Validity of Confidence Intervals in Case of Model Selection Uncertainty

von

Helena Sophia Domes

zur Erlangung des Grades

M.Sc. Pflanzenbiotechnologie

der Naturwissenschaftlichen Fakultät
der Leibniz Universität Hannover

vorgelegt von

Helena Sophia Domes

Geb. am 30.05.1989 in Berlin

Erstprüfer: Dr. Frank Schaarschmidt

Zweitprüfer: Prof. Dr. Thomas Debener

Hannover, Dezember 05, 2016

Abstract

The primary objectives of this thesis are the extension of multimodel inference to multiple comparisons and the construction of model-averaged simultaneous confidence intervals (MASCI). This approach is intended to help dealing with model selection uncertainty, an issue which arises when there is uncertainty in regard to several influencing factors in an experiment. Two cases were considered in this thesis: a linear model setting with unknown influence of several covariables and a setting with different correlation structures of linear mixed effects models.

For this purpose, two formulas by Burnham and Anderson (2002) for calculating the standard error of model-averaged parameters were modified. By changes in the quantiles they were made suitable for multiple comparisons. This approach was tested on the mentioned settings for a multiple comparison by simulation studies. Model-averaging was compared to the often used AICc selection of one model.

The simulation studies showed that the proposed calculation of model-averaged simultaneous confidence intervals works. Especially for small sample sizes of 5 or 10 model-averaging performed better than using a single model after AICc selection. For instance, AICc selection achieved a coverage probability of 70 – 90 % in case of sample size $n_i = 5$, instead of the expected 95 %. Although MASCIs could not reach the 95 %, too, the coverage was always over 90 % and consistently better than the result of only one model. As expected, these asymptotic methods work better for high sample sizes.

The calculation of MASCI was also applied in two examples.

The calculation of model-averaged simultaneous confidence intervals can now easily be made with a function implemented in R called "`masci`". How to use it and the additional code can be found in the Appendix.

Keywords: multimodel inference, model-averaging, simultaneous confidence intervals, multiple comparisons, model selection uncertainty

Content

List of Figures

List of Tables

1. Introduction: The Problem with Model Selection Uncertainty

Statistical analyses for significance, estimation, standard error or confidence intervals are based on data and a statistical model. Usually, the model is not known a priori. Additionally, in some situations of data analysis several models are plausible. This creates uncertainty about which model should be selected for inference. A common problem, even with the same data, is the production of diverse results for significance etc. by different statistical models.

The issue of model selection uncertainty and the usage of multimodel inference will be the core of this thesis. The approach of model-averaging will be extended on multiple comparisons. To illustrate the issue, some example data are shown in the beginning. Thereafter, some statistical basics are explained and a new method is introduced. After two simulation studies, the results of the following data are shown.

1.2 Example "Bodenfauna": Influence of Genetically Modified Crops on Three Soil-Living Species

This field trail deals with the influence of genetically modified crops on three soil-living species. For this, four plant varieties ("variety") were cultivated in a complete block design (8 blocks) and the total number of the three soil-living invertebrate species of interest was counted over the entire vegetation period ("species"). The genetically modified ("GM"), a near-isogenic ("ISO") and a conventional variety ("B") were compared with a second conventional variety "A". Furthermore six soil characteristics were collected ("pH", grain-size fraction 1-3 ("KF1", "KF2", "KF3"), content of organic substance ("C_{org}") and Nitrogen ("N")).

Table 1.1: Short overview of the structure and variables of the "Bodenfauna" data

Target variable	soil-living species	Sum of three counted species
Treatment groups	four plant varieties	GM
		ISO
		A
		B
Covariables	block design	8 blocks
	soil characteristics	pH
		KF1, KF2, KF3
		C_{org}
		N

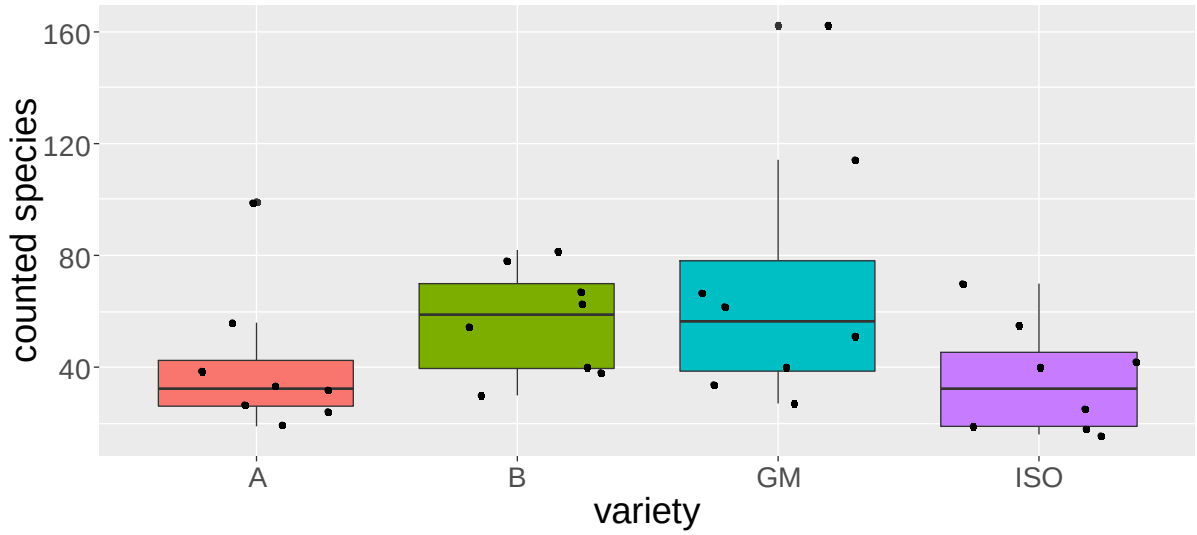

Figure 1.1: "Bodenfauna" data with number of counted species for each variety (A, B, GM, ISO)

The next step for the evaluation was the formation of possible models. An influence of the mentioned covariables and also of the block design on the invertebrates seemed plausible. Consequently, six linear models were built. Model `lm1´ is the standard model for analysing block design. `lm1-3´ could play a role due to soil biology. Here, `KF1´ was selected as an example for the grain-size fraction. After that, a full model with all covariables (`lm5´) and finally a model with none of them (`lm6´) was built.

lm1<- lm(species ~ variety + block)
lm2<- lm(species ~ variety + pH)
lm3<- lm(species ~ variety + KF1)
lm4<- lm(species ~ variety + C_{org})
lm5<- lm(species ~ variety + block + pH + KF1 + KF2 + C_{org})
lm6<- lm(species ~ variety)

All of these models can be justified logically. So which one explains the data best and should therefore be chosen for the multiple comparisons? To illustrate how the use of different models influences further evaluation, the calculated confidence intervals of a Dunnett comparison are shown in Figure 1.2. The values of estimated mean and standard error and therefore the confidence bounds vary.

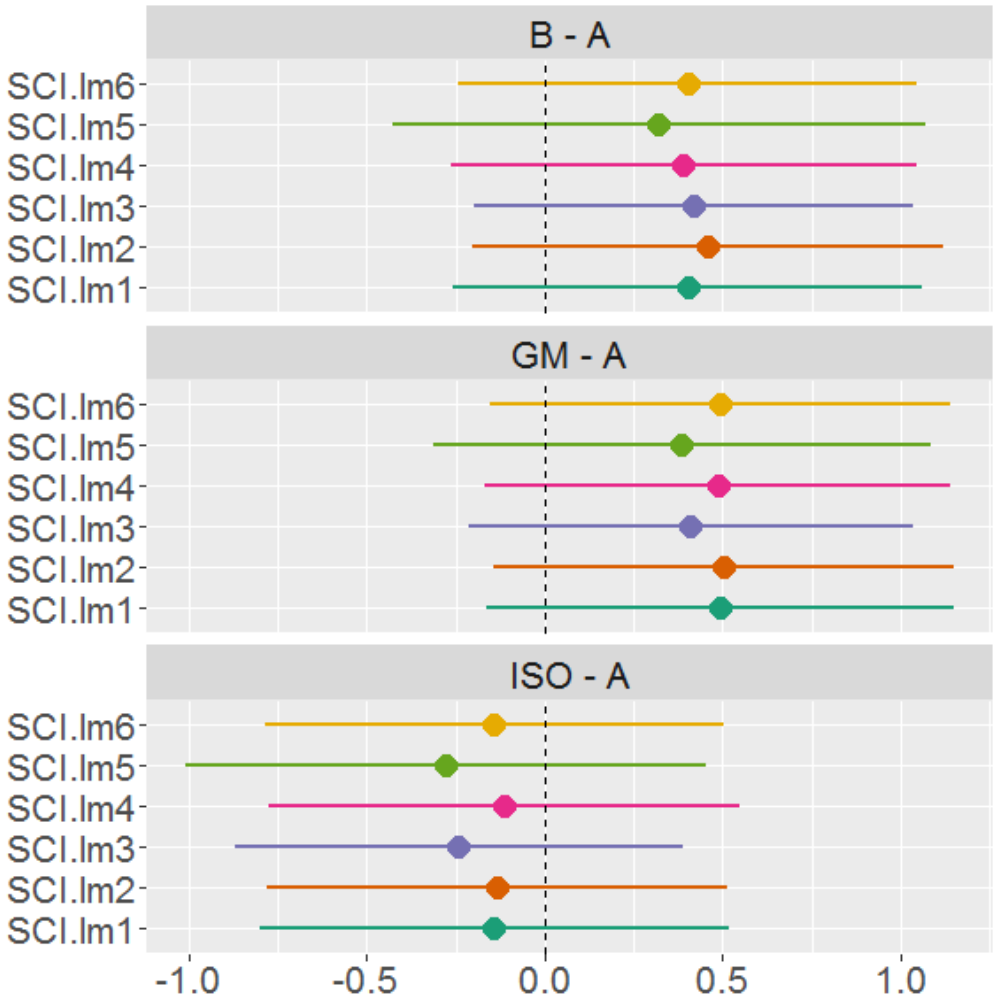

Figure 1.2: Calculated simultaneous confidence intervals (SCI) based on six different models for three comparisons (Dunnett test).

How to solve this evaluation using model-averaging is shown in chapter 7.

1.2 Example: Bradykinin Receptor Antagonism

The next example deals with postoperative bleeding after cardiopulmonary bypass (CPB). That is often caused by fibrinolysis and you can use D-dimers as a biomarker. CPB promotes fibrinolysis via a peptide called bradykinin and the associated bradykinin B_2 receptor. In a randomized, double blind trail, two drugs (HOE 140 and ε-aminocaproic acid (EACA)) were compared to a placebo control (saline) (Balaguer et al. 2013). HOE 140 is a specific bradykinin B_2 receptor antagonist and EACA a well-known antifibrinolytic drug. 37 (EACE and placebo) and 38 (HOE140) patients were assigned to the treatment groups. The concentration of the D-dimers in blood samples was tested at these five time points:

- prior to surgical incision (baseline)
- after 30 minutes on a heart-lung machine (“on-pump”)
- after 60 minutes “on-pump”
- after separation from the heart-lung machine (post-bypass)
- on the first postoperative day

The goal was to test which treatments (HOE140, EACA) reduce D-dimers and therefore fibrinolysis compared to the control.

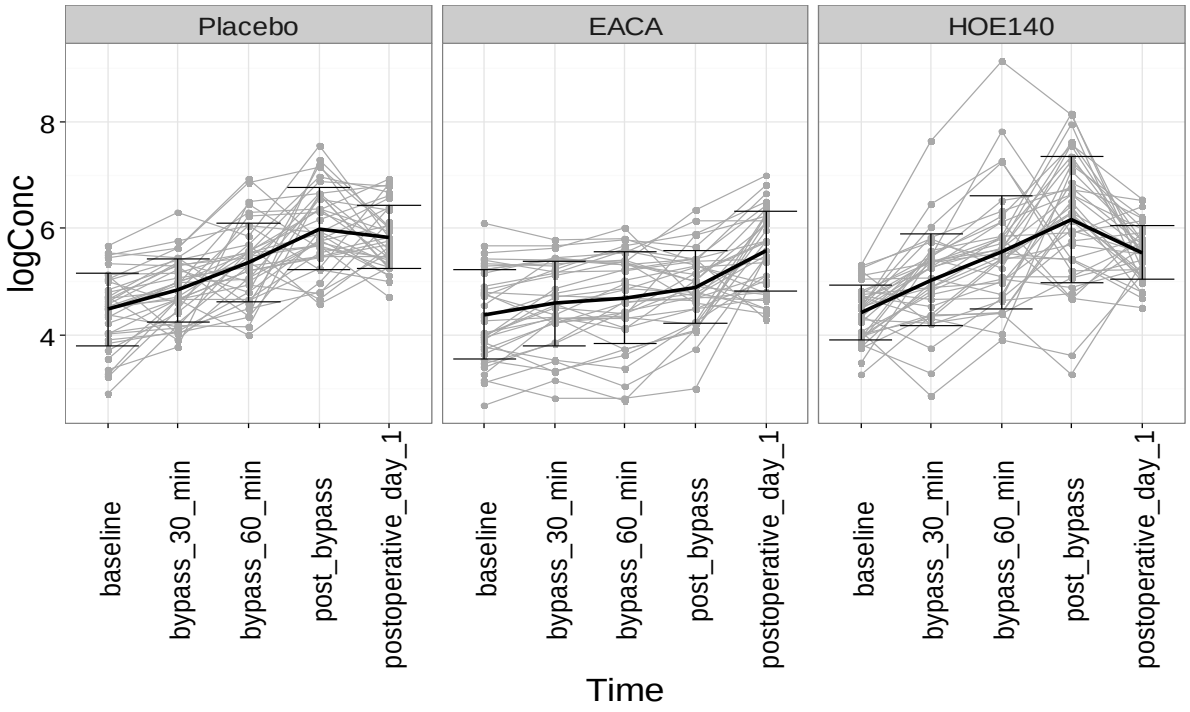

Figure 1.3: Bradykinin data set; individual patient observations (grey) and mean (black) of log-concentration of D-dimers per treatment.

The different observations over the time can be correlated among each other. By using linear mixed effects models, these dependencies can be taken into account. They can be adapted to the data by different correlations structures (read more in chapter 5). Normally, an AIC selection of the candidate models is then performed. In Chapter 7 the evaluation is also carried out using model-averaging.

2. Model Selection Procedures

In this chapter, some methods of model selection are reviewed, focused on the AIC(c) selection.

2.1 AIC Model Selection

The following presents results according to Burnham and Anderson (2002) for a given set of *R* candidate models. After building a set of *R* candidate models, information criteria help to compare the adequacy of these models. A popular criterion is the Akaike information criterion (AIC, Akaike 1973). It estimates the expected, relative Kullback-Leiber (K-L; Kullback & Leiber 1951) divergence (*I(f,g)*) and therefore the information loss between the unknown truth *f(x)* and the approximating model $g(x|\theta)$ with data *x* and parameter θ.

$$I(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x|\theta)}\right) dx$$

For $\log\left(\mathcal{L}\left(\hat{\theta}|data\right)\right)$ as the maximized log-likelihood of model g_i and K as the number of estimable parameters, the AIC is calculated as follows.

$$AIC = -2\log\left(\mathcal{L}\left(\hat{\theta}|y\right)\right) + 2K$$

Comparing the AIC values of all models of your set, the one with the lowest AIC is to be preferred due to being the best approximation to *f(x)*. However, since AIC is a relative value, it is only comparable between models fitting the same data. Additionally, one should note that AIC only chooses the best model out of the built set, thus, these models should be chosen carefully. Furthermore, it is assumed that none of the candidate models is the "true model" (Burnham & Anderson 2002).

For small sample sizes the AICc, following the underneath formula is recommended (Sugiura 1978; Hurvich & Tsai 1989). Burnham and Anderson (2002) mentioned the limit $\frac{n}{K} < 40$ for AICc use with *n* being the sample size. And in 2004 Burnham and Anderson described that AIC*c* converges to AIC, as *n* gets large, so AICc should be used in practice. Therefore, the AICc is also used in this thesis. For cases with overdispersed count data, there is another extension of AIC called QAIC(c) (Anderson et al. 1994), but it is not used in this thesis.

$$AICc = \mathrm{AIC} + \frac{2\,\mathrm{K}\,(\mathrm{K}+1)}{n-K-1}$$

To describe model probabilities, Burnham & Anderson (2002) named the following model weights "Akaike weights". These weights w_i sum up to 1 and have to be built for all models in the set, too. They can be interpreted as a quantified probability of a model g_i to be the actual K-L best model in this set (e.g., Burnham & Anderson, 2001). Usingw_i, the model with the highest probability should be chosen. After each change in the set of *R* models, the weights have to be calculated again. Instead of AIC, other information criteria like AICc or BIC can be used.

$$w_i = \frac{\exp\left(-\frac{1}{2}\,\Delta\, AIC_i\right)}{\sum_{r=1}^{R}\exp\left(-\frac{1}{2}\,\Delta\, AIC_r\right)}$$

2.2 The Bayesian Information Criterion

Another information criterion used for model selection is the Bayesian Information Criterion (BIC). Like the AIC it is based on the likelihood function but differs in the penalty term. According to Schwarz (1978) the BIC is defined as

$$BIC = -2\log\left(\mathcal{L}\left(\hat{\theta}\middle|y\right)\right) + K\cdot\log n$$

and the model with lowest BIC is chosen. Alternatively, the BIC can be used to calculate Akaike weights.

3. Model Averaging/Multimodel Inference:

In case of model selection uncertainty, the described methods won´t lead to one clear best model, e.g. with an Akaike weight of over 95 % compared to the other models in the set. Instead of choosing only one model nonetheless and therewith risking poorer results (e.g. Chatfield 1995), model averaging allows to use the information of the entire set for further calculation. This often leads to better precision and reduced bias (Burnham & Anderson 1998; chapter 4+5)**.**

Additionally, it is reported that the decision for a single model leads to confidence intervals (CI) with too high error rates (Hurvich & Tsai 1990, Lukacs et al. 2010) whereas model-averaged CIs can have coverage rates close to the nominal level and can be narrower than those from the full model (Fletcher & Dillingham 2011). Several approaches for a model-averaged variant were made (Buckland et al. (1997), Burnham & Anderson (2002), Hjort & Claeskens (2003), Claeskens & Hjort (2008), Fletcher & Turek (2011), Turek & Fletcher (2012)). This thesis focuses on the publications of Buckland (1997) and Burnham & Anderson (2002, 2004).

Regardless of the advantages of model-averaging, a thoughtful selection of the candidate models is unavoidable. Burnham and Anderson (2003) recommend a small model set. Fletcher and Dillingham (2011) did not find that the differences between methods were greatly affected by the choice of model sets. They used a set of 19 models. Feas et al. (2007) found stabilized results using a large set and mentioned an example with 36 candidate models.

3.1 Burnham & Anderson

The model-averaged estimate of a parameter θ is calculated according to Buckland et al. (1997) as follows. It is important that the interpretation of $\theta$is consistent across all *R* candidate models.

$$\hat{\theta} = \sum_i w_i \hat{\theta}_i$$

For model-averaged confidence interval construction a simple 95 % 1-parameter interval can be used whereas *z* is the 1- $\frac{\alpha}{2}$th percentile of the standard normal distribution.

$$\hat{\theta} \pm z_{1-\propto/2} \cdot \widehat{se}(\hat{\theta})$$

As an estimator of the variance of estimator $\hat{\theta}$ under consideration of the model weights w_i, Buckland et al. (1997) recommend the following calculation with $\hat{\bar{\theta}} = \sum_{i=1}^{R} w_i \hat{\theta}_i$:

$$\widehat{var}\left(\hat{\bar{\theta}}\right) = \left[\sum_{i=1}^{R} w_i \sqrt{\widehat{var}(\hat{\theta}_i|g_i) + (\hat{\theta}_i - \hat{\bar{\theta}})^2}\right]^2$$

Burnham and Anderson (2002) extend this formula by a term including the quotient of the 97.5th percentile of the *t*-distribution with *v* degrees of freedom and the standard normal distribution. They called it *"adjusted standard error" (ase)* in their publications and furthermore mentioned two different variations of calculation.

The first calculation for $\widehat{ase}$ is described in Burnham & Anderson 2002 (section 4.3.3, page 164) and will be named $\widehat{ase}_1$ hereinafter.

$$\widehat{ase}_1(\hat{\theta}_i) = \sum_{i=1}^{R} w_i \sqrt{\left(\frac{t_{v_i,1-\alpha/2}}{z_{1-\alpha/2}}\right)^2 \widehat{var}(\hat{\theta}_i|g_i) + (\hat{\theta}_i - \hat{\bar{\theta}})^2}$$

The second variant is also mentioned in Burnham & Anderson (2002; p.345, Eq. (6.12)) and will be called $\widehat{ase}_2$ from now on.

$$\widehat{ase}_2(\hat{\theta}_i) = \sqrt{\sum_{i=1}^{R} w_i \left[\left(\frac{t_{v_i,1-\alpha/2}}{z_{1-\alpha/2}} \right)^2 \widehat{var}(\hat{\theta}_i | g_i) + (\hat{\theta}_i - \hat{\bar{\theta}})^2 \right]}$$

4. Multiple Comparisons

If more than two groups are considered in an experiment, many-to-one or all-pairs tests can be used for their comparison. The Dunnett-test (Dunnett 1955) as a many-to-one test and the Tukey-test (Tukey 1953) for all-pair comparisons are widely used. Therefore, the data should be continuous, independent, normally distributed and have homogeneous variance. The associated contrast matrices **C** for 4 groups are shown below. The rows contain the single comparison and the columns the different treatments. Whereas the "-1" and "1" label the treatments chosen for the comparison. E.g. the first row of the Dunnett contrast matrix imply the comparison of treatment 1 ("1") against the control ("-1").

$$\boldsymbol{C}^{Dunnett} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}$$

$$\boldsymbol{C}^{Tukey} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

Our parameter of interest is the matrix product $\theta = \mathbf{C}\mu$, where μ is the column vector of treatment means $\mu = (\mu_1, \mu_2, \mu_3, \mu_4)$.

$$\theta = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} = \begin{matrix} -\mu_1 + \mu_2 + 0 + 0 \\ -\mu_1 + 0 + \mu_3 + 0 \\ -\mu_1 + 0 + 0 + \mu_4 \end{matrix}$$

4.1 Simultaneous Confidence Intervals

The corresponding test statistic T for a single contrast *h* with $\hat{\theta}$ as difference of treatment means is

$$T_h = \frac{\widehat{\theta_h}}{\widehat{se}(\hat{\theta}_h)}$$

For e.g. Dunnett comparison of three groups to control, **T** = (T_1, T_2, T_3) follows the multivariate *t*-distribution.

The calculation of an approximate simultaneous confidence interval (SCI) for a chosen contrast uses a two-sided quantile of the multivariate *t*- distribution with α = 0.05, *v* degrees of freedom and the correlation **R** among the test statistics (Hothorn et al. 2008).

$$\hat{\theta} \pm t_{1-\alpha}^{two}(v, \boldsymbol{R}) \cdot \widehat{se}(\hat{\theta})$$

4.2 Model-Averaged Simultaneous Confidence Intervals MASCI

The next step combines the model-averaging formulas for $\widehat{ase}$ (1+2) by Burnham and Anderson (2002) and the calculation of SCIs for using multimodel inference for multiple comparisons, too. Therefore, the additional term containing the *t*- and *z*-quantiles was modified. The new formula replaces the quantile of the univariate *t*-distribution by one of the multivariate *t*-distribution **(**with α = 0.05 and correlation matrix $\hat{R}$ **)**.

The simultaneous confidence intervals computed using $\widehat{se}_1$ will be called *model-averaged simultaneous confidence intervals 1 (MASCI1)* and using $\widehat{se}_2$ leads to *MASCI2*:

$$\widehat{se}_1(\hat{\theta}_i) = \sum_{i=1}^{R} w_i \sqrt{\left(\frac{t_{1-\alpha/2}(v, \boldsymbol{R})}{z_{1-\alpha/2}}\right)^2 \widehat{var}(\hat{\theta}_i | g_i) + (\hat{\theta}_i - \hat{\bar{\theta}})^2}$$

$$\widehat{se}_2(\hat{\theta}_i) = \sqrt{\sum_{i=1}^{R} w_i \left[\left(\frac{t_{1-\alpha/2}(v, \boldsymbol{R})}{z_{1-\alpha/2}}\right)^2 \widehat{var}(\hat{\theta}_i | g_i) + (\hat{\theta}_i - \hat{\bar{\theta}})^2\right]}$$

The performance of these model-averaged simultaneous confidence intervals was tested in two simulation studies. First, a linear model setting with unknown influence of several covariables (Chapter 5) and second, a setting with linear mixed-effect models with unknown random effect structure (Chapter 6) was simulated.

5. Simulation Study 1: Linear Models with Unknown Influence of Covariables

The next two chapters deal with simulation studies. But why do we need them? After the made assumption about the calculation of model-averaged simultaneous confidence intervals, there are still questions to be asked. Does the adaption of 1-parameter confidence intervals to multiple comparisons work this way? The derivation of ase_1 and ase_2 by Burnham and Anderson (2002) was rather heuristic. Moreover, the used methods are asymptotically, i.e. only valid for large sample sizes *n*. So for which sample sizes does the method work and which *n* can be chosen in practice? Are there any improvements through this more complex method against simpler approaches? And is this method worth the effort? Simulation studies are therefore a tool to verify the previous suggested calculation for model-averaged simultaneous confidence intervals.

As a criterion for the comparison of different methods, the coverage probability was chosen. For this, a large number of data sets are simulated for a special parameter setting and then the SCIs are calculated. How many of these SCIs contain the true difference of mean? The coverage probability should be 95%.

$$coverage\ probability = \frac{SCI\ with\ true\ difference\ of\ mean}{number\ of\ all\ calculated\ SCI}$$

In the first simulation study the coverage probabilities of confidence intervals calculated with different methods were compared. As model averaging method intervals were built according to both variants of 4.2 ("*Model-Averaged Simultaneous Confidence Intervals*, MA-SCI1/2"). The underlying set of models was made up from all possible models (using a maximum of four covariables in one model) without any selection (Tab. 5.2). The other intervals were

calculated first after model selection using AICc ("AIC best"), then with all ("full model") and last with none ("smallest model") of the covariables.

The computed situation was a multiple comparison of three treatment groups (control, T1, T2) with different effect of covariables. For the comparison, a Dunnett test was chosen (T1-control; T2-control). As an illustration, the data of one simulated parameter setting is shown in Figure 5.1.

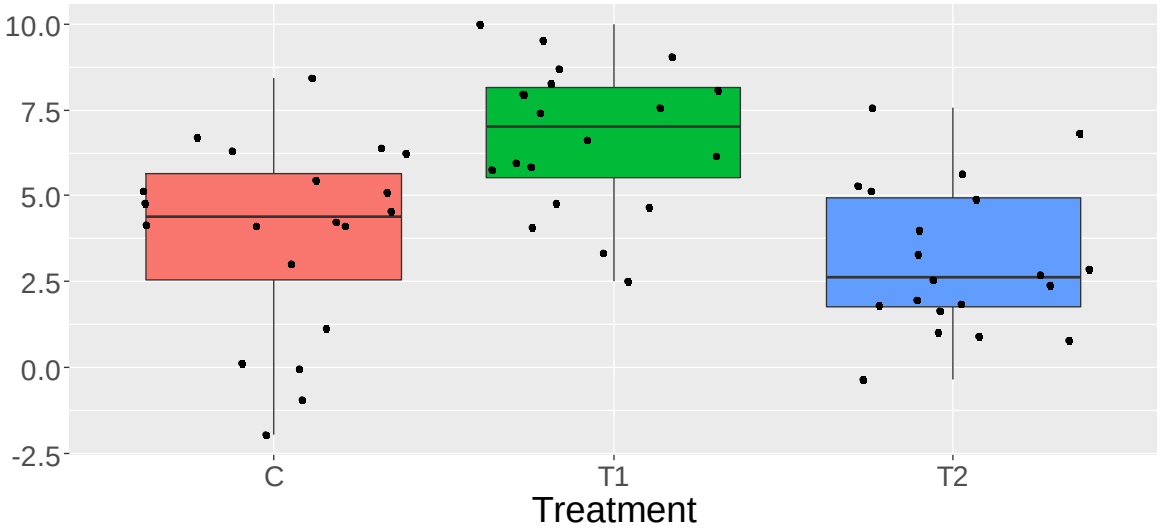

Figure 5.1: Data of one exemplary chosen parameter setting of simulation study 1.

The goal was to simulate model uncertainty and to identify which method deals best with it.

5.1 Parameter Settings

The data generation followed the formula underneath

$$yij = ai + b1*x1ij + b2*x2ij + \ldots + b10*x10ij + eij$$

i	treatment group 1-3
j	1,..., n[*i*] Observations per group
yij	observations
x1ij,..., x10ij	values of covariable 1 to 10
ai	Intercept of group i=1,2,3; parameter of interest is the difference of these two intercepts d
b1,...,b10	slope of the covariables x1ij,..., x10ij
eij	Normal distributed residuals of the linear model, with eij~ N(0, sde2), and sde is the standard deviation of the residues

The variation of parameters n_i (sample size per group), b and sde is shown in Table 5.1 and 5.2. Therefore, 48 different settings were tested. The true difference between treatment 1 and the control was fixed at 2 und treatment 2 had no difference to control. There was no correlation among the covariables.

Table 5.1: Fixed and varied values of parameter n_i, sde, b and d.

Parameter		values
n	Sample size	5 |10 | 20 | 50
sde	Standard error of residuals	0.1 | 3 | 10
b	Number of covariables	4 | 6 | 8 | 10
	Correlation between covariables	No (fixed)
d	True difference of mean	T1 – C: 2 (fixed)
		T2 – C: 0 (fixed)

For a better understanding of the model building process, Table 5.2 shows the number of all candidate models at the different covariable settings. Remember that a maximum of four covariables were in one model. Additionally, the values of the parameter b are listed.

Table 5.2: Model building at the different covariable settings and values of b.

Number of covariables	models in the set	value of each b
4	16	1, 0, -2, 0
6	57	1, 0, -2, 0, 0, 2
8	163	1, 0, 0, -1, 0, 0, 2, 0
10	386	0, 0, 0, 1, 0, 0, -1, 0, 2, 0

The simulation had 1000 repetitions.

5.2 Control with Two Groups

As a control for the new model averaged method, the whole simulation study was repeated comparing only two groups. Therefore, the two original equations (see chapter 3) from Burnham & Anderson (2002) were used.

5.3 Shannon-Wiener Index

As an indication of model uncertainty in the single settings of this simulation study, the Shannon-Wiener index was calculated for every repetition and the mean for each setting was outputted. The Shannon-Wiener index is often used to describe biodiversity, but here it reflects the weight distribution of the models.

$$H' = -\sum_i p_i \cdot \ln p_i \qquad \text{mit } p_i = \frac{n_i}{N}$$

N sum of all weights → 1

n_i Akaike weight of model$_i$

If all models are equally weighted, the Shannon-Wiener index will reach its maximum (*log* of the number of models). Because of the different numbers of models (16, 57, 163 or 386) in the settings you can´t compare the indices automatically. For the comparison between all settings the following quotient was used:

$$Shannon_{Quotient} = \frac{Shannon\ index\ H'\ of\ the\ setting}{Shannon_{max}}$$

5.4 Result

As you see in figure 5.2 and 5.3, especially the standard error (sde) and the number of observations (*n*) and covariables affect the coverage probability of the model-averaged and AIC selected methods. The coverage probability of the other two methods (usage of "full" and "smallest model") constantly remained around 0.95.

For $n \geq 20$ all methods work equally well, but smaller sample sizes showed poorer results. With an increasing number of covariables and with higher standard errors the coverage probabilities deteriorate even more. This effect was observed most clearly for "AICbest". At settings with b = 10, sde = 3 or 10 and sample sizes of 5, the coverage probability went down to 70 %. The model averaged methods performed better than the AIC selected (see also Figure 5.4).

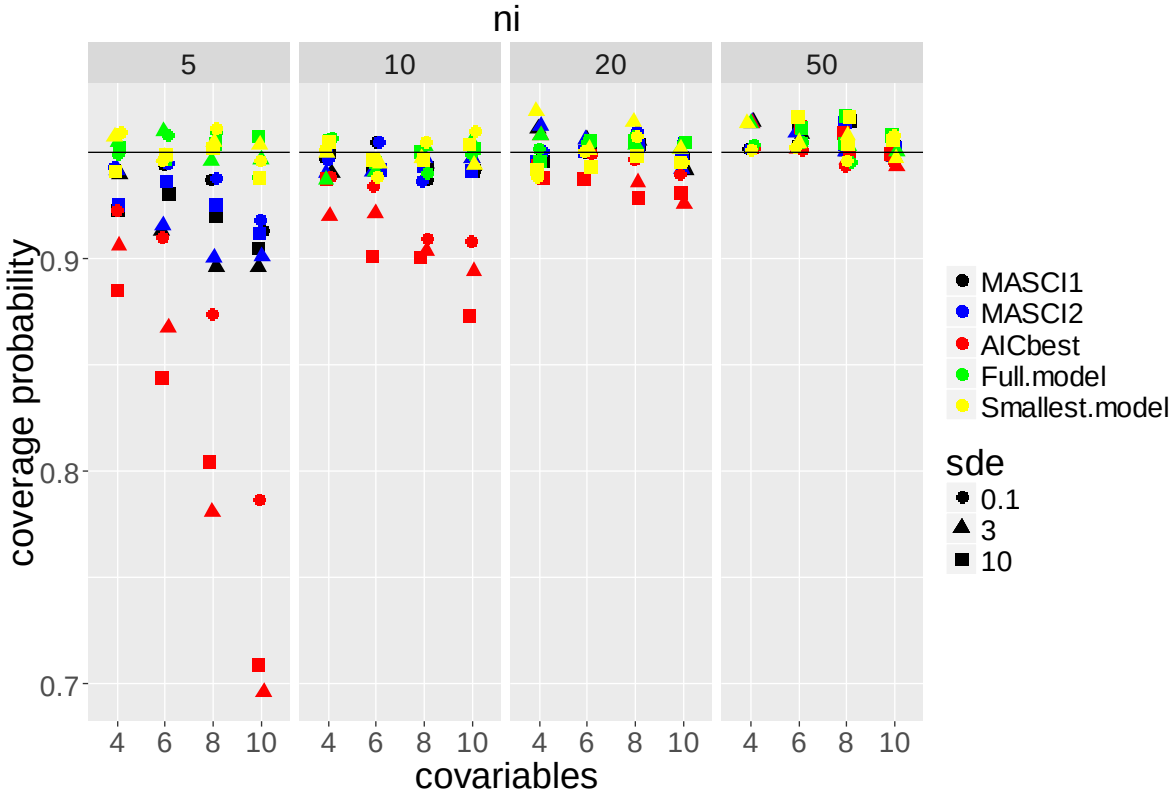

Figure 5.2: Coverage probability of two-sided 95% simultaneous confidence intervals calculated according to different methods and at diverse parameter settings.

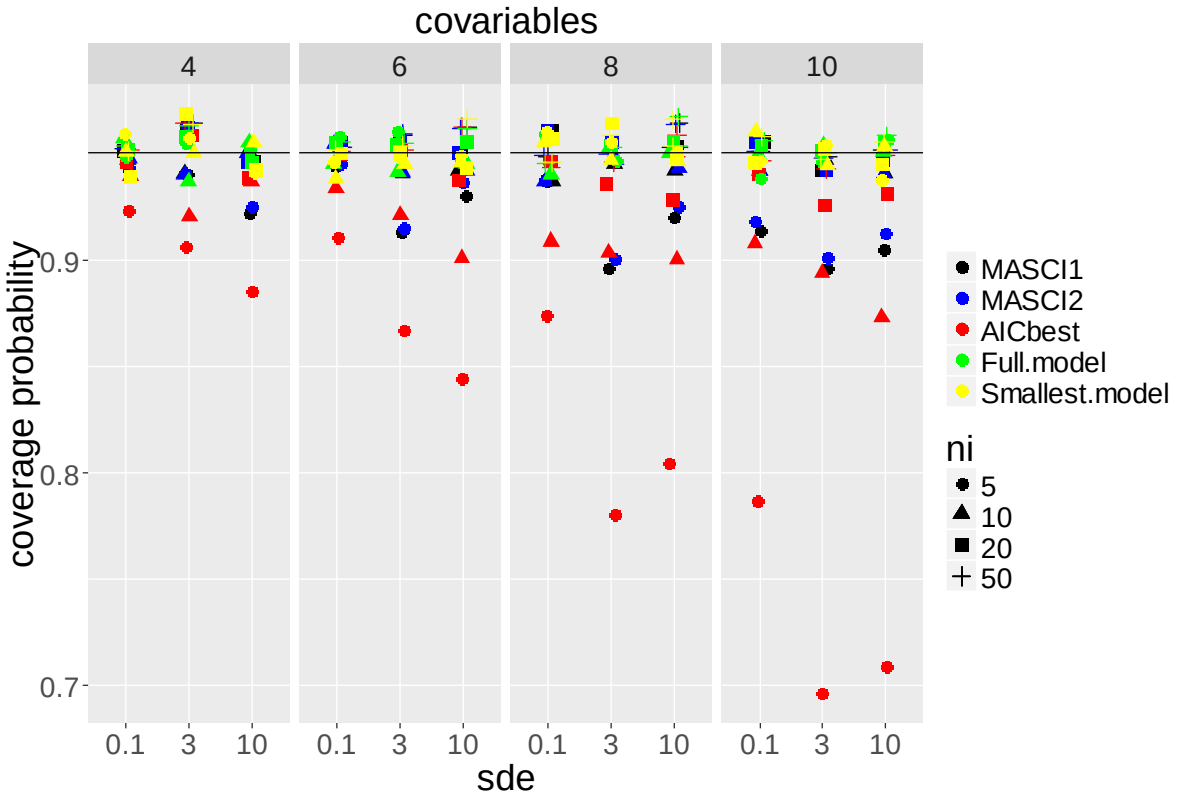

Figure 5.3: Coverage probability of two-sided 95% simultaneous confidence intervals calculated according to different methods and at diverse parameter settings.

In the direct comparison there are hardly any differences between the two variants of model-averaged simultaneous confidence intervals (Figure 5.3). In some settings MASCI2 reached marginal better results. Furthermore, it can be seen that the model-averaged methods achieve better results than the interval after AIC selection.

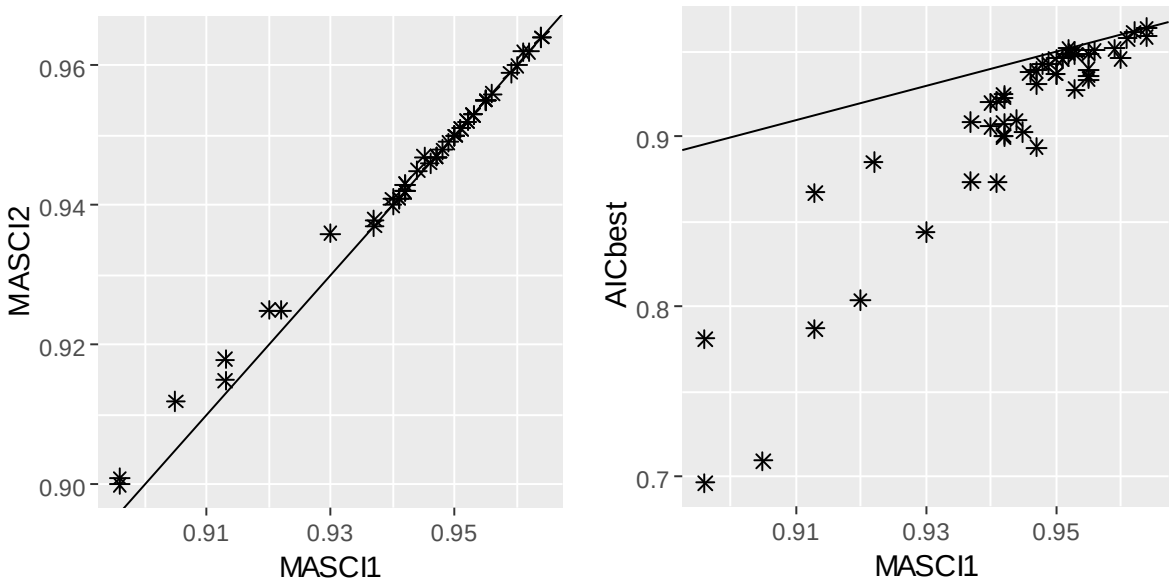

Figure 5.4: Comparison of the coverage probabilities of the two model-averaged methods (MASCI 1+2) and of the MASCI1 to the AIC selected method. The solid line would implicate equal coverage probabilities of the methods.

Figure 5.5 shows the Shannon quotient at different parameter settings. Higher Shannon quotients reflect a greater uniformity of model weights and therefore more obvious model uncertainty. It can be seen that the Shannon quotient is lower with small and becomes larger with increasing standard error. Small sample sizes did not lead to higher uncertainty and an effect of the number of covariables is not uniform.

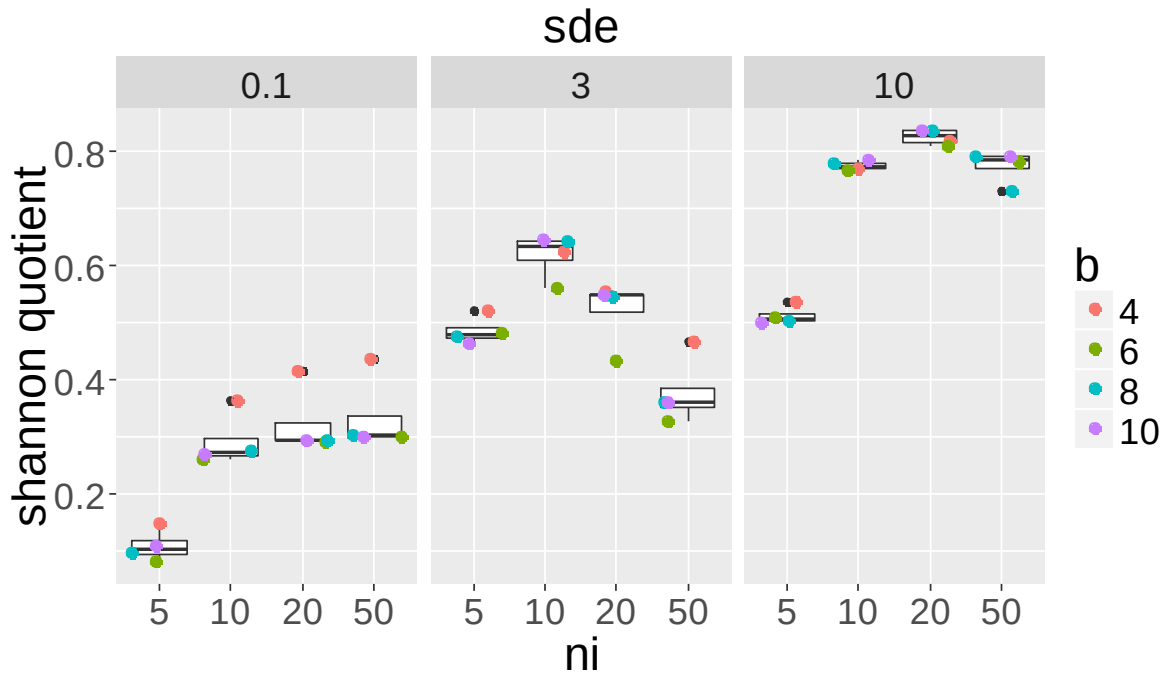

Figure 5.5: Shannon quotient for different parameter settings with influence if parameters n_i, sde and b.

As a further illustration the mean of the three highest model weights is shown in Figure 5.6.

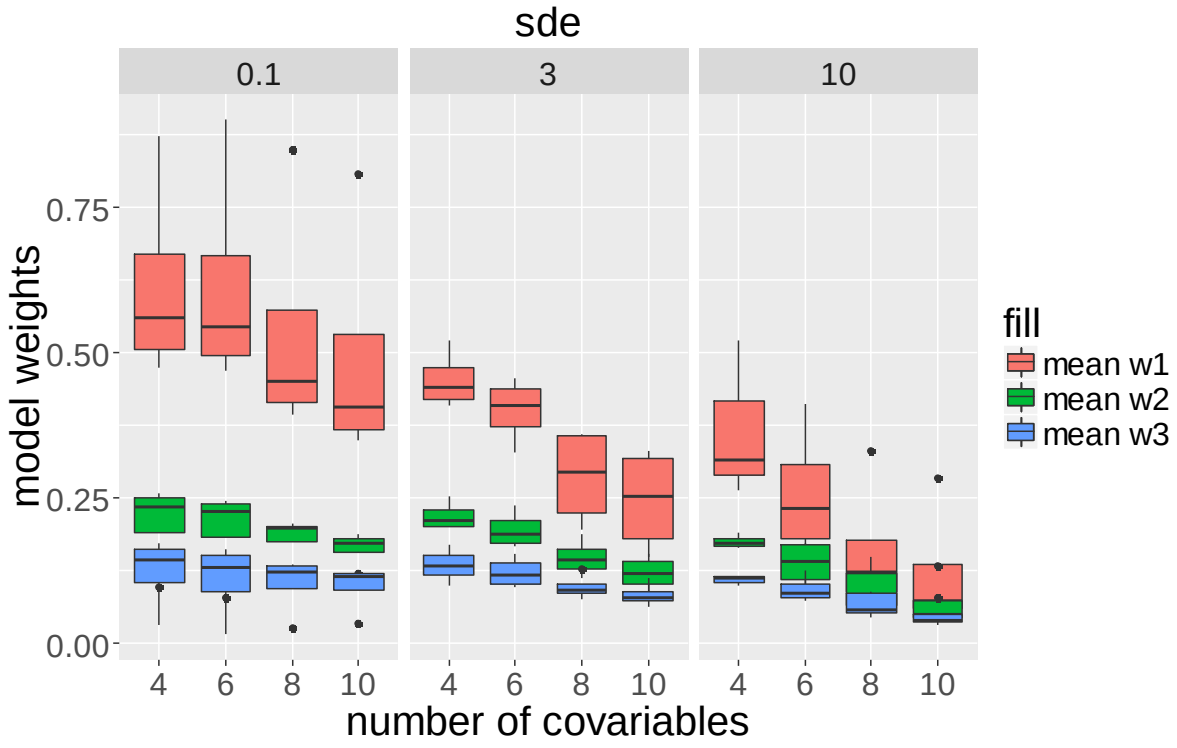

Figure 5.6: The mean of the three highest weighted models ("w1", "w2", "w3") for different parameter settings.

5.4.1 Results from control

The control results roughly follow the pattern of the multiple comparisons (Fig. 5.6). Apart from the setting with b≥ 8 and n_i = 5, where the full model and "AICbest" couldn´t be fitted. As a reminder: all in all, only ten observations were simulated in these settings and not 15 as in the comparison of three groups.

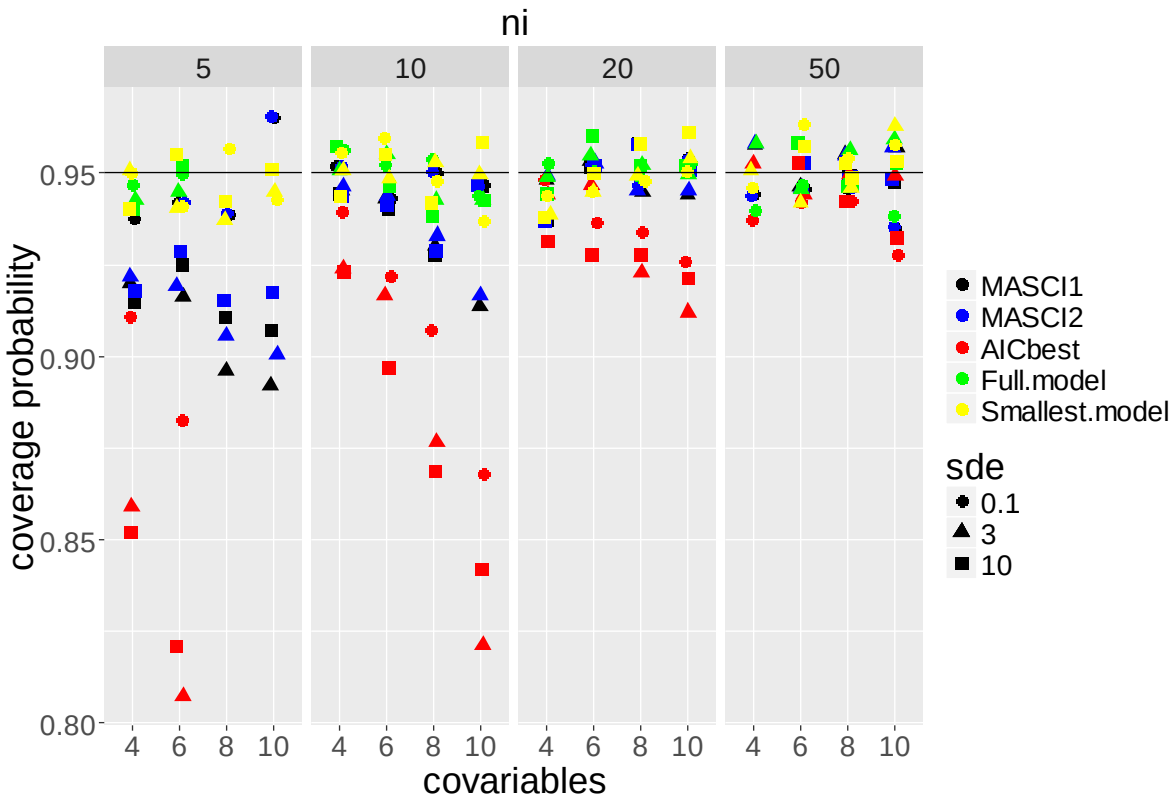

Figure 5.6: Coverage probability of two-sided 95% confidence intervals calculated according to different methods and at diverse parameter settings.

As you see in Figure 5.7, the model-averaged methods work equally and both achieve better results than the one using AICc selection. In some settings using ase_2 reached marginal better results.

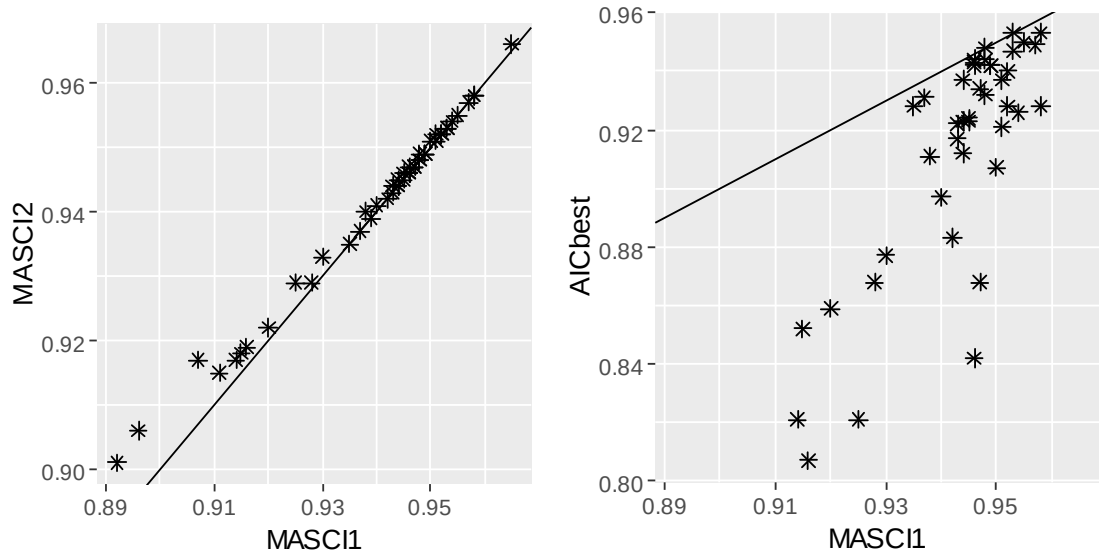

Figure 5.7: Comparison of the coverage probabilities of the two model-averaged methods and of the model-averaged (1) to the AIC selected method. The solid line would implicate equal coverage probabilities of the methods.

5.5 Discussion

The Shannon quotient and the Akaike weights should provide information about the model selection uncertainty of the single settings. Although these were the means from the 1000 repetitions per setting, the uncertainty about model selection is evident (Figure 5.5, 5.6). Thus, a good starting point for evaluation with multimodel inference was created.

Moreover, the results show that using a model-averaged confidence interval improves the coverage probability in contrast to the usage of only the "AICbest" model. This applies to single as to multiple comparisons. In the simulated multiple comparison, MASCIs reached with sample sizes ni $\geq$ 10 the desired approximately coverage of 0.95. In case of smaller sample sizes, the coverage probability decreased. However, this method also achieves better results than AICc selection, even with smaller sample sizes. For instance, AICc selection achieved a coverage probability of 70 – 90 % in case of sample size n_i = 5, instead of the expected 95 %. This means that in the worst case α is 0.3 instead of the assumed and required α = 0.05.

Comparing the two groups, the coverage probability reached 95 % only with sample size $n_i \geq 20$, because the total number of observations was lower (e.g. $n_c+n_{T1} = 40$) than in the case of the multiple comparison (e.g. $n_{c+} n_{T1+} n_{T3} =60$).

The two MASCI variants worked mostly the same. In some settings, MASCI2 reached marginal better results. This observation can also be found in the literature. Turek and Fletcher (2012) explain this with the narrower interval of MASCI1. The interval width was not documented in this simulation study. Fletcher and Dillingham reported in 2011 that both equations performed equally well using AIC criterions (AIC, AICc, AIC*c) but for BIC the use of ase_2 provided slightly better coverage.

The formation of the model set was a little different than in practice because of the automatic simulation. All possible covariable combinations were computed without any well-considered choice as recommended (Burnham & Anderson 2002) for building a model set. Furthermore, the maximum number of covariables was set to 4. The extent to which these facts have had a negative impact is difficult to predict.

Surprising was the good performance of the smallest and largest model. But to therefore always choose these simple solutions is not a recommended alternative. Covariables with effect on the observation should be taken into account in the evaluation. In this simulation study, only a few of the generated covariables had an influence (Table 5.2). And these 2-3 covariables per setting had values from -2 to 2. Perhaps, settings should be simulated with a larger variation of the covariables. But why not choose a model with all the factors considered important? Burnham and Anderson (2011) refer to the limited information content of a data set and that an increasing number of factors eventually make the fitted model unstable and uninformative.

6. Simulation Study 2: Linear Mixed Effects Models with Random Effects

In this simulation study an experiment with crossover design was computed. Those designs are especially important in health care. A varying number of individuals receive a sequence of different treatments at five time points. Ultimately, all individuals received all the treatments. As an illustration, an example dataset was generated and plotted (Figure 6.1).

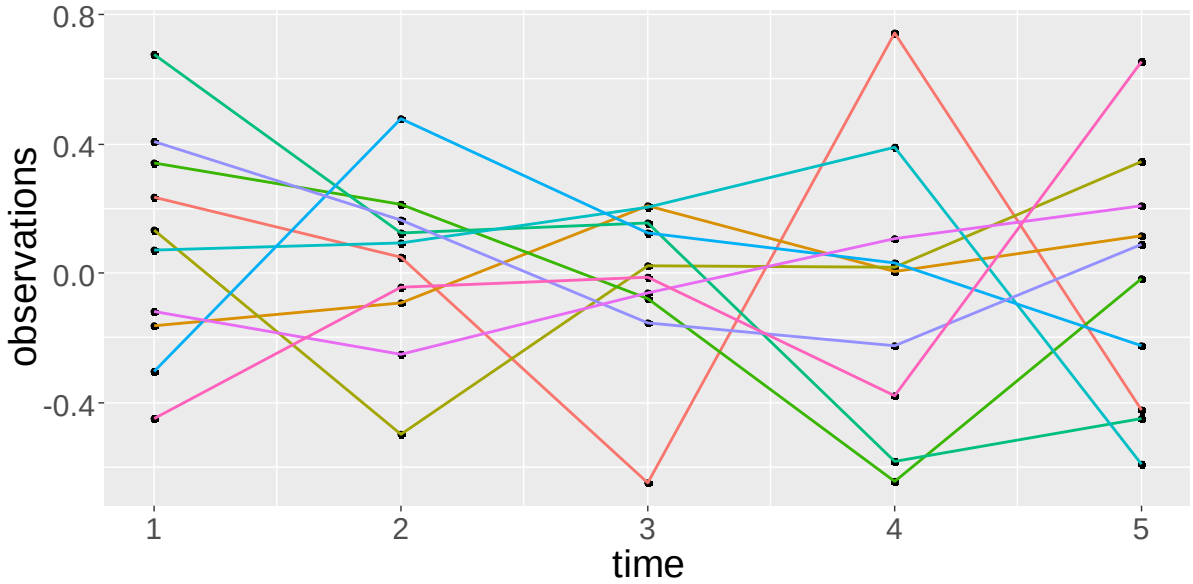

Figure 6.1: Simulated data in crossover design with 10 individuals (colored lines) and five different treatments at five time points.

As in simulation study 1, a Dunnett test was chosen to compare the five treatments (treatment A-D to control). The simultaneous confidence intervals were built using both model-averaged equations and with the "best" model after AICc selection. AIC selection is recommended by Zuur et al. (2009, chapter 7) and also Pinheiro and Bates (2000)**.** The coverage probability of each method was evaluated.

6.1 Parameter Settings

In experimental setups with repeated measurements, linear mixed-effects models can be used for evaluation. The model can be written as seen below with y_i as a known vector of observations, X_i and Z_i are known design matrices for the fixed and random effects and β is an unknown vector of the fixed effects. Additionally to fixed-effect parameters, the residual error ϵ_i and random effects parameters b_i can be included.

$$y_i = X_i\beta + Z_i b_i + \epsilon_i \ , \quad i = 1, \dots, n$$

In this simulation study X_i codes the treatment groups, such that the fixed effect parameter β estimates the mean differences between the treatment groups (parameter of interest in multiple comparisons). Z_i codes, which observation belongs to which individual. $b_i \sim \mathrm{N}(0, \sigma^2_{\mathrm{ind}})$ is the random effect representing the variance between individuals. ϵ_i is the vector of residuals over time belong the same individuals. Repeated measurement over time may invoke different correlations among these residuals, depending on time. These correlation structures are subject to model selection.

In this simulation study, some frequently used correlation structures for the random effects were chosen (for background information read Pinheiro & Bates 2000, chapter 5.3). Eventually, eight different data generating settings with the "true" correlation structure (Table 6.1) were simulated.

Table 6.1: Used correlation structures for data generating. Since they are symmetric, only their upper triangles are displayed.

Independence ("NULL")	$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 \\ & & 1 & 0 & 0 \\ & & & 1 & 0 \\ & & & & 1 \end{pmatrix}$	
First-order autoregressive (AR1)	$\begin{pmatrix} 1 & 0.8 & 0.64 & 0.51 & 0.41 \\ & 1 & 0.8 & 0.64 & 0.51 \\ & & 1 & 0.8 & 0.64 \\ & & & 1 & 0.8 \\ & & & & 1 \end{pmatrix}$	"AR1" corAR1(0.8, ~1|id)

	$\begin{pmatrix} 1 & 0.3 & 0.09 & 0.03 & 0.01 \\ & 1 & 0.3 & 0.09 & 0.03 \\ & & 1 & 0.3 & 0.09 \\ & & & 1 & 0.3 \\ & & & & 1 \end{pmatrix}$	"AR1low" corAR1(0.3, ~1|id)
Autocorrelation-moving average (ARMA)	$\begin{pmatrix} 1 & 0.49 & 0 & 0 & 0 \\ & 1 & 0.49 & 0 & 0 \\ & & 1 & 0.49 & 0 \\ & & & 1 & 0.49 \\ & & & & 1 \end{pmatrix}$	"ARMA1" corARMA(0.8, ~1|id, p=0, q=1)
	$\begin{pmatrix} 1 & 0.28 & 0 & 0 & 0 \\ & 1 & 0.28 & 0 & 0 \\ & & 1 & 0.28 & 0 \\ & & & 1 & 0.28 \\ & & & & 1 \end{pmatrix}$	"ARMA1low" corARMA(0.3, ~1|id, p=0, q=1)
	$\begin{pmatrix} 1 & 0.8 & 0.48 & 0.29 & 0.17 \\ & 1 & 0.8 & 0.48 & 0.29 \\ & & 1 & 0.8 & 0.48 \\ & & & 1 & 0.8 \\ & & & & 1 \end{pmatrix}$	"ARMA2" corARMA(c(0.6,0.8), ~1|id, p=1, q=1)
	$\begin{pmatrix} 1 & 0.36 & 0.04 & 0.004 & 0 \\ & 1 & 0.36 & 0.04 & 0.004 \\ & & 1 & 0.36 & 0.04 \\ & & & 1 & 0.36 \\ & & & & 1 \end{pmatrix}$	"ARMA2low" corARMA(c(0.1,0.3), ~1|id, p=1, q=1)
Unstructured	$\begin{pmatrix} 1 & 0.8 & 0.4 & 0.2 & 0 \\ & 1 & 0.8 & 0.4 & 0 \\ & & 1 & 0.8 & 0 \\ & & & 1 & 0 \\ & & & & 1 \end{pmatrix}$	corSymm(c(0.8,0.4,0.2,0.0, 0.4,0.2,0.0, 0.2,0.0, 0.0), ~1|id)

Beside the random effects, some parameters (n_i, sdi, sde) vary per setting; others were fixed. A compilation can be found in Table 6.2.

Table 6.2: Values of fixed and varied parameters n_i, sdi, sde and differences of means.

n_i	number of individuals	5 | 10 | 20 |100
sdi	standard deviation between individuals	0.01 | 0.5 | 1 | 5
sde	residual standard deviation	0.5 | 1 | 10
	Number of measurement points	5 (fixed)
	Differences of means	A – control: 0
		B – control: 1
		C - control: 0
		D – control: 0

The model set was built of six models with different random effects. The structures followed the ones mentioned in Table 6.1 (Independence, AR1, ARMA) but differ in their values. The underlying "true" model was never included in the set except the assumption of independence. In some cases, a model could not be fitted to the generated data. Then, the model set was smaller.

A total of 384 settings were simulated with 2000 repetitions.

6.2 Results

The coverage probabilities were calculated for each setting and method (Figure 6.2). The model-averaging methods achieved better results than using AICc selection except setting n_i = 100. At this high sample size, the methods worked equally well and reached an approximate coverage of 0.95. With decreasing n_i, the AICc selected model produced increasingly poorer results up to a coverage probability of under 90 % (n_i = 5 and 10). At the same time, model-averaging performed much better. Thus, the coverage probability can be positively influenced by sample size and calculation method. An influence by sde and sdi on the coverage was not found.

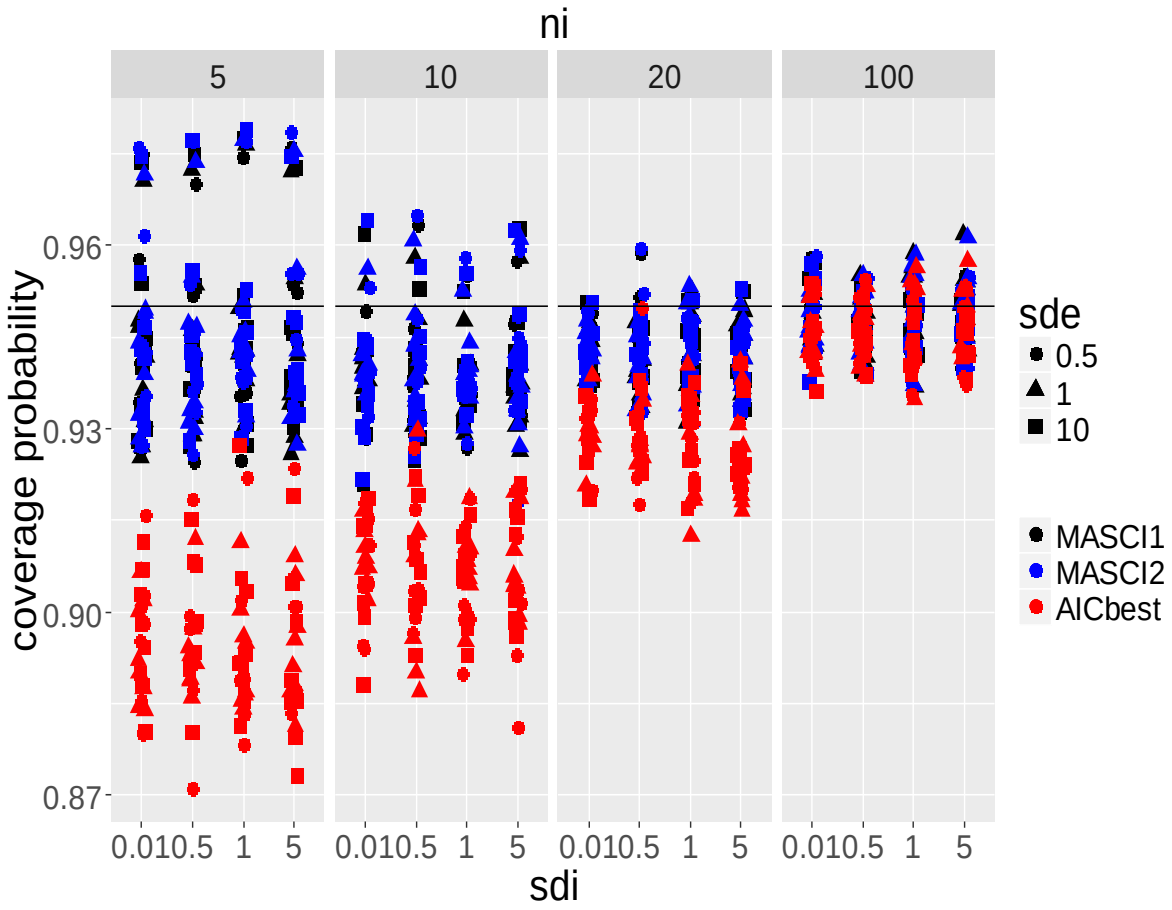

Figure 6.2: Coverage probabilities from simultaneous confidence intervals calculated by different methods (model-averaged (MASCI 1/2) and AICc selection) at several parameter settings. n_i is the number of individuals, sdi the standard deviation between individuals and sde the residual standard deviation.

As shown in Figure 6.3, the Shannon quotient as an illustration for model selection uncertainty was similar in most settings. It fluctuated around 0.75, whereas a value of 1 would mean that all models are equally weighted. The mean of the highest model weight was around 0.4. These values indicate an existing model selection uncertainty in the settings. However, a complete independence of the n_i effect and the model uncertainty could not be achieved. Yet, the lower coverage probability at smaller sample sizes (Figure 6.2) cannot be explained by greater uncertainty. On the contrary, smaller sample sizes often reached lower Shannon quotients and therefore lower model uncertainty. As an exception, settings with an "ARMA2" correlation structure and $n_i = 50$ and 100 led to the lowest Shannon quotients by far.

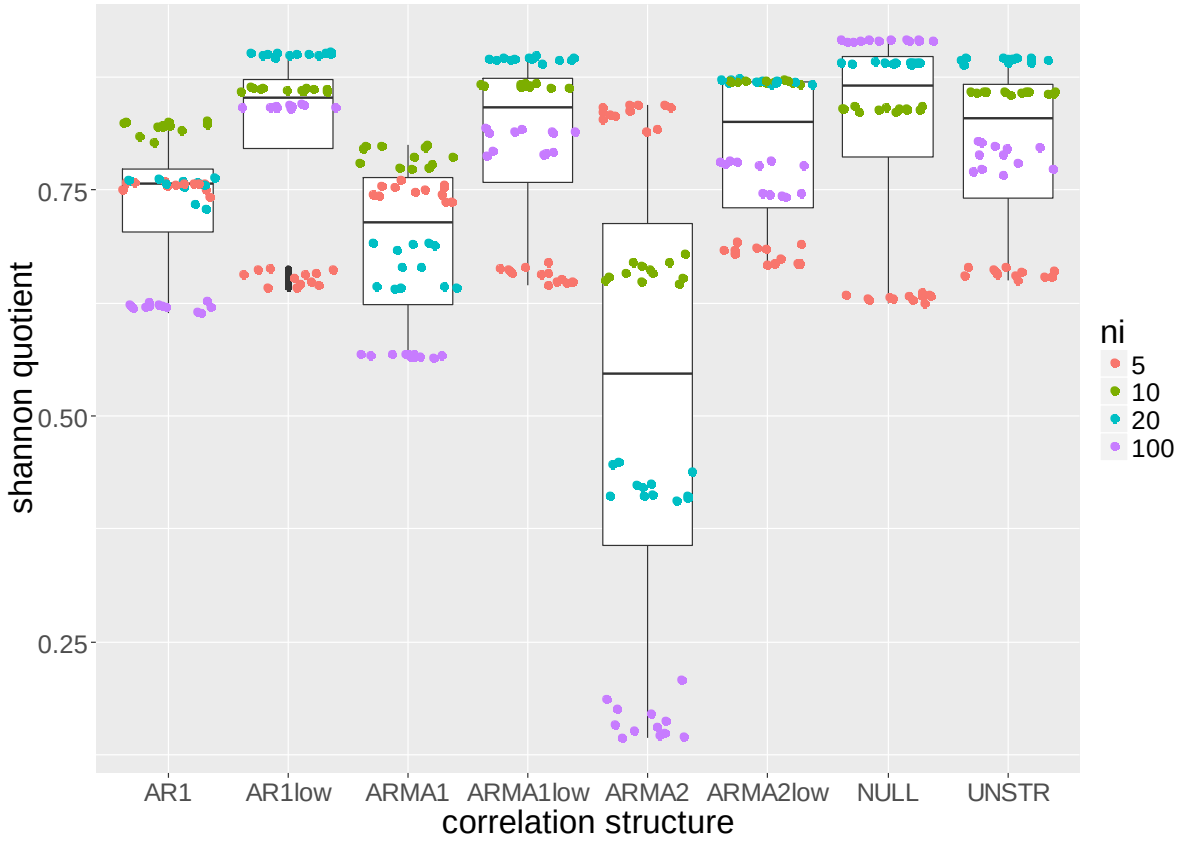

Figure 6.3: Mean of the Shannon quotient per setting and in dependence of the correlation structure and n_i.

In addition, the selection uncertainty did not negatively affect the coverage (Figure 6.4). Lower coverage probabilities can be better clarified with the selected method.

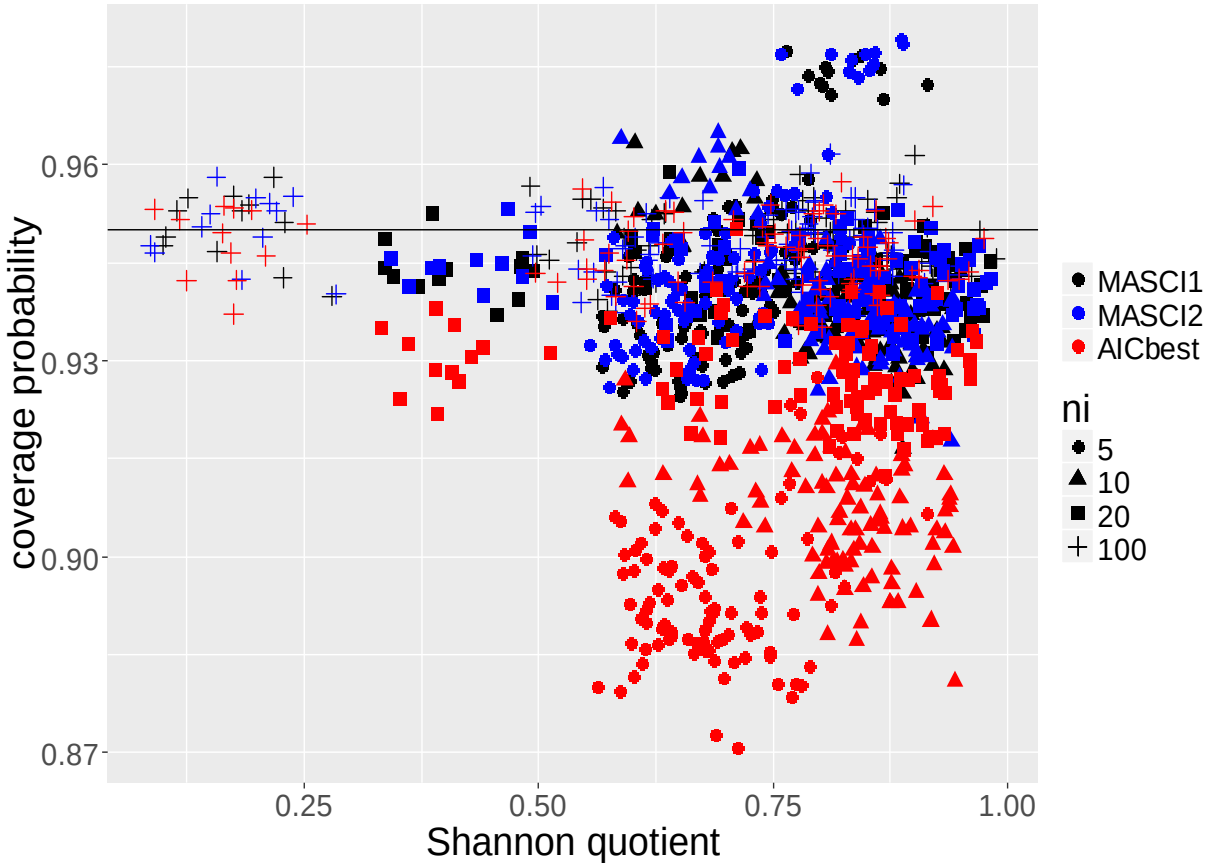

Figure 6.4: Coverage probability depending on the Shannon quotient for model-averaging (MASCI 1/2) and AICc selection method. Additionally parameter n_i is represented.

Between the two model-averaging methods, no serious difference could be found (see especially Figure 6.5), if at all MASCI2 performed slightly better. However, the difference to the AICc selection method was clear.

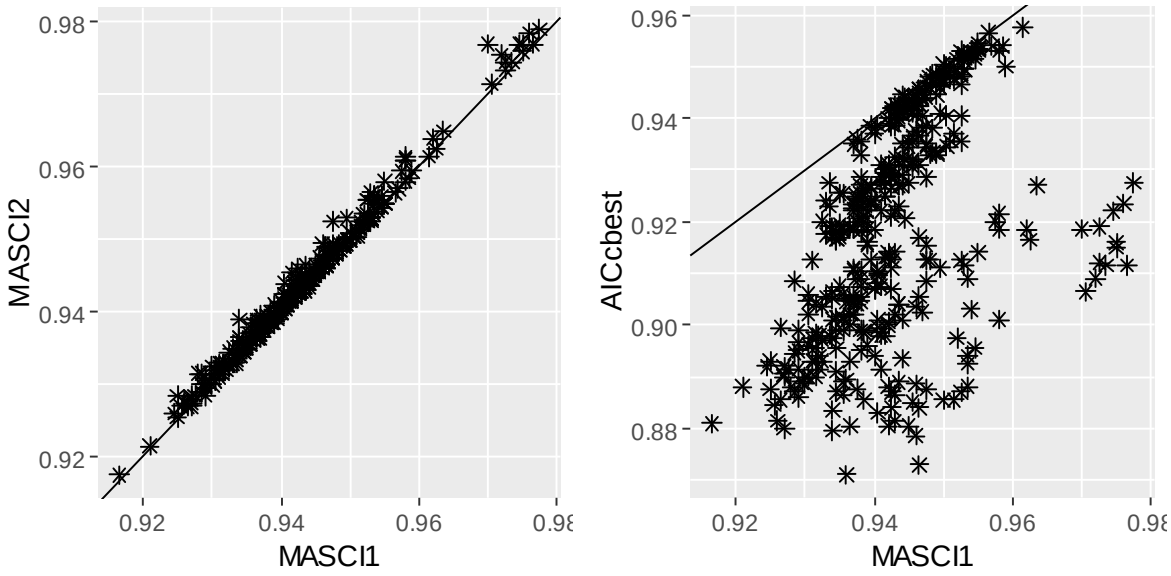

Figure 6.5: Direct comparison of coverage probabilities with model-averaged simultaneous confidence intervals 1/2 (MASCI1/2) and from the best model after AICc selection

Apart from some outliers, the final number of models in the sets approached approximately six. An influence of a lower number of models on the coverage probability cannot be seen.

6.3 Discussion

As in simulation study 1, an existing model selection uncertainty can be assumed in the various settings.

The coverage probability could be improved with MASCIs. Only for sample sizes $n_i = 100$ the AICc selected model performed equally well (Figure 6.2 and 6.4). Once more, MASCI2 achieved slightly better results than MASCI1 (Figure 6.5).

Unlike simulation study 1, no control was performed using the original equations (3.1). A comparable setting would be a comparison of two treatments at only two time points. However, a simulation with only two time points makes no sense since the selection among different correlation structures between several time points was the goal of this study.

7. Results Examples

This Chapter presents the results of the example data “Bodenfauna” (7.1) and “Bradykinin” (7.2) and each a brief diskussion.

7.1 Results Example „Bodenfauna“

As an illustration for the model selection uncertainty, the Akaike weights were calculated and shown in figure 7.1:

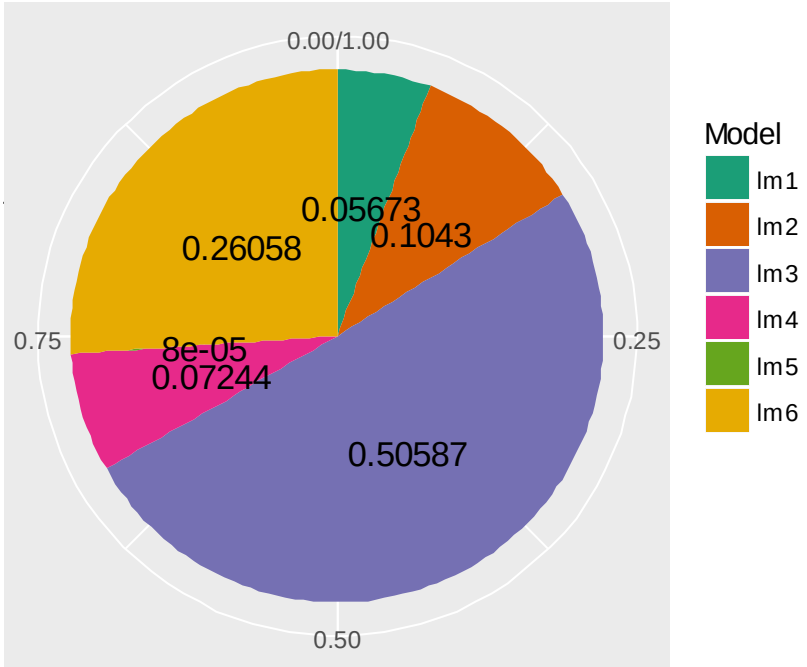

Figure 7.1: The calculated Akaike weights (see 2.1) of the set of candidate models

As you see in the pie chart (Figure 7.1), none of the built models reach really high weights (e.g > 0.95). To deal with this uncertainty, model-averaged simultaneous confidence intervals

(MASCI 1+2) according to chapter 4.2 were calculated for a Dunnett test. Figure 7.2 shows MASCI 1+2 in addition to the SCIs based on each single model for the tree comparisons.

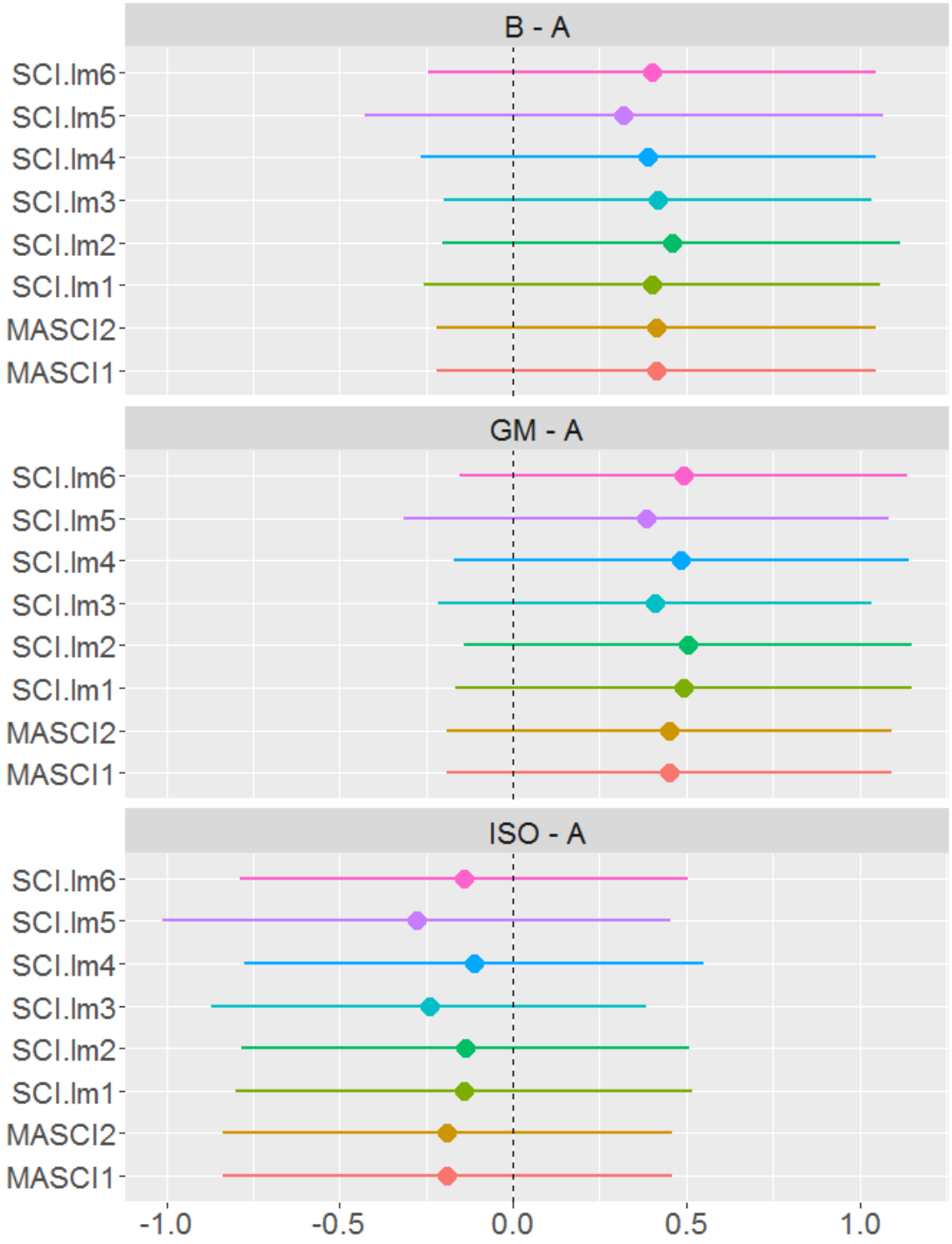

Figure 7.2: Model-averaged (MASCI 1+2) and single model (lm1-lm6) simultaneous confidence intervals of the multiple comparisons (Dunnett test) of the number of species in different plant varieties ("B – A", "GM – A", "ISO – A").

With none of the methods a difference in the number of species could be detected. The two model-averaged variants achieved same confidence intervals. For detailed values see Table 7.1.

Table 7.1: This table contains the values of several simultaneous confidence intervals of the multiple comparisons (Dunnett test) of the number of species in different plant varieties ("B – A", "GM – A", "ISO – A"). These were calculated for each single candidate model (SCI.lm1,..., SCI.lm6) and then model-averaged according to chapter 4.2 (MASCI1&2).

Comparison	Method/Model	Estimate	Lower bound	Upper bound
B - A	SCI.lm1	0,40095	-0,2574	1,0593
B - A	SCI.lm2	0,4583	-0,2024	1,119
B - A	SCI.lm3	0,41765	-0,1979	1,0332
B - A	SCI.lm4	0,3901	-0,2653	1,0455
B - A	SCI.lm5	0,3197	-0,4285	1,0679
B - A	SCI.lm6	0,40095	-0,2437	1,0456
B - A	MASCI1	0,4145847	-0,2196306	1,0488
B - A	MASCI2	0,41458335	-0,2197333	1,0489
GM - A	SCI.lm1	0,49315	-0,1652	1,1515
GM - A	SCI.lm2	0,5046	-0,1419	1,1511
GM - A	SCI.lm3	0,4098	-0,2142	1,0338
GM - A	SCI.lm4	0,48615	-0,1685	1,1408
GM - A	SCI.lm5	0,3853	-0,3139	1,0845
GM - A	SCI.lm6	0,4931	-0,1515	1,1377
GM - A	MASCI1	0,4516326	-0,1898348	1,0931
GM - A	MASCI2	0,45167455	-0,1897509	1,0931
ISO - A	SCI.lm1	-0,14125	-0,7996	0,5171
ISO - A	SCI.lm2	-0,13555	-0,7816	0,5105
ISO - A	SCI.lm3	-0,2409	-0,8686	0,3868
ISO - A	SCI.lm4	-0,11335	-0,7762	0,5495
ISO - A	SCI.lm5	-0,2781	-1,0096	0,4534
ISO - A	SCI.lm6	-0,14125	-0,7859	0,5034
ISO - A	MASCI1	-0,1890507	-0,8359014	0,4578
ISO - A	MASCI2	-0,1890608	-0,8358216	0,4577

7.1.1 Discussion "Bodenfauna"

As illustrated in Figure 7.1, none of the candidate models reached an Akaike weight of $w_i \geq 0.95$. Therefore, we can assume model selection uncertainty. Nevertheless, the conclusion for all methods are the same. But the different confidence bounds are clearly visible (Figure 7.2,

Table 7.1). Unlike the simulation studies, MASCI1 and MASCI2 achieved equal results. The width of the calculated intervals differed only after the third decimal (Table 7.1).

With a sample size of 8, this field trail was set in a range between the two n_i settings of simulation study 1 (n_i = 5 and 10). Sample sizes of 10 reached approximately 95 % coverage but n_i = 5 performed worse. When exactly the 95 % are reached, the study did not cover.

7.2 Results „Bradykinin"

For evaluation five models with different correlation structures were fitted. The calculated Akaike weights are shown in Figure 7.3. Model `fdt1´ reached only *w* = 0.00003 and therefore cannot be seen on the pie chart. The highest Akaike weight was achieved by model `fdtARMA02´ with *w* = 0.53.

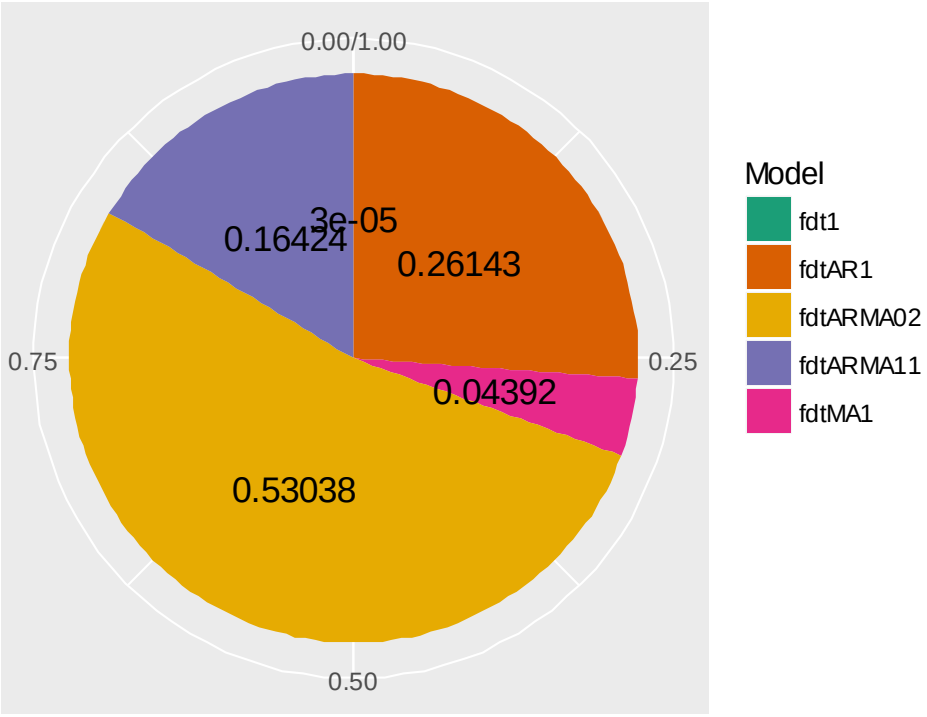

Figure 7.3: The calculated Akaike weights of the set of candidate models.

All treatments were compared at each time point, except baseline. The corresponding SCIs were calculated for each single model and using model-averaging. In Figure 7.4, the

calculated SCIs from the model with the lowest (fdt1) and highest (fdtARMA02) weight are shown together with both MASCIs.

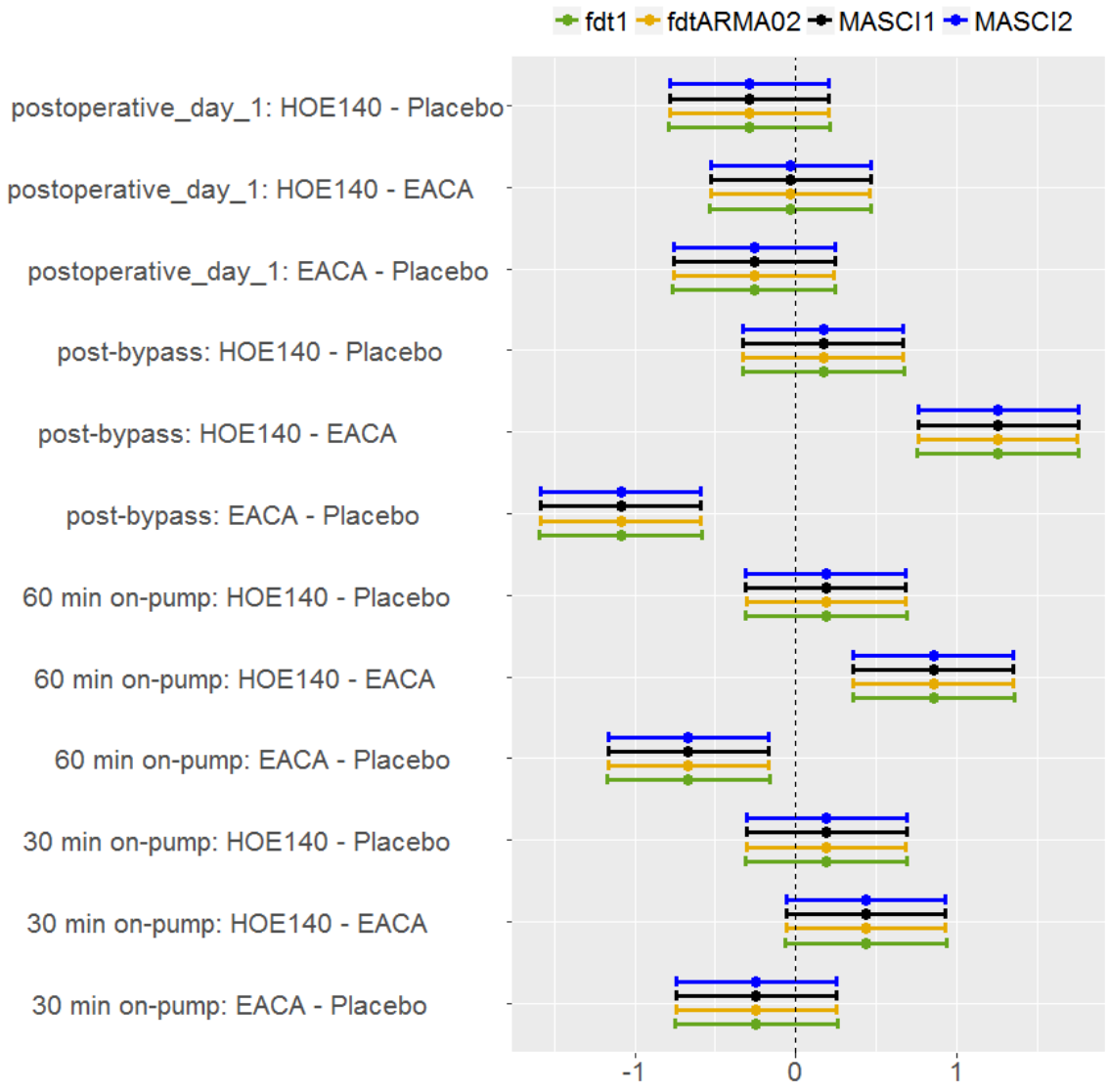

Figure 7.4: SCIs for all-pair comparison at different time points. For the two MASCIs and the SCIs based on the lowest (fdt1) and highest (fdtARMA02) weighted models.

7.2.1 Discussion

Despite assumable model selection uncertainty (Figure 7.3), there are hardly any differences between the confidence bounds (Figure 7.4). One reason could be the high sample size of 37 or 38 patients. In simulation study 2, only sample sizes of n_i = 5, 10, 20 or 100 were observed. The range around 37 was insufficiently covered. Consequently, we do not know how the difference between the methods at this point is or whether they behave as with n_i = 100.

MASCI1 and 2 performed equally well.

8. General Discussion

Extending model-averaging on multiple comparisons and finding a construction for model-averaged simultaneous confidence intervals was the goal of this thesis. The simulation studies showed that the proposed calculation of model-averaged simultaneous confidence intervals works. Two cases were considered: a linear model setting with unknown influence of several covariables and a setting with different correlation structures of linear mixed effects models. Model selection uncertainty could be assumed in both simulations. Model-averaging was compared against the often used AICc selection of one model. Especially for small sample sizes of 5 or 10 model-averaging performed better than using a single model after AICc selection. For instance, AICc selection achieved a coverage probability of 70 – 90 % in case of sample size $n_i = 5$, instead of the expected 95 %. Although MASCIs could not reach the 95 %, too, the coverage was always over 90 % und always better than the result of only one model. As is not to be expected otherwise, these asymptotic methods work better for high sample sizes.

This improvement over AICc selection has been descripted in the literature (e.g. Fletcher & Dillingham 2011). However, the confidence intervals were partly calculated differently. Fletcher and Turek (2011) recommend model-averaging a set of profile likelihood-intervals to improve coverage rate and avoid the assumption of normality of the estimates. Their simulation studies confirm an improved coverage compared to intervals using ase_1. For simulation, they made the assumption that the "true" model is the largest in the set. In 2012, Turek and Fletcher took a different approach averaging the tail areas of the distribution of the estimators of the possible models and building model-averaged tail area (MATA) intervals. For dose response modelling, Faes et al. (2007) and Jensen&Ritz (2015) used model-averaging methods, too.

Between MASCI 1 and 2, only marginal differences could be found in the simulation studies. In the examples, they differed only after several decimals.

Unlike the simulation studies, the results of model-averaged and single model SCIs did not differ in the examples. Naturally, it remains important to note that substantial fewer comparisons were considered. In the example "Bradykinin", 12 comparisons and in "Bodenfauna", only 3 comparisons were made for one data set. By contrast, in the simulation studies the multiple comparisons were made for 48 (study 1) or 384 (study 2) parameter settings with each 1000 or 2000 repetitions. This means a calculation of overall 48000 or 768000 different data sets and therefore multiple comparisons.

The model-averaging method seems to be promising and should be tested in further settings.

The calculation of MASCIs can now easily be made with a function implemented in R called "`masci`". How to use it and the additional code can be found in the Appendix.

9. Appendix

9.1 R Function for MASCI calculation

The above explained calculation for model-averaged simultaneous confidence intervals (MASCI) was implemented in an R function called "`masci`" for simplified usage. You can choose whether MASCI1 or 2 is calculated. The function requires the R-packages `MuMIn` (Barton 2015), `multcomp` (Hothorn et al. 2015) and `plyr` (Wickham 2015).

Step 1: Build a set of candidate models based on the data and previous knowledge. The entire set is used for further analysis. The K-L best models will reach the highest weights and if the "true" model is included its weight would approach one.

```
lm1<- lm(SP ~ Sorte + Block , data=Bodenfauna)
...
lm6<- lm(SP ~ Sorte, data = Bodenfauna)
```

Step 2: Chose a multiple comparison and carry it out for all models. So far it only works with R-packet `multcomp`. The results (e.g. the `glht` objects) must be collected in a list.

```
lm1.mc <- glht(lm1,linfct = mcp(Sorte = "Dunnett"))
...
lm6.mc <- glht(lm6,linfct=mcp(Sorte="Dunnett"))
FITLIST.mc <- list(lm1.mc, lm2.mc, lm3.mc, lm4.mc, lm5.mc, lm6.mc)
```

Step 3: The implemented R-function "`masci`" only needs this list for calculating the averaged simultaneous confidence intervals according to 4.2. As default the function calculate MASCI2, if `masci2 = FALSE` the MASCI1 will be calculated.

```
MASCI1.Bodenfauna <- masci(glht.list = FITLIST.mc,   masci2=FALSE)
MASCI2.Bodenfauna <- masci(glht.list=FITLIST.mc)
```

The output contains the `Estimate` of treatment differences and the lower (`lwr`) and upper (`upr`) confidence bounds.

```
> MASCI1.Bodenfauna
           Estimate        lwr       upr
B - A     0.4146017 -0.2195325 1.0487358
GM - A    0.4516512 -0.1897356 1.0930380
ISO - A -0.1890597 -0.8358023 0.4576828
```

masci code

```
masci <- function(glht.list, masci2= TRUE)
{
  require("MuMIn")
  require("multcomp")
  require("plyr")

  RES <- matrix(NA, ncol=3, nrow=length((Fitlist.mc [[1]]$linfct)[,1]))
  length.mc <- length((glht.list [[1]]$linfct)[,1])
  Fitlist<-lapply(glht.list, function(x){x$model})
{
  W1   <- Weights(do.call(what="AICc", args=Fitlist))
  a <- 1-((1-(1-0.05))/2)

  x.Fit <- laply(.data=glht.list,.fun=function(x){coefficients(x)})
  se.Fit <- laply(.data=glht.list,.fun=function(x){sqrt(diag(vcov(x)))})

  confint.glht.list <- lapply(glht.list, function(x){confint(x, level = 0.95)})
  t.Fit <- unlist(lapply(confint.glht.list, function(x){attr(x$confint,
which="calpha")}))

  wx <- aaply(x.Fit, 2, function(x){weighted.mean(x, W1, na.rm = TRUE)})
  x.sqdiff <- aaply(x.Fit, 1, function(x){x -wx})
  z <- c(t.Fit/qnorm(a))^2
  xvar <- se.Fit^2
  x.for.ase<-(xvar*z)+x.sqdiff^2

if (masci2){
  ase <- aaply(x.for.ase, 2, function(x){sqrt(weighted.mean(x, W1, na.rm = TRUE))})
  }

else{
  sqrt.x.for.ase <- aaply(x.for.ase, 1, function(x){sqrt(x)})
  ase <- aaply(sqrt.x.for.ase, 2, function(x){weighted.mean(x, W1, na.rm = TRUE)})
  }

  ci <- qnorm(a,lower.tail=TRUE)* ase

  lwr <- as.numeric(wx-ci)
  upr <- as.numeric(wx+ci)
  CI <- cbind(lwr, upr)

  glht.object1 <-(FITLIST.mc[1])
  sumglht1<-summary(glht.object1[[1]])
  testglht1<-sumglht1$test
  coefglht1 <-testglht1$coefficients
  namesMC <- names(coefglht1)
  RES <- cbind(wx,CI)
}
rownames(RES)<-namesMC
colnames(RES)<-c("Estimate", "lwr", "upr")
return (RES)
}
```

9.2 R Code Simulation Studies

```
##############################################################################

###### Simulation Study 1: Covariables ###################################
##############################################################################

library(MuMIn)
library(multcomp)
library(MASS)
library(plyr)
library(MCPAN)

# ni: integer vector, elements= sample size per group, length=number of
groups
# names(ni)
# a: single numeric value (intercept in group 1)
# d: vector of numeric values; differences to group 1

#### Data Generation

dancovaCOV<-function(ni, a, d, b, cov, sde)
{
  require(mvtnorm)
  if(length(d)+1 != length(ni)){warning("Length of 'ni' and length of 'd'
do not fit")}
  if(is.null(names(ni))){TRT   <-   c("C",   paste("T",   1:(length(ni)-1),
sep=""))}else{TRT <- names(ni)}
  N <- sum(ni)
  ncov<-nrow(cov)
  X <- rmvnorm(n=N, mean=rep(0, ncov), sigma=cov)
  xnam <- paste("x", 1:ncov, sep="")
  colnames(X) <- xnam
  bvec <- rep(b, length.out=ncov)
  fb <- factor(rep(TRT, times=ni))
  dat <- data.frame("Treat"=fb, X)
  form1 <- as.formula(paste(" ~ Treat + ", paste(xnam, collapse=" + ")))
  #print(form1)
  mm    <-    model.matrix(form1,    data=dat)    #    statt:    mm    <-
model.matrix(formula=form1, data=dat)
  para <- matrix(c(a, d, bvec), ncol=1)
  epsilon <- rnorm(n=N, mean=0, sd=sde)
  y<-mm %*% para + epsilon

  return(data.frame(y=y, Treat=fb,  X))
}

#### Model Set
# 1)
indexlist <- function(nox, maxx){
  indx <- 1:nox
  indlist <- NULL
  for(i in 1:maxx){
    indlist <- c(indlist, as.list(data.frame(combn(x=indx,m=i))))
  }
  return(indlist)
}

# 2) all possible models were built

formulist <- function(indexlist, xnames=NULL, trtname="Treat", yname="y",
add0=TRUE){
  if(is.null(xnames)){XNAM    <-    paste("x",    1:max(unlist(indexlist)),
sep="")}else{XNAM <- xnames}
  FORMLIST <- lapply(indexlist, FUN=function(x){
```

```
    paste(yname, " ~ ", trtname, " + ", paste(XNAM[x], collapse=" +
"),sep="")})
  if(add0){FORMLIST <- c(list("x0"= paste(yname, " ~ ", trtname, sep="")),
FORMLIST)}
  names(FORMLIST)<-NULL
  return(FORMLIST)
}


#### Simulation Code
SIMmulti<- function(nsim=nsim, ni=ni, a, d, b, cov, sde, maxx=maxx)
{
  extractcoef <- function(fitlist, which){ unlist(lapply(fitlist,
FUN=function(x){coefficients(x)[which]})) }
  extractcoefE <- function(fitlist, which){ unlist(lapply(fitlist,
FUN=function(x){coefficients(summary(x))[which,2]})) }
  extract.probs <- function(res, nsim){colSums(res)/nsim}
  RES<-matrix(NA, ncol=11, nrow=nsim)
  NOX <- length(b)
  ALLI <- indexlist(nox=NOX,maxx=maxx)
  FORMI <- formulist(indexlist=ALLI)

  FORMULA2 <- FORMI[[1]]                 # Ohne Kovariablen

  for(i in 1:nsim)
  {
    DAT  <- dancovaCOV(ni=ni, a=a, d=d, b=b, cov=cov, sde=sde)


    FORMIall<-paste( "y ~ Treat+", paste(paste("x", 1:length(b), sep=""),
collapse=" + "))

    FIT1 <- lm(as.formula(FORMIall), data=DAT)
    FIT2 <- lm(as.formula(FORMULA2), data=DAT)

    FITLIST <- lapply(FORMI, FUN=function(x){lm(as.formula(x), data=DAT)})

    FITLIST.mc <- llply(FITLIST, .fun=function(FITLIST){glht(FITLIST,linfct
= mcp(Treat = "Dunnett"))})

    models <- length(FITLIST)

    W   <- Weights(do.call(what="AICc", args=FITLIST))
    W1 <- sort((round(W, 4)), decreasing=TRUE)[1]
    W2 <- sort((round(W, 4)), decreasing=TRUE)[2]
    W3 <- sort((round(W, 4)), decreasing=TRUE)[3]
    Shannon.wi <- Shannon(W)

    max.Shannon <- log(length(FITLIST))

    masci1 <- masci(glht.list = FITLIST, masci2=FALSE)
    masci2 <- masci(glht.list = FITLIST)

    CIall <- confint(glht((aov(FIT1, data=DAT)),linfct = mcp(Treat =
      "Dunnett")))
    FITstepAIC <- stepAIC(FIT1, scope = list(lower = ~ Treat), trace=0)

    CIstepAIC <- confint(glht(FITstepAIC,linfct = mcp(Treat = "Dunnett")))
    CIign <- confint(glht((aov(FIT2, data=DAT)),linfct = mcp(Treat =
      "Dunnett")))
    enthaltenmasci1 <- all(masci1[,1]<=d & masci2[,2]>=d)
    enthaltenmasci2 <- all(masci2[,1]<=d & masci2[,2]>=d)
    enthaltenSTEPAIC <- all(CIstepAIC$confint[,2]<=d &
      CIstepAIC$confint[,3]>=d)
    enthaltenALL <- all((CIall$confint[,2]<=d & CIall$confint[,3]>=d))
    enthaltenIGN <- all((CIign$confint[,2]<=d & CIign$confint[,3]>=d))

RES[i,]<-c(enthaltenmasci1, enthaltenmasci2, enthaltenSTEPAIC,
enthaltenALL, enthaltenIGN,W1,W2,W3, Shannon.wi, max.Shannon, model
  }
```

```
  RES <- as.data.frame(RES)
  coverage.probabilities <- extract.probs(res=RES[1:5],nsim=nsim)
  MeanW <- colMeans(RES[6:11])

  names(coverage.probabilities) <-c("MASCI1","MASCI2", "STEP", "ALL","IGN")
  names(MeanW) <- c("MW W1", "MW W2", "MW W3", "MW Shannon", "max Shannon",
      "models")

  return(c(coverage.probabilities, MeanW))

}

### Parameter Setting

Settings <- expand.grid(ni=c(5,10,20,50), sde=c(0.1,3,10),
b=c("b4","b6","b8", "b10"))

b4=c(1,0,-2,0)
b6=c(1,0,-2,0,0,2)
b8=c(1,0,0,-1,0,0,2,0)
b10=c(0,0,0,1,0,0,-1,0,2,0)




##############################################################################
###### Simulation Study 2: Repeated Measurements ############################

##############################################################################

library(nlme)
library(mvtnorm)
library(ggplot2)
library(multcomp)
library(MuMIn)
library(plyr)
library(MCPAN)


# repeated measures: cross-over-design
# jedem Individuen werden über die Zeit alle Behandlungen von Interesse
zugewiesen,
# die zeitliche Reihenfolge der Behandlungen

# ni: Anzahl Individuen
# nt: Anzahl Zeipunkte=Anzahl Behandlungen
# m Mittelwert in Behandlung A
# d Mittelwertsdifferenzen zu Behandlung A, B-A, C-A, ...
# sdi: stddev. between individuale
# sde: residual stddev
# cor ... noch nicht: Korrelastionsstruktur wie in Pinheiro und Bates(2000)


#Data Generation:

drepco <- function(ni, nt, m, d, sdi, sde, cor=NULL, print=FALSE)
{
  require("mvtnorm")
  require("nlme")

  id <- paste("ID", 1:ni, sep="")
  idf <- factor(id, levels=id)
  trt <- LETTERS[1:nt]
  trtf <- factor(trt, levels=trt)
  idf2 <- rep(idf, each=nt)
  trtf2 <- as.vector(replicate(n=ni, sample(trtf)))
  time <- 1:nt
  time2 <- rep(time, times=ni)
```

```
  mmX <- model.matrix(~trtf2)
  MTRT2  <- mmX %*% matrix(c(m,d),ncol=1)
  UID <- rnorm(n=ni,0, sd=sdi)
  UID2 <- rep(UID, each=nt)

  DF2 <- data.frame(resp=MTRT2, trt=trtf2, id=idf2, time=time2)

  if(is.null(cor)){COR <- corCompSymm(0, ~1|id)}else{COR <- cor}

  SDE <- diag(rep(sde, length=nt))
  CORLIST2 <- corMatrix(Initialize(COR, data=DF2))
  SIGLIST2 <- lapply(CORLIST2,function(x){SDE %*% x %*% SDE})
  RESLIST2 <- lapply(SIGLIST2, function(x){rmvnorm(n=1, sigma=x)})


  if(print){
    print(CORLIST2[[1]])
    print(SIGLIST2[[1]])
  }
  RES2 <- unlist(RESLIST2)

  DF2$resp <- MTRT2 + UID2 + RES2

  return(DF2)

}

# Simulation Code

SIMrepeatedMeasures<- function(nsim=nsim, ni, nt, m, d, sdi, sde,
cor,print)
{
  extractcoef <- function(fitlist, which){ unlist(lapply(fitlist,
FUN=function(x){coefficients(x)[which]})) }
  extractcoefE <- function(fitlist, which){ unlist(lapply(fitlist,
FUN=function(x){coefficients(summary(x))[which,2]})) }
  extract.probs <- function(res, nsim){colSums(res)/nsim}

  RES<-matrix(NA, ncol=9, nrow=nsim)


  for(i in 1:nsim)
  {
    DAT  <- drepco(ni=ni, nt=nt, m=m, d=d, sdi=sdi, sde=sde, cor)

    #fit <- try(lme(resp ~ trt, data=DAT, random= ~1|id))

    # Modellfit

    FITLISTpos <- lapply(corlist, FUN=function(x){try(lme(resp ~ trt,
data=DAT, random= ~1|id, correlation=x), silent=TRUE)})
    FITclass <-unlist(lapply(FITLISTpos, class))
    wsucc <- which(FITclass == "lme")
    FITLIST <- FITLISTpos[wsucc]# Liste aller Modellfits

    FITLIST.mc <- llply(FITLIST, .fun=function(FITLIST){glht(FITLIST,linfct
= mcp(trt = "Dunnett"), df=FITLIST$fixDF$terms[2])})

    models <- length(FITLIST)

    W   <- Weights(do.call(what="AICc", args=FITLIST))
    W1 <- sort((round(W, 4)), decreasing=TRUE)[1]
    W2 <- sort((round(W, 4)), decreasing=TRUE)[2]
    W3 <- sort((round(W, 4)), decreasing=TRUE)[3]
    Shannon.wi <- Shannon(W)

    max.Shannon <- log(length(FITLIST))

    masci1 <- masci(glht.list = FITLIST.mc, masci2=FALSE)
    masci2 <- multi.par.avg.2004(glht.list=FITLIST.mc)
```

```
    #AICc best
    mstAICc <- model.sel(FITLIST) #model selection table
    FitAICcbest <- get.models(mstAICc, subset = 1)[[1]]

    AICcSCI <- confint(glht(FitAICcbest,linfct = mcp(trt = "Dunnett")))

    enthaltenmasci1 <- all(masci1[,1]<=d & masci1[,2]>=d)
    enthaltenmasci2 <- all(masci2[,1]<=d & masci2[,2]>=d)
    enthaltenAIC <- all(AICcSCI$confint[,2]<=d & AICcSCI$confint[,3]>=d)


    RES[i,]<-c(enthaltenmasci1, enthaltenmasci2,enthaltenAIC, W1,W2,W3,
      Shannon.wi, max.Shannon, models)

  }
  RES <- as.data.frame(RES)

  coverage.probabilities <- extract.probs(res=RES[1:3],nsim=nsim)
  MeanW <- colMeans(RES[4:9])

  names(coverage.probabilities) <- c("MASCI1","MASCI2","AICcbest")
  names(MeanW) <- c("MW W1", "MW W2", "MW W3", "MW Shannon", "max Shannon",
      "models")

  return(c(coverage.probabilities, MeanW))

}

# Correlation Structures for Data Generation
"NULL"    <-corNULL <- NULL
"AR1"     <-corAR1(0.8, ~1|id)
"AR1low"  <-corAR1(0.3, ~1|id)
"ARMA1"   <-corARMA(0.8,  ~1|id, p=0, q=1)
"ARMA1low"<-corARMA(0.3, ~1|id, p=0, q=1)
"ARMA2"   <-corARMA(c(0.6,0.8), ~1|id, p=1, q=1)
"ARMA2low"<-corARMA(c(0.1,0.3), ~1|id, p=1, q=1)
"UNSTR"   <-corSymm(c(0.8,0.4,0.2,0.0, 0.4,0.2,0.0, 0.2,0.0, 0.0), ~1|id)

# Correlation Structures for Model Building
cor1<-NULL
cor2<-corAR1(0.5, ~1|id)
cor3<-corAR1(0.2, ~1|id)
cor4<-corARMA(0.4, ~1|id, p=0, q=1)
cor5<-corARMA(c(0.7, 0.7), ~1|id, p=0, q=2)
cor6<-corARMA(c(0.3,0.5), ~1|id, p=1, q=1)

corlist <- list(cor1, cor2, cor3, cor4, cor5, cor6)

# Parameter settings
paradat <-  expand.grid(ni=c(5, 10, 20, 100), sdi=c(0.01, 0.5, 1, 5),
sde=c(0.5, 1, 10))
```

9.3 Example data

Table 9.1: "Bodenfauna" data

Block	Variety	Species	pH	KF1	KF2	KF3	C_{org}	N
1	ISO	19	5,3	66,6	25,9	7,5	0,69	0,062
2	B	38	5,4	53,9	36,4	9,6	0,78	0,070
3	A	99	5,6	49,7	39,8	10,5	0,81	0,077
4	GM	27	5,8	47,1	41,9	11,1	0,83	0,074
5	ISO	18	5,8	49,7	40,3	10,0	0,71	0,067
6	GM	40	5,3	48,5	41,0	7,7	0,67	0,065
7	A	56	5,6	48,9	40,7	8,8	0,72	0,070
8	ISO	16	5,4	60,0	29,1	10,9	0,95	0,087
1	B	40	5,2	68,2	25,6	6,2	0,77	0,067
2	A	19	5,5	63,8	27,7	8,6	0,95	0,084
3	GM	114	5,7	47,1	41,7	11,2	0,85	0,081
4	ISO	40	5,7	47,2	41,5	11,3	0,84	0,073
5	B	30	5,6	54,5	36,3	9,2	0,77	0,080
6	A	33	5,3	59,1	32,8	8,1	0,72	0,069
7	GM	34	5,3	57,1	34,3	8,6	0,83	0,080
8	B	63	5,2	57,9	33,7	8,4	1,00	0,092
1	A	24	5,1	68,1	24,8	7,1	0,86	0,079
2	GM	66	5,5	56,3	33,8	10,0	0,83	0,081
3	ISO	70	5,6	46,2	44,0	9,8	0,82	0,073
4	B	55	5,7	57,6	33,9	8,5	0,95	0,081
5	A	38	5,6	51,6	38,0	10,4	0,79	0,076
6	ISO	25	5,3	55,5	35,6	8,8	0,77	0,076
7	B	82	5,4	55,2	35,5	9,3	0,74	0,074
8	GM	62	5,1	63,2	27,7	9,2	0,97	0,089
1	GM	51	5,4	57,2	33,0	9,7	0,80	0,076
2	ISO	55	5,5	44,8	41,8	16,9	0,84	0,078
3	B	78	5,4	59,8	30,7	9,5	1,02	0,097
4	A	27	5,9	58,9	31,5	9,6	1,02	0,099
5	GM	162	5,8	59,2	31,2	9,5	0,98	0,087
6	B	67	5,2	56,7	34,8	8,5	0,78	0,073
7	ISO	42	5,4	61,1	30,5	8,4	0,69	0,069
8	A	32	5,5	59,0	29,7	11,2	0,80	0,076

Table 9.2: Bradykinin data

ID	Drug	Baseline	bypass_30_min	bypass_60_min	post_bypass	postoperativ_day
1	Placebo	5,0328	5,3976	5,7022	6,9740	6,6857
2	Placebo	3,2558	4,6634	4,9376	4,7945	6,1281
3	Placebo	4,3291	4,9985	5,3642	4,6745	5,6563
4	Placebo	5,0445	5,6254	6,8620	7,1506	6,0146
5	Placebo	4,6330	5,3544	5,0056	5,7522	5,4053
6	Placebo	4,4899	4,7122	5,5447	5,9043	5,1270
7	Placebo	5,5365	5,4413	4,5876	6,6356	5,2526
8	Placebo	4,7983	5,0799	5,2646	6,6610	6,8176

9	Placebo	5,6800	6,3052	5,8627	7,5499	5,8322
10	Placebo	4,2362	4,6818	5,1980	5,4522	5,7017
11	Placebo	4,6783	5,3148	5,4520	6,3755	5,9469
12	Placebo	4,6946	4,8387	6,0578	5,2295	5,0096
13	Placebo	5,0146	4,6146	5,7012	4,5909	5,2785
14	Placebo	2,9159	4,1731	4,1696	6,6341	5,2684
15	Placebo	4,7121	3,9182	5,0801	5,7692	6,0655
16	Placebo	4,5448	4,9799	5,0404	6,5512	6,3418
17	Placebo	3,7075	4,4197	5,4676	5,6720	4,7158
18	Placebo	4,0451	4,2258	5,2906	5,4650	5,7607
19	Placebo	4,7075	5,4494	6,2947	6,3584	5,4620
20	Placebo	4,5409	4,6756	5,2056	5,4788	5,4775
21	Placebo	5,4544	5,6356	6,9267	5,7817	6,3965
22	Placebo	3,5622	4,6621	4,6821	5,5313	5,8226
23	Placebo	4,0237	4,1471	4,4944	4,8059	5,5482
24	Placebo	4,0242	4,6737	4,0138	4,9760	6,3968
25	Placebo	4,8555	5,2929	5,5312	6,8874	6,7966
26	Placebo	4,4740	4,9425	6,4587	5,6305	5,4560
27	Placebo	3,3500	3,7805	6,5075	6,6543	5,6083
28	Placebo	5,4953	5,7717	6,0502	5,9109	6,5771
29	Placebo	4,3722	4,1659	5,5897	6,1666	6,6083
30	Placebo	3,2259	4,4792	6,2273	6,3012	5,8045
31	Placebo	3,9048	4,0390	4,3449	5,9357	5,3479
32	Placebo	5,1379	5,1654	5,5610	6,3409	6,9182
33	Placebo	4,4870	4,9854	5,5398	5,6722	5,2836
34	Placebo	3,8467	4,0678	4,7957	7,1809	5,4543
35	Placebo	5,3440	5,2924	4,3547	7,2827	6,0409
36	Placebo	5,1262	4,0154	4,6562	5,2560	6,8244
37	Placebo	4,5716	5,2354	4,8145	5,7797	5,1237
38	EACA	3,6417	4,2443	4,3455	4,1988	5,9690
39	EACA	5,2347	5,4318	5,3544	5,0166	6,3620
40	EACA	2,6799	3,1637	2,7691	4,1240	5,7665
41	EACA	4,7959	4,7436	4,6865	4,6506	6,1241
42	EACA	5,2384	5,3986	5,3641	6,3473	6,9881
43	EACA	3,5313	3,3329	3,6316	4,2515	6,2711
44	EACA	5,4374	5,4060	5,3116	5,1482	5,7317
45	EACA	4,7497	5,2387	5,6755	5,9061	6,4901
46	EACA	3,9994	4,5213	4,4003	4,9266	4,7635
47	EACA	4,8602	5,1202	5,3856	4,6021	4,4033
48	EACA	3,9622	4,8755	5,2243	5,4294	6,4295
49	EACA	3,9181	4,7215	4,8748	4,8122	4,4851
50	EACA	3,3999	3,8630	4,6350	4,7032	4,8127
51	EACA	4,9221	5,1586	5,5928	5,3106	4,7746
52	EACA	5,2920	5,3436	5,4094	5,0067	4,4851
53	EACA	4,5566	5,4543	5,3121	5,4264	5,7615
54	EACA	4,7541	4,3777	4,9523	4,7395	5,4966
55	EACA	5,4079	5,1810	5,2174	5,3912	6,6474

56	EACA	5,1635	5,4551	5,5464	6,1374	6,1371
57	EACA	3,9581	4,3298	4,1104	5,0314	6,0722
58	EACA	3,4613	3,8509	4,4477	4,7059	4,6487
59	EACA	3,7405	4,2339	4,4179	5,1570	5,7681
60	EACA	5,6767	5,6736	5,7947	5,6614	6,8070
61	EACA	4,3312	4,7918	5,0311	5,2973	4,9566
62	EACA	4,3916	4,5870	4,1308	4,4881	5,0201
63	EACA	4,7908	4,2328	4,6065	4,7723	5,5932
64	EACA	4,0540	4,3076	4,3098	4,4834	4,6896
65	EACA	3,1456	2,8166	2,8126	2,9901	5,5030
66	EACA	3,7806	3,5349	3,3830	4,2426	5,1109
67	EACA	3,1163	3,5123	3,0519	3,7246	5,4771
68	EACA	5,3308	5,2923	5,6081	5,4353	4,3036
69	EACA	3,8724	3,3093	3,7313	4,0622	6,2560
70	EACA	4,2644	4,3283	3,7426	4,0981	6,1142
71	EACA	6,1016	5,7934	5,8101	5,3895	5,4321
72	EACA	4,0072	4,5687	4,8483	5,2821	5,3571
73	EACA	5,5426	5,6376	6,0201	4,6867	5,1257
74	EACA	3,2589	4,3041	4,3615	5,8699	6,2428
75	HOE140	4,5445	4,8873	5,4660	6,3073	5,5362
76	HOE140	3,9803	4,8262	5,4041	6,2802	5,1087
77	HOE140	3,9903	4,3823	5,4635	5,4369	4,8006
78	HOE140	3,7487	4,3809	4,4025	5,0761	5,4158
79	HOE140	4,5056	6,4541	7,2489	5,2038	5,2371
80	HOE140	4,7335	4,5281	4,9957	5,9588	5,5508
81	HOE140	3,8898	5,5183	6,1211	4,6945	4,5179
82	HOE140	5,0462	5,0754	4,7602	6,8555	6,1951
83	HOE140	4,2530	5,3592	5,8218	7,3369	5,5629
84	HOE140	5,2032	4,1962	4,6478	4,7834	5,6296
85	HOE140	4,8197	4,6761	4,6860	7,1969	6,2184
86	HOE140	3,8338	3,2804	4,4123	3,2687	5,3690
87	HOE140	4,3278	5,4680	7,8066	5,9094	6,4212
88	HOE140	4,7870	5,4737	6,2346	7,5477	5,5188
89	HOE140	5,3246	7,6445	9,1348	8,1383	5,4224
90	HOE140	4,4823	5,2752	5,6341	7,2637	6,0082
91	HOE140	4,8353	5,3661	5,7152	7,9531	4,9737
92	HOE140	4,0152	4,3153	4,0224	7,0902	5,7680
93	HOE140	5,2540	5,8040	5,2497	7,5609	6,5343
94	HOE140	4,5069	5,4298	5,0879	5,0988	5,7403
95	HOE140	5,2905	5,7389	5,3995	6,4169	4,7768
96	HOE140	3,7707	4,9903	5,8444	5,9301	6,1398
97	HOE140	4,5504	6,0234	7,2443	5,4204	5,8115
98	HOE140	3,4947	5,4485	6,5225	6,7342	6,5223
99	HOE140	4,7115	4,3043	5,4028	5,6985	5,5124
100	HOE140	4,2983	4,2313	4,4258	6,1352	5,1057
101	HOE140	3,9304	4,8874	4,9906	6,4953	4,9962
102	HOE140	4,7710	4,8861	5,6303	6,6411	5,9891

103	HOE140	4,6426	5,3947	5,6087	7,6143	5,7604
104	HOE140	3,2668	4,3548	4,4293	6,2312	5,1774
105	HOE140	4,0637	2,8750	3,9118	3,6200	5,1403
106	HOE140	5,1109	5,3725	5,5433	5,4071	4,6849
107	HOE140	4,0725	4,7567	5,0518	4,8867	5,2484
108	HOE140	4,6810	5,2553	5,8034	4,6827	5,9189
109	HOE140	3,9481	5,5836	5,8453	8,1567	5,2388
110	HOE140	4,3871	5,2834	6,1268	6,6257	5,6175
111	HOE140	4,5206	5,8390	6,3119	6,7712	5,6876
112	HOE140	4,3511	3,7497	4,6832	5,7740	5,9268

10 References

Anderson, D.R., Burnham, K.P., and White, G.C. (1994). AIC model selection in overdispersed capture–recapture data. *Ecology* **75**, 1780–1793.

J. M. Balaguer, C. Yu, J. G. Byrne, S. K. Ball, M. R. Petracek, N. J. Brown, and M. Pretorius (2013). Contribution of endogeneous bradykinin to _brinolysis, inammation, and blood product transfusion following cardiac surgery: a randomized clinical trial. *Clinical Pharmacology and Therapeutics*, 93(4):326{334.

Barton, K.(2015). MuMIn: Multi-model inference, 2015. URL http://CRAN.R-project.org/package=MuMIn. R package version 1.15.1.

Buckland, S.T., Burnham, K.P., and Augustin, N.H. (1997). Model selection: an integral part of inference. Biometrics 53, 603–618.

Burnham,Kenneth P. and David R. Anderson. 1998. *Model Selection and Inference: A Practical Information-Theoretical Approach*. New York: Springer-Verlag.

Burnham, K. P ., & Anderson, D. R. (2001). Kullback–Leibler information as a basis for strong inference in ecological studies. *Wildlife Research*, **28**, 111-119.

Burnham KP, Anderson DR (2002). *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd edn. Springer, New York

Burnham KP, Anderson DR (2004) Multimodel inference: understanding AIC and BIC in model selection. *Sociol Methods* Res 33:261–304

Chatfield, C., (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society*. Series A (Statistics in Society) 158 (3), 419–466.

Claeskens, G., and Hjort, N. L. (2008), *Model Selection and Model Averaging*. *Cambridge Series on Statistical and Probabilistic Mathematics*, Vol. 27. Cambridge: Cambridge University Press, xvii, 312.

Dunnett, C. W. (1955), A multiple comparison procedure for comparing several treatments with a control, *Journal of the American Statistical Association*, 50, 1096{1121.

Faes C, Aerts M, Geys H, Molenberghs G. (2007). Model averaging using fractional polynomials to estimate a safe level of exposure. *Risk Analysis*, 27:111–123.

Fletcher, D., Dillingham, P., (2011). Model-averaged confidence intervals for factorial experiments. *Computational Statistics & Data Analysis* 55: 3041-3048.

Fletcher, D. and Turek, D. (2011), "Model-Averaged Profile Likelihood Intervals," *Journal of Agricultural, Biological, and Environmental Statistics*, 17 (1), 38–51.

Hjort, N. L., and Claeskens, G. (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98 (464), 879–899.

Hothorn, T., Bretz, F., Westfall, P., 2008. Simultaneous inference in general parametric models. *Biometrical Journal* 50, 346–363.

Hothorn, T., Bretz, F., Westfall, P., Heiberger, R.M., Schützenmeister, A. and Scheibe, S. (2015) multcomp: Simultaneous inference in general parametric models, 2015. URL http://CRAN.R-project.org/package=multcomp. R package version 1.4-1.

Hurvich, C. M., and C.-L. Tsai. (1989). Regression and time series model selection in small samples. *Biometrika* 76:297–307.

Hurvich, C. M., and Tsai, C. L. (1990), "The Impact of Model Selection on Inference in Linear Regression," *The American Statistician*, 44 (3), 214–217.

Hurvich, C. M., Tsai, C.-L. (1991) Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika* **78**, 499–509.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79-86.

Lukacs, P. M., Burnham, K. P., and Anderson, D. R. (2010), Model Selection Bias and Freedman's Paradox, *Annals of the Institute of Statistical Mathematics*, 62 (1), 117–125.

Pinheiro, J. C., Bates, D. M. (2000). *Mixed-effect models in S and S-PLUS*. Springer Verlag, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6:461–464.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. Commun. *Stat. Theory Methods* A7:13–26.

Tukey, J. W. (1953), The Problem of Multiple Comparisons, unpublished manuscript reprinted in: The Collected Works of John W. Tukey, Volume 8, 1994, H. I. Braun (Ed.), Chapman and Hall, New York. 100.

Turek, D., Fletcher, D. (2012), "Model-averaged Wald confidence intervals," *Computational Statistics and Data Analysis*, 56, 2809–2815.

Wickham, H. (2015) *plyr: Tools for Splitting, Applying and Combinding Data, 2015.* URL: http://CRAN.R-project.org/package=plyr. R package version: 1.8.3

Wickham, H. and W. Chang (2015). *ggplot2: An implementation of the Grammar of Graphics*, 2015. URL http://CRAN.R-project.org/package=ggplot2. R package version 1.0.1.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., Smith, G.M. (2009) *Mixed Effects Models and Extentions in Ecology with R*, Statistics for Biology and Health, Springer Science+Business Media, LLC 2009

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht sind und dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt wurde.

Hannover, den 05.12.2016 _____