponential kernel defined in Equation (3.2.2) as shown in Chapter 3 and V_k is a Gaussian kernel, that can be introduced as follows:

$$V_k(y_{il}, y_{ij}) = \frac{1}{v_k\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_{il} - y_{ij}}{v_k}\right)^2\right] \tag{4.2.3}$$

where $i = 1, \ldots, T$, $j \in \{1, \ldots, J\}$, $l = 1, \ldots, J$ and v_k is a vertical bandwidth for component k.

The bandwidth h_k for the weight kernel W_k is denoted horizontal bandwidth, because it controls the size of the local neighbourhood horizontally, and this provides model information from the data about trends. However, the bandwidth v_k was called vertical, because it controls the symmetric local neighbourhood around a target point y_{Tj} vertically, in order to provide information about similar time series in the data pattern. The larger the local neighbourhood size, then the more information that is provided from other time series to a target time point, and this contributes to improving the prediction of certain time series. The weight function, which is used to control the weights of the observation points y_{ij}, as used in the local neighbourhood plays a role in prediction for a given time series. It gives more weights to observations close to a given point y_{Tj} from a target time series. Therefore, we can assume that for the observation y_{ij}, the value of G_j is known, and this gives us the "complete "data (y_{ij}, G_{ij1},

$\ldots, G_{ijK})$, with the local probability, as follows:

$$P(y_{ij}, G_{ij1}, \ldots, G_{ijK}) = \prod_{k=1}^{K} (f_{ijk}\pi_{kj})^{G_{ijk}W_k(t_i,t_T)V_k(y_{il},y_{ij})}$$

The corresponding local likelihood function $\mathcal{L}^*$, called complete local likelihood, can be denoted as follows:

$$\mathcal{L}^*(\mathbf{\Phi}_c|y_{1j}, \ldots, y_{Tj}, G_{ij1}, \ldots, G_{ijK}) = \prod_{i=1}^{T}\prod_{l=1}^{J}\prod_{k=1}^{K} (f_{ijk}\pi_{kj})^{G_{ijk}W_k(t_i,t_T)V_k(y_l,y_{ij})}$$

where $\mathbf{\Phi}_c = (\pi_{1j}, \ldots, \pi_{(K-1)j}, \beta_{1j}, \ldots, \beta_{Kj}, \sigma_j)$ is a vector containing all the parameters in the mixture model MLCV. Then, the local log-likelihood function ℓ^* can be denoted as follows:

$$\begin{aligned}
\ell^*(\mathbf{\Phi}_c|y_{1j}, \ldots, y_{Tj}, G_{ij1}, \ldots, G_{ijK}) &= \log \mathcal{L}^*(\mathbf{\Phi}_c|y_{1j}, \ldots, y_{Tj}, G_{ij1}, \ldots, G_{ijK}) \\
&= \sum_{i=1}^{T}\sum_{l=1}^{J}\sum_{k=1}^{K} G_{ijk}W_k(t_i,t_T)V_k(y_{il},y_{ij}) \log \pi_{kj} \\
&+ G_{ijk}W_k(t_i,t_T)V_k(y_{il},y_{ij}) \log f_{ijk}
\end{aligned}$$

If we interpret the π_{kj} as 'prior' probability of class membership, then the posterior probabilities of class membership can be produced using Bayes' theorem. As the G_{ijk} are in fact unknown, they can be replaced with conditional expectations as follows:

$$r_{ijk} = E(G_{ijk}|y_{ij}) = P(G_{ijk} = 1|y_{ij}) = P(G_j = k|y_{ij})$$

Using Bayes' theorem, the following can be applied:

$$r_{ijk} = P(G_j = k|y_{ij}) = \frac{P(G_j = k)P(y_{ij}|G_j = k)}{\sum_\ell P(G_j = \ell)P(y_{ij}|G_j = \ell)} = \frac{\pi_{kj}f_{ijk}}{\sum_\ell \pi_{\ell j}f_{ij\ell}}$$

In the v-th cycle of the EM–algorithm iteration, we have the estimates $\pi_{kj}^{(v)}$, $\beta_{kj}^{(v)}$ and $\sigma_j^{(v)}$. Then, in the $(v+1)$-th cycle, using the estimates $\pi_{kj}^{(v)}$, $\beta_{kj}^{(v)}$ and $\sigma_j^{(v)}$, the posterior probabilities $r_{ijk}^{(v+1)}$ are equivalent to the following:

$$r_{ijk}^{(v+1)} = P(G_{ijk} = 1|y_{ij}) = \frac{\pi_{kj}^{(v)} \exp\left(-\frac{1}{2}(\frac{y_{ij}-\beta_{kj}^{(v)}}{\sigma_j^{(v)}})^2\right)}{\sum_{\ell=1}^{K} \pi_{\ell j}^{(v)} \exp\left(-\frac{1}{2}(\frac{y_{ij}-\beta_{\ell j}^{(v)}}{\sigma_j^{(v)}})^2\right)} \tag{4.2.4}$$

Equation (4.2.4) is identical to the E–step of the EM–algorithm. In the M–step, for

the $\pi_{kj}^{(v+1)}$, a Lagrange multiplier is applied by setting the following:

$$\partial\left(Q(\boldsymbol{\Phi}_c|\boldsymbol{\Phi}_c^{(v)}) - \lambda(\sum_{k=1}^{K}\pi_{kj}^{(v+1)} - 1)\right)/\partial\pi_{kj}^{(v+1)} = 0, \quad k = 1, \ldots, K$$

since $\sum_{k=1}^{K}\pi_{kj}^{(v+1)} = 1$, the following can be obtained:

$$\pi_{kj}^{(v+1)} = \frac{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})}{\sum_{i=1}^{T}\sum_{l=1}^{J}\sum_{k=1}^{K} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})} \tag{4.2.5}$$

By setting $\partial\ell^*/\partial\beta_{kj}^{(v+1)} = 0$ and $\partial\ell^*/\partial\sigma_j^{(v+1)} = 0$, it is possible to obtain estimates as follows:

$$\beta_{kj}^{(v+1)} = \frac{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij}) y_{ij}}{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})} \tag{4.2.6}$$

$$\sigma_j^{2(v+1)} = \frac{\sum_{i=1}^{T}\sum_{l=1}^{J}\sum_{k=1}^{K} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(y_{ij} - \beta_{kj}^{(v+1)})^2}{\sum_{i=1}^{T}\sum_{l=1}^{J}\sum_{k=1}^{K} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})} \tag{4.2.7}$$

For more details about the calculations of Equations (4.2.6) and (4.2.7), see Appendix D.

It is necessary to iteratively update the E–step and M–step using different initial values at each grid point locally until the algorithm converges. The iteratively updates the E–step and M–step locally until $\hat{\boldsymbol{\Phi}}_c^{(v)} \approx \hat{\boldsymbol{\Phi}}_c^{(v+i)}, i = 1, 2, \ldots$. According to this stopping criterion, the EM–algorithm stopped at $v = 200$ in this study. As a result, the double-kernel-weighted version of the EM–algorithm is considered locally, in order to estimate π_{kj}, β_{kj} and σ_j for each component k and for the j-th time series at a given time point t_T. Once the estimates of π_{kj}, β_{kj} and σ_j in Equations 4.2.5–4.2.7 are obtained, different approaches of forecasting can be proposed to find the m-step-ahead forecasts for the j-th time series and at a given time point t_T. These approaches will be presented later in Section 4.5.2 of this chapter.

4.3 Mixture models using local linear kernel estimators and vertical kernels (MLLV)

In this section, the model MLCV was generalized using local linear estimators rather than local constant estimators to carry out the localised estimation step, and as such,

is named the MLLV model. The motivation behind using local linear estimators was to improve prediction from time series data, which has a linear trend, as discussed previously in Chapter 3. For a given time series y_{ij} and at a time point t_T, the k-th non-parametric regression function was approximated as $m_k(t_i) \approx m_k(t_T)+m_k^{(1)}(t_T)(t_i-t_T)$, and this motivated the localised model as follows

$$y_{ij}(t_i) = \begin{cases} \beta_{01j}(t_T) + \beta_{11j}(t_T)(t_i - t_T) + \epsilon_{ij1}, & \text{with proportion } \pi_{1j}(t_T) \\ \vdots & \\ \beta_{0Kj}(t_T) + \beta_{1Kj}(t_T)(t_i - t_T) + \epsilon_{ijK}, & \text{with proportion } \pi_{Kj}(t_T) \end{cases} \tag{4.3.1}$$

where the intercepts β_{0kj} and the slopes β_{1kj} are fixed unknown coefficients, that depend implicitly on a fixed time point t_T, whereas $\pi_{kj}(t_T)$ is the proportion of the k-th component, such that $0 \leq \pi_{kj} \leq 1$ for $k = 1, \ldots, K$ and $\sum_{k=1}^{K} \pi_{kj} = 1$, K is the number of components, such that $K < J$ and the errors $\epsilon_{ijk} \sim N(0, \sigma_j^2)$ are independently distributed.

For the j-th time series and at a time point t_T, data is weighted by exponential and normal kernels for each component, which are defined as in Equations (3.2.2) and (4.2.3). In the estimation step, the EM–algorithm was used to estimate $\mathbf{\Phi}_l = (\pi_{1j}, \ldots, \pi_{(K-1)j}, \beta_{01j}, \ldots, \beta_{0Kj}, \beta_{11j}, \ldots, \beta_{1Kj}, \sigma_j)$, which is a vector containing all the parameters in the mixture model MLLV. Let G_j be a random vector, which is defined as in Equation (4.2.2). Then, we have $P(G_j = k) = \pi_{kj}$ and we denote

$$f_{ik} \equiv P(y_i|G_j = k) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(y_{ij} - \beta_{0k} - \beta_{1k}(t_i - t_T))^2}{2\sigma_j^2}\right)$$

Then the following applies:

$$P(y_{ij}, G_j = k) = P(y_{ij}|G_j = k)P(G_j = k) = \pi_{kj} f_{ijk}$$

Therefore, we assume now that, for an observation y_{ij} of the j-th time series, the value of G_j is known. This gives the "complete" data $(y_{ij}, G_{ij1}, \ldots, G_{ijK})$, with local probability as follows:

$$P(y_{ij}, G_{ij1}, \ldots, G_{ijK}) = \prod_{k=1}^{K} (f_{ijk}\pi_k)^{G_{ijk}W_k(t_i,t_T)V_k(y_{il},y_{ij})}$$

Then, the corresponding local likelihood function $\mathcal{L}^*$, which is called complete local

likelihood [62], is as follows:

$$L^*(\mathbf{\Phi}_l|y_{1j},\ldots,y_{Tj},G_{ij1},\ldots,G_{ijK}) = \prod_{i=1}^{T}\prod_{l=1}^{J}\prod_{k=1}^{K}(f_{ijk}\pi_k)^{G_{ijk}W_k(t_i,t_T)V_k(y_l,y_{ij})}$$

Therefore, the log local likelihood function ℓ^* is as follows:

$$\begin{aligned}\ell^*(\mathbf{\Phi}_l|y_{1j},\ldots,y_{Tj},G_{ij1},\ldots,G_{ijK}) &= \log L^*(\mathbf{\Phi}_l|y_{1j},\ldots,y_{Tj},G_{ij1},\ldots,G_{ijK})\\ &= \sum_{i=1}^{T}\sum_{l=1}^{J}\sum_{k=1}^{K}G_{ijk}W_k(t_i,t_T)V_k(y_{il},y_{ij})\log\pi_k\\ &+ G_{ijk}W_k(t_i,t_T)V_k(y_{il},y_{ij})\log f_{ijk} \qquad (4.3.2)\end{aligned}$$

As the G_{ijk} are in fact unknown, we replace them by their conditional expectations as follows

$$r_{ijk} \equiv E(G_{ijk}|y_{ij}) = P(G_{ijk}=1|y_{ij}) = P(G_j=k|y_{ij})$$

Using Bayes' theorem, one has

$$r_{ijk} = P(G_j=k|y_{ij}) = \frac{P(G_j=k)P(y_{ij}|G_j=k)}{\sum_\ell P(G_j=\ell)P(y_{ij}|G_j=\ell)} = \frac{\pi_k f_{ijk}}{\sum_\ell \pi_\ell f_{ij\ell}}$$

Then, in the $(v+1)$-th cycle of the EM–algorithm, using the estimates $\pi_{kj}^{(v)}$, $\beta_{0kj}^{(v)}$, $\beta_{1kj}^{(v)}$ and $\sigma_j^{(v)}$, the posterior probabilities $r_{ijk}^{(v+1)}$ can then be given as follows

$$r_{ijk}^{(v+1)} = \frac{\pi_{kj}^{(v)}\exp\left(-\frac{1}{2}(\frac{y_{ij}-\beta_{0kj}^{(v)}-\beta_{1kj}^{(v)}(t_i-t_T)}{\sigma_j{}^{(v)}})^2\right)}{\sum_{\ell=1}^{K}\pi_{\ell j}^{(v)}\exp\left(-\frac{1}{2}(\frac{y_{ij}-\beta_{0\ell j}^{(v)}-\beta_{1\ell j}^{(v)}(t_i-t_T)}{\sigma_j^{(v)}})^2\right)} \qquad (4.3.3)$$

For the M–step, the estimates of $\pi_{kj}^{(v+1)}$, $\beta_{0kj}^{(v+1)}$,$\beta_{1kj}^{(v+1)}$ and $\sigma_j^{(v+1)}$ are found as follows: One needs to apply a Lagrange multiplier for $\pi_{kj}^{(v+1)}$ since $\sum_{k=1}^{K}\pi_{kj}^{(v+1)}=1$. Setting

$$\partial\left(Q(\mathbf{\Phi}_l|\mathbf{\Phi}_l^{(v)}) - \lambda(\sum_{k=1}^{K}\pi_k^{(v+1)}-1)\right)/\partial\pi_k^{(v+1)} = 0, \quad k=1,\ldots,K$$

One obtains

$$\pi_{kj}^{(v+1)} = \frac{\sum_{i=1}^{T}\sum_{l=1}^{J}r_{ijk}^{(v+1)}W_k(t_i,t_T)V_k(y_{il},y_{ij})}{\sum_{i=1}^{T}\sum_{l=1}^{J}\sum_{k=1}^{K}r_{ijk}^{(v+1)}W_k(t_i,t_T)V_k(y_{il},y_{ij})} \qquad (4.3.4)$$

In addition, the estimators of $\beta_{0k}^{(v+1)}$,$\beta_{1k}^{(v+1)}$ and $\sigma_j^{(v+1)}$ are as follows:

$$\beta_{0kj}^{(v+1)} = \frac{S_{k,T,j,2}S_{k,T,j,0}^* - S_{k,T,j,1}S_{k,T,j,1}^*}{S_{k,T,j,2}S_{k,T,j,0} - S_{k,T,j,1}^2} \qquad (4.3.5)$$

$$\beta_{1kj}^{(v+1)} = \frac{S_{k,T,j,0}S^*_{k,T,j,1} - S_{k,T,j,1}S^*_{k,T,j,0}}{S_{k,T,j,2}S_{k,T,j,0} - S^2_{k,T,j,1}} \tag{4.3.6}$$

where

$$S_{k,T,j,s} = \sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_{T+m}) V_k(y_{il}, y_{ij})(t_i - t_T)^s$$

and

$$S^*_{k,T,j,s} = \sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_{T+m}) V_k(y_{il}, y_{ij})(t_i - t_T)^s y_{ij}$$

$$\sigma_j^{2(v+1)} = \frac{\sum_{i=1}^{T}\sum_{l=1}^{J}\sum_{k=1}^{K} r_{ijk}^{(v+1)} W_k(t_i, t_{T+m}) V_k(y_{il}, y_{ij})(y_{ij} - \beta_{0kj}^{(v+1)} - \beta_{1kj}^{(v+1)}(t_i - t_T))^2}{\sum_{i=1}^{T}\sum_{l=1}^{J}\sum_{k=1}^{K} r_{ijk}^{(v+1)} W_k(t_i, t_{T+m}) V_k(y_{il}, y_{ij})} \tag{4.3.7}$$

For more details about the calculations of Equations (4.3.5)- (4.3.7), see Appendix D.

The double-kernel-weighted version of the EM–algorithm applies iteration Equations (4.3.4)–(4.3.7) until convergence occurs. In addition, the stopping criterion of the EM–algorithm used is the same as the stopping criterion used for the MLCV model in Section 4.2.Once the bandwidths h_k and v_k for each component k are set, and the estimates from Equations (4.3.3)–(4.3.7) are found, then an approach for prediction is used to predict a future observation. This approach will be shown in Section 4.5.2 of this chapter later.

4.4 Model selection

Model selection for the MLCV and MLLV models includes the selection of the number of components K and the bandwidths h_k and v_k. Choosing the number of components is a very important issue when using mixture models. In the applications of this chapter, the number of components was fixed at 2, especially for the real data applications. However, further research should be considered in relation to the selection of the number of components for double-localised mixture models.

There is no doubt that the selection of the pairs of bandwidths, the horizontal and vertical bandwidths (h_k, v_k), for each component k for MLCV and MLLV models play a central role in making the model efficient for prediction. The better the bandwidths used, then the more accurate predictions will be. In this research, different choices of fixed horizontal and vertical bandwidths were used on real data. However, a com-

prehensive simulation study for prediction using MLCV and MLLV models based on optimal pairs of bandwidths has been put to once side for future study. The selection of K and the bandwidths h_k and v_k might affect each other, and so this property should be taken into account in a simulation study.

4.5 Forecasting

In this section, different approaches of prediction based on models in Sections 4.1, 4.2, and 4.3 are presented for a given time series and at a time point t_T. Once the bandwidths (h_1, h_2) and (v_1, v_2) are determined, different approaches of prediction based on the selected bandwidths are suggested for the MLC and MLL as in Section 4.1 and the MLCV and MLLV models. These approaches produce m-step-ahead forecasts as follows:

4.5.1 Forecasting using localised mixture models for multi-valued regression data

This section explores the forecast equation for the model, as discussed in Section 4.1 of this chapter. A new approach towards m-step-ahead forecasts for the j-th time series at a time point t_T using historical information from multiple time series at t_T is proposed. Once the MLC and MLL are fitted locally at a target point t_T, posterior probabilities taken from the E–step of the EM–algorithm can be derived for the j-th time series at a time point t_T, which can be denoted $r_{Tjk}, k = 1, \ldots, K$. In addition, the fitted estimates of $\mathbf{\Phi}_c = (\pi_1, \ldots, \pi_{K-1}, \beta_1, \ldots, \beta_K, \sigma_j)$ for the MLC model and $\mathbf{\Phi}_l = (\pi_1, \ldots, \pi_{K-1}, \beta_{01}, \ldots, \beta_{0K}, \beta_{11}, \ldots, \beta_{1K}, \sigma_j)$ for the MLL model are obtained. Then, we obtain the following m-step-ahead forecast equation using the MLC model for the j-th time series at t_T as follows:

$$\hat{y}^{\text{MLC}}_{(T+m)j} = \sum_{k=1}^{K} r_{Tjk}\hat{\beta}_k \tag{4.5.1}$$

In addition, once the MLL model is fitted for the j-th time series at t_T, the m-step-ahead forecasts equation is as follows:

$$\hat{y}^{\text{MLL}}_{(T+m)j} = \sum_{k=1}^{K} r_{Tjk} \left(\hat{\beta}_{0k} + \hat{\beta}_{1k}(t_{T+m} - t_T) \right) \tag{4.5.2}$$

4.5.2 Forecasting using double localised mixture models

In this section, different approaches towards prediction based on pairs of bandwidths are suggested for the double-localised mixture models MLCV and MLLV. This is done in order to predict at a given time point t_T and for the j-th time series using other time series.

In the MLCV model, the forecast was calculated directly from historical data as a local average of observed past values, with the sizes of the local neighbourhoods and the specific weights of the values defined by exponential and Gaussian kernels. Two approaches can be proposed to forecast future observations of the j-th time series at a time point t_{T+m} using the MLCV method. In the first approach, the m-step-ahead forecast equation is obtained by solving the minimisation problem outlined as follows:

$$\hat{y}^{\text{MLCV}^{(1)}}_{(T+m)j} = \min_{a} \sum_{i=1}^{T} \sum_{k=1}^{K} \sum_{\ell=1}^{J} r_{ijk} W_k(t_i, t_{T+m}) V_k(y_{il}, y_{ij})(y_{ij} - a)^2$$

From this we get, the following m-step-ahead forecast equation:

$$\hat{y}^{\text{MLCV}^{(1)}}_{(T+m)j} = \frac{\sum_{i=1}^{T} \sum_{l=1}^{J} \sum_{k=1}^{K} r_{ijk} W_k(t_i, t_{T+m}) V_k(y_{il}, y_{ij}) y_{ij}}{\sum_{i=1}^{T} \sum_{l=1}^{J} \sum_{k=1}^{K} r_{ijk} W_k(t_i, t_{T+m}) V_k(y_{il}, y_{ij})} \tag{4.5.3}$$

In the second approach, the fitted MLCV was used for prediction, and this gave the following forecast equation

$$\hat{y}^{\text{MLCV}^{(2)}}_{(T+m)j} = \sum_{k=1}^{K} \hat{\pi}_{kj} \hat{\beta}_{kj} \tag{4.5.4}$$

where $\hat{\pi}_{kj}$ and $\hat{\beta}_{kj}$ are the fitted parameters of MLCV.

For the MLLV model, a new approach was presented for prediction based on the fitted MLLV model. The m-step-ahead forecast equation was applied as follows:

$$\hat{y}^{\text{MLLV}}_{(T+m)j} = \sum_{k=1}^{K} \hat{\pi}_{kj} \left[\hat{\beta}_{0kj} + \hat{\beta}_{1kj}(t_{T+m} - t_T) \right] \tag{4.5.5}$$

where $\hat{\pi}_{kj}$, $\hat{\beta}_{0kj}$ and $\hat{\beta}_{0kj}$ are the fitted parameters at a time point t_T and for the j-th time series. For the rest of this chapter, notations are used for ease of explanation. For example, MLCV$^{(i)}(h,\, v)$, $i = 1, 2$ (MLLV$(h,\, v)$) refers to forecasting based on the horizontal and vertical vectors $h = (h_1, \ldots, h_K)$, $v = (v_1, \ldots, v_K)$, where $i = 1$ and $i = 2$ indicate forecast approaches for the MLCV, which are presented by Equations (4.5.3) and (4.5.4), respectively.

4.6 Applications

In this section, real data examples are presented to show the performance of the MLCV$^{(i)}, i = 1, 2$ and MLLV models for forecasting, compared with other time series models, such as the ARIMA and Holt models. In addition, a comparison is undertaken between the MLCV$^{(i)}, i = 1, 2$ (the MLLV model) models and the MLC (the MLL) model in section 4.1. The data used in these examples is the same as the real data used in Section 2.3.1 of Chapter 2, which represents the annual energy use of 134 countries between 1995 and 2011. We choose three countries with representative patterns for this presentation as follows: Figure 4.4 displays the time series of log of energy use for the Ivory Coast, Albania and Lithuania. It can be seen from Figure 4.4 (left) that the time series for the Ivory Coast has two features: it shows an overall increasing linear trend, but there is still considerable variability, especially between 2002 and 2004. In this period, there is a sensible sharp increase in energy use consumption. The time series for Albania has a quite consistent linear trend with little variability, see Figure 4.4 (middle). However, in the case of Lithuania, we have a volatile time series without any linear trend, as shown in Figure 4.4 (right). Further insight is provided by Figure 4.5, which shows a sequence of 2–boxplots of log of energy use data with a time series for each country from 1995 to 2011. We can see from Figures 4.5 (left and right), that it is clear, that the Ivory Coast and Albania belong to the low energy use group of countries. However, Lithuania appears as a developed country with a high energy use consumption, as shown in Figure 4.5 (bottom).

The log of energy use data for these countries is fitted at the target points of $t_T = 2000, \ldots, 2008$, in order to obtain the m-step-ahead forecasts ($m = 1, \ldots, 3$) for each

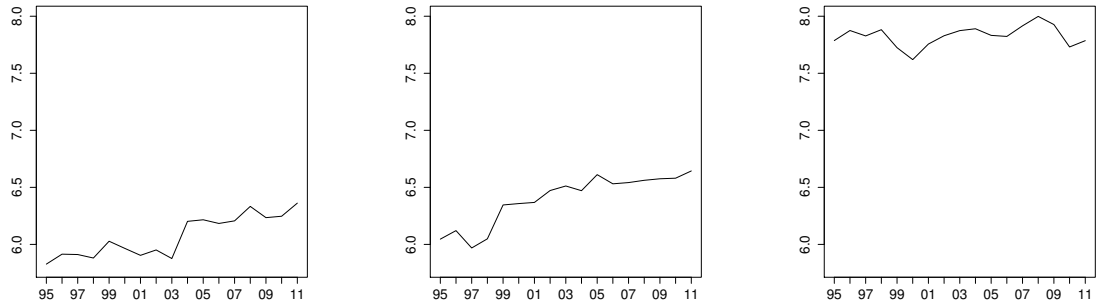

Figure 4.4: Time series of log of energy use data for the Ivory Coast, Albania and Lithuania (from left to right). The horizontal axis denotes the calendar year (from 1995 to 2011), and the vertical axis gives the annual energy use (natural log of kg oil equivalent per capita).

time point t_T using the different models. Hence, 9 forecasts were obtained for each model and a forward lag.

For the MLCV and MLLV models, $K = 2$ components were used to fit the data. To assess the performance of the forecasts using these models, the SSRE and SARE were considered for the m-step-ahead forecasts as defined in Equations (3.7.2) and (3.9.1), as shown in Chapter 3. In our analysis, a and b in Equations (3.7.2) and (3.9.1) take the values $a = 2000$ and $b = 2008$, respectively. Tables 4.1–4.6 summarise the results[1] of the m-step-ahead forecasts based on selected bandwidths, according to the SSRE criterion for MLCV$^{(i)}, i = 1, 2$ and the MLLV models for energy use in the Ivory Coast, Albania and Lithuania, respectively. Different settings for horizontal bandwidths (h_1, h_2) and vertical bandwidths (v_1, v_2) are considered, in order to capture different short and long-term trends, with small and large vertical bandwidths prevailing in these data sets as shown in Tables 4.1–4.6. In Tables 4.1, 4.3, and 4.5, the same horizontal bandwidths were used: $(h_1 = 1, h_2 = 3)$. However, the horizontal bandwidths used in Tables 4.2, 4.4 and 4.6 are $(h_1 = 1, h_2 = 5)$. The reason for using the different second horizontal bandwidths was to investigate the influence of the size of the local neighbourhood horizontally on the second components in the prediction performance for the MLCV$^{(i)}, i = 1, 2$ and MLLV models.

For all Tables 4.1–4.6, five different cases of vertical bandwidth selection are shown.

[1] All values of SSRE snd SARE in tables are multiplied by 1000.

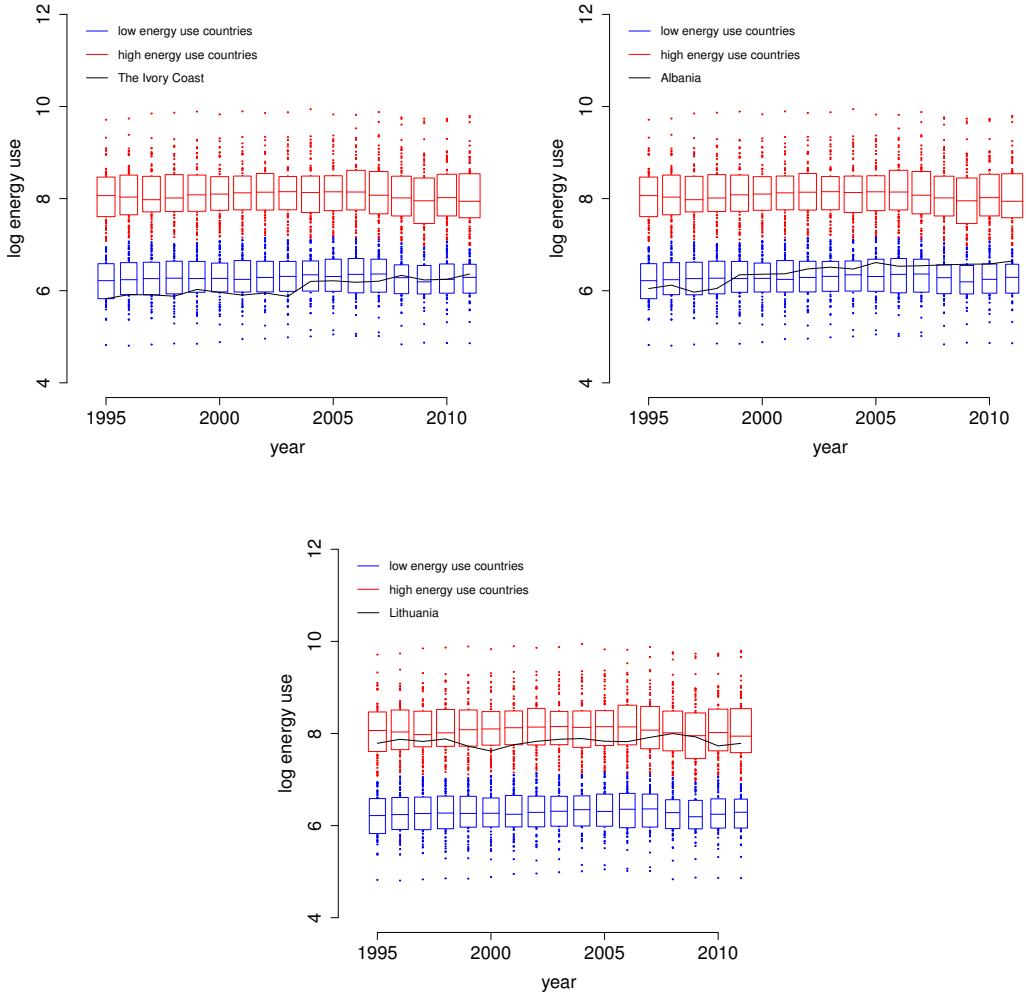

Figure 4.5: 2–boxplots of log of energy use data between 1995 to 2011, with time series for the Ivory Coast (left), Albania (right) and Lithuania (bottom)

The first and third rows of these tables show the case where the first component fits short-term horizontal and vertical bandwidths, and the second component fits long-term horizontal and vertical bandwidths together. This choice allows the first component to use more recent information for all countries, with information from close countries shown in the data patterns for target country. The difference between the two cases in the first and third rows of these tables is the size of the vertical bandwidths, which become larger for both vertical bandwidths components in the second case. In the second and fourth rows of the tables, for the first component, a short-term trend bandwidth with a large-term vertical bandwidth is used to fit data affecting all countries. In addition, the MLCV$^{(i)}, i = 1, 2$ and the MLLV models have

(v_1, v_2)	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(1)	SSRE(2)	SSRE(3)
				$(h_1, h_2) = (1, 3)$					
	MLCV$^{(1)}$			MLCV$^{(2)}$			MLLV		
$(0.3, 3)$	**0.044**	0.042	0.046	**0.394**	0.588	0.814	0.609	1.131	1.833
$(3, 0.3)$	0.257	0.254	0.280	0.549	0.776	1.089	**0.399**	0.589	0.743
$(0.5, 1.5)$	0.045	0.043	0.047	0.395	0.590	0.816	0.586	1.075	1.665
$(1.5, 0.5)$	0.143	0.092	0.090	0.455	0.664	0.943	0.427	0.692	0.982
$(0.3, 0.3)$	**0.044**	0.042	0.046	**0.394**	0.589	0.816	0.554	1.013	1.459

Table 4.1: The SSRE of forecasting for the Ivory Coast from 1995 to 2008.

(v_1, v_2)	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(1)	SSRE(2)	SSRE(3)
				$(h_1, h_2) = (1, 5)$					
	MLCV$^{(1)}$			MLCV$^{(2)}$			MLLV		
$(0.3, 3)$	**0.044**	0.042	0.046	**0.394**	0.588	0.814	0.586	1.080	1.770
$(3, 0.3)$	0.424	0.431	0.484	0.700	0.943	1.304	**0.381**	0.533	0.651
$(0.5, 1.5)$	0.045	0.043	0.047	0.395	0.590	0.816	0.560	1.003	1.579
$(1.5, 0.5)$	0.103	0.081	0.088	0.462	0.670	0.947	0.451	0.766	1.064
$(0.3, 0.3)$	0.045	0.041	0.046	0.395	0.589	0.815	0.518	0.918	1.39

Table 4.2: The SSRE of forecasting for the Ivory Coast from 1995 to 2008.

fitted the energy use for all countries to a large amount of historical information about energy consumption, with information ranging from close countries to a target country, using a large-term trend bandwidth h_1 with small-term bandwidth v_1 for the second component. This component of MLCV$^{(i)}, i = 1, 2$ and MLLV models affects the closest countries to a target country. The last row in all the tables shows results when using equal vertical bandwidths for both components. In this case, the effect of short and large-term trends for similar countries in the data pattern to a target country is shown.

Tables B.1, B.2 and B.3 in Appendix B show the percentages of countries, that influence the prediction for the target countries, the Ivory Coast, Albania and Lithuania, using different sizes of vertical local neighbourhoods. For example, it is clear from Table B.1 that $v = 0.3$ in all countries, that the percentage range is between 16% and 21%. This includes, which information for 22 to 29 countries around the Ivory Coast. However, $v = 3$ takes into account information from 93% to 98% countries, in order to predict energy use of the Ivory Coast at a given time points $t_i, i = 2000$ to 2008. This means, that 125 to 132 countries are classed as being the local vertical neighbourhood.

Tables 4.1 and 4.2 show the results of the SSRE for m-step-ahead forecasts for models MLCV$^{(i)}, i = 1, 2$ and MLLV using energy use data for the Ivory Coast. The errors in these tables are, overall, of larger magnitude than for Albania and Lithuania, due to the larger variation in the data for the Ivory Coast. However, it can be seen that

the MLCV$^{(1)}$ model performed well for all forward lags, and produced a smaller errors margin than the MLCV$^{(2)}$ and MLLV models. It had the best performance when $v = (v_1 = 0.3, v_2 = 3)$ and $v = (v_1 = 0.3, v_2 = 0.3)$ for both cases $h = (h_1 = 1, h_2 = 3)$ and $h = (h_1 = 1, h_2 = 5)$. This means that the MLCV$^{(1)}$ model tends to choose the smallest horizontal and vertical bandwidths together, and this picks very recent information from similar countries to the Ivory Coast in the data pattern.

Further insight is provided in Figure 4.6 which shows the fitted mixture probabilities for $t_i, i = 2000, \ldots, 2008$ for one-step-ahead forecasts for the MLCV$^{(i)}, i = 1, 2$ (left) and MLLV (right) models. One can observe that in all cases, the proportion of the short-term horizontal and vertical component settles at 100%. As a result, the choice of the paired bandwidth $(h, v) = (1, 0.3)$ is strongly recommended for the MLCV$^{(i)}, i = 1, 2$ models. Moreover, the MLLV model shows a slightly better performance than the MLCV$^{(2)}$ model for $(h_1 = 1, h_2 = 5), (v_1 = 3, v_2 = 0.3)$ only. As can be seen from Figure 4.6 (right), it is clear that the proportion of the long-term trend horizontal bandwidths and the small-term vertical bandwidths settles at 100%, which is plausible since the MLLV model has the ability to model long-term linear trends for a data set. In this case, the MLLV model is beneficial for prediction when using the information from the last 5 years of a given time point for similar countries to the Ivory Coast in the data pattern. Table B.1 shows that MLLV used information from 17.58% of countries on average, which are closest to the Ivory Coast in the data pattern. This supports the conclusion that the MLLV model can pick long-term linear trend components. As a result, the choice $(h, v) = (5, 0.3)$ is favourable for the MLLV model in this example.

For the Albania data, the conclusion obtained previously is confirmed, in that the MLCV$^{(1)}$ model leads generally to favourable results in comparison with the MLCV$^{(2)}$ and MLLV models. It produced small errors when using $(v_1 = 0.3, v_2 = 3)$, $(v_1 = 0.5, v_2 = 1.5)$ and $(v_1 = 0.3, v_2 = 0.3)$ for both $(h_1 = 1, h_2 = 3)$ and $(h_1 = 1, h_2 = 5)$ as shown in Tables 4.3 and 4.4. In this example, the MLLV$((h_1 = 1, h_2 = 3), (v_1 = 3, v_2 = 0.3))$ model performs worse than the MLCV$^{(2)}$ model for prediction when SSRE=0.128×10^{-3}, as shown in Table 4.3. This result suggests that the MLLV model is not useful for prediction when the data has only a small amount of variability.

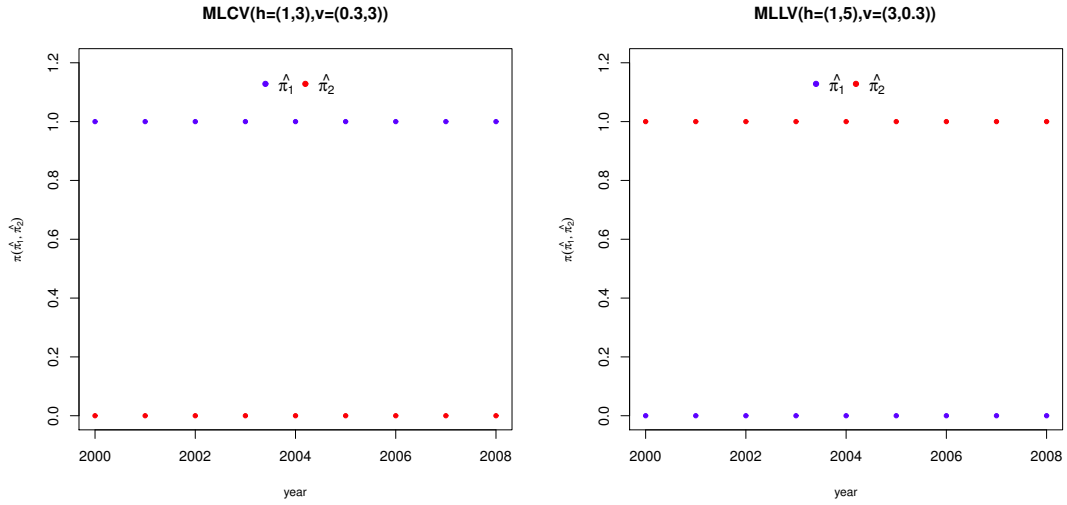

Figure 4.6: For data from the Ivory Coast, fitted parameters $\hat{\pi}_k(t_T)$ (left) using MLCV model and fitted parameters $\hat{\pi}_{0k}(t_T)$ (right) using MLLV model .

(v_1, v_2)	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(1)	SSRE(2)	SSRE(3)
				$(h_1, h_2) = (1, 3)$					
	MLCV$^{(1)}$			MLCV$^{(2)}$			MLLV		
$(0.3, 3)$	**0.016**	0.015	0.011	**0.119**	0.195	0.269	0.147	0.177	0.266
$(3, 0.3)$	0.191	0.168	0.130	0.394	0.570	0.702	**0.128**	0.220	0.418
$(0.5, 1.5)$	**0.016**	0.011	0.016	0.120	0.197	0.272	0.157	0.209	0.321
$(1.5, 0.5)$	0.052	0.046	0.031	0.197	0.316	0.410	0.142	0.221	0.405
$(0.3, 0.3)$	**0.016**	0.015	0.012	0.120	0.197	0.272	0.153	0.208	0.334

Table 4.3: The SSRE of forecasting for Albania from 1995 to 2008.

Looking at the data from Lithuania, it can be seen that it is not suitable for the MLLV model for prediction due to the nature of the time series, which shows a non-linear data structure. Here, the ability to model local constant trends with small-term vertical bandwidths plays a strong role in enhancing prediction. This continues to hold for forecast using $(h, v) = (1, 0.3)$ for the MLCV$^{(i)}, i = 1, 2$ model and using $(h, v) = (5, 0.3)$ for the MLLV model as shown in Tables 4.5 and 4.6. It is clear that the MLLV model produces a poor performance compared to the other models, especially for a short-term horizontal and vertical components and large-term horizontal and vertical components.

In summary, the results for Tables 4.1–4.6 provides insight into the choice of paired bandwidths for double-localised mixture models. From the examples above, it is clear that the MLCV$^{(i)}, i = 1, 2$ models produce good predictions based on SSRE of m-step-ahead forecasts for short-term horizontal bandwidths and short-term vertical bandwidths. This results are reasonable since the MLC$^{(i)}, i = 1, 2$ models as discussed in

(v_1, v_2)	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(1)	SSRE(2)	SSRE(3)
				$(h_1, h_2) = (1, 5)$					
		MLCV$^{(1)}$			MLCV$^{(2)}$			MLLV	
$(0.3, 3)$	**0.016**	0.015	0.011	**0.119**	0.195	0.269	0.154	0.180	0.268
$(3, 0.3)$	0.385	0.346	0.292	0.631	0.841	1.003	0.156	0.283	0.500
$(0.5, 1.5)$	**0.016**	0.015	0.011	**0.119**	0.195	0.269	0.155	0.191	0.304
$(1.5, 0.5)$	0.039	0.034	0.024	0.165	0.262	0.348	**0.147**	0.214	0.357
$(0.3, 0.3)$	**0.016**	0.015	0.011	**0.119**	0.195	0.269	0.154	0.211	0.330

Table 4.4: The SSRE of forecasting for Albania from 1995 to 2008.

(v_1, v_2)	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(1)	SSRE(2)	SSRE(3)
				$(h_1, h_2) = (1, 3)$					
		MLCV$^{(1)}$			MLCV$^{(2)}$			MLLV	
$(0.3, 3)$	**0.011**	0.021	0.019	**0.085**	0.253	0.274	0.186	0.771	1.251
$(3, 0.3)$	0.045	0.074	0.074	0.092	0.185	0.197	**0.142**	0.429	0.578
$(0.5, 1.5)$	0.012	0.021	0.018	0.087	0.256	0.277	0.166	0.703	1.133
$(1.5, 0.5)$	0.017	0.031	0.028	0.088	0.221	0.240	0.153	0.531	0.788
$(0.3, 0.3)$	**0.011**	0.021	0.019	**0.085**	0.253	0.274	0.161	0.663	1.045

Table 4.5: The SSRE of forecasting for Lithuania from 1995 to 2008.

(v_1, v_2)	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(1)	SSRE(2)	SSRE(3)	SSRE(1)	SSRE(2)	SSRE(3)
				$(h_1, h_2) = (1, 5)$					
		MLCV$^{(1)}$			MLCV$^{(2)}$			MLLV	
$(0.3, 3)$	**0.011**	0.021	0.019	**0.085**	0.253	0.274	0.180	0.737	1.190
$(3, 0.3)$	0.064	0.091	0.090	0.103	0.180	0.194	**0.141**	0.385	0.512
$(0.5, 1.5)$	0.012	0.021	0.018	0.087	0.256	0.277	0.162	0.675	1.080
$(1.5, 0.5)$	0.013	0.029	0.020	0.089	0.232	0.258	0.143	0.543	0.837
$(0.3, 0.3)$	**0.011**	0.021	0.019	**0.085**	0.253	0.274	0.154	0.639	1.008

Table 4.6: The SSRE of forecasting for Lithuania from 1995 to 2008.

Chapter 3 are superior for higher lags and smaller historical bandwidths. In all examples, the MLCV$^{(i)}(1, 0.3), i = 1, 2$ models are recommended, in order to produce accurate m-step-ahead forecasts. In addition, small vertical bandwidths provide the MLCV$^{(i)}, i = 1, 2$ models with a lot of information from related countries to the target country, that contributes to improving the prediction for the target country. However, it appears that the MLLV model can only be recommended for large-term horizontal bandwidths and small-term vertical bandwidths thanks to its ability to fit linear trends well, and this is supported the results obtained for the MLL model as shown in Chapter 3. From these examples, MLLV(5,0.3) is suggested for prediction for the energy use of the Ivory Coast and Lithuania, because it is best suited to picking long-term trends for this data. However, for the Albania data, the MLLV(3,0.3) is adequate, since the nature of data has on a small amount of variability. The above examples show that there is no need to use more than a pair of bandwidths (h, v), in order to obtain good predictions based on the SSRE of m-step ahead forecasts. However, this does not necessarily mean that a mixture of double-localised regression over-fits model, because there is a

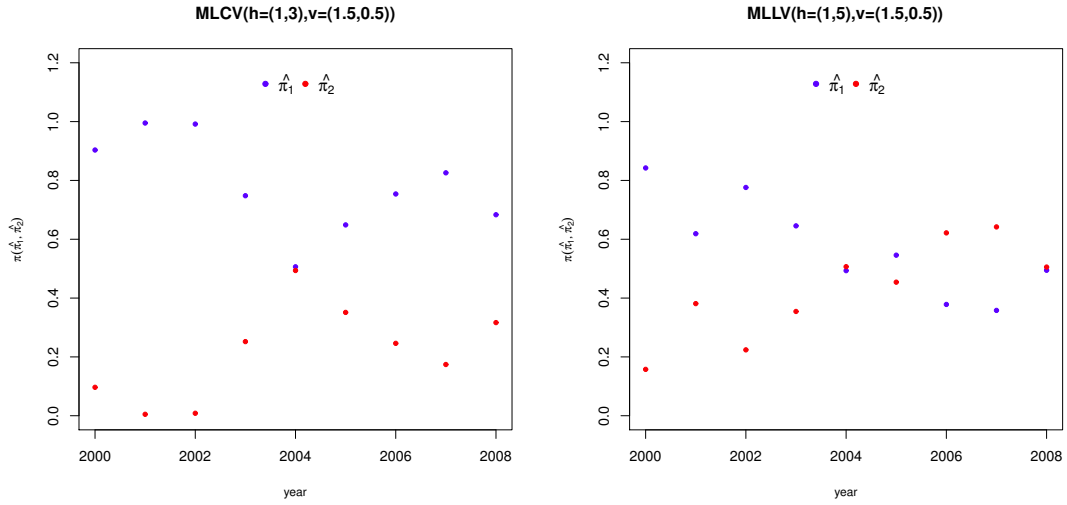

Figure 4.7: Data from the Ivory Coast, fitted parameters $\hat{\pi}_k(t_T)$ (left) using MLCV model and fitted parameters $\hat{\pi}_{0k}(t_T)$ (right) using MLLV model .

need for more than one component at some single time point. For example, from Figure 4.7 (left), it is clear that the fitted proportions $\hat{\pi}_1$ and $\hat{\pi}_2$ for both components at 2004 equal 0.5. When fitting the data for the MLCV$((h_1 = 1, h_2 = 3), (v_1 = 1.5, v_2 = 0.5))$ model two components have the same importance for prediction. In addition, Figure 4.7 (right) shows, that the MLLV$((h_1 = 1, h_2 = 5), (v_1 = 1.5, v_2 = 0.5))$ takes into account two components with the same fitted proportions $\hat{\pi}_1 = \hat{\pi}_2 = 0.5$. MLCV$((h_1 = 1, h_2 = 3), (v_1 = 1.5, v_2 = 0.5))$ and MLLV$((h_1 = 1, h_2 = 5), (v_1 = 1.5, v_2 = 0.5))$ produce a poor performance in predictions based on the SSRE of m-step-ahead forecasts when using these pairs of bandwidths for all time points from 2000 to 2008, as shown in Tables 4.1 and 4.2.

Tables 4.7– 4.9 show the results based on the SSRE and SARE for the three countries under study. In these tables, the results of MLCV$^{(i)}, i = 1, 2$ and MLLV models are produced from Tables 4.1–4.6 based on the best choices of pairs of bandwidths (h_1, h_2) and (v_1, v_2), which give the smallest SSRE. In addition, from these tables, it was possible to evaluate the performance of MLCV$^{(i)}, i = 1, 2$, MLLV, MLC and MLL models, for multiple time series, as discussed in Section 4.1, used in the forecasting compared to other time series models, such as the ARIMA and Holt models for all countries under study.

For the Ivory Coast data, we can see that the MLCV$^{(1)}(h = 1, v = 0.3)$ produced

Model	SSRE(1)	SSRE(2)	SSRE(3)	SARE(1)	SARE(2)	SARE(3)
$MLCV^{(1)}$	0.044	0.042	0.046	4.957	4.512	5.064
$MLCV^{(2)}$	0.394	0.588	0.814	14.219	17.658	24.310
MLC	3.196	2.475	1.789	19.580	15.257	11.110
MLLV	0.381	0.533	0.651	14.840	18.798	21.714
MLL	3.469	2.88	2.307	21.256	17.784	14.329
ARIMA	0.546	0.751	0.938	16.884	20.941	24.797
Holt	0.475	0.912	1.338	2.528	3.674	4.967

Table 4.7: The SSRE and SARE of forecasting for the Ivory Coast from 1995 to 2008.

Model	SSRE(1)	SSRE(2)	SSRE(3)	SARE(1)	SARE(2)	SARE(3)
$MLCV^{(1)}$	0.016	0.015	0.011	2.989	2.741	2.395
$MLCV^{(2)}$	0.119	0.195	0.269	8.120	10.975	13.214
MLC	0.144	0.203	0.282	0.936	1.328	1.856
MLLV	0.128	0.220	0.418	9.737	13.163	18.824
MLL	0.095	0.119	0.159	0.618	0.778	1.042
ARIMA	0.234	0.381	0.466	11.429	13.270	15.427
Holt	0.202	0.333	0.555	1.817	2.489	3.347

Table 4.8: The SSRE and SARE of forecasting for Albania from 1995 to 2008.

Model	SSRE(1)	SSRE(2)	SSRE(3)	SARE(1)	SARE(2)	SARE(3)
$MLCV^{(1)}$	0.011	0.021	0.019	3.039	3.774	3.387
$MLCV^{(2)}$	0.085	0.253	0.274	8.396	13.504	14.564
MLC	0.375	0.517	0.622	2.956	4.066	4.891
MLLV	0.141	0.385	0.512	10.463	16.041	18.402
MLL	0.607	0.767	0.923	4.777	6.040	7.256
ARIMA	0.114	0.190	0.176	8.237	12.006	12.013
Holt	0.184	0.730	1.201	1.274	2.521	3.335

Table 4.9: The SSRE and SARE of forecasting for Lithuania from 1995 to 2008.

generally favourable results, with MLLV($h = 5, v = 0.3$) showing itself to be superior to all models except $MLCV^{(1)}$. In addition, forecasting using MLC and MLL models produced poor performance in comparison to using other models, as shown in Table 4.7. From Table 4.8, the picture obtained previously was confirmed: the $MLCV^{(1)}(h = 1, v = 0.3)$ model performed well for all forward lags, and produced a smaller error margin than all the other models. However, MLLV($h = 3, v = 0.3$) model showed better performance than the ARIMA and Holt models only. In addition, using the MLL model was better than the MLLV model, which means that the vertical kernels, in this case was not useful. For the data from Lithuania, the situation was similar to that of the Ivory Coast. However, as seen in Table 4.9, the ARIMA model performed

better than the MLLV$^{(1)}(h = 5, v = 0.3)$ model.

In conclusion, the examples provided in this chapter have given evidence for the superiority of the MLCV$^{(1)}$ model, with a pair of small horizontal and vertical bandwidths, especially for higher lags. In respect of the MLCV$^{(i)}, i = 1, 2$ method in general, the pair of bandwidths $(h = 1, v = 0.3)$ produced better results generally based on the SSRE of the m-step-ahead forecasts. Using the MLLV model with the pair of bandwidths $(h = 5, v = 0.3)$ for variable data, as in the Ivory Coast and Lithuania, were produced better results for prediction than the MLL, ARIMA and Holt models, except in the case of Lithuania, where the ARIMA model showed the best performance for prediction among these models.

It appears that the MLLV method can only be recommended for data of high variability with linear trends thanks to its ability to fit the linear trend for a long-term horizontal bandwidth as shown in the MLL method outlined in Chapter 3. The MLCV$^{(2)}$ model showed worse performance for prediction in comparison with the MLCV$^{(1)}$ and MLLV models for very strong variability, with overall increasing linear trends only. However, the MLCV$^{(2)}$ model was superior to the ARIMA and Holt models in performance for prediction for all three countries.

4.7 Conclusions

In conclusion, this chapter has presented a novel approach towards forecasting based on double-localised mixtures of non-parametric regressions. In this chapter, non-parametric regression allows forecasts to be calculated horizontally from historical data, as a local average of observed past values over time. In addition Gaussian kernels provide weights to all multiple data sets vertically, at a given time point and around a given data point from a given time series simultaneously over time. In the first model, which is named the MLCV model, local constant estimators were used to carry out the localised estimation step using a pair of bandwidths: both horizontal and vertical bandwidths were anchored at a given time point and for a target time series. In the second model, which is referred to as the MLLV model, the MLCV model was generalized using local linear estimators.

Estimation for these models was achieved using a double-kernel-weighted version of the EM-algorithm, using exponential kernels with different horizontal bandwidths as weight functions of the historical data and Gaussian kernels, with different bandwidths for the vertical data at a target time point. In addition, double-localised mixture models could be considered as a bandwidth selection tool for horizontal and vertical bandwidths. For selected pairs of bandwidths, the double-localised mixture technique determined the proportions used for each mixture's components, as related to a pair bandwidths. The high proportion of a component informs the high effect of the pair of bandwidths used. As a result, double-localised mixture model could help the data analyst support the decision on the selection of the pair of bandwidths. In order to undertake forecasting, several approaches for prediction at the time t_{T+m}, and for a given data set, using these models were investigated, as shown above in three representative patterns of data. It is clear that for the selected pairs of bandwidths, the three examples provide insight into, which MLCV and MLLV models gave the best performance for prediction, compared to using traditional models, such as the ARIMA and Holt models. In addition, the performance of the MLCV and MLLV models for prediction was compared with the MLC and MLL models for multi-valued regression data.

In the first example, the data showed very strong variability in relation to linear trends. The data in the second example had little variability, but in the final example, the linear trend was eliminated, and the variability of data was observed. The results suggest that only the MLLV model can improve predictions from time series data, in comparison to the ARIMA and Holt models, for pairs of long-term horizontal bandwidths and small-term vertical bandwidth with short forward lag, as in the case of very volatile time series. In addition, the MLVC$^{(1)}$ model showed good performance for prediction with a pair of short-term trend horizontal bandwidths and a small-term vertical bandwidth. However, further forecasting methods should be investigated to enhance this comparison.

In the real data applications, there was no need to fit the data using more than one component, in order to produce good prediction results. In addition, the MLCV$^{(1)}$ approach for prediction using Equation (4.5.3) provided very good prediction in almost all cases. This suggests that prediction is a rather simple problem, where the latest

observations provide the most useful information for forecasting and the use of complex models, in order to reduce bias or variance will generally struggle to compete with this. Indeed, double-localised mixture models are very powerful methodologies, but the best application for double-localised mixture models when $K > 1$ has perhaps, yet to be found.

Although, this study was restricted to pairs of bandwidths, that was not optimised, the results appeared competitive and challenging arguments, and in favour of the MLCV$^{(1)}$ and MLLV models, in comparison to the other models used for prediction. This suggests that further study is needed on the performance of the MLCV$^{(i)}, i = 1, 2$ and MLLV models for prediction when all bandwidths are optimised. A measure of degree of localisation relating to a combination of a pair of bandwidths should be developed. In addition, it is strongly recommended, that researchers test the number of components of double-localised regression models. These features could make additional advantage of the MLLV and MLCV models for prediction.

Chapter 5

Conclusions and future research

5.1 Conclusions

The first contribution of this thesis was a new powerful graphical tool, that can be used to visualise and analyse data stemming from a mixture of K distributions. This plot was named the K–boxplot. It is a developed version of the traditional boxplot, and is used especially for finding additional information regarding the location and spread of individual groups in mixture data, which are ignored by a traditional boxplot. It is worth mentioning, that the K–boxplot cannot be used as an inference tool, that can make automated decisions about the distribution or the number of components in mixture data. However, it is a helpful and useful tool to support the data analyst in this respect. K–boxplots are implemented in the function `kboxplot`, which is made available as part of the R package **UEM**. The examples presented in this thesis are listed in the R Documentation files of the R package **UEM**. The methodology of the K–boxplot can be implemented in codes, as part of any statistical software package.

The thesis proceded with the development of prediction techniques from time series data using a mixture of local regression model, named MLC and MLL models. A novel approach to forecasting based on localised mixtures of non-parametric regressions is proposed. A new approach of bandwidth selection for prediction has been developed. This new methodology contributes towards improving prediction for MLL and NLL models, especially compared to other models under study. The results suggest that only the MLL model can improve predictions from linear and variable time series data,

in comparison with the Holt and ARIMA models for long-term component and short forward lag. In addition, the MLC$^{(2)}$ model showed good performance in the prediction of short-term components and non-linear data trends. The results provide competitive and challenging arguments in favour of MLC$^{(2)}$ and MLL models, in comparison to other models used for prediction, even although one of the bandwidths is not optimised. Further study on the performance of MLC$^{(2)}$ and MLL models for prediction when all bandwidths are optimised is recommended. Bandwidth selection is implicit to the MLC and MLL approaches for prediction. This feature makes additional advantage of the MLC and MLL methodologies for prediction. Although, the bandwidths of MLC$^{(1)}$ model was not optimized, the MLC$^{(1)}$ model became superior in comparison with all other models especially for small value of bandwidths.

This thesis has also presented a novel approach to forecasting based on double-localised mixtures of non-parametric regressions. It is clear, for selected pairs of bandwidths, the performance of MLCV and MLLV models for prediction can be compared with MLC and MLL models used for multiple time series. The MLLV model can only improve predictions from time series data, in comparison with the ARIMA and Holt models, for a pair of long-term horizontal bandwidths and small-term vertical bandwidths with a short forward lag. In addition, the MLVC$^{(1)}$ model showed good performance for prediction, using a pair of short-term horizontal bandwidth and small-term vertical bandwidths. In the real data applications, there was no need to fit the data using more than one component to achieve good predictions. In addition, using the MLCV$^{(1)}$ approach in combination with Equation (4.5.3) provided very good prediction in almost all cases. This suggests that prediction is a rather simple problem, where the latest observations provide the most useful information for forecasting and the use of complex models, in order to reduce bias or variance will generally struggle to compete with this. It can still be concluded that double-localised mixture models are very powerful methodologies, but that the best application to use for double-localised mixture models when $K > 1$ has perhaps yet to be found. The results appear competitive and challenging arguments in favour of MLCV$^{(1)}$ and MLLV models compared to other models used for prediction, although this study was restricted to examining pairs of bandwidths, that were not optimised.

5.2 Future research

In Section 1.4.1 of Chapter 1, different Bayesian approaches for estimation mixture models are discussed. Hence, it is a good idea to use one of these approaches to estimate the parameters of localised mixture models in Chapter 3 and double-localised mixture models in Chapter 4, in order to investigate the effects of this approach on the performance of prediction. In Section 1.4.2 of Chapter 1, popular methodologies of testing the number of components for mixture models are viewed, which can be useful, in further study, to investigate the number of components in the proposed models for prediction in Chapters 3 and 4. In addition, there is a need to develop a new methodology, in order to find optimal pairs of bandwidths for the proposed models for prediction in Chapter 4.

A new approach to estimate the proposed models for prediction in Chapters 3 and 4 with direct consideration of prediction could be taken into account for further research. The MLC and MLL models in Chapter 3 and the MLCV and MLLV models in Chapter 4 can be developed in future research by providing them with additional information. For example, the seasonality of data, which make these models adequate for seasonal time series. In addition, an interesting topic related to the robustness of the proposed models for prediction in Chapters 3 and 4 can be considered in further research.

Bibliography

[1] A. H. Abuzaid, I. B. Mohamed, and A. G. Hussin. Boxplot for circular variables. *Computational Statistics*, 27(3):381–392, 2012.

[2] M. Aitkin, B. Francis, J. Hinde, and R. Darnell. *Statistical Modelling in R*. Oxford University Press Oxford, 2009.

[3] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[4] G. Aneiros-Pérez and P. Vieu. Nonparametric time series prediction: a semi-functional partial linear modeling. *Journal of Multivariate Analysis*, 99(5):834–857, 2008.

[5] J.-P. Baudry and G. Celeux. EM for mixtures. *Statistics and Computing*, 25(4):713–726, 2015.

[6] Y. Benjamini. Opening the box of a boxplot. *The American Statistician*, 42(4):257–262, 1988.

[7] C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3):561–575, 2003.

[8] D. Böhning, C. Hennig, G. J. McLachlan, and P. D. McNicholas. The 2nd special issue on advances in mixture models. *Computational Statistics & Data Analysis*, 71:1–2, 2014.

[9] C. Bruffaerts, V. Verardi, and C. Vermandele. A generalized boxplot for skewed and heavy-tailed distributions. *Statistics & Probability Letters*, 95:110–117, 2014.

[10] R. J. Carroll, D. Ruppert, and A. H. Welsh. Local estimating equations. *Journal of the American Statistical Association*, 93(441):214–227, 1998.

[11] R. M. Cassie. Some uses of probability paper in the analysis of size frequency distributions. *Marine and Freshwater Research*, 5(3):513–522, 1954.

[12] J. Chen and J. D. Kalbfleisch. Penalized minimum-distance estimates in finite mixture models. *Canadian Journal of Statistics*, 24(2):167–175, 1996.

[13] M.-Y. Cheng, J. Fan, and J. S. Marron. On automatic boundary corrections. *The Annals of Statistics*, 25(4):1691–1708, 1997.

[14] D. B. H. Cline. Admissibile kernel estimators of a multivariate density. *The Annals of Statistics*, 16(4):1421–1427, 1988.

[15] J. Ćwik and J. Koronacki. A combined adaptive-mixtures/plug-in estimator of multivariate probability densities. *Computational Statistics & Data Analysis*, 26(2):199–218, 1997.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

[17] J. Diebolt and C. P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, 56(2):363–375, 1994.

[18] J. Einbeck, D. Bonetti, and N. M. Qarmalah. *UEM: Estimating and Updating Mixtures via EM*, 2017. R package version 0.1-7.

[19] J. Einbeck and J. Taylor. A number-of-modes reference rule for density estimation under multimodality. *Statistica Neerlandica*, 67(1):54–66, 2013.

[20] J. Fan. Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21(1):196–216, 1993.

[21] J. Fan, M. Farmen, and I. Gijbels. Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society. Series B*, 60(3):591–608, 1998.

[22] J. Fan and I. Gijbels. Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, 20(4):2008–2036, 1992.

[23] J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. New York, Chapman & Hall, 1996.

[24] J. Fan, I. Gijbels, and M. King. Local likelihood and local partial likelihood in hazard regression. *The Annals of Statistics*, 25(4):1661–1690, 1997.

[25] J. Fan, N. E. Heckman, and M. P. Wand. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, 90(429):141–150, 1995.

[26] J. Fox. *Nonparametric Simple Regression: Smoothing Scatterplots*. Sage, 2000.

[27] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

[28] R. Fried, J. Einbeck, and U. Gather. Weighted repeated median smoothing and filtering. *Journal of the American Statistical Association*, 102(480):1300–1308, 2007.

[29] M. Frigge, D. C. Hoaglin, and B. Iglewicz. Some implementations of the boxplot. *The American Statistician*, 43(1):50–54, 1989.

[30] S. Frühwirth-Schnatter. Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96(453):194–209, 2001.

[31] S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer, 2006.

[32] E. S. Gardner. Exponential smoothing: the state of the art. *Journal of Forecasting*, 4(1):1–28, 1985.

[33] S. Gelper, R. Fried, and C. Croux. Robust forecasting with exponential and Holt–Winters smoothing. *Journal of Forecasting*, 29(3):285–300, 2010.

[34] I. Gijbels, A. Pope, and M. P. Wand. Understanding exponential smoothing via kernel regression. *Journal of the Royal Statistical Society: Series B*, 61(1):39–50, 1999.

[35] S. M. Goldfeld and R. E. Quandt. A markov model for switching regressions. *Journal of Econometrics*, 1(1):3–15, 1973.

[36] P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[37] P. J. Green and S. Richardson. Hidden markov models and disease mapping. *Journal of the American Statistical Association*, 97(460):1055–1070, 2002.

[38] J. P. Harding. The use of probability paper for the graphical analysis of polymodal frequency distributions. *Journal of the Marine Biological Association of the United Kingdom*, 28(1):141–153, 1949.

[39] W. K. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer, 2012.

[40] J. Henna. On estimating of the number of constituents of a finite mixture of continuous distributions. *Annals of the Institute of Statistical Mathematics*, 37(1):235–240, 1985.

[41] C. Hennig. Identifiablity of models for clusterwise linear regression. *Journal of Classification*, 17(2):273–296, 2000.

[42] J. Hinde, S. Ingrassia, T. I. Lin, and P. McNicholas. The third special issue on advances in mixture models. *Computational Statistics & Data Analysis*, 93:2–4, 2016.

[43] A. C. A. Hope. A simplified monte carlo significance test procedure. *Journal of the Royal Statistical Society. Series B*, 30(3):582–598, 1968.

[44] M. Huang, R. Li, and S. Wang. Nonparametric mixture of regression models. *Journal of the American Statistical Association*, 108(503):929–941, 2013.

[45] M. Huang and W. Yao. Mixture of regression models with varying mixing proportions: a semiparametric approach. *Journal of the American Statistical Association*, 107(498):711–724, 2012.

[46] M. Hubert and E. Vandervieren. An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, 52(12):5186–5201, 2008.

[47] M. Hurn, A. Justel, and C. P. Robert. Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, 12(1):55–79, 2003.

[48] C. M. Hurvich, J. S. Simonoff, and C.-L. Tsai. Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society: Series B*, 60(2):271–293, 1998.

[49] R. Hyndman and Y. Khandakar. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(1):1–22, 2008.

[50] R. J. Hyndman, M. Akram, and B. C. Archibald. The admissible parameter space for exponential smoothing models. *Annals of the Institute of Statistical Mathematics*, 60(2):407–426, 2008.

[51] R. J. Hyndman, A. B. Koehler, R. D. Snyder, and S. Grose. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, 18(3):439 – 454, 2002.

[52] C. F. Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[53] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

[54] C. T. Kelley. *Iterative Methods for Optimization*. SIAM, 1999.

[55] M. Krzywinski and N. Altman. Points of significance: visualizing samples with box plots. *Nature Methods*, 11(2):119–120, 2014.

[56] B. G. Leroux. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20(3):1350–1360, 1992.

[57] B. G. Lindsay. Mixture models: theory, geometry and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 5:i–163, 1995.

[58] B. G. Lindsay and K. Roeder. Residual diagnostics for mixture models. *Journal of the American Statistical Association*, 87(419):785–794, 1992.

[59] J. S. Marron and D. Nolan. Canonical kernels for density estimation. *Statistics & Probability Letters*, 7(3):195–199, 1988.

[60] R. McGill, J. W. Tukey, and W. A. Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.

[61] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. New York, Wiley, 2007.

[62] G. McLachlan and D. Peel. *Finite Mixture Models*. New York, Wiley, 2004.

[63] M. Miloslavsky and M. J. van der Laan. Fitting of mixtures with unspecified number of components using cross validation distance estimate. *Computational Statistics & Data Analysis*, 41(3):413–428, 2003.

[64] E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

[65] K. Pearson. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 187:253–318, 1896.

[66] A. Polymenis and D. M. Titterington. On the determination of the number of components in a mixture. *Statistics & Probability Letters*, 38(4):295–298, 1998.

[67] N. M. Qarmalah, J. Einbeck, and F. P. A. Coolen. k-boxplots for mixture data. *Statistical Papers*, 2016, doi=10.1007/s00362-016-0774-7.

[68] N. M. Qarmalah, J. Einbeck, and F. P. A. Coolen. Mixture models for prediction from time series, with application to energy use data. *Archives of Data Science Series A*, 2(1):1–15, 2017.

[69] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[70] P. E. Rossi and G. M. Allenby. Bayesian statistics and marketing. *Marketing Science*, 22(3):304–328, 2003.

[71] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[72] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *The R Journal*, 8(1):289, 2016.

[73] J. S. Simonoff. *Smoothing Methods in Statistics*. Springer, 2012.

[74] A. F. M. Smith and G. O. Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B*, 55(1):3–23, 1993.

[75] J. L. Solka, E. J. Wegman, C. E. Priebe, W. L. Poston, and G. W. Rogers. Mixture structure analysis using the akaike information criterion and the bootstrap. *Statistics and Computing*, 8(3):177–188, 1998.

[76] M. Stephens. Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. *Annals of Statistics*, 24(2):40–74, 2000.

[77] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B*, 36(2):111–147, 1974.

[78] J. W. Taylor. Exponential smoothing with a damped multiplicative trend. *International Journal of Forecasting*, 19(4):715–725, 2003.

[79] R. Tibshirani and T. Hastie. Local likelihood estimation. *Journal of the American Statistical Association*, 82(398):559–567, 1987.

[80] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, 1985.

[81] J. W. Tukey. A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, 2:448–485, 1960.

[82] J. W. Tukey. *Exploratory Data Analysis*. Reading, Pearson, 1977.

[83] W. Vandaele. Robust estimation of arima models. *Journal of Econometrics*, 16(1):163, 1981.

[84] G. Wahba and Y. Wang. When is the optimal regularization parameter insensitive to the choice of the loss function?. *Communications in Statistics-Theory and Methods*, 19(5):1685–1700, 1990.

[85] M. P. Wand and M. C. Jones. *Kernel Smoothing*. London, Chapman & Hall, 1994.

[86] P. Wang, M. L. Puterman, I. Cockburn, and N. Le. Mixed Poisson regression models with covariate dependent rates. *Biometrics*, 52(2):381–400, 1996.

[87] Z. Wang and D. Bellhouse. A diagnostic tool for regression analysis of complex survey data. *Statistical Papers*, 56(4):1041–1053, 2015.

[88] G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

[89] M. Wedel and W. S. DeSarbo. A latent class binomial logit methodology for the analysis of paired comparison choice data. *Decision Sciences*, 24(6):1157–1170, 1993.

[90] M. Wedel and W. S. DeSarbo. A mixture likelihood approach for generalized linear models. *Journal of Classification*, 12(1):21–55, 1995.

[91] D. S. Young and D.R. Hunter. Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics & Data Analysis*, 54(10):2253 – 2266, 2010.

Appendix A

Brief guide to notation

MLC	mixture of local constant regression model
MLL	mixture of local linear regression model
NLC	local constant regression model
NLL	local linear regression model
MLCV	mixture of local constant regression model with vertical kernels
MLLV	mixture of local linear regression model with vertical kernels
t_i	predictor or regressor variable
y_i	response variable
K	number of components
J	number of time series
W_k	exponential weight function
V_k	normal vertical kernel
h_k	horizontal bandwidth of component k
v_k	vertical bandwidth of component k
$\mathcal{L}(\cdot)$	likelihood function
$\ell(\cdot)$	log likelihood function

Appendix B

v\year	2000	2001	2002	2003	2004	2005	2006	2007	2008
0.3	15.67	16.42	17.16	15.67	19.40	18.66	17.16	17.16	20.90
0.5	26.12	25.37	27.61	23.13	31.34	32.84	27.61	29.10	29.10
1.5	57.46	54.48	56.72	54.48	61.94	61.19	60.45	60.45	62.69
3	94.78	93.29	93.29	92.54	97.01	96.27	96.27	96.27	97.76

Table B.1: The percentage of countries included different local neighbourhood v around the Ivory Coast at target years from 2000 to 2008.

v\year	2000	2001	2002	2003	2004	2005	2006	2007	2008
0.3	20.90	20.15	20.15	20.15	20.90	20.15	20.15	20.90	20.15
0.5	32.84	32.84	33.58	32.84	34.33	32.90	32.90	32.90	29.85
1.5	68.66	68.66	69.40	69.40	69.40	72.39	71.64	71.64	72.39
3	99.25	99.25	99.25	99.25	99.25	99.25	99.25	98.51	97.76

Table B.2: The percentage of countries included different local neighbourhood v around Albania at target years from 2000 to 2008.

v\year	2000	2001	2002	2003	2004	2005	2006	2007	2008
0.3	17.16	17.16	17.16	20.15	20.15	16.42	17.16	20.90	18.66
0.5	26.87	28.36	29.10	29.10	30.60	30.60	29.10	32.09	32.09
1.5	80.60	75.37	73.13	74.63	76.12	76.12	77.61	75.37	70.90
3	100	100	100	100	100	100	100	100	99.25

Table B.3: The percentage of countries included different local neighbourhood v around Lithuania at target years from 2000 to 2008.

Appendix C

Proof of Equation 3.2.5

Setting $\partial \ell^* / \partial \beta_k^{(l+1)} = 0$, one obtains

$$\sum_{i=1}^{T} \sum_{k=1}^{K} G_{ik} W_k(t_i, t_T) \frac{(y_i - \beta_k^{(l+1)})}{\sigma^{2(l)}} = 0$$

Setting $G_{ik} = r_{ik}^{(l+1)}$, we have the following

$$\beta_k^{(l+1)} = \frac{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T) y_i}{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)}$$

Proof of Equation 3.2.6

Setting $\partial \ell^* / \partial \sigma^{(l+1)} = 0$, one obtains

$$\sum_{i=1}^{T} \sum_{k=1}^{K} G_{ik} W_k(t_i, t_T) \left[-\sigma^{2^{(l+1)}} + (y_i - \beta_k^{(l+1)})^2 \right] = 0$$

Setting $G_{ik} = r_{ik}^{(l+1)}$, we have the following

$$\sigma^{2^{(l+1)}} = \frac{\sum_{i=1}^{T} \sum_{k=1}^{K} r_{ik}^{(l+1)} W_k(t_i, t_T)(y_i - \beta_k^{(l+1)})^2}{\sum_{i=1}^{T} \sum_{k=1}^{K} r_{ik}^{(l+1)} W_k(t_i, t_T)}$$

Proof of Equation 3.3.4

Setting $\partial \ell^* / \partial \beta_{0k}^{(l+1)} = 0$, we obtain

$$-\sum_{i=1}^{T} \sum_{k=1}^{K} G_{ik} W_k(t_i, t_T)(y_i - \beta_{1k}^{(l+1)}(t_i - t_T)) + \sum_{i=1}^{T} \sum_{k=1}^{K} G_{ik} W_k(t_i, t_T) \beta_{0k}^{(l+1)} = 0$$

Setting $G_{ik} = r_{ik}^{(l+1)}$

$$\beta_{0k}^{(l+1)} = \frac{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(y_i - \beta_{1k}^{(l+1)}(t_i - t_T))}{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)} \tag{C.0.1}$$

Setting $\partial \ell^* / \partial \beta_{1k}^{(l+1)} = 0$, we obtain

$$-\sum_{i=1}^{T}\sum_{k=1}^{K} G_{ik} W_k(t_i, t_T)(y_i - \beta_{0k}^{(l+1)})(t_i - t_T) + \sum_{i=1}^{T}\sum_{k=1}^{K} G_{ik} W_k(t_i, t_T)(t_i - t_T)^2 \beta_{1k}^{(l+1)} = 0$$

Setting $G_{ik} = r_{ik}^{(l+1)}$

$$\beta_{1k}^{(l+1)} = \frac{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(y_i - \beta_{0k}^{(l+1)})(t_i - t_T)}{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T)^2} \tag{C.0.2}$$

By solving Equation(C.0.1) and Equation(C.0.2)

$$\beta_{0k}^{(l+1)} = \frac{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T) y_i}{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)} - \beta_{1k}^{(l+1)} \frac{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T)}{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)}$$

From Equation(C.0.2)

$$\begin{aligned} \beta_{0k}^{(l+1)} &= \frac{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T) y_i}{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)} - \frac{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(y_i - \beta_{0k}^{(l+1)})(t_i - t_T)}{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T)^2} \\ &\left[\frac{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T))}{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)} \right] \end{aligned}$$

Let

$$S_{k,T,j} = \sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T)^j$$

and

$$S_{k,T,j}^* = \sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T)^j y_i$$

$$\beta_{0k}^{(l+1)} = \frac{S_{k,T,0}^*}{S_{k,T,0}} - \left[\frac{S_{k,T,1}^* - \beta_{0k}^{(l+1)} S_{k,T,1}}{S_{k,T,2}} \right] \frac{S_{k,T,1}}{S_{k,T,0}}$$

Then

$$\beta_{0k}^{(l+1)} = \frac{S_{k,T,2} S_{k,T,0}^* - S_{k,T,1} S_{k,T,1}^*}{S_{k,T,2} S_{k,T,0} - S_{k,T,1}^2}$$

On the other hand,

$$\beta_{1k}^{(l+1)} = \frac{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T) y_i (t_i - t_T)}{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T)^2} - \beta_{0k}^{(l+1)} \frac{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T)}{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T)^2}$$

From Equation(C.0.1)

$$\beta_{1k}^{(l+1)} = \frac{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T) y_i (t_i - t_T)}{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T)^2} - \frac{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(y_i - \hat{\beta}_{1k}(t_i - t_T))}{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)} \left[\frac{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T)}{\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T)^2} \right]$$

$$\beta_{1k}^{(l+1)} = \frac{S_{k,T,1}^*}{S_{k,T,2}} - \left[\frac{S_{k,T,0}^* - \beta_{1k}^{(l+1)} S_{k,T,1}}{S_{k,T,0}} \right] \frac{S_{k,T,1}}{S_{k,T,2}}$$

Then

$$\beta_{1k}^{(l+1)} = \frac{S_{k,T,0} S_{k,T,1}^* - S_{k,T,1} S_{k,T,0}^*}{S_{k,T,2} S_{k,T,0} - S_{k,T,1}^2}$$

In addition, we can estimate $\beta_{0k}^{(l+1)}$ and $\beta_{1k}^{(l+1)}$ by solving the least square problem

$$\sum_{i=1}^{T} \left[y_i - \beta_{0k}^{(l+1)} - \beta_{1k}^{(l+1)} (t_i - t_T) \right]^2 r_{ik}^{(l+1)} W_k(t_i, t_T)$$

Let $\mathbf{X}$ denote the $n \times 2$ matrix with i-th row $(1, t_i - t_T)$ and $\mathbf{W}$ denote the $n \times n$ diagonal matrix with i-th diagonal element $w_i(t_T) = r_{ik}^{(l+1)} W_k(t_i, t_T)$, then the local linear estimator defined by the coefficients $\beta_{0k}^{(l+1)}$ and $\beta_{1k}^{(l+1)}$ is

$$\beta_{0k}^{(l+1)} = e_1^T \left[\mathbf{X}^T \mathbf{W} \mathbf{X} \right]^{-1} \mathbf{X}^T \mathbf{W} \boldsymbol{y}$$

and

$$\beta_{1k}^{(l+1)} = e_2^T \left[\mathbf{X}^T \mathbf{W} \mathbf{X} \right]^{-1} \mathbf{X}^T \mathbf{W} \boldsymbol{y}$$

where $e_1 = (1, 0)^T$, $e_2 = (0, 1)^T$ and $\mathbf{y} = (y_1, \ldots, y_T)$.

$$\mathbf{X}^T \mathbf{W} \mathbf{X} = \begin{bmatrix} \sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T) & \sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T) \\ \sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T) & \sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T)^2 \end{bmatrix}$$

$$det(\mathbf{X}^T \mathbf{W} \mathbf{X}) = \sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T) \sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T)^2 - \left[\sum_{i=1}^{T} r_{ik}^{(l+1)} W_k(t_i, t_T)(t_i - t_T) \right]^2$$

Then

$$\mathbf{X}^T \mathbf{W} \mathbf{X} = \begin{bmatrix} S_{k,T,0} & S_{k,T,1} \\ S_{k,T,1} & S_{k,T,2} \end{bmatrix}$$

$$det(\mathbf{X}^T \mathbf{W} \mathbf{X}) = S_{k,T,0} S_{k,T,2} - S_{k,T,1}^2$$

$$\left[\mathbf{X}^T\mathbf{W}\mathbf{X}\right]^{-1} = \begin{bmatrix} \frac{S_{k,T,2}}{S_{k,T,0}S_{k,T,2}-S_{k,T,1}^2} & \frac{-S_{k,T,1}}{S_{k,T,0}S_{k,T,2}-S_{k,T,1}^2} \\ \frac{-S_{k,T,1}}{S_{k,T,0}S_{k,T,2}-S_{k,T,1}^2} & \frac{S_{k,T,0}}{S_{k,T,0}S_{k,T,2}-S_{k,T,1}^2} \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{W}\mathbf{y} = \begin{bmatrix} S_{k,T,0}^* \\ S_{k,T,1}^* \end{bmatrix}$$

Then

$$\beta_{0k}^{(l+1)} = \frac{S_{k,T,2}S_{k,T,0}^* - S_{k,T,1}S_{k,T,1}^*}{S_{k,T,2}S_{k,T,0} - S_{k,T,1}^2}$$

and

$$\beta_{1k}^{(l+1)} = \frac{S_{k,T,0}S_{k,T,1}^* - S_{k,T,1}S_{k,T,0}^*}{S_{k,T,2}S_{k,T,0} - S_{k,T,1}^2}$$

Proof of Equation 3.3.5

Setting $\partial\ell^*/\partial\sigma^{(l+1)} = 0$, one obtains

$$\sum_{i=1}^{T}\sum_{k=1}^{K} G_{ik}W_k(t_i,t_T)\left[\sigma^{2(l+1)} + (y_i - \beta_{0k}^{(l+1)} - \beta_{1k}^{(l+1)}(t_i,t_T))^2\right] = 0$$

Setting $G_{ik} = r_{ik}^{(l+1)}$, we have the following

$$\sigma^{2(l+1)} = \frac{\sum_{i=1}^{T}\sum_{k=1}^{K} r_{ik}^{(l+1)}W_k(t_i,t_T)(y_i - \beta_{0k}^{(l+1)} - \beta_{1k}^{(l+1)}(t_i,t_T))^2}{\sum_{i=1}^{T}\sum_{k=1}^{K} r_{ik}^{(l+1)}W_k(t_i,t_T)}$$

Appendix D

Proof of Equation 4.2.6

Setting $\partial \ell^* / \partial \beta_{kj}^{(v+1)} = 0$, one obtains

$$\sum_{i=1}^{T} \sum_{l=1}^{J} \sum_{k=1}^{K} G_{ijk} W_k(t_i, t_T) V_k(y_{il}, y_{ij}) \frac{(y_{ij} - \beta_{kj}^{(v+1)})}{\sigma_j^{2(v)}} = 0$$

Setting $G_{ijk} = r_{ijk}^{(v+1)}$, we have the following

$$\beta_{kj}^{(v+1)} = \frac{\sum_{i=1}^{T} \sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij}) y_{ij}}{\sum_{i=1}^{T} \sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})}$$

Proof of Equation 4.2.7

Setting $\partial \ell^* / \partial \sigma_j^{(v+1)} = 0$, one obtains

$$\sum_{i=1}^{T} \sum_{l=1}^{J} \sum_{k=1}^{K} G_{ijk} W_k(t_i, t_T) V_k(y_{il}, y_{ij}) \left[-\sigma_j^{2(v+1)} + (y_{ij} - \beta_{kj}^{(v+1)})^2 \right] = 0$$

Setting $G_{ijk} = r_{ijk}^{(v+1)}$, we have the following

$$\sigma_j^{2(v+1)} = \frac{\sum_{i=1}^{T} \sum_{l=1}^{J} \sum_{k=1}^{K} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij}) (y_{ij} - \beta_{kj}^{(v+1)})^2}{\sum_{i=1}^{T} \sum_{l=1}^{J} \sum_{k=1}^{K} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})}$$

Proof of Equations 4.3.5 and 4.3.6

Setting $\partial \ell^* / \partial \beta_{0k}^{(v+1)} = 0$ one obtains

$$\begin{aligned} & - \sum_{i=1}^{T}\sum_{l=1}^{J}\sum_{k=1}^{K} G_{ijk} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(y_{ij} - \beta_{1k}^{(v+1)}(t_i - t_T)) \\ & + \sum_{i=1}^{T}\sum_{l=1}^{J}\sum_{k=1}^{K} G_{ijk} W_k(t_i, t_T) V_k(y_{il}, y_{ij}) \beta_{0k}^{(v+1)} = 0 \end{aligned}$$

Setting $G_{ijk} = r_{ijk}^{(v+1)}$

$$\beta_{0k}^{(v+1)} = \frac{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(y_{ij} - \beta_{1k}^{(v+1)}(t_i - t_T))}{\sum_{i=1}^{T} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})} \tag{D.0.1}$$

Setting $\partial \ell^* / \partial \beta_{1k}^{(v+1)} = 0$, we obtain

$$\begin{aligned} & - \sum_{i=1}^{T}\sum_{l=1}^{J}\sum_{k=1}^{K} G_{ijk} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(y_{ij} - \beta_{0k}^{(v+1)})(t_i - t_T) \\ & + \sum_{i=1}^{T}\sum_{l=1}^{J}\sum_{k=1}^{K} G_{ijk} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(t_i - t_T)^2 \beta_{1k}^{(v+1)} = 0 \end{aligned}$$

Setting $G_{ik} = r_{ijk}^{(v+1)}$

$$\beta_{1k} = \frac{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(y_{ij} - \beta_{0k}^{(v+1)})(t_i - t_T)}{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(t_i - t_T)^2} \tag{D.0.2}$$

By solving Equation(D.0.1) and Equation(D.0.2)

$$\begin{aligned} \beta_{0k}^{(v+1)} &= \frac{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij}) y_{ij}}{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})} \\ &- \beta_{1k}^{(v+1)} \frac{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(t_i - t_T)}{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})} \end{aligned}$$

From Equation(D.0.2)

$$\begin{aligned} \beta_{0k}^{(v+1)} &= \frac{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij}) y_{ij}}{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} k(t_i, t_T) V_k(y_{il}, y_{ij})} \\ &- \frac{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(y_{ij} - \beta_{0k}^{(v+1)})(t_i - t_T)}{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(t_i - t_T)^2} \\ & \left[\frac{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(t_i - t_T)}{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})} \right] \end{aligned}$$

Let

$$S_{k,T,j,s} = \sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(t_i - t_T)^s$$

and

$$S^*_{k,T,j,s} = \sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(t_i - t_T)^s y_{ij}$$

$$\beta_{0k}^{(v+1)} = \frac{S^*_{k,T,j,0}}{S_{k,T,j,0}} - \left[\frac{S^*_{k,T,j,1} - \beta_{0k}^{(v+1)} S_{k,T,j,1}}{S_{k,T,j,2}}\right]\frac{S_{k,T,j,1}}{S_{k,T,j,0}}$$

Then

$$\beta_{0k}^{(v+1)} = \frac{S_{k,T,j,2}S^*_{k,T,j,0} - S_{k,T,j,1}S^*_{k,T,j,1}}{S_{k,T,j,2}S_{k,T,j,0} - S^2_{k,T,j,1}}$$

On the other hand,

$$\begin{aligned}\beta_{1k}^{(v+1)} &= \frac{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij}) y_{ij}(t_i - t_T)}{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(t_i - t_T)^2} \\ &- \beta_{0k}^{(v+1)} \frac{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(t_i - t_T)}{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(t_i - t_T)^2}\end{aligned}$$

From Equation(D.0.1)

$$\begin{aligned}\beta_{1k}^{(v+1)} &= \frac{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij}) y_{ij}(t_i - t_T)}{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(t_i - t_T)^2} \\ &- \frac{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(y_{ij} - \beta_{1k}^{(v+1)}(t_i - t_T)}{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})} \\ &\left[\frac{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(t_i - t_T)}{\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})(t_i - t_T)^2}\right]\end{aligned}$$

$$\beta_{1k}^{(v+1)} = \frac{S^*_{k,T,j,1}}{S_{k,T,j,2}} - \left[\frac{S^*_{k,T,j,0} - \beta_{1k}^{(v+1)} S_{k,T,j,1}}{S_{k,T,j,0}}\right]\frac{S_{k,T,j,1}}{S_{k,T,j,2}}$$

Then

$$\beta_{1k}^{(v+1)} = \frac{S_{k,T,j,0}S^*_{k,T,j,1} - S_{k,T,j,1}S^*_{k,T,j,0}}{S_{k,T,j,2}S_{k,T,j,0} - S^2_{k,T,j,1}}$$

In addition, we can estimate $\beta_{0k}^{(v+1)}$ and $\beta_{1k}^{(v+1)}$ by solving the least square problem

$$\sum_{i=1}^{T}\sum_{l=1}^{J}\left[y_{ij} - \beta_{0k}^{(v+1)} - \beta_{1k}^{(v+1)}(t_i - t_T)\right]^2 r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})$$

Let $\mathbf{X}$ denote the $n \times 2$ matrix with i-th row $(1, t_i - t_T)$ and $\mathbf{W}$ denote the $n \times n$ diagonal matrix with i-th diagonal element $w_i(t_T) = r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})$, then

the local linear estimator defined by the coefficients $\beta_{0k}^{(v+1)}$ and $\beta_{1k}^{(v+1)}$ is

$$\beta_{0k}^{(v+1)} = e_1^T \left[\mathbf{X}^T\mathbf{W}\mathbf{X}\right]^{-1} \mathbf{X}^T\mathbf{W}\boldsymbol{y}$$

and

$$\beta_{1k}^{(v+1)} = e_2^T \left[\mathbf{X}^T\mathbf{W}\mathbf{X}\right]^{-1} \mathbf{X}^T\mathbf{W}\boldsymbol{y}$$

where $e_1 = (1,0)^T$, $e_2 = (0,1)^T$ and $\mathbf{y} = (y_1, \ldots, y_T)$.

$$\mathbf{X}^T\mathbf{W}\mathbf{X} = \begin{bmatrix} \sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i,t_T)V_k(y_{il},y_{ij}) & \sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i,t_T)V_k(y_{il},y_{ij})(t_i-t_T) \\ \sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i,t_T)V_k(y_{il},y_{ij})(t_i-t_T) & \sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i,t_T)V_k(y_{il},y_{ij})(t_i-t_T)^2 \end{bmatrix}$$

$$\begin{aligned} det(\mathbf{X}^T\mathbf{W}\mathbf{X}) &= \sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i,t_T)V_k(y_{il},y_{ij}) \sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i,t_T)V_k(y_{il},y_{ij})(t_i-t_T)^2 \\ &- \left[\sum_{i=1}^{T}\sum_{l=1}^{J} r_{ijk}^{(v+1)} W_k(t_i,t_T)V_k(y_{il},y_{ij})(t_i-t_T)\right]^2 \end{aligned}$$

Then

$$\mathbf{X}^T\mathbf{W}\mathbf{X} = \begin{bmatrix} S_{k,T,j,0} & S_{k,T,j,1} \\ S_{k,T,j,1} & S_{k,T,j,2} \end{bmatrix}$$

$$det(\mathbf{X}^T\mathbf{W}\mathbf{X}) = S_{k,T,j,0}S_{k,T,j,2} - S_{k,T,j,1}^2$$

$$\left[\mathbf{X}^T\mathbf{W}\mathbf{X}\right]^{-1} = \begin{bmatrix} \frac{S_{k,T,j,2}}{S_{k,T,j,0}S_{k,T,j,2}-S_{k,T,j,1}^2} & \frac{-S_{k,T,j,1}}{S_{k,T,j,0}S_{k,T,j,2}-S_{k,T,j,1}^2} \\ \frac{-S_{k,T,j,1}}{S_{k,T,j,0}S_{k,T,j,2}-S_{k,T,j,1}^2} & \frac{S_{k,T,j,0}}{S_{k,T,j,0}S_{k,T,j,2}-S_{k,T,j,1}^2} \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{W}\mathbf{y} = \begin{bmatrix} S_{k,T,j,0}^* \\ S_{k,T,j,1}^* \end{bmatrix}$$

Then

$$\beta_{0k}^{(v+1)} = \frac{S_{k,T,j,2}S_{k,T,j,0}^* - S_{k,T,j,1}S_{k,T,j,1}^*}{S_{k,T,j,2}S_{k,T,j,0} - S_{k,T,j,1}^2}$$

and

$$\beta_{1k}^{(v+1)} = \frac{S_{k,T,j,0}S_{k,T,j,1}^* - S_{k,T,j,1}S_{k,T,j,0}^*}{S_{k,T,j,2}S_{k,T,j,0} - S_{k,T,j,1}^2}$$

Proof of Equation 4.3.7

Setting $\partial \ell^* / \partial \sigma_j^{(v+1)} = 0$, one obtains

$$\sum_{i=1}^{T} \sum_{l=1}^{J} \sum_{k=1}^{K} G_{ijk} W_k(t_i, t_T) V_k(y_{il}, y_{ij}) \left[-\sigma_j^{2(v+1)} + (y_{ij} - \beta_{0kj}^{(v+1)} - \beta_{1kj}^{(v+1)}(t_i - t_T))^2 \right] = 0$$

Setting $G_{ijk} = r_{ijk}^{(v+1)}$, we have the following

$$\sigma_j^{2(v+1)} = \frac{\sum_{i=1}^{T} \sum_{l=1}^{J} \sum_{k=1}^{K} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij}) (y_{ij} - \beta_{0kj}^{(v+1)} - \beta_{1kj}^{(v+1)}(t_i - t_T))^2}{\sum_{i=1}^{T} \sum_{l=1}^{J} \sum_{k=1}^{K} r_{ijk}^{(v+1)} W_k(t_i, t_T) V_k(y_{il}, y_{ij})}$$