2010
2011

FACULTY OF SCIENCES

Master of Statistics: Epidemiology & Public Health Methodology

Masterproef

Development of a new semi-parametric mixture model for interval censored data, with applications in antimicrobial resistance

Promotor :
Prof. dr. Marc AERTS

Stijn Jaspers

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Epidemiology & Public Health Methodology

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Epidemiology & Public Health Methodology

Masterproef

Development of a new semi-parametric mixture model for interval censored data, with applications in antimicrobial resistance

Promotor :
Prof. dr. Marc AERTS

Stijn Jaspers

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Epidemiology & Public Health Methodology

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Acknowledgment

This thesis is the endpoint of my two-year journey through the Master in Statistics program at Hasselt University. During these years, I met a lot of people who made the journey easier and who contributed to the fact that time has flown by. Therefore, some words of appreciation are in place.

First of all, I would like to express my sincere gratitude to my supervisor, Professor Marc Aerts, who has been a great support throughout the thesis-writing months. His guidance and good ideas were extremely helpful for the preparation of this thesis and he kept me motivated to bring the job to a good end. Despite his busy schedule, he managed to find time for our weekly appointments and I could always count on him when a problem showed up. He also brought me in contact with Ruth Nysen, a PhD student who is working on a related topic. I also need to thank her for providing me with her opinions and kind words of encouragement.

I am indebted to my fellow students in the Master program. They all have been very kind and understanding, always prepared to help out when necessary. Special thanks goes to four friends I met during my Bachelor in mathematics at Hasselt University and who have accompanied me in the Master program: Steven, Yannick, Vicky and Katrien. They brought fun into the lunch breaks and provided some welcome moments of distraction and entertainment.

A final thank you is addressed to my other friends, family and of course to my parents for their loving support and their never ending belief in me. Their good advice and perfect care make it easier for me to fulfill my goals and dreams.

Summary

Due to the excessive and sometimes inappropriate usage of antimicrobials, there has been a continuing development and spread of pathogens with acquired resistance mechanisms. This antimicrobial resistance (AMR) has become one of the largest public health burdens of the last decades and it is therefore extremely important to study and keep track of the emerging of the resistance isolates.

In this thesis, interest was in exploring a new mixture model for AMR data such as Minimum Inhibition Concentration (MIC) values. Mixture models were ought to be ideally suited in this setting as they offer a natural framework for modeling the unobserved population heterogeneity of wild-type and resistant isolates. Since MIC values are often obtained using dilution type laboratory experiments, the proposed methods needed to account for the additional data complexity of interval censoring.

Besides the existing method of Turnidge et al. (2006) for estimating the parameters of the first component, three alternative methods were introduced. First, an adjustment was made to the existing method to make the transition between the wild-type and resistant component more graduate. Nevertheless, both the original and adjusted version suffered from the same shortcoming of not providing a direct means to identify the most suited distribution for the first component. Therefore, two approaches were developed that are encompassed in the more general maximum likelihood framework, namely the likelihood and multinomial based methods. Especially the latter, in combination with the averaged AIC selection procedure, was found to perform very well.

In order to incorporate information on the resistant isolates into the mixture model, the penalized mixture approach by Kauermann and Schellhase (2009) was presented. This semi-parametric density estimation routine was slightly adjusted in two ways with the aim of creating a full semi-parametric mixture model that is able to describe the MIC distribution. First of all, the censored nature of the MIC data was taken into account by using Gaussian distribution functions to construct the basis instead of the corresponding densities. The second adjustment is related to the wild-type component, which can be assumed to be of a well-known parametric form. More specifically, the first basis distribution is assumed to be equal to the wild-type component that was found via the methods described above. The according weight is estimated without posing a related penalty to the likelihood, while the remaining weights are estimated using the penalized approach.

In conclusion, the resulting semi-parametric mixture model does not only provide an estimate for the prevalence of resistant isolates, but also gives a good indication of their particular distribution. Hence, the developed model provides valuable information in the field of antimicrobial resistance and is considered to be useful in the important task of monitoring the emergence of resistance.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Ever since the discovery of penicillin in the late 1920's, antimicrobial agents have had a major impact on human and animal mortality and morbidity caused by microbial infections. Diseases that previously caused mortality and morbidity on a large scale were brought under control, causing several generations to grow up without the fear for infectious diseases that their forebears knew (Drusano, 2003). Unfortunately, due to an excessive and sometimes inappropriate usage, there has been a continuing development and spread of pathogens that have become resistant to antimicrobials. The emerging of resistance has already been thoroughly described by many authors for distinct kinds of microorganisms in both nosocomial and community settings. Tenover (2006) and references therein mainly deal with the mechanisms of antimicrobial resistance in bacteria. Examples include resistance of Escherichia coli against aminopenicillins and the well known methicillin-resistant Staphylococcus aureus (MRSA). Perea et al. (2001) on the other hand paid attention to resistance among fungi, in particular the Candida species, that has become evident since the introduction and widespread use of azole antifungals. Finally, a discussion of drug resistance in herpesvirusses and hepatitis B can be found in Strasfeld and Chou (2010). It should be clear that antimicrobial resistance (AMR) has become a global public health problem, posing a major threat to the successful use of antimicrobial agents in both human and veterinary medicine.

Antimicrobial resistance does not only result into increased morbidity and mortality, but has also a large impact on the costs of health care. In order to reduce these adverse effects, there is a need for prevention of the emergence and spread of resistant microorganisms. There is evidence that wiser use of antimicrobials may diminish the rate at which resistance emerges (Drusano, 2003). Keeping this in mind, medical care givers should properly reflect on the selection, dosing and duration of the treatment they prescribe. Nevertheless, prudent use of antimicrobials is not the only strategy for fighting AMR. Other infection control practices include good hand hygiene as well as screening and isolation of infected patients (ECDC, 2010).

Because of the major public health burden, it is very important to keep track of how the organisms are distributed in the population and whether they are resistant to a particular type of antimicrobial. In this perspective, the European Food Safety Authority (EFSA) analyses data from distinct member states (MS) on antimicrobial resistance in both zoonotic (e.g. Salmonella and Campylobacter) and non-zoonotic (e.g. Escherichia coli) microorganisms. Aerts et al. (2011) illustrate the use of statistical models for the analysis of temporal trends in the antimicrobial resistance data. To study the time trend,

a distinction is made between qualitative and quantitative data. Quantitative data are referring to the minimum inhibition concentration (MIC) values, as they are available for different MS. They can be treated in their original format or as binary outcomes, after dichotomization based on harmonized threshold values called Epidemiological Cut-Off values (ECOFFs). These binary outcomes, the wild-type and resistant isolates, are representing the qualitative data. Temporal trends based on the latter kind of data can be identified through the application of models for binary data, such as logistic regression. However, one of the drawbacks of this approach is the use of the threshold value, which might in some cases be disputable. In addition, trends in the MIC distribution above the threshold will not be observed, as all data are collapsed into the single category of resistant isolates. Whereas the MIC distribution to the left of the threshold is not expected to change over time (the wild-type distribution), one does expect to observe changes over time in the MIC distribution of the resistant isolates. To detect such changes over time one needs to consider the full ordinal or quantitative scale of the MIC distribution. Possible ways of analysis include the use of generalized logit models and mixture models. Both the quantitative and qualitative methods have their strengths and weaknesses, which are detailed upon in Aerts et al. (2011).

Interest in describing MIC distributions in a quantitative manner emerged during the process of European harmonization of breakpoints, such as the aforementioned ECOFFs, under the supervision of the European Committee on Antimicrobial Susceptibility Testing (EUCAST), see for example Kahlmeter et al. (2003) and Kahlmeter and Brown (2004). However, although attempts have been made, estimation of the wild-type distribution is still not solved as it is complicated by the fact that it is part of a mixture. Therefore, a first objective of this Master thesis is to study an existing method for estimation of this first components distribution. In addition, several new methods are presented and discussed on their performance. In a next stage, interest is in the development of a new semi-parametric mixture model for AMR data, such as MIC values, that takes into account both the wild-type and resistant component. With regard to the second component, Kauermann and Schellhase (2009) recently proposed a semi-parametric method to estimate the unknown density of a continuous variable, using a penalized mixture approach. Since MIC values are typically left or interval censored, the method of Kauermann and Schellhase (2009) will be extended to allow for this additional complexity. Based on this adjusted method, the new semi-parametric model will be developed.

The remainder of the thesis is organised as follows. Chapter 2 will provide more information on MIC values and how these fit into the mathematical framework of mixture models. A current approach in determining the wild-type distribution will be discussed in chapter 3, together with some own developed methods. In order to estimate the density of the second component, chapter 4 pays attention to the method of Kauermann and Schellhase (2009) and its extension toward censored data. The final semi-parametric mixture model will also be presented in that chapter and the proposed methods are applied to a real life dataset in chapter 5. Finally, a discussion will end the thesis in chapter 6.

Chapter 2

Mathematical Framework

In order to familiarize the reader with the topic of AMR data, this chapter will provide more information on the Minimum Inhibition Concentration (MIC) values. More specifically, the focus is here on how these values are obtained and attention is paid to their associated complexities. Secondly, it will be shown how the MIC values fit into the mathematical framework of mixture models, the further development of which will be the main challenge of the remainder of the thesis.

2.1 Minimum Inhibition Concentration

The Minimum Inhibitory Concentration (MIC) is the lowest concentration of an antimicrobial agent that will inhibit the visible growth of a microorganism. When analysing the antimicrobial susceptibility of certain populations of microbes, it is conventional to tabulate the number of isolates at specific MIC values, determined by employing serial twofold dilutions between selected maximum and minimum concentrations. The following example, based on Wu et al. (2008), gives a notion of how such a dilution experiment is carried out. In this example, E.coli and Salmonella bacterial cultures were tested against an array of 17 antimicrobial agents at a variety of concentrations in a microtitre plate. The output from testing one sample is a vector of MIC values for each antimicrobial agent, as determined by the broth microdilution method. Other techniques in determining MIC values exist (e.g. Jorgensen and Ferraro, 1998), but are not discussed here. Consider now an array of a single antimicrobial agent at concentrations of 0.125, 0.25, 0.5, 1, 2, 4 and 8 mg/L. A particular bacterial isolate may show inhibition of growth at 4 and 8 mg/L but growth at lower concentrations. The reported MIC value would then be 4 mg/L. This value means that a concentration of 4 mg/L showed inhibition, whereas a concentration of 2 mg/L did not. Consequently, the true inhibition is between 2 and 4 mg/L. Similarly, if the reported MIC value is less than 0.125 mg/L (i.e the lowest reading on the plate), all that is known is that the true inhibitory concentration is between 0 and 0.125 mg/L. Equally, a reported MIC value that is greater than 8 mg/L indicates that no concentrations on the plate showed inhibition of growth. Hence, the actual concentration that inhibited bacterial growth is greater than 8 mg/L. From this example it is clear that the used method only provides censored readings. The MIC value is only known to be either below the minimum concentration tested, between two concentrations or above the maximum concentration tested in the array for that antimicrobial agent. A more detailed

description of the preparation of plates in dilution experiments can be found in Schöne et al. (2009). It is clear that the additional complexity of censored readings needs to be taken into account when developing methods for the estimation of the MIC distribution. Methods developed throughout the paper are based on left- and interval censoring. However, extensions toward right-censoring are straightforward.

2.2 Mixture Models

The approach that will be followed here in regard to estimating the MIC distribution is based on mixture models, which have experienced increased interest in a variety of fields over the last decades. As such they have become well-recognized and popular methods. For example, Lindsay (1995) outlines several applications that indicate the wide scope of mixture models. Also in the field of antimicrobial resistance, they seem to be ideally suited since they offer a natural framework for modeling unobserved population heterogeneity. The term heterogeneity refers to the situation where the population of interest consists of various subpopulations. In the current setting, the interest is in a mixture model with the following hierarchical structure: at the first level, the model consists of two components:

$$f(x) = f_1(x|\theta_1)(1-p) + f_2(x|\theta_2)p. \tag{2.1}$$

The first component refers to the so-called wild-type distribution of MIC values, whereas the second component coincides with the resistant isolates distribution. The probability p is a parameter of major interest in the field of AMR since it refers to the prevalence of resistant isolates. Subject-matter knowledge suggests that the first component is of a well-known parametric form, such as the log-normal or gamma distribution (Lee and Whitmore, 1999 ; Turnidge et al., 2006). The second component on the other hand is often multimodal, as it is itself a mixture of different resistant strains. Therefore, this second component will be modeled using a second mixture:

$$f_2(x|\theta_2) = \sum_{l=1}^{m} f_l(x|\theta_{2l})p_l. \tag{2.2}$$

Depending on the interest and the typical application, the second component could be modeled as a classical mixture, allowing one to clearly distinguish the contributing components. Alternatively, it can be modeled as being a single density following the penalized mixture approach of Kauermann and Schellhase (2009). The latter will be further detailed upon in chapter 4. Since the method described here combines a parametric first component with a semi-parametric second component, the resulting mixture model will be of a semi-parametric nature.

Chapter 3

Estimation of the First Component

In the previous chapter, it was seen that the MIC data fit naturally into the mathematical framework of mixture models. The current chapter will study the wild-type first component in more detail. The main issue is to determine which distribution and according parameters are most suited to describe this first component, valuable information that is necessary to find the full mixture in a later stage. Initial attention will be paid to a method introduced by Turnidge et al. (2006) and an adjusted version thereof. Finally, two new methods, encompassed in a more general likelihood framework, are presented. Based on a small simulation study, the performance of the proposed methods is compared.

3.1 Existing Method with new Adjustment

Turnidge and colleagues (2006) developed a method for characterizing the wild-type MIC distribution from which one can derive the epidemiological cut-off values (ECOFFs). These ECOFFs separate microorganisms without (wild-type isolates) and with acquired resistance mechanisms (resistant isolates) to the agent in question. The proposed method is based on the assumption that the wild-type component of the MIC distribution follows a log-normal curve. Therefore, the logarithmic transformation was applied to the data, rendering a normally distributed first component. The idea is now to perform a non-linear least squares regression on the cumulative counts for a range of data subsets. Starting with the subset that includes values that are one dilution higher than the first mode, the cumulative counts are fitted to the cumulative normal curve. Three parameters will be estimated, namely the mean (μ) and standard deviation (σ) of the normal distribution function and the total number of observations (N) in the presumed subset. The used model can be described as follows:

$$y_i = f(x_i) + \epsilon_i = N * pnorm(x_i, \mu, \sigma) + \epsilon_i, \tag{3.1}$$

where y_i represents the cumulative counts upto group i, indicated by x_i, and $pnorm(.)$ the Gaussian cdf. After having obtained model estimates for the first subset of the data, the procedure is repeated with an augmented subset (i.e. MIC values that are 1 dilution higher are added to the previous subset). Figure 3.1 was obtained from Turnidge et al. (2006) and provides a visual representation of the proposed method. The authors argued that

the optimum fit was obtained when the difference between the observed and estimated number of isolates in the fitted subset was minimal.

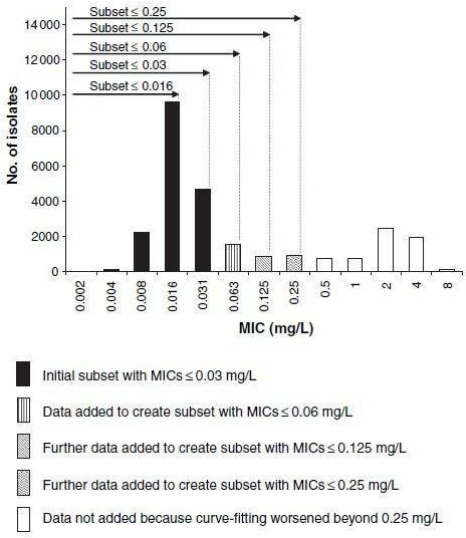

Figure 3.1: Iterative procedure for analysing increasing subsets of the data as found in Turnidge et al. (2006).

Since it is known that there is overlap between the wild-type and resistant component, there will be contamination of the first component by the second component from a certain point onwards. Therefore, the authors opted to estimate N, rather than taken it fixed. Nevertheless, differences between estimated and observed N can still be quite large, especially when the region of overlap is considerable. A straightforward idea is to extent the original method by adding pointwise new observations to the cumulative counts. Hence, a more graduate transition is made in the region of overlap, reducing the influence of contamination to the proposed method. The overall procedure and rule for selecting the optimal fit remain the same as originally presented. However, more iterations are involved since the cumulative counts are build up unit by unit. The latter implies that the adjusted method is slightly more computationally intensive compared to the original method.

3.2 Likelihood Based Method

Despite the fact that the method discussed in the previous section seems to perform well in practice (Turnidge et al., 2006), there remain some strong remarks about the assumptions that are made. First of all, the strongest assumption is about the distributional form of the wild-type component. The authors suggested the use of the log-normal cumulative distribution function, implying that the wild-type component belongs to a log-normal distribution. This assumption was derived through fitting the non-linear regression model using several bell-shaped distributions and applying a goodness-of-fit test. It could well be that other distributions (such as the gamma) are more appropriate and hence should be preferred. Therefore, it would be more desirable if a direct comparison could be made between distinct distributional assumptions, for example using the Akaike Information Criterion (Akaike, 1974). In addition, the non-linear least squares regression approach

can be considered an ad hoc method, requiring appropriate starting values to ensure convergence of the used algorithm. The selection of these starting values can often be quite time consuming, rendering this method less attractive. For these reasons, another approach is suggested that is encompassed in the more general framework of maximum likelihood. Of course, since the area of application stays the same, there still remain the complexities of having censored data and a region of overlap between the wild-type and resistant components. The latter difficulty is addressed with an idea similar to that of Turnidge et al. (2006), namely constructing the likelihood in a cumulative fashion.

A first important remark that needs to be made is about the nature of the used data. While the method discussed in the previous subsection requires the cumulative counts of the number of isolates in the distinct MIC categories, the likelihood based method considered below is based on the individual data points. Therefore, in this section, the notation x_i will refer to the ith observed MIC value. In case there are no censored values, the full likelihood can be expressed as $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$, where θ represents the parameters of the density function of interest. For instance, in case of the normal density function, θ corresponds to the vector (μ, σ^2), whereas the important parameters of a gamma distribution are its shape and scale. However, since it is known that the observed data are probably a sample from a mixture density, maximizing this full likelihood directly will not result into the desired estimates for the first component. Therefore, it might be useful to construct the likelihood stepwise, only using data points upto a certain MIC subgroup (cfr. figure 3.1). Denoting by n_k the number of observations smaller than or equal to the highest MIC category (k) under consideration, the following likelihood results:

$$L_k(\theta) = \prod_{i=1}^{n_k} \frac{f(x_i;\theta)}{1 - P(X > k)} = \prod_{i=1}^{n_k} \tilde{f}_k(x_i;\theta),$$

where the observations are ordered in increasing order. Note that since only part of the possible domain of data values is considered, the general density function cannot be used anymore. Rather, a truncated version of the density function of interest is considered, explaining the entrance of the denominator in the likelihood above. Indeed, adding that term renders a density function that only allows values that are smaller than or equal to k:

$$\int_{-\infty}^{k} \tilde{f}_k(x;\theta)\mathrm{dx} = \int_{-\infty}^{k} \frac{f(x;\theta)}{1 - P(X > k)}\mathrm{dx} = \frac{\int_{-\infty}^{k} f(x;\theta)\mathrm{dx}}{\int_{-\infty}^{k} f(x;\theta)\mathrm{dx}} = 1$$

The final adjustment that is to be made to the likelihood specified above is to account for the left- or interval censored nature of the data of interest. Therefore, the truncated cumulative distribution function will be used instead of the truncated density function. More specifically, it is specified that the observed MIC value is either contained within the interval spanned by two consecutive categories or below the lowest observable category (MIC_{min}):

$$L_k(\theta) = \prod_{i=1}^{n_k} \tilde{F}_k(x_i;\theta) I(x_i \in MIC_{min}) + (\tilde{F}_k(x_i;\theta) - \tilde{F}_k(x_i - 1;\theta)) I(x_i \notin MIC_{min}).$$

This way, all information that is available from the sample is used in the analysis. This approach to deal with censoring has been applied with success before in, amongst others, the related field of exposure assessment (Hewett and Ganser, 2007). In conclusion, the

final loglikelihood that will be maximized to find the optimal parameters of interest is:

$$l_k(\theta) = \sum_{i=1}^{n_k} \left\{ \log \left[\tilde{F}_k(x_i;\theta) \right] I(x_i \in MIC_{min}) + \log \left[\tilde{F}_k(x_i;\theta) - \tilde{F}_k(x_i - 1;\theta) \right] I(x_i \notin MIC_{min}) \right\}. \tag{3.2}$$

Maximization of the above loglikelihood will be performed for increasing subsets of the data, starting with the first four MIC categories. For each of the consecutive model fits, a Pearson chi-squared test statistic will be considered in order to determine the optimal fit (Agresti, 2002). The test statistic employed here can be used for testing the hypothesis $H_0 : F = F_0$ versus $H_1 : F \neq F_0$ and can formally be stated as

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}.$$

For each j, the observed counts are the total number of observations in the respective MIC category (e.g. corresponding to the frequencies indicated in figure 3.1). Summing up these subtotals provides the observed total counts after k categories, denoted by N_k. The expected counts are calculated based on the estimated parameters in the model fit corresponding to the first k MIC groups under consideration, i.e. $E_j = \hat{p}_j * N_k = \left[(\tilde{F}_k(j;\hat{\theta}) - \tilde{F}_k(j-1;\hat{\theta}) \right] * N_k$, j=1,...,k. The resulting test statistic has a chi-squared distribution with degrees of freedom equal to the number of MIC categories used in the fit lowered by one and the number of estimated parameters. For each of the considered subsets, the test statistic and corresponding p-value are calculated. Typically, the p-value will be larger than the presumed significance level $\alpha = 0.05$ for some initially fitted subsets. However, after a certain point, contamination by the second component will lead to the rejection of the nullhypothesis. The optimum fit and hence the most suited parameters correspond to the last fit that did not reject H_0.

Since the method presented here is based on optimizing a given likelihood function, less accurate starting values are needed compared to the non-linear least squares fit that is used in the method of Turnidge et al. (2006). In addition, no restriction is laid on the parametric form of the used density function. Hence, several distributional assumptions can be made and the resulting fits can be compared based on the well-known AIC criterion (Akaike, 1974). Note however that this criterion requires the models to be fit on the same data. For this reason, a drawback is that only comparisons can be made for model fits that were based on the same subset of the data. Two solutions to this problem will be presented in the next section and in chapter 4, making use of the multinomial distribution and the full mixture model respectively.

3.3 Multinomial Based Method

The likelihood based approach presented in the previous section was developed with the aim of creating a general framework for the comparison of several distributional assumptions about the first component. This goal was only partly reached as it was argued that in order to apply the AIC criterion, the models need to be fit to the same data. Hence, the desired comparison could only be made between models that were fit to the same subset of the data. Therefore, this section presents a first alternative which allows for a direct comparison between all model fits under consideration. A central role in this approach is put aside for the multinomial distribution.

3.3.1 Outline of the Proposed Method

As a consequence of the clustered nature of the data of interest, observations are grouped into several categories corresponding to their respective MIC values. Instead of assuming directly a continuous distribution for these outcomes, the groupings can initially be considered as being possible outcomes for a random variable following a multinomial distribution. This way, a saturated model can be constructed using the multinomial distribution as the core of the likelihood function. Consider for example an experiment with n independent trials in which k possible outcomes can be attained. Each of these outcomes has an according probability p_i such that $\forall i \in \{1, \ldots, k\} : 0 \leq p_i \leq 1$ and $\sum_{i=1}^{k} p_i = 1$. When the experiment is carried out, an array of random variables is observed $(X_1, \ldots, X_k)$, where X_i represents the number of times outcome i was observed over the n trials. This array follows a multinomial distribution with parameters n and $(p_1, \ldots, p_k)$, for which the probability function is given by

$$\begin{aligned} P(X_1 = x_1, \ldots, X_k = x_k) &= f(x_1, \ldots, x_k; n, p_1, \ldots, p_k) \\ &= \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}. \end{aligned}$$

The parameters of interest are the probabilities p_i for observing the distinct outcomes. Since the data x_i and the total number of trials n are fixed, the first term of the probability function above can be omitted when constructing the likelihood. Hence, the resulting simplified loglikelihood is given by $l(p_1, \ldots, p_k) = \sum_{i=1}^{k} x_i \log p_i$, maximization of which leads to the observed relative frequencies $\frac{X_i}{n}$ as being the maximum likelihood estimators for the probabilities p_i.

Nevertheless, interest remains in finding the density and corresponding parameters that best describe the wild-type component. This can be achieved by making use of the fact that the observed groupings are actually the result of the censored readings of the dilution experiment. Hence, the multinomial probabilities corresponding to a certain outcome i can be rewritten as

$$\begin{cases} \tilde{p}_i = F(u_i; \theta) \ldots \text{if i=1} \\ \tilde{p}_i = F(u_i; \theta) - F(l_i; \theta) \ldots \text{if otherwise,} \end{cases} \tag{3.3}$$

where u_i and l_i are the respective upper and lower values of the ith MIC category and $F(.)$ represents the cumulative distribution function under consideration, with θ the according parameters. The number of parameters that are contained in θ (e.g. two for a normal distribution, two for a gamma distribution,...) determines with how many MIC categories the procedure starts. More specifically, compared to the number of parameters used in the specified distribution, one additional category is needed in order to obtain an unsaturated model. For example, in case of the normal distribution function, the first three MIC categories form the starting point. The idea is now to tentatively replace some of the multinomial probabilities with their parametric counterparts in (3.3). The probabilities of the remaining outcomes are left unchanged and are thus to be estimated similar to those of the saturated model. The resulting sequence of likelihoods is specified in (3.4), where k_j indicates how many of the original multinomial probabilities are replaced: $k_j = \#$ of parameters used in the assumed distribution $+ j$. This sequence can be maximized to obtain several proposal estimates for the parameters of interest. Note that as a result of the parametric assumption, less parameters are used in the construction of the likelihood

when j increases. Since all and hence the same data are used in every step, the AIC criterion can be applied to select the most appropriate parameter estimates.

$$l_j(p_1, \ldots, p_k) = \sum_{i=1}^{k_j} x_i \log \tilde{p}_i + \sum_{i=k_j}^{k} x_i \log p_i, \quad j = 1, ..., k-2, \tag{3.4}$$

However, the approach described above can only be used in case there is no contamination by a second component, so when the data are not sampled from a mixture distribution. Therefore, in order to render the method useful for the situations of interest to this paper, some small changes are to be performed. Instead of using the probabilities as specified in (3.3), an additional parameter representing the mixing proportion needs to be specified:

$$\begin{cases} \tilde{p}_i = \pi * F(u_i; \theta) \ldots \text{if i=1} \\ \tilde{p}_i = \pi * [F(u_i; \theta) - F(l_i; \theta)] \ldots \text{if otherwise.} \end{cases} \tag{3.5}$$

Since an additional parameter is entered in the likelihood function, it is also necessary to augment the number of starting categories to render an unsaturated model. Hence, the sequence of likelihoods in (3.4) can still be applied, making use of the new parametric counterparts of the probabilities as specified in (3.5) and augmenting the used categories by one, representing the additional mixing parameter: $k_j = \#$ of parameters used in the assumed distribution $+j+1$. Again here, the AIC criterion can be used to select the most optimal parameter estimates.

3.3.2 Determination of Most Suited Parameters

In the previous section, it was argued to make use of the AIC criterion to select the most optimal model fit. Usually, when applying this criterion, the most suited parameter estimates are selected according to the minimum AIC value. However, when some of these values are relatively close together, one can also apply the model averaged approach as found in Burnham and Anderson (2002). The idea is to combine the obtained estimates into an averaged estimate as follows:

$$\hat{\theta}_a = \sum_{i=1}^{K} w_i \hat{\theta}_i,$$

where the w_i represent the so-called Akaike weights, which are defined as

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{j=1}^{K} \exp(-\frac{1}{2}\Delta_j)}.$$

These Akaike weights make use of the difference between the AIC value corresponding to a specific estimate $\hat{\theta}_i$ and the minimum observed AIC value: $\Delta_i = \text{AIC}_i - \text{AIC}_{min}$. When this difference is small, the Akaike weight is large and the corresponding estimate is allowed to contribute more to the averaged estimate. On the other hand, large differences are an indication of a less suited estimate and hence the small Akaike weight reduces the influence of that estimate to the averaged one.

Still another alternative to the AIC criterion is to apply a Deviance goodness-of-fit test, which compares the consecutive model fits to a saturated model:

$$D^2(M) = -2\left[\log(L(M)) - \log(L(Sat))\right],$$

which follows a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the saturated and reduced model. Large values for the Deviance indicate a lack of fit, hence suggesting that too many of the original probabilities are replaced by their alternatives in (3.5). Finally, the most parsimonious model that does not reject the Deviance GOF test will be selected as being most optimal. It is however possible that all of the considered models result into a test statistic that is too large and that none of them is deemed suited. In the latter situation, the estimates from another saturated model can be selected, i.e. the model where only the multinomial probabilities in the first s categories are replaced by their counterparts in (3.5), where s represents the number of parameters used in those adjusted probabilities.

3.4 Application to Artificial Datasets

In order to determine their respective behavior, the methods discussed in the previous sections are applied to two artificially created datasets. More specifically, with the aim of representing the study setting of interest, a random sample $X = (X_1, \ldots, X_n)$ was taken from two mixtures of three normal densities. The first dataset consists of 1000 sampled values from a mixture with a relatively small region of overlap and equally contributing components:

$$X_i \sim (1 - 0.5)\mathcal{N}(-2, 1) + 0.5\left\{0.5\mathcal{N}(3, 1.5^2) + 0.5\mathcal{N}(5, 1.5^2)\right\}. \tag{3.6}$$

The second dataset on the other hand was chosen in such a way that there is a considerable overlapping region, with a first component that has a larger weight compared to the second component:

$$X_i \sim (1 - 0.3)\mathcal{N}(-2, 1) + 0.3\left\{0.6\mathcal{N}(1, 1) + 0.4\mathcal{N}(5, 1)\right\}. \tag{3.7}$$

Again here, 1000 sampled values were obtained. The notation of the mixtures presented above represents the structure that is of basic interest in this paper (cfr. expressions (2.1) and (2.2)). For both sampled sets of data, the separate component densities can be found in figure 3.2, together with the resulting mixture density. Note that the first component of both mixtures corresponds to a normal density with mean equal to -2 and standard deviation equal to 1. These values are of primary interest in this chapter and will initially be estimated using the original and adjusted method of Turnidge et al. (2006). Consecutively, the obtained results are compared to the newly developed likelihood and multinomial based methods. The analysis is performed using R version 2.10.0 (code in appendix C). For simplicity, the sampled values will be referred to as being MIC values.

It has been argued that due to the nature of the laboratory experiments, MIC data are often left- or interval censored. For this reason, the sampled datasets are modified to account for this additional data complexity by rounding each value within a one-unit interval to the upper bound of that interval, i.e. if $c < x_i \leq c + 1$, the true value x_i is replaced by the integer value $c + 1$. For the two mixtures above, it is assumed that the lowest observable MIC value is equal to -4, hence the corresponding category will be interpreted as being left-censored, meaning that in fact the true MIC value for every observation in that category is smaller than or equal to -4. The resulting situation is depicted in the histograms in figure 3.3, a frequently used representation to visually study MIC data and breakpoints.

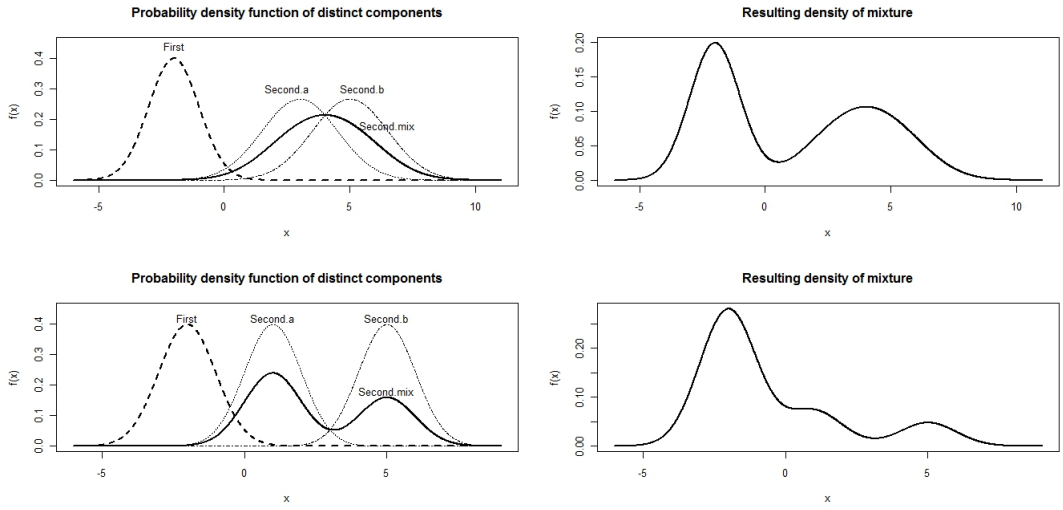

Figure 3.2: Density function and distinct component densities of the artificial data set of mixture (3.6) (top row) and mixture (3.7) (bottom row)

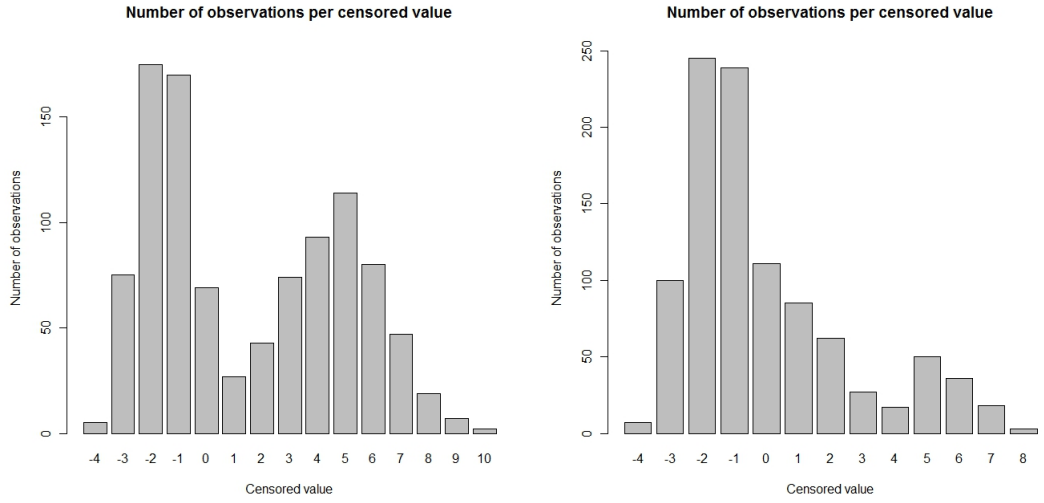

Figure 3.3: Histogram representing the distribution of censored sample from mixture (3.6) (left) and (3.7) (right).

3.4.1 Method of Turnidge et al. (2006)

Recall that the method of Turnidge et al. (2006) assumes that the first component follows a log-normal distribution. The samples in the two artificial examples are taken from a mixture with a normal first component and hence no log-transformation is needed anymore. Rather, model (3.1) can be applied directly to the data in table 3.1.

From the histograms in figure 3.3, it can be seen that the first mode is obtained at the MIC value of -2 for both samples. This implies that all values that are smaller than or equal to -1 are included in the first regression step of the procedure presented by Turnidge et al. (2006). Since the optimal fit can be found where the difference between

Table 3.1: Counts and cumulative counts per considered group, corresponding to the histogram plots in figure 3.3.

	Distribution mixture (3.6)												
	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
Counts	5	75	175	170	69	27	43	74	93	114	80	47	19
Cumulative	5	80	255	425	494	521	564	638	731	845	925	972	991
	9	10											
Counts	7	2											
Cumulative	998	1000											
	Distribution mixture (3.7)												
	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
Counts	7	100	245	239	111	85	62	27	17	50	36	18	3
Cumulative	7	107	352	591	702	787	849	876	893	943	979	997	1000

the estimated and true number of observations in the fitted subset was minimal, table 3.2 presents the results from the non-linear least squares regression fit, together with an additional column representing the aforementioned difference. For the first mixture, it

Table 3.2: Parameter estimates of the non-linear least squares regression approach.

	Estimates mixture (3.6)							
	Number of observations				*Mean*		*Standard deviation*	
Subset	True	Est.	Diff.	Std. Err.	Est.	Std.Err.	Est.	Std.Err.
-1	425	490.53	65.53	19.69	-2.05	0.07	0.95	0.06
0	494	503.16	9.16	5.84	-2.01	0.03	0.98	0.03
1	**521**	**515.37**	**-5.63**	**6.26**	**-1.97**	**0.04**	**1.03**	**0.05**
2	564	537.86	-26.14	13.22	-1.88	0.09	1.13	0.12
3	638	578.41	-59.59	25.02	-1.70	0.18	1.36	0.24
4	731	653.53	-77.47	43.01	-1.28	0.33	1.93	0.43
	Number of observations				*Mean*		*Standard deviation*	
Subset	True	Est.	Diff.	Std. Err.	Est.	Std.Err.	Est.	Std.Err.
-1	591	681.89	90.89	22.54	-2.04	0.05	0.94	0.05
0	702	717.70	15.70	10.73	-1.96	0.03	1.00	0.04
1	**787**	**772.28**	**-14.72**	**23.07**	**-1.82**	**0.09**	**1.15**	**0.11**
2	849	817.23	-31.77	26.12	-1.69	0.11	1.29	0.15
3	876	843.58	-32.42	23.41	-1.61	0.12	1.39	0.15
4	893	860.61	-32.39	20.63	-1.55	0.11	1.46	0.15

can be seen that the optimum fit corresponds to an estimated mean equal to -1.97 and estimated standard deviation of 1.03. Since there is only a small overlapping region, the estimated parameters are close to the true values. For the second mixture on the other hand, the large overlapping region results into less accurate estimates. The mean is estimated to be -1.82 and the corresponding estimated standard deviation is 1.15. Relative to the first mixture, the difference between observed end estimated number of observations in the optimal fitted subset was much larger for the second mixture (14.72 vs. 5.63). In

addition, taking a closer look at the results for the second mixture, it can be noted that the second most optimal fit (corresponding to a difference of 15.70) provides more accurate estimates for the parameters of interest. This again implies that improvements can be made with regard to the originally proposed method.

3.4.2 Adjustment to Method of Turnidge et al. (2006)

A first way to improve to previous method was to consider new cumulative counts in a pointwise fashion, hence reducing the influence of contamination by the second component. The exact procedure of how to perform this method becomes more clear through an application to the created datasets. Therefore, consider again the values observed in table 3.1. For the sample from mixture (3.6), it is seen that there are 27 observations with a value between 0 and 1. On the other hand, from figure 3.2, it is clear that it is unlikely that all of these 27 values are sampled from the first component. Despite the fact that the original method of Turnidge et al. (2006) takes this into account via the estimation of N rather than fixing these observations, the estimates for the mean and standard deviation are definitely influenced. The idea of the adjusted method is to add in each step one of the 27 observations and perform the analysis with the adjusted cumulative counts (498, 499,. . . , 521). This can be done for each of the original groups and the minimum difference per group can be compared. The results of this procedure can be found in table 3.3.

Table 3.3: Parameter estimates of the adjusted non-linear least squares regression approach.

		Estimates mixture (3.6)							
		Number of observations				*Mean*		*Standard deviation*	
Subset	Added	True	Est.	Diff.	Std. Err.	Est.	Std.Err.	Est.	Std.Err.
-1	1	256	256.25	0.25	4.95	-2.84	0.09	0.33	0.18
0	42	467	473.68	6.68	6.04	-2.10	0.03	0.90	0.04
1	**10**	**504**	**504.04**	**0.04**	**3.00**	**-2.01**	**0.02**	**0.99**	**0.02**
2	1	522	518.22	-3.78	4.49	-1.96	0.03	1.04	0.04
3	1	565	547.44	-17.56	11.47	-1.84	0.09	1.17	0.12
4	1	639	601.33	-37.67	23.57	-1.59	0.19	1.52	0.26
		Estimates mixture (3.7)							
		Number of observations				*Mean*		*Standard deviation*	
Subset	Added	True	Est.	Diff.	Std. Err.	Est.	Std.Err.	Est.	Std.Err.
-1	1	353	353.27	0.27	6.96	-2.84	0.12	0.32	0.22
0	59	650	659.00	9.00	7.56	-2.09	0.03	0.89	0.03
1	**18**	**720**	**720.01**	**0.01**	**5.38**	**-1.95**	**0.02**	**1.01**	**0.03**
2	1	788	780.71	-7.29	15.17	-1.80	0.07	1.17	0.09
3	1	850	831.44	-18.56	20.21	-1.65	0.10	1.34	0.13
4	1	877	854.93	-22.07	19.26	-1.57	0.11	1.43	0.14

Compared to table 3.2, an additional column is added to the table 3.3 representing how many observations are added pointwise to the previous cumulative count. For instance, the optimal fit for the first mixture (boldface) was obtained when adding 10 observations from the category of MIC values equal to one to the previous subset. This

way, the true number of observations in that optimal subset is equal to 494+10=504. The adjusted method seems to provide a considerable improvement to the estimated mean and standard deviation of the first component. Especially the estimates for the second mixture are closer to reality compared to the results in table 3.2. This can be explained by the simple fact that there was more contamination by the resistant component in case of the second mixture and therefore a larger impact could be observed compared to the results of the first mixture. Hence, in a situation similar to the second mixture, the adjusted method provides a very useful and more appropriate alternative to the original approach. Nevertheless, despite the near equality of observed and estimated number of observations, estimates are not perfect. This is probably due to sampling variability and indicates that one needs to be cautious about the obtained results and should not put blind thrust in them.

3.4.3 Likelihood Based Method

Despite the improved estimates, the adjusted method still suffers from the shortcoming of not being able to identify the most suited density to describe the first component. Therefore, the results obtained above are compared to those from the new likelihood based method, using a normal cumulative density function as proposed density. Table 3.4 presents the estimates that are obtained via maximizing the loglikelihood detailed in (3.2) with the employed endpoint specified in the corresponding first column of the table.

Table 3.4: Parameter estimates of the likelihood based method.

	Estimates mixture (3.6)					
	Mean		*Standard deviation*			
Endpoint	Est.	Std.Err.	Est.	Std.Err.	Likelihood	p-value
-1	-2.09	0.08	0.90	0.06	-464.80	0.0973
0	**-2.00**	**0.05**	**0.96**	**0.04**	**-665.23**	**0.1444**
1	-1.91	0.05	1.06	0.04	-777.08	0.0019
2	-1.64	0.06	1.38	0.05	-977.97	< 0.0001
3	-1.09	0.08	1.94	0.07	-1289.35	< 0.0001
4	-0.31	0.12	2.58	0.09	-1629.31	< 0.0001
	Estimates mixture (3.7)					
	Mean		*Standard deviation*			
Endpoint	Est.	Std.Err.	Est.	Std.Err.	Logikelihood	p-value
-1	-2.07	0.07	0.89	0.05	-642.28	0.0917
0	**-1.95**	**0.04**	**0.98**	**0.04**	**-950.47**	**0.0534**
1	-1.69	0.05	1.21	0.04	-1244.70	< 0.0001
2	-1.47	0.05	1.44	0.04	-1496.44	< 0.0001
3	-1.36	0.05	1.56	0.04	-1629.54	< 0.0001
4	-1.28	0.06	1.67	0.04	-1729.74	< 0.0001

It is observed that the current likelihood method provides more realistic estimates compared to the original method of Turnidge et al. (2006) and performs similar to the adjusted method. Only a minor difference is observed for the optimal estimates compared to those obtained in table 3.3, mainly with regard to the standard deviation parameter.

Recall that the loglikelihood that is given in the sixth column may not be used to compare between the fits corresponding to different endpoints since the respective model fits are based on different data points. Rather, it can be used for comparing between different proposal densities. Therefore, the procedure should be carried out in an equivalent way using the other proposal density and a within-row comparison based on AIC can give an indication of the most suited density.

3.4.4 Multinomial Based Method

Initial interest is in determining how the multinomial based method developed in this chapter performs in estimating the parameters of the first component based on the two mixture examples introduced before. Secondly, another mixture is considered in which the first component follows a gamma distribution. Hence in the latter example, not only the estimation of the parameters is important. Rather, the main task there is to select the most appropriate distribution and of course the according parameters.

With the initial aim of determining the estimation performance of the multinomial based method, the data in table 3.1 are again considered. Note that the current method only makes use of the observed counts per MIC category, so the cumulative counts can be disregarded. As can be seen, the sample resulting from mixture (3.6) consists of 15 MIC categories. Therefore, next to fitting the saturated multinomial model, the procedure also fits 12 additional models. The first of these additional models applies the adjusted probabilities in (3.5) to the first four categories. The final model under consideration replaces all multinomial probabilities with the respective alternatives in (3.5). The parameter estimates, together with the observed AIC value for all of these models are presented in table 3.5. In addition, the p-values corresponding to the Deviance GOF test can also be retrieved from the table. It can be seen that results are given for two saturated models (Sat 1 and Sat 2). The former corresponds to the multinomial model, whereas the second model results from replacing the probabilities only in the first three categories. Since three parameters are used in the replacement, the corresponding model is still saturated. Equivalently, the sample from mixture (3.7) consists of 13 MIC categories, leading to the fit of ten additional models. The resulting parameter estimates and corresponding AIC and p-values can again be found in table 3.5. For comparison purposes, recall that the true parameters of interest were π=0.5, mean=-2 and sd=1 for mixture (3.6) and π=0.7, mean=-2 and sd=1 for mixture (3.7).

After investigating the results for the first example mixture (3.6), several remarks can be made. First of all, disregarding the saturated models, the smallest AIC value is observed for the second additional model. The coinciding parameter estimates are 0.50, -2.00 and 0.96 for the mixing proportion π, the mean and standard deviation respectively. These parameter estimates are also obtained when using the Deviance criterion for selecting the most optimal model. However, since the AIC value of the first additional model does only differ slightly from the minimum, the model-averaged estimate is also influenced by this less appropriate estimate. The respective weights for additional models 1 and 2 are 0.41 and 0.58. The remaining models have a negligible impact on the final model-averaged estimate as their AIC values are too distinct from the minimum. Regarding the second example mixture (3.7), it can be seen that the AIC value increases with the considered additional models. The minimum AIC value is hence obtained for the first of these additional models, resulting into 0.67, -2.07 and 0.89 as estimates for the mixing proportion π, the mean

Table 3.5: Parameter estimates for mixtures (3.6) and (3.7), applying the multinomial based method. P-values refer to the Deviance test statistic.

	Mixture (3.6)					*Mixture* (3.7)				
Model	AIC	π	Mean	Sd	p-value	AIC	π	Mean	Sd	p-value
Sat 1	4775.34	-	-	-	-	4231.17	-	-	-	-
Sat 2	4775.34	0.33	-2.53	0.68	-	4231.17	0.47	-2.48	0.70	-
Add 1	4776.23	0.48	-2.09	0.90	0.0893	4232.06	0.67	-2.07	0.89	0.0893
Add 2	4775.56	0.50	-2.00	0.96	0.1211	4233.55	0.72	-1.95	0.98	0.0411
Add 3	4784.86	0.52	-1.91	1.06	0.0014	4281.19	0.80	-1.69	1.21	< 0.0001
Add 4	4880.65	0.57	-1.64	1.38	< 0.0001	4338.82	0.86	-1.47	1.44	< 0.0001
Add 5	5043.52	0.65	-1.09	1.94	< 0.0001	4361.95	0.88	-1.36	1.56	< 0.0001
Add 6	5164.31	0.77	-0.31	2.58	< 0.0001	4392.00	0.89	-1.28	1.67	< 0.0001
Add 7	5254.35	0.94	0.73	3.29	< 0.0001	4594.78	0.95	-0.96	2.10	< 0.0001
Add 8	5270.69	1.00	1.10	3.50	< 0.0001	4696.77	0.98	-0.72	2.41	< 0.0001
Add 9	5279.55	1.00	1.07	3.44	< 0.0001	4733.14	1.00	-0.60	2.57	< 0.0001
Add 10	5281.87	1.00	1.06	3.41	< 0.0001	4735.98	1.00	-0.59	2.58	< 0.0001
Add 11	5282.33	1.00	1.05	3.40	< 0.0001					
Add 12	5284.64	1.00	1.05	3.40	< 0.0001					
Model averaged		0.49	-2.04	0.93			0.68	-2.03	0.92	

and standard deviation respectively. Again, these values correspond to those obtained through application of the Deviance criterion. Nevertheless, the p-value corresponding to the second additional model fit is not highly significant. Rather, it could be termed as borderline, which would result into 0.72, -1.95 and 0.98 as being the estimates for the respective parameters of interest. It seems that in the current situation, the model-averaged estimates are closest to reality as they are 0.68, -2.03 and 0.92 respectively.

Next to the fact that the multinomial based method seems to provide adequate estimates for the parameters regarding the first component, its full strength becomes clear from the example to be discussed next. Interest in this example is namely to determine which distribution is most suited for describing the wild-type component. With this purpose, a new mixture is considered, with a first component corresponding to a gamma density with shape and scale parameters equal to 4 and 0.5 respectively. Hence the corresponding mean and standard deviation of the first component are 2 and 1. The second component is again assumed to be a mixture of two normal densities. Hence, the sample under investigation can be represented as

$$X_i \sim (1 - 0.5)\text{Gamma}(4, 0.5) + 0.5\left\{0.5\mathcal{N}(7, 1) + 0.5\mathcal{N}(10, 1)\right\}. \tag{3.8}$$

The described procedure is now applied two times, using respectively the normal and gamma cdf to obtain the probabilities in (3.5). The results are summarized in table 3.6 and the model fits for both procedures can be compared based on their AIC values.

Taking a first glance at the outcomes for the normal cumulative density function reveals that the most optimal parameters for the mean and standard deviation of the first component are equal to 1.87 and 0.84. The mixing proportion is estimated to be 0.48 and the corresponding AIC value is equal to 4663.38. Regarding the outcomes from the gamma cumulative density function, it is seen that the three mentioned selection criteria result into three distinct parameter estimates. Based on the minimum of the AIC values, the

Table 3.6: Parameter estimates for mixture (3.8) applying the multinomial based method with a normal and gamma cdf. P-values refer to the Deviance test statistic.

	Normal cdf					*Gamma cdf*				
Model	AIC	π	Mean	Sd	p-value	AIC	π	Shape	Scale	p-value
Sat 1	4662.22	-	-	-	-	4662.22	-	-	-	-
Sat 2	4662.22	0.46	1.79	0.75	-	4662.22	0.53	3.95	0.53	-
Add 1	4663.38	0.48	1.87	0.84	0.0755	4661.74	0.49	4.46	0.44	0.2183
Add 2	4688.43	0.50	1.97	0.99	< 0.0001	4662.71	0.51	4.07	0.50	0.1064
Add 3	4790.77	0.54	2.17	1.34	< 0.0001	4703.37	0.54	3.01	0.76	< 0.0001
Add 4	5018.19	0.65	2.93	2.23	< 0.0001	4853.62	0.70	1.71	2.06	< 0.0001
Add 5	5087.28	0.76	3.69	2.87	< 0.0001	4897.73	0.91	1.34	3.99	< 0.0001
Add 6	5107.01	0.83	4.19	3.25	< 0.0001	4905.23	1.00	1.28	4.80	< 0.0001
Add 7	5178.06	1.00	5.26	3.97	< 0.0001	5016.75	1.00	1.46	3.87	< 0.0001
Add 8	5212.59	1.00	5.19	3.81	< 0.0001	5116.40	1.00	1.58	3.40	< 0.0001
Add 9	5247.39	1.00	5.15	3.72	< 0.0001	5190.80	1.00	1.65	3.19	< 0.0001
Add 10	5253.45	1.00	5.15	3.71	< 0.0001	5202.27	1.00	1.65	3.17	< 0.0001
Model averaged		0.48	1.87	0.84			0.50	4.31	0.46	

estimates are found to be 4.46 for the scale and 0.44 for the shape. The mixing weight is estimated to be 0.49 and the corresponding AIC value is 4661.74. Applying the Deviance selection criterion, the estimates are somewhat closer to reality, namely shape equal to 4.07, scale equal to 0.50 and a mixing weight of 0.50. The AIC value corresponding to this model fit is 4662.71. Note that both of these AIC values are smaller than that of the optimal fit for the normal cdf. Hence, as it should be, the gamma distribution is deemed more suitable. Finally, the model averaged estimates for the gamma cdf are 0.50, 4.31 and 0.46 for the mixing weight, shape and scale parameters. In summary, the multinomial based approach seems to result into adequate estimates, with the additional advantage of providing a direct comparison between distinct distributional assumptions.

3.4.5 Comparison with Midpoint Approach

In order to take into account the censored nature of the data when using the likelihood based approach, the truncated distribution function was employed rather than the corresponding density function. It could be questioned whether this extension was worth the effort. For comparison purposes, the midpoint approach could be applied, where the interval censored observations are replaced by the midpoint of the according interval. Next, one can act as if these values were the true observations and apply the likelihood procedure with the truncated density. The outcomes for the two mixtures introduced at the beginning of the section can be found in table 3.7. It is seen that the obtained estimates are relatively close to those obtained in table 3.4. Regarding the second mixture, it is noted that none of the p-values obtained for the midpoint approach were larger than 0.05. Nevertheless, the first fitted subset yielded a borderline p-value and was selected as being most optimal. Based on the current observations, it seems that the censored adjusted version is somewhat more appropriate. However, since these results are only based on one single scenario of the mixtures of interest, a better comparison can be made through a simulation study, which is carried out in the next section.

Table 3.7: Parameter estimates of the midpoint approach.

	Estimates mixture (3.6)					
	Mean		*Standard deviation*			
Endpoint	Est.	Std.Err.	Est.	Std.Err.	Likelihood	p-value
-1	-2.07	0.08	0.96	0.05	-466.80	0.0581
0	**-1.99**	**0.05**	**1.02**	**0.04**	**-670.02**	**0.0785**
1	-1.91	0.05	1.11	0.04	-782.80	0.0052
2	-1.64	0.06	1.41	0.05	-982.40	< 0.0001
3	-1.09	0.08	1.96	0.07	-1293.06	< 0.0001
4	-0.31	0.12	2.59	0.09	-1633.18	< 0.0001
	Estimates mixture (3.7)					
	Mean		*Standard deviation*			
Endpoint	Est.	Std.Err.	Est.	Std.Err.	Logikelihood	p-value
-1	**-2.05**	**0.07**	**0.96**	**0.05**	**-645.12**	**0.0424**
0	-1.93	0.05	1.05	0.04	-956.56	0.0207
1	-1.69	0.05	1.26	0.04	-1250.66	< 0.0001
2	-1.46	0.05	1.48	0.04	-1502.23	< 0.0001
3	-1.36	0.05	1.59	0.04	-1635.62	< 0.0001
4	-1.28	0.06	1.70	0.04	-1735.86	< 0.0001

3.5 Simulation Study for Methods First Component

In the previous sections, several methods were introduced for estimating the parameters that characterize the wild-type component of the MIC distribution. Based on two artificial examples, their respective performances were examined. Nevertheless, the outcomes of the distinct methods discussed above are only based on one possible realization of the respective example mixtures. As a results, it could well be that the obtained insights are due to chance only. Therefore table 3.8 presents the results of a small simulation study, in which for each of the two mixtures of interest 200 samples were drawn. Consecutively, the mean and standard deviation of the first component were calculated for all of these samples, using the procedures proposed in the previous sections, namely the original and adjusted method of Turnidge et al. (2006) as well as the new likelihood and multinomial based approach. In addition to the first component parameters, the multinomial based method, that selects the most optimal parameters using AIC, Deviance and the Akaike Weights, also provides an estimate for the mixing weight π. The comparison between the applied methods is performed using the following quantities:

- $\text{Bias}(\hat{\theta}) = \text{E}(\hat{\theta}) - \theta$
- $\text{Var}(\hat{\theta}) = \text{E}[(\hat{\theta} - \text{E}(\hat{\theta}))^2]$
- $\text{MSE}(\hat{\theta}) = \text{E}[(\hat{\theta} - \theta)^2]$

The mean squared error (MSE) is of most interest since it combines information on bias and variance into one measure: $\text{MSE}(\hat{\theta}) = \text{Bias}^2(\hat{\theta}) + \text{Var}(\hat{\theta})$. In order for a direct comparison between the distinct methods, also the relative quantities are presented, with the smallest of the observed measures set equal to one.

Table 3.8: Simulation study for checking the performance of the discussed methods when estimating the parameters of interest for the first component in mixtures (3.6) and (3.7).

		Mixture (3.6)					
Method	Parameter	Bias	Variance	MSE	Rel. Bias	Rel. Var	Rel. MSE
Turnidge	mean	0.0595	0.0035	0.0071	3.1316	**1.0000**	1.0441
	Sd	0.0569	0.0034	0.0066	5.1727	**1.0000**	1.6923
Adjusted	mean	0.0223	0.0106	0.0111	1.1737	3.0286	1.6324
	Sd	0.0191	0.0087	0.0090	1.7364	2.5588	2.3077
ML	mean	0.0509	0.0063	0.0089	2.6789	1.8000	1.3088
	Sd	0.0401	0.0054	0.0070	3.6455	1.5882	1.7949
Midpoint	mean	0.0524	0.0075	0.0102	2.7579	2.1429	1.5000
	Sd	0.1084	0.0049	0.0166	9.8545	1.4412	4.2564
AIC	π	0.0058	0.0004	0.0004	**1.0000**	1.3333	1.3333
	mean	0.0190	0.0075	0.0078	**1.0000**	2.1429	1.1471
	Sd	0.0110	0.0042	0.0044	**1.0000**	1.2353	1.1282
Deviance	π	0.0148	0.0031	0.0033	2.5517	10.3333	11.0000
	mean	0.0513	0.0177	0.0204	2.7000	5.0571	3.0000
	Sd	0.0398	0.0075	0.0090	3.6182	2.2059	2.3077
Averaged	π	0.0066	0.0003	0.0003	1.1379	**1.0000**	**1.0000**
	mean	0.0230	0.0063	0.0068	1.2105	1.8000	**1.0000**
	Sd	0.0150	0.0037	0.0039	1.3636	1.0882	**1.0000**
		Mixture (3.7)					
Method	Parameter	Bias	Variance	MSE	Rel. Bias	Rel. Var	Rel. MSE
Turnidge	mean	0.1804	0.0037	0.0362	4.1281	**1.0000**	4.5823
	Sd	0.1586	0.0042	0.0293	6.8957	1.4000	8.3714
Adjusted	mean	0.0929	0.0078	0.0165	2.1259	2.1081	2.0886
	Sd	0.0720	0.0066	0.0118	3.1304	2.2000	3.3714
ML	mean	0.0682	0.0060	0.0107	1.5606	1.6216	1.3544
	Sd	0.0391	0.0031	0.0047	1.7000	1.0333	1.3429
Midpoint	mean	0.0871	0.0257	0.0332	1.9931	6.9459	4.2025
	Sd	0.0981	0.0500	0.0596	4.2652	16.6667	17.0286
AIC	π	0.0213	0.0010	0.0015	1.0047	1.6667	1.3636
	mean	0.0437	0.0081	0.0100	**1.0000**	2.1892	1.2658
	Sd	0.0230	0.0041	0.0047	**1.0000**	1.3667	1.3429
Deviance	π	0.0273	0.0035	0.0043	1.2877	5.8333	3.9091
	mean	0.0579	0.0186	0.0220	1.3249	5.0270	2.7848
	Sd	0.0335	0.0063	0.0074	1.4565	2.1000	2.1143
Averaged	π	0.0212	0.0006	0.0011	**1.0000**	**1.0000**	**1.0000**
	mean	0.0437	0.0060	0.0079	**1.0000**	1.6216	**1.0000**
	Sd	0.0231	0.0030	0.0035	1.0043	**1.0000**	**1.0000**

From table 3.8, some interesting observations can be made. Initially, the focus is put on the results from the first mixture example. Based on the MSE values for the procedures introduced in sections 3.1 and 3.2, the original method of Turnidge et al. (2006) seems to outperform the adjusted and maximum likelihood based method for the estimation of the mean as well as the standard deviation. This is mainly due to a reduced

variance compared to those other methods. Nevertheless, in terms of bias, the adjusted and maximum likelihood method seem to perform better. Comparing the selection procedures for the multinomial based approach introduced in section 3.3, it is seen that the MSE is lowest when an averaged estimate of the desired parameters is considered. The selection based on minimum AIC performs similar to the averaged method, whereas the Deviance criterion has much higher variance and MSE, especially for the mean parameter. The bias corresponding to the AIC selection criterion is lowest, closely followed by the averaged approach.

A slightly different observation can be made for the second mixture. Recall that this mixture had a larger region of overlap and hence a larger improvement could be seen between the estimates for the original and adjusted method of Turnidge et al. (2006) in tables 3.2 and 3.3. This is again reflected in this simulation study, since the adjusted method is found to be less biased and has a smaller MSE compared to the original method. A similar remark can be made about the ML approach, which even performs better compared to the adjusted method. Among the selection procedures for the multinomial based method, again the averaged approach performs very promising, closely followed by the AIC selection criterion. These two selection procedures outperform all of the considered methods, with a small preference for the averaged approach, mainly when regarding the mean parameter.

Previously, a comparison was made between the likelihood based method and the midpoint approach. Recall that the idea of the latter was to replace the interval censored values with the midpoints of the corresponding intervals and simply apply the truncated densities rather than the according distributions. Again there, it was argued that a simulation study would provide deeper insights in the performance of both methods. Therefore, the results for the midpoint approach can also be found in table 3.8. Comparing the MSE values, it is observed that the likelihood approach performs better when the censoring was taken into account. Whereas the effects on the estimate of the mean parameter are only minor, a relatively larger effect is seen for the estimate of the standard deviation. Regarding the first mixture, the MSE for the estimate of the standard deviation is 2.4 times larger for the midpoint approach. An even more pronounced difference is observed for the second mixture, where the MSE of the midpoint approach is 12.68 times larger compared to the censored adjusted approach. This huge difference for the second mixture is mainly due to an inflated variance for the midpoint approach when estimating the standard deviation parameter.

Some final notes have to be made with respect to the results of this simulation study. Similar to the remark made in section 3.7, 21 out of the 200 samples of the first mixture did not have any p-value that was larger than 0.05 when the midpoint procedure was employed. This number was even higher for the second mixture, namely 41 out of the 200 samples. Hence, for these samples, estimates were selected based on the highest observed p-value. Although this issue did not occur with the censored version of the likelihood based method in section 3.4.3, the simulation study revealed that 4 samples of the first and 7 samples of the second mixture had comparable problems. A second remark concerns the Deviance selection criterion. In the same 4 of the 200 samples from mixture (3.6) and 7 out of the 200 from mixture (3.7), none of the p-values were larger than 0.05. Hence, in these cases, estimates from the saturated model were employed as discussed in section 3.3.2.

3.6 Concluding Remarks

This chapter presented an overview of distinct methods to estimate the first component of a mixture distribution. Next to the existing method of Turnidge et al. (2006) and an adjustment to that approach, two new approaches were provided. Based on a simulation study, their respective performances were discussed. The adjustment that was made to the original method of Turnidge et al. (2006) was found to be a useful alternative when there was a considerable region of overlap. However, since this method suffers from the same shortcomings as the original method, interest was mainly in the maximum likelihood and multinomial based procedures. Although the former procedure caused some problems, this method is believed to perform well as its MSE is relatively close to the minimum. Nevertheless, it was the multinomial based procedure, with an averaged or minimum AIC selection criterion, that looks very promising since it performed best based on bias, variance and MSE, especially when regarding the second mixture.

Taking all of the former into account, it is believed that the adjusted method, the ML procedure and the multinomial based approach (using the AIC or averaged selection criteria) provide appropriate alternatives to the method of Turnidge et al. (2006). In addition to the performance outcomes found above, the multinomial based method has the additional advantage that it is encompassed in the more general likelihood framework, allowing for a straightforward comparison between distinct distributional assumptions regarding the first component.

Chapter 4

Semi-Parametric Mixture Model

The approaches discussed in the previous chapter provide different pathways for estimating the parameters for the first component of the MIC distribution. Nevertheless, main interest remains in obtaining an estimate for the full mixture MIC distribution. However, since little is known about the distribution of the resistant isolates, a more general form of density estimation needs to be applied to obtain more insight in this second component. After providing some background information on density estimation, the penalized mixture procedure (Kauermann and Schellhase, 2009) will be discussed in more detail. Consecutively, a censored-adjusted version of the method will be applied when estimating the full mixture with a semi-parametric mixture model.

4.1 Estimation of the Second Component

As already addressed above, little is known about the distribution of the resistant component of the MIC distribution. Most likely, this second component is itself a mixture as it is composed of several resistant strains of the microorganisms under investigation. Several approaches for estimating an unknown density exist and are presented in this section. Primary focus is on the method by Kauermann and Schellhase (2009).

4.1.1 Background Information on Density Estimation

Different density estimation routines exist and can roughly be categorized into four partly overlapping categories. One of the most prominent methods is nonparametric kernel density estimation, introduced by Rosenblatt (1956) and Parzen (1962). Formally this estimator can be represented as:

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \qquad x \in \mathbb{R}.$$

The window-width (or bandwidth) h_n is the parameter that determines the amount of smoothing. Different functions can be used as a kernel K(.), the choice of which does not pose any problems. However, to make sure that the expression above is a density function, the kernel function needs to integrate to one. Typically, it is assumed that the kernel function is non-negative and symmetric around zero. On the other hand, the determination of the optimal window-width is more important, but also more challenging.

Different approaches to determine the amount of smoothing exist and are discussed in for example Jones et al. (1996). A second technique for estimating an unknown density results by writing this density as its logistic density transform (Leonard, 1978)

$$\hat{f}(x) = \frac{\exp\left[\eta(x)\right]}{\int \exp\left[\eta(z)\right] \mathrm{dz}},$$

with $\eta(.)$ an unknown but smooth function, estimated using spline technology. More information on this method can be found in Gu and Wang (2003). A third approach is based on extending and smoothing the classical histogram as originally suggested by Boneva et al. (1971). Following this idea, Lindsey (1974a, 1974b) suggested to use a regression estimation scenario with the number of observations per bin in the histogram as Poisson count. Equivalently, Eilers and Marx (1996) make use of the same idea, but employ penalized spline smoothing instead. The spline approach and Poisson approach are thereby closely related which results by approximating the integral above with a rectangular method.

Next to the three approaches that are briefly discussed above, there also exists a fourth line of density estimation: using a mixture approach. In this case, the unknown density results by finite mixtures of component densities. These components can be either unknown, such as in the classical mixture models (e.g. McLachlan and Peel, 2000), or the components can be known. In the latter case, Ghidey et al. (2004) have proposed to use a finite but penalized mixture of Gaussian densities to estimate a random effects distribution in a linear mixed model. This idea has been adopted repeatedly and is further described in, amongst others, Komarek and Lesaffre (2008). Kauermann and Schellhase (2009) also follow the proposed method of penalized density estimation, focusing on both Gaussian densities and standardized B-splines as corresponding component densities. The latter approach will be further detailed upon in this section as it is believed to provide a meaningful way to estimate the unknown resistant distribution.

4.1.2 Penalized Mixture Approach

In order to obtain a semi-parametric estimate for the unknown density of the resistant isolates, the penalized mixture approach of Kauermann and Schellhase (2009) can be applied. Let X denote the univariate random variable of interest (i.e. the MIC value), with true density function f. The main idea is to approximate f as a mixture of densities:

$$f_K(x) = \sum_{k=-K}^{K} c_k \phi_k(x), \tag{4.1}$$

where the $\phi_k(x)$ are referred to as the basis densities and the c_k will be called the weights. In order for (4.1) to actually represent a density function, the weights need to be larger than zero and sum up to one. With the aim of avoiding constrained maximization, the weights are reparametrized:

$$c_k(\beta) = \frac{\exp(\beta_k)}{\sum_{k=-K}^{K} \exp(\beta_k)},$$

with $\beta_0 \equiv 0$ for identifiability and the $\beta = (\beta_{-K}, \ldots, \beta_{-1}, \beta_1, \ldots, \beta_K)$ such that $\int f_K(x)\mathrm{dx}$ =1. The current approach hence assumes that the basis densities are known and fixed

density functions with specified parameters. It is assumed that all the $\phi_k(x)$ are continuous on their support and converge toward zero at their boundary. One possible choice for the basis densities was already applied in for example Ghidey et al. (2004) and Komarek and Lesaffre (2008), namely to make use of Gaussian densities with fixed mean (μ_k) and variance (σ_k^2). The mean values can be referred to as being the knots of the basis. Kauermann and Schellhase (2009) also pay attention to the alternative B-spline densities, as they are numerically more stable and theoretically more appealing. B-spline densities are standard B-splines (de Boor, 1978) that are normed to be densities. A short introduction to B-splines and how they are standardized can be found in appendix A.

For convenience, the knots at which the basis densities are located will be denoted by μ_k, with k running from -K to K. The assumption is made that the knots cover the range of the observed values of the random variable X and that their location is fixed. In the typical setting, it is most easy to consider equidistant knots and hence this will also be assumed subsequently. Several options exist for the value of the variance of the Gaussian basis densities. Kauermann and Schellhase (2009) opted for a standard deviation that equals half the interval spanned by two consecutive knots: $\tau_1 = \frac{1}{2}(\mu_j - \mu_{j-1})$. On the other hand, the standard deviation employed by Ghidey et al. (2004) was slightly larger: $\tau_2 = \frac{2}{3}(\mu_j - \mu_{j-1})$. This was argumented by the fact that a Gaussian density which extends over $\mu \pm \tau_2$ can be approximated by a B-spline of degree 3 that extends over 4 equidistant knots. In the remainder of the thesis, the choice of Ghidey et al. (2004) will be followed.

It appears that the number of knots plays an important role in terms of bias and variance. When a fine grid for the knots of the basis densities is involved (large K), there will be overfitting which results into a density estimate that is too wiggly. When the grid is not chosen fine enough, meaning that there are only a limited number of component densities (small K), the result will be a relatively smooth but biased density estimate. For this reason, one needs to find a compromise between smoothness and unbiasedness. To accomplish this, the approach of Eilers and Marx (1996) will be followed. This means that a relatively large number of basis functions is considered, but the loglikelihood is penalized for overfitting via introducing a penalty term based on the finite differences of adjacent coefficients. Hence, the penalty is put on the basis coefficients β_k by penalizing the variation of c_k over k. Assuming an independent sample $x_i, i = 1, \ldots, n$, the loglikelihood can be written as

$$l(\beta) = \sum_{i=1}^{n} \log \sum_{k=-K}^{K} c_k(\beta)\phi_k(x_i).$$

As already mentioned, this loglikelihood is supplemented by adding a quadratic penalty term, yielding the penalized likelihood

$$l_p(\beta, \lambda) = l(\beta) - \frac{1}{2}\lambda\beta^T D_m \beta,$$

where the penalty matrix D_m implies smoothness and λ is the penalty parameter. In order to penalize the variation of the weights c_k, it suffices to restrict the difference between the coefficient β_k and β_{k-1} or β_{k+1} respectively. Therefore, mth-order differences are penalized. To obtain the final penalty matrix of interest, some intermediate results need to be presented. Let $\tilde{L}_m$ denote a $(\tilde{K} - m) \times \tilde{K}$ dimensional matrix representing the (m+1)th order differences, with $\tilde{K} = 2K + 1$. For example, the second order difference

matrix $\tilde{L}_1$ is equal to

$$\tilde{L}_1 = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{pmatrix}.$$

By definition, $\beta_0 \equiv 0$, implying that the linear combination with this coefficient can be omitted. Therefore, the $(\tilde{K}-m)\times 2K$ dimensional matrix L_m is derived from the difference matrix $\tilde{L}_m$ by omitting the redundant middle column corresponding to β_0. Finally, the penalty matrix D_m is obtained as $L_m^T L_m$.

In order to maximize the penalized likelihood, a Newton-Raphson approach will be followed. Hence, the first and second derivative are needed. In this regard, denote with $\mathcal{C}(\beta)$ the $(2K+1)\times(2K)$ dimensional matrix with elements

$$\frac{\partial c_k(\beta)}{\partial \beta_j},\ k=-K,\ldots,K,\quad j=-K,\ldots,-1,1,\ldots,K.$$

This matrix results as $\mathcal{C}(\beta) = \left(\text{diag}(\tilde{c}) - \tilde{c}\tilde{c}^T\right)[,\{-K,\ldots,-1,1,\ldots,K\}]$, where $[,A]$ refers to extracting the columns given by the index set A and $\tilde{c} = (c_{-K}(\beta),\ldots,c_0(\beta),\ldots,c_K(\beta))^T$. With these notations, the first derivative of the penalized loglikelihood with respect to β now equals

$$s_p(\beta,\lambda) = \frac{\partial l(\beta)}{\partial \beta} - \lambda D_m \beta = \sum_{i=1}^{n} \frac{\mathcal{C}^T(\beta)\tilde{\boldsymbol{\phi}}_i}{f(x_i)} - \lambda D_m \beta,$$

with $\tilde{\boldsymbol{\phi}}_i = (\phi_{-K}(x_i),\ldots,\phi_0(x_i),\ldots,\phi_K(x_i))^T$ and f(x) as defined in (4.1). The negative second order derivative of the penalized loglikelihood may be approximated by

$$J_p(\beta,\lambda) = -\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta} + \lambda D_m \approx \sum_{i=1}^{n} \frac{\mathcal{C}^T(\beta)\tilde{\boldsymbol{\phi}}_i\tilde{\boldsymbol{\phi}}_i^T\mathcal{C}(\beta)}{f(x_i)^2} + \lambda D_m.$$

In order to find the maximum likelihood estimate for β, Newton-Raphson scoring will be performed using a fixed λ:

$$\beta_{t+1} = \beta_t - 2^{-v}\{s_p(\beta_t,\lambda)(J_p(\beta_t,\lambda))^{-1}\}.$$

At the start of the process, v is set equal to zero. When no new maximum is achieved for a current v, this value is increased step by step and hence the step size is bisected. This iterative procedure is repeated until the new loglikelihood value (corresponding to β_{t+1}) differs less than 0.1 loglikelihood points from the loglikelihood corresponding to the old estimate.

While having kept the penalty parameter λ fixed when updating the estimate of β, it is equally important to obtain a good estimate for this parameter as well since it steers the amount of smoothing. Estimation of λ can be obtained in two ways. First of all, the best penalty parameter can be seen as the one that minimizes the Akaike Information Criterion (Akaike, 1974). For a given λ, $\text{AIC}(\lambda) = -2l_p(\beta,\lambda) + 2\text{dim}(\beta|\lambda)$, where $\text{dim}(\beta|\lambda)$ represents the effective degrees of freedom depending on λ. Applying a method by Gray (1992), these effective degrees of freedom can be determined using the observed Fisher information matrices based on the penalized and unpenalized likelihoods: $\text{dim}(\beta|\lambda) =$

trace$(J_p^{-1}(\beta,\lambda)J_p(\beta,\lambda=0))$. Note that when λ equals zero, no penalty is put on the loglikelihood and hence dim$(\beta|\lambda=0)$ reduces to 2K, the number of used parameters. However, selecting λ using this first method requires a grid search and fitting the density for a set of λ values, which is usually quite time consuming. Therefore, Kauermann and Schellhase (2009) proposed an alternative method making use of the link with mixed models. Adopting a Bayesian viewpoint, they derived an estimating equation from which one could derive the value of λ:

$$\hat{\lambda}^{-1} = \frac{\hat{\beta}^T D_m \hat{\beta}}{dim(\hat{\beta}|\hat{\lambda}) - (m-1)}.$$

The derivation of the formula above is rather technical and one is referred to Kauermann and Schellhase (2009) for more information. Since both sides of the equation depend on the parameter of interest, an iterative solution is possible by fixing λ on the right hand side and update the parameter on the left hand side. Consecutive iterations can be performed after updating again the right hand side. These iterations are terminated when the denominator is smaller than some prespecified threshold (0.01) or when the new λ is approximately converged (i.e. the new value differs by only 0.001 times the old value from the old value of λ). In case none of those two criteria are met, the iteration for λ is terminated after eleven steps.

The fitting of the presented procedure requires a number of practical settings which are implemented int the R package pendensity (Schellhase, 2009). A uniform distribution is assumed as starting values for the β parameters, i.e. the β_k are set equal to zero in order to start the Newton procedure. To avoid terminating the algorithm in a local instead of global maximum, it is advisable to fit the density for a number of different starting values and take the fit with the maximum value for the likelihood. Even more important in avoiding the problem of a local maximum is the choice of the penalty parameter, which should be large enough. It is therefore recommendable to start the Newton procedure with a large λ (e.g. 50000). As a default, the number of knots is set equal to 41, mimicking the rule of thumb suggested in Ruppert (2002). However, the number of knots does not seem as influential on the fit as the amount of smoothing, which is determined by the selection of the penalty parameter. Finally, the procedure repeatedly iterates between the Newton-Raphson step and the estimating equation above for estimating respectively β and λ, until the value for λ has converged. From the resulting estimates for β, the weights c_k can be calculated and substituted in (4.1) to obtain an estimate for the density of interest.

4.1.3 Extension to Censored Data

The penalized mixture approach was introduced with the aim of obtaining a suitable way to estimate the density of the resistant component of the MIC density. However, as a result of the dilution type laboratory experiments, MIC data are censored and the penalized mixture approach of Kauermann and Schellhase (2009) cannot be applied directly. Rather, a small adjustment needs to be made regarding the used basis functions. More specifically, in total equivalence with the idea used when developing the likelihood based approach in section 3.2, the original basis density functions are replaced with their corresponding distribution functions when constructing the likelihood:

$$l(\beta) = \sum_{i=1}^{n} \log \sum_{k=-K}^{K} c_k \Phi_k(x),$$

where $\Phi_k(.)$ can be either the Gaussian or B-spline cdf (for the latter, see appendix A). The penalization and optimization of the likelihood are done similar to the original procedure. The resulting estimates can again be substituted in equation (4.1) to obtain an estimate of the desired density.

4.2 Semi-Parametric Mixture Model

After having presented approaches to estimate the wild-type and resistant component, the developed ideas will be combined here to create a full semi-parametric mixture model for the estimation of the MIC density. Based on the resulting mixture density, one can derive the prevalence of the resistant isolates and perform for instance model based classification to determine whether the isolates under investigation belong to the wild-type or resistant class.

Recall from chapter 2 that the general form of the mixture distribution of the minimum inhibition concentration values is given by

$$f(x) = f_1(x|\theta_1)(1-p) + f_2(x|\theta_2)p.$$

The first component can be assumed to be of a fixed parametric form, for which parameter estimates were found in chapter 3. However, major interest remains in finding an estimate for the prevalence p of the resistant isolates. The multinomial based method already provided an estimate for this proportion. However, the latter method assumed a multinomial distribution for the outcomes of the resistant isolates, an assumption that could be improved. The only available information about the second component is that it is itself a mixture of several resistant strains. Therefore, an elegant way to incorporate this information into the model is through the application of the penalized mixture approach of Kauermann and Schellhase (2009).

4.2.1 Outline of the Proposed Method

The idea is to assume that the parameters of the first component, θ_1, are known and equal to their estimates that were found via the corresponding developed methods. The information on the second component will be introduced by the application of a method that is similar to the method of Kauermann and Schellhase (2009). More specifically, the estimator for the density of the MIC values is based on (4.1), to which one additional component is added:

$$f_K(x) = (1-p)f_1(x;\theta_1) + p\sum_{k=-K}^{K} c_k\phi_k(x) = \sum_{k=-(K+1)}^{K} \tilde{c}_k\tilde{\phi}_k(x). \tag{4.2}$$

This component represents the wild-type component and will not be penalized for as it is assumed to be of a fixed parametric form. This means that $f_1(x;\theta_1)$ is taken to be equal to the estimated density found in chapter 3. Note that the density estimator above has been rewritten in terms of (4.1), where the weights have been rescaled such that the first weight corresponds to $(1-p)$. The according first basis density is assumed to be equal to $f_1(x;\theta_1)$. The estimator in (4.2) is again used to construct the penalized likelihood. Note however that, in contrast to the approach in the previous chapter, there is no need

to attach a penalty to the parameter of the first basis function. Therefore, the penalty matrix D_m, introduced earlier, is supplemented with an initial row and column of zeros. Optimization occurs equivalently to the method in section 4.1.2.

Of course, one needs to account for the fact that the estimated parameters of the first component are assumed to be fixed. In this regard, a bootstrap procedure (Efron and Tibshirani, 1994) could be carried out that repeats both steps of the analysis, namely the estimation of the wild-type component and the fitting of the corresponding semi-parametric mixture model. This way, the variability related to the estimated prevalence through the latter model is taken appropriately into account.

One of the main difficulties with the proposed approach is the placement of the knots that are needed to find the penalized mixture estimate of the resistant component. This is mainly due to the fact that these knots need to span the region of MIC values that belong to resistant isolates, information that is unknown. Using a too wide region might lead to fitting unexplained sampling variability that was not captured by the fixed first component, resulting into random fluctuations in the neighborhood where in fact no resistant MIC values were observed. On the other hand, a range that is too small results into an inappropriate estimate in the region of overlap between first and second component, a region of particular interest. One of the most plausible options to solve this issue was to locate the base densities at equidistant knots ranging between the mean of the first component and the maximum value in the observed dataset. As a result, observations that are smaller than or equal to the estimated mean of the first component are contributed solely to the wild-type component and do not influence the penalized mixture estimate of the resistant component. Nevertheless, it is highly recommended to assume different knot ranges and compare the corresponding model fits based on the AIC criterion.

4.2.2 Obtaining Estimates for First Component

The approach, as it is described above, largely depends on the estimate of the first component parameters that were found using the procedures described earlier. Instead, the semi-parametric mixture model can be used as an alternative approach to obtain suitable estimates for the parameters of the wild-type component. The idea is very similar to the multinomial based approach that was introduced in section 3.3, where the multinomial distribution was employed as a means to incorporate information on the resistant component. A natural adjustment is to replace the multinomial distribution by the density estimate that is found using the penalized mixture approach of Kauermann and Schellhase (2009). Hence, to determine the value for θ_1, a grid of possible values for this parameter vector can be considered and the most suited among them can be selected based on the Akaike Information Criterion. Note that the same criterion was also addressed as a possible way to select the most optimal penalty parameter. However, since it was time consuming, the faster estimating equation was preferred. A similar remark can be made here as the grid search will be more time consuming compared to for example the multinomial based methods in section 3.3.

4.2.3 Model Based Classification

Next to having obtained an estimate for the prevalence of the resistant isolates, the semi-parametric mixture model has another nice feature. The resulting estimate for the full

mixture density of the MIC data can be used to determine to which of the two classes a specific isolate belongs: wild-type or resistant. This way, the mixture provides a useful alternative to the epidemiological cut off values (ECOFFs) that are typically used to determine resistance. Whereas the latter approach can be termed deterministic (the isolate is either resistant or not), the mixture approach allows for a probabilistic interpretation. More specific, the probability for a certain isolate to be in the wild-type class is equal to

$$P(\text{wild-type}|\text{MIC} = x) = \frac{(1-p)f_1(x;\theta_1)}{(1-p)f_1(x;\theta_1) + p\sum_{k=-K}^{K} c_k\phi_k(x)}.$$

Classification can be made according to the majority vote, in which case the isolate is classified to the wild-type strain if the probability specified above is larger than 50 %. For more information, see e.g. Hastie et al. (2009).

4.3 Application to Artificial Datasets

The simulation study that was performed in the paper by Kauermann and Schellhase (2009) revealed a promising result for the original approach compared to the alternatives in section 4.1.1. In order to determine how the censored-adjusted version developed in this chapter behaves, the procedure was applied to the two artificial examples from chapter 3. While the method is actually introduced to capture the density of the resistant isolates, it will be applied here to model the total mixture, hence representing the full MIC density of interest. Instead of providing the exact estimates of the weights c_k, a visual comparison is preferred. In this regard, figure 4.1 presents the resulting density estimates for mixtures (3.6) and (3.7).

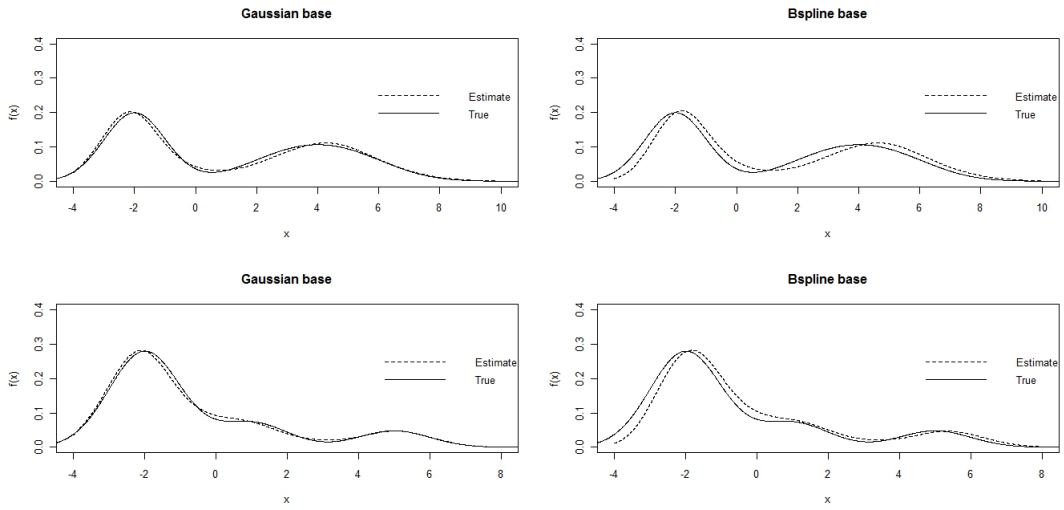

Figure 4.1: Graphical representation of estimated density for mixtures (3.6) (top row) and (3.7) (bottom row) using Gaussian and B-spline bases.

It is seen that the true density (black solid line) is approximated nearly perfect when a Gaussian base is employed. The estimate is able to capture both peaks of mixture (3.6) and the region of overlap in mixture (3.7) does not result into large problems. On the other hand, when a B-spline basis is used, the resulting plot is slightly shifted to the right,

probably because of the censored nature of the data. From these two examples, it seems that a B-spline basis is not that well suited, in contrast to its magnificent performance in the original method of Kauermann and Schellhase (2009).

In addition, the techniques developed in the previous section are also applied to the two artificial examples. Estimates for the parameters of the first component were already obtained through the application of the six procedures developed in chapter 3. The idea is now to consider each of these estimates as being the fixed first component parameters (θ_1) and substitute the corresponding wild-type density ($f_1(x;\theta_1)$) into expression (4.2). Consecutively, the censored-adjusted version of the procedure by Kauermann and Schellhase (2009) is used to estimate the weights $\tilde{c}_k$. Since the first of these weights represents the proportion of wild-type isolates, one minus this proportion reflects the prevalence of the resistant isolates. For each of the six possible scenarios, table 4.1 presents the estimate for this prevalence together with the AIC value of the corresponding model fit. Gaussian cumulative distribution functions were used to construct the base.

Table 4.1: Estimate for prevalence of resistant isolates using the full mixture approach with estimates for the parameters of the first component as found by the methods discussed in chapter 3.

	Mixture (3.6)			*Mixture* (3.7)		
Method	Estimates θ_1	Prevalence	AIC	Estimates θ_1	Prevalence	AIC
Turnidge	-1.97; 1.03	0.497	4765.13	-1.82; 1.15	0.258	4252.95
Adjusted	-2.01; 0.99	0.502	4762.22	-1.95; 1.01	0.293	4235.87
ML	-2.00; 0.96	0.504	4761.36	-1.95; 0.98	0.298	4234.63
AIC	-2.00; 0.96	0.504	4761.36	-2.07; 0.89	0.342	4232.62
Deviance	-2.00; 0.96	0.504	4761.36	-2.07; 0.89	0.342	4232.62
Averaged	-2.04; 0.93	0.509	4761.00	-2.03; 0.92	0.325	4232.72

Regarding the first mixture, all parameter settings result into estimates for the prevalence of resistant isolates that closely approximate the true value of 0.5. In addition, the corresponding AIC values are fairly close to each other and all parameter estimates, except those obtained from the original method of Turnidge et al. (2006), seem to be equally plausible. Based on this criterion, the results from the multinomial based method with averaged AIC selection criterion are most optimal. Similar observations can be made for the second mixture, except that the prevalence estimates are somewhat more distinct. The smallest AIC value can be found for the multinomial based approach with minimum AIC and Deviance as selection criteria. However, the solution for the averaged selection criterion has only a marginally larger AIC value and is preferred based on the simulation study in chapter 3.

Next to having obtained an estimate for the prevalence of resistant isolates, the procedure also provides an estimate for the density of the resistant component. Instead of showing the exact estimates for the weights, a graphical representation is ought to be more convenient. Therefore, figure 4.2 displays the resulting densities of the full mixture and of the resistant isolates for both mixtures under consideration. Graphs are shown for the results corresponding to the estimates who are indicated as being most suited based on the AIC values in table 4.1.

For the first mixture, an excellent fit is observed for both the full mixture and the corresponding second component density. On the contrary, a different observation

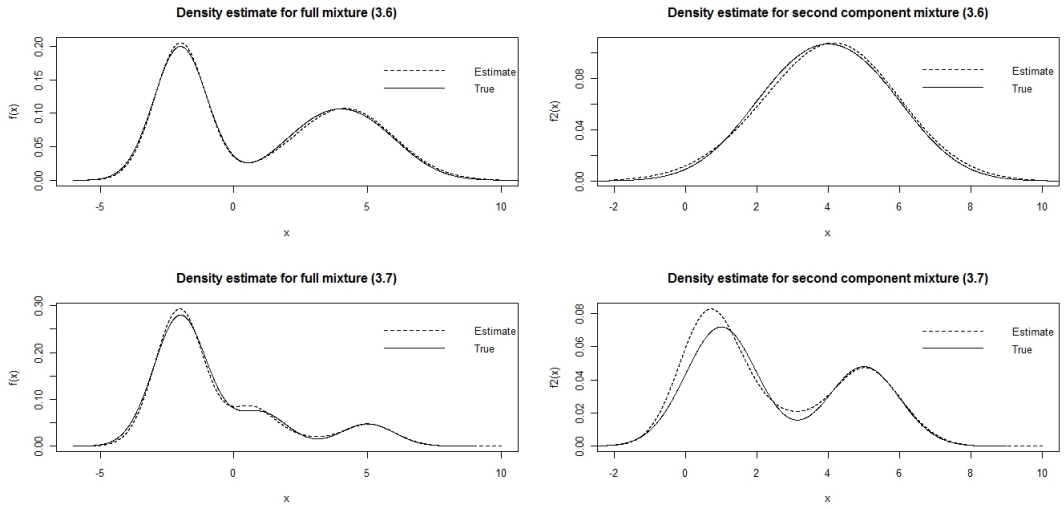

Figure 4.2: Graphical representation of estimated density for the entire mixture (3.6) (top row) and (3.7) (bottom row), with corresponding second component densities using the parameters corresponding to the minimum of the AIC values in table 4.1, using Gaussian basis densities.

is made for the second mixture. Whereas the fit for the full mixture is reasonably well, the corresponding second component density estimate is less accurate. The procedure is still able to identify two peaks for this resistant component, but the first peak is shifted slightly to the left. This is probably due to the fact that the second component needed to compensate for the less appropriate estimates for the first component (which were in fact too small). In addition, there is also variability that is inherent to working with samples and this definitely has an influence on the obtained results.

4.4 Concluding Remarks

In this chapter, next to providing a means to estimate the density of the resistant isolates, a semi-parametric mixture model to capture the full MIC distribution was presented as well. Gaussian basis densities were deemed most appropriate since the alternative B-spline densities resulted into a shifted estimate. The resulting semi-parametric mixture model provided an excellent fit for the artificial example with only a small region of overlap. Both the density of the entire mixture and the corresponding second component were estimated close to their true counterparts. When the region of overlap was larger, less accurate results for the first component parameters resulted into a less appropriate resistant density estimate, while the fit for the entire mixture was still satisfactory.

In order to take the sampling variability into account, the same curve-fitting procedure was applied to the 200 samples from the simulation study in the previous chapter. Although the output is not presented above, similar observations were made and the aforementioned conclusions were confirmed.

Chapter 5

Application to Real Data

The methodology developed in the previous chapters for estimating the MIC distribution seemed to perform promising when applied to the artificial examples. However, although it was aimed for, it is not sure that the created examples represent real AMR data. Therefore, in order to determine how the developed methods behave in such a real life situation, they will be applied to an original dataset in this chapter. Estimation of the first component is performed using the original method of Turnidge et al. (2006), as well as the newly developed multinomial based approach. These estimates are consecutively used to construct the full mixture density of the MIC data.

5.1 Description of Data

The European Committee on Antimicrobial Susceptibility Testing (EUCAST) is an organization that deals with breakpoints and technical aspects of phenotypic in vitro antimicrobial susceptibility testing. Most antimicrobial MIC breakpoints (e.g. epidemiological cut-off values) in Europe have been harmonised by EUCAST. On their website, the committee provides MIC distributions for a wide range of organisms and antimicrobial agents. These distributions are based on collated data from a total of close to 20000 MIC distributions from worldwide sources. The distributions include MIC values from national and international studies such as resistance surveillance programs, as well as MIC distributions from published articles, the pharmaceutical industry, veterinary programs and individual laboratories. The developed methods in chapter 3 are applied to one antibiotic-bacterium combination, namely the resistance of Escherichia coli against ampicillin.

The resulting MIC distribution consists of 39220 isolates that were obtained from 48 distinct sources. All data were collected in the form of number of isolates with different MIC values on the power of 2 scale. MIC values ranged from 0.125 mg/L to 512 mg/L, with the first mode being located around the value of 2 mg/L. A graphical representation of the data is given by the barplot in figure 5.1. Two large peaks are clearly visible at the values of 2 and 4, probably representing the center of the wild-type component. Toward the larger MIC values, two smaller peaks are located at the values of 64 and 256, which could represent distinct strains of resistant isolates. The EUCAST has set the harmonized epidemiological cut-off value equal to 8 mg/L, meaning that isolates with a larger MIC value are referred to as being resistant.

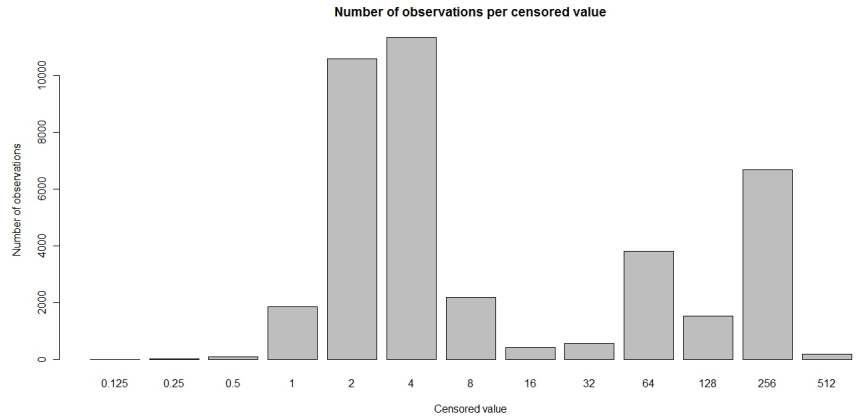

Figure 5.1: Dataset regarding the AMR data for ampicillin - E. coli example.

5.2 Application of Proposed Methods

In first instance, the data were subjected to the original method of Turnidge et al. (2006) to obtain an estimate for the parameters of the first component. As outlined in section 3.1, non-linear least squares regression was employed to fit the cumulative $\log_2$-transformed MIC distribution to the normal cumulative function for increasing subsets of the data. The most optimal parameters are those who minimize the difference between the observed and estimated cumulative count in the fitted subset. Table 5.1 presents the outcomes for the initial three fitted subsets. The first one of these subsets included the isolates that had an MIC value (on the $\log_2$ scale) smaller than or equal to 2, corresponding to the second observed mode of the distribution under investigation. It is seen that the difference between estimated and true number of isolates is minimal for the second subset. Hence, the estimates for the mean and standard deviation were selected from this subset. The mean MIC is estimated to be 1.04 and the corresponding standard deviation is 0.71 (both on the $\log_2$ scale). From the third fitted subset onwards, the absolute difference became larger again.

Whereas the original method makes use of the normal cdf for modeling the subsequent subsets, it could well be that another parametric distribution function is more appropriate. Even though the approach of Turnidge et al. (2006) does not allow for a direct comparison between two distributional assumptions, the procedure was carried out a second time, using the gamma cdf instead. Since a gamma distributed random variable is strictly positive, the original MIC scale is used here. The results of the second procedure are also summarized in table 5.1. Selection of the most optimal parameters is done in total similarity as above, leading to an estimate of 4.65 for the shape and 0.48 for the scale parameter.

Although the difference between the estimated and true number of observations in the optimal subset (266.87) is somewhat larger compared to when the normal cdf was employed, this does not necessarily imply that the latter is most suited. Rather, a direct comparison of the two assumptions should be carried out, using for instance the multinomial based method in combination with the AIC criterion. The results of this approach, assuming both a normal and gamma first component, can be found in table 5.2. On the $\log_2$ scale the AIC and model averaged selection procedures result into an

Table 5.1: Parameter estimates according to the method of Turnidge et al. (2006), applied to the ampicillin - E. coli data.

Normal cdf								
	Number of observations				*Mean*		*Standard deviation*	
Subset	True	Est.	Diff.	Std. Err.	Est.	Std.Err.	Est.	Std.Err.
1	23846	26267.86	2421.86	67.25	1.04	0.00	0.72	0.00
2	26027	26123.42	96.42	33.74	1.04	0.00	0.71	0.00
3	26450	26289.19	-160.81	76.55	1.04	0.01	0.72	0.01
Gamma cdf								
	Number of observations				*Shape*		*Scale*	
Subset	True	Est.	Diff.	Std. Err.	Est.	Std.Err.	Est.	Std.Err.
1	23846	24699.58	853.58	22.46	5.54	0.03	0.38	0.00
2	26027	25760.13	-266.87	290.51	4.65	0.42	0.48	0.05
3	26450	26089.21	-360.79	246.59	4.42	0.41	0.51	0.05

estimated mean of 1.51 and 1.50 respectively, and a standard deviation equal to 0.91 for the Gaussian wild-type component. Note however that the mixing weight π is estimated to be 1, implying that all isolates belong to a single normal density. Since this is very unlikely, a better selection criterion would be to compare the AIC values for the model fits where the corresponding estimate of π is smaller than 1, leading to an estimated mean of 1.06 and sd equal to 0.74. In case of a gamma first component, the AIC selection procedure results into estimates of 5.56 and 0.86 for the shape and scale parameter respectively, corresponding to a mean of 2.11 and a standard deviation of 0.80 (on the original scale). The results of the averaged approach are only marginally different since the shape is estimated to be 5.54 and the scale 0.40. Based on the AIC values in the table, it is observed that the gamma distribution performs slightly better compared to the Gaussian.

Table 5.2: Parameter estimates of the multinomial based method with a normal and gamma cdf, applied to the ampicillin - E. coli data.

	Normal cdf					*Gamma cdf*				
Model	AIC	π	Mean	Sd	p-value	AIC	π	Shape	Scale	p-value
Sat 1	143007.6	-	-	-	-	143007.6	-	-	-	-
Sat 2	143007.6	0.03	0.57	1.14	-	143007.6	0.01	4.92	0.13	-
Add1	143044.9	1.00	1.51	0.91	< 0.0001	143026.3	1.00	5.41	0.45	< 0.0001
Add2	143050.1	1.00	1.40	0.86	< 0.0001	143027.0	0.61	5.66	0.37	< 0.0001
Add3	143086.7	0.68	1.06	0.74	< 0.0001	143025.1	0.63	5.56	0.38	< 0.0001
Add4	143098.2	0.67	1.03	0.72	< 0.0001	143321.7	0.66	4.51	0.51	< 0.0001
Add5	143572.8	0.67	1.06	0.77	< 0.0001	145284.7	0.67	3.67	0.66	< 0.0001
Model averaged		1.00	1.50	0.91			0.73	5.54	0.40	

Whereas the multinomial approach already provides a first comparison between the distinct distributional assumptions, it does not take into account the full distribution of the second component. Therefore, a second way to compare between the assumptions for the first component is to make use of the semi-parametric mixture model. This way,

information on the distribution of the resistant isolates is incorporated as well via the semi-parametric density estimate of Kauermann and Schellhase (2009). A comparison is made between all pairs of parameter estimates for the first component that were obtained above. Each of these pairs are substituted in expression (4.2) and assumed to represent the true parameters of the wild-type component. Consecutively, optimisation of the resistant component is performed using the penalized mixture approach. Figure 5.2 gives a visual comparison of the fitted models, overlaid on the barplot of the original data. Note that all graphs are made on the $\log_2$ scale.

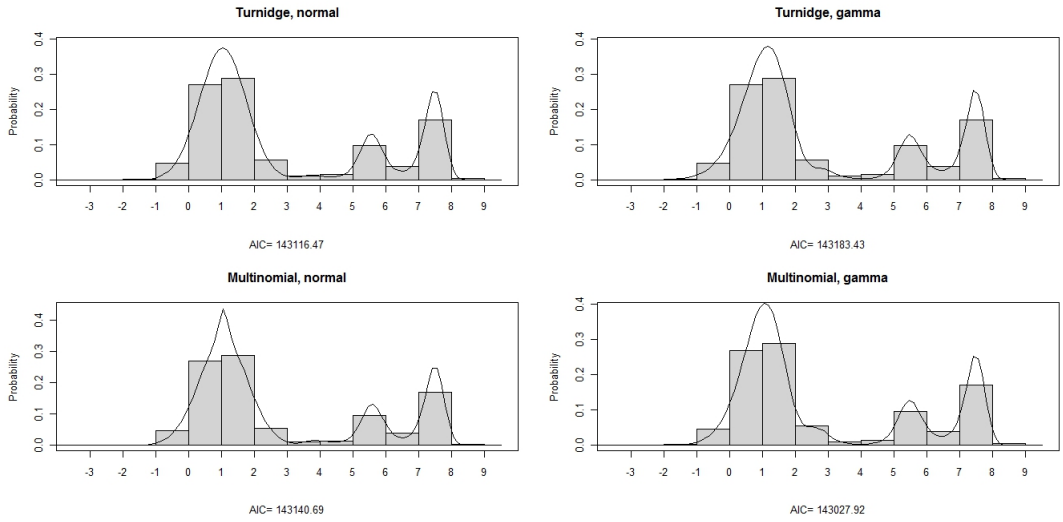

Figure 5.2: Fitted semi-parametric models for AMR data regarding the ampicillin - E. coli example.

The smallest AIC value (143028) is observed for the estimates of the multinomial based method with a gamma first component (bottom, right). The second most optimal estimates were obtained via the method of Turnidge et al. (2006) with a normal cdf (143116). Using the updated selection criterion above, the multinomial based approach with a normal first component resulted into similar estimates as those from the method of Turnidge et al. (2006). Nevertheless, the AIC value for the former is considerably higher, namely 143141. This difference in fit is also observed in figure 5.2. It is seen there that for the multinomial approach, the normal first component is not able to explain all observations in the MIC category of 2 mg/L. The semi-parametric density estimate for the resistant isolates therefore shows a first peak at the respective category, explaining the departure from normality there. Finally, the least suited estimates were obtained via the application of Turnidge et al. (2006) with a gamma first component. For comparison purposes, the AIC value when the full MIC density was estimated using the penalized mixture approach is equal to 143017, hence relatively close to the most optimal solution.

Recall that the placement of the knots for the penalized mixture approach was one of its major issues. In the fits above, the knot range was determined by the mean of the first component and the maximum MIC value in the dataset. Other knot ranges were considered and compared to the fits above. In case of a gamma first component, one could use the mode rather than the mean. In terms of AIC, this resulted into an improved model fit when applied to the estimates of the method of Turnidge et al. (2006), namely 143065. Nevertheless, assuming different knot ranges for the other procedures as

well, the solution regarding the multinomial approach in combination with a gamma first component remained most optimal.

The resulting semi-parametric mixture found above can now be employed to determine some properties of interest. First of all, the initial weight of the mixture corresponds to the prevalence of wild-type isolates. For the ampicillin - E.coli combination, 63% (bootstrap se: 0.041) of the isolates is estimated to belong to the wild-type component and hence the remaining 37% are resistant. In order to determine to which class a specific isolate belongs, model based classification can be performed as described in section 4.2.3. Applying the majority rule, isolates with an AIC value less than or equal to 4 would be classified as being wild-type organisms. This cut-off value for classification is however not in accordance with the harmonized ECOFF as stated by the EUCAST, which is equal to 8 mg/L. According to the model based classification, the probability to belong to the wild-type class when the MIC value is equal to 8 mg/L is only 0.4%. In the current example, this means that only 9 out of the 2181 obserations in the latter category are wild-type. On the contrary, classification based on the second most optimal fit (assuming a normal first component) coincided with the harmonized cut-off value.

5.3 Concluding Remarks

In this chapter, the method of Turnidge et al. (2006) and the new multinomial based approach were applied to an AMR dataset regarding the resistance of E.coli against ampicillin. For both approaches, two distributional assumptions for the first component were compared based on the AIC values resulting from the full semi-parametric mixture model. Via the multinomial approach, the gamma distribution with shape and scale parameter equal to 5.56 and 0.38 turned out to be most suited. Based on the resulting mixture, the prevalence of resistant isolates was estimated to be 37%. This value needs to be interpreted with caution since data from distinct time periods and countries have been aggregated.

Due to convergence problems, results for the likelihood based approach were not included. Different starting values for the parameters of interest were assumed, but did not provide a solution to the problem. In addition, strict application of the minimum difference rule for selecting the most optimal subset for the adjusted method of Turnidge et al. (2006) lead to irrelevant results and hence these were also discarded in the previous section. However, as can be seen from the related output in appendix B, the second most optimal fit resulted in results similar to the original method. As a final note, a new issue was encountered in the current application, namely a mixing weight estimated to be 1. Since it is known that the MIC distribution is in fact a mixture, the selection criterion was adapted in such a way that these implausible estimates are discarded.

In conclusion, it can be stated that the newly developed multinomial method provides an adequate alternative to the method of Turnidge et al. (2006); not only performing promising in artificially created examples, but in real life situations as well. Whereas the likelihood and adjusted method also seemed to work well when applied to the artificial mixtures, some problems occurred in this chapter. Nevertheless, it is believed that all the developed methods can provide valuable information in distinct situations, with the multinomial based approach having the most broad range of application.

Chapter 6

Discussion and Further Research

In this thesis, interest was in exploring a new mixture model for antimicrobial resistance data such as Minimum Inhibition Concentration (MIC) values. Mixture models were ought to be ideally suited in this setting as they offer a natural framework for modeling the unobserved population heterogeneity of wild-type and resistant isolates. Since MIC values are often obtained using dilution type laboratory experiments, the proposed methods needed to account for the additional data complexity of interval censoring. In first instance, the focus was put on estimating the wild-type component of the mixture, which is assumed to be of a well-known parametric form. The existing method of Turnidge et al. (2006) makes use of the log-normal distribution and aims at determining the most optimal parameters through the application of non-linear least squares regression to a range of data subsets. An adjustment to this method was presented to make the transition between the wild-type and resistant component more graduate. Nevertheless, both the original and adjusted version suffered from the same shortcoming of not providing a direct means to identify the most suited distribution for the first component. Therefore, the likelihood and multinomial based methods were developed, approaches that are encompassed in the more general maximum likelihood framework. The former method is related to the method of Turnidge et al. (2006) in the sense that it uses increasing subsets of the data to obtain optimal parameter estimates. Hence, different distributional assumptions can only be compared within a particular subset since the AIC criterion requires the models to be fit on the same data. On the other hand, the multinomial based approach uses all data at hand and therefore does not suffer from this restriction. In addition to being most suited to compare between distinct distributional assumptions, a simulation study also revealed that the multinomial procedure (in combination with the minimum or averaged AIC selection criteria) outperformed the other proposed methods in terms of MSE.

Whereas the wild-type component could be assumed to be of a parametric form, less is known about the resistant component of the mixture distribution. Since it is believed that the latter component consists of several resistant strains, it was modeled using a second mixture. Instead of estimating the resistant component as a classical mixture, a censored-adjusted version of the penalized mixture approach of Kauermann and Schellhase (2009) was preferred. This density estimation routine was considered to be an elegant way of incorporating information on the resistant isolates. Hence, the final semi-parametric mixture model was created via extending the penalized mixture approach in such a way that it was able to account for a fixed first component. Due to the fact that this first component was actually estimated, a bootstrap procedure was proposed to

take into account the additional variability when estimating the prevalence of the resistant isolates. However, other alternatives exist and should be further explored.

Application of the proposed methods to a real life dataset provided some additional interesting insights into their respective behavior. One of the most important observations was related to the selection criteria for the multinomial approach. Compared to the artificial examples, the mixing weight was estimated to be 1 for the initial model fits. A situation that is not realistic since the data are believed to arise from a mixture distribution. Therefore, the new selection criterion only included model fits that provided estimates for the mixing weight smaller than 1. Other remarks concerned the convergence problems for the likelihood approach and the inadequate estimates as a result of the too strict application of the minimum difference rule for the adjustment to the method of Turnidge et al. (2006). However, because of their promising performance in the artificial examples, it is believed that also these latter methods can provide useful alternatives to the original method of Turnidge et al. (2006). Regarding the full semi-parametric mixture model approach for selecting the most optimal parameters of the first component, it was noted that the AIC values for some of the fitted models were very close together (difference of 0.1). In such situations, it is very important not to rely solely on this selection criterion, but also to keep in mind the results of the simulation study that was carried out in chapter 3. More specifically, when a choice needs to be made between several close model fits, a slight preference is given to the multinomial based procedure with the averaged AIC selection criterion.

Both Gaussian densities and standardized B-splines were used as corresponding basis densities for the penalized mixture approach. Nevertheless, in contrast to its magnificent performance in the original method, the B-spline basis did not provide adequate estimates when used in the adjusted version. Especially in the initial range of observations, the resulting density estimate was shifted. The discrepancy between the Gaussian and B-spline basis became dramatic when the latter was employed in the full semi-parametric mixture model, as it failed to produce appropriate estimates for the first, and hence also for the second, component. For this reason, the Gaussian base was considered to be most suited for subsequent modeling. However, because of their excellent performance in the original setting, it is worthwhile to further explore the behavior of these B-splines in combination with censored data. Initial efforts have already been undertaken but were not conclusive in determining the particular cause of the shifted estimate. Therefore, future research in this interesting field is highly recommended.

References

Aerts, M., Faes, C. and Nysen, R. (2011). Request for assistance in development of statistical analyses of methods for evaluation of data on antimicrobial resistance in zoonotic agent isolates from animals and food . *Scientific report submitted to EFSA*.

Agresti, A. (2002). *Categorical Data Analysis*. New Jersey: John Wiley and Sons.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19,** 716-723.

Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodel Inference : A Practical Information-Theoretic Approach*. New York: Springer.

de Boor, C. (1978). *A Practical Guide to Splines*. Berlin: Springer.

Drusano, G.L. (2003). Prevention of resistance: a goal for dose selection for antimicrobial agents. *Clinical Infectious Diseases*, **36**, 42-50.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman and Hall/CRC.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**,89-121.

European Centre for Disease Prevention and Control. Antimicrobial resistance surveillance in Europe 2009. Annual Report of the European Antimicrobial Resistance Surveillance Network (EARS-Net). Stockholm: ECDC; 2010.

European Committee on Antimicrobial Susceptibility Testing. Data from the EUCAST MIC distribution website, last accessed 25/08/2011. http://www.eucast.org.

Ghidey, W., Lesaffre, E. and Eilers, P.H.C. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics*, **60**, 945-953.

Gu, C. and Wang, J. (2003). Penalized likelihood density estimation: direct cross-validation and scalable approximation. *Statistica Sinica*, **13**,811-826.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning, Data Mining, Inference, and Prediction.* New York: Springer.

Hewett, P. and Ganser, G.H. (2007). A comparison of several methods for analyzing censored data. *The Annals of Occupational Hygiene*, **51**, 611-632.

Jones, M.C., Marron, J.S. and Sheather, S.J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, **91**, 401-407.

Jorgensen, J.H. and Ferraro, M.J. (1998). Antimicrobial susceptibility testing: general principles and contemporary practices. *Clinical Infectious Diseases*, **26**, 973-980.

Kahlmeter, G., Brown, D.F., Goldstein, F.W., MacGowan, A.P., Mouton, J.W., Osterlund, A., Rodloff, A., Steinbakk, M., Urbaskova, P. and Vatopoulos, A. (2003). European harmonization of MIC breakpoints for antimicrobial susceptibility testing

of bacteria. *Journal of Antimicrobial Chemotherapy*, **52**, 145-148.

Kahlmeter, G. and Brown, D.F. (2004). Harmonization of antimicrobial breakpoints in Europe, can it be achieved?. *Clinical Microbiology Newsletter*, **26**, 187-192.

Kauermann, G. and Schellhase, C. (2009). Density estimation with a penalized mixture approach. Technical Report, Centre for Statistics, Bielefeld University.

Komárek, A. and Lesaffre, E. (2008). Generalized linear mixed model with a penalized Gaussian mixture as random effects distribution. *Computational Statistics and Data Analysis*, **52**, 3441-3458.

Lee M.-L.T. and Whitmore G.A (1999). Statistical inference for serial dilution assay data. *Biometrics*, **55**, 1215-1220.

Leonard, T. (1978). Density estimation, stochastic processes and prior information. *Journal of the Statistical Society Series B*, **73**, 113-146.

Lindsey, J.K. (1974a). Comparison of probability distributions. *Journal of the Royal Statistical Society Series B*, **36**, 38-47.

Lindsey, J.K. (1974b). Construction and comparison of statistical models. *Journal of the Royal Statistical Society. Series B*, **36**, 418-425.

Lindsay, B.G. (1995). *Mixture Models: Theory, Geometry and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 5*. Hayward: Institute of Mathematical Statistics.

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*. New York: John Wiley.

Parzen, E. (1962). On the estimation of a probability density and the mode. The Annals of Mathematical Statistics, **33**,1965-1976.

Perea, S., Lpez-Ribot, J.L., Kirkpatrick, W.R., McAtee, R.K., Santillan,R.A., Martinez, M., Calabrese, D., Sanglard, D. and Patterson , T.F. (2001). Prevalence of resistance mechanisms to azole antifungal agents in Candida albicans isolates displaying high-level fluconazole resistance from HIV-infected patients. *Antimicrobial Agents Chemotherapy*, **45**, 2676-2684.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the USA*, **42**, 43-47.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, **11**, 735-757.

Schellhase, C. (2009). *pendensity: Density Estimation with a Penalized Mixture Approach*. R package version 0.2.

Schön, T., Juréen, P., Giske, C.G., Chryssanthou, E., Sturegård, E., Werngren, J., Kahlmeter, G., Hoffner, S.E. and Ängeby, K.A. (2009). Evaluation of wild-type MIC distributions as a tool for determination of clinical breakpoints for Mycobacterium tuberculosis. *Journal of Antimicrobial Chemotherapy*, **64**, 786-793.

Strasfeld, L. and Chou, S. (2010). Antiviral drug resistance: mechanisms and clinical implications. *Infectious Disease Clinics of North America*, 24, 413-437.

Tenover, F.C. (2006). Mechanisms of antimicrobial resistance in bacteria. *The American Journal of Medicine*, **119**, 3-10.

Turnidge, J., Kahlmeter, G. and Kronvall, G. (2006). Statistical characterisation of bacterial wild-type MIC value distributions and the determination of epidemiological cut-off values. *Clinical Microbiology and Infection*, **12**, 418-425.

Wu. H., Schwarz, C. J., and de With, N. (2008). The analysis of minimum inhibitory concentration (MIC) data for anti-microbial agents from complex experiments. *Submitted to Journal of Applied Statistics*.

Appendix A

B-spline Basis Functions

Kauermann and Schellhase (2009) pay attention to two types of basis density functions. The first one are the well-known Gaussian densities, which do not require further explanation. The second type, the so-called B-splines, are possibly less understood in general. Therefore, this appendix provides some more information on how how these B-spline functions look like, how they are constructed and in what way they can be transformed to represent densities and distributions.

In this regard, Eilers and Marx (1996) give a very brief but gentle and general description of B-splines. In short, a B-spline consists of polynomial pieces that are connected in a special way. For example, a B-spline of degree 1 is depicted in figure A.1 (top row on the left). As can be seen, this spline consists of two linear pieces ($x_1 - x_2$ and $x_2 - x_3$) that are joined at the middle knot. To the left of x_1 and to the right of x_3, this B-spline is zero. By introducing more knots, a larger set of B-splines can be constructed as is shown in figure A.1 (top row on the right). Similarly, the bottom row of figure A.1 (on the left) shows a B-spline of degree two, consisting of three quadratic pieces that are joined at two inner knots. In total, the B-spline is based on four adjacent knots. Next to the fact that the values of the two polynomial pieces match at the joining point, also their first derivatives are equal there. Again here, considering more knots provides a larger set of B-splines. Note that these B-splines overlap each other. More specifically, first-degree B-splines have overlap with two neighbors, whereas second-degree B-splines overlap with four neighbors.

Based on the two examples above, more general properties can be derived that characterize a B-spline of degree q:

- it consists of q+1 polynomial pieces, each of degree q;
- the polynomial pieces join at q inner knots;
- at the joining points, derivatives upto order q-1 are continuous;
- the B-spline is positive on a domain spanned by q+2 knots; everywhere else it is zero;
- except at the boundaries, it overlaps with 2q polynomial pieces of its neighbors;
- at a given x, q+1 B-splines are nonzero.

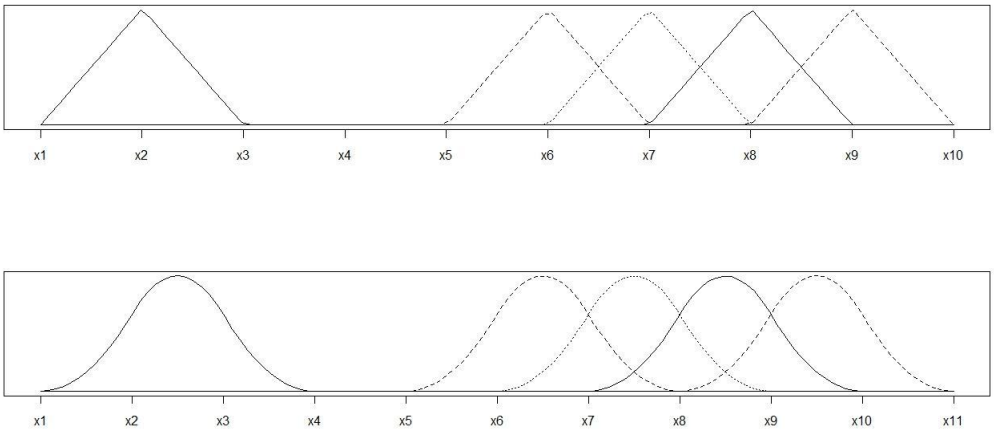

Figure A.1: Graphical representation of B-splines of degree 1 (top row) and degree 2 (bottom row).

Kauermann and Schellhase (2009) prefer to work with the order of the B-spline, following the original definition of de Boor (1977), the order corresponds to the degree+1.

In order to give a formal definition of the B-spline basis functions, let $\tau = \{\tau_i\}_{i=1}^{M}$ denote the sequence of knots that are used in the construction of the spline functions. Let accordingly $B_{i,m}(x)$ denote the ith B-spline basis function of order m for the knot-sequence τ, where τ_i is the left most knot of the B-spline function. These B-spline functions are defined recursively in terms of divided differences as follows:

$$B_{i,1}(x) = \begin{cases} 1 \ldots \text{if } \tau_i \leq x < \tau_{i+1}, \\ 0 \ldots \text{if otherwise,} \end{cases}$$

for i = 1,...,M-1. These are also known as Haar basis functions and can be used to define the B-spline basis functions of higher order m $\leq$ M:

$$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x),$$

for i = 1,...,M-m. The functions specified above do not represent densities yet, since they do not integrate to one. The normalizing constant is equal to $\frac{m}{\tau_{i+m} - \tau_i}$, which is constant in case of equidistant knots. Hence, the B-spline density functions (of order m) that make up the base used in the penalized mixture approach are given by

$$f_{k,bspline}(x) = \frac{m}{\tau_{k+m} - \tau_k} B_{k,m}(x).$$

However, when the sample under investigation is known to be censored, these density functions need to be replaced by their corresponding distribution functions. These are defined by making use of B-spline functions that have one additional order:

$$F_{k,bspline}(x) = \sum_{j \geq k} B_{j,m+1}(x).$$

Appendix B

Application to Real Data

In chapter 5, two of the discussed methods for estimating the parameters of the first component were applied to real AMR data. This appendix presents the output for the remaining two methods, namely the adjusted approach to the method of Turnidge et al. (2006) and the likelihood based approach. With the adjusted version of the method of Turnidge et al. (2006), it is aimed for to get a closer approximation of the true number of isolates. With this purpose, new cumulative counts are considered via pointwise addition of new observations to the original cumulative counts. The results of this procedure are presented in table B.1, where the second column indicates how many observations are added to the previous cumulative counts.

Table B.1: Parameter estimates of the adjusted non-linear least squares regression approach, applied to the ampicillin - E. coli data.

		Number of observations				*Mean*		*Standard deviation*	
Subset	Added	True	Est.	Diff.	Std. Err.	Est.	Std.Err.	Est.	Std.Err.
1	1	12516	12516.01	0.01	57.16	0.21	0.88	0.21	0.87
2	2819	26665	26739.65	74.65	87.45	1.07	0.01	0.75	0.01
3	97	26124	26123.92	-0.08	21.29	1.04	0.00	0.71	0.00

The minimum difference (0.01) is observed for the first fitted subset (MICs ≤ 2 on $\log_2$ scale), resulting into estimates of the mean and standard deviation equal to 0.21 and 0.21 respectively. These values seem to be less plausible when regarding the barplot in figure 5.1. Exploring the output further, it is observed that the second most optimal fit occurs in the third subset. The difference in this subset is only -0.08 and hence also negligibly small. Estimates corresponding to this fit are 1.04 for the mean and 0.71 for the standard deviation, i.e. the same as those from the original method.

Problems occurred with the likelihood based approach, as it failed to converge when more than three MIC categories were included in the fit. The procedure was fit using both the normal as well as gamma truncated distribution functions. Because of the failure of this procedure, the resulting estimates (see table B.2) are not appropriate.

Table B.2: Parameter estimates of the likelihood based method with truncated normal and gamma distribution, applied to the ampicillin - E. coli data.

	Estimates using truncated normal					
	Mean		*Standard deviation*			
Endpoint	Est.	Std.Err.	Est.	Std.Err.	Likelihood	p-value
0	477.90	1106.18	12.85	14.85	-437.67	< 0.0001
1	1.90	0.20	1.00	0.05	-5852.77	< 0.0001
	Estimates using truncated gamma					
	Shape		*Scale*			
Endpoint	Est.	Std.Err.	Est.	Std.Err.	Logikelihood	p-value
1.0	4.18	0.13	898806.07	209114692.99	-437.61	< 0.0001
2.0	5.66	0.29	0.37	0.04	-5846.64	< 0.0001

Appendix C

R Code for Proposed Methods

In this appendix, the R code can be found for the methods that were introduced in chapter 3. The same code was applied in the simulation study that was carried out at the end of that same chapter, with seeds equal to 1,...,200. The example mixtures were created using the following function:

```
Mixture<-function(n=1000,p1=0.5,p2=0.5,mu1=-2,mu2=3,mu3=5,sd1=1,sd2=1.5,
sd3=1.5)
{
set.seed(1988)
y1<-rnorm(n*p1,mu1,sd1)
y2<-rnorm(n*(1-p1)*p2,mu2,sd2)
y3<-rnorm(n*(1-p1)*(1-p2),mu3,sd3)
y4<-c(y1,y2,y3)
data<-as.data.frame(cbind(y4,c(rep(1,length(y1)),rep(2,length(y2)),
rep(3,length(y3)))))
names(data)<-c("yvalue","class")
x2<-vector()
x1<-sort(data$yvalue)
y.min<-min(data$yvalue)
y.max<-max(data$yvalue)
groups<-seq(-4,ceiling(y.max),1)
for(i in 1:length(x1)){
j<-1
calc<-TRUE
while(calc){
if(x1[i]<=groups[j] ) {
x2<-c(x2,groups[j])
calc<-FALSE
}
else j<-j+1
}
}
# How many observations per group
counts<-vector()
for(i in 1:length(groups)){
```

```
counts<-c(counts,length(x2[x2==groups[i]]))
}
cum.counts<-vector()
for(i in 1:length(counts)){
cum.counts<-c(cum.counts,sum(counts[1:i]))
}
data.cumul<-as.data.frame(cbind(groups,cum.counts,counts))
names(data.cumul)<-c("Group","Counts.Cumul","Counts")
return(list("data"=data.cumul,"sample"=x1,"censored.sample"=x2))
}

set1<-Mixture(1000,0.5,0.5,-2,3,5,1,1.5,1.5)
```

Original method of Turnidge et al. (2006)

```
estimates.original<-matrix(0,1,ncol=9)
for(i in 5:length(set1$data$Group)){
help.matrix<-matrix(nrow=1,ncol=9)
fit<-nls(Counts.Cumul~N*pnorm(Group,mu,sigma^2),data=set1$data[1:i,],
start=list(N=1000,mu=-3,sigma=1))
help.matrix[1,1]=i
help.matrix[1,2]=round(set1$data$Counts.Cumul[i],2)
help.matrix[1,3]=round(coef(fit)[1],2)
help.matrix[1,4]=round(-set1$data$Counts.Cumul[i]+coef(fit)[1],2)
help.matrix[1,5]=round(summary(fit)$parameters[1,2],2)
help.matrix[1,6]=round(coef(fit)[2],2)
help.matrix[1,7]=round(summary(fit)$parameters[2,2],2)
help.matrix[1,8]=round(coef(fit)[3]^2,2)
help.matrix[1,9]=round(2*coef(fit)[3]*summary(fit)$parameters[3,2],2)
estimates.original<-rbind(estimates.original,help.matrix)
}
solution.original<-estimates.original[abs(estimates.original[,4])==
min(abs(estimates.original[,4])),]
```

Adjustment to method of Turnidge et al. (2006)

```
estimates.help1<-matrix(0,1,9)
estimates.minimum.per.group1<-matrix(0,1,9)
for(i in 5:length(set1$data$Group)){
estimates.improved1<-matrix(nrow=set1$data$Counts[i],ncol=9)
for (j in 1:set1$data$Counts[i]){
use.set<-set1
Cumul<-use.set$data$Counts.Cumul[i-1]+j
use.set$data$Counts.Cumul[i]<-Cumul
fit<-nls(Counts.Cumul~N*pnorm(Group,mu,sigma^2),data=use.set$data[1:i,],
start=list(N=1000,mu=-3,sigma=1),control = list(maxiter = 10000))
estimates.improved1[j,1]=j
estimates.improved1[j,2]=round(use.set$data$Counts.Cumul[i],2)
estimates.improved1[j,3]=round(coef(fit)[1],2)
```

```
estimates.improved1[j,4]=round(-use.set$data$Counts.Cumul[i]+coef(fit)[1],2)
estimates.improved1[j,5]=round(summary(fit)$parameters[1,2],2)
estimates.improved1[j,6]=round(coef(fit)[2],2)
estimates.improved1[j,7]=round(summary(fit)$parameters[2,2],2)
estimates.improved1[j,8]=round(coef(fit)[3]^2,2)
estimates.improved1[j,9]=round(2*coef(fit)[3]*summary(fit)$parameters[3,2],2)
}
estimates.help1<-rbind(estimates.help1,estimates.improved1)
estimates.minimum.per.group1<-rbind(estimates.minimum.per.group1,
estimates.improved1[abs(estimates.improved1[,4])==
min(abs(estimates.improved1[,4])),])
}
estimates.new1<-estimates.help1[-1,]
per.group.minimum1<-estimates.minimum.per.group1[-1,]
solution.new1<-estimates.new1[abs(estimates.new1[,4])==
min(abs(estimates.new1[,4])),]
```

Likelihood Based Approach

```
Minusloglik<-function(parms=c(-1.95,1.01),endpoint=0,
data=full.data$censored.sample){
y=data
nn <- matrix(1:length(which(y<=endpoint)))
z<-seq(min(y),max(y),1)
component1<- apply(nn, 1, function(i,y){
if(y[i]==min(y))pnorm(y[i],parms[1],parms[2]^2)
else pnorm(y[i],parms[1],parms[2]^2)-pnorm(z[y[i]==z+1],
parms[1],parms[2]^2)}, y)
component1<-component1/pnorm(endpoint,parms[1],parms[2]^2)
loglik<-sum(log(component1))
minusloglik<- -loglik
}
full.data<-set1
endpoints<-c(-2,-1,0,1,2,3,4,5,6,7)
estimates.final<-matrix(0,1,7)
for(i in 1:length(endpoints)){
estimates<-matrix(0,1,7)
test.vector<-vector()
fit<-nlm(Minusloglik,c(-3,2),hessian=TRUE,endpoint=endpoints[i],
data=full.data$censored.sample)
estimates[1,1]<-endpoints[i]
estimates[1,2]<-round(fit$estimate[1],2)
estimates[1,3]<-round(sqrt(diag(solve(fit$hessian)))[1],2)
estimates[1,4]<-round(fit$estimate[2]^2,2)
estimates[1,5]<-round(sqrt(4*estimates[1,4]*diag(solve(fit$hessian)))[2],2)
estimates[1,6]<- round(-fit$minimum,2)
use.groups<-full.data$data$Group[full.data$data$Group<=endpoints[i]]
for(j in 1:length(use.groups)){
if(j==1){
```

```
observed<-full.data$data$Counts[full.data$data$Group==use.groups[j]]
expected<-pnorm(use.groups[j],estimates[1,2], estimates[1,4])/
pnorm(use.groups[length(use.groups)],estimates[1,2],
estimates[1,4])*full.data$data$Counts.Cumul[full.data$data$Group
==use.groups[length(use.groups)]]
test.vector<-c(test.vector,(observed-expected)^2/expected)
}
else{
observed<-full.data$data$Counts[full.data$data$Group==use.groups[j]]
expected<-((pnorm(use.groups[j],estimates[1,2], estimates[1,4])-
pnorm(use.groups[j-1],estimates[1,2], estimates[1,4]))/
pnorm(use.groups[length(use.groups)],estimates[1,2],
estimates[1,4]))*full.data$data$Counts.Cumul[full.data$data$Group
==use.groups[length(use.groups)]]
test.vector<-c(test.vector,(observed-expected)^2/expected)
}
}
estimates[1,7]<-round(1-pchisq(sum(test.vector),length(use.groups)-3),4)
estimates.final<-rbind(estimates.final,estimates)
}
```

Multinomial Based Approach

```
Loglik.multi<-function(parms,y=data){
loglikelihood=0
parms.new<-c(parms,0)
probs<-exp(parms.new)/sum(exp(parms.new))
for(i in 1:length(y)) loglikelihood<-loglikelihood+y[i]*log(probs[i])
return(-loglikelihood)
}

Loglik<-function(parms,y=data,nr.groups=3){
loglikelihood=0
z<-unique(y[,1])
p<-vector()
for(i in 1:nr.groups){
if (i==1) p<-c(p,pnorm(min(z),parms[2],parms[3]^2))
else p<-c(p,pnorm(y[i,1],parms[2],parms[3]^2)-pnorm(z[y[i,1]==(z+1)],
parms[2],parms[3]^2))
}
p<-parms[1]*p
parms.new<-c(p,parms[-c(1,2,3)],0)
probs<-c(parms.new[1:length(p)],(1-sum(p))*exp(parms.new[(length(p)+1):
length(parms.new)])/sum(exp(parms.new[(length(p)+1):length(parms.new)])))
for(i in 1:nrow(y)) loglikelihood<-loglikelihood+y[i,2]*log(probs[i])
return(-loglikelihood)
}

data<-set1$data
```

```
output<-matrix(0,1,5)
for(i in 1:(nrow(data))){
if (i<3) test<-nlm(Loglik.multi,rep(1/(nrow(data)-1),nrow(data)-1),y=data[,2])
else{
if (i<nrow(data)) test<-nlminb(c(0.5,0,1,rep(1/(nrow(data)-i-1),
(nrow(data)-i-1))),Loglik,y=data,nr.groups=i,lower=c(0,-Inf,0,
rep(-Inf,nrow(data)-i-1)),upper=c(1,Inf,Inf,rep(Inf,nrow(data)-i-1)),
control=list(iter.max=10000))
else test<-nlminb(c(0.5,1,1),Loglik,y=data,nr.groups=i,lower=c(0,-Inf,0),
upper=c(1,Inf,Inf),control=list(iter.max=10000))
}
if(i<3) AIC<-2*test$minimum+2*length(test$estimate)
else AIC<-2*test$objective+2*length(test$par)
if(i<3){
parms.new<-c(test$estimate,0)
probs<-exp(parms.new)/sum(exp(parms.new))
}
else{
z<-unique(y[,1]);nr.groups=i;p<-vector()
for(i in 1:nr.groups){
if (i==1) p<-c(p,test$par[1]*pnorm(min(z),test$par[2],test$par[3]^2))
else p<-c(p,test$par[1]*(pnorm(y[i,1],test$par[2],test$par[3]^2)-
pnorm(z[y[i,1]==(z+1)],test$par[2],test$par[3]^2)))
}
parms.new<-c(p,test$par[-c(1,2,3)],0)
probs<-c(parms.new[1:length(p)],(1-sum(p))*exp(parms.new[(length(p)+1):
length(parms.new)])/sum(exp(parms.new[(length(p)+1):length(parms.new)])))
}
if(i<3) output<-rbind(output,c(AIC,999,999,999,test$minimum))}
else{
if(i<nrow(data)) output<-rbind(output,c(AIC,test$par[1],test$par[2],
test$par[3]^2,test$objective))
else output<-rbind(output,c(AIC,test$par[1],test$par[2],
test$par[3]^2,test$objective))
}
}
pvalues<-vector()
for(i in 4:nrow(output[-1,])){
pvalues<-c(pvalues,1-pchisq(2*(output[-1,][i,5]-output[-1,][1,5]),i-3))
}
min<-min(output[c(5:nrow(output)),1])
diff<-output[c(5:nrow(output)),1]-min
weights<-exp(-1/2*diff)/sum(exp(-1/2*diff))
sum(weights*output[c(5:nrow(output)),2])
sum(weights*output[c(5:nrow(output)),3])
sum(weights*output[c(5:nrow(output)),4])
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Development of a new semi-parametric mixture model for interval censored data, with applications in antimicrobial resistance

Richting: **master of Statistics-Epidemiology & Public Health Methodology**
Jaar: **2011**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Jaspers, Stijn

Datum: **12/09/2011**