Chapter 3

Using Akaike's Information Theoretic Criterion in Mixed-Effects Modeling of Pharmacokinetic Data: A Simulation Study*

We first define population data as a set of one or more measurements in two or more individuals *e.g.*, patients, volunteers, or animals). Such data may be characterized by mixed-effects models, where the mixed effects consist of fixed and random effects. Fixed effects (or fixed effect factors) are, for example, the times at which the measurements are obtained, and covariates such as demographic characteristics of the individuals. Due to random effects (or random effect factors), the model output may vary between measurements, and between individuals. When mixed-effects models are fitted to population data, the question arises as to how many of those effects should be incorporated in the model. This is the so-called problem of variable selection.[40]

One strategy is to observe the change in goodness-of-fit by adding one more parameter and test the significance of that change.[16] In the maximum likelihood approach, the objective function value (OFV), being the minus two logarithm of the likelihood function, is minimized. To attain a p-value of *e.g.*, 0.05 or less, the decrease in OFV, when adding one parameter, should be 3.84 or more.[16]

Another strategy is to apply Akaike's information theoretic criterion (AIC), which can be written as

$$\text{AIC} = \text{OFV} + 2 \cdot D, \tag{3.1}$$

where D is the number of parameters in the model.[1,16,18,40] The model with the lowest value of AIC is considered the best one. In the case of just adding one parameter, the OFV needs to decrease only 2 points or more to be incorporated in the model, so the associated p-value > 0.05 seems too high to justify this strategy.

When additional model parameters are incorporated, the significance of one model parameter might change, but the interpretation of AIC does not.[18] However, when multiple significance tests are performed, the significance level of each individual test should be corrected to a lower value, so a decrease of 2 points for one parameter does again seem to be too low.

[*] E Olofsen, A Dahan, *F1000Research* 2015; 2:71

Even if the strategy of using AIC leads to optimal variable selection, the question arises as to whether this is also the case when using mixed-effects models. In theory, the model that is best according to AIC is the one that minimizes prediction error[1,96] and this is also true for a mixed-effects model when predicting data for individuals for which no data have been obtained so far.[96]

In the literature, many simulation studies have assessed the performance of AIC, but to our knowledge these were never done in selecting the model with minimal prediction error for population data. In this article, we will define a toy pharmacokinetic model and observe the performance of AIC when adding fixed effects to this model, as well as when adding interindividual variability.

3.1 Methods

3.1.1 A Hypothetical Pharmacokinetic Model

Consider the following function $y(t)$, an infinite sum of exponentials, and its relationship with a (negative) power of time:[57]

$$y(t) = \int_0^\infty \exp(-\lambda t)d\lambda = -\frac{1}{t}\exp(-\lambda t)\Big|_0^\infty = \frac{1}{t} \text{ for } t > 0. \tag{3.2}$$

Figure 3.1A shows that this function looks like a typical pharmacokinetic profile after bolus administration. This model is to be regarded as a toy model, because we do not expect it to adequately describe pharmacokinetic data, although variations of power functions of time have been shown to fit pharmacokinetic data well.[57] We approximate $y(t) = 1/t$ by the following sum of M exponentials:

$$\hat{y}(t_j; \boldsymbol{\alpha}, \boldsymbol{\lambda}) = \sum_{m=1}^{M} \alpha_m \exp(-\lambda_m t_j). \tag{3.3}$$

The M parameters $\boldsymbol{\lambda}$ and measurement time instants t_j are fixed and are set to have distinct values as described in the next subsections. The coefficients $\boldsymbol{\alpha}$ (related to how a drug dose is distributed across compartments) are parameters to be estimated. Let the number of α_m that are not fixed to zero be denoted by K. Then the above approximation has the property that while the fits of models to the data would improve with increasing K, we would need no less than $K = M$ exponentials to obtain a perfect fit. Moreover, with noisy data, it might be that for $K < M$ an optimal fit is obtained in the sense that then the associated prediction error of the model is minimal. Figure 3.1B shows how eleven (in this case error-free) samples from this function can be approximated by sums of exponentials.

3.1.2 Individual Data Modeling and Simulation

In the following, the time instants t_j, $j = 1, \cdots, M$, centered around 1, were chosen within $[1/t_{\max}, t_{\max}]$ according to

$$t_j = \left(\frac{j}{M+1-j}\right)^{\gamma}, \tag{3.4}$$

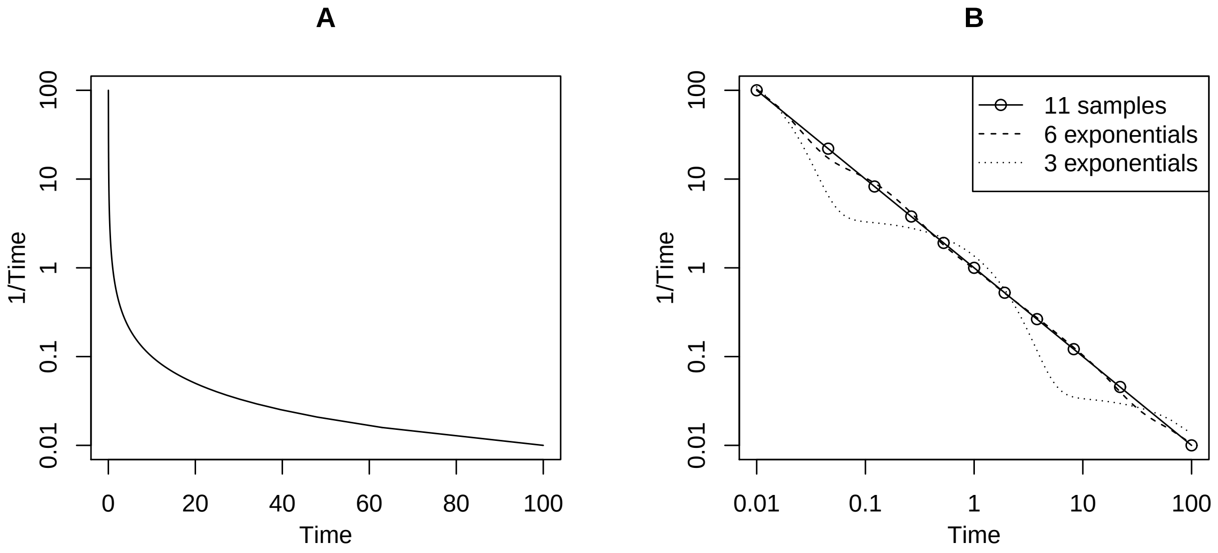

Figure 3.1: A: function $y(t) = 1/t$, and B: approximations obtained by fitting six and three exponentials to the depicted eleven samples. Note the log-lin and log-log scales for panels A and B, respectively. Time has arbitrary units.

with $\gamma = \log(t_{\max})/\log(M)$; $t_{\max}$ was set to 100 (see the time axis of Figure 3.1B for an example with $M = 11$). Simulated data with constant proportional error were generated *via*

$$y(t_j) = \frac{1}{t_j}(1 + \epsilon_j), \tag{3.5}$$

where ϵ_j denotes Gaussian measurement noise with variance σ^2. The M time constants $\boldsymbol{\lambda}$ were fixed according to $\lambda_m = 1/t_m$, $m = 1, \cdots, M$. In this setting the model eq. (3.3) can be fitted to simulated data using weighted linear least squares regression, with weight factors $w(t_j) = 1/t_j$ (note that no precaution is needed against $\epsilon \leq -1$). Linear least squares regression is very fast and robust, so it allows for the evaluation of many simulation scenarios.

3.1.3 Population Data Modeling and Simulation

Population data consisting of N individuals were simulated *via*

$$y_i(t_j) = \frac{1}{t_j} \cdot (\exp(\eta_i) + \epsilon_{ij}) \text{ with } i = 1, \cdots, N, \tag{3.6}$$

where η_i denotes interindividual variability with variance ω^2. The random effect η_i influences the overall magnitude of the values of y_i, but not the shape of the function in time, so this is similar to a random effect that influences pharmacokinetic volume of distribution. The nonlinear mixed-effects model for the population data was then written as:

$$\hat{y}_i(t_j;\boldsymbol{\alpha},\boldsymbol{\lambda}) = \sum_{m=1}^{M} \alpha_m \exp(-\lambda_m t_j + \eta_i). \tag{3.7}$$

Note that with $N > 1$, a perfect fit is no longer obtained with $K = M$ nonzero coefficients $\boldsymbol{\alpha}$, because the ϵ_{ij} are generally different for different i (individuals). Just one different η_i for each individual i cannot compensate for M different ϵ_{ij}.

3.1.4 Statistical Analysis

Simulation data were generated *via* eq. (3.6) in R.[85] Model fitting was also done in R, with function “lm()” from package “stats”, except for nonlinear mixed-effects model fitting for simulated data with $\omega^2 > 0$, which was done in NONMEM version 7.3.0.[9] Parameters $\boldsymbol{\alpha}$ (see eq. (3.7)) were either fixed to zero or free. Although the $\boldsymbol{\alpha}$ are expected to be positive with pharmacokinetic data, they were not constrained to be positive. So it was not possible for parameters to become essentially fixed to zero due to that constraint, which would reduce the dimensionality of the model. Prediction error (ν^2) was calculated with

$$\nu^2 = \frac{1}{N \cdot M} \sum_{i=1}^{N} \sum_{j=1}^{M} \left(\frac{z_i(t_j) - \hat{y}_i(t_j)}{w(t_j)} \right)^2, \tag{3.8}$$

using predictions based on eq. (3.7) with the random effects $\eta_i = 0$. Validation data $z_i(t_j)$ were also generated *via* eq. (3.6), but with different realizations of ϵ_{ij} and η_i. Error terms weighted with $w(t_j) = 1/t_j$ are homoscedastic, which is an assumption underlying regression analysis and allows for the interpretation of ν^2 as independent of time. The objective function OFV was also calculated at the estimated parameters using the validation data, denoted OFV$_v$, which should on average be approximately equal to Akaike’s criterion (see Supplementary material). OFV$_v$ was compared with AIC and also with Akaike’s criterion with a correction for small sample sizes (AIC$_c$[18])

$$\text{AIC}_c = \text{OFV} + 2 \cdot D \cdot \left(1 + \frac{D+1}{N \cdot M - D - 1}\right). \tag{3.9}$$

The above criteria were normalized by dividing them by the number of observations (see Supplementary material for motivation), and averaged over 1000 runs (unless otherwise stated; and runs where NONMEM’s minimization was not successful were excluded). For plotting purposes, 95% confidence intervals or confidence regions for means were determined using R’s packages “gplots” and “car”, under the assumption that averages over 1000 variables are normally distributed. Model selection frequencies were calculated based on optimal models according to AIC$_c$ as determined for each simulation data set.

3.1.5 Selection of Parameter Values

Simulation parameters M and σ^2 are expected to determine the number of exponentials K; if the number of measurements M increases and/or the measurement error σ^2 decreases, K will increase. Without interindividual variance, so $\omega^2 = 0$, the information in the data increases as N increases, so also in that case K is expected to increase. With $N = 2$, $M = 11$ and $\sigma^2 = 0.5$, pilot simulations indicated a $K \approx 4$. When $\omega^2 > 0$, prediction error will increase, but it is less easy to predict what its effect will be on K. For ω^2

Table 3.1: Selecting $K = 1, \cdots, M = 11$ evenly spaced rate constants from $\boldsymbol{\lambda}$: 0 and 1 denote α_m to be fixed to zero, and a free parameter to be estimated, respectively (see eq. (3.7)).

K	$m:1$	2	3	4	5	6	7	8	9	10	11
1	0	0	0	0	0	1	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	1
3	1	0	0	0	0	1	0	0	0	0	1
4	1	0	0	1	0	0	0	1	0	0	1
5	1	0	0	1	0	1	0	1	0	0	1
6	1	0	1	0	1	0	1	0	1	0	1
7	1	0	1	1	0	1	0	1	1	0	1
8	1	1	0	1	1	0	1	1	0	1	1
9	1	1	0	1	1	1	1	1	0	1	1
10	1	1	1	1	1	0	1	1	1	1	1
11	1	1	1	1	1	1	1	1	1	1	1

values of 0, 0.1, and 0.5 were selected - values that are encountered in practice. Because there is only one random effect in the mixed-effects model, the relatively low number of individuals $N = 5$ was selected.

For a certain choice of M, there are $2^M - 1$ possible combinations of λs to choose for the terms $\exp(-\lambda_m t_j)$ in the sum of exponentials (excluding the case of a model without exponentials). Because accurate evaluation of all models at different parameter values is not feasible with respect to computer time, the set of possible combinations was reduced to one with evenly spaced λs. Table 3.1 gives an example for the case $M = 11$.

3.2 Results

Figure 3.2 shows the averaged prediction error *versus* number of exponentials for all possible choices of $\boldsymbol{\lambda}$, with $N = 2$, $M = 11$, $\sigma^2 = 0.5$, and $\omega^2 = 0$. From the figure it is clear that prediction error may indeed increase if the number of exponentials selected is too large. The bigger solid circles correspond to the models chosen in Table 3.1; in general the evenly spaced selection of exponents resulted in models with smallest prediction error.

Figure 3.3 shows simulation results using the model set defined in Table 3.1, starting from $K = 4$, with parameters $N = 5$, $M = 11$, $\sigma^2 = 0.5$, and $\omega^2 = 0$. The model with $K = 6$ exponentials had minimal mean AIC$_c$, and also minimal mean OFV$_\nu$ and minimal mean squared prediction error (ν^2). With $N = 5, M = 11$, there are still visible differences between AIC$_c$ and AIC; although AIC would in this case also select the optimal model, AIC appears to favor more complex models. Note that the sizes of the confidence intervals and confidence regions can be made arbitrarily small by choosing the number of runs to be higher than the selected number of 1000 (at the expense of computer time).

Figure 3.4 shows simulation results with $\omega^2 = 0.1$; mixed-effects analysis was used

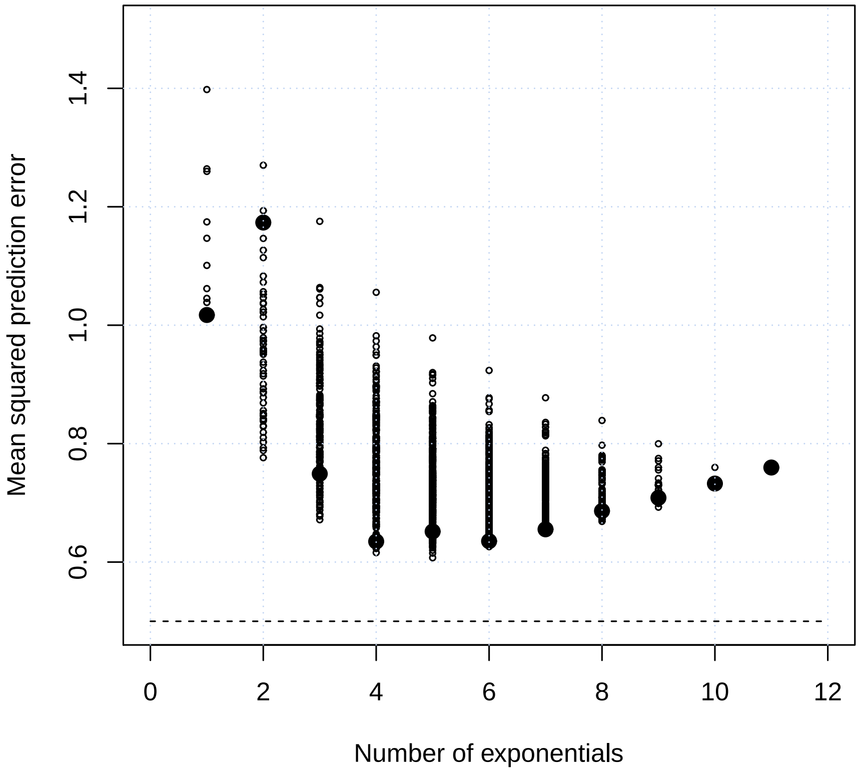

Figure 3.2: Mean squared prediction error ν^2 (eq. (3.8)) as a function of the number of exponentials, with 2047 models, averaged over 100 runs, $N = 2$, $M = 11$, $\sigma^2 = 0.5$, $\omega^2 = 0$. The dashed line represents the prediction error from the true model, so that $\nu^2 = \sigma^2$. The bigger solid circles correspond to the models chosen in Table 3.1.

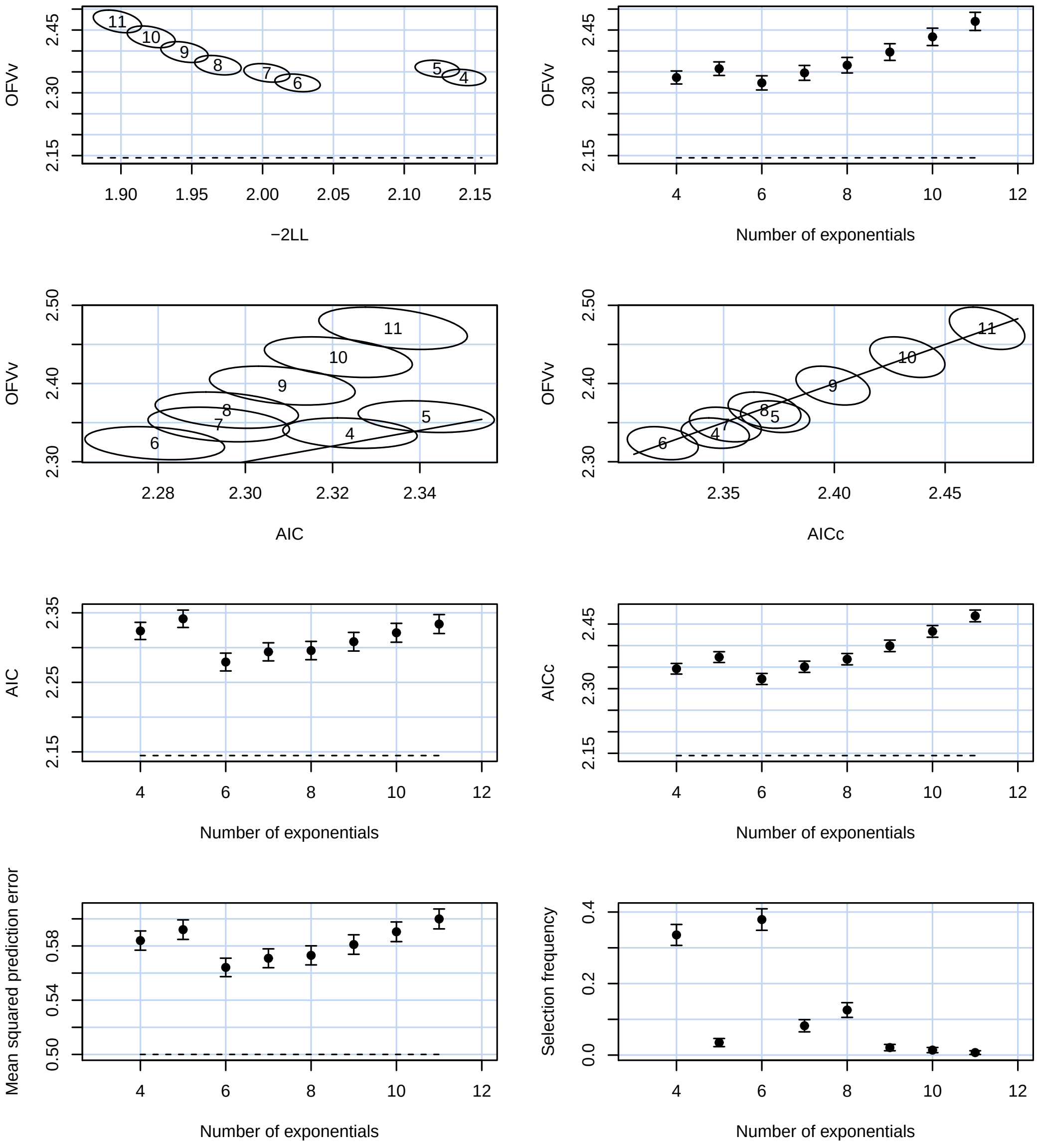

Figure 3.3: Mean OFV$_v$ as a function minus of two log-likelihood (-2LL), the number of exponentials, AIC and AIC$_c$ (top four panels), and AIC, AIC$_c$, prediction error v^2, and model selection frequencies as a function of the number of exponentials (lower four panels), averaged over 1000 runs, $N = 5$, $M = 11$, $\sigma^2 = 0.5$, $\omega^2 = 0$. The dashed lines represent the theoretical values for an infinite amount of data (see Appendix). Error bars and ellipses denote 95% confidence intervals and confidence regions, respectively. The numbers in the confidence regions denote the number of exponentials. The solid lines in the middle upper panels are lines of identity.

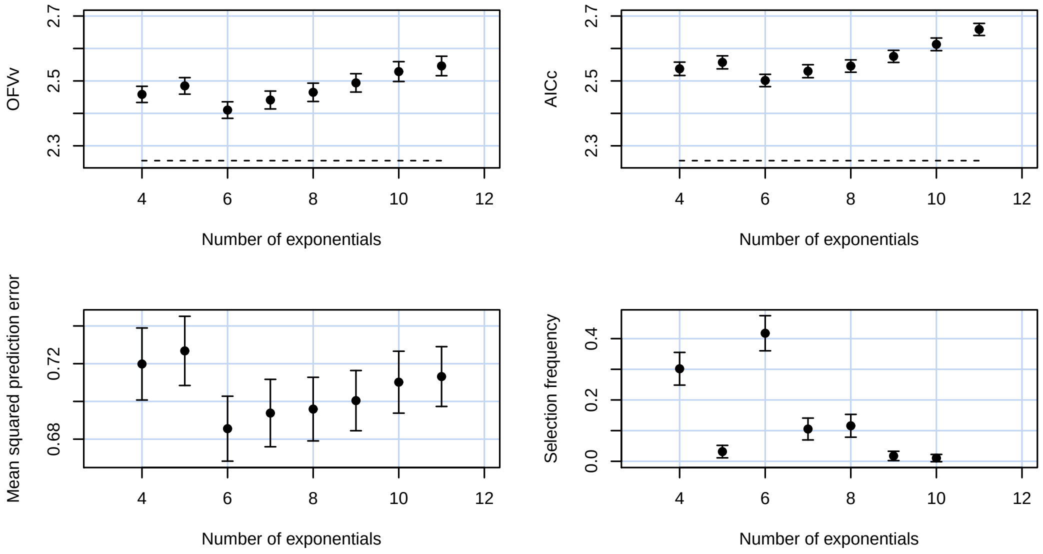

Figure 3.4: Mean OFV$_{\upsilon}$, AIC$_c$ prediction error ν^2, and model selection frequencies as a function of the number of exponentials, for $\omega^2 = 0.1$; parameters otherwise identical to those for Figure 3.3.

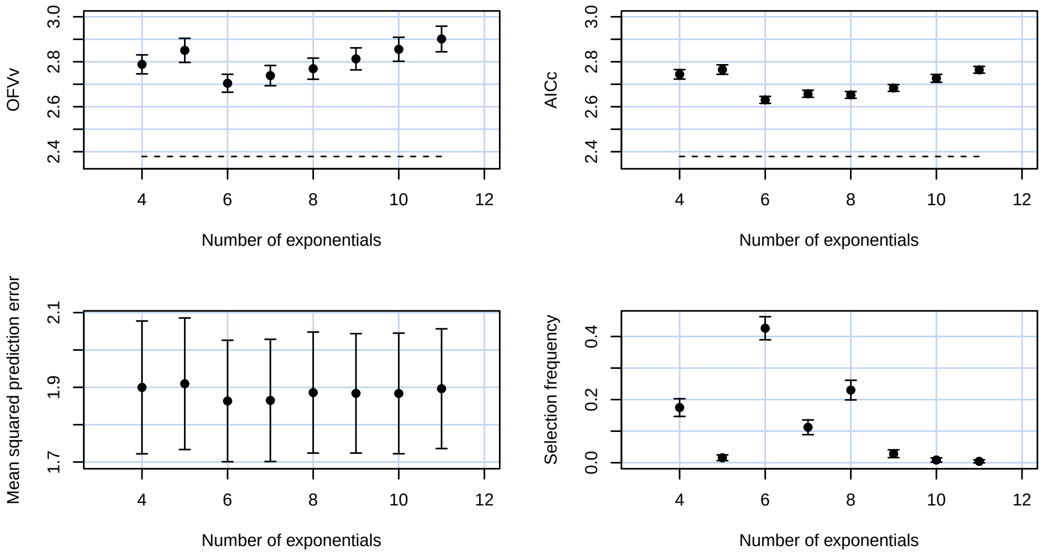

Figure 3.5: Mean OFV$_{\upsilon}$, AIC$_c$ prediction error ν^2, and model selection frequencies as a function of the number of exponentials, for $\omega^2 = 0.5$; parameters otherwise identical to those for Figure 3.3.

to fit the population data. The main difference with the results of data with $\omega^2 = 0$ is the overall increase in OFV_v and AIC_c. The optimal number of exponentials remained $K = 6$.

Figure 3.5 shows simulation results with ω^2 set at the higher value of 0.5. The main differences with the results of data with $\omega^2 = 0.1$ are again the overall increase in OFV_v, AIC_c and prediction error, and also in the variability in the prediction error. The optimal number of exponentials remained $K = 6$, although AIC_c begins to favor the models with larger K (a simulation with N increased to 7, both OFV_v and AIC_c favored larger models; data not shown).

3.3 Discussion

With the objective of creating a simulation context resembling pharmacokinetic analysis where concentration data are approximated by a sum of exponentials, the toy model $y(t) = 1/t$ was chosen. In this setting, reality - the reality of the toy model - is always underfitted. When mixed-effects models were fitted to the simulated data, mean AIC_c was approximately equal to the validation criterion mean OFV_v. The minima of mean AIC_c and mean OFV_v coincided. With large interindividual variability, mean expected prediction error (ν^2, see eq. (3.8)), with random effects fixed to zero), was less discriminative between models, so that it becomes less suitable as a validation criterion; it does not take into account whether estimated interindividual variability matches the variability in the validation data.

3.3.1 Akaike's *versus* the Conditional Akaike Information Criterion

Vaida and Blanchard proposed a conditional Akaike information criterion to be used in model selection for the "cluster focus".[96] It is important to stress that their definition of cluster focus is the situation where data are to be predicted of a cluster that was also used to build the predictive model. In that case, the random effects have been estimated, and then the question arises how many parameters that required. In our situation, a cluster is the data from an individual; AIC was used in the situation of predicting population data consisting of individual data that were not used to build the model. This would seem to be the most common situation in clinical practice. Furthermore, AIC for the population focus is asymptotically equivalent with leave-one-individual-out cross-validation; AIC for the individual focus with leave-one-observation-out cross-validation.[31]

3.3.2 Akaike's *versus* the Bayesian Akaike Information Criterion

We chose to perform simulations using the model given by eq. (3.2) because approximating data with a sum of exponentials is daily practice in pharmacokinetic analysis where data are obtained from "infinitely complex" systems, and we cannot hope to find the "correct" model. The Bayesian information criterion (BIC) is consistent in the sense that it selects the correct model, given an infinite amount of data.[18] The reason that AIC can be used in "real-life" problems is that as the amount of data goes to infinity, the complexity, or dimension, of the model that should be applied should also go infinity.[19] Burnham and Anderson show that it is possible to choose the prior for BIC in such a way that it incorporates the knowledge that more complex models should be favored if the amount of data increases, and so that the BIC "reduces" to AIC.[18,19] In the situation that the correct model set belongs to the set of evaluated models, a selection criterion that both finds the correct model and minimizes prediction error would be preferable - but Yang concluded that this may not be possible.[98]

3.3.3 Model Selection Criterion AIC and Predictive Performance

It should be noted that minimizing AIC has a more general interpretation than just minimizing prediction error ν^2 as given for example by eq. (3.8). The interpretation of minimizing AIC is minimizing the difference between the the information contained in the data and captured by the model.[18] Independent or future population data z are not just predicted by $\hat{y}$; also the distributions of the expected random effects ϵ and η are characterized by $\hat{\sigma}^2$ and $\hat{\omega}^2$. That is why OFV_ν (and not ν^2) is the criterion to be used to assess the predictive performance of a model.

3.3.4 Regression Weights as Functions of the Model Output

The simulated data were analyzed using weighted (non)linear regression, see eq. (3.6)), where measurement noise was weighted according to the exact function value. In practice, when the weights are unknown, the model output may be used to weight the data. In that case simulated data should be generated (*cf.* eq. (3.6)) *via*

$$y_i(t_j) = \frac{1}{t_j} \cdot \exp(\eta_i) \cdot (1 + \epsilon_{ij}). \tag{3.10}$$

The likelihood function and AIC are both still well-defined if the model output $\hat{y}_i(t_j) \neq 0$. Prediction errors are to be calculated with

$$\nu^2 = \frac{1}{N \cdot M} \sum_{i=1}^{N} \sum_{j=1}^{M} \left(\frac{z_i(t_j) - \hat{y}_i(t_j)}{\hat{y}_i(t_j)} \right)^2 , \tag{3.11}$$

where where $\hat{y}$ possibly becomes arbitrarily close to zero for less than optimal models, and ν^2 may be based on long-tailed distributed numbers. To be able to compare prediction errors from different models, the weight factors could be chosen identical for all K to the model output of the largest model - see Appendix for further analysis.

3.3.5 Model Selection Uncertainty

Theoretically, and in the discussed simulations, minimum mean AIC is related to best mean predictive performance, where the mean is taken across multiple studies and prespecified models. This holds independent of the number of models. However, in practice, we have data from one study and the task of specifying the models to consider. As soon as there is more than one model, there is a nonzero probability that the model selected based on AIC would have, on average, a larger prediction error than the optimal one. Also, if we were able to repeat the study, the average prediction error based on the models with minimum AIC would be larger than optimal. With many models, model selection is called unstable in the sense that each time a study is repeated it would lead to the selection of another model.

The figure panels with the model selection frequencies (Figure 3.3, Figure 3.4, and Figure 3.5) show: 1) there is a relationship between the model with highest selection probability and minimum mean prediction error, but this relationship is not one-to-one; 2) there can be an almost as large selection probability for a model that is not associated with minimum mean prediction error; but 3) in that case, their minimum mean prediction errors are comparable.

Models with equal mean predictive properties may have different properties in different extrapolation scenarios. Model averaging,[18] where model parameters or their predictions are averaged, reduces model selection instability and hence may be used to avoid model specific inference which discards model selection uncertainty. Data dredging[18] refers to the situation where there is an increasingly large set of models which are not prespecified. At the point the data dredging is stopped (by the investigator, or by the computer), the best model is at high risk to fit only the data at hand, and hence cannot be used for prediction.[21]

3.3.6 Limitations of the Study

We recognize the following limitations of our study:

- The simulation model contained only one random effect to describe interindividual variability, and therefore the number of random effect (co)variances was fixed to one in the model set used for fitting. While the number of (co)variance parameters should be counted as ordinary parameters,[96] at least in well behaved situations,[36] we did not investigate the process of optimizing this part of a random effects model.

- The nonlinearity in the mixed-effects model was simply due to a multiplicative factor $\exp(\eta)$ in the model output. Usually, random effects in pharmacokinetic models have more complex influence on the model output. However, the lognormal nature of $\exp(\eta)$ is a characteristic property of both our toy model and general pharmacokinetic models.

- The characteristics of the exponentials incorporated in the regression models were evenly spaced, and the values of the rate constants $\boldsymbol{\lambda}$ were fixed. We expect that with more freedom in the specification of the set of models, prediction errors with overfitted models may be worse. However, the agreement between AIC_c and prediction error should persist.

- We did not evaluate all possible models within their definition, but only those listed in Table 3.1, and it makes sense to limit the model set to reduce model selection instability.[18,98] We did not address how to optimally select the rate constants $\boldsymbol{\lambda}$. Stepwise selection methods have their disadvantages.[83] With stepwise forward selection, AIC_c may even perform worse than AIC.[60]

- We did not evaluate the process of covariate selection. However, the set of exponentials may be viewed as a number of (somewhat correlated) predictors. It is therefore expected that the present findings also hold for other types of covariates.

3.4 Conclusion

In conclusion, the present simulation study demonstrated that, at least in a relatively simple mixed-effects modeling context with a set of prespecified models, minimum mean AIC$_c$ coincided with best predictive performance, also in the presence of interindividual variability.

Acknowledgment: The authors would like to thank J. de Goede for many fruitful discussions.

3.A Appendix: Supplementary Material

In the following, we summarize theory on the maximum likelihood approach and AIC relevant for this paper. We start with the situation for data from one individual and show how AIC is related to OFV$_v$. Subsequently we discuss the situation for population data.

Suppose the model for measured data y_j, $j = 1, \cdots, M$ is given by (*cf.* eqs ((3.5), (3.6), and (3.10))

$$y_j = \hat{y}_j + w_j \cdot \epsilon_j,$$

where $\hat{y}_j$ is the model output, w_j are weight factors, and ϵ_j are independent normally distributed with mean zero and variance σ^2. The likelihood function L for this data set is then given by

$$L(y;\boldsymbol{\theta}) = \prod_{j=1}^{M} \frac{1}{w_j \sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_j - \hat{y}_j}{w_j \sigma}\right)^2\right], \tag{3.12}$$

where the set of parameters $\boldsymbol{\theta}$ contains σ^2 and those needed to calculate $\hat{y}$. The objective function value (OFV) is defined as minus two times the natural logarithm of the likelihood:

$$\text{OFV} = -2\log(L(y;\boldsymbol{\theta}) = \sum_{j=1}^{M} \log(w_j^2) + M\log(\sigma^2) + M\log(2\pi) + \frac{1}{\sigma^2}\sum_{j=1}^{M}\left(\frac{y_j - \hat{y}_j}{w_j}\right)^2. \tag{3.13}$$

Note that in writing “OFV”, the data and parameters it depends on have been omitted. Now maximum likelihood is obtained when OFV is minimal; constant terms such as $M\log(2\pi)$ may then be discarded (for example, in NONMEM’s calculation of the the objective function). The minimum is attained for certain values of parameters of $\hat{y}$, and for the parameter value of σ^2, when the derivative of OFV with respect to that parameter is zero:

$$\frac{\partial \text{OFV}}{\partial \sigma^2} = \frac{M}{\sigma^2} - \frac{1}{(\sigma^2)^2}\sum_{j=1}^{M}\left(\frac{y_j - \hat{y}_j}{w_j}\right)^2 = 0,$$

so the maximum likelihood estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{y_j - \hat{y}_j}{w_j}\right)^2 .$$

By subsituting this estimate in eq. (3.13), we obtain

$$\mathrm{OFV} = \sum_{j=1}^{M}\log(w_j^2) + M\log(\hat{\sigma}^2) + M\ \log(2\pi) + M. \tag{3.14}$$

By substituting this result in eq. (3.1), we have

$$\mathrm{AIC} = \sum_{j=1}^{M}\log(w_j^2) + M\log(\hat{\sigma}^2) + M\ \log(2\pi) + M + 2D.$$

The term $2D$ arises from the fact that in minimizing the Kullback-Leibler information, *i.e.*, a measure of the distance between reality and the best approximating model, expectations have to be taken over a data space leading to estimates of parameters $\boldsymbol{\theta}$ (and hence $\hat{y}$, and possibly w (see below)) and over a second independent data space y.[18] So AIC as defined above should on average be approximately equal the value of OFV (eq. (3.13)), with estimated values for the parameters and validation data z_j denoted OFV_v:

$$\mathrm{OFV}_v = \sum_{j=1}^{M}\log(w_j^2) + M\log(\hat{\sigma}^2) + M\log(2\pi) + \frac{1}{\hat{\sigma}^2}\cdot\sum_{j=1}^{M}\left(\frac{z_j - \hat{y}_j}{w_j}\right)^2 . \tag{3.15}$$

So when OFV and AIC are both minimized, the latter term - the sum of squared weighted prediction errors - should also be minimal. For the plots in this paper, the measures OFV, OFV_v, AIC, and AIC_c, were normalized by dividing them by the number of data samples. With an infinite amount of data, and $\hat{\sigma}^2 = \sigma^2$, the normalized criteria should attain the value of $\log(\sigma^2) + \log(2\pi) + 1$.

Note that if the weights w_j are taken as in subsection “Data simulation”, the term $\sum\log(w_j^2)$ vanishes (this is a just a curiosity of that choice of weights); if the w_j are taken as the measurements y_j, the expectation of this term is the same for every K (for every model considered here). However, if the weights are taken as the model output $\hat{y}_j$, the expectation of the term will not vanish for a less than perfect model, and will differ between different models. To compare their ν^2, the weights for all models could be fixed to the model output of the best model - but since that is unknown at this point - to the output of the largest model.

For population data, the likelihood function is the product across individual marginal likelihoods where the random effects have been integrated out. For one individual i, and the model given by eq. (3.6), the likelihood L_i is

$$L_i = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^M \cdot \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\sum_{j=1}^{M}\left(\frac{\exp(\eta_i) + \epsilon_{ij} - \exp(\eta')}{\sigma}\right)^2\right]\cdot\frac{1}{\omega\sqrt{2\pi}}\cdot\exp\left[-\frac{1}{2}\left(\frac{\eta'}{\omega}\right)^2\right]d\eta'.$$

The ϵ_{ij} have on average mean zero and variance σ^2, and NONMEM’s first-order condi-

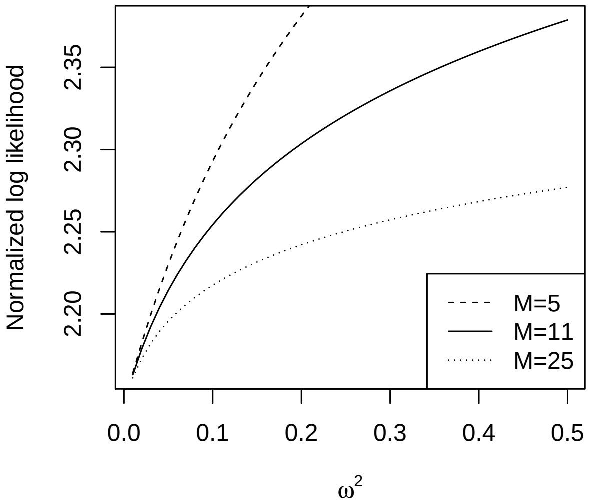

Figure 3.6: Theoretical values of the normalized log-likelihood with $\sigma^2 = 0.5$, as a function of ω^2 (interindividual variability), for different values of M (the number of observations per individual).

tional estimation method linearizes around the empirical Bayesian estimate of η_i, so that $\exp(\eta') = \exp(\hat{\eta}_i) \cdot (1 + \eta')$. The equation then reduces to

$$L_i = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^M \cdot \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2} \cdot M \cdot \left(1 + \left(\frac{\eta' \cdot \exp(\hat{\eta}_i)}{\sigma}\right)^2\right)\right] \cdot \frac{1}{\omega\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}\left(\frac{\eta'}{\omega}\right)^2\right] d\eta';$$

next some algebra gives for the expected value of minus two log L_i:

$$-2\log L_i = M \cdot (\log(\sigma^2) + \log(2\pi) + 1) + \log(M \cdot \exp(2\eta_i) \cdot \omega^2/\sigma^2 + 1). \quad (3.16)$$

The minus two log-likelihood for the population data is the sum of N individual $-2\log L_i$. Now let the expected normalized likelihood be $\overline{NL}$, which is the expected population minus two log-likelihood divided by $N \cdot M$, taking into account that the η_i have on average mean zero and variance ω^2:

$$\overline{NL} = \log(\sigma^2) + \log(2\pi) + 1 + \frac{1}{M}\int_{-\infty}^{\infty} \log(M \cdot \exp(2\eta') \cdot \omega^2/\sigma^2 + 1) \cdot \exp\left[-\frac{1}{2}\left(\frac{\eta'}{\omega}\right)^2\right] d\eta'. \quad (3.17)$$

Figure 3.6 depicts the normalized log-likelihood (with eq. (3.17) evaluated numerically) as a function of ω^2, for $\sigma^2 = 0.5$ and three values of M. For large M, the last term in eq. (3.17) (the integral divided by M) goes to zero, and the uncertainty left in the data is determined only by σ^2. Values for $M = 11$, and $\omega^2 = 0$, 0.1, and 0.5 were used as “target” values in Figures 3.3 - 3.5. The observed averaged normalized log-likelihoods will be larger, because the models used do not fit perfectly, and the parameters are estimated instead of set to their true values.

The context of AIC is also the one where the ηs have been integrated out (but with the parameters at their estimated values), which is to be done when all data are acquired. So while the characteristics of the set of (validation) data are optimally captured, this context is different from the case where prediction errors are calculated with the random effects set to zero instead of integrated out. In that case, the above AIC and OFV_v criteria do not match, as the components of the likelihood in eq. (3.12) are no longer independent (they can only independent if the true values of η for the individuals are also zero). Note however, that from the higher perspective of optimally characterizing a future set of population data, this is a less important case.