Biostatistics (2015), *pp.* 1–14
doi:10.1093/biostatistics/kxv030

Application of the Vertex Exchange Method to estimate a semi-parametric mixture model for the MIC density of *Escherichia coli* isolates tested for susceptibility against ampicillin

STIJN JASPERS*

Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Diepenbeek 3590, Belgium

stijn.jaspers@uhasselt.be

GEERT VERBEKE

Interuniversity Institute for Biostatistics and Statistical Bioinformatics, KU Leuven, Leuven B-3000, Belgium

DANKMAR BÖHNING

Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton SO17 I BJ, UK

MARC AERTS

Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Diepenbeek 3590, Belgium

SUMMARY

In the last decades, considerable attention has been paid to the collection of antimicrobial resistance data, with the aim of monitoring non-wild-type isolates. This monitoring is performed based on minimum inhibition concentration (MIC) values, which are collected through dilution experiments. We present a semi-parametric mixture model to estimate the entire MIC density on the continuous scale. The parametric first component is extended with a non-parametric second component and a new back-fitting algorithm, based on the Vertex Exchange Method, is proposed. Our data example shows how to estimate the MIC density for *Escherichia coli* tested for ampicillin and how to use this estimate for model-based classification. A simulation study was performed, showing the promising behavior of the new method, both in terms of density estimation as well as classification.

Keywords: Antimicrobial resistance; Censoring; Model-based classification; Semi-parametric; Vertex Exchange Method.

*To whom correspondence should be addressed.

1. INTRODUCTION

Finite mixture models are very popular in various scientific areas since they provide a powerful tool to model unobserved population heterogeneities and latent structures. An extensive overview of the theory of these models can be found in McLachlan and Peel (2000), while Titterington *and others* (1985) provide a wide range of data applications. The application of interest in this paper concerns one of the main public health burdens of the last decades, namely the field of antimicrobial resistance (AMR) and antimicrobial susceptibility testing. In order to study and monitor the emergence of isolates with reduced susceptibility against antimicrobials, minimum inhibitory concentrations (MIC) are collected. MIC is the lowest concentration of an antimicrobial that inhibits the visible growth of a microorganism after overnight incubation. The MIC is commonly measured via an agar or broth dilution method, in which a standardized amount of the isolate is exposed to successive 2-fold concentrations of the antimicrobial (i.e. 0.25, 0.5, 1, 2 mg/L, . . .). The MIC is defined as the lowest concentration with no visible growth after a prescribed incubation period. Consider, for example, a bacterial isolate that is subjected to an antimicrobial at concentrations 0.5, 1, 2, and 4 mg/L. In case the isolate shows inhibition of growth at values of 2 and 4 mg/L, but growth at lower values, the reported MIC value is equal to 2 mg/L. However, the true inhibition occurs between the concentrations of 1 and 2 mg/L, so the obtained MIC value is interval censored. See, for example, Andrews (2001) and Wiegand *and others* (2008) for a detailed description of how to obtain these MIC values.

The common way of representing a dilution experiment is by drawing an MIC distribution, i.e. the frequency of occurrence of each given MIC plotted against the MIC value. Figure 1(a) shows an MIC distribution determined for 1890 isolates of *Escherichia coli* tested for susceptibility against ampicillin in the year 2010. For a given bacterial species, the multi-modal pattern of the MIC distribution can usually enable the separation of the wild-type population of microorganisms from those non-wild-type populations which show a reduced susceptibility to the antimicrobial in question. The wild-type susceptible population, typically located on the left of the MIC distribution, is assumed to have no acquired or mutational resistance. It commonly shows a uni-modal distribution reflecting a slight biological variability around a mode whose value will not be altered by changing circumstances over time. The second component, representing the non-wild-type isolates, is often multi-modal since it represents different non-wild-type sub-populations which are characterized by different degrees of reduced susceptibility conferred by different mechanisms. However, the number of these non-wild-type sub-populations as well as their distributions are unknown *a priori*. This becomes apparent when regarding Figure 1(b), which also shows an MIC distribution for the combination *E. coli*–ampicillin. It can be noted that while the first component is comparable to that of Figure 1(a), the non-wild-type component of the distribution shows an additional mode and a higher spread.

Let the univariate random variable X represent the MIC value with probability density function $f(x)$. In our context, a two-component mixture

$$f(x) = \pi f_1(x; \boldsymbol{\theta}_1) + (1 - \pi) f_2(x; \boldsymbol{\theta}_2) \tag{1.1}$$

is assumed, in which f_1 and f_2, respectively, represent the wild-type and non-wild-type component of the MIC distribution and the prevalence of wild-type isolates is denoted by π. Despite the importance of analyzing AMR data, the statistical literature regarding this topic is rather limited. It is current practice to classify an isolate into the wild-type or non-wild-type sub-population based on an epidemiological cut-off (ECOFF) value, defined as the upper limit of the wild-type distribution. According to the guidelines of the European Committee on Antimicrobial Susceptibility Testing (EUCAST), the ECOFF can be determined based on visual inspection of the histogram resulting from the dilution experiment or, alternatively, it can be statistically calculated using the approach of Turnidge *and others* (2006). The latter approach aims at providing an estimate for the wild-type density function (f_1), from which the ECOFF is derived as the 99.9th percentile. In a similar fashion, Jaspers *and others* (2014) also adopt a local view, focusing on the wild-type first component only. They proposed an improved likelihood-based procedure, called the

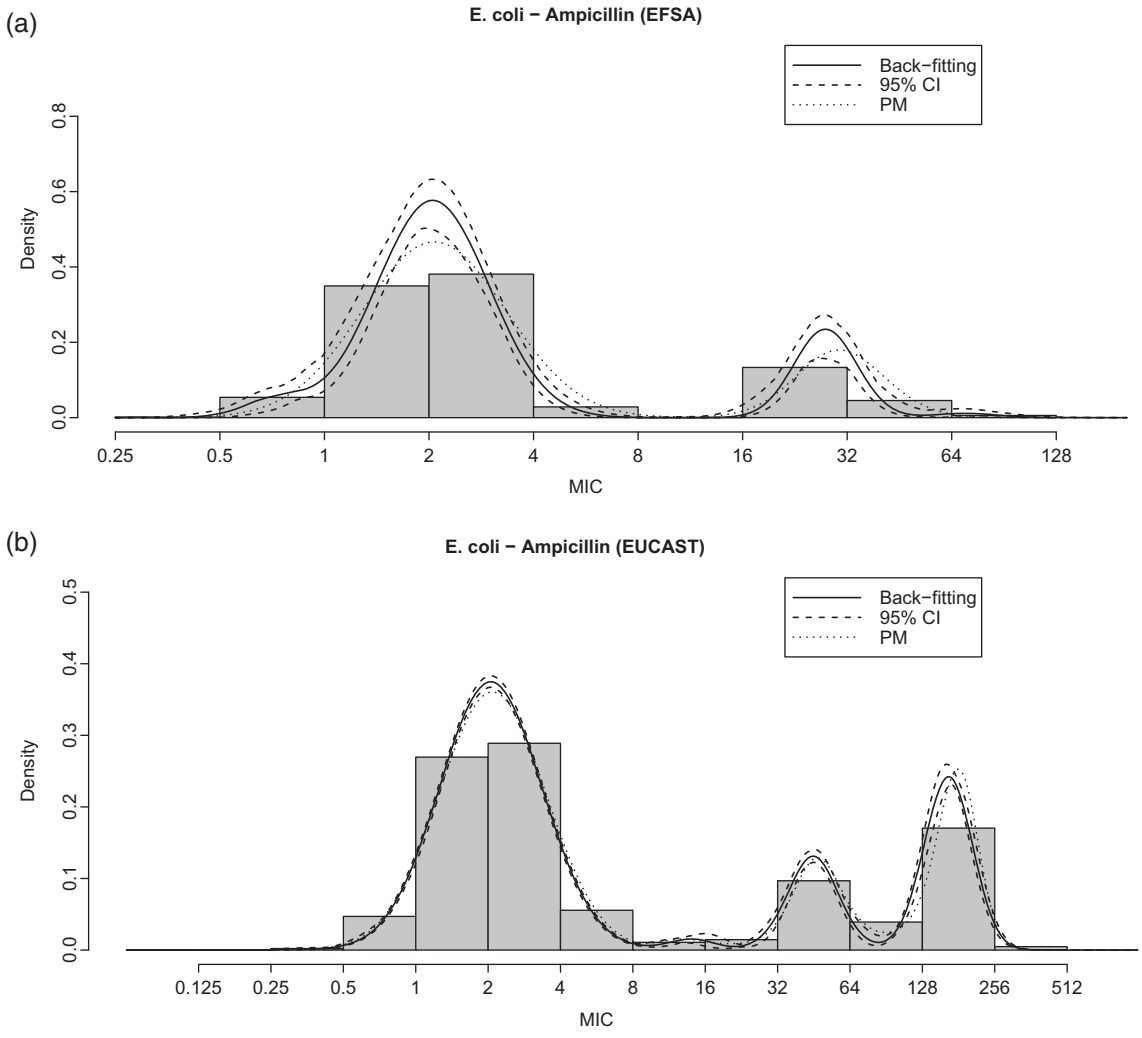

Fig. 1. Histogram of *E. coli* isolates tested for susceptibility against ampicillin—sources EFSA (a) and EUCAST (b). Overlaid are the estimated density using the back-fitting algorithm (solid) with its simultaneous 95% bootstrap confidence limits (dashed) and the estimate resulting from the PM approach (dotted).

multinomial based method (MBM) to identify the most suitable distribution of the first component and to estimate its parameters.

Model-based classification is a valuable alternative for determining the sub-population of a specific isolate. With this technique, isolates are classified to the wild-type sub-population when the posterior probability

$$\frac{\pi f_1(x;\boldsymbol{\theta}_1)}{\pi f_1(x;\boldsymbol{\theta}_1) + (1-\pi) f_2(x;\boldsymbol{\theta}_2)} \tag{1.2}$$

is larger than 0.5. It is clear that this option requires an estimate for the entire mixture density f. Craig (2000) suggested to approximate the entire density f in (1.1) by a mixture of Gaussian density functions.

This approach was followed by Annis and Craig (2005), who assumed two fixed components, representing the wild-type and non-wild-type sub-populations. However, no *a priori* information is available on the number of components for the non-wild-type density, nor on the shape of these component density functions. Therefore, a non-parametric second component seems more appealing. Jaspers *and others* (2014) provide a two-stage semi-parametric mixture model to estimate the mixture density of interest. The first stage determines the estimates of the first component using the MBM. Fixing the so-obtained estimates as being the true parameters of the wild-type component, i.e. $\boldsymbol{\theta}_1$, the density of the second component is determined using an extended version of the penalized mixture (PM) approach by Schellhase and Kauermann (2012). Nevertheless, a drawback of this two-stage procedure is that the parameters of the first component are not updated, but kept fixed at the initial estimates provided by the MBM. This provides inadequate estimates of the standard errors. In addition, there seemed to be some kind of discontinuity in the region of overlap between the first and second component, resulting from the used two-stage approach. Hence, although attempts have been made, there still is no satisfactory methodology which is able to estimate the mixture distribution of interest, taking into account the underlying complexities.

The major aim of this paper is to present a new method which is able to jointly estimate the parametric first component, f_1, and the non-parametric second component, f_2. In this respect, we will consider a non-parametric maximum likelihood estimate (NPMLE) for the second component, using the Vertex Exchange Method (VEM), introduced by Böhning (1986). This directional derivative-based algorithm will be adjusted in such a way that it can cope with the censored nature of the MIC data. Although it was already used for survival data, the presented approach is a new application regarding the VEM in a censored setting. The new approach can play a profound role in the monitoring of AMR data, with emphasis on prevalence estimation and model-based classification. In addition, an adequate description of the distribution of the non-wild-type component is necessary when the goal is to discover shifts over time.

Some data examples are presented in Section 2. In Section 3, we will give a small review of the VEM and present the final estimation procedure of the semi-parametric mixture model. An application of the new method can be found in Section 4 and a simulation study shows its performance in Section 5. Finally, a discussion ends the paper in Section 6.

2. DATA

Escherichia coli is a Gramm-negative bacterium that is usually a commensal of humans and animals. Nevertheless, pathogenic variants can cause intestinal and extra-intestinal infections, including urinal tract infections and meningitis. The preferred treatment depends on the nature of the infection and antimicrobial treatment is not recommended for every type of infection (Igarashi *and others*, 1999). Several studies have shown that resistance in *E. coli* isolates is relatively high and has been emerging over the last decades (Kronvall, 2010; Tadesse *and others*, 2012). In this report, we will focus on the susceptibility of *E. coli* against ampicillin, with the major aim of estimating the MIC value density.

Across Europe, several institutions are concerned with collecting data on AMR and identifying possible threats to the human health. On a yearly basis, the European Centre for Disease Prevention and Control and the European Food Safety Authority (EFSA) jointly prepare an annual European Union Summary Report on AMR in zoonotic and indicator bacteria from humans, animals, and food. Since 2010, data are collected from EU member states on an isolate-based level. For the purpose of this modeling study, an exemplary MIC distribution summarizing the results of ampicillin susceptibility testing of indicator *E. coli* isolates in 2010 has been provided by EFSA. Four member states provided information regarding this antibiotic–bacterium combination, resulting into a subset of 1890 isolates. A graphical representation of the MIC value distribution can be found in Figure 1(a).

Another important organization within the field of AMR is the EUCAST. This organization is mainly concerned with breakpoints and technical aspects of phenotypic *in vitro* antimicrobial susceptibility testing. Most antimicrobial MIC breakpoints (e.g. ECOFF values) in Europe have been harmonized by EUCAST. An interesting collection of MIC distributions can be found on their website. These distributions are based on collated data from a total of almost 20 000 MIC distributions from worldwide sources. For comparison purposes, the same antibiotic–bacterium combination has been selected for our analysis: ampicillin–*E. coli*. The resulting MIC distribution consists of 39 220 isolates that were obtained from 48 distinct sources. The observed MIC values ranged from 0.125 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 histogram in Figure 1(b). Two large peaks are clearly visible at the values of 2 and 4 mg/L, probably representing the center of the wild-type component. Towards the larger MIC values, two smaller peaks are located at the values of 64 and 256 mg/L, which could represent distinct strains of the non-wild-type isolates.

3. Methodology

With X denoting the MIC value of a certain isolate, the probability density function of interest is represented by $f(x;\boldsymbol{\Phi}) = \pi f_1(x;\boldsymbol{\theta}_1) + (1-\pi) f_2(x)$. As argued above, the first component, f_1, can be assumed to be of a fixed parametric form depending on certain parameters $\boldsymbol{\theta}_1$. In this paper, the primary focus for the first component will be on the log-normal assumption, with $\boldsymbol{\theta}_1$ representing the mean and standard deviation of the first component. The proportion of isolates corresponding to this wild-type component is reflected in the prevalence parameter π. Because less information is present on the second component, a non-parametric estimate will be considered. Bordes and Vandekerkhove (2010), Hohmann and Holzmann (2013), Xiang *and others* (2014), and Ma and Yao (2015) all consider a similar model in which the second component is assumed to be symmetric but unknown. Nevertheless, in the AMR setting, this symmetry does not apply. We therefore consider a basis-function representation of f_2, given by $f_2(x;\boldsymbol{\Psi}) = \sum_{c=1}^{k} \pi_c g(x;\lambda_c)$, where $\boldsymbol{\Psi}$ denotes the mixing distribution in which weights $\pi_1, \ldots, \pi_k$ are given to support points $\lambda_1, \ldots, \lambda_k$. For identifiability reasons, some assumptions on f_2 are required (Bordes *and others*, 2006; Ma and Yao, 2015). In this paper, g is assumed to be log-normally distributed and the support points λ_c correspond to the means of these log-normal densities. In order to obtain enough flexibility to estimate the unknown second component, a generous number of log-normal densities will be employed. These densities are located at fixed equidistant support points which are assumed to cover the range of observed MIC values. However, since we do not want to overfit the data, the standard deviations of these basis functions will be kept fixed at a predetermined common value. A grid search will be performed to determine the right amount of smoothing based on the Akaike Information Criterion (Akaike, 1974): $\mathrm{AIC} = -2l + 2\mathrm{K}$, with K the number of parameters. In conclusion, the density of the MIC value X is given by $f(x;\boldsymbol{\Phi}) = \pi f_1(x;\boldsymbol{\theta}_1) + (1-\pi)\sum_{c=1}^{k} \pi_c g(x;\lambda_c)$, where $\boldsymbol{\Phi} = (\pi, \boldsymbol{\theta}_1, \boldsymbol{\Psi})$.

The typical application of maximum likelihood estimation concerns a random sample of individual data points of X. In practice, however, it frequently occurs that the collected data are only present in grouped form, meaning that the frequencies of observations in fixed intervals are reported. This is also the case for MIC data, a direct consequence of the collection using dilution experiments. Suppose therefore that the sample space is partitioned into m mutually exclusive intervals, for which the boundaries are denoted by $a_0, a_1, \ldots, a_m$. The reported data are the number of isolates n_i falling into the interval $[a_{i-1}, a_i]_{i=1,\ldots,m}$, where $n = \sum_{i=1}^{m} n_i$ corresponds to the total number of independently collected MIC values. In what follows, $y_i = (a_i, n_i)$ denotes the observed grouped data. In order to construct the log-likelihood to be optimized, the distribution function $F(x)$ of the MIC values is required. In this regard, let $F_1(x;\boldsymbol{\theta}_1) = \int_{-\infty}^{x} f_1(t;\boldsymbol{\theta}_1)\,\mathrm{d}t$ and $F_2(x,\boldsymbol{\Psi}) = \sum_{c=1}^{k} \pi_c G(x;\lambda_c)$ with $G(x;\lambda_c) = \int_{-\infty}^{x} g(t;\lambda_c)\,\mathrm{d}t$, for $c = 1, \ldots, k$.

Then, $F(x) = \pi F_1(x; \boldsymbol{\theta}_1) + (1-\pi)F_2(x; \boldsymbol{\Psi})$. With the m-dimensional vectors $\mathbf{A}$ and $\mathbf{N}$ represented by $\mathbf{A} = \pi \mathbf{P}_1 + (1-\pi)\mathbf{P}_2$, with $\mathbf{P}_j = [F_j(a_i) - F_j(a_{i-1})]_{i=1,\ldots,m}$ for $j = 1, 2$ and $\mathbf{N} = [n_i]_{i=1,\ldots,m}$, the final log-likelihood is given by

$$l(\pi, \boldsymbol{\theta}_1, \boldsymbol{\Psi}|y) = \sum_{i=1}^{m} n_i \log\{F(a_i) - F(a_{i-1})\} = \mathbf{N}\log\{\mathbf{A}^{\mathrm{T}}\}. \tag{3.1}$$

3.1 *The VEM algorithm*

The VEM is a directional derivative-based method, used to obtain the NPMLE of a mixing distribution. The main idea of the method is to search in each iteration for the direction that maximizes the log-likelihood increase $\Delta = l(\boldsymbol{\Psi}^{(it+1)}) - l(\boldsymbol{\Psi}^{(it)})$, where $l(\boldsymbol{\Psi})$ is a shorthand notation for $l(\boldsymbol{\Psi}|\pi, \boldsymbol{\theta}_1)$, while $\boldsymbol{\Psi}^{(it)}$ and $\boldsymbol{\Psi}^{(it+1)}$ refer to the current and updated estimates of $\boldsymbol{\Psi}$ at iterations (it) and $(it+1)$, respectively. Once the optimal direction is found, weights are exchanged between the support points that contribute the most and the least to this difference. These points are identified based on the definition of the directional derivative of the log-likelihood from one distribution $\boldsymbol{\Psi}^{(it)}$ to the other $\boldsymbol{\Psi}^{(it+1)}$. When $\boldsymbol{\Psi}^{(it+1)}$ is degenerate at λ_c (i.e. $\pi_c = 1$) then $\boldsymbol{\Psi}^{(it+1)} = \boldsymbol{\Psi}_{\lambda_c}$. In particular, the directional derivative $D(\boldsymbol{\Psi}^{(it)}, \boldsymbol{\Psi}_{\lambda_c})$ of $l(\boldsymbol{\Psi})$ at $\boldsymbol{\Psi}^{(it)}$ in the direction of $\boldsymbol{\Psi}_{\lambda_c}$ for interval-censored data is defined as

$$\begin{aligned} D(\boldsymbol{\Psi}^{(it)}, \boldsymbol{\Psi}_{\lambda_c}) &= \lim_{s \to 0+} \frac{l((1-s)\boldsymbol{\Psi}^{(it)} + s\boldsymbol{\Psi}_{\lambda_c}) - l(\boldsymbol{\Psi}^{(it)})}{s} \\ &= \left.\frac{\partial l(\boldsymbol{\theta}_1, (1-s)\boldsymbol{\Psi}^{(it)} + s\boldsymbol{\Psi}_{\lambda_c})}{\partial s}\right|_{s=0} \\ &= \sum_{i=1}^{m} \left[n_i \frac{F(a_i; \lambda_c) - F(a_{i-1}; \lambda_c)}{F(a_i; \boldsymbol{\Psi}^{(it)}) - F(a_{i-1}; \boldsymbol{\Psi}^{(it)})} \right] - n. \end{aligned} \tag{3.2}$$

A single step of the VEM consists of finding the directions that, respectively, minimize and maximize this directional derivative. More specifically, after evaluating the directional derivative for each of the k support points λ_c, we can identify $\lambda^- = \mathrm{argmin}_{\lambda_c} D(\hat{\boldsymbol{\Psi}}^{(it)}, \boldsymbol{\Psi}_{\lambda_c})$ and $\lambda^+ = \mathrm{argmax}_{\lambda_c} D(\hat{\boldsymbol{\Psi}}^{(it)}, \boldsymbol{\Psi}_{\lambda_c})$ and update their weights as $\hat{\pi}_{\lambda^-}^{(it+1)} = (1 - s^\star)\hat{\pi}_{\lambda^-}^{(it)}$ and $\hat{\pi}_{\lambda^+}^{(it+1)} = s^\star \hat{\pi}_{\lambda^-}^{(it)} + \hat{\pi}_{\lambda^+}^{(it)}$ with the step length $s^\star \in [0, 1]$ defined as $s^\star = \mathrm{argmax}_s [l(\boldsymbol{\Psi}^{(it+1)}(s)) - l(\boldsymbol{\Psi}^{(it)})]$. In order to obtain a complete NPMLE, the EM algorithm is required to refine the location of the support points, i.e. the λ_c's. In this regard, one can follow the approach by McLachlan and Jones (1988). This is especially of importance when interest is in calculating standard errors. Indeed, it was noted that the improvement in the obtained point estimate after performing an additional EM step is rather limited. Therefore, when interest is only in point estimation, one could restrict to the VEM only (Tsonaka *and others*, 2009).

3.2 *Back-fitting algorithm*

We are required to estimate two sets of parameters. First of all, there are the parameters inherent to the parametric distribution assumed for the first component, i.e. $\boldsymbol{\theta}_1$. In addition, we also need to estimate the weights attached to the distinct components, i.e. $\tilde{\boldsymbol{\pi}} = (\pi, \pi_1, \ldots, \pi_k)$. In this respect, we adopt a back-fitting algorithm, based on the work of Tsonaka *and others* (2009). The grid of basis densities consists of log-normal density functions, with their support points placed at equidistant values (on the $\log_2$-scale) that cover the observed range of MIC values. A dense grid was found to be most optimal, meaning that in the following applications the distance between two support points, i.e. $\lambda_j - \lambda_{j-1}$, will equal 0.1. Regarding

the choice of the standard deviation, we adopt a model-based approach in which the procedure is repeated for several fixed values for the standard deviation and the most optimal value is determined with the AIC criterion. Consecutively, the grid of basis functions for the second component is extended with a component density for the first component, which is placed at the initial estimate according to the MBM, i.e. at $\hat{\boldsymbol{\theta}}_1 = (\hat{\mu}^{\text{MBM}}, \hat{\sigma}^{\text{MBM}})$. The weights attached to these basis components are determined using the VEM algorithm, while the parameters of the first component are updated using maximum likelihood. For notational simplicity, all weight parameters are assumed to be within $\boldsymbol{\Psi}$. The procedure is detailed upon in the following points:

1. *Obtaining initial values*. Regarding the first component parameters, we can obtain initial values using the MBM as described in Jaspers *and others* (2014). This procedure provides an initial estimate for the prevalence parameter π as well. In order to obtain initial values for the other weights, the VEM algorithm is applied until the maximal value of the directional derivative in each direction is smaller than $\epsilon_1 = 1\text{e} - 3$ or terminated in case the step size $s^\star$ is reduced below $1\text{e}{-}7$. In this initial procedure, the values of π and $\boldsymbol{\theta}_1$ are kept fixed. At the start of the VEM, a zero weight is attached to the support points which are located below the initial value of the first component mean.
2. *Updating of the weights*. For fixed $\hat{\boldsymbol{\theta}}_1^{(it)}$, we perform one step of the VEM algorithm to obtain $\hat{\boldsymbol{\Psi}}^{(it+1)}$, which contains the weights for all the components. Note that compared with the procedure for obtaining the initial estimates, we also allow π to be updated.
3. *Updating of the first component parameters*. Keeping the weights fixed at their estimate from step (1), i.e. $\hat{\boldsymbol{\Psi}}^{(it+1)}$, we update the parameters of the first component by maximizing $l(\boldsymbol{\theta}_1 | \hat{\boldsymbol{\Psi}}^{(it+1)})$, to obtain $\hat{\boldsymbol{\theta}}_1^{(it+1)}$.
4. *Convergence*. The steps described in (1) and (2) are repeated until $l(\hat{\boldsymbol{\theta}}_1^{(it)}, \hat{\boldsymbol{\Psi}}^{(it)}) - l(\hat{\boldsymbol{\theta}}_1^{(it-1)}, \hat{\boldsymbol{\Psi}}^{(it-1)}) < \epsilon_2(1 + |l(\hat{\boldsymbol{\theta}}_1^{(it-1)}, \hat{\boldsymbol{\Psi}}^{(it-1)})|)$, with $\epsilon_2 = 1\text{e} - 8$.

Applying the above back-fitting algorithm, it was noted that upon convergence, the weight of the first component π was occasionally decreased to zero and replaced with several non-parametric components. In order to resolve this apparent identifiability problem, we impose a little penalty on the standard deviation of the first component. More specifically, prior knowledge on the parameter is present from the MBM and we state that the final estimate cannot differ too much from this initial estimate by subtracting, from the likelihood in (3.1), the following penalty term: $\tau * (\sigma - \hat{\sigma}^{\text{MBM}})^2$. A large value for τ implies a great prior belief in the initial estimate and does not allow the new estimate to differ from that initial estimate. On the other hand, a small value for τ still allows the updated estimate to deviate in some extent from the initial value. Based on the performed simulation studies and the considered data examples, $\tau = 2$ was found to be an acceptable penalty. In addition, we determine AIC values at every iteration step of the algorithm: $\text{AIC}^{(it)} = -2l(\hat{\boldsymbol{\theta}}_1^{(it)}, \hat{\boldsymbol{\Psi}}^{(it)}) + 2K$. The number of parameters K corresponds to the number of non-zero weights, added with the number of parameters of the first component. In this way, models with more components of the non-parametric second component are penalized more heavily.

3.3 *Estimation of standard errors*

Consider the observed Hessian of $l(\boldsymbol{\theta}_1, \boldsymbol{\Psi})$, denoted by $H = \begin{smallmatrix} H_{\boldsymbol{\theta}_1\boldsymbol{\theta}_1} & H_{\boldsymbol{\theta}_1\tilde{\pi}} \\ H_{\tilde{\pi}\boldsymbol{\theta}_1} & H_{\tilde{\pi}\tilde{\pi}} \end{smallmatrix}$, where $H_{\boldsymbol{\theta}_1\boldsymbol{\theta}_1} = \partial^2 l(\boldsymbol{\theta}_1, \boldsymbol{\Psi}) / \partial\boldsymbol{\theta}_1 \partial\boldsymbol{\theta}_1$ and analogous definitions for $H_{\boldsymbol{\theta}_1\tilde{\pi}}$ and $H_{\tilde{\pi}\tilde{\pi}}$. Regarding the estimation of standard errors, the negative of H needs to be inverted and evaluated at $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\Psi}})$. Nevertheless, H can become substantially large and can therefore suffer from singularity. In this paper, we will consider a generalized inverse of H in order to report standard errors for the parameter estimates of particular interest, i.e. $\hat{\boldsymbol{\theta}}_1$ and $\hat{\pi}$. On the contrary,

problems still occur for the second component weights which are located near the boundary of the parameter space. Since these distinct component weights are not of direct interest, we consider simultaneous bootstrap confidence intervals for the entire estimated density $\hat{f}(x)$. In this respect, $B = 2000$ bootstrap samples are taken and the estimated densities are evaluated in a finite number of grid points $\{x_1, \ldots, x_I\}$. Let $r_i^{(t)}$ denote the rank of the tth bootstrap estimate at grid point x_i. Define t_l as the lth order statistic of $[\max\{\max_{1\leqslant i\leqslant I}(r_i^{(t)}); B + 1 - \min_{1\leqslant i\leqslant I}(r_i^{(t)})\}; t = 1, \ldots, B]$. Then, by construction, the interval $[\hat{f}^*(x_i)^{[B+1-t_l]}; \hat{f}^*(x_i)^{[t_l]}]$ has a global confidence level of at least $100(l/B)\%$. In the examples below, 95% confidence intervals are considered.

4. DATA ANALYSIS

The newly developed back-fitting algorithm is applied to the two sets of MIC data introduced in Section 2. The MBM will not be detailed upon, but the reader is referred to Jaspers *and others* (2014) for a detailed description and analysis. Rather, the focus will be on the estimation of the entire mixture distribution for the combination *E. coli*–ampicillin.

4.1 *Data from EFSA*

From the visual investigation of the histogram in Figure 1(a), we could identify a first mode around the value of 2 mg/L. Application of the MBM resulted into an estimated prevalence for the wild-type component of 0.86 (s.e. 1.70e−2). The mean and standard deviation were estimated to be 1.05 (s.e. 2.25e−2) and 0.69 (s.e. 1.81e−2) on the $\log_2$-scale. These values can be used as starting values for the developed back-fitting algorithm. The final density estimates are shown on Figure 1(a). It is observed that the estimated density consists of two modes. The final estimates for the mean and standard deviation of the first component were (on the $\log_2$-scale) equal to 1.04 (s.e. 2.08e−2) and 0.54 (s.e. 2.01e−2), respectively. Therefore, the first mode is located at an MIC value of 2 mg/L. This wild-type component receives an estimated weight of 0.78 (s.e. 2.20e−2), corresponding to the estimated prevalence of wild-type isolates. It appears that there is only a unique sub-population of non-wild-type isolates in this dataset, with the mode of their distribution at an MIC value of 32 mg/L. For comparison purposes, the final estimate from the PM approach is also added to the plot in Figure 1(a). In this two-stage procedure, the mean and standard deviation are fixed at the initial estimates from the MBM and the wild-type prevalence is estimated at 0.81 (s.e. 1.76e−2). The model-based classification rule in (1.2) identifies isolates with an MIC value larger than 8 mg/L as being non-wild-type, which coincides with the harmonized ECOFF. A classification line is plotted in Figure 2, with the simultaneous bootstrap confidence limits overlaid. The limits are rather wide in the region of overlap, which is probably a consequence of the rather scarce amount of data in that specific region.

4.2 *Data from EUCAST*

Similar to the analysis above, we could again identify a wild-type distribution around the value of 2 mg/L. Figure 1(b) presents graphically the result for the EUCAST data. The mean and standard deviation (on the $\log_2$-scale) of the wild-type components are estimated at 1.04 (s.e. 7.5e−3) and 0.70 (s.e. 7.2e−3), respectively, and the prevalence of this sub-population is estimated to be 0.66 (s.e. 4.9e−3). In contrast to the EFSA data, the non-wild-type component is now composed of three sub-groups. A first, relatively small, mode is observed at an MIC value of 16 mg/L, while two larger sub-groups can be identified at 64 and 256 mg/L. Due to an increased sample size, the bootstrap confidence intervals are also more narrow compared with the previous example. Application of the model-based classification rule identifies again 8 mg/L as being the MIC value separating the wild-type from the non-wild-type component.

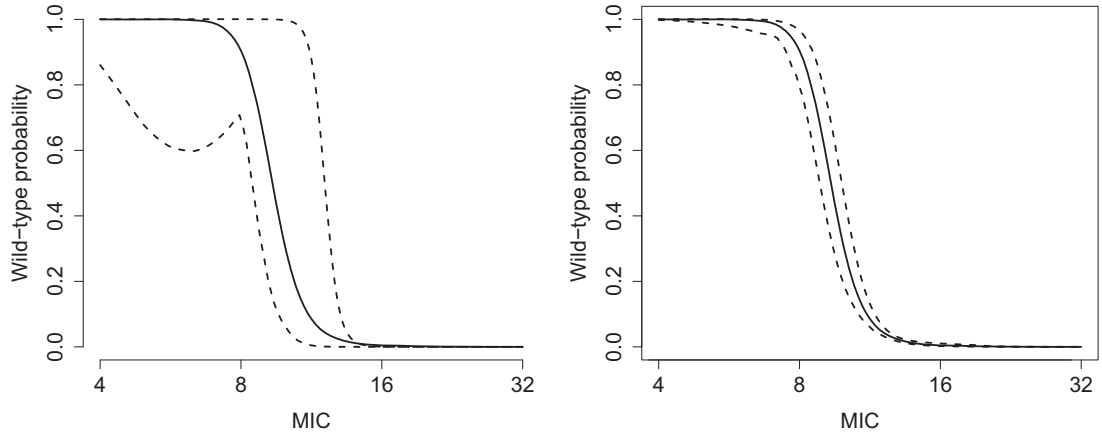

Fig. 2. Probability to belong to the wild-type class for EFSA (left) and EUCAST (right) data.

The probability for an isolate to belong to the wild-type component, shown in Figure 2, gradually diminishes with increasing MIC values.

5. Simulation study

A small simulation study is performed to assess the performance of the presented method. In general, we considered a mixture distribution with two main components, reflecting the two major sub-populations of the isolates of interest. Two different scenarios were investigated. In the first one, data are generated from a special case of the proposed model. The wild-type component is assumed to be log-normally distributed with mean 2 and standard deviation 0.8. The non-wild-type component is a 50:50 mixture of two log-normal densities with (on the $\log_2$-scale) means equal to 4.5 and 7.5, respectively, and standard deviations equal to 0.7 and 0.6, respectively. On the other hand, the second scenario considers a gamma first component with shape and scale equal to 3 and 1.6, respectively. The non-wild-type component is a 50:50 mixture of two slightly skewed t-distributions. For both scenarios, the prevalence of wild-type isolates is set to 0.6, resulting into the following mixture densities:

$$X_1 \sim 0.6 \log \mathcal{N}(2, 0.8) + 0.4\{0.5 \log \mathcal{N}(4.5, 0.7) + 0.5 \log \mathcal{N}(7.5, 0.6)\}, \tag{5.1}$$

$$X_2 \sim 0.6\Gamma(3, 1.6) + 0.4\{0.5 \text{ st}(4, 1, 1, 10) + 0.5 \text{ st}(7.5, 0.8, -1, 10)\}. \tag{5.2}$$

The considered sample sizes are 500, 1000, and 5000. In each case, the 1000 obtained samples were censored in order to resemble real-life datasets as closely as possible. The developed back-fitting algorithm is compared with the two-stage PM approach presented in Jaspers *and others* (2014).

Figures 3 and 4 present the results for mixtures (5.1) and (5.2), respectively. The estimates for each of the samples are shown in gray, with the mean estimate (dashed) and the true density (solid) overlaid. The plots on the left correspond to the PM approach, while the results of the new back-fitting algorithm can be found on the right. An important difference between the two approaches is the placement of the basis densities related to the second component. While the newly developed method puts no restriction on these locations, the two-stage PM approach allows only basis densities after a certain starting point. More details on how this starting point is determined can be found in Jaspers *and others* (2014). A direct consequence of this difference is the improved estimate in the region of overlap. This is observed most clearly on the

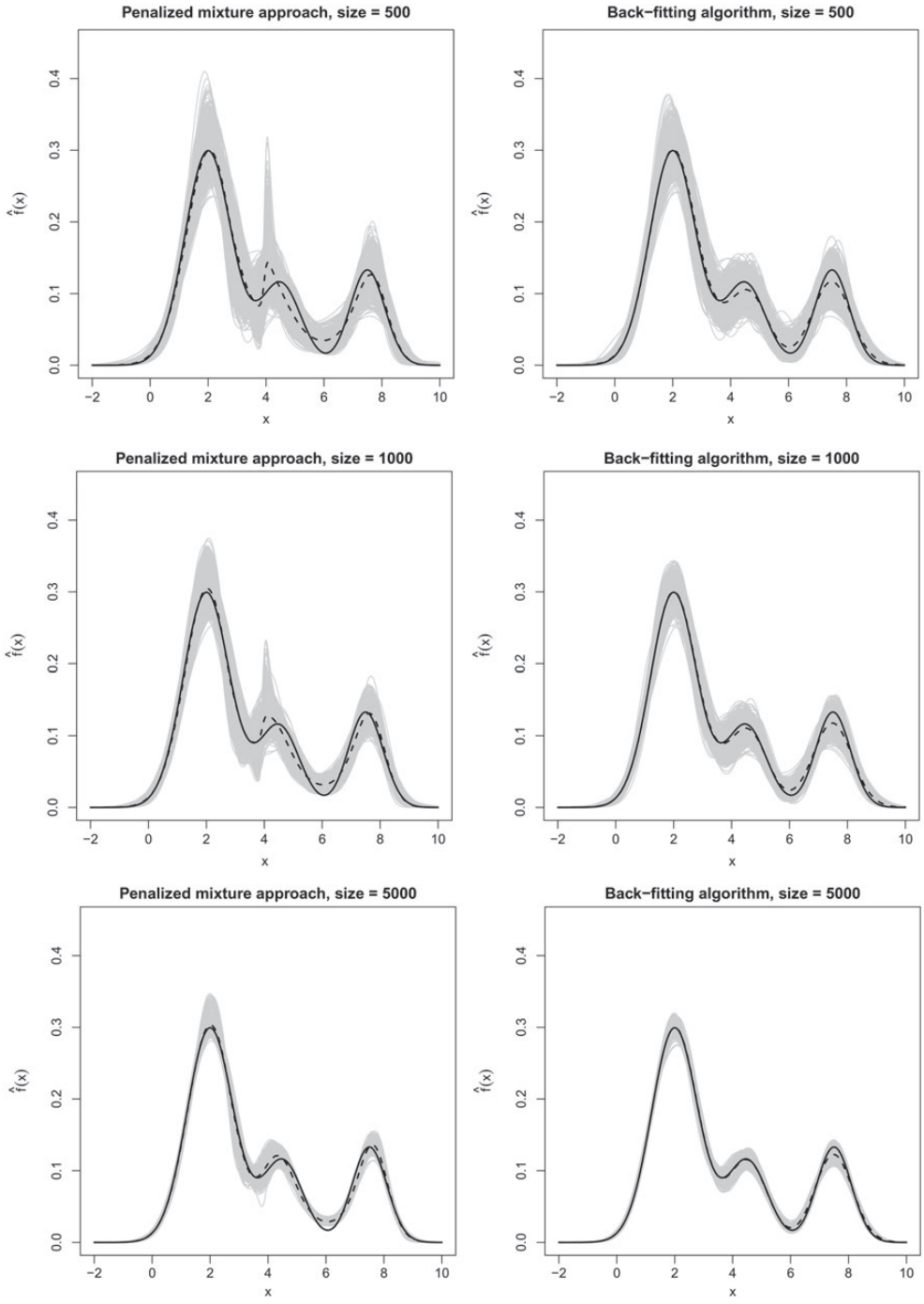

Fig. 3. Graphical representation of the simulation results for mixture (5.1). Figures on the left correspond to the PM approach, while the figures on the right resulted from the back-fitting algorithm. The individual estimates are represented in gray-scale, with the true density (full), and averaged estimate (dashed) overlaid. Sample sizes: 500 (top), 1000 (middle), 5000 (bottom).

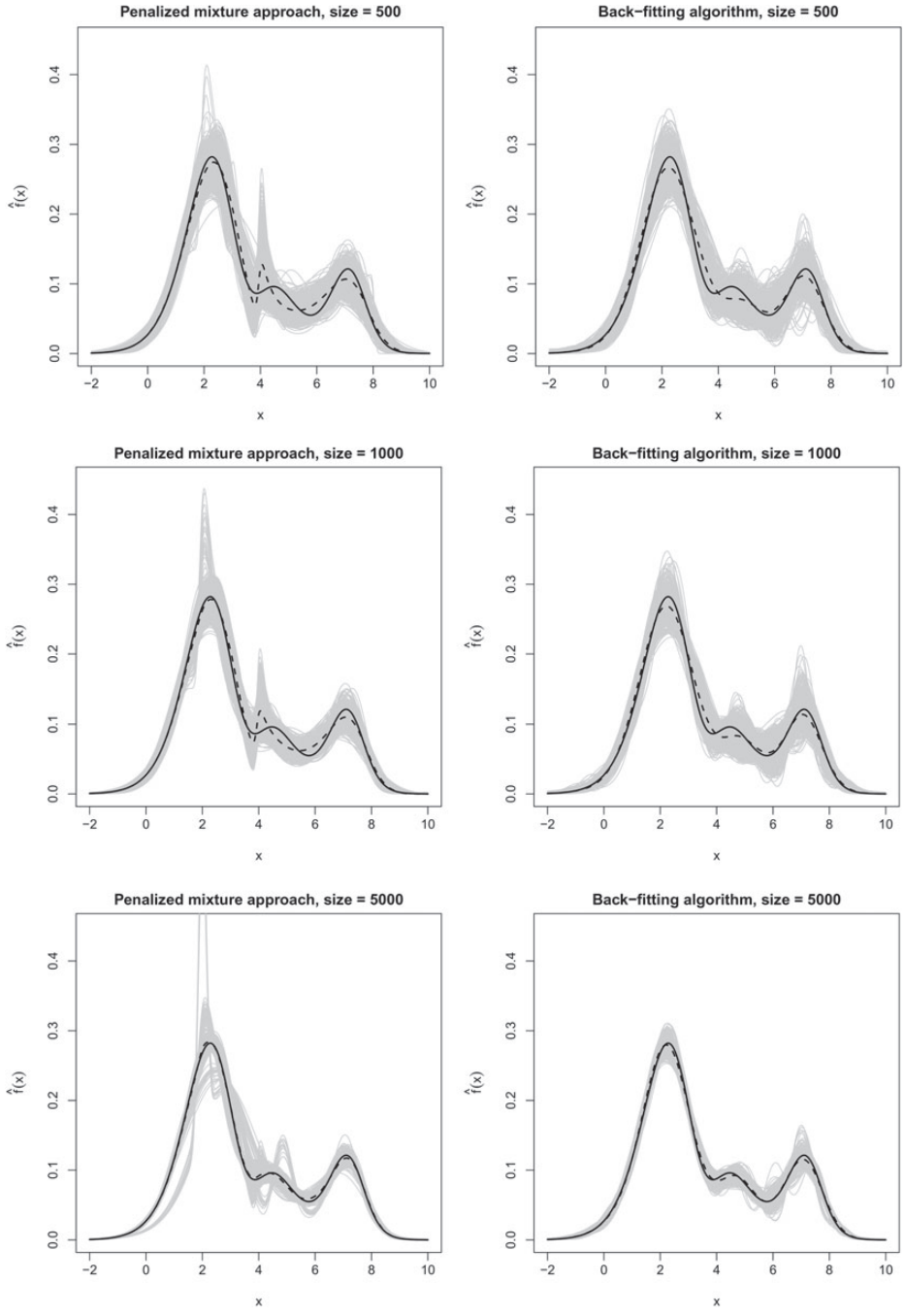

Fig. 4. Graphical representation of the simulation results for mixture (5.1). Figures on the left correspond to the PM approach, while the figures on the right resulted from the back-fitting algorithm. The individual estimates are represented in gray-scale, with the true density (full), and averaged estimate (dashed) overlaid. Sample sizes: 500 (top), 1000 (middle), 5000 (bottom).

Table 1. *IMSE values and mis-classification percentages (MP) for the two simulation scenarios*

	IMSE ($\times 10^{-5}$)		MP ($\times 10^{-2}$)	
Sample size	PM	VEM	PM	VEM
Scenario 1: mixture (5.1)				
500	34.086	21.027	7.011	6.709
1000	19.218	11.578	6.682	7.029
5000	5.675	3.064	6.648	6.648
Scenario 2: mixture (5.2)				
500	25.771	23.543	7.890	5.958
1000	16.279	15.085	8.863	5.848
5000	7.295	3.802	10.475	7.598

plots in Figure 3. For sample size 500, the PM approach results into some extreme estimates in the region of overlap, while these do not occur for the back-fitting algorithm. As sample size increases, these extreme estimates become less pronounced and both methods show a similar behavior. Nevertheless, there seems to be a slight advantage to the method introduced in this paper. A similar observation can be made for mixture (5.2). In Figure 4, the most apparent result can be seen for size 5000. The PM approach seems to provide an appropriate mean estimate, but there is a lot of variability in the first component. Again here, the back-fitting algorithm is more stable and provides better results.

Finally, for all grid values $x_i, i = 1, \ldots, I$, the MSE is calculated as $\text{MSE}_{\hat{f}(x_i)} = \text{Bias}^2_{\hat{f}(x_i)} + \text{Var}_{\hat{f}(x_i)}$, with $\text{Bias}_{\hat{f}(x_i)} = E[\hat{f}(x_i)] - f(x_i)$, and $\text{Var}_{\hat{f}(x_i)} = E[(\hat{f}(x_i) - f(x_i))^2]$. A numerical comparison between the methods can be made based on the integrated mean squared error (MSE), defined as $\text{IMSE}_{\hat{f}} = (1/I)\sum_{i=1}^{I} \text{MSE}_{\hat{f}(x_i)}$. The results are summarized in Table 1. In addition, this table also shows the mis-classification probabilities after application of the model-based classification in (1.2). It is observed that the newly introduced back-fitting algorithm outperforms the PM approach for both summary measures under consideration.

6. DISCUSSION

In this paper, we presented a new method to estimate a semi-parametric mixture model representing a continuous MIC density. The developed procedure was applied to *E. coli* isolates tested for susceptibility against ampicillin. The considered population of *E. coli* isolates was assumed to be composed of two large sub-groups, termed as the wild-type and non-wild-type sub-populations, respectively. A log-normal assumption was made for the underlying density of the wild-type component. On the other hand, very few restrictions are put on the second component density, since less information is present regarding the related sub-group. Therefore, a flexible mixture with a large number of log-normal component densities was assumed to model the density of the non-wild-type isolates in a non-parametric way. Based on the introduced back-fitting algorithm, optimal estimates were obtained for the weights of the distinct component densities and for the parameters related to the parametric first component. A key role within this method is put aside for the VEM (Böhning, 1986). Once the density estimate is obtained, model-based classification can be used to determine class-membership of the isolates under investigation. In this respect, the method provides a valuable alternative to standard methods such as visual investigation of histograms of MIC values or the statistical method provided by Turnidge *and others* (2006), which is based on an estimate for the first component only.

The approach presented above is an extended and improved version of the two-stage strategy introduced in Jaspers *and others* (2014). This latter method fixed the first component to its initial estimate and obtained a non-parametric density estimate for the second component using the PM approach (Schellhase and Kauermann, 2012). With the new approach, the initial values for the first component are updated as well. In addition, with the PM approach, there were some difficulties with the location of the component densities constituting the non-parametric second component. These issues are resolved with the new procedure. As a result, the simulation study revealed a clear improvement in the obtained estimates, especially in the region of overlap between the wild-type and non-wild-type components. Also the evolution of MSE values over the fitted range shows an advantage of the new approach compared with its competitor.

In order to incorporate the uncertainty related to the estimation process, simultaneous bootstrap confidence intervals for the density estimates were considered. Due to a singular hessian, no exact confidence intervals could be obtained. However, these singularities only prevailed for weights at the boundary of the parameter space and the generalized inverse could be used to calculate standard errors for the most important parameters, i.e. the prevalence π and the parameters related to the first component.

Finally, although we focused on a log-normal assumption for the first component, the method can be extended to incorporate other parametric assumptions. Different densities can be compared and the most the optimal one can be selected based on the AIC values. As discussed above, monitoring of AMR data is of high importance. Therefore, further research includes the estimation of time trends in the MIC densities, which can aid in the identification of possible public health risks. In addition, we are also studying the use of Bayesian methods to obtain a semi-parametric estimate of the MIC density (Jaspers *and others*, 2015).

Acknowledgments

Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged. The authors are grateful to EFSA for the approval to use the *Ampicillin* data. Finally, the first author is grateful for the hospitality during his visit to the Southampton Statistical Science Research Institute. *Conflict of Interest*: None declared.

Funding

The research of the first author was supported by the Research Foundation Flanders (FWO) [grant number 11E2913N]. For the simulations and bootstraps, we used the infrastructure of the VSC - Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government - department EWI.

References

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

Andrews, J. M. (2001). Determination of minimum inhibition concentrations. *Journal of Antimicrobial Chemotherapy* **48**, S1.5–S1.16.

Annis, D. H. and Craig, B. A. (2005). Statistical properties and inference of the antimicrobial MIC test. *Statistics in Medicine* **24**, 3631–3644.

Böhning, D. (1986). A vertex-exchange-method in D-optimal design theory. *Metrika* **33**, 337–347.

BORDES, L., DELMAS, C. AND VANDEKERKHOVE, P. (2006). Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian Journal of Statistics* **33**(4), 733–752.

BORDES, L. AND VANDEKERKHOVE, P. (2010). Semiparametric two-component mixture model with a known component: an asymptotically normal estimator. *Mathematical Methods of Statistics* **19**(1), 22–41.

CRAIG, B. A. (2000). Modeling approach to diameter breakpoint determination. *Diagnostic Microbiology and Infectious Disease* **36**, 193–202.

HOHMANN, D. AND HOLZMANN, H. (2013). Semiparametric location mixtures with distinct components. *Statistics* **47**(2), 348–362.

IGARASHI, T., INATOMI, J., WAKE, A., TAKAMIZAWA, M., KATAYAMA, H. AND IWATA, T. (1999). Failure of prediarrheal antibiotics to prevent hemolytic uremic syndrome in serologically proven *Escherichia coli* o157:h7 gastrointestinal infection. *The Journal of Pediatrics* **135**, 768–769.

JASPERS, S., AERTS, M., VERBEKE, G. AND BELOEIL, P. A. (2014a). Estimation of the wild-type minimum inhibitory concentration value distribution. *Statistics in Medicine* **33**, 289–303.

JASPERS, S., AERTS, M., VERBEKE, G. AND BELOEIL, P. A. (2014b). A new semi-parametric mixture model for interval censored data, with applications in the field of antimicrobial resistance. *Computational Statistics and Data Analysis* **71**, 30–42.

JASPERS, S., LAMBERT, P. AND AERTS, M. (2015). A Bayesian approach to the semi-parametric estimation of a MIC distribution. *Technical Report*, I-Biostat.

KRONVALL, G. (2010). Antimicrobial resistance 1979–2009 at Karolinska Hospital, Sweden: normalized resistance interpretation during a 30-year follow-up on *Staphylococcus aureus* and *Escherichia coli* resistance development. *APMIS* **118**, 621–639.

MA, Y. AND YAO, W. (2015). Flexible estimation of a semiparametric two-component mixture model with one parametric component. *Electronic Journal of Statistics* **9**, 444–474.

MCLACHLAN, G. J. AND JONES, P. N. (1988). Fitting mixture models to grouped and truncated data via the em algorithm. *Biometrics* **44**, 571–578.

MCLACHLAN, G. AND PEEL, D. (2000) *Finite Mixture Models*. New York: Wiley.

SCHELLHASE, C. AND KAUERMANN, G. (2012). Density estimation and comparison with a penalized mixture approach. *Computational Statistics* **27**, 757–777.

TADESSE, D. A., ZHAO, S., TONG, E., AYERS, S., SINGH, A., BARTHOLOMEW, M. J. AND MCDERMOTT, P. F. (2012). Antimicrobial drug resistance in *Escherichia coli* from Humans and Food Animals, United States, 1950–2002. *Emerging Infectious Diseases* **18**(5), 741–749.

TITTERINGTON, D. M., SMITH, A. F. M. AND MAKOV, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

TSONAKA, R., VERBEKE, G. AND LESAFFRE, E. (2009). A semi-parametric shared parameter model to handle nonmonotone nonignorable missingness. *Biometrics* **65**, 81–87.

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.

WIEGAND, I., HILPERT, K. AND HANCOCK, R. E. (2008). Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nature Protocols* **3**(2), 163–175.

XIANG, S., YAO, W. AND WU, J. (2014). Minimum profile Hellinger distance estimation for a semiparametric mixture model. *Canadian Journal of Statistics* **42**(2), 246–267.

[*Received March 7, 2015; revised June 17, 2015; accepted for publication July 14, 2015*]