Estimation of an MIC distribution using a two-stage semi-parametric mixture model

Estimation of an MIC distribution using a two-stage semi-parametric mixture model

Jaspers Stijn[1], Aerts Marc[1], Verbeke Geert[2]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium
[2] Interuniversity Institute for Biostatistics and statistical Bioinformatics, KU Leuven, Leuven, Belgium

E-mail for correspondence: `stijn.jaspers@uhasselt.be`

Abstract: Antimicrobial resistance has become one of the main public health burdens of the last decades, and monitoring the development and spread of non-wild-type isolates has therefore gained increased interest. Monitoring is performed, based on the Minimum Inhibition Concentration (MIC) values, which are collected through the application of dilution experiments. A semi-parametric mixture model is presented, which is able to estimate the full continuous MIC distribution. The model is based on an extended and censored-adjusted version of the penalized mixture approach often used in density estimation. A data application and simulation study are presented in which the promising behaviour of the new method is illustrated.

Keywords: Antimicrobial resistance; Censoring; Penalized mixture approach; Semi-parametric

1 Introduction

Antimicrobial resistance (AMR) is the main undesirable side effect of antimicrobial use in both humans and animals. Due to the continuous positive selection of resistant bacterial clones, the population structure of microbial communities is modified. AMR has become one of the main public health burdens of the last decades and it is therefore extremely important to study and monitor the emergence of isolates with reduced susceptibility against antimicrobials. This may be performed by determining the minimum inhibition concentration (MIC), which is commonly measured via a broth dilution method. A standardized amount of the isolate is exposed to successive two-fold concentrations of the antimicrobial and the MIC is defined as the smallest concentration of the antimicrobial substance that inhibits the visible growth of the microorganism. Figure 1 shows an MIC distribution determined for 5190 isolates of *E. coli* tested for susceptibility against nalidixic acid. Note that, as a result of the dilution type laboratory experiments, MIC data are censored.

Our interest is in identifying the full continuous MIC distribution. In this regard, mixture models are ideally suited as they offer a natural framework for modelling unobserved population heterogeneity. In our context, a two-component mixture

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

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. The first component, representing the wild-type isolates, is assumed to be of a fixed parametric form and can hence be modelled parametrically. The second component, representing the non-wild-type isolates, is often multi-modal, and in this case, it is itself a mixture of different non-wild-type subpopulations. In order to impose as little constraints as possible, the second component will be left completely unspecified and a non-parametric estimate will be considered.

2 The semi-parametric mixture model

2.1 Estimation of the first component

Denoting by Y_i the number of times MIC value i was observed over the n trials, the observed MIC groupings can be seen as possible outcomes of $Y = (Y_1, \ldots, Y_k) \sim \text{Mult}(n,p)$, where k represents the number of different MIC categories and $p = (p_1, \ldots, p_k)$ such that $p_1 + \ldots + p_k = 1$. The maximum likelihood estimates for the multinomial probabilities p_i are just the observed relative frequencies $\frac{Y_i}{n}$. Nevertheless, the main interest remains in identifying the most suited parameters of the continuous wild-type distribution rather than those of the discrete multinomial distribution. This can be achieved by exploiting the fact that the observed groupings are actually the result of the censored readings of the dilution experiment. Hence the multinomial probabilities corresponding to a certain outcome i can be rewritten as

$$\begin{cases} \tilde{p}_i = \pi * F(u_i;\theta) \text{ if } i = 1, \\ \tilde{p}_i = \pi * [F(u_i;\theta) - F(l_i;\theta)] \text{ otherwise,} \end{cases} \tag{2}$$

where u_i and l_i are the respective upper and lower values of the ith MIC category and $F(.)$ represents the cumulative distribution function under consideration, with θ its corresponding parameters. In addition, the unknown parameter π accounts for the fact that the true MIC distribution is a mixture of the wild-type and non-wild-type component.
The idea is to replace an increasing number of the multinomial probabilities with their parametric counterparts in (2). The probabilities of the remaining outcomes are left unchanged and are thus to be estimated similar to

those of the saturated model (i.e. the observed relative frequencies). The resulting sequence of likelihoods is specified by

$$l_j(\tilde{p}_1, \ldots, \tilde{p}_{k_j}, p_{k_j+1}, \ldots, p_k) = \sum_{i=1}^{k_j} y_i \log \tilde{p}_i + \sum_{i=k_j+1}^{k} y_i \log p_i, \qquad (3)$$

with $j = 1, \ldots, k-2$ and where k_j indicates how many of the original multinomial probabilities are replaced: $k_j = j$ + length of ϕ, with $\phi = (\theta, \pi)$. This sequence can be maximized to obtain several proposal estimates for the parameters of interest. Note that as a result of the parametric assumption, less parameters are used in the construction of the likelihood when j increases. The AIC criterion can be applied to select the most appropriate parameter estimates or averaged estimates can be considered using the Akaike weights.

2.2 Estimation of the second component

The second component will be estimated using a censored-adjusted version of the penalized mixture approach by Schellhase and Kauermann (2012). Let X denote the univariate random variable of interest (i.e. the MIC value), with true density function f. The main idea is to approximate f as a mixture of densities:

$$f_K(x) = \sum_{k=-K}^{K} c_k \phi_k(x), \qquad (4)$$

where the $\phi_k(x)$ are the basis densities and the c_k are called the weights. In order to avoid constrained maximization, the weights are reparametrized:

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

with $\beta_0 \equiv 0$ for identifiability. The basis densities are assumed to be Gaussian density functions, which are located at a fixed number of knots, corresponding to their respective means.
The number of knots plays an important role in terms of bias and variance. A compromise between smoothness and unbiasedness is obtained through the approach of Eilers and Marx(1996): a large number of basis functions is considered, but the log-likelihood is penalized for overfitting via a penalty term based on the finite differences of adjacent coefficients.
Assuming an independent sample $x_i, i = 1, \ldots, n$, the final log-likelihood to be optimized can be written as

$$l_p(\beta, \lambda) = \sum_{i=1}^{n} \log \sum_{k=-K}^{K} c_k \phi_k(x_i) - \frac{1}{2} \lambda \beta^T D_m \beta,$$

where D_m is the penalty matrix and λ is the smoothing parameter. In order to obtain estimates for the β parameters, Newton-Raphson scoring is performed, while the penalty parameter λ is updated using an estimating equation. For further details, see Schellhase and Kauermann (2012).
In order to take the censoring into account, the original basis density functions are replaced by their corresponding distribution functions:

$$l(\beta) = \sum_{i=1}^{n} \log \sum_{k=-K}^{K} [c_k \Phi_k(x_i) I(x_i \in MIC_{min}) + c_k \{\Phi_k(x_i) - \Phi_k(x_i - 1)\} I(x_i \notin MIC_{min})] .$$

Penalization and optimization of the likelihood are done similar to the original procedure.

2.3 The semi-parametric mixture model

The idea is to fix f_1 to the estimate obtained in the initial phase, using the method in Section 2.1. Information on the second component is then introduced through the censored-adjusted penalized mixture approach. More specifically, the estimator for the density of the MIC values is based on (4), to which one additional component is added:

$$f_K(x) = \pi f_1(x; \theta_1) + (1 - \pi) \sum_{k=-K}^{K} c_k \phi_k(x) = \sum_{k=-(K+1)}^{K} \tilde{c}_k \tilde{\phi}_k(x). \qquad (5)$$

The additional component represents the wild-type component and will not be penalized as it is assumed to be fixed. Regarding the placement of the knots for the second component, recall that the model based on the likelihood in (3) is fitted to increasing subsets of the data. The fit with the smallest AIC value identifies the subset of the data that most likely belongs to the wild type component. In addition, due to the interval censoring, the highest MIC value in this subset identifies a possible lower bound for the MIC values of the non-wild-type isolates and will hence be used as the first knot of the basis. Finally, the estimator in (5) is used to construct the penalized likelihood. The adjustment for censored observations and the optimization occur in full similarity as above.

3 Application to real data

The two-stage procedure described in Section 2 is applied to MIC data obtained from the EUCAST website. The data concern the susceptibility of *E. coli* against nalidixic acid. Both a log-normal and a gamma distribution were assumed for the first component. The optimal mean and standard

deviation for the former were (on the $\log_2$-scale) 1.04 (se = 0.01) and 0.58 (se = 0.01), respectively. In case of a gamma first component, the shape and scale were 8.32 (se = 0.31) and 0.25 (se = 0.01), respectively. Figure 1 shows the result of the semi-parametric mixture model. The fixed gamma first component results into the lowest AIC and the estimated mixing weight ($\hat{\pi}$) is 0.87 (se = 0.01).

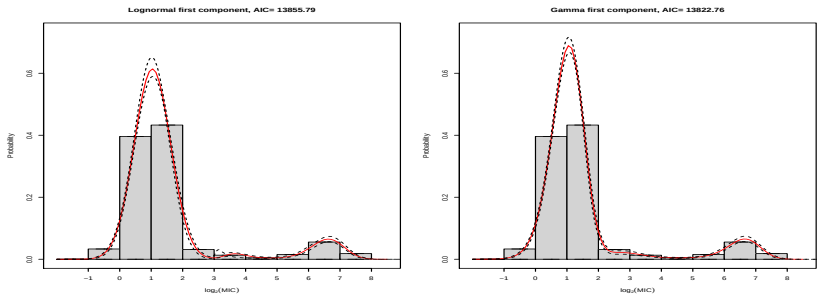

FIGURE 1. Barplot and estimates of the distribution of Minimum Inhibitory Concentrations (MIC) in *E. coli* isolates tested for susceptibility against nalidixic acid - source: EUCAST website.

4 Simulation study

Samples are taken from a mixture distribution with two main components. 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. The prevalence of wild-type isolates is taken to be 0.6. The multinomial based method is compared to the semi-parametric mixture model, assuming both a log-normal and gamma first component. The performance of the methods are compared based on the MSE values for the estimate of the prevalence of wild-type isolates. In addition, the Kullback-Leibler distance indicates the performance of the semi-parametric mixture model when estimating the entire mixture density. The considered sample sizes are 500, 1000 and 5000.

Since in real-life applications, the true underlying distribution of the first component is unknown, the averaged estimates of the two approaches should be regarded. From Table 1, it is seen that the semi-parametric mixture model outperforms the multinomial based method when estimating the prevalence of wild-type isolates. This is most pronounced in case of the larger sample size. The KL distance also indicates a promising behaviour of the semi-parametric mixture model.

TABLE 1. Summary of the simulation study. Presented are the averaged Kullback-Leibler distance (KL) and the MSE values when estimating π using the multinomial-based method (MSE_1) and the semi-parametric-mixture model (MSE_2), when assuming a log-normal and gamma first component, as well as for the most optimal fit using AIC.

Sample size	Assumed f_1	KL (s.e.)	MSE_1	MSE_2
500	Log-normal	0.022 (0.008)	0.0027	0.0020
	Gamma	0.023 (0.011)	0.0021	0.0022
	AIC	0.022 (0.008)	0.0028	0.0023
	Averaged	-	0.0017	0.0015
1000	Log-normal	0.012 (0.005)	0.0014	0.0010
	Gamma	0.015 (0.006)	0.0020	0.0024
	AIC	0.013 (0.005)	0.0016	0.0013
	Averaged	-	0.0010	0.0009
5000	Log-normal	0.004 (0.001)	0.0002	0.0002
	Gamma	0.009 (0.002)	0.0075	0.0094
	AIC	0.004 (0.001)	0.0018	0.0003
	Averaged	-	0.0014	0.0003

5 Discussion

A two-stage semi-parametric mixture model to estimate a continuous MIC distribution from censored observations was presented. In addition to an estimate for the prevalence of wild-type isolates, the model also provides an estimate for the non-wild-type distribution. Regarding the susceptibility of *E. coli* isolates against nalidixic acid, the prevalence of wild-type isolates was estimated to be 0.87 (0.01). Finally, a simulation study indicated a promising behaviour of the new method. Our ongoing research includes the simultaneous estimation of the first and second component.

Acknowledgments: Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged. The research of the first author was supported by the Research Foundation Flanders (FWO), grant 11E2913N.

References

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

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