The Annals of Applied Statistics
2016, Vol. 10, No. 2, 906–924
DOI: 10.1214/16-AOAS918

A BAYESIAN APPROACH TO THE SEMIPARAMETRIC ESTIMATION OF A MINIMUM INHIBITORY CONCENTRATION DISTRIBUTION[1]

BY STIJN JASPERS[*,2], PHILIPPE LAMBERT[†,‡] AND MARC AERTS[*]

Hasselt University, Université de Liège† and Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Université Catholique de Louvain‡*

Bacteria that have developed a reduced susceptibility against antimicrobials pose a major threat to public health. Hence, monitoring their distribution in the general population is of major importance. This monitoring is performed based on minimum inhibitory concentration (MIC) values, which are collected through dilution experiments. We present a semiparametric mixture model to estimate the MIC density on the full continuous scale. The wild-type first component is assumed to be of a parametric form, while the nonwild-type second component is modelled nonparametrically using Bayesian P-splines combined with the composite link model. A Metropolis within Gibbs strategy was used to draw a sample from the joint posterior. The newly developed method was applied to a specific bacterium–antibiotic combination, that is, *Escherichia coli* tested against ampicillin. After obtaining an estimate for the entire density, model-based classification can be performed to check whether or not an isolate belongs to the wild-type subpopulation. The performance of the new method, compared to two existing competitors, is assessed through a small simulation study.

1. Introduction. Antimicrobials are substances used to kill microorganisms or to inhibit their growth. The accidental discovery and isolation of penicillin by Sir Alexander Fleming marks the start of modern day antibiotics. Nevertheless, it soon became clear that bacteria could develop antibiotic resistance whenever too little penicillin was used or when it was used for a too short period. New antimicrobial agents have been developed ever since, but, unfortunately, so has antimicrobial resistance (AMR) [Palumbi (2001)]. An excessive and sometimes inappropriate usage of antimicrobials has led to an increasing amount of bacterial isolates that are able to withstand antimicrobial treatments. Since isolates with a reduced susceptibility to antimicrobials pose a major threat to public health, it is important to study and monitor their distribution.

Across Europe, several institutions are concerned with collecting data on antimicrobial resistance and identifying possible threats to human health. On a yearly

Received November 2014; revised June 2015.

[1]Supported by the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy).

[2]Supported in part by the Research Foundation Flanders (FWO), Grant 11E2913N.

Key words and phrases. Antimicrobial resistance, Bayesian, composite link model, interval-censored, semiparametric.

basis, the European Centre for Disease Prevention and Control (ECDC) and the European Food Safety Authority (EFSA) jointly prepare an annual European Union Summary Report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food. Since 2010, data are collected from EU member states on an isolate-based level. EFSA coordinates the annual reporting of antimicrobial resistance data from the member states, analyses the data collected and issues the results of this analysis. AMR data typically constitute minimum inhibitory concentration (MIC) values, which are collected through broth agar dilution methods. In these experiments, a standardized amount of the isolate is exposed to successive twofold concentrations of an antimicrobial (i.e., 0.5, 1, 2 mg/l, ...). The MIC is defined as the lowest concentration of the antimicrobial 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, Hilpert and Hancock (2008) for a detailed description of the collection of MIC values.

When analysing the obtained AMR data, two data complexities need to be taken into account. First of all, due to the setup of the dilution experiments, MIC data are typically censored. Indeed, 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. Second, it is unknown a priori whether an isolate belongs to the wild-type or nonwild-type subpopulation. Therefore, the analysis needs to account for unobserved population heterogeneity.

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

$$f(x) = \gamma f_1(x;\theta_1) + (1-\gamma) f_2(x;\theta_2) \tag{1}$$

is assumed, in which f_1 and f_2 represent the wild-type and nonwild-type component, respectively, of the MIC distribution and the prevalence of wild-type isolates is denoted by γ. 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 unimodal distribution reflecting a slight biological variability around a mode which is not altered by changing circumstances over time [Finch et al. (2010)]. Therefore, the first component in (1) can be assumed to be of a fixed parametric form, such as the log-normal or gamma distribution [Lee and Whitmore (1999), Turnidge, Kahlmeter and Kronvall (2006)]. The second component, representing the nonwild-type isolates, is often multi-modal, suggesting that it is itself a mixture of different nonwild-type subpopulations which are characterised

by different degrees of reduced susceptibility conferred by different mechanisms. To allow proper modelling of the different possible characteristics of the nonwild type distribution, a nonparametric approach will be considered.

Despite the importance of analysing AMR data, the statistical literature regarding this topic is rather limited. It is current practice to classify an isolate into the wild-type or nonwild-type subpopulation based on an epidemiological cutoff value (ECOFF), 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 [Kahlmeter et al. (2003)] or, alternatively, it can be statistically calculated using the approach of Turnidge, Kahlmeter and Kronvall (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 et al. (2014a) also adopt a local view, focussing on the wild-type first component only. They proposed an improved likelihood-based procedure, called the 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 subpopulation of a specific isolate. With this technique, isolates are classified to the wild-type subpopulation when the posterior probability

$$\frac{\gamma f_1(x;\theta_1)}{\gamma f_1(x;\theta_1) + (1-\gamma) f_2(x;\theta_2)} \tag{2}$$

is larger than 0.5. It is clear that this option requires an estimate for the entire mixture density f. Craig (2000) suggested that one may approximate the entire density f in (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 nonwild-type subpopulations. However, no a priori information is available on the number of components for the nonwild-type density, nor on the shape of these component density functions. Therefore, a nonparametric second component seems more appealing. Jaspers et al. (2014b) provide a two-stage semiparametric 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, that is, θ_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. These drawbacks were circumvented by the back-fitting algorithm presented in Jaspers et al. (2016). In their paper, the authors proposed a likelihood based method for the estimation of both the wild-type and nonwild-type component. The second, nonwild-type component was modelled using a generous number of normal densities for which the weights were determined using the Vertex Exchange Method [Böhning (1986)]. Although the estimator provides good estimates for the MIC density of interest, it was found to be less trivial to extend it to include covariates. The inclusion of a time component especially seems very appealing when interest is in developing a monitoring tool that is able to detect trends over distinct years in the MIC distribution. In this paper, an alternative to the back-fitting algorithm is presented. The approach will follow Lambert and Eilers (2009) and adopt a Bayesian viewpoint for the estimation of the MIC density of interest. More specifically, a combination of the composite link model (CLM) with roughness penalties is considered.

In Section 2, we will discuss the Bayesian composite link model approach by Lambert and Eilers (2009) and explain how it can be modified to fit within the antimicrobial resistance context. An application of the new method to two data examples can be found in Section 3 and a simulation study shows its performance in Section 4. Finally, a discussion ends the paper in Section 5.

2. Mathematical framework. As argued above, AMR data typically constitute MIC values obtained from dilution experiments. Since these experiments deliver only censored readings, we are dealing with grouped continuous data, meaning that the frequencies of observations in fixed intervals are reported. The concept of estimating a density from grouped continuous data was addressed by Lambert and Eilers (2009) from a Bayesian viewpoint. In this section, we will first elaborate on this original method and, in a second stage, expand it to our data setting.

2.1. *Bayesian composite link model.* Lambert and Eilers (2009) present a Bayesian approach to density estimation from grouped continuous data. The authors propose a combination of the composite link model with roughness penalties to estimate smooth continuous densities from such data. In this regard, assume that one is interested in estimating a discrete representation of a continuous density function $f_X(\cdot)$ of a random variable X on a fine grid on (a_0, a_J). This fine grid consists of many grid points (say, 100 or more) that partition (a_0, a_J) into I consecutive intervals $l_i = (\chi_{i-1}, \chi_i)$ of equal width Δ with midpoints u_i $(i = 1, \ldots, I)$, $\chi_0 = a_0$ and $\chi_I = a_J$. With this notation, the quantities of interest are $\pi_i = \int_{l_i} f_X(t)\,dt \approx f_X(u_i)\Delta$.

Let m_j $(j = 1, \ldots, J)$ denote the number of observations belonging to each of the given nonoverlapping wide bins $\mathcal{J}_j = (a_{j-1}, a_j)$ partitioning (a_0, a_J). For simplicity, it is assumed that the limits of these wide bins constitute a subset of $\{\chi_0, \ldots, \chi_I\}$. The relationship between the wide bins and the initial grid is coded by the $J \times I$ matrix $C = [c_{ji}]$, where $c_{ji} = 1$ if $l_i \subset \mathcal{J}_j$ and 0 otherwise.

The probability that an observation belongs to the jth wide bin, $\mathcal{J}_j$, can now be modelled as $\kappa_j = \sum_{i=1}^{I} c_{ji}\pi_i$, or, in vector notation, $\boldsymbol{\kappa} = C\boldsymbol{\pi}$. In case the only available data are the frequencies associated with the wide bins, the estimation of the π_i's is an ill-conditioned problem. Therefore, π is requested to be smooth, following the P-spline approach presented by Eilers and Marx (1996) and Eilers (2007).

More specifically, consider a basis $\{b_k(\cdot) : k = 1, \ldots, K\}$ of B-splines associated to equidistant knots on (a_0, a_J). Denote by $(B)_{ik} = b_{ik} = b_k(u_i)$ the $I \times K$ matrix giving the basis functions evaluated at the midpoints u_i $(i = 1, \ldots, I)$. We now model π_i by

$$\pi_i = \pi_i(\boldsymbol{\phi}) = \frac{e^{\eta_i}}{e^{\eta_1} + \cdots + e^{\eta_I}},$$

with $\boldsymbol{\eta} = B\boldsymbol{\phi}$ and the identifiability constraint $\sum_{k=1}^{K} \phi_k = 0$. The P-spline penalty is based on rth order differences, $\Delta^r\boldsymbol{\phi}$, of the spline coefficients $\boldsymbol{\phi}$. In a Bayesian setting, this translates into a prior distribution on the spline coefficients [Lang and Brezger (2004)]:

$$\Delta^r \phi_k \sim \mathcal{N}\bigl(0, \tau^{-1}\bigr).$$

Therefore, the (improper) prior for the B-spline coefficients is assumed to be

$$p(\boldsymbol{\phi}|\tau) \propto \tau^{\mathcal{R}(p)/2} \exp\left\{-\frac{\tau}{2}\boldsymbol{\phi}' P \boldsymbol{\phi}\right\}.$$

In the formula above, $\mathcal{R}(P)$ denotes the rank of P (in general, $K - r$) and $P = D'D$ is the penalty matrix such that $\sum_k (\Delta^r \phi_k)^2 = \boldsymbol{\phi}' P \boldsymbol{\phi}$. A gamma prior $\mathcal{G}(a, b)$ is usually advocated to express prior ignorance about suitable values for τ.

Hence, the model for π, apart from the penalty, is an ordinary generalised linear model (GLM), whereas the model for $\boldsymbol{\kappa}$ is a composite link model [Thompson and Baker (1981)].

Using Bayes' theorem, the joint posterior distribution is found to be

$$\begin{aligned} p(\boldsymbol{\phi}, \tau|\mathcal{D}) &= \frac{p(\mathcal{D}|\boldsymbol{\phi}, \tau)\, p(\boldsymbol{\phi}, \tau)}{p(\mathcal{D})} \\ &\propto p(\mathcal{D}|\boldsymbol{\phi}, \tau)\, p(\boldsymbol{\phi}, \tau) \\ &\propto \left\{\prod_{j=1}^{J} \kappa_j^{m_j}\right\} \tau^{a+0.5\mathcal{R}(P)-1} \exp\bigl\{-\tau\bigl(b + 0.5\boldsymbol{\phi}' P \boldsymbol{\phi}\bigr)\bigr\}, \end{aligned}$$

where $\mathcal{D}$ denotes the observed frequencies associated with the wide bins, that is, $\{(m_j, \mathcal{J}_j), j = 1, \ldots, J\}$. The conditional posterior distributions for τ and $\boldsymbol{\phi}$ are, respectively,

$$(\tau|\boldsymbol{\phi}, \mathcal{D}) \sim \mathcal{G}\bigl(a + 0.5\mathcal{R}(P), b + 0.5\boldsymbol{\phi}' P \boldsymbol{\phi}\bigr), \tag{3}$$

$$p(\boldsymbol{\phi}|\tau, \mathcal{D}) \propto \left\{\prod_{j=1}^{J} \kappa_j^{m_j}\right\} \exp\left\{-\frac{\tau}{2}\boldsymbol{\phi}' P \boldsymbol{\phi}\right\}. \tag{4}$$

Markov chain Monte Carlo methods can be used to draw a sample from the posterior. The authors propose a Metropolis-within-Gibbs sampling strategy with a Gibbs step for τ [see equation (3)] and a Metropolis step through the modified Langevin–Hastings algorithm for $\boldsymbol{\phi}$ with equation (4). In the latter algorithm, proposals for the B-spline coefficients are generated from a multivariate normal density. With $p(\boldsymbol{\phi}|\mathcal{D})$ the posterior distribution used as shorthand notation for $p(\boldsymbol{\phi}|\mathcal{D}, \tau^{(m-1)})$ and $\boldsymbol{\phi}^{(m-1)}$ the state of the chain at iteration $(m-1)$, the proposal for $\boldsymbol{\phi}$ at iteration m is obtained by taking a random sample from

$$N\bigl(\boldsymbol{\phi}^{(m-1)} + 0.5\delta\Sigma\nabla\log p\bigl(\boldsymbol{\phi}^{(m-1)}|\mathcal{D}\bigr), \delta\Sigma\bigr),$$

where $\delta > 0$ and Σ is ideally an approximation to the second order dependence structure of the conditional posterior. The acceptance probability is defined as

$$\alpha\bigl(\boldsymbol{\phi}^{(m-1)}, \boldsymbol{\phi}\bigr) = \min\biggl\{1, \frac{p(\boldsymbol{\phi}|\mathcal{D})}{p(\boldsymbol{\phi}^{(m-1)}|\mathcal{D})}\frac{q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(m-1)})}{q(\boldsymbol{\phi}^{(m-1)}, \boldsymbol{\phi})}\biggr\},$$

where

$$\frac{q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(m-1)})}{q(\boldsymbol{\phi}^{(m-1)}, \boldsymbol{\phi})} = \exp\biggl\{-\frac{1}{2}\bigl(G + G^{(m-1)}\bigr)'\biggl(\bigl(\boldsymbol{\phi} - \boldsymbol{\phi}^{(m-1)}\bigr) + \frac{\delta\Sigma}{4}\bigl(G - G^{(m-1)}\bigr)\biggr)\biggr\},$$

with

$$G = \nabla\log p(\boldsymbol{\phi}|\mathcal{D}) \quad \text{and} \quad G^{(m-1)} = \nabla\log p\bigl(\boldsymbol{\phi}^{(m-1)}|\mathcal{D}\bigr).$$

The value of δ can be tuned to reach the optimal acceptance rate of 0.57. For more details, the reader is referred to Haario, Saksman and Tamminen (2001), Atchadé and Rosenthal (2005) and Lambert and Eilers (2009).

2.2. *Adaptation to cope with parametric first component.* Although the methodology presented by Lambert and Eilers (2009) provides a very nice, nonparametric estimate of a continuous density from grouped data, it still needs updates to fit within the purpose of this paper. Indeed, a separate estimate for the parametric first component is required in order to perform model-based classification. Additional interest is also in the prevalence of nonwild-type isolates and in the characteristics of the wild-type component density. Therefore, we will now present the extension to the existing methodology.

Let

$$\pi_i = P(X \in l_i) = \gamma\underbrace{\bigl[f_1(u_i; \theta_1)\Delta\bigr]}_{\pi_i^{(1)}} + (1-\gamma)\underbrace{\bigl[f_2(u_i)\Delta\bigr]}_{\pi_i^{(2)}}$$

and κ_j be as defined before.

The wild-type component is commonly accepted to be unimodally distributed. Hence, the first part of the mixture can be assumed to be of a parametric form.

The focus in this paper will be on the log-normal assumption, although other assumptions can be implemented as well. More specifically, after log-transforming the data, we will model $\pi_i^{(1)}$ using the cumulative normal distribution function, Φ, as follows:

$$\pi_i^{(1)} = \Phi(\chi_i; \mu_1, \sigma_1) - \Phi(\chi_{i-1}; \mu_1, \sigma_1),$$

where χ_i and χ_{i-1} are the upper and lower bound of the small bin l_i and μ_1 and σ_1 denote the mean and standard deviation of the log-normal first component, respectively.

Regarding the small-bin probabilities related to the second component, we will follow the B-spline approach presented in Lambert and Eilers (2009):

$$\pi_i^{(2)}(\boldsymbol{\phi}) = \frac{\exp(\eta_i)}{\exp(\eta_1) + \cdots + \exp(\eta_I)},$$

with $\boldsymbol{\eta} = B\boldsymbol{\phi}$ and $\sum \phi_k = 0$.

In addition to the B-spline coefficients $\boldsymbol{\phi}$ and the penalty parameter τ, we need to estimate three parameters, that is, the mixing weight γ and the mean (μ_1) and standard deviation (σ_1) related to the log-normal first component. Assuming a normal prior for the mean μ_1 of the first component (with hyperparameters μ_{11} and μ_{12} denoting the mean and standard deviation, respectively), an inverse gamma prior for σ_1(with hyperparameters denoted by σ_{11} and σ_{12}) and a beta prior for the mixing weight γ (with hyperparameters α and β), one gets as joint posterior

$$\begin{aligned}
&p(\gamma, \mu_1, \sigma_1, \boldsymbol{\phi}, \tau | \mathcal{D}) \\
&\quad \propto \left\{ \prod_{j=1}^{J} \kappa_j^{m_j} \right\} p(\gamma) p(\mu_1) p(\sigma_1) p(\boldsymbol{\phi}|\tau) p(\tau) \\
&\quad \propto \left\{ \prod_{j=1}^{J} \kappa_j^{m_j} \right\} \gamma^{\alpha-1} (1-\gamma)^{\beta-1} \exp\left(-\frac{(\mu_1 - \mu_{11})^2}{2\mu_{12}^2} \right) \sigma_1^{-\sigma_{11}-1} \exp\left(-\frac{\sigma_{12}}{\sigma_1} \right) \\
&\qquad \times \tau^{a+0.5R(p)-1} \exp\left(-\tau\left[b + 0.5\boldsymbol{\phi}' P \boldsymbol{\phi}\right]\right).
\end{aligned}$$

As will become clear in Section 3.1, priors were taken to be relatively informative, with their means corresponding to the estimates of the multinomial-based method [Jaspers et al. (2014a)].

From the joint posterior, we obtain the following conditional posterior distributions:

$$(\tau | \gamma, \theta_1, \boldsymbol{\phi}, \mathcal{D}) \sim \mathcal{G}\left(a + 0.5R(p) - 1; b + 0.5\boldsymbol{\phi}' P \boldsymbol{\phi}\right),$$

$$p(\boldsymbol{\phi} | \gamma, \theta_1, \tau, \mathcal{D}) \propto \prod_{j=1}^{J} \kappa_j^{m_j} \exp(-0.5\tau \boldsymbol{\phi}' P \boldsymbol{\phi}),$$

$$p(\gamma|\theta_1, \boldsymbol{\phi}, \tau, \mathcal{D}) \propto \prod_{j=1}^{J} \kappa_j^{m_j} \gamma^{\alpha-1}(1-\gamma)^{\beta-1},$$

$$p(\mu_1|\sigma_1, \gamma, \boldsymbol{\phi}, \tau, \mathcal{D}) \propto \prod_{j=1}^{J} \kappa_j^{m_j} \exp\left(-\frac{(\mu_1-\mu_{11})^2}{2\mu_{12}^2}\right),$$

$$p(\sigma_1|\mu_1, \gamma, \boldsymbol{\phi}, \tau, \mathcal{D}) \propto \prod_{j=1}^{J} \kappa_j^{m_j} \sigma_1^{-\sigma_{11}-1} \exp\left(-\frac{\sigma_{12}}{\sigma_1}\right).$$

Inference is performed in full similarity to the original method (see Section 2.1). Hence, MCMC methods are used to draw a sample $\{(\boldsymbol{\phi}^{(m)}, \tau^{(m)}, \mu_1^{(m)}, \sigma_1^{(m)}, \gamma^{(m)}), m = 1, \ldots, M\}$ from the posterior. In this respect, a Gibbs step is used for sampling τ, while the Langevin–Hastings algorithm is employed to sample from the remaining posteriors.

3. Data analysis: *Escherichia coli* vs. ampicillin. *Escherichia coli* is a Gram-negative bacterium that is commonly present in the digestive tracts of humans and animals. Although it is a commensal, pathogenic variants can cause intestinal and extra-intestinal infections, including urinary 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 et al. (1999)]. Several studies have shown that resistance in *E. coli* isolates is relatively high and has been emerging over the last decades. For example, a 30-year follow up study performed in Sweden showed an increasing resistance trend for ampicillin, sulfonamide, trimethoprim and gentamicin [Kronvall (2010)]. Similarly, a retrospective study during 1950–2002 performed in different US states revealed a significant upward trend in resistance for ampicillin, sulfonamide and tetracycline [Tadesse et al. (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. For the purpose of this modelling study, data were obtained from two major European institutions concerned with the collection and analysis of AMR data, that is, EFSA and EUCAST.

3.1. *Data from EFSA.* Since 2010, data concerning antimicrobial resistance are collected from European Union member states on an isolate-based level. EFSA coordinates the annual reporting of AMR data from the member states, analyses the data collected and issues the results of this analysis. An exemplary MIC distribution summarising 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 in a subset of 1890 isolates. A graphical representation of the MIC value distribution can be found in Figure 1. The mode of the wild-type component is located between the

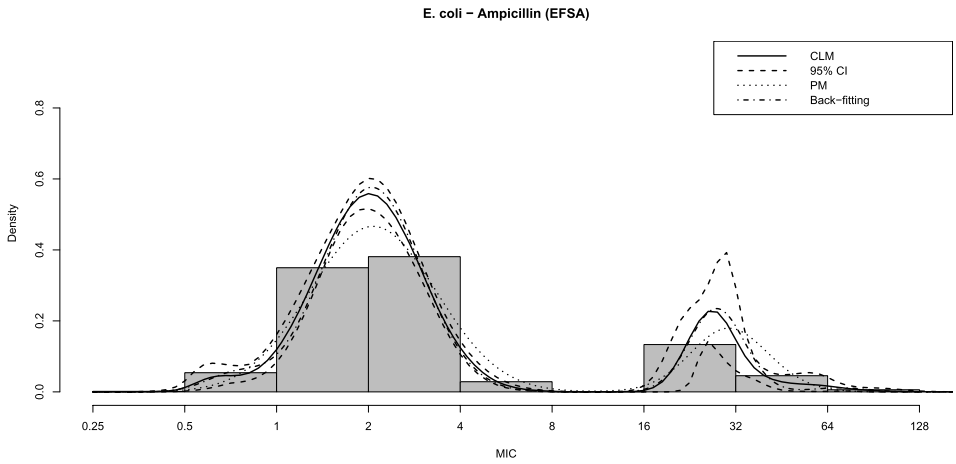


FIG. 1. *Barplot of E. coli isolates tested for susceptibility against ampicillin—source*: *EFSA*. *Overlaid are the estimated density using the updated Bayesian CLM* (*solid*) *with its* 95% *credible interval* (*dashed*), *the estimate resulting from the PM approach* (*dotted*) *and the VEM estimate* (*dash-dotted*).

values of 2 and 4 mg/l and there is presumably a unique nonwild-type population, for which the modal MIC is located at 32 mg/l.

Initial estimates for the parameters of the first component were obtained using the MBM. On the $\log_2$ scale, the mean was estimated to be 1.05 (0.02), while the standard deviation of the Gaussian first component was estimated at 0.69 (0.02). The estimated mixing weight corresponding to this first component was 0.86 (0.02). These initial values were used to construct the priors that are required for the updated Bayesian CLM. More specifically, the employed hyperparameters are constructed such that the mean of the prior distribution corresponds to the point estimates from the MBM, with variances equal to $\sqrt{0.05}$ for μ_1, 0.05 for σ_1 and 0.0005 for the mixing weight γ. As a result, the prior distributions of the first component parameters are relatively informative, but still allow some variability in the vicinity of the initial estimates. This information is essential for identifiability reasons. In addition, 40 equally spaced cubic B-splines were considered as the basis for the nonparametric second component and a third order penalty was employed.

Figure 1 shows the estimated density (black solid line), computed from a MCMC chain of length 200,000, accompanied with the 95% pointwise credible interval (black dashed line). Two major modes can be identified. The first mode, representing the wild-type population, corresponds to the mean of the parametric first component. The mean value for this parameter is 1.01 (0.03), while the standard deviation of the Gaussian first component is estimated at a mean value of 0.57 (0.03). The prevalence of wild-type isolates is estimated to be 0.80 (0.01). In addition to this first component, we also obtain an estimate for the entire MIC distribution. This is especially convenient when interest is in performing model-based

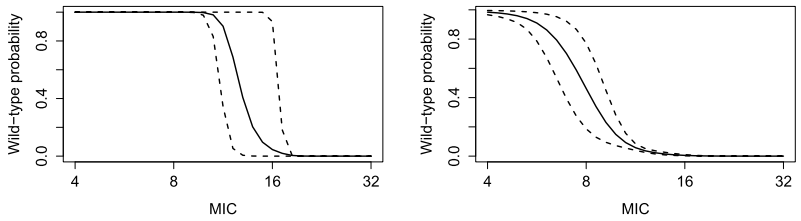


FIG. 2. *Probability to belong to the wild-type class for EFSA* (*left*) *and EUCAST data* (*right*).

classification. Figure 2 shows the probability to belong to the wild-type subpopulation. It is observed that the breakpoint is located between 8 and 16 mg/l. This value corresponds to the harmonised ECOFF proposed by EUCAST. For comparison purposes, the estimates obtained with the back-fitting algorithm and the PM approach are overlaid on Figure 1.

3.2. *Data from EUCAST.* Another important organization within the field of AMR is the European Committee on Antimicrobial Susceptibility Testing (EUCAST). This organization is mainly concerned with breakpoints and technical aspects of phenotypic in vitro antimicrobial susceptibility testing. Most antimicrobial MIC breakpoints (e.g., epidemiological cutoff 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 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 3. 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 nonwild-type isolates.

Again, here Figure 3 shows the estimated MIC density for the three methods. A similar conclusion can be made from the different approaches, but focus is again on the newly introduced Bayesian CLM. The mean of the first component is estimated at 1.06 (0.01), while the estimated standard deviation is 0.68 (0.01). The prevalence of wild-type isolates corresponding to the mixing weight of the first component is estimated at 0.63 (0.01). It is, however, noteworthy that these data from EUCAST are collected over different time periods and geographical regions, such that these latter estimates are only exemplary. Finally, Figure 2 shows the classification probability. Also, for this dataset, the MIC value of 8 mg/l can be termed as the cutoff between wild-type and nonwild-type isolates.

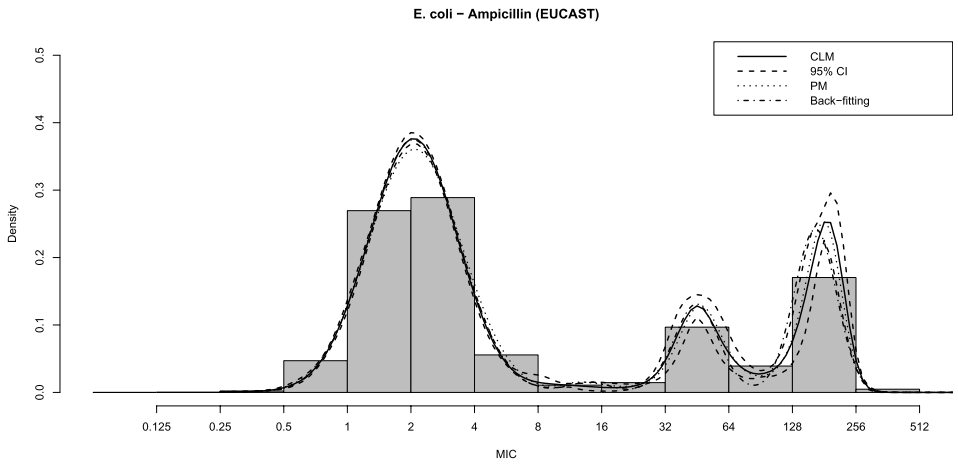


FIG. 3. *Barplot of E. coli isolates tested for susceptibility against ampicillin—source*: *EUCAST* (*website*: *mic.eucast.org/Eucast2/regShow.jsp?Id=1529*). *Overlaid are the estimated density using the updated Bayesian CLM* (*solid*) *with its* 95% *credible interval* (*dashed*), *the estimate resulting from the PM approach* (*dotted*) *and the VEM estimate* (*dash-dotted*).

4. 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 subpopulations of the isolates of interest. Two different scenarios were investigated. In the first one, the wild-type component is assumed to be log-normally distributed with mean 2 and standard deviation 0.8. Note that this means that the correct distributional assumption for the first component will be made in the Bayesian CLM approach. The nonwild-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 nonwild-type component is a 50:50 mixture of two slightly skewed t-distributions. Hence, this second scenario shows what will happen if we make an incorrect assumption about the first component. For both scenarios, the prevalence of wild-type isolates is set to 0.6, resulting in the following mixture densities:

$$X_1 \sim 0.6 \log \mathcal{N}(2, 0.8) + 0.2 \log \mathcal{N}(4.5, 0.7) + 0.2 \log \mathcal{N}(7.5, 0.6), \tag{5}$$

$$X_2 \sim 0.6 \mathcal{G}(3, 1.6) + 0.2\,\mathrm{st}(4, 1, 1, 10) + 0.2\,\mathrm{st}(7.5, 0.8, -1, 10), \tag{6}$$

where st denotes the skewed t-distribution as described in Azzalini and Capitanio (2003). 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. In the absence of a golden standard, we decided to compare between three existing methods. The adjusted CLM approach is compared to the

back-fitting algorithm and to the two-stage penalized mixture (PM) approach presented in Jaspers et al. (2014b, 2016), respectively.

The plots in Figure 4 relate to the first scenario. The results from the PM approach are located in the left column, those from the back-fitting algorithm in the middle and the estimates from the Bayesian CLM can be found in the right column. It is immediately clear that the back-fitting algorithm and the Bayesian CLM outperform the PM approach, especially in the region of overlap between the first and second components. This could be expected since these methods allow an update of the parameters of the first component, while the PM approach considers them fixed. On the other hand, the results of the back-fitting algorithm and those from the Bayesian CLM are comparable. It seems that the newly introduced method has a somewhat higher variability related to the estimate of the second component and slightly overestimates the valley between the last two modes in the case of the largest sample size. However, recall that this first scenario consists of three log-normal densities and that the basis employed for the back-fitting algorithm consists of the same type of densities. As such, this first scenario is a special, simplified case of the model considered in the back-fitting algorithm and it is therefore not a surprise that method slightly outperforms the fully nonparametric Bayesian CLM.

Figure 6 shows the evolution of the mean squared error (MSE) values for the estimated density resulting from mixture (5). More specifically, for all grid values $x_i, i = 1, \ldots, I$, the MSE is calculated as follows:

$$\mathrm{MSE}_{\hat{f}(x_i)} = \mathrm{Bias}^2_{\hat{f}(x_i)} + \mathrm{Var}_{\hat{f}(x_i)},$$

with

$$\mathrm{Bias}_{\hat{f}(x_i)} = E\big[\hat{f}(x_i)\big] - f(x_i),$$

$$\mathrm{Var}_{\hat{f}(x_i)} = E\big[\big(\hat{f}(x_i) - f(x_i)\big)^2\big].$$

The conclusions made from Figure 4 are confirmed in this plot. Apart from some deviations for the smaller sample sizes, the dotted line (Bayesian CLM) is almost everywhere located between the solid (PM approach) and dashed lines (back-fitting algorithm). Hence, in terms of MIC, we could state that the newly introduced CLM approach performs similarly to the back-fitting algorithm, with a slight advantage for the latter method in this scenario.

A numerical comparison can be made based on the integrated MSE:

$$\mathrm{IMSE}_{\hat{f}} = \frac{1}{I}\sum_{i=1}^{I} \mathrm{MSE}_{\hat{f}(x_i)}.$$

The same conclusions can be made from the numerical counterparts, which can be found in Table 1. For all sample sizes, the PM approach is outperformed by both the back-fitting algorithm and the Bayesian CLM. In addition, the IMSE values for

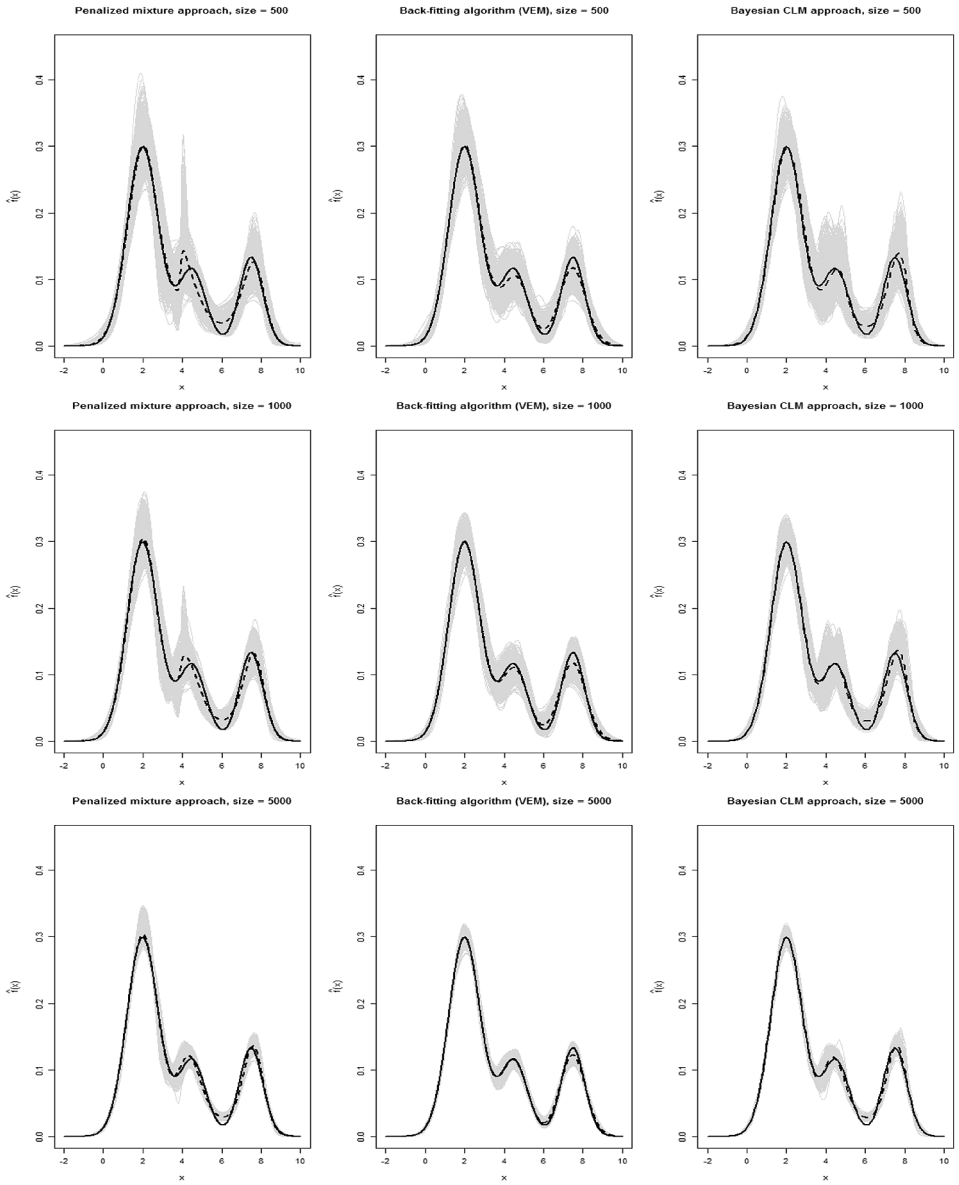


FIG. 4. *Graphical representation of the simulation results for mixture* (5). *The left column corresponds to the PM approach, the middle column to the back-fitting algorithm and the right column shows results for the new CLM approach. The individual estimates are represented in grey scale, with the true density* (*full*) *and averaged estimate* (*dashed*) *overlaid. Sample sizes*: 500 (*top*), 1000 (*middle*), 5000 (*bottom*).

TABLE 1
Integrated mean squared error (IMSE) values and Misclassification percentages (MP) for the two simulation scenarios

	IMSE ($\times 10^{-5}$)			MP ($\times 10^{-2}$)		
Sample size	**PM**	**VEM**	**CLM**	**PM**	**VEM**	**CLM**
			Scenario 1: Mixture (5)			
500	34.086	21.027	29.729	7.011	6.709	6.074
1000	19.218	11.578	17.274	6.682	7.029	6.413
5000	5.675	3.064	4.453	6.648	6.648	6.648
			Scenario 2: Mixture (6)			
500	25.771	23.543	23.865	7.890	5.958	5.830
1000	16.279	15.085	14.522	8.863	5.848	5.620
5000	7.295	3.802	4.641	10.475	7.598	6.449

the newly introduced method are intermediate between its two competitors. For all three methods, IMSE values decrease when sample size increases.

Similarly, the plots in Figure 5 show the individual and mean estimates for the second scenario under investigation. Recall that, in this scenario, the underlying first component is not Gaussian, so an incorrect distributional assumption is made for both the back-fitting algorithm and the Bayesian CLM. Nevertheless, they both still outperform the PM approach. Even for the smaller sample sizes, the newly introduced method successfully identifies the three modes and therein performs better than the back-fitting algorithm.

From Figure 6, one can also claim that both the back-fitting algorithm and the Bayesian CLM perform better than the PM approach and that the Bayesian CLM is a valuable alternative to the back-fitting algorithm. The IMSE values for the latter two methods tend to be more similar than in scenario 1, again with a larger IMSE for the PM. The back-fitting algorithm resulted in the smallest IMSE values for the largest sample size (5000), but was comparable to the Bayesian CLM for sizes 500 and 1000.

Finally, Table 1 also presents the misclassification errors made when applying the model-based classification described in equation (2) to the sampled datasets. It is observed that, for both scenarios of interest, the Bayesian CLM outperforms the two older methods, which is an additional advantage for the newly developed method.

5. Discussion. In this paper, we introduced a new method for the estimation of a minimum inhibitory concentration (MIC) value distribution. Since we need to deal with unobserved population heterogeneity, a mixture model approach was considered. The mixture consists of two main components, termed as the wild-type

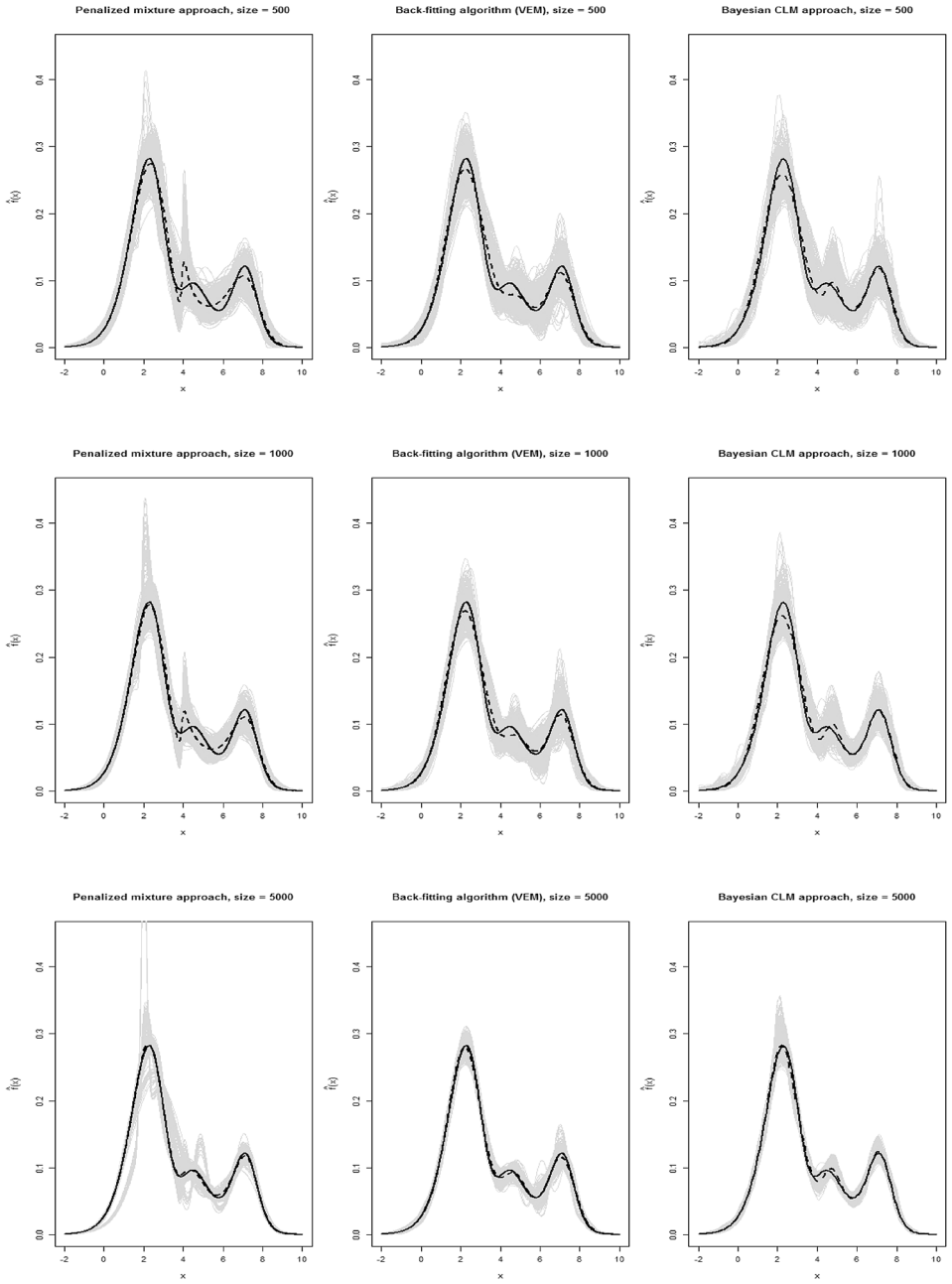
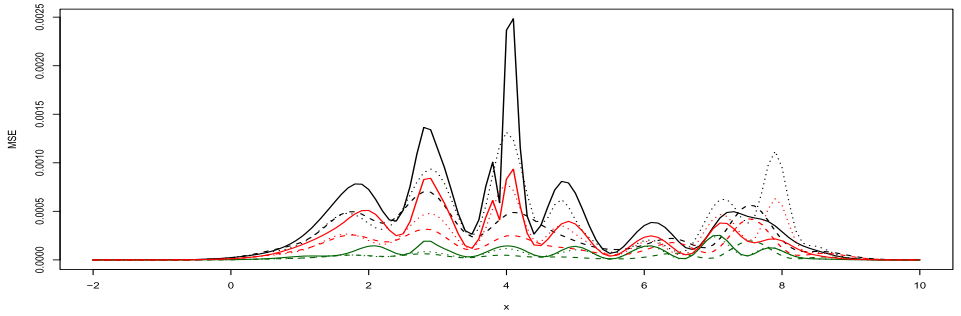
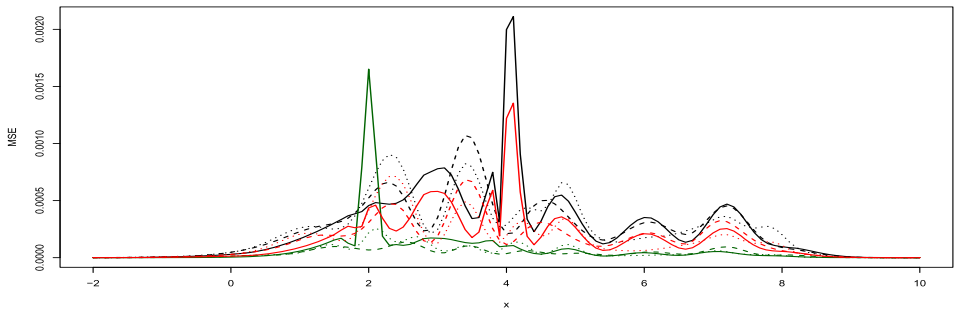


FIG. 5. *Graphical representation of the simulation results for mixture* (6).*The left column corresponds to the PM approach, the middle column to the back-fitting algorithm and the right column shows results for the new CLM approach. The individual estimates are represented in grey scale, with the true density* (*full*) *and averaged estimate* (*dashed*) *overlaid. Sample sizes*: 500 (*top*), 1000 (*middle*), 5000 (*bottom*).



(a) Evolution of MSE for mixture (5)



(b) Evolution of MSE for mixture (6)

FIG. 6. *Evolution of the MSE values of the estimated densities for sample sizes* 500 (*black*), 1000 (*red*) *and* 5000 (*green*) *regarding mixtures* (5) *and* (6). *The solid lines refer to the penalized mixture approach*, *the dashed lines refer to the back-fitting algorithm and the dotted lines represent the CLM approach*.

first component and the nonwild-type second component. The wild-type subpopulation is assumed to have no acquired or mutational resistance and commonly shows a unimodal distribution reflecting a slight biological variability around a mode which is not altered by changing circumstances over time. Therefore, a simple parametric specification seems appropriate. Focus in this paper was on the log-normal assumption, although other distributions can be implemented as well. On the other hand, we have no detailed information on the distribution of the nonwild-type isolates. This subpopulation is probably further subdivided into different types of nonsusceptible isolates, each with their respective distribution around a different mode. Since we do not want to pose any restrictions on this second component density, we employed a nonparametric approach, following Lambert and Eilers (2009). Therefore, the new method is an updated version of the Bayesian composite link model. The new approach was applied on two datasets concerning antimicrobial susceptibility of *E. coli* against ampicillin. In addition to obtaining the MIC density estimate, the procedure provides separate estimates for the first com-

ponent, which can be used for model-based classification. All computations were done with R and the employed code can be obtained from the first author.

The developed method is able to deal with the tree types of censoring that are related to dilution experiments (i.e., left-, right- and interval-censoring). Nevertheless, a drawback in real-life applications is that data can be collected from different laboratories across different countries, yielding different dilution ranges. The method is not able to deal with these nonuniform ranges. Fortunately, formal rules for the collection of AMR data, including the selection of ranges, have been postulated on the European Union level, which will reduce the amount of nonuniform data ranges in the future [2013/652/EU (2013)]. For this reason, we made the assumption of constant ranges in our data examples, with interval-censored observations throughout, except for the largest assumed to be right-censored.

From the IMSE values obtained in the simulation study in Section 4, we could observe that the newly introduced method outperformed the PM approach in both scenarios under investigation. In scenario 1, the CLM was slightly outperformed by the back-fitting algorithm, but scenario 2 suggests that the Bayesian CLM is a better option when simulated data do not strictly conform to the back-fitting algorithm hypothesis. This second scenario is more likely to occur in practice since continuing shifts in the unknown underlying distribution can easily result into skewness. In addition, the misclassification rates resulting from the Bayesian CLM were smallest. It should be noted that, in both scenarios, there was a limited amount of overlap between wild-type and nonwild-type isolates. This was based on the two data examples where the two subpopulations were relatively well segregated. However, increasing the amount of overlap reduces the performance properties of the new method (as well as that of its competitors) since the identifiability of both components will diminish. Nevertheless, based on both the simulation study and the data analysis, we believe the new Bayesian CLM can have wide applicability in the field of antimicrobial resistance (AMR). Moreover, the method can be adjusted to incorporate covariates (e.g., time, country, ...). Indeed, specific attention in the field of AMR is the detection of possible shifts over time in the distribution. While we know that the wild-type component is stable over time, it is possible that the prevalence of wild-type isolates increases or decreases. Therefore, a time-dependent mixing weight $\gamma(t)$ could be introduced in the model. In addition, it is also possible that the nonwild-type component distribution shifts over time. The B-spline coefficients can therefore be made time-dependent to see how this distribution evolves. The identification of time shifts could be an important trigger for public health organisations to take appropriate measures. These extensions with covariates are part of our ongoing research. This is an additional advantage over the back-fitting approach, where the inclusion of covariates was found to be less trivial. In this way, the developed model can be a nice alternative to the current practice of using standard regression models for AMR monitoring. These models rely heavily on the ECOFF and, in the case of binary data, trends above the ECOFF cannot be detected.

In summary, we can conclude by stating that the field of antimicrobial resistance testing is an important and quickly evolving area of interest. Developing tools for monitoring the MIC distributions of certain high-risk bacteria is extremely important and we believe this new method is a first promising step in that direction. Besides the inclusion of a time trend, it can also be of interest to consider multivariate models that are able to jointly estimate the MIC distribution of two or more antimicrobials to model their co-resistance patterns.

Acknowledgements. We wish to express our thanks to the "Projet d'Actions de Recherche Concertée (ARC) 27/16-039" from the "Communauté française de Belgique," granted by the "Academie universitaire de Louvain." 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. The authors are grateful to EFSA for the approval to use the *Ampicillin* data.

REFERENCES

2013/652/EU (2013). Commission Implementing Decision of 12 November 2013 on the monitoring and reporting of antimicrobial resistance in zoonotic and commensal bacteria (notified under document C(2013) 7145). Text with EEA relevance.

ANDREWS, J. M. (2001). Determination of minimum inhibition concentrations. *J. Antimicrob. Chemother.* **48** S1.5–S1.16.

ANNIS, D. H. and CRAIG, B. A. (2005). Statistical properties and inference of the antimicrobial MIC test. *Stat. Med.* **24** 3631–3644. MR2212304

ATCHADÉ, Y. F. and ROSENTHAL, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* **11** 815–828. MR2172842

AZZALINI, A. and CAPITANIO, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 367–389. MR1983753

BÖHNING, D. (1986). A vertex-exchange-method in D-optimal design theory. *Metrika* **33** 337–347. MR0868043

CRAIG, B. A. (2000). Modeling approach to diameter breakpoint determination. *Diagn. Microbiol. Infect. Dis.* **36** 193–202.

EILERS, P. H. C. (2007). Ill-posed problems with counts, the composite link model and penalized likelihood. *Stat. Model.* **7** 239–254. MR2749992

EILERS, P. H. C. and MARX, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11** 89–121. MR1435485

FINCH, R. G., GREENWOOD, D., WHITLEY, R. J. and NORRBY, S. R. (2010). *Antibiotic and Chemotherapy*. Saunders, Elsevier.

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504

IGARASHI, T., INATOMI, J., WAKE, A., TAKAMIZAWA, M., KATAYAMA, H. and IWATA, T. (1999). Failure of pre-diarrheal antibiotics to prevent hemolytic uremic syndrome in serologically proven *Escherichia coli* O157:H7 gastrointestinal infection. *J. Pediatr.* **135** 768–769.

JASPERS, S., AERTS, M., VERBEKE, G. and BELOEIL, P.-A. (2014a). Estimation of the wild-type minimum inhibitory concentration value distribution. *Stat. Med.* **33** 289–303. MR3146764

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. *Comput. Statist. Data Anal.* **71** 30–42. MR3131952

JASPERS, S., VERBEKE, G., BÖHNING, D. and AERTS, M. (2016). 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. *Biostatistics* **17** 94–107. MR3449853

KAHLMETER, G., BROWN, D. F. J., 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.

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.

LAMBERT, P. and EILERS, P. H. C. (2009). Bayesian density estimation from grouped continuous data. *Comput. Statist. Data Anal.* **53** 1388–1399. MR2657099

LANG, S. and BREZGER, A. (2004). Bayesian P-splines. *J. Comput. Graph. Statist.* **13** 183–212. MR2044877

LEE, M. L. T. and WHITMORE, G. A. (1999). Statistical inference for serial dilution assay data. *Biometrics* **55** 1215–1220.

PALUMBI, S. R. (2001). Humans as the world's greatest evolutionary force. *Science* **293** 1786–1790.

SCHELLHASE, C. and KAUERMANN, G. (2012). Density estimation and comparison with a penalized mixture approach. *Comput. Statist.* **27** 757–777. MR3041856

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. *Emerg. Infect. Dis.* **18 (5)** 741–749.

THOMPSON, R. and BAKER, R. J. (1981). Composite link functions in generalized linear models. *J. Roy. Statist. Soc. Ser. C* **30** 125–131. MR0629493

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. *Clin. Microbiol. Infect.* **12** 418–425.

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

S. JASPERS
M. AERTS
INTERUNIVERSITY INSTITUTE FOR BIOSTATISTICS AND STATISTICAL BIOINFORMATICS
HASSELT UNIVERSITY
AGORALAAN GEBOUW D
3590 DIEPENBEEK
BELGIUM
E-MAIL: stijn.jaspers@uhasselt.be

P. LAMBERT
INSTITUT DES SCIENCES HUMAINES ET SOCIALES
MÉTHODES QUANTITATIVES EN SCIENCES SOCIALES
UNIVERSITÉ DE LIÈGE
BOULEVARD DU RECTORAT 7 (B31)
B-4000 LIÈGE
BELGIUM