Frequentist Model-Averaged Confidence Intervals

Daniel B. Turek

A thesis submitted for the degree of

Doctor of Philosophy

at the University of Otago, Dunedin,

New Zealand

April 2013

Abstract

Model averaging is a technique used to account for model uncertainty in the process of multimodel inference. A frequentist model-averaged estimator is defined as a weighted average of the single-model estimates resulting from each candidate model. We may also desire a model-averaged confidence interval, which similarly accounts for the uncertainty in model selection. Preexisting constructions for such an interval have assumed a normal sampling distribution for model-averaged estimators, and thereby construct Wald intervals centered around the model-averaged estimate. This approach is problematic, since the assumption of normality is generally incorrect. Furthermore, it relies upon accurate estimation of the standard error term, the form of which is not well understood.

We propose and study a new approach to the construction of frequentist model-averaged confidence intervals, which is analogous to that of a credible interval in Bayesian model averaging. This new construction is called a model-averaged tail area (MATA) interval, since it involves a weighted average of single-model nominal error rates. Through a variety of simulation studies, we compare MATA intervals against the preexisting approach of centering a Wald interval around a model-averaged estimate, and also against Bayesian model averaging. Intervals are assessed in terms of their achieved coverage rate and relative width. We consider several information criteria for the construction of frequentist model weights, and a variety of Bayesian prior distributions for the candidate models and parameters.

The frequentist MATA interval was observed to have the best coverage properties in the normal linear setting. In addition, constructing model weights using Akaike's information criterion (AIC) appeared to benefit the performance of frequentist model-averaged intervals. A different result was observed in non-normal settings, where the Bayesian approach more fully accounted for model uncertainty. Bayesian model averaging produced wider intervals with superior coverage rates, relative to any frequentist approach. Furthermore, the use of a data-dependent prior probability mass function for the set of candidate models resulted in Bayesian intervals with coverage rates nearest to the nominal value.

Acknowledgements

First and foremost, I would like to thank my supervisor, David Fletcher. His patience, encouragement, guidance, and friendship have been continuous since day one. I cannot imagine a better supervisor to learn from, and to work with. Thank you, David, for making my entire PhD experience both educational and enjoyable.

I must also thank Otago's Department of Mathematics and Statistics, and the Head of Department, Richard Barker. Richard has been completely supportive of my program, not to mention kind and encouraging. The department has treated me more as a member of the faculty than as a student, providing me the opportunity to lecture, and to participate in departmental matters. I am grateful to the entire department for this overwhelmingly positive experience.

My officemates Tim, Ella, Peter, and Jimmy have also made my PhD experience as fulfilling as it has been. We've had great camaraderie through lively office banter, tutoring stats papers, lunchtime games and quizzes, dinners and drinks, rugby, twilight cricket, golf, and finding that we like the people with whom we work. It's been a fun three years, and I'm genuinely sad that our days in the office together are nearly over.

Finally, I'll reserve a very special thank you for my parents, and for Jenn. All of you have been incredibly supportive and encouraging throughout my PhD program, and I'm fortunate to have such wonderful family behind me. I am grateful in a fundamental way to each of you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the process of statistical inference, a researcher may typically be presented with a set of data, and have the task of determining the most appropriate model for explaining the observed values. Accurate specification of this generating model will allow for increased understanding of the data presented, as well as aid in the prediction of future events. These inferences depend upon selection of an appropriate model for the underlying data-generating process.

The accurate identification and specification of such a model is not a trivial task. At the highest level, selection of the model distribution may be intuitive. The Poisson model is a common choice for modeling count data, and the normal linear model may be used to represent a wide range of naturally occurring continuous measurements. However, further specification of the details of any model allow for a wide range of possibilities. In the regression context, the decision of which covariates to include in the model is a common question. A researcher may have as many covariates available as data points, in which case incorporating all covariates into the model is not realistic, since this would over-fit the available data. Additional levels of model complexity can

quickly arise by considering temporal patterns, individual heterogeneity, the use of mixture models, or a wide variety of other modeling techniques.

These considerations allude to the principle of parsimony: the trade-off between a model using fewer parameters, to provide inference with low variance but potentially high bias, versus a highly parameterized model, giving rise to reduced bias but a larger variance. Statisticians seek a parsimonious middle ground, providing a useful compromise between low bias estimates, also with sufficiently low variance. Achieving this middle ground by identifying which covariates to include, and more generally which model to use, is not an easy task. This process is known as model selection.

We first consider a variety of approaches towards performing model selection, and discuss the consequences of this selection process. When the primary objective is prediction of future observations or parameter estimation, the alternative approach of model averaging may be used. This technique averages the predictions or parameter estimates across the models under consideration, and in doing so, hedges against the potential consequences of model selection.

1.1 Model Selection

There exist a wide variety of methods for performing model selection. We first briefly describe four of the methods which are most commonly used in practice. Next, the information-theoretic concept of Kullback-Leibler distance between models is introduced. This allows us to provide additional detail about the information criteria method for performing model selection.

1.1.1 Methods

Hypothesis Testing

The selection between nested models can be posed in the hypothesis testing framework, in which models are compared through use of a null hypothesis. One approach using this hypothesis testing framework is a step-forward selection process. Covariates are individually tested against a null value, and those which are deemed significant (at a predetermined critical level) are added sequentially to the model. When no further covariates are deemed significantly different from their null values, the final model has been obtained. This method, although easily performed in statistical software packages, is problematic when many null hypothesis tests are necessary, or when the candidate models are not nested (Westfall and Young, 1993).

Cross-Validation

Model selection may be performed through a cross-validation procedure (Stone, 1974; Geisser, 1975). This involves dividing the data set into two partitions. The first partition is known as the training data, and generally consists of the majority of the data set. Each model under consideration is independently fit to the training data. The second partition, known as the testing data, consists of the remainder of the data set. This testing data is used to assess the performance of each fitted model, through some criterion such as mean squared error. This process of data partitioning, model fitting, and model testing is repeated many times, and the final model is selected as that which performs best on the testing data set most often (Shao, 1993; Ronchetti *et al.*, 1997). The most common implementation of this is "leave one out" cross-validation, in which the testing data set consists of only a single observation. This technique has been widely studied, and is frequently used for performing model selection.

Bootstrapping

A bootstrapping procedure (Efron, 1979; Mooney and Duval, 1993) may also be used for model selection. Bootstrapping involves generating multiple independent resamples of the observed data by sampling with replacement, and performing analysis on each new sample. This generates samples from the empirical distribution of the observed data, and inferences on each independent sample are representative of those based on the true, underlying distribution function. Shao (1996) describes the method by which bootstrapping may be used to perform model selection. On each bootstrap resample, each candidate model is assessed in terms of a preselected loss function, for example mean-squared error. The final model selected is determined as that which performed best, on average, across all bootstrap resamples. Bootstrapping is computationally intensive relative to other methods, but is an accepted method for performing model selection (Burnham and Anderson, 2002, p.90-94).

Information Criteria

Perhaps the most widely practiced method for performing frequentist model selection is through the use of an information criterion. This method provides an approach to model selection derived from the information-theoretic quantity of Kullback-Leibler distance. Under this technique, a numerical value is calculated for each model. This value measures the relative Kullback-Leibler distance from that model to the true (unknown) generating distribution, relative to the other models under consideration. Information criteria values incorporate a measure of the complexity of candidate models, and hence attain a balance between bias and variance, in accordance with the principle of parsimony. The model with the most desirable value of the information criterion is deemed as the "best model," and this model is selected from those under consideration. A variety of information criteria exist and are used in practice, which

vary in their definitions, underlying assumptions, and asymptotic properties.

1.1.2 Kullback-Leibler Distance

An important component of frequentist model selection theory is the Kullback-Leibler (KL) distance (Kullback and Leibler, 1951). The KL distance is an analytical measure of the information *lost* when a particular probability distribution (model) is used to approximate the true, generating distribution. This may also be thought of as the *distance* between the true and the generating distributions (Kullback, 1959), or a measure of the *inefficiency* when a particular distribution is used to approximate truth (Cover and Thomas, 1991).

Assume the function $f(x)$ is the true probability distribution function for a quantity of interest. Generally, we wish to approximate this distribution by some other distribution function, $g(x)$. These functions may take any form, and may each depend on any number of parameters which are suppressed in this notation. A measure of the information *lost* through this approximation is given by the KL distance from g to f, given by

$$\mathrm{KL}(f, g) = \int_X f(x) \log\left(\frac{f(x)}{g(x)}\right) dx,$$

where X is the domain of $g(x)$. The KL distance is equal to zero if and only if the approximating distribution is identically equal to the true distribution, and otherwise $\mathrm{KL}(f, g)$ is strictly positive. However, the KL distance is not a true "distance" function, in that it does *not* have the property of symmetry between f and g. That is, in general $\mathrm{KL}(f, g) \neq \mathrm{KL}(g, f)$, as noted by Burnham and Anderson (2002, p.50-54).

It is important to note that the KL distance between two models has a firm, defensible foundation in information theory. Knowledge of the technical derivation of KL distance is not necessary for understanding its role in statistical model selection, and is

herein omitted. Thorough discussions of KL distance, its derivation, and relationship to the broader field of information theory may be found in either Cover and Thomas (1991) or Kullback (1959). It is worth noting that the calculation of KL distance is closely related to Shannon's theory of information (Shannon, 1948; Shannon and Weaver, 1949), and also to Boltzmann's entropy (Boltzmann, 1896). KL distance provides grounds for an information-theoretic approach to the process of model selection.

1.1.3 Information Criteria

A variety of information criteria exist, varying in their mathematical formulation and their motivation. Below are descriptions of some of the most commonly used information criteria. We consider a single model, parameterized in terms of the vector θ of length p. We also assume an observed data set y, consisting of n observations. The value $\hat{L}$ is the maximized value of the likelihood function $L(\theta|y)$ of the model under consideration; that is, $\hat{L} = L(\hat{\theta}|y)$, where $\hat{\theta}$ is the maximum likelihood estimate of the parameter vector θ.

Akaike's Information Criterion

Perhaps the most popular and widely utilized information criterion is Akaike's information criterion (AIC; Akaike, 1974, 1981, 1987). Akaike was the first person to relate the problem of model selection to the theoretical KL distance between models. He showed that the maximized log-likelihood function of a model is a biased estimator of the relative KL distance to the generating model. Further, the bias term may be approximated by the number of estimated parameters in the given model. Interestingly, to calculate this estimator of relative KL distance to truth, knowledge of truth itself is not necessary. The AIC value of a model is given as $\text{AIC} = -2\log\hat{L} + 2p$. The

equation for AIC expressly shows the trade-off between the maximized log-likelihood of a model fit, and the number of parameters used by the model to achieve this fit. Note that *lower* values of AIC correspond to a more *favorable* model. The definitions of most other information criteria are similar in form to that of AIC.

Since AIC is an (approximately) unbiased estimator of the *relative* KL distance to truth, only differences in AIC values are of interest. The AIC value of one particular model alone tells us nothing; only when compared relative to AIC values from other models may we infer which model is nearer to truth. AIC also has the desirable property of *efficiency*. This means that the model selected by AIC will asymptotically perform as well as the theoretically best model in the candidate set, measured in terms of mean squared error. A more precise description of *efficiency* appears in Claeskens and Hjort (2008, p.108-112).

Corrected AIC

AIC is known to favor complex models, sometimes selecting models which are overly complex relative to the data available (Sugiura, 1978). A small-sample correction to AIC was formalized by Hurvich and Tsai (1989), which induces a preference towards simpler models when the available data set is small, and hence not necessarily able to support a more complex model. This "corrected" version of AIC is known as $\mathrm{AIC_c}$, and the value for a particular model is given by $\mathrm{AIC_c} = -2\log\hat{L} + 2p(\frac{n}{n-p-1})$. $\mathrm{AIC_c}$ introduces an increased penalty for more complex models when n is small, and asymptotically converges to AIC. Some authors, for example Burnham and Anderson (2002), advocate the use of $\mathrm{AIC_c}$ rather than AIC whenever model selection is performed.

Bayesian Information Criterion

The Bayesian information criterion (BIC) was developed by Schwarz (1978), and in some texts is referred to as the Schwarz criterion. BIC employs a penalty term for model complexity which is dependent upon the sample size, and value of the criterion is given as $\text{BIC} = -2 \log \hat{L} + p \log n$. This asymptotic relationship with n affords BIC the desirable property of *consistency*. This means that BIC will asymptotically select the model with the minimum KL distance to the true model (from among those models under consideration), and if several candidate models achieve this minimum distance, BIC will select that with the fewest parameters. AIC, in contrast, only satisfies the condition of *weak consistency*. In the event where several candidate models achieve the minimal KL distance to truth, asymptotically, AIC may select any of these models (Claeskens and Hjort, 2008, p.99-102). Interestingly, it is impossible for an information criterion to be fully consistent *and* efficient, as was proved by Yang (2005). Another useful property of BIC relates to Bayesian analyses, and is a direct consequence of the formulation of this criterion: BIC values may be used to asymptotically approximate Bayesian posterior model probabilities, when equal model prior probabilities are used. A useful discussion of this property appears in Link and Barker (2006).

Deviance Information Criterion

The deviance information criterion (DIC; Spiegelhalter *et al.*, 2002) is a variation of BIC, which is used in Bayesian analyses. DIC differs in its penalty term, which estimates the *effective* number of parameters, p_D, in a Bayesian analysis. This quantity can be thought of as the number of unconstrained parameters in the model, where fractional values represent the degree to which a parameter is unconstrained. Each parameter contributes a quantity between zero and one to the value of p_D. The maximum possible value of one is contributed to p_D when a parameter estimate is generated solely

from the data, and is completely unconstrained by a prior distribution. The minimum possible value of zero is contributed to p_D when a parameter is fully constrained by the model or its prior distribution, and the observed data has no effect whatsoever on the parameter estimate. Otherwise, a value between zero and one is added to p_D, when a parameter estimate is affected by both a prior distribution and the data (Gelman *et al.*, 2004, p.182). The value of this information criterion is given by $\text{DIC} = -2\log\hat{L}_D + 2p_D$, where $\hat{L}_D$ is the value of the likelihood function evaluated at the mean of the parameter posterior distribution. Details regarding the calculation of DIC may be found in Claeskens and Hjort (2008, p.90-94).

Focused Information Criterion

Generally, information criteria aim to select a single model from the candidate set, which is deemed the "best" overall choice for all subsequent inferences. Following a different approach, the focused information criterion (FIC) aims to select the best model for estimation of a parameter or quantity of interest (Claeskens and Hjort, 2003; Hjort and Claeskens, 2006). This may be a mean, variance, regression coefficient, or more generally any function of the model parameters. Mathematically, FIC will select the model which gives rise to the minimum mean squared error, when estimating the parameter of interest. A rigorous formulation of FIC appears in Claeskens and Hjort (2008, p.145-154). When the candidate set contains only two models, which are nested and differ only by a single parameter of interest, then AIC and FIC are asymptotically equivalent. This is not true, however, in more complex situations.

Quasi-AIC

The phenomenon of overdispersion occurs when the observed variance of a data set exceeds the theoretical variance specified by a generalized linear model. Many statistical models for count data have a theoretical variance which is directly related to the mean. Under the Poisson model, for example, the variance is identically equal to the mean. In the presence of overdispersion, the estimation of model parameters remains unbiased, however the resulting estimate of variance will be overly optimistic. The quasi-likelihood function (Wedderburn, 1974) can be used for analyzing overdispersed data sets. The variance inflation factor ϕ is estimated, which indicates the degree of overdispersion of the observed data set relative to the model-theoretical variance. See Cox and Snell (1989) for the details of calculating $\hat{\phi}$. The quasi-AIC information criterion (QAIC) is a modified version of AIC for use on overdispersed data, and is given as $\text{QAIC} = -2(\log \hat{L})/\hat{\phi} + 2p$ (Lebreton *et al.*, 1992; Anderson *et al.*, 1994). Note that the number of model parameters, p, must be incremented by one, to account for the estimation of ϕ.

Corrected QAIC

A small-sample correction to QAIC may be readily defined. This is analogous to AIC_c, the small-sample corrected form of AIC. This corrected information criterion would be used for small samples of overdispersed count data. The value of this criterion is defined as $\text{QAIC}_\text{c} = -2(\log \hat{L})/\hat{\phi} + 2p(\frac{n}{n-p-1})$, as described in Burnham and Anderson (2002, p.69-70). Similar to AIC_c, Burnham and Anderson (2002) recommend using QAIC_c whenever overdispersion is present. As one would expect, QAIC_c asymptotically converges to QAIC as the sample size increases.

Takeuchi's Information Criterion

The calculation of AIC includes an approximation which simplifies its final formulation. Specifically, the bias correction in the estimated relative KL distance to the true model f is approximated as the dimension of the parameter vector of the approximating model g. This approximation is nearly exact in many situations, allowing for easy calculation of AIC values. Following the initial publication of AIC, Takeuchi (1976) derived the exact bias correction, producing an information criterion which is an asymptotically unbiased estimator of the KL distance to the true model (Burnham and Anderson, 2002, p.65-66).

Takeuchi's information criterion (TIC) is given by TIC $= -2\log\hat{L} + 2p^*$, where $p^* = \text{tr}(J^{-1}K)$. This calculation involves $p \times p$ matrices J and K, which are calculated using the first and second-order derivatives of the log-likelihood function. See Claeskens and Hjort (2008, p.23-27,43-44) for a thorough description of calculating p^*. In practice, estimation of J and K from the observed data, and also numerical inversion of J are unstable operations, making TIC difficult to calculate. For most practical applications, AIC is sufficiently accurate, and far simpler to calculate than TIC.

1.2 Model Selection Uncertainty

When analyzing a set of data, a researcher may be provided with a set of candidate models, or generate this set through careful consideration of the data and the natural processes thought to govern it. We typically consider this candidate set to be a fixed set of R models, $\{M_1, \ldots, M_R\}$. Through the process of model selection (perhaps using an information criterion), a particular model is selected from this candidate set, say M_k, where $1 \leq k \leq R$. We discard the remaining $R-1$ models, and perform the

subsequent inference using the selected model, M_k. The model selected by this process is referred to as the "best model."

This "best model" technique is known to have several inherent problems. The data set is used twice in the analysis: first to perform the selection of model M_k, and a second time in any inferences which follow. Inferences may include estimation of model parameters, predictions of future observations, assessing goodness-of-fit, and the construction of confidence intervals. The data set is analyzed as though it were a fresh set of observations generated from M_k, whereas in fact, M_k was selected as a result of the particular data set. The distribution of the data set *conditional* upon selection of M_k may vary substantially from the unconditional distribution of the data set, if it were actually generated from M_k. This implicit conditioning is known to produce biases in the resulting inferences (Ye, 1998; Miller, 2002).

In addition, the "best model" technique involves discarding the remaining $R - 1$ models following selection of M_k. At this point, the previous uncertainty as to which model is correct is also discarded. Inference does not reflect the uncertainty in the model selection process, which results in under-estimates of standard errors, and thereby produces overly narrow confidence intervals (Burnham and Anderson, 2002, p.46). Model selection also adversely affects the estimation of error variance subsequent to model selection.

These phenomena are collectively known as model selection uncertainty, and were famously referred to by Breiman (1992) as the "quiet scandal" of statistics, wherein following model selection a researcher assumes the selected model was given *a priori*. Chatfield (1995) provides a thorough background and historical perspective on the issue of model selection uncertainty, noting that the additional uncertainty resulting from model selection may even be the dominant source of uncertainty in regression analyses. Summarizing the issue, Chatfield (1995, p.422) says:

> "...when a model is formulated and fitted to the same data, inferences made from it will be biased and overoptimistic when they ignore the data analytic actions which preceded the inference. Statisticians must stop pretending that model uncertainty does not exist and begin to find ways of coping with it."

Model selection uncertainty may be thought of as having two components: bias introduced into post-model selection estimators, and additional variability in these estimators which results from uncertainty in the model selection process.

1.2.1 Bias

Estimating model parameters subsequent to model selection is often known to introduce bias into the parameter estimates. This can be understood through a thought exercise. Say we perform model selection in the context of variable selection for a regression analysis, and suppose that a particular covariate, x_1, is strongly related to the dependent variable, y. The data observed is very likely to represent this trend, and the selected model M_k will thus include x_1. Further, the associated regression coefficient estimate, $\hat{\beta}_1$, will approximately represent a random observation from the sampling distribution of this estimator, since x_1 is likely to be included in M_k on repeated sampling of the observed data set. Thus, inferences about β_1 will adhere to theory derived from the sampling distribution of $\hat{\beta}_1$ (Burnham and Anderson, 2002, p.43-45).

In contrast, suppose a second covariate x_2 is, in truth, weakly correlated with y. If, through some degree of chance, x_2 is determined to be "significant" in the observed data set, then x_2 will be included in M_k. In this case, the estimated regression coefficient $\hat{\beta}_2$ is likely to be biased away from zero, since x_2 is not likely to have been included in the selected model *unless* the value $\hat{\beta}_2$ comes from an extremity of its sampling

distribution. Consequently, the estimated regression coefficients for covariates weakly correlated with the response variable are likely to be biased away from zero, since otherwise the model selection procedure would not have included these covariates.

Conversely, say a third covariate x_3 is also weakly correlated with y, and x_3 is *not* included in M_k. In this situation, the associated regression coefficient $\hat{\beta}_3$ will be biased towards zero – in fact, the estimated regression coefficient $\hat{\beta}_3$ under M_k is identically zero, since x_3 was deemed "insignificant" in the model selection process. Miller (1984) provides a detailed description of these biases arising in estimates of regression coefficients.

In addition, any number of covariates which in fact are completely uncorrelated with y may have been incorrectly included by model selection. When this occurs, these covariates will necessarily have exaggerated estimates of their significance, or estimated regression coefficients biased away from zero. This is known as "Freedman's paradox," where covariates with no relation to the response are included through model selection (Freedman, 1983; Freedman and Pee, 1989; Lukacs *et al.*, 2010). Given a single data set, a researcher has no means to distinguish these spurious effects from variables which genuinely have a weak correlation to the response.

To summarize, covariates strongly correlated with the response variable are likely to be selected, and the corresponding regression coefficient estimates will be approximately unbiased samples from their sampling distributions. However, covariates weakly correlated with the response variable are likely to be excluded entirely, or when included, to have regression coefficient estimates biased away from zero, serving to exaggerate their true effect. Similarly, genuinely insignificant predictor variables may be included in the final model, with biased estimates of their effect. As a result of conditioning on the validity of model M_k, inferences produced using M_k will generally be biased. This bias propagates into all derived estimators, including estimates of the

mean, and future forecasts. This phenomenon is known as model selection bias.

1.2.2 Variance

The process of model selection introduces a problem for estimating standard error terms, as well. Model selection conditions on the selected model (say M_k) for calculating the variance of an estimator, which does not estimate the true, unconditional variance. Assume the quantity of interest θ has a common interpretation under each model, and the estimate of θ under M_k is $\hat{\theta}_k$. Under model selection the estimator is taken as $\hat{\theta} = \hat{\theta}_k$, and the associated variance is estimated using $\hat{\text{var}}(\hat{\theta}) = \hat{\text{var}}(\hat{\theta}_k)$, both of which have conditioned on M_k being true.

The law of total variance (Weiss *et al.*, 2005, p.385) provides an expression for the true, *unconditional* variance of $\hat{\theta}$. Letting K represent a random variable specifying the index of the selected model, the unconditional variance is given as

$$\text{var}(\hat{\theta}_K) \;=\; \text{E}\left[\text{var}(\hat{\theta}_K|M_K)\right] \;+\; \text{var}\left(\text{E}[\hat{\theta}_K|M_K]\right).$$

When variances are calculated conditional on M_k (using model selection), the expectation over K has been omitted from the first term of the unconditional variance, and the second term is entirely absent. Since this second term provides a non-negative contribution to the total variance, model selection results in variance estimates which are generally smaller than the true, unconditional variance; this is a direct consequence of disregarding model selection uncertainty. Thereby, confidence intervals for $\hat{\theta}$ are generally too narrow, and are likely to produce coverage rates below the nominal level (Burnham and Anderson, 2002, p.45-47).

The inaccuracy in post-model selection (conditional) variances has been quantified

in many studies. Hurvich and Tsai (1990) used a simulation study to show that confidence intervals for linear regression parameters, conditional upon a selected model, were far below the nominal levels. Following model selection using either AIC or BIC, confidence intervals designed for 95% coverage were in fact found to achieve a coverage rate as low as 79%.

Through analytical study, Kabaila (1995) considered prediction intervals in the setting of linear regression. For finite sample sizes and appropriate choices of regression coefficient values, Kabaila (1995) observed that prediction intervals following model selection may attain arbitrarily low coverage probabilities. In essence, depending on the true nature of the underlying generating process, the performance of prediction intervals subsequent to model selection may be arbitrarily poor.

The under-estimation of post-model selection variances also affects estimates of error variance. This is known to produce inflated values of R^2, a measure for model goodness-of-fit. Using a step-forward variable selection procedure, Rencher and Pun (1980) observed that R^2 values under the selected model were significantly inflated, relative to the true, theoretical values – oftentimes by a factor of two or more. This implies that careless use of model selection procedures may result in severely over-estimating the goodness-of-fit of a model.

Many such studies exist, and universally find that ignoring the effects of model selection uncertainty (introducing biases and incorrect estimation of variance) can result in nontrivial adverse effects on inferences following model selection. The issues and consequences of model selection are well recognized in statistical literature. The difficulty lies in accurately incorporating the effects of model selection into the process of inference; resolution will require stepping back from model selection entirely.

1.3 Multimodel Inference

In order to resolve these issues resulting from model selection, we must reconsider our approach towards inference. Instead of performing inference using a single model subsequent to model selection, we instead begin formal inference at the stage of having an observed data set, and set of candidate models $\{M_1, \ldots, M_R\}$ which has been specified independently of the observed data. Inference will proceed beginning with the entire set of R candidate models, which is known as multimodel inference (Burnham and Anderson, 2002, p.149-167).

1.3.1 Model Weights

The key idea behind multimodel inference is that we do not fully accept a single model, then reject all other models, as was done using model selection. Instead, we will acknowledge our uncertainty regarding which model is truth, and accept all candidate models *to some degree*. We quantify our degree of belief, or the relative strength of the evidence in support of each model, through use of numerical model weights. This approach implicitly requires the assumption that *truth is in the model set*, which we are willing to accept.

For each of the R candidate models, we assign a weight w_i to model M_i, for all $i = 1, \ldots, R$. The values of these weights will follow certain conventions, which aid in their interpretation. We restrict $w_i \in [0, 1]$ for all i, and also impose the constraint that $\sum_{i=1}^{R} w_i = 1$. Under these restrictions, model weights may be thought of as "probabilities" associated with each model, although we must exercise caution in our exact interpretation. If $w_i < w_j$, then in some sense model M_j is more likely, or more plausible than the competing model M_i.

An expression for model weights may be written using the AIC values for each candidate model. The motivation behind this form is to express a likelihood, considering the data and the set of candidate models, for the plausibility of each *model*, given the data. Note the contrast to the usual likelihood function under each individual model, which represents the plausibility of model parameter values, given the data (Burnham and Anderson, 2002, p.74-77). Each candidate model is fit to the data, producing the maximized value of each model's likelihood function, which are used to calculate the AIC value AIC_i for each model M_i. The weight w_i associated with model M_i is calculated as

$$w_i \;=\; \frac{\exp(-\frac{1}{2}\text{AIC}_i)}{\sum_{j=1}^{R}\exp(-\frac{1}{2}\text{AIC}_j)}. \tag{1.1}$$

The model weights defined in equation (1.1) are also sometimes called *Akaike* weights, since they are calculated using Akaike's information criterion. The numerator of this expression was suggested by Akaike (1983), and again by Bozdogan (1987), for providing the relative likelihood of a candidate model. The denominator of equation (1.1) provides a normalizing constant, which ensures that $\sum_{i=1}^{R} w_i = 1$. Buckland *et al.* (1997) formalized the definition of model weights, by including the normalization constant.

Recall that AIC values provide only a *relative* measure for each model, as described in Section 1.1.3. Any numerical constant may be added or subtracted from the AIC values of all candidate models, to no effect. Bearing this in mind, for numerical stability when evaluating the exponential function, the minimum AIC value among all candidate models is typically subtracted away from all AIC values. This gives rise to a ΔAIC value for each model, denoted as ΔAIC_i for model M_i, where $\Delta\text{AIC}_i \equiv \text{AIC}_i - \min_j(\text{AIC}_j)$. These ΔAIC values may be used in place of AIC values for the calculation of model weights, giving rise to identical values of w_i. This computationally stable construction of w_i is given on the final line of equation (1.2), where $\text{AIC}_{\min} \equiv \min_j(\text{AIC}_j)$, and

is generally taken as the formal definition of Akaike model weights (Burnham and Anderson, 2002, p.75).

$$\begin{aligned} w_i &= \frac{\exp(-\frac{1}{2}\mathrm{AIC}_i)}{\sum_{j=1}^{R}\exp(-\frac{1}{2}\mathrm{AIC}_j)} \\ &= \frac{\exp(-\frac{1}{2}(\Delta\mathrm{AIC}_i + \mathrm{AIC}_{\min}))}{\sum_{j=1}^{R}\exp(-\frac{1}{2}(\Delta\mathrm{AIC}_j + \mathrm{AIC}_{\min}))} \\ &= \frac{\exp(-\frac{1}{2}\Delta\mathrm{AIC}_i)}{\sum_{j=1}^{R}\exp(-\frac{1}{2}\Delta\mathrm{AIC}_j)} \end{aligned} \tag{1.2}$$

One must use caution when interpreting model weights. The value of a particular weight w_i is only meaningful relative to the R candidate models under consideration. We might say the weight w_i represents the evidence that M_i is the most parsimonious model among the candidate set, or that M_i has the minimum KL distance to the true, generating distribution. One must not make any claims that w_i is "the probability that M_i is truth," since this implicitly relies on our assumption that "truth is in the model set." Since we have no concrete reason to believe the true model is contained in our set of candidate models, interpretations should extend only to statements regarding KL distances to truth, and only relative to the candidate models under consideration.

The construction of model weights given in equation (1.2) may be applied using other information criteria. The model weights constructed using AIC values will thereby be called AIC weights, and those constructed using BIC, for example, will be called BIC weights. These alternate constructions of model weights have been considered in the statistical literature. Among others, Claeskens and Hjort (2008) study the use of FIC weights, while Buckland *et al.* (1997) considers both BIC weights and $\mathrm{AIC_c}$ weights. Model weights may be constructed using any information criterion, depending on the assumptions and approach of a researcher.

1.3.2 Model Averaging

The process of multimodel inference uses a smoothing, or averaging over all models in the candidate set, to mitigate the biases which arise from model selection, and correctly account for the components of variance. This is known as model averaging. The use of model averaging has recently gained popularity in frequentist statistics (see, for example, Buckland *et al.*, 1997; Claeskens and Hjort, 2003; Holländer *et al.*, 2006; Lee and Hugall, 2006; Fletcher and Dillingham, 2011; Nakagawa and Freckleton, 2011). Interestingly, however, the concept of model averaging arises more naturally in the Bayesian paradigm, which likewise has a long history of study (e.g., Raftery *et al.*, 1997; Hoeting *et al.*, 1999; Wasserman, 2000; Wintle *et al.*, 2003).

Frequentist model averaging makes use of the model weights w_i to achieve this smoothing over the discrete set of candidate models. Assume that we have a set of R candidate models $\{M_1, \ldots, M_R\}$, and we are interested in prediction of a quantity θ. Denote the predicted value of θ under model M_i as $\hat{\theta}_i$, for each i, and the set of AIC weights as $\{w_i\}$ for our candidate models. The model-averaged (MA) estimator $\hat{\theta}$ of the quantity θ is given as the weighted average of the estimates produced under each model (Burnham and Anderson, 2002, p.150):

$$\hat{\theta} \;=\; \sum_{i=1}^{R} w_i \, \hat{\theta}_i. \tag{1.3}$$

This construction of the MA estimate in equation (1.3) makes intuitive use of the AIC weights. If a particular model M_k achieves the majority of the weighting (say $w_k = 0.9$), then the MA estimate will be heavily influenced towards $\hat{\theta}_k$, the estimate made under model M_k. Likewise, if two contending models equally share the majority of the model weight (say $w_j \approx w_k \approx 0.5$), then the MA estimate $\hat{\theta}$ will approximately equal the mean of the estimates from each of these two models. And as we would expect,

if a particular candidate model M_k has an AIC value substantially larger (worse) than the contending models, then the associated model weight $w_k \approx 0$, and the estimate $\hat{\theta}_k$ from model M_k will have virtually no impact on $\hat{\theta}$.

The application of (1.3) can be extended to any quantity that has a common interpretation across all candidate models. We may consider the expectation of the probability distribution associated with each model, $\theta_i \equiv \mathrm{E}[Y_i]$, where Y_i is a random observation generated from model M_i. In this case, the MA expectation $\hat{\theta}$ is the weighted average of each single-model expectation, $\hat{\theta}_i$. Note that each model may be parameterized differently, utilizing different structural parameters, and generally containing different numbers of parameters; for example, considering the exponential and the Weibull distributions as the two candidate models. Alternatively, in linear regression each candidate model might incorporate a different subset of the explanatory variables.

When all candidate models have some parameters or derived quantities in common, we may use equation (1.3) to find a MA estimate. For example in an analysis of covariance setting, we may wish to estimate the expected value of the response variable corresponding to a particular treatment group, but be uncertain about which covariates to adjust for. Using model averaging, a researcher may produce a MA estimator for the treatment group of interest, without taking a firm stance on the subset of covariates to include in the regression. Likewise, again considering the exponential and Weibull distributions, model averaging may be used to produce a MA estimate for the scale parameter common to these two distributions. This may be useful when the observed data provides comparable support for both models.

The situation is slightly more complicated when model averaging over regression coefficients, in which some models may include a particular covariate, while others do not. Say we are considering the regression coefficient β_p for an explanatory variable

x_p. Without loss of generality, say models $M_1, \ldots, M_n$ do not regress on x_p, while models $M_{n+1}, \ldots, M_R$ include x_p, and hence estimate the associated coefficient β_p. For $i = n+1, \ldots, R$, let us denote $\hat{\beta}_{p,i}$ as the estimate of β_p under model M_i.

One approach is to find a MA estimate of β_p over *those models which include* x_p. We first calculate new model weights w_i' for those models which include x_p, which are proportional to the original model weights, but also satisfy $\sum_{j=n+1}^{R} w_j' = 1$. The new weights w_i' are scalings of the original model weights. Then, we model-average the estimates of β_p across all models which include it, using these new weights. The corresponding formulae are given in equation (1.4).

$$\begin{aligned} w_i' &= \frac{w_i}{\sum_{j=n+1}^{R} w_j}, \quad (i = n+1, \ldots, R) \\ \hat{\beta}_p &= \sum_{i=n+1}^{R} w_i' \, \hat{\beta}_{p,i} \end{aligned} \tag{1.4}$$

Using this approach, $\hat{\beta}_p$ represents the MA coefficient of x_p, *given* x_p *is included in the model*. Provided one assumes that x_p should be included as an explanatory variable, then the MA estimate of its associated coefficient is given by equation (1.4). One must be cautious in using this approach, since models which include x_p may exhibit model selection bias, and hence produce a biased MA estimate of β_p (see Section 1.2.1).

An alternate approach to calculating MA regression coefficients is to model-average over *all models*: those which include x_p, and those which do not. Under models which do not include x_p, and hence do not include the parameter β_p, the estimate of this coefficient is identically zero. That is, for each model M_i where $i = 1, \ldots, n$, we use $\hat{\beta}_{p,i} = 0$ for calculation of the MA estimator $\hat{\beta}_p$. Using $\{w_i\}$, the original set of model weights,

$$\hat{\beta}_p = \sum_{i=1}^{R} w_i \, \hat{\beta}_{p,i}. \tag{1.5}$$

As this averages across all models, the MA estimator $\hat{\beta}_p$ given in (1.5) is not conditional upon x_p being included in the final model. This estimate takes into account the uncertainty as to whether x_p is a significant predictor of the dependent variable. Accordingly, some degree of the model selection bias inherent to the non-zero estimates of β_p is averaged out, since the summation in (1.5) includes models where $\hat{\beta}_{p,i} = 0$. By including these models, this MA estimator incorporates model uncertainty, and may produce a more reliable *unconditional* estimator for the effect of x_p (Burnham and Anderson, 2002, p.150-153). In addition, this formulation for MA regression coefficients is consistent with our definition for MA estimation of the mean, using equation (1.3) when $\theta_i \equiv \mathrm{E}[Y_i]$, and Y_i is a random observation from model M_i.

1.3.3 Model-Averaged Confidence Intervals

Model averaging provides a logical and effective way of calculating MA expectations, parameter estimates, or generally model averaging any quantity which has a common interpretation across all candidate models. This method corrects for some degree of model selection bias, and thus incorporates uncertainty regarding which model is correct into the resulting estimators. The question remains of how to accurately account for model uncertainty in the *uncertainty* of MA estimators. An accurate representation for the uncertainty (or variance) of these MA estimators would allow us to bound these quantities, through the use of model-averaged confidence intervals.

Buckland *et al.* (1997) take an analytical approach towards estimating the variance of model-averaged estimators; that is, estimation of $\mathrm{var}(\hat{\theta})$, with $\hat{\theta}$ defined as in (1.3). Their approach accounts for model selection bias, since the estimator $\hat{\theta}_i$ under model M_i will generally be a biased estimate of θ. Further, to calculate a worst-case variance, it is assumed that estimators under different models are perfectly correlated, or that $\mathrm{corr}(\hat{\theta}_i, \hat{\theta}_j) = 1$ for all i and j. Following this approach, Buckland *et al.* (1997) arrive

at the first estimator for the unconditional variance of a MA estimate,

$$\hat{\text{var}}_1(\hat{\theta}) \;=\; \left[\sum_{i=1}^{R} w_i \sqrt{\hat{\text{var}}\left(\hat{\theta}_i\right) + \left(\hat{\theta}_i - \hat{\theta}\right)^2} \right]^2, \tag{1.6}$$

where each $\hat{\theta}_i$ is the estimate of θ under model M_i, $\hat{\text{var}}(\hat{\theta}_i)$ is the estimate of the conditional variance of $\hat{\theta}_i$ given M_i is true, $\hat{\theta}$ is the model-averaged estimate of θ defined in (1.3), and w_i are the model weights defined in (1.2).

If one is willing to make the assumption of asymptotic normality for the sampling distribution of $\hat{\theta}$, then equation (1.6) may be used to construct confidence intervals for MA estimators. Since this interval uses the first unconditional variance estimator $\hat{\text{var}}_1(\hat{\theta})$, it will be referred to as the first model-averaged Wald (MAW_1) interval. The $(1-2\alpha)100\%$ MAW_1 confidence interval for the MA estimator $\hat{\theta}$ is defined as any Wald interval: $\hat{\theta} \pm z_\alpha \hat{\text{var}}_1(\hat{\theta})^{1/2}$, where z_α is the $(1-\alpha)$ quantile of the standard normal distribution (Brazzale *et al.*, 2007, p.5-7). If the MA estimator $\hat{\theta}$ followed a normal sampling distribution, and if we could specify an asymptotically correct estimator for $\text{var}(\hat{\theta})$, this would be a theoretically correct formulation for constructing MA confidence intervals.

However, neither of these conditions are generally true. The sampling distribution of $\hat{\theta}$ can be dramatically non-normal (see Claeskens and Hjort, 2008, p.196-197, and Figure 7.1 on p.198), due to the random nature of the model weights. In addition, the first unconditional variance estimator $\hat{\text{var}}_1(\hat{\theta})$ given in (1.6) is an approximation. It requires several assumptions regarding the model weights, and the correlations between single-model parameter estimates. These assumptions are described in Buckland *et al.* (1997).

It is conceptually tempting to assume an asymptotically normal sampling distribution for $\hat{\theta}$, considering the property of large sample normality typical of maximum

likelihood estimators (see, for example, Davison, 2003, p.118-120). As $\hat{\theta}$ is generally constructed as a weighted average R single-model maximum likelihood estimators, and considering the fact that linear combinations of normal random variables are themselves normal (De Veaux *et al.*, 2008, p.396-400), one may be quick to attribute this same property to the model-averaged estimator, $\hat{\theta}$.

The problem with this approach is that the model weights are random quantities themselves, so that when we consider $\hat{\theta}$ as a linear combination of normally distributed random variables, the coefficients in this linear combination are, in fact, random variables. Consequently, $\hat{\theta}$ does not follow a normal sampling distribution. Buckland *et al.* (1997) side-steps this problem when deriving the unconditional variance estimator in (1.6), stating "Consider first the unrealistic case that ... the weights w_k are known constants." Their derivation of the unconditional variance proceeds following this assumption.

Burnham and Anderson (2002) build upon the results of Buckland *et al.* (1997), and propose alternate approaches toward estimating the unconditional variance of MA estimators. A heuristic argument relating to Bayesian model averaging in Burnham and Anderson (2002, p.345) gives rise to the second estimate for the unconditional variance:

$$\hat{\text{var}}_2(\hat{\theta}) \;=\; \sum_{i=1}^{R} w_i \left[\hat{\text{var}}\left(\hat{\theta}_i\right) + \left(\hat{\theta}_i - \hat{\theta}\right)^2 \right]. \tag{1.7}$$

It follows from Jensen's inequality that $\hat{\text{var}}_1(\hat{\theta}) \leq \hat{\text{var}}_2(\hat{\theta})$, and hence this second unconditional variance estimator will always produce a more conservative (larger positive) estimate for the variance of $\hat{\theta}$. A Wald confidence interval constructed using the second variance estimator, $\hat{\text{var}}_2(\hat{\theta})$, will be referred to as the second model-averaged Wald (MAW_2) interval. A MAW_2 interval is centered around the same MA point estimate, $\hat{\theta}$, and is at least as wide as a MAW_1 interval.

Furthermore, Burnham and Anderson (2002) suggest a modification to either estimate of unconditional variance, which takes into account the error degrees of freedom under each model. This modification is termed the "*adjusted* [variance] estimator" (Burnham and Anderson, 2002, p.164), which uses quantiles from the normal and t-distributions to scale the conditional variance estimates under each model. Despite the numerous suggested expressions for the unconditional variance (or standard error) of a model-averaged estimator, Burnham and Anderson (2002, p.345) state:

> "We have not studied this matter further; however, the issue of the unconditional variance and covariance for AIC [model-averaged estimators] are subject areas worthy of more research."

The study of MA confidence intervals continues in Hjort and Claeskens (2003), and later in Claeskens and Hjort (2008). These references consider the performance of the MAW_1 and MAW_2 confidence intervals, constructed using the unconditional variance estimators $\hat{\text{var}}_1(\hat{\theta})$ and $\hat{\text{var}}_2(\hat{\theta})$, and are found to result in coverage rates far below the nominal values (Hjort and Claeskens, 2003, p.885-886). They proceed to consider alternative approaches to MA confidence intervals, following a "local misspecification framework." This particular setting imposes specific assumptions about the nested structure of the candidate models, in which the structure of these models depends upon the true values of underlying model parameters. Under this particular framework, alternate approaches to MA confidence intervals are shown to result in improved coverage properties (Claeskens and Hjort, 2008, p.206-211). However, no further progress is made towards the general setting of model averaging, which does not impose any framework for the structure of the candidate model set.

1.4 Thesis Outline

Through this thesis, we extend the existing methodology for the construction of model-averaged confidence intervals. A new approach is proposed, referred to as the model-averaged tail area (MATA) construction, which is fundamentally different from the methods already described. The MATA approach does not attempt or require estimation of the unconditional variance of a model-averaged estimator, which was the direction of study taken by Burnham and Anderson (2002). Nor does it rely upon any specific framework for the structure of the candidate models, as was done by Claeskens and Hjort (2008).

In the subsequent chapters, we define several MATA constructions for model-averaged confidence intervals, and use simulation to study the performance of this new methodology in a variety of settings. Chapter 2 develops the MATA technique in the context of profile likelihood confidence intervals, motivated using an analogy to Bayesian model averaging. An analytical development of the MATA construction is presented in Chapter 3, in the context of Wald intervals for normal linear and non-normal models. Chapter 4 presents a comparison between these MATA constructions and Bayesian model averaging, and Chapter 5 provides a complete summary of the research and conclusions relevant to all chapters.

Chapters 2, 3, and 4 were written as stand-alone research articles. Each chapter provides a detailed study of a particular aspect of the MATA construction. Indeed, Chapters 2 and 3 represent articles published during the course of the postgraduate research program, and Chapter 4 has recently been submitted for publication. For this reason, there exists some amount of repetition between chapters, particularly in the introductory material.

Chapter 2

Model-Averaged Profile Likelihood Intervals [1]

2.1 Introduction

Parameter estimation has traditionally been based on a single model, possibly after a model selection process, and has therefore ignored model uncertainty (Chatfield, 1995; Draper, 1995). It is well documented that use of a single "best model" can lead to problems, in particular that the resulting confidence intervals will generally have error rates above the nominal level (Hurvich and Tsai, 1990; Lukacs *et al.*, 2010).

Model averaging has been proposed as a means of making some allowance for model uncertainty, from both a frequentist and a Bayesian perspective (Buckland *et al.*, 1997; Raftery *et al.*, 1997; Volinsky *et al.*, 1997; Hoeting *et al.*, 1999; Burnham and Anderson, 2002; Claeskens and Hjort, 2008). Our focus is on frequentist methods, but we make

[1] The content of this chapter has been published in the *Journal of Agricultural, Biological, and Environmental Statistics*; see Fletcher and Turek (2012).

use of ideas inherent in the Bayesian approach. In the frequentist setting, a model-averaged estimate of a parameter is a weighted mean of the estimates from each of the candidate models, the weights being chosen using an information criterion or the bootstrap (Buckland *et al.*, 1997).

Methods for calculating a confidence interval around a model-averaged estimate have been considered by Buckland *et al.* (1997), Burnham and Anderson (2002), Hjort and Claeskens (2003), and Claeskens and Hjort (2008). All of these methods provide Wald intervals, in that they assume approximate normality of the model-averaged estimate and require estimation of its variance. In many situations, we would not expect a Wald interval to perform well, regardless of whether we use a single model or model averaging; for example, when the data are skewed and the sample size is not large enough to assume normality of the single-model estimates.

As in the single-model setting, we might overcome this problem by transforming the parameter, the hope being that the sampling distribution of the corresponding transformed estimate is approximately normal. A confidence interval is then obtained by back-transforming the endpoints of the Wald interval for the transformed parameter. We refer to this as a transformed Wald interval. An example would be the use of a logit-transformation when the parameter is a probability, a transformation that naturally arises when we fit a logistic regression model. Indeed, it will often be the case that a transformed Wald interval will be a more natural choice than a Wald interval. However, even if the estimate of the transformed parameter is approximately normal for each model, the model-averaged estimate may not be, due to the weighting and the need to estimate the weights.

An alternative approach in the single-model setting is to use a profile likelihood interval. It is well known that a profile likelihood interval will generally provide a better coverage rate than a Wald interval, particularly for skewed data (Brown *et al.*, 2003;

Cox and Hinkley, 1974, p.343). The purpose of this chapter is to propose a method for calculating a model-averaged profile likelihood interval. In doing so, we make use of the fact that in many settings an appropriate choice of priors for the parameters leads to a Bayesian credible interval that can be regarded as an approximation to a profile likelihood interval (Severini, 1991). This connection allows us to make use of results for Bayesian model averaging, and to define a model-averaged profile likelihood interval by analogy with a model-averaged Bayesian credible interval (Hoeting *et al.*, 1999).

In Sections 2.2 and 2.3 we describe the current methods for model averaging, then derive the form of the model-averaged profile likelihood interval. In Section 2.4 we illustrate use of this new model-averaged interval in an example involving negative binomial regression. In Sections 2.5 and 2.6 we describe and give the results from two simulation studies, which assess the coverage properties of the model-averaged profile likelihood interval compared to model-averaged versions of the Wald interval and the transformed Wald interval. We conclude with a discussion in Section 2.7.

2.2 Current Methods

Suppose we wish to estimate a parameter θ using model averaging. Typically, model averaging is most useful for prediction, as θ then has the same interpretation in all models. We therefore suppose that θ is the expected value of a response variable Y given the values of a set of predictor variables. Suppose our set of models is $\{M_1, \ldots, M_R\}$. Following Buckland *et al.* (1997), a model-averaged estimate of θ is

$$\hat{\theta} \;=\; \sum_{i=1}^{R} w_i \, \hat{\theta}_i, \tag{2.1}$$

where w_i is the weight given to M_i and $\hat{\theta}_i$ is the estimate under M_i. A $(1-2\alpha)100\%$ model-averaged Wald interval for θ has the form

$$\hat{\theta} \pm z_\alpha \hat{\text{var}}(\hat{\theta})^{1/2}, \tag{2.2}$$

where $\hat{\text{var}}(\hat{\theta})$ is an estimate of the variance of $\hat{\theta}$, and $\Phi(z_\alpha) = 1-\alpha$, where $\Phi(\cdot)$ is the standard normal distribution function. There have been a number of suggestions as to an appropriate formula for $\hat{\text{var}}(\hat{\theta})$, which appear in Buckland *et al.* (1997), Burnham and Anderson (2002), and Claeskens and Hjort (2008). We consider the second formula proposed by Burnham and Anderson (2002, p.345), which is motivated by analogy with the variance of a model-averaged posterior distribution in the Bayesian setting (Claeskens and Hjort, 2008, p.217),

$$\hat{\text{var}}_2(\hat{\theta}) \;=\; \sum_{i=1}^{R} w_i \left[\hat{\text{var}}\left(\hat{\theta}_i\right) + \left(\hat{\theta}_i - \hat{\theta}\right)^2 \right], \tag{2.3}$$

where $\hat{\text{var}}(\hat{\theta}_i)$ is an estimate of the variance of $\hat{\theta}_i$ given M_i. A Wald confidence interval for the model-averaged estimator $\hat{\theta}$, constructed using this second variance estimator proposed by Burnham and Anderson (2002), $\hat{\text{var}}_2(\hat{\theta})$, will be referred to as the second model-averaged Wald (MAW_2) interval.

A *transformed* model-averaged Wald interval for θ, based on the one-to-one transformation $\phi = g(\theta)$, is given by

$$g^{-1}\left(\hat{\phi} \pm z_\alpha \hat{\text{var}}(\hat{\phi})^{1/2}\right),$$

where $\hat{\phi}$ and $\hat{\text{var}}(\hat{\phi})$ are obtained using equations (2.1) and (2.3), with θ replaced by ϕ. When constructed using the variance estimator in equation (2.3), this *transformed* Wald interval will be referred to as a MAW_2^* interval.

We focus on the use of AIC weights to perform model averaging (Buckland *et al.*, 1997), but the method we derive can be used with any other information criterion, such as $\mathrm{AIC_c}$ or BIC. We do not consider use of the bootstrap, as our aim is to consider methods which are not computationally intensive. The AIC weight for M_i is calculated from the equation

$$w_i \;=\; \frac{\exp\left(-\frac{1}{2}\,\Delta\mathrm{AIC}_i\right)}{\sum_{j=1}^{R}\exp\left(-\frac{1}{2}\,\Delta\mathrm{AIC}_j\right)}, \tag{2.4}$$

where $\Delta\mathrm{AIC}_i = \mathrm{AIC}_i - \mathrm{AIC}_{\min}$, and $\mathrm{AIC}_{\min}$ is the minimum AIC value among all candidate models. The AIC value of model M_i is given as

$$\mathrm{AIC}_i \;=\; -2\log\hat{L}_i + 2p_i,$$

where $\hat{L}_i$ is the maximized value of the likelihood under M_i, and p_i is the number of parameters in M_i.

2.3 Model-Averaged Tail Area Profile Likelihood

In many single-model settings, with an appropriate choice of prior distributions for the parameters, a Bayesian credible interval can be regarded as an approximation to a profile likelihood interval (Severini, 1991). Although it might seem odd to regard a Bayesian interval in this manner, especially from a Bayesian perspective, it is useful to do so here simply as a means of deriving a model-averaged profile likelihood interval. This suggests that a model-averaged Bayesian (MAB) interval might be regarded as an approximation to a model-averaged profile likelihood interval. In order to see this, we first consider the forms of a profile likelihood interval and a Bayesian credible interval. For simplicity of presentation, we focus on the lower confidence limit, as the upper limit is obtained by replacing α with $(1-\alpha)$ throughout the following equations. Given a model M, the lower limit of a $(1-2\alpha)100\%$ profile likelihood interval is defined as the

value of θ_L which satisfies

$$\Phi(r(\theta_L)) \;=\; \alpha, \tag{2.5}$$

where $r(\theta)$ is the signed likelihood ratio statistic, given by

$$r(\theta) \;=\; \text{sign}(\hat{\theta} - \theta)\, \sqrt{2\left(\log L_p(\hat{\theta}) - \log L_p(\theta)\right)},$$

using the maximum likelihood estimate $\hat{\theta}$ of the parameter θ, and the profile likelihood function for θ, given by

$$L_p(\theta) \;=\; \max_{\lambda} L(\theta, \lambda).$$

Here, $L(\theta, \lambda)$ is the likelihood function for model M, parametrized in terms of θ and the remaining (nuisance) parameters λ.

In the single-model setting, an equal-tailed $(1-2\alpha)100\%$ Bayesian credible interval has a lower limit given by the value of θ_L which satisfies

$$\int_{-\infty}^{\theta_L} p(\theta|M, y)\, d\theta \;=\; \alpha, \tag{2.6}$$

where $p(\theta|M, y)$ is the posterior distribution for θ given M, and y represents the observed data.

The fact that a Bayesian interval can provide a close approximation to a profile likelihood interval (for an appropriate choice of priors for the parameters) means that we can obtain an approximate $(1-2\alpha)100\%$ profile likelihood interval by solving equation (2.6) rather than (2.5), even though the philosophies underlying the two equations are quite different. This suggests that we should be able to derive a model-averaged profile likelihood interval using an analogous approximation. We now consider how an equal-tailed MAB interval for θ is calculated. The model-averaged posterior distribu-

tion for θ (Hoeting *et al.*, 1999) is given by

$$p(\theta|y) \;=\; \sum_{i=1}^{R} p(M_i|y)\; p(\theta|M_i, y), \tag{2.7}$$

where $p(M_i|y)$ is the posterior probability for model M_i. By analogy with the single-model Bayesian credible interval in (2.6), the lower limit of a $(1-2\alpha)100\%$ MAB interval is given by the value of θ_L which satisfies

$$\int_{-\infty}^{\theta_L} p(\theta|y)\, d\theta \;=\; \alpha. \tag{2.8}$$

Substituting the expression in equation (2.7) into (2.8), and reversing the order of integration and summation (since R is finite), leads to

$$\sum_{i=1}^{R} p(M_i|y) \int_{-\infty}^{\theta_L} p(\theta|M_i, y)\; d\theta \;=\; \alpha. \tag{2.9}$$

The form of (2.9) implies that the lower limit, θ_L, is such that a weighted mean over all R models of the posterior tail areas associated with θ_L is equal to α, where the weights are the posterior model probabilities.

The forms of equations (2.5) and (2.6) suggest that we can obtain an approximation to the lower limit of a $(1-2\alpha)100\%$ model-averaged profile likelihood interval by finding the value of θ_L which satisfies

$$\sum_{i=1}^{R} p(M_i|y)\; \Phi(r_i(\theta_L)) \;=\; \alpha, \tag{2.10}$$

where $r_i(\cdot)$ is defined analogously under model M_i. Finally, if we replace the posterior model probabilities $p(M_i|y)$ with AIC model weights, as given in (2.4), we obtain the following definition for the lower limit of a $(1-2\alpha)100\%$ model-averaged profile

likelihood interval:

$$\sum_{i=1}^{R} w_i \, \Phi(r_i(\theta_L)) \;=\; \alpha. \tag{2.11}$$

By analogy with the interpretation of a MAB interval, equation (2.11) implies that the lower limit (θ_L) of a $(1-2\alpha)100\%$ model-averaged profile likelihood interval is such that a weighted mean over all R models of the nominal error rates associated with θ_L is equal to α. As this involves averaging the "tail areas" of the sampling distributions of single-model statistics, this interval will be referred to as the model-averaged tail area profile likelihood (MATA-PL) interval.

Unlike a profile likelihood interval for a single model, our derivation of a MATA-PL interval is based on an analogy with Bayesian model averaging, rather than large-sample distribution theory. This type of approach was also used by Buckland *et al.* (1997) to determine a formula for calculating model weights based on an information criterion. In addition, the use of equation (2.1) can be motivated by an analogy with the mean of the model-averaged posterior distribution when vague priors are used for the parameters, with the model weights again taking the place of the posterior model probabilities and the single-model parameter estimates replacing the corresponding posterior means (Claeskens and Hjort, 2008, p.217). Furthermore, as mentioned in Section 2.2, the choice of equation (2.3) is based on an analogy with the variance of a model-averaged posterior distribution.

2.4 Example: Ecklonia Abundance

We illustrate use of the different types of model-averaged confidence intervals using data from a study of the relationship between abundance of a seaweed (*Ecklonia radiata*; hereafter *Ecklonia*) and that of a sea urchin (*Evechinus chloroticus*; hereafter

Evechinus) in Fiordland, New Zealand. The primary aim of the study was to predict the increase in *Ecklonia* abundance that would arise as a consequence of a specified reduction in *Evechinus* abundance, the aim being to assess the potential impacts of commercial harvesting of *Evechinus*.

At each of 103 locations, abundance of both *Ecklonia* and *Evechinus* were measured using 25m^2 quadrats. In addition, a number of physical variables were measured, as some of these were thought to influence the relationship between abundance of *Ecklonia* and *Evechinus*. For simplicity of presentation we consider only one of these variables, the distance to the mouth of the fiord, which was regarded as an index of exposure to sea swell. Further information on the study and the complete dataset can be found in Fletcher *et al.* (2005). The subset of the full dataset which we consider is shown in Table 2.1.

We make use of a negative binomial regression to predict *Ecklonia* abundance, using the two predictors *Evechinus* abundance and distance to the mouth of the fiord. Let Y_i be the *Ecklonia* density (plants per quadrat) observed in location i, for $i = 1, \ldots, n$. We assume that Y_i has a negative binomial distribution with mean μ_i and dispersion parameter k $(\mu_i, k > 0)$. Note that k is assumed to be the same for all observations. We consider the following three models for μ_i, where we omit the subscript i for simplicity of presentation:

M_1: $\log \mu = \alpha + \beta_1 x_1$

M_2: $\log \mu = \alpha + \beta_1 x_1 + \beta_2 x_2$

M_3: $\log \mu = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$

where x_1 is *Evechinus* density (individuals per quadrat), and x_2 is the distance to the mouth of the fiord (km).

Table 2.1: Data for use in the *Ecklonia* abundance negative binomial regression example, recorded in Fiordland, New Zealand. Data include *Ecklonia* density (y), *Evechinus* density (x_1) and distance to the mouth of the fiord (x_2).

y	x_1	x_2	y	x_1	x_2	y	x_1	x_2
0	0	7	2	35	12	45	0	12
0	0	7	3	14	11	46	28	4
0	0	12	3	115	9	48	4	13
0	3	11	6	22	5	49	0	9
0	5	10	8	15	7	56	29	8
0	5	11	8	53	7	57	2	12
0	6	10	11	64	7	57	11	7
0	11	9	12	28	11	57	15	11
0	13	11	12	52	9	58	0	11
0	18	9	15	14	11	58	28	7
0	18	9	16	0	11	59	0	14
0	19	8	19	11	11	59	9	8
0	19	7	20	11	6	61	0	12
0	27	7	20	14	8	64	1	12
0	27	7	20	20	10	67	0	14
0	29	7	22	7	11	70	0	12
0	31	8	24	5	7	72	0	9
0	32	6	25	0	12	73	5	9
0	33	10	26	12	11	79	2	10
0	34	11	28	20	5	81	3	11
0	38	13	30	11	9	81	8	11
0	39	5	30	11	13	84	0	10
0	50	6	31	20	6	87	0	10
0	78	6	33	13	14	88	25	12
0	110	5	33	19	11	89	2	12
0	133	7	33	29	8	94	28	12
0	220	11	33	52	5	96	2	11
1	15	10	34	24	6	96	4	7
1	15	6	35	25	7	103	29	11
1	23	6	38	0	13	118	13	9
2	0	4	38	19	11	137	0	11
2	15	6	39	0	13	157	4	12
2	16	11	40	35	4	168	5	12
2	19	11	44	0	9	203	50	3
2	34	11						

Figure 2.1 (a) and (b) show estimates of the relationship between *Ecklonia* and *Evechinus* abundance for $x_2 = 7$ and $x_2 = 11$, where the models have been fitted using maximum likelihood. The fitted line for M_3 differs from the other two models, particularly for lower values of x_1 when $x_2 = 7$. Models M_1 and M_2 provide very similar fitted lines, for both values of x_2.

Figure 2.1 (c) and (d) show the model-averaged estimates of the relationship for $x_2 = 7$ and $x_2 = 11$, together with the three types of model-averaged confidence intervals, based on the AIC weights of $w_1 = 0.33$, $w_2 = 0.15$ and $w_3 = 0.52$, and using a logarithmic transformation for constructing the MAW_2^* interval. The intervals differ most for the higher values of x_1, for which there are fewer observations (see Table 2.1). The lower and upper limits for the MAW_2 interval are always lower than the corresponding limits for the MAW_2^* and MATA-PL intervals, as would be expected from its behavior in single-model settings. For the lower limit, the MAW_2^* and MATA-PL intervals are almost identical, while for the upper limit, the MAW_2^* interval is smaller than the MATA-PL interval when $x_2 = 7$ but higher when $x_2 = 11$, particularly for the higher values of x_1.

2.5 Simulation Study

In order to assess the performance of the different methods of model averaging, we performed two simulation studies. The first was based on the example in Section 2.4, while the second was based on a simpler setting which allowed a more extensive investigation of the differences between the methods.

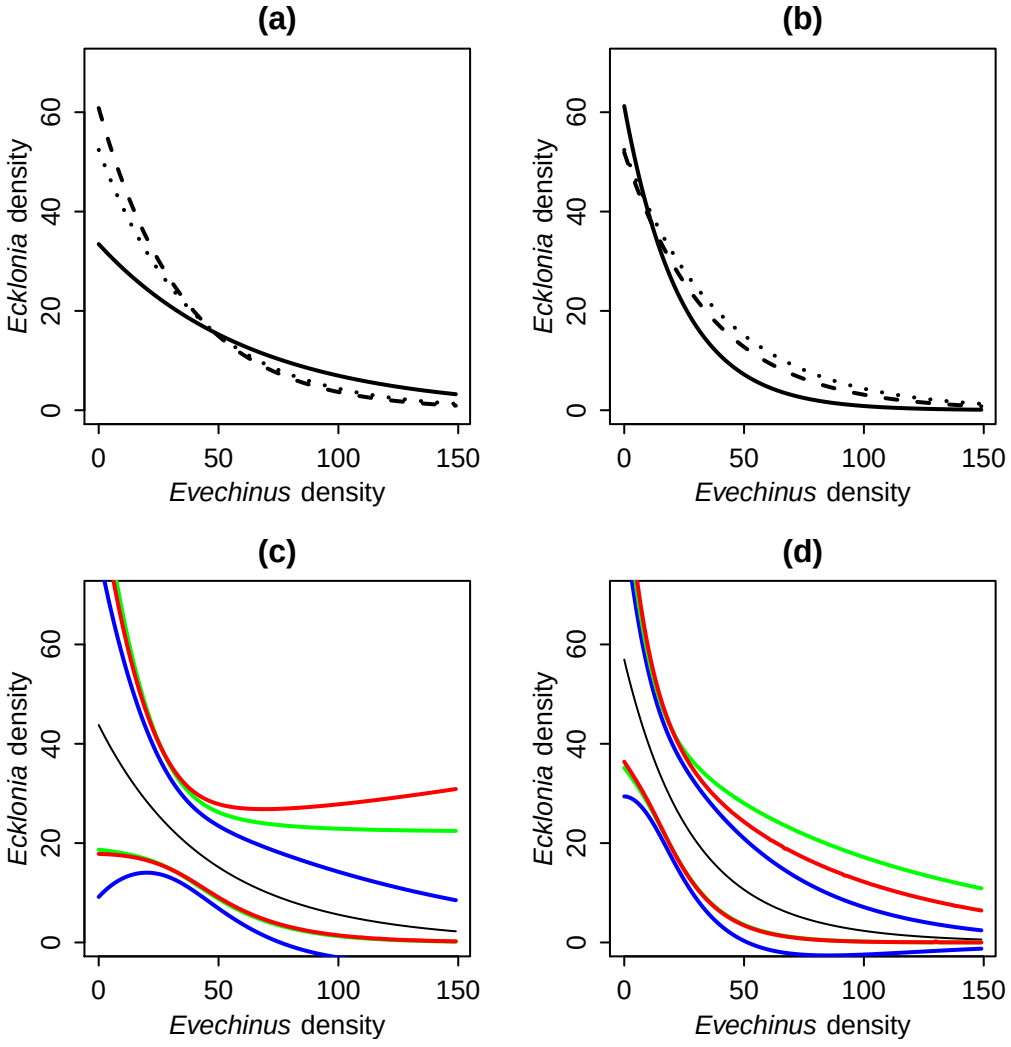

Figure 2.1: Graphs (a) and (b) show predicted *Ecklonia* density versus *Evechinus* density for models M_1 (dotted), M_2 (dashed) and M_3 (solid), when the distance to the mouth of the fiord is 7km or 11km. Graphs (c) and (d) show the corresponding model-averaged predictions (black lines), with MAW_2 (blue), MAW_2^* (green) and MATA-PL (red) confidence intervals.

2.5.1 Negative Binomial

We focused on estimation of mean *Ecklonia* density for four combinations of *Evechinus* density and distance to the fiord mouth: $x_1 = 25, 100$ and $x_2 = 7, 11$. These values were chosen to cover the range observed for the predictor variables; in particular, they allowed us to consider cases where estimation of μ would be relatively precise ($x_1 = 25$) versus those where it would not ($x_1 = 100$). We set the sample size equal to that of the real dataset ($n = 103$), and considered the three models described in Section 2.4. We used 10,000 simulations, giving standard errors for the lower and upper error rates of at most 0.5%. The data were generated from the largest model, M_3, with the parameter values set equal to the maximum likelihood estimates for the real data. We specified the likelihood in terms of μ by expressing the intercept term α as a function of μ and the remaining parameters.

2.5.2 Lognormal

In this study we considered observations Y_{ij} with $\log Y_{ij} \sim N(\beta_i, \sigma^2)$, for $i = 1, 2$, and $j = 1, \ldots, n$. We wish to estimate each $\mu_i \equiv \mathrm{E}[Y_{ij}] = \exp(\beta_i + \frac{\sigma^2}{2})$. Model averaging was based on the following two models:

M_1: $\beta_1 = \beta_2$
M_2: β_1 and β_2 unspecified

The data were generated from the largest model, M_2. As the results will be influenced by the difference between β_1 and β_2, and by the values of β_1 and β_2 relative to σ^2, we arbitrarily set $\beta_1 = 0$ and $\sigma^2 = 1$. We considered $\beta_2 = 0, 0.1, \ldots, 3$ and sample sizes of $n = 10, 25, 50$. We used 25,000 simulations, giving standard errors for the lower and upper error rates of at most 0.3%. We specified the likelihood in terms of μ_i by

expressing each β_i as $\log \mu_i - \frac{\sigma^2}{2}$.

2.5.3 Computations

All calculations were performed in R Version 2.9.0 (2009). We used the quasi-Newton method in the *optim* function to maximize the likelihood and *uniroot* to solve equation (2.11). For both studies, we used the log-transformation for the MAW_2^* interval. In the lognormal simulation, estimates of the asymptotic standard errors of maximum likelihood estimators were obtained from the hessian matrix, which was obtained as part of the optimization procedure. In the negative binomial simulation study, the hessian matrix was obtained using the method *hessian* from R library *numDeriv*. This matrix was found to be singular in 0.3% of the simulation runs; results for these runs were discarded.

For both simulation studies, we summarized the performance of each method by calculating the lower and upper error rates, i.e. the proportion of simulation runs for which $\theta_L > \theta$ and for which $\theta_U < \theta$, where θ_L and θ_U are the lower and upper confidence limits, respectively. In addition, we calculated the mean of the lower and upper relative half-widths, which were respectively defined as $\frac{\theta_L - \theta}{\theta}$ and $\frac{\theta_U - \theta}{\theta}$.

2.6 Simulation Results

2.6.1 Negative Binomial

Table 2.2 shows the error rates and mean half-widths, for each combination of the predictor variables and for each method of model averaging. Both the MAW_2^* and

Table 2.2: Summary of results from the the negative binomial simulation study. The error rates and relative half-widths for the MAW_2, MAW_2^* and MATA-PL confidence intervals are shown, for each of four combinations of *Evechinus* density (x_1) and distance to the mouth of the fiord (x_2). The true mean *Evechinus* density (μ) is also shown.

				Error Rates		Half-Widths	
x_1	x_2	μ	Interval	Lower	Upper	Lower	Upper
25	7	22.6	MAW_2	0.7	6.0	-0.41	0.46
25	7	22.6	MAW_2^*	2.5	3.6	-0.33	0.56
25	7	22.6	MATA-PL	3.0	2.6	-0.32	0.60
100	7	7.0	MAW_2	0.0	21.9	-1.17	0.98
100	7	7.0	MAW_2^*	0.5	8.7	-0.71	1.93
100	7	7.0	MATA-PL	1.2	6.7	-0.68	2.54
25	11	21.0	MAW_2	1.0	5.2	-0.48	0.61
25	11	21.0	MAW_2^*	3.8	2.9	-0.36	0.76
25	11	21.0	MATA-PL	3.9	2.0	-0.36	0.77
100	11	0.8	MAW_2	0.0	4.0	-4.28	8.39
100	11	0.8	MAW_2^*	12.3	2.7	-0.61	19.68
100	11	0.8	MATA-PL	8.5	2.2	-0.67	15.06

MATA-PL intervals clearly outperform the MAW_2 interval, as expected. Where estimation of the interval is difficult, such as the upper limit when $x_1 = 100$ and $x_2 = 7$, or the lower limit when $x_1 = 100$ and $x_2 = 11$, the MATA-PL interval clearly performs better than the MAW_2^* interval, even if it does not come close to the nominal level. Otherwise, there is little to differentiate between the MAW_2^* and MATA-PL intervals, any differences in performance being roughly within simulation error.

The relative half-widths show that the MAW_2 interval has limits that are consistently lower than those of the other two methods, as expected from the error rates. Likewise, when the MATA-PL interval outperforms the MAW_2^* interval, it has limits that are higher than those of the MAW_2^* interval, the difference being most apparent for the upper limit when $x_1 = 100$ and $x_2 = 7$. When $x_1 = 100$ and $x_2 = 11$, the MATA-PL interval provides an upper limit that is generally lower than that for the MAW_2^* interval, even though the error rates for the two methods are both close to the nominal level.

2.6.2 Lognormal

We focus on the results for μ_2, as those for μ_1 were qualitatively similar. The left column of Figure 2.2 shows the lower and upper error rates plotted against β_2, for each value of the sample size. The MATA-PL interval clearly outperforms the other two methods, with an error rate consistently closer to the nominal level. As expected, the MAW_2 interval is also consistently worse than the MAW_2^* interval.

As β_2 increases, the performance of each method worsens, as a consequence of model uncertainty increasing as β_2 approaches the value of σ^2. The exception to this is the lower error rate for the MAW_2 interval, which is consistently poor. The relationship with β_2 is clearest for the upper error rate, where the worst performance occurs when β_2

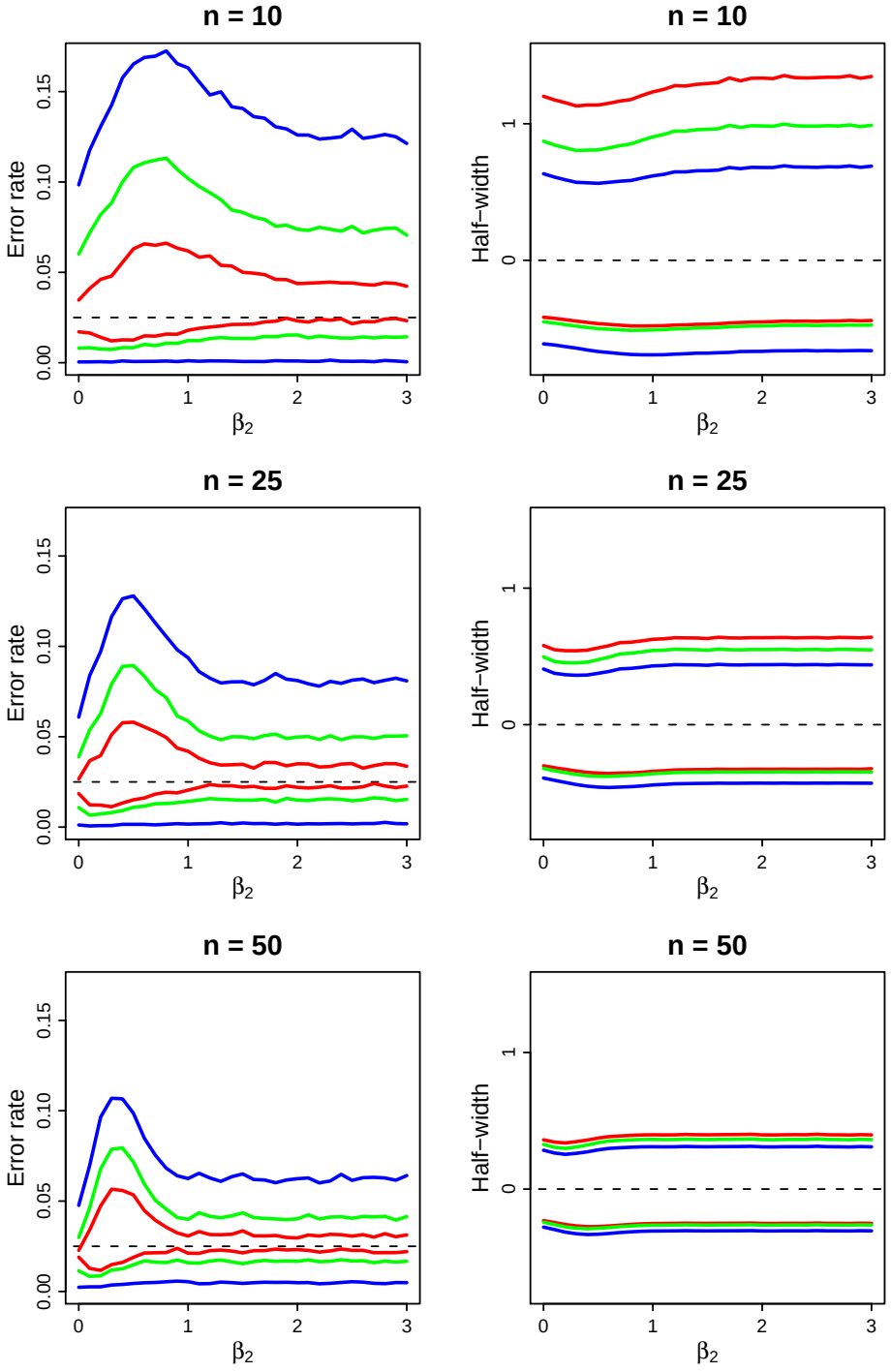

Figure 2.2: Error rates (left) and mean relative half-widths (right) showing the lognormal simulation results for the MAW_2 (blue), MAW_2^* (green) and MATA-PL (red) confidence intervals. The dashed reference lines are the nominal 2.5% error rate (left) and zero (right).

lies between zero and one. Figure 2.3 shows the mean AIC weight for model M_2 plotted as a function of β_2, for each value of the sample size. The mean M_2 weight increases as β_2 increases, and attains the value $w_2 = 0.5$ for values of β_2 roughly between 0.25 and 0.5. This agrees with the patterns seen in Figure 2.2, as model weights of $w_1 = w_2 = 0.5$ correspond to the greatest model uncertainty.

As would be expected, the error rates approach the nominal level as n increases. All the methods perform poorly on the upper error rate when n is small, although the MATA-PL interval is still a clear improvement on the other two methods, mirroring the results of the negative binomial simulations. As n increases, the range of values of β_2 for which the methods perform poorly decreases, due to an overall decrease in model uncertainty. However, there is still a degradation in performance as β_2 increases from zero, the worst performance occurring at smaller values of β_2 as n increases, as would be predicted from Figure 2.3.

The right column of Figure 2.2 shows the mean relative half-widths plotted against β_2, for each sample size considered. The superior performance of the MATA-PL interval is a consequence of it having limits that are higher than those of the other methods, the difference being most apparent for the upper limit. The MAW_2^* interval has lower limits that are approximately the same as those of the MATA-PL interval, even though it does not achieve the same performance on the lower error rate, again mirroring some results from the negative binomial simulations.

2.7 Discussion

The aim of this chapter has been to provide a method for calculating model-averaged profile likelihood confidence intervals. Each endpoint of the interval is obtained by

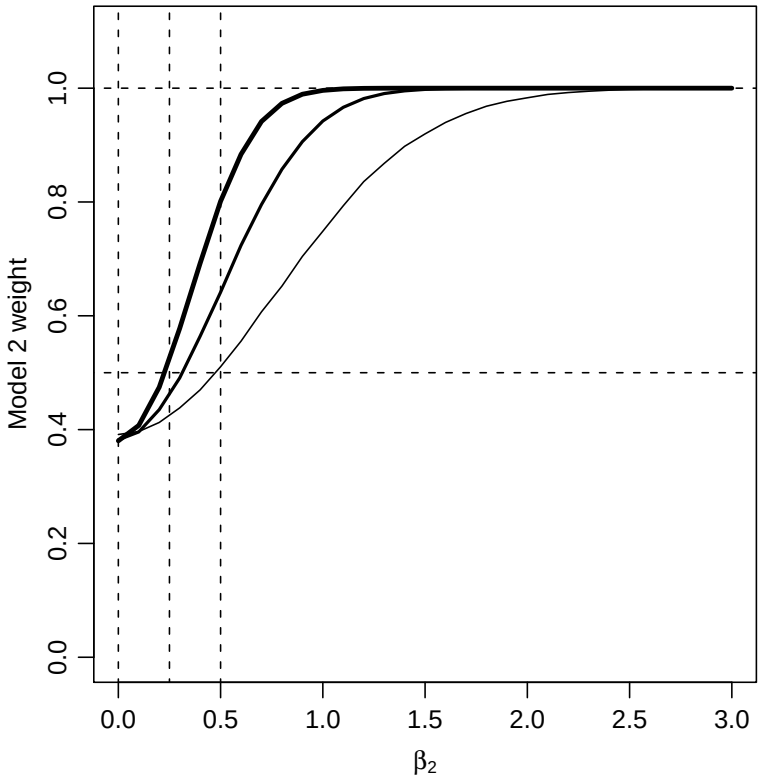

Figure 2.3: Mean AIC weight for model M_2 (w_2) versus β_2, for sample sizes $n = 10$ (thin), $n = 25$ (medium), and $n = 50$ (thick). The dashed horizontal lines at $w_2 = 0.5,\ 1$ and the dashed vertical lines at $\beta_2 = 0,\ 0.25,\ 0.5$ are plotted for ease of reference.

ensuring that a weighted mean, over all models, of the nominal error rates associated with that endpoint is equal to the required error rate, weighting with respect to the AIC model weights. Note that this is not the same as taking a weighted mean of the profile likelihood confidence interval limits for each model, an approach that would not be justified given the non-linear relationship between a confidence limit and its error rate.

We would expect the MATA-PL interval defined here to work well whenever the single-model profile likelihood interval works well for each candidate model. Our simulation results certainly suggest that use of the MATA-PL interval is preferable to use of model-averaged Wald intervals, with or without using a transformation of the parameter.

The method we have proposed is based on the idea that a model-averaged Bayesian credible interval, with an appropriate choice of prior distributions for the parameters, should provide a good approximation to the MATA-PL interval. As discussed in Section 2.2, the frequentist model averaging literature contains examples of similar reasoning. Indeed, Buckland *et al.* (1997) derived their original formula for model weights by analogy with Bayesian methods for comparing models.

We have not made use of asymptotic distribution theory in deriving the method, unlike that used to justify a single-model profile likelihood interval. If there is a single model that is closest to "truth" (in terms of Kullback-Leibler distance) among all candidate models, the MATA-PL interval will converge asymptotically to a profile likelihood interval for this model, since the AIC weight for that model converges in probability to one (Claeskens and Hjort, 2008, p.99-102). Any derivation of results on convergence of the error rates for the MATA-PL interval will be complicated by the weighting and the need to estimate the model weights.

We have chosen to focus on settings in which the data are skewed, as we would expect the sampling distribution of a single-model estimate to be non-normal. However, even for the normal linear model, the sampling distribution of a model-averaged estimate may be sufficiently non-normal to cause a model-averaged Wald interval to perform poorly. Furthermore, this suggests that a MATA-PL interval might generally be expected to perform better than a MAW_2^* interval, as the latter relies on the sampling distribution of the model-averaged estimate being approximately normal after transformation of the parameter. Even if the transformation achieves exact normality of the sampling distribution for a single-model estimate, it does not guarantee that a MAW_2^* interval will perform well. Furthermore, use of a MAW_2^* interval requires consideration of which transformation might be appropriate, whereas a MATA-PL interval does not.

In our simulations, we have assumed that "truth is in the model set," and that it is the largest of those models. The first assumption does not imply that we believe real data to be generated by a statistical model. We have assumed this solely in order to simplify the simulation procedure. We could have generated data from a more complex model, but there is no reason to believe this would lead to any greater insight. Clearly, all the methods would be expected to perform worse if none of the models are close to "truth." In addition, there would be some arbitrariness in the choice of a more complex model. The second assumption is based on our belief that "truth" is generally complex, and therefore it is unrealistic to generate the data from one of the simpler models. Having said that, the case of $\beta_2 = 0$ in the lognormal simulation corresponds to models M_1 and M_2 both being "truth."

It is worth noting that substitution of the AIC weights for the posterior model probabilities can be thought of as making use of implicit priors for the models (Burnham and Anderson, 2004). A further discussion of this point, and the implied prior

distributions appears in Link and Barker (2006).

Although we have used ideas from Bayesian model averaging, we have not considered direct use of a model-averaged Bayesian interval here, as it is computationally more intensive than a MAW or MATA-PL interval. For the same reason, we have not considered the use of bootstrapping, either to construct a confidence interval or to estimate model weights.

We have focused on the use of AIC weights, but the method for calculating a MATA-PL interval can be used with any information criterion, such as AIC_c or BIC. For the reasons outlined above, we would expect a MATA-PL interval to outperform the other methods regardless of which criterion is used. It would be of interest to compare the performance of a MATA-PL interval using different information criteria, as previous comparisons of information criteria have tended to focus on model selection and predictive mean squared error, rather than coverage rates of confidence intervals (Burnham and Anderson, 2004; Claeskens and Hjort, 2008).

Chapter 3

Model-Averaged Wald Intervals[2]

3.1 Introduction

Statistical inference has traditionally been based on a single model, which would typically be chosen through a process of model selection. Such an approach ignores model uncertainty, which can introduce bias, and lead to overestimating the precision of the resulting inferences (Chatfield, 1995).

Model averaging is one approach to allow for model uncertainty (Burnham and Anderson, 2002; Claeskens and Hjort, 2008; Raftery *et al.*, 1997). Throughout this chapter we focus on frequentist model averaging. Suppose that we wish to obtain an estimate for a parameter θ and that we have a set of candidate models $\{M_1, \ldots, M_R\}$. Instead of selecting a "best" model M_i and using the parameter estimate $\hat{\theta}_i$ under that model, we calculate a model-averaged estimate as the weighted sum of single-model estimates, i.e. $\hat{\theta} = \sum_{i=1}^{R} w_i \hat{\theta}_i$, where w_i is the weight for M_i, which is usually based

[2] The content of this chapter has been published in *Computational Statistics & Data Analysis*; see Turek and Fletcher (2012).

on an information criterion such as AIC (Buckland *et al.*, 1997). Note that model averaging only makes sense when θ has a consistent interpretation across all candidate models.

The sampling distribution of $\hat{\theta}$ is more complex than in the single-model setting, being a mixture of the sampling distributions of each $\hat{\theta}_i$. This leads to complications in the calculation of a model-averaged confidence interval for θ. One simple approach is to calculate a model-averaged $(1-2\alpha)100\%$ confidence interval using a Wald interval of the form $\hat{\theta} \pm z_\alpha \hat{\text{var}}(\hat{\theta})^{1/2}$, where z_α is the $(1-\alpha)$ quantile of the standard normal distribution. Even for this simple approach there is the difficulty of determining a suitable estimate for $\text{var}(\hat{\theta})$. Several options were put forward in Buckland *et al.* (1997) and further discussed in Burnham and Anderson (2002). Claeskens and Hjort (2008, p.206-207) noted the difficulty in estimating $\text{var}(\hat{\theta})$ and showed that such intervals will often perform poorly in terms of coverage rate. The issue of estimating $\text{var}(\hat{\theta})$ aside, use of such an interval also incorrectly assumes a normal sampling distribution for $\hat{\theta}$.

We propose a new method for calculating a model-averaged confidence interval, which we refer to as a model-averaged tail area Wald (MATA-Wald) interval. This interval is based on single-model Wald intervals rather than an overall Wald interval. It does not involve the assumption of normality for $\hat{\theta}$, nor an estimate of $\text{var}(\hat{\theta})$. It is based on the assumption that exactly one model among the set of candidate models is true, and that each $\hat{\theta}_i$ is normally distributed *given* that M_i is true. As in many settings, this assumption of normality might be more closely satisfied after a suitable transformation of θ.

In Section 3.2, we describe the current methodology for calculating model-averaged Wald confidence intervals. Section 3.3 develops the MATA-Wald confidence interval, for both normal and non-normal data. In Sections 3.4 and 3.5, we describe a simulation study to compare the coverage performance of the MATA-Wald interval against the

existing methods, in the setting of linear regression. We conclude with a discussion in Section 3.6.

3.2 Current Methods

We consider two existing methods for constructing model-averaged Wald confidence intervals, both of the form $\hat{\theta} \pm z_\alpha \hat{\text{var}}(\hat{\theta})^{1/2}$. The first model-averaged Wald interval (MAW_1) uses the first estimate of $\text{var}(\hat{\theta})$ proposed by Burnham and Anderson (2002, p.164), which here includes a scaling adjustment for the uncertainty in estimating each $\text{var}(\hat{\theta}_i)$. This is the *adjusted* variance estimator, as described in Section 1.3.3:

$$\hat{\text{var}}_1(\hat{\theta}) \;=\; \left[\sum_{i=1}^{R} w_i \sqrt{(t_{\nu_i}/z_\alpha)^2 \,\hat{\text{var}}\left(\hat{\theta}_i\right) + \left(\hat{\theta}_i - \hat{\theta}\right)^2} \;\right]^2,$$

where t_{ν_i} is the $(1-\alpha)$ quantile of the t-distribution with ν_i degrees of freedom, ν_i is the error degrees of freedom associated with model M_i, and the summation is over the R models in the candidate set. The second model-averaged Wald interval (MAW_2) uses the second estimate of $\text{var}(\hat{\theta})$ proposed by Burnham and Anderson (2002, p.345), which is again adjusted for the uncertainty in each $\hat{\text{var}}(\hat{\theta}_i)$:

$$\hat{\text{var}}_2(\hat{\theta}) \;=\; \sum_{i=1}^{R} w_i \left[(t_{\nu_i}/z_\alpha)^2 \,\hat{\text{var}}\left(\hat{\theta}_i\right) + \left(\hat{\theta}_i - \hat{\theta}\right)^2 \right].$$

By Jensen's inequality, $\hat{\text{var}}_1(\hat{\theta}) \leq \hat{\text{var}}_2(\hat{\theta})$, so the MAW_2 interval will always be at least as wide as the MAW_1 interval.

3.3 Model-Averaged Tail Area Wald

We develop the MATA-Wald interval initially for the simplest case of the normal linear model, and then extend it to other settings.

3.3.1 Normal Linear Model

Suppose that exactly one of the candidate models is true and that we wish to determine a lower $(1-2\alpha)100\%$ confidence limit (θ_L) for θ. By definition, we require that the sampling distribution of θ_L is such that

$$\Pr(\theta < \theta_L) \;=\; \alpha.$$

Let c_i $(i = 1, \ldots, R)$ be an unknown indicator variable which assumes the value one if model M_i is true, and zero otherwise. Then, by definition

$$\Pr(\theta < \theta_L) \;=\; \sum_{i=1}^{R} c_i \,\Pr(\theta_i < \theta_L),$$

where θ_i is the value of θ if model M_i is true. Since $\hat{\theta}_i$ is normally distributed if M_i is true, we have

$$\Pr(\theta_i < \theta_L) \;=\; \Pr(T_i > t_{L,i}) \;=\; 1 - F_{\nu_i}(t_{L,i}),$$

where $T_i = (\hat{\theta}_i - \theta_i)/\hat{\text{var}}(\hat{\theta}_i)^{1/2}$, $t_{L,i} = (\hat{\theta}_i - \theta_L)/\hat{\text{var}}(\hat{\theta}_i)^{1/2}$, $F_\nu(\cdot)$ is the distribution function of the t-distribution with ν degrees of freedom, and ν_i is the error degrees of freedom associated with model M_i. Combining the above results leads to defining θ_L as the value satisfying

$$\sum_{i=1}^{R} c_i \,(1 - F_{\nu_i}(t_{L,i})) \;=\; \alpha. \tag{3.1}$$

As the c_i are unknown, we estimate them using model weights w_i, based on some criterion for comparing the models. This leads to specification of a lower $(1-2\alpha)100\%$ MATA-Wald confidence limit as the value of θ_L satisfying the equation

$$\sum_{i=1}^{R} w_i \left(1 - F_{\nu_i}(t_{L,i})\right) \;=\; \alpha. \tag{3.2}$$

An upper MATA-Wald confidence limit can be defined similarly, as the value of θ_U which satisfies

$$\sum_{i=1}^{R} w_i \, F_{\nu_i}(t_{U,i}) \;=\; \alpha, \tag{3.3}$$

where $t_{U,i} = (\hat{\theta}_i - \theta_U)/\hat{\text{var}}(\hat{\theta}_i)^{1/2}$.

Equations (3.2) and (3.3) can be solved via numerical methods. Use of the term "model-averaged tail area" (MATA) is suggested by the form of these equations, as they each involve a model-weighted sum of tail areas of t-distributions.

3.3.2 Non-Normal Models

It is straightforward to calculate the MATA-Wald confidence interval in other settings, so long as we can specify a transformation $\phi = g(\theta)$ for which the sampling distribution of $\hat{\phi}_i = g(\hat{\theta}_i)$ is approximately normal when M_i is true. An example would be the use of $\phi = \text{logit}(\theta)$ when θ is a probability.

The only change required in the interval construction is use of the standard normal distribution function in place of that of the t-distribution. In the non-normal setting, a $(1-2\alpha)100\%$ MATA-Wald confidence interval for θ is given by the values of θ_L and

θ_U which satisfy the pair of equations

$$\sum_{i=1}^{R} w_i \left(1 - \Phi(z_{L,i})\right) = \alpha,$$
$$\sum_{i=1}^{R} w_i \, \Phi(z_{U,i}) = \alpha,$$

where $z_{L,i} = (\hat{\phi}_i - \phi_L)/\hat{\text{var}}(\hat{\phi}_i)^{1/2}$, $z_{U,i} = (\hat{\phi}_i - \phi_U)/\hat{\text{var}}(\hat{\phi}_i)^{1/2}$, $\phi_L = g(\theta_L)$, $\phi_U = g(\theta_U)$, and $\Phi(\cdot)$ is the standard normal distribution function.

As with all model-averaged confidence intervals, theoretical study of the asymptotic properties of the MATA-Wald interval is complicated by the randomness of the model weights. We therefore make use of a simulation study to assess its performance relative to the existing methods.

3.4 Simulation Study

We carried out a simulation study to compare the performance of the MATA-Wald interval with the two existing model-averaged Wald interval constructions (MAW_1 and MAW_2). We considered a normal linear regression setting involving two predictor variables, x_1 and x_2. We focused on five candidate models, all having $Y_i \sim N(\mu_i, \sigma^2)$, with μ_i specified as follows:

M_1: $\mu_i = \beta_0$

M_2: $\mu_i = \beta_0 + \beta_1 x_{1,i}$

M_3: $\mu_i = \beta_0 + \beta_2 x_{2,i}$

M_4: $\mu_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}$

M_5: $\mu_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_{12} x_{1,i} x_{2,i}$

We arbitrarily set $\sigma^2 = 1$, and chose $\beta_0 = 1$, $\beta_1 = \beta_2 = 0.3$, $\beta_{12} = 0.1$, such that all five models are assigned non-trivial weights for a range of sample sizes.

Rather than choosing fixed sets of values for x_1 and x_2, the covariates were randomly generated for each simulation. The sets x_1 and x_2 were composed of realizations from $X_{1,i} \overset{iid}{\sim} N(0, 1)$ and $X_{2,i} \overset{iid}{\sim} \Gamma(\alpha{=}2, \beta{=}1)$, which simplified the process of defining these sets for each sample size, which took the values $n = 15, 20, \ldots, 100$.

The parameter of interest (θ) was chosen to be the value of μ for a variety of points in the (x_1, x_2) prediction space. These points were selected as quantiles of the x_1 and x_2 generating distributions, providing insight as to how location in the covariate space may affect confidence interval performance. Each combination of the 10%, 30%, 50%, 70% and 90% quantiles of both distributions were analyzed, for a total of 25 prediction points.

For the first set of simulations we generated data exclusively under the largest model, M_5. For the second set, we allowed the generating model to vary randomly, with each candidate model having an equal probability of being selected prior to each simulation run. The motivation behind this was to provide a simple assessment of the performance of the methods for a range of situations, including those in which any of the candidate models represented truth.

We used each of following three information criteria for calculating model weights:

$$\begin{aligned} \text{AIC} &= -2 \log \hat{L} + 2p \\ \text{AIC}_\text{c} &= -2 \log \hat{L} + 2p \, (\tfrac{n}{n-p-1}) \\ \text{BIC} &= -2 \log \hat{L} + p \log n, \end{aligned}$$

where $\hat{L}$ is the maximized likelihood and p is the number of parameters. Model weights

were then calculated as in Buckland *et al.* (1997),

$$w_i \;\propto\; \exp\left(-\tfrac{1}{2}\,\Delta\mathrm{IC}_i\right),$$

where $\Delta\mathrm{IC}_i \equiv \mathrm{IC}_i - \min_j(\mathrm{IC}_j)$, and IC_i is the value of the information criterion for model M_i.

We assessed the performance of each method by calculating the lower and upper error rate for that interval, i.e. the proportion of simulations for which $\theta_L > \theta$ or $\theta_U < \theta$. Results were averaged over 100,000 simulation runs, thereby ensuring that the standard error for each error rate was at most 0.1%. We also calculated the mean relative half-widths, where the lower and upper relative half-widths were defined as $\frac{\theta-\theta_L}{\theta}$ and $\frac{\theta_U-\theta}{\theta}$ respectively.

All calculations were performed in R, version 2.12.0 (2010). Numerical solutions to equations (3.2) and (3.3) were obtained using the root-finding command *uniroot*.

3.5 Simulation Results

For simplicity, we focus on results for prediction of μ at the 50% and 90% quantiles of the x_1 and x_2 generating distributions, respectively, as results from all prediction points were qualitatively similar in terms of comparing the model-averaged intervals. The main effect of the prediction point location was to improve error rate performance as either x_1 or x_2 approached the median of their respective generating distributions.

3.5.1 Largest Model is Truth

Figure 3.1 provides a summary of the results for the first set of simulations, in which the data were generated exclusively under the largest model, M_5. The leftmost column shows the lower and upper error rates plotted against the sample size, for each method and each choice of model weights. There is little variation among the lower error rates, which are plotted as negative values for display purposes. All of these are close to or just below the nominal level; the only discernible difference is for small sample sizes, where the MATA-Wald interval has a slightly smaller error rate than the other two methods.

There is substantially more variation among the upper error rates. The MATA-Wald interval has an upper error rate that is consistently closer to the nominal level than that of the other methods, while the MAW_1 interval consistently performs worse than the MAW_2 interval, presumably a consequence of being narrower. Regardless of the method, the upper error rates are generally smaller when using AIC model weights. Interestingly, use of AIC_c leads to worse upper error rates as n decreases, even though it is usually preferred to AIC for model selection in the case of small samples. As would be expected, all the error rates approach the nominal level as n increases. Use of the true model (M_5) leads to the error rates being exactly equal to the nominal level, as would be expected in the normal linear setting.

The center column of Figure 3.1 shows the error rates plotted against the relative half-widths, with points of the same color corresponding to different sample sizes (larger sample sizes corresponding to smaller half-widths). We can see that the superior performance of the MATA-Wald interval (compared to the MAW_1 and MAW_2 intervals) in the upper error rate is achieved with only a small (in some cases negligible) increase in the half-width. For the lower limit, the MATA-Wald interval achieves a smaller error

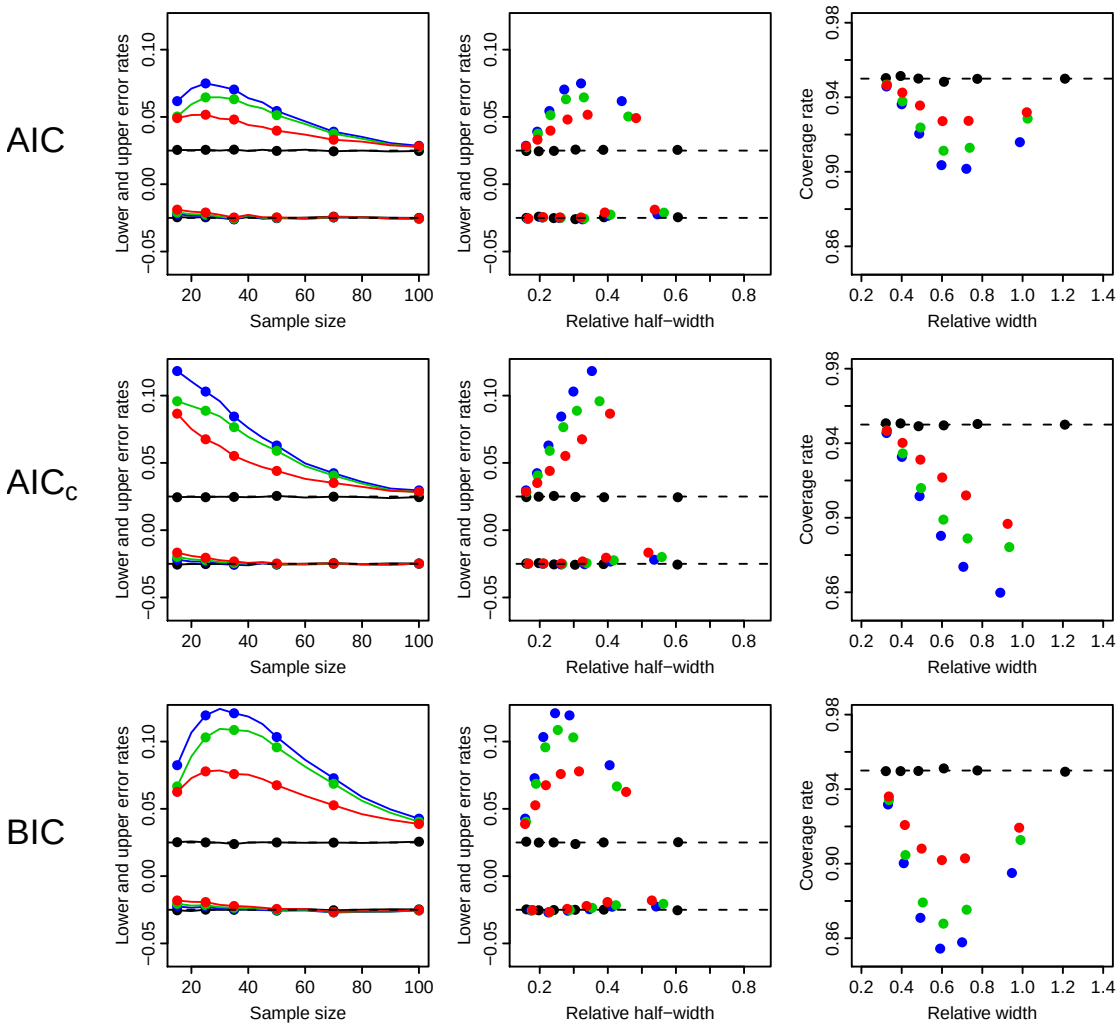

Figure 3.1: Performance of the MATA-Wald (red), MAW_1 (blue), and MAW_2 (green) confidence intervals for prediction of the mean, at the 50% and 90% quantiles of the x_1 and x_2 generating distributions, respectively. The first, second, and third rows use AIC, AIC_c, and BIC weights. Nominal rates are shown as dashed lines. The data are generated under M_5, and the black points show the performance of the Wald interval based on this model.

rate with a *smaller* half-width. The superior coverage performance of the true model comes at a substantial cost in terms of interval width: for small sample sizes the upper half-width is as much as 30% longer than that of the MATA-Wald interval, and the lower half-width up to 20% longer.

The rightmost column of Figure 3.1 provides a similar summary, in terms of the overall coverage rate and the total (lower + upper) relative width. We see that the MATA-Wald interval consistently outperforms the MAW_2 interval in both respects, being narrower and having better coverage. The MAW_1 interval is occasionally narrower than the MATA-Wald interval, but provides a substantially worse coverage rate. Once again, use of the true model achieves the nominal coverage rate, however this comes at the expense of being up to 25% wider than the MATA-Wald interval.

3.5.2 Random Generating Model

Figure 3.2 provides a similar summary for the second set of simulations, in which the true model was equally likely to be any of the candidate models. The comparisons between methods and between information criteria are similar to those for the first set of simulations, the main differences being that the upper error rates are closer to the nominal level and the half-widths are larger. This appears to arise because in 40% of the simulations the data are generated under M_1 or M_2. These models do not include x_2, so use of this predictor is unnecessary and the majority of the candidate models will achieve the nominal coverage rate. This leads to improved coverage for all model-averaged intervals. In addition, as β_2 is positive, both M_1 and M_2 produce lower values for μ than M_5, which leads to a larger *relative* half-width. Another difference from the first set of simulations is that use of the true model leads to substantially smaller relative half-widths, as a consequence of some of the data being generated under smaller models.

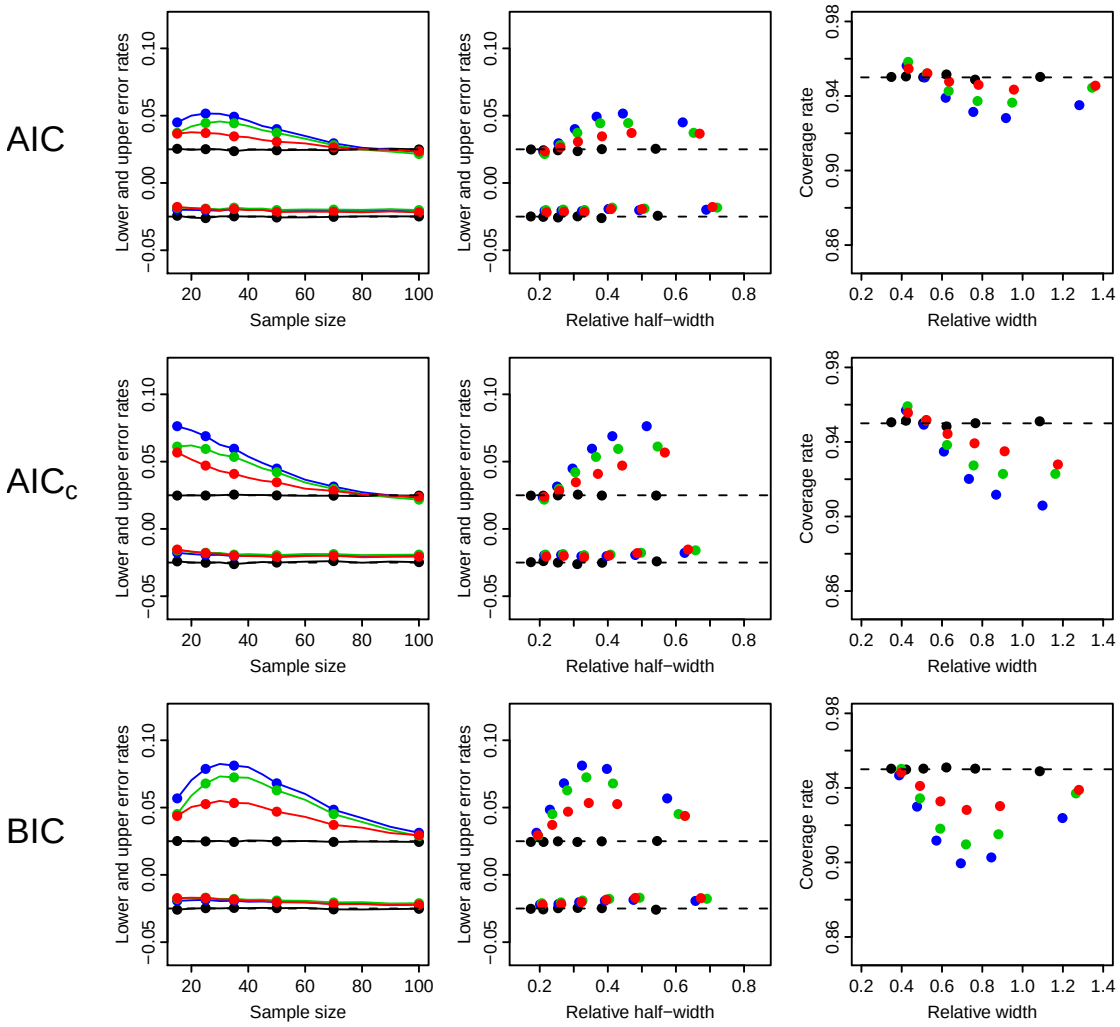

Figure 3.2: Performance of the MATA-Wald (red), MAW_1 (blue), and MAW_2 (green) confidence intervals for prediction of the mean, at the 50% and 90% quantiles of the x_1 and x_2 generating distributions, respectively. The first, second, and third rows use AIC, AIC_c, and BIC weights. Nominal rates are shown as dashed lines. The data are generated from a randomly selected model at the onset of each simulation, and the black points show the performance of the Wald interval based on this model.

Figure 3.3 shows the mean values of the model weights for each set of simulations. When the data are generated under M_5, the weight for that model approaches one as n increases, albeit quite slowly for BIC. Choice of a relatively small interaction term ($\beta_{12} = 0.1$) accounts for the high values of the M_4 model weight. For the second set of simulations, the mean weights cluster in the neighborhood of 0.2. As would be expected, AIC favors larger models compared to AIC_c or BIC. This explains the superior performance, in terms of error rates, of the AIC weights in the first set of simulations, where the true model is the largest. However, this superior performance also holds in the second set of simulations, where the true model is generally not the largest. We also carried out a set of simulations in which M_3 was the true model, and obtained similar results in terms of comparing the three criteria. The results for this simulation, as well as results for data generation under M_2, appear in Appendix B.1. As noted above, it is interesting that AIC_c produces overly narrow intervals for small sample sizes. It would appear that the AIC_c correction term degrades interval performance in this setting, regardless of the data generation scheme.

3.6 Discussion

We have proposed a new method for calculating a model-averaged Wald confidence interval. It differs from the existing model-averaged Wald interval constructions, which rely upon estimation of the variance of $\hat{\theta}$, which is complicated by the randomness of the model weights. Burnham and Anderson (2002, p.345) admit that estimation of this variance term is "worthy of more research." In addition, the existing model-averaged interval constructions rely upon the incorrect assumption that $\hat{\theta}$ follows a normal sampling distribution.

The MATA-Wald interval is based on the simple idea that if we knew the true

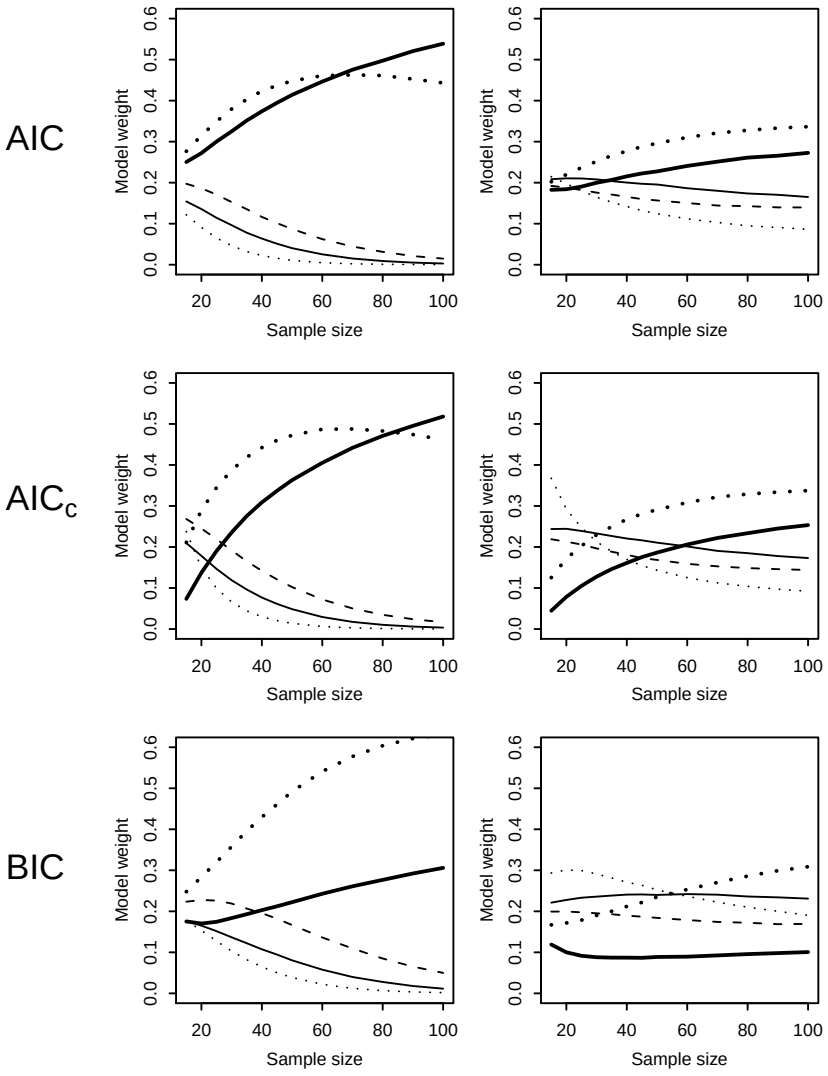

Figure 3.3: Mean AIC, AIC_c, and BIC model weights plotted versus sample size. Models are represented as M_1 (small dotted), M_2 (dashed), M_3 (thin line), M_4 (bold dotted), and M_5 (bold line). The first column corresponds to data generation under the largest model, M_5, while the second column uses random selection of the generating model.

model we would calculate a confidence interval based on that model. It involves model averaging of single-model Wald intervals, thereby avoiding the need to make any assumption about the distribution of $\hat{\theta}$, or to estimate its variance. As in the Bayesian approach to model averaging, it involves assuming that one of the candidate models is the true model. A MATA-Wald interval can be used whenever a simple Wald interval is likely to have good properties for each of the candidate models. As in the single-model setting, this may require a transformation of the parameter.

Theoretical study of the asymptotic properties of model-averaged confidence intervals is complicated by the random nature of the model weights. However, if the true model is among the set of candidate models, the weight for the true model will converge towards one as n increases, for each of the information criteria considered here (Claeskens and Hjort, 2008, p.99-102). Thus, the MATA-Wald interval will converge to a simple Wald interval based on the true model. Even if the true model is not among the candidate models, the MATA-Wald interval will converge to a simple Wald interval based on the model with minimum Kullback-Leibler distance to the generating model (Burnham and Anderson, 2002).

Our simulation study suggests that the MATA-Wald interval will perform better than the existing methods in terms of coverage, often with little or no increase in interval width (Figures 3.1 and 3.2). For both sets of simulations we have provided a benchmark in terms of coverage, by considering use of the true model. Clearly, this benchmark cannot be achieved in practice without knowledge of the true model. Even if this were possible, it may not be ideal. For example, in the first set of simulations the MATA-Wald interval provides a substantial reduction in interval width compared to the largest model, while incurring only a small reduction in coverage performance.

We recognize that simulation studies are limited in scope. However, our derivation

of the MATA-Wald interval is completely general with regard to the setting in which it is applied. We do not feel our assumption that "truth is in the model set" is a philosophical problem (see, for example, Link and Barker, 2006), nor is it likely to affect our comparison of the different methods.

It is well known that use of the "best model" (selected using an information criterion) can lead to poor performance in terms of coverage rate, in the presence of model uncertainty (Hurvich and Tsai, 1990; Fletcher and Dillingham, 2011). This was confirmed in our simulations (results not shown), where the best-model confidence intervals performed substantially worse in terms of coverage than model-averaged intervals. It is interesting to note that for construction of a MATA-Wald interval we estimate the c_i indicator variables in equation (3.1) using model weights, while use of a best-model interval corresponds to setting $\hat{c}_j = 1$ and $\hat{c}_{i \neq j} = 0$, where M_j is the best model.

Use of AIC weights was found to be preferable to AIC_c or BIC in terms of coverage rates, regardless of whether or not the largest model was truth. This was the case even for small samples, when one might have expected AIC_c to outperform AIC (Sugiura, 1978; Hurvich and Tsai, 1989). This finding agrees with those of Fletcher and Dillingham (2011), regarding the performance of existing model-averaged Wald interval constructions.

Chapter 4

Comparison of Bayesian and Frequentist Intervals

4.1 Introduction

Historically, parameter estimation has been based on a single model selected from the set of candidate models, with no allowance made for model uncertainty (Chatfield, 1995; Draper, 1995). Recently, model averaging has gained popularity as a technique to incorporate model uncertainty into the process of inference (Raftery *et al.*, 1997; Burnham and Anderson, 2002; Claeskens and Hjort, 2008).

Model averaging is a natural extension in the Bayesian paradigm, where the choice of model is introduced as a discrete-valued parameter. A prior probability mass function is specified for this parameter, defining the prior probability of each candidate model. Bayesian multimodel inference proceeds in a manner analogous to single-model inference. Posterior model probabilities are defined by the posterior distribution of the discrete-valued model parameter, and posterior distributions for model parameters

naturally account for model uncertainty (Hoeting *et al.*, 1999). In practice, Bayesian model averaging is achieved by allowing a Gibbs sampler to traverse this augmented parameter space, which generates approximations to the posterior distributions of interest.

In the frequentist setting, a model-averaged parameter estimate $\hat{\theta}$ is defined as the weighted sum of single-model estimates: $\hat{\theta} = \sum_{i=1}^{R} w_i \hat{\theta}_i$, where $\hat{\theta}_i$ is the parameter estimate under model M_i, and model weights w_i are determined from an information criterion such as AIC. Several different approaches to frequentist model-averaged confidence intervals have been suggested. Wald intervals of the form $\hat{\theta} \pm z_\alpha \, \hat{\text{var}}(\hat{\theta})^{1/2}$, where z_α is the $(1-\alpha)$ quantile of the standard normal distribution, rely on accurate estimation of $\text{var}(\hat{\theta})$. Estimation of this term is complicated by the fact that *both* the model weights and the single-model parameter estimates are random variables. Burnham and Anderson (2002) have suggested a variety of forms for $\hat{\text{var}}(\hat{\theta})$, which are studied in Claeskens and Hjort (2008, p.206-207) and in Turek and Fletcher (2012). In each of these studies, model-averaged Wald intervals of this form were found to perform poorly in terms of coverage rate.

An alternate type of model-averaged Wald interval, the model-averaged tail area Wald (MATA-Wald) interval, is proposed in Turek and Fletcher (2012). Here, each confidence limit is defined as the value for which a weighted sum of the resulting single-model Wald interval error rates is equal to the desired error rate. The model weights, w_i, are used to calculate this weighted sum. In a simulation study, this type of interval outperformed model-averaged intervals of the form $\hat{\theta} \pm z_\alpha \, \hat{\text{var}}(\hat{\theta})^{1/2}$ (Turek and Fletcher, 2012). This weighted-sum approach to model-averaged interval construction has also been applied to profile likelihood intervals, to produce a model-averaged tail area profile likelihood (MATA-PL) confidence interval (Fletcher and Turek, 2012).

In this chapter, we compare the performance of model-averaged Bayesian credible

intervals with the frequentist MATA intervals of Turek and Fletcher (2012) and Fletcher and Turek (2012). The effect of using various model prior probabilities and parameter prior distributions on Bayesian intervals is considered. We also study the use of several information criteria to calculate frequentist model weights. A theoretical study of the asymptotic properties of these intervals is complicated by the random nature of the model weights. For this reason, we assess the performance of these intervals through a simulation study.

In Sections 4.2 and 4.3, we define the Bayesian and frequentist model-averaged intervals. The differences between these intervals are highlighted in Section 4.4, through an example involving cloud seeding. We describe the simulation study used to compare these intervals in Section 4.5, and present the results of this study in Section 4.6. We conclude with a discussion in Section 4.7.

4.2 Bayesian Model Averaging

Assume a set of candidate models $\{M_i\}$, where the parameter of interest (θ) is common to all models. For data y, let model M_i have likelihood function $L_i(\theta, \lambda_i)$, parameterized in terms of θ and the nuisance parameter λ_i, which may be vector-valued. The Bayesian model-averaged posterior distribution for θ is

$$p(\theta|y) \;=\; \sum_{i=1}^{R} p(\theta|M_i, y)\; p(M_i|y), \tag{4.1}$$

where $p(\theta|M_i, y)$ is the posterior distribution of θ under model M_i, $p(M_i|y)$ is the posterior probability of M_i, and the summation is taken over the set of R candidate models (Hoeting *et al.*, 1999). An equal-tailed $(1-2\alpha)100\%$ model-averaged Bayesian (MAB) credible interval is defined as the α and $(1-\alpha)$ quantiles of $p(\theta|y)$.

Each posterior distribution $p(\theta|M_i, y)$ in equation (4.1) may be expressed through integration of the joint posterior, as

$$\begin{aligned} p(\theta|M_i, y) &= \int p(\theta, \lambda_i|M_i, y)\, d\lambda_i \\ &\propto \int L_i(\theta, \lambda_i)\, p(\theta, \lambda_i|M_i)\, d\lambda_i, \end{aligned} \tag{4.2}$$

following from Bayes' theorem, where $p(\theta, \lambda_i|M_i)$ is the joint prior distribution for parameters θ and λ_i under M_i. The posterior model probabilities in (4.1) may be expressed as $p(M_i|y) \propto p(y|M_i)\, p(M_i)$, where $p(M_i)$ is the prior probability of model M_i, and $p(y|M_i)$ is the integrated likelihood under M_i, given by

$$p(y|M_i) = \iint L_i(\theta, \lambda_i)\, p(\theta, \lambda_i|M_i)\, d\theta\, d\lambda_i. \tag{4.3}$$

Evaluation of the integrals in equations (4.2) and (4.3) is generally difficult in practice, and Markov chain Monte Carlo (MCMC) simulation is used to approximate the posterior distributions of interest. In the multimodel case, this is implemented using the reversible jump MCMC (RJMCMC) algorithm (Green, 1995).

4.3 Frequentist Model Averaging

The frequentist MATA confidence intervals are constructed in a manner analogous to Bayesian model averaging. Confidence limits are defined such that the weighted sum of error rates under each single-model interval will produce the desired overall error rate. This utilizes model weights w_i, which are derived from an information criterion.

We initially focus on the information criterion AIC $= -2 \log \hat{L} + 2p$ to define model weights, where $\hat{L}$ is the maximized likelihood and p is the number of parameters. Model

weights are calculated as $w_i \propto \exp(-\Delta\text{AIC}_i/2)$, where $\Delta\text{AIC}_i \equiv \text{AIC}_i - \min_j(\text{AIC}_j)$, and AIC_i is the value of the information criterion for model M_i (Buckland *et al.*, 1997). Other choices of information criteria for defining model weights are addressed in the discussion in Section 4.7.

4.3.1 MATA Wald

In the case of normal data, the limits θ_L and θ_U of a single-model $(1-2\alpha)100\%$ Wald interval satisfy the equations

$$\begin{aligned} 1 - F_\nu(t_L) &= \alpha \\ F_\nu(t_U) &= \alpha, \end{aligned}$$

where $F_\nu(\cdot)$ is the distribution function of the t-distribution with ν degrees of freedom, ν is the error degrees of freedom associated with the model, $t_L = (\hat{\theta} - \theta_L)/\hat{\text{var}}(\hat{\theta})^{1/2}$, $t_U = (\hat{\theta} - \theta_U)/\hat{\text{var}}(\hat{\theta})^{1/2}$, and $\hat{\text{var}}(\hat{\theta})$ is the estimated variance of $\hat{\theta}$. A MATA-Wald interval is constructed using a weighted sum of the single-model error rates. The lower and upper confidence limits of a MATA-Wald interval, θ_L and θ_U, are defined as the values satisfying

$$\begin{aligned} \sum_{i=1}^{R} w_i \, (1 - F_{\nu_i}(t_{L,i})) &= \alpha \\ \sum_{i=1}^{R} w_i \, F_{\nu_i}(t_{U,i}) &= \alpha, \end{aligned} \tag{4.4}$$

where each model M_i has ν_i error degrees of freedom, $t_{L,i} = (\hat{\theta}_i - \theta_L)/\hat{\text{var}}(\hat{\theta}_i)^{1/2}$, $t_{U,i} = (\hat{\theta}_i - \theta_U)/\hat{\text{var}}(\hat{\theta}_i)^{1/2}$, and $\hat{\theta}_i$ is the estimate of θ under model M_i.

The MATA-Wald interval may be generalized to non-normal data, assuming we can specify a transformation $\phi = g(\theta)$ for which the sampling distribution of $\hat{\phi}_i = g(\hat{\theta}_i)$

is approximately normal when M_i is true. For example, $\phi = \text{logit}(\theta)$ when θ is a probability. In this case, the MATA-Wald interval confidence limits θ_L and θ_U are the values satisfying the pair of equations

$$\begin{aligned} \sum_{i=1}^{R} w_i \, (1 - \Phi(z_{L,i})) &= \alpha \\ \sum_{i=1}^{R} w_i \, \Phi(z_{U,i}) &= \alpha, \end{aligned} \tag{4.5}$$

where $\Phi(\cdot)$ is the standard normal distribution function, $z_{L,i} = (\hat{\phi}_i - \phi_L)/\hat{\text{var}}(\hat{\phi}_i)^{1/2}$, $z_{U,i} = (\hat{\phi}_i - \phi_U)/\hat{\text{var}}(\hat{\phi}_i)^{1/2}$, $\phi_L = g(\theta_L)$, and $\phi_U = g(\theta_U)$, as set out in Turek and Fletcher (2012).

4.3.2 MATA Profile Likelihood

Assuming a single model with likelihood function $L(\theta, \lambda)$, the limits θ_L and θ_U of the $(1 - 2\alpha)100\%$ profile likelihood interval for θ satisfy the equations

$$\begin{aligned} \Phi\left(r(\theta_L)\right) &= \alpha \\ 1 - \Phi\left(r(\theta_U)\right) &= \alpha, \end{aligned}$$

where $r(\theta)$ is the signed likelihood ratio statistic, defined as

$$r(\theta) = \text{sign}(\hat{\theta} - \theta) \sqrt{2 \left(\log L_p(\hat{\theta}) - \log L_p(\theta) \right)}, \tag{4.6}$$

and $L_p(\theta) = \max_{\lambda} L(\theta, \lambda)$ is the profile likelihood function for θ (Davison, 2003, p.126-129). The lower and upper confidence limits, θ_L and θ_U, of the MATA-PL interval are

defined as the values which satisfy

$$\begin{aligned} \sum_{i=1}^{R} w_i \, \Phi\left(r_i(\theta_L)\right) &= \alpha \\ \sum_{i=1}^{R} w_i \, \left(1 - \Phi\left(r_i(\theta_U)\right)\right) &= \alpha, \end{aligned} \tag{4.7}$$

where $r_i(\theta)$ is defined in terms of the corresponding likelihood function $L_i(\theta, \lambda_i)$, as in equation (4.6), and as described in Fletcher and Turek (2012).

4.4 Example: Cloud Seeding

We use a study of cloud seeding to illustrate the differences between these methods of model averaging. There is clear evidence that seeding clouds causes an increase in the mean volume of rainfall (Simpson *et al.*, 1971; Simpson, 1972; Rosenfeld and Woodley, 1993). However, the size of this effect may depend on the pattern of motion of the clouds. As rainfall volume has agricultural impacts, the results may affect the practicality and focus of cloud seeding operations. The data we consider come from testing conducted by the Experimental Meteorology Laboratory in Florida, USA. Total rainfall volume was measured for 27 stationary clouds, 16 of which were seeded and 11 of which were unseeded. The full data set appears in Biondini (1976), and the subset relevant to our analysis is in Table 4.1.

Suppose we aim to predict the expected rainfall from seeded, stationary clouds. The lognormal distribution can provide a good model for total rain volume (Biondini, 1976). Denote the volume of rainfall from seeded, stationary clouds as R_S, where $\log R_S \sim N(\beta_S, \sigma^2)$, and the volume of rainfall resulting from unseeded, stationary clouds as R_U, where $\log R_U \sim N(\beta_U, \sigma^2)$. Let the quantity of interest be the expected rain volume resulting from the seeded clouds, $\theta_S \equiv \mathrm{E}[R_S] = \exp(\beta_S + \frac{\sigma^2}{2})$, and we

Table 4.1: Rain volume data for use in the cloud seeding example, recorded by the Experimental Meteorology Laboratory in Florida, USA, in 1968 and 1970. All clouds are stationary, and are categorized as seeded or unseeded. Rain volume is measured in thousands of cubic meters ($10^3\,\mathrm{m}^3$).

Seeded clouds Rain Volume	Unseeded clouds Rain Volume
160.32	32.29
38.84	32.53
3396.34	397.33
605.02	1026.84
147.21	427.38
248.27	1487.62
339.80	45.28
339.80	6.06
1209.79	6.06
245.66	201.63
870.11	26.84
146.34	
315.44	
142.63	
40.46	
50.23	

consider the following two models:

M_1: $\beta_S = \beta_U$
M_2: β_S and β_U unspecified

In the Bayesian analyses, we used a vague $N(0, \sigma^2{=}100^2)$ prior distribution for parameters β_S and β_U, a uniform prior distribution on the interval $(0, 100)$ for σ (Gelman, 2006), and an equal prior probability for each model. We ran an MCMC algorithm for 300,000 iterations, with a 5% burn-in period. Convergence was assessed using the Brooks-Gelman-Rubin (BGR) diagnostic on two parallel chains (Gelman and Rubin, 1992; Brooks and Gelman, 1998). This indicated convergence for each model, with all BGR values less than 1.008.

Frequentist models were fit using maximum likelihood. Since we are interested in prediction of θ_S, each likelihood function was reparameterized using $\log \theta_S - \frac{\sigma^2}{2}$ in place of β_S, and $\log \theta_U - \frac{\sigma^2}{2}$ in place of β_U. The MATA-Wald interval was constructed using equation (4.5) and the MATA-PL interval using (4.7), both of which used AIC weights for w_i.

The resulting Bayesian posterior model probabilities were $p(M_1|y) = p(M_2|y) = 0.50$, which were equal to the model prior probabilities to two decimal places. The AIC weights slightly favored M_2, with $w_1 = 0.38$ and $w_2 = 0.62$. Figure 4.1 shows the predicted mean rain volume $\hat{\theta}_S$ from seeded, stationary clouds, with 95% confidence intervals. Predictions and confidence intervals are shown for single-model inferences under M_1 and M_2, as well as using model averaging.

The Bayesian posterior mean and the maximum likelihood estimate for predicted rainfall are reasonably similar, with the Bayesian estimate being approximately 15% higher under each model. As expected, all estimates under M_2 (where seeding may

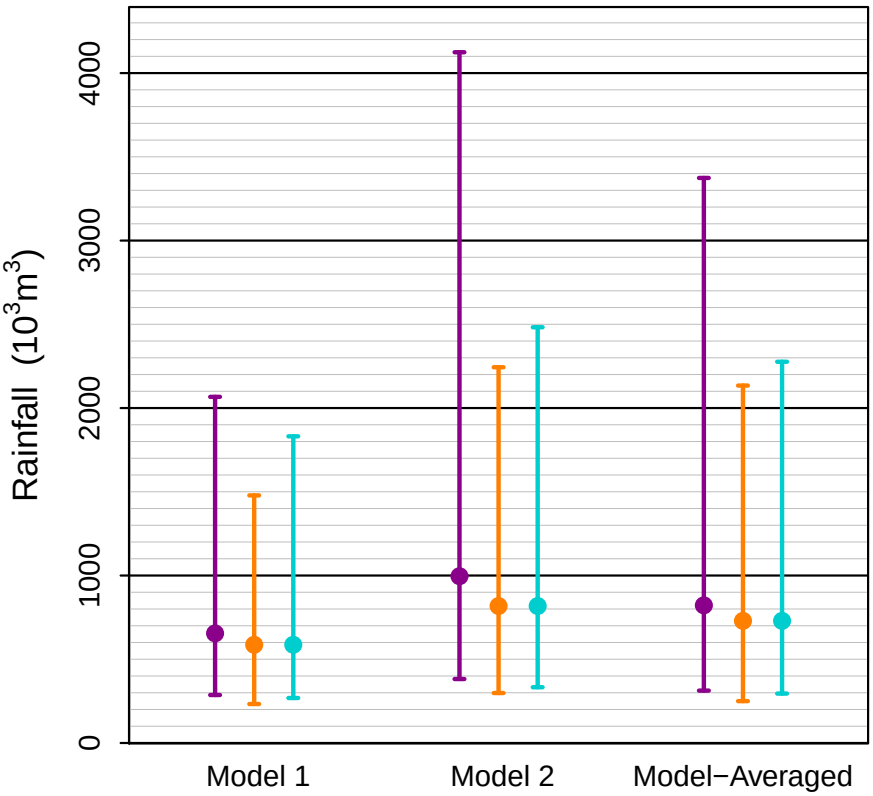

Figure 4.1: Expected mean rainfall for seeded, stationary clouds, under each model and using model averaging. Vertical bars show 95% intervals for each prediction. Intervals shown: Bayesian and MAB (purple), Wald and MATA-Wald (orange), profile likelihood and MATA-PL (blue).

cause increased rainfall) are greater than those under M_1.

The differences between methods are highlighted by confidence intervals for the expected rainfall. All lower limits are reasonably similar, while the upper limits from the Bayesian analyses are significantly higher than those from the frequentist analyses. This is particularly true under M_2 and also when model averaging, where the MAB interval is 62% wider than the MATA-Wald interval. The MAB interval produces a visually appealing compromise between the single-model Bayesian intervals, especially when considering the high degree of model uncertainty.

Each profile likelihood interval is slightly more asymmetric than the corresponding Wald interval, as one would expect. The frequentist MATA-Wald and MATA-PL intervals again produce a pleasing compromise between the separate inferences under each model. In light of the model uncertainty present, it would seem appropriate to use one of the model-averaged intervals to summarize the results of this analysis.

4.5 Simulation Study

Based on the example in Section 4.4, we considered a two-sample setting for the simulation study, using both normal and lognormal data. Observations were generated as either $Y_{ij} \sim N(\beta_i, \sigma^2)$, or $\log Y_{ij} \sim N(\beta_i, \sigma^2)$, for $i = 1, 2$ and $j = 1, \ldots, n$. We fixed $\beta_1 = 0$, $\beta_2 = 1$, and $\sigma^2 = 1$, and varied the sample size n between 10 and 100. We focused on prediction of $\theta_i \equiv \mathrm{E}[Y_{ij}]$, for $i = 1, 2$. In the lognormal case, $\theta_i = \exp(\beta_i + \frac{\sigma^2}{2})$, so the likelihood was again reparameterized using $\log \theta_i - \frac{\sigma^2}{2}$ in place of β_i. The two models considered were:

M_1: $\beta_1 = \beta_2$

M_2: β_1 and β_2 unspecified

The performance of each method was measured by the lower and upper error rates, defined as the proportion of simulations for which $\theta_L > \theta$ or $\theta_U < \theta$. We averaged results over 20,000 simulations, ensuring a standard error for the error rates less than 0.3%. In addition, we calculated the mean lower and upper relative interval half-widths, defined as $\frac{\theta-\theta_L}{\theta}$ and $\frac{\theta_U-\theta}{\theta}$, respectively. All calculations were performed in R, version 2.13.0 (2011).

4.5.1 Bayesian Intervals

Three sets of prior probabilities were considered, for the construction of three distinct model-averaged Bayesian intervals. The standard model-averaged Bayesian (MAB) interval used equal prior probabilities for each model, and "flat" prior distributions for the parameters: $\beta_i \sim N(0, \sigma^2{=}100^2)$, and $\sigma \sim \text{Uniform}(0, 100)$, as suggested in Gelman (2006). The Jeffreys model-averaged Bayesian (MAB_J) interval used equal model prior probabilities, and improper Jeffreys prior distributions (Jeffreys, 1946) for the parameters: $p(\beta_i) \propto 1$, and $p(\sigma) \propto \frac{1}{\sigma}$ (see, for example, Box and Tiao, 1973). The Kullback-Leibler model-averaged Bayesian (MAB_KL) interval used flat prior distributions for the parameters, and the Kullback-Leibler (KL) prior probability for each model, defined as:

$$p(M_i) \propto \exp\left(p_i\left(\tfrac{1}{2}\log n - 1\right)\right),$$

where p_i is the number of parameters in model M_i (Burnham and Anderson, 2002, p.302-305). The KL model prior is a Bayesian counterpart to frequentist AIC model weights, being designed to produce posterior model probabilities asymptotically equal to AIC model weights.

A Gibbs sampler was implemented in R, using the RJMCMC algorithm. Convergence of two parallel chains was again assessed using the BGR convergence diagnostic.

Simulations which failed to converge after 100,000 iterations (BGR > 1.1) were discarded. In total, 99.7% of the simulations were retained, with a maximum BGR value of 1.099, and a mean BGR value of 1.007. The initial 5% of each simulation was discarded as the burn-in period.

4.5.2 Frequentist Intervals

Frequentist model-averaged intervals were constructed using AIC weights. For the normal linear simulation, the MATA-Wald interval was calculated using equation (4.4), and the lognormal MATA-Wald interval using (4.5). The MATA-PL interval was defined according to (4.7), using the reparameterized likelihood in the lognormal case. Numerical solutions to these equations were found using the R command *uniroot*.

4.6 Simulation Results

In the normal linear setting, the results for θ_1 and θ_2 are identical by symmetry. In addition, in the lognormal setting the results were qualitatively similar for θ_1 and θ_2. Therefore, for simplicity we focus on the results for θ_2. Figure 4.2 (top) shows the estimated lower and upper error rates for the MATA-Wald, MATA-PL, and MAB intervals. The MATA-Wald interval performs best on the upper error rate, in particular for small sample sizes, followed by the MATA-PL and MAB intervals. All intervals asymptotically approach the nominal rates. We would expect the MATA-Wald interval to perform well, since M_2 is the generating model, and the Wald interval based on this model will achieve exact nominal coverage in this setting. To observe the trade-off between coverage rate and interval width, error rates are also plotted against the corresponding half-widths. The MATA-Wald interval requires only a small increase in

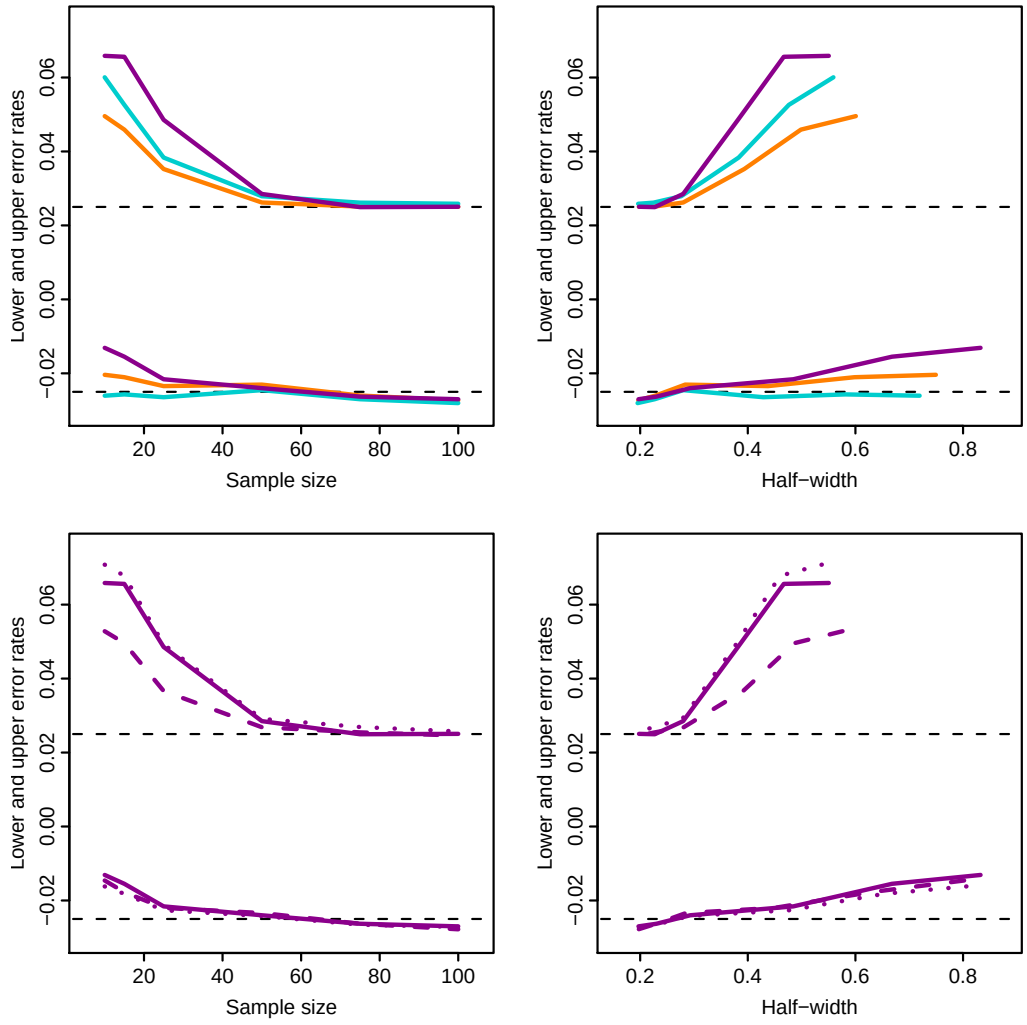

Figure 4.2: Confidence interval performance for prediction of the mean, θ_2, in the normal linear simulation. Top row: MAB (purple), MATA-Wald (orange), and MATA-PL (blue) intervals. Bottom row: MAB (solid), MAB_{J} (dotted), and MAB_{KL} (dashed) Bayesian intervals. Nominal error rates are shown as dashed lines, and lower error rates are plotted as negative values.

upper half-width to achieve notably improved performance.

Figure 4.2 (bottom) provides the same comparison for the Bayesian MAB, MAB_{J}, and MAB_{KL} intervals. The MAB_{KL} interval provides a noticeable improvement in performance on the upper limit, as compared to the MAB and MAB_{J} intervals, each of which uses equal model prior probabilities. Use of the KL prior probability for models in the MAB_{KL} interval provides an improvement of almost 2% in the upper error rate, for small sample sizes. This improvement comes at the expense of a negligible increase in the upper half-width. In addition, the use of Jeffreys prior distributions for the parameters slightly degrades the performance of the Bayesian interval, relative to the use of flat prior distributions.

Figure 4.3 provides analogous comparisons in the lognormal setting. Here, the MAB interval outperforms the frequentist MATA intervals in the upper error rate, although this comes at the cost of a substantial increase in upper half-width. The MAB interval remains within 1.5% of the nominal error rate for all sample sizes, while the frequentist MATA intervals deviate by as much as 4%. The MATA-PL interval performs slightly better than the MATA-Wald interval, which performs poorly overall, as might be expected of any Wald interval in this setting.

Comparison of the Bayesian intervals in the lognormal setting is qualitatively similar to that of the normal linear setting. The use of the KL prior probability for models in the MAB_{KL} interval provides a substantial improvement over the use of equal prior probabilities, and here the use of Jeffreys prior distributions for the parameters severely degrades performance, relative to the use of flat prior distributions. Overall, the Bayesian interval using KL model prior probabilities outperforms all other model-averaged interval constructions in the lognormal setting.

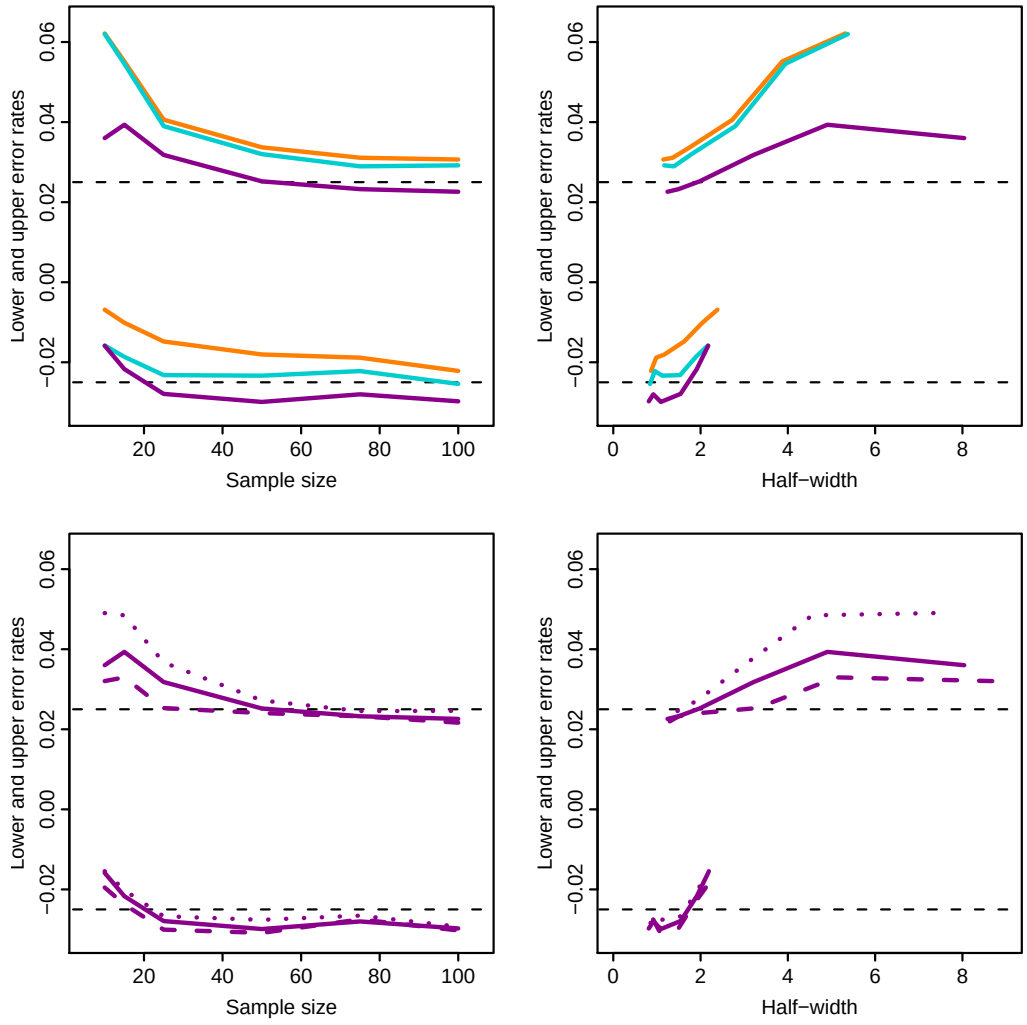

Figure 4.3: Confidence interval performance for prediction of the mean, θ_2, in the lognormal simulation. Top row: MAB (purple), MATA-Wald (orange), and MATA-PL (blue) intervals. Bottom row: MAB (solid), MAB_J (dotted), and MAB_KL (dashed) Bayesian intervals. Nominal error rates are shown as dashed lines, and lower error rates are plotted as negative values.

4.7 Discussion

The aim of this chapter has been to compare the performance of Bayesian and frequentist model-averaged confidence intervals. The frequentist MATA intervals are based upon model averaging the *error rates* of single-model intervals, rather than basing an interval around a model-averaged parameter estimate. This construction is analogous to Bayesian model averaging, and in fact the idea was initially motivated by analogy to a model-averaged Bayesian interval (Fletcher and Turek, 2012). The same frequentist construction was studied further in Turek and Fletcher (2012), where it is shown that asymptotically, this model-averaged interval will converge to the single-model interval based upon the candidate model with minimum Kullback-Leibler distance to the true, generating model.

Through simulation, the frequentist MATA-Wald interval produced the best coverage properties in the normal linear setting, where we would expect Wald intervals to perform well. In the lognormal setting, Bayesian intervals produced substantial improvement over the frequentist intervals. A Bayesian analysis allows fully for parameter uncertainty, and does not rely on the asymptotic distributions of parameter estimates. So long as we are willing to accept the prior distributions for the parameters, we might expect the Bayesian approach to be better suited for non-normal settings. In contrast, when the assumptions of Wald intervals are satisfied exactly (as with normal data), use of the frequentist MATA-Wald interval resulted in improved coverage performance.

In both settings, the use of KL prior probabilities provided a noteworthy improvement in the performance of the Bayesian interval, when compared to the use of equal model prior probabilities. The KL model prior is designed to produce posterior model probabilities approximately equal to frequentist AIC model weights. This agreement between posterior probabilities and model weights was observed in our simulation.

Burnham and Anderson (2002) describe prior probabilities which depend upon sample size and model complexity, such as the KL prior, as "savvy priors," and argue in favor of their use. Larger data sets have the potential to support more complex models, which may justify assigning model prior probabilities dependent upon the data available and the relative complexity of the models being considered.

In contrast, Link and Barker (2006) argue that for large sample sizes the data ought to completely dominate the priors, and the use of prior probabilities which depend upon the sample size may prevent this from occurring. They also argue that prior probabilities should represent one's beliefs *prior to data collection*, and have no dependence upon the data observed. This is consistent with Box and Tiao (1973), where a prior distribution is defined as "what is known about [a parameter] without knowledge of the data." This discrepancy in what a prior probability may represent is interesting, especially considering that data-dependent priors were seen to be advantageous for Bayesian model averaging.

Thus far, we have presented results for frequentist model-averaged intervals constructed using AIC model weights. Two alternate information criteria were also considered: AIC_c (Sugiura, 1978) and BIC (Schwarz, 1978). AIC_c was derived as a small-sample correction to AIC, and in certain contexts may be favorable for use in model selection (Hurvich and Tsai, 1989). BIC provides an asymptotic approximation to Bayes factors, and may also be used to approximate the posterior model probabilities which result from equal model priors (Link and Barker, 2006).

In our study, the frequentist intervals based upon AIC_c and BIC weights were consistently inferior to those using AIC weights. This was true in both simulation settings, and also for small sample sizes, when one may have expected AIC_c to perform best. The results for AIC_c weights and BIC weights appear in Appendices B.2 and B.3. This finding is consistent with the results of Fletcher and Dillingham (2011), in which

model-averaged intervals constructed using AIC weights yielded improved coverage properties over a variety of other information criteria, including both AIC_c and BIC.

Our study has used the assumption that "truth is in the model set." This same assumption is used in the derivations of both the MATA-Wald and MATA-PL intervals, as well as generally in Bayesian multimodel inference. We do not feel that this assumption undermines our conclusions, since all model averaging techniques would be adversely affected when this assumption is not met.

Our simulation has also followed the assumption that "the largest model is truth." Philosophically this may not pose a problem, as Burnham and Anderson (2002) believe that nature is arbitrarily complex, and it is unrealistic to assume we might fully characterize the underlying process. From this viewpoint, model selection attempts to identify the most parsimonious approximating model to truth, given the finite nature of the observed data. This assumption may in part explain the superior performance of AIC model weights, since AIC is known to favor increased model complexity (Kass and Raftery, 1995). However we don't consider this an issue, since results from Fletcher and Turek (2012) indicate that intervals using AIC weights perform at least as well as those using other information criteria when the most complex model is *not* the generating model.

Any simulation study is inherently limited in scope. Herein, we have considered both normal and non-normal data, as well as a wide range of sample sizes, and observed consistent patterns throughout. Bayesian model averaging was better suited for the non-normal setting, and the frequentist MATA-Wald interval performed best in the normal linear setting. In addition, the performance of model-averaged Bayesian intervals was improved through use of the KL model prior, a data-dependent prior probability.

This result raises consideration of exactly what model prior probabilities represent; in particular, whether or not knowledge of the *size* of an observed sample provides grounds to update model prior probabilities. This question warrants additional study, and future research could investigate situations where the true model is not contained in the candidate set, or the true model is nested within the candidate set.

Chapter 5

Discussion and Conclusions

5.1 Findings

Through simulation studies, a variety of constructions for model-averaged confidence intervals were compared. These included both Bayesian and frequentist constructions, with particular attention given to the new MATA construction for frequentist model-averaged intervals. All confidence intervals were assessed in terms of coverage rate and interval width, each of which was decomposed, respectively, into upper and lower error rates, and upper and lower half-widths. A variety of patterns were observed between the relative performances of these model-averaged interval constructions.

5.1.1 Normal Linear Model

In the normal linear regression setting, the MATA-Wald interval was compared against the preexisting MAW methodology for the construction of model-averaged intervals. By nature, MAW intervals are Wald confidence intervals centered around a model-averaged

point estimate. We might suppose each of these constructions would perform well in the normal linear setting, since in the absence of model uncertainty, a Wald confidence interval will exactly achieve the nominal coverage rate (Davison, 2003, p.370-373). Any reduction we observe in the achieved coverage rate is a direct consequence of model uncertainty. For each sample size considered, the MATA-Wald interval achieved a coverage rate nearer to the nominal value of 95% than either of the MAW constructions. In addition, the MATA-Wald interval was not significantly wider than the MAW intervals; the improved coverage rate did *not* come at the expense of a wider interval (Turek and Fletcher, 2012). This would suggest that in the normal linear setting, the MATA-Wald interval provides a fundamental improvement upon the existing techniques for model-averaged interval construction.

These results for the MATA-Wald and MAW intervals, as presented in Chapter 3, were for one particular point in the covariate space. This point was selected as a particular quantile of the covariate distribution. The same assessment of confidence intervals was also performed at eight other points, which were defined using a wide range of quantiles to adequately "explore" the covariate space. Similar results were observed in each case: the MATA-Wald interval consistently outperformed the MAW intervals in terms of coverage rate, without incurring a substantial increase in interval width. For brevity, these results have not been shown, as they are qualitatively identical to those presented in Chapter 3.

Our study of the normal linear setting continued in Chapter 4, using simulation to compare MATA intervals to MAB credible intervals. Again, we might expect the MATA-Wald interval to perform well in this context. The MATA-Wald interval performed more favorably than the MATA-PL interval, and also more favorably than the standard MAB interval constructed using equal model prior probabilities. However, when using the data-dependent KL prior probability for models, the performance of

the resulting MAB_{KL} interval was nearly identical to that of the MATA-Wald interval constructed using AIC model weights.

The close agreement between the MATA-Wald interval and the Bayesian MAB_{KL} interval is interesting, since KL prior probabilities are constructed to produce Bayesian posterior model probabilities that are asymptotically equal to frequentist AIC weights (Burnham and Anderson, 2002, p.302-305). The resulting model weights and posterior probabilities are shown in Figure 5.1. We observe the expected agreement between the frequentist and the Bayesian results, in particular for large sample sizes. In effect, the frequentist MATA-Wald interval is equivalent to the MAB_{KL} interval when the Bayesian model priors are chosen appropriately to "agree" with AIC model weights. In that case, the resulting model-averaged intervals are nearly identical. Furthermore, the frequentist MATA-Wald interval using AIC weights and the MAB_{KL} interval yielded the best coverage performance, among all model-averaged intervals considered in the normal linear setting.

5.1.2 Non-Normal Models

The various constructions of model-averaged confidence intervals were also compared in the context of non-normal data distributions. In Chapter 2, a two-sample lognormal simulation was used to compare the MATA-PL interval to the previously existing MAW construction of model-averaged intervals. We independently varied both the sample size, and the difference between the means of the underlying distributions, to provide an understanding of interval performance in different situations. For all values of the simulation parameters considered, the MATA-PL interval produced error rates and an overall coverage rate nearest to the nominal levels (Fletcher and Turek, 2012). This result would suggest that the MATA-PL interval provides improved coverage rates over the preexisting MAW interval construction, in the case of non-normal data.

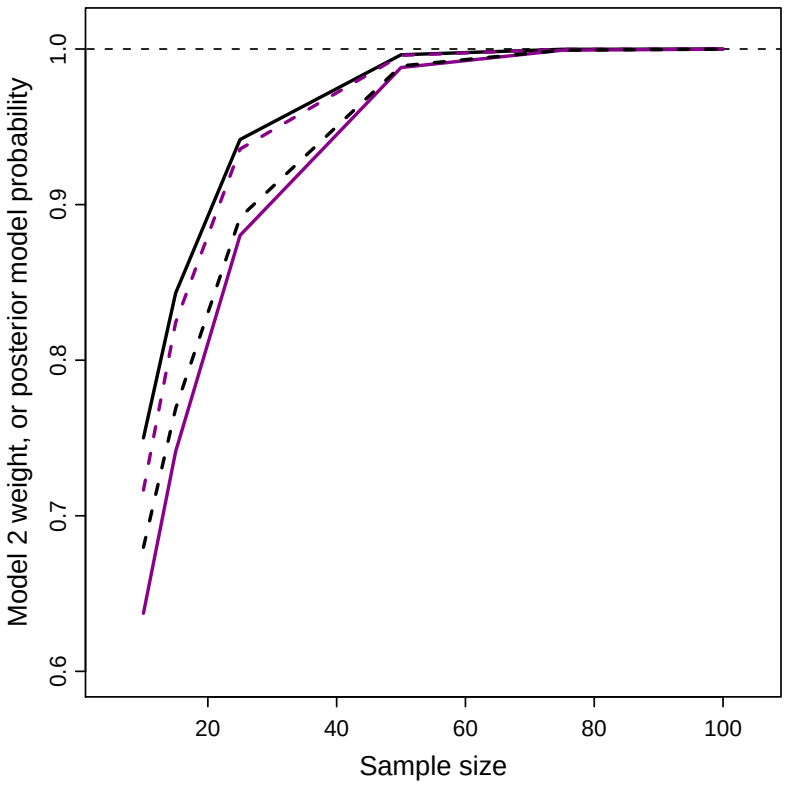

Figure 5.1: Mean frequentist model weights and Bayesian posterior model probabilities plotted versus sample size, from the normal linear simulation study described in Section 4.5. Weights are for Model 2, the generating model. Frequentist model weights: AIC (black solid) and BIC (black dashed). Bayesian posterior model probabilities resulting from: equal model priors (purple solid) and KL model priors (purple dashed).

As we generally expect, the performance of all model-averaged intervals deteriorated most severely when model uncertainty was highest. This occurred when the AIC model weights were approximately equal for each of the two models, and resulted from specific values of the difference between population means, for each sample size considered. At these "worst case" points, which exhibited the largest model uncertainty, all model-averaged intervals experienced degraded performance. However, the MATA-PL interval still performed nearest to the nominal coverage rate.

Analysis of the lognormal setting continued in Chapter 4, where the MATA-Wald and MATA-PL intervals were compared against model-averaged Bayesian credible intervals. In this non-normal setting, we may not expect the MATA-Wald interval to perform as well as in the normal linear case, since it relies upon the asymptotically normal sampling distribution of single-model estimators, which will only be accurate in large samples (Davison, 2003, p.118-125). Similarly, the MATA-PL interval relies on the asymptotic distribution of the likelihood ratio test statistic (Davison, 2003, p.126-131), and hence also may not perform well.

The standard MAB interval was seen to outperform both MATA interval constructions, for each sample size considered. The Bayesian approach fully allows for uncertainty in model parameters, and also uncertainty in the model selection process (Raftery *et al.*, 1997; Hoeting *et al.*, 1999). All three MAB intervals were noticeably wider than the MATA intervals for all values of the sample size, but this increased interval width was an accurate representation of the total (unconditional) uncertainty in parameter estimates. In this non-normal setting, model-averaged Bayesian intervals universally produced the best coverage rate performance, and in fact were quite near to the nominal level, even for small sample sizes. This result suggests that the use of model-averaged Bayesian intervals is the best choice, in the presence of non-normal data distributions. However, it is also worth noting that the construction of

Bayesian intervals required a substantial increase in computational time, relative to the frequentist methods.

5.1.3 Frequentist Model Weights

The construction of a frequentist model-averaged confidence interval requires a set of model weights, typically derived from an information criterion (Buckland *et al.*, 1997). Traditional MAW interval constructions involve weighting the single-model estimators and associated variances (Burnham and Anderson, 2002), while the MATA interval construction employs a weighting of single-model confidence interval error rates (Turek and Fletcher, 2012). The set of model weights is calculated using an information criterion, for which there exists a wide variety of possible constructions. The choice of which information criterion to use is determined by a researcher's assumptions and beliefs, and depending on the specific analysis at hand, any particular information criterion may be justified. Further, the use of different information criteria will generally result in different model weights, and hence result in different confidence limits for model-averaged intervals.

In each simulation study presented, the use of AIC, BIC, and AIC_c were all considered for the construction of frequentist model weights. This produced three versions of each frequentist model-averaged interval, corresponding to the three information criteria. These three constructions resulted in different error rates and half-widths for each interval. Across all simulations, the use of AIC weights for the construction of model-averaged intervals produced error rates rates nearest to the nominal values. AIC weights also typically resulted in wider intervals.

Consider again the lognormal simulation study in Chapter 2. Results were presented for AIC weights, since these produced the best coverage properties for each particular

interval. The full results from this simulation include interval performances resulting from the use of AIC, BIC, and AIC_c weights. Figure 5.2 presents these results for the MATA-PL interval, showing the error rate and half-width performance of this interval, when constructed using AIC, BIC, and AIC_c weights.

For each sample size and value of β_2, AIC weights resulted in error rates nearest to the nominal value of 2.5%. In large samples ($n = 50$), the results for AIC and AIC_c weights are nearly identical; however for small samples ($n = 10$), AIC_c weights perform noticeably poorer, and in fact, nearly identical to BIC weights. In addition, the use of AIC weights produced only slightly wider intervals. Qualitatively similar results for relative performances of AIC, BIC, and AIC_c weights were observed in all simulation studies.

In each simulation, the use of AIC weights outperformed AIC_c weights even in the case of small samples. This may be a surprise, since AIC_c is designed as a small sample correction to AIC (Hurvich and Tsai, 1989). However, in the context of model-averaging, the use of AIC weights resulted in improved coverage properties of model-averaged intervals. This finding is consistent with the results of Fletcher and Dillingham (2011), regarding the relative performances of AIC, BIC, and AIC_c weights. These results suggest that the use of AIC weights produces model-averaged intervals with the best coverage properties, among the information criteria studied. However, implicit in these results is the assumption that the largest model is truth. The use of this assumption may bias simulation results to favor AIC, since AIC is known to favor model complexity more so than other information criteria (Sugiura, 1978). We further discuss this assumption of the largest model being truth in Section 5.2.

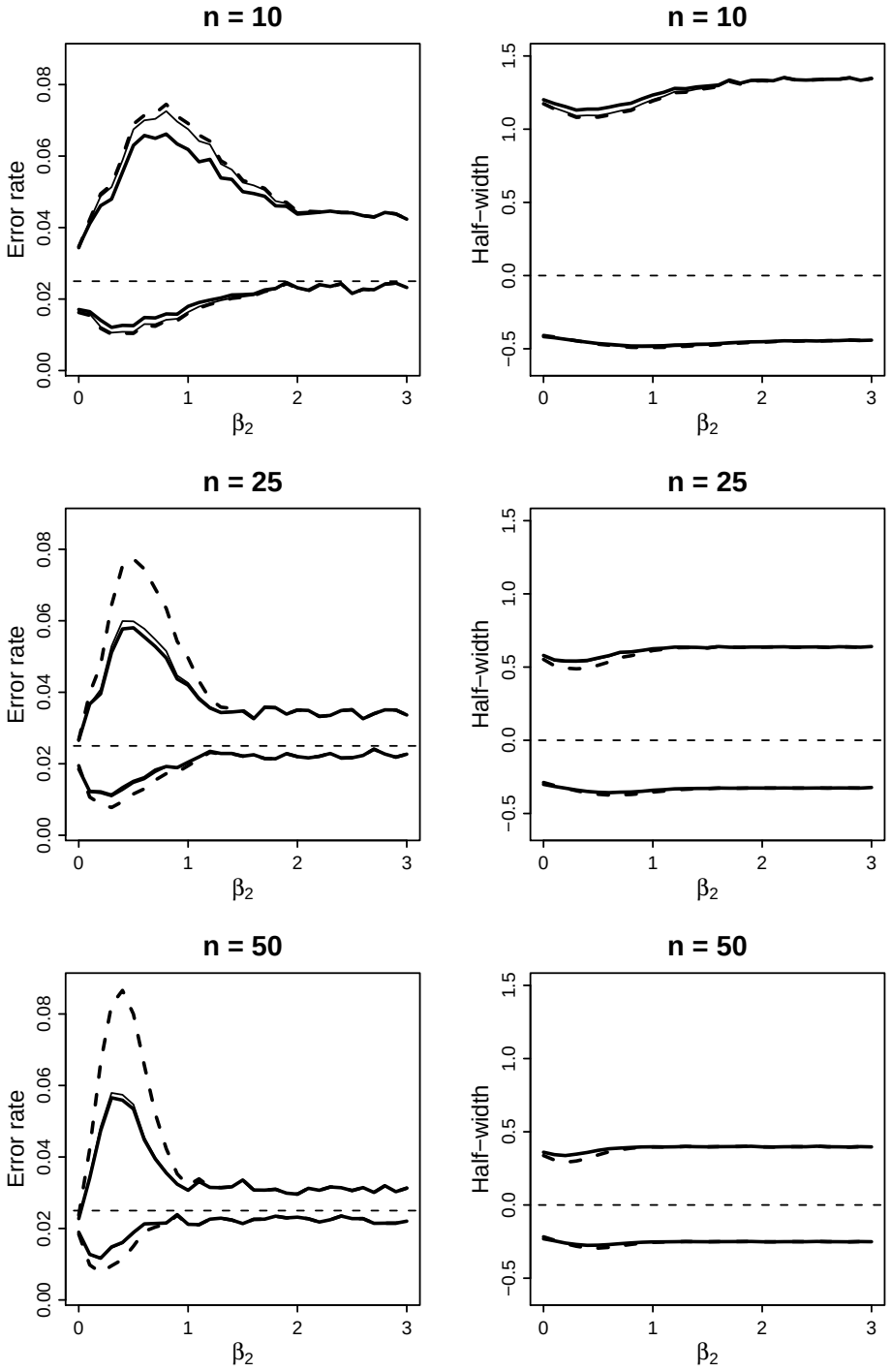

Figure 5.2: Error rates (left) and mean relative half-widths (right) showing the lognormal simulation results, as described in Section 2.5. Results are for the MATA-PL interval constructed using AIC weights (thick solid), BIC weights (dashed), and AIC_c weights (thin solid). The dashed reference lines are the nominal 2.5% error rate (left) and zero (right).

5.1.4 Prior Distributions in Bayesian Model Averaging

A broad comparison between various methods for constructing model-averaged confidence intervals was presented in Chapter 4. This included the frequentist MATA intervals developed herein, and also Bayesian model-averaged credible intervals. In any Bayesian analysis, prior distributions for the model parameters must be specified. In the context of Bayesian model averaging, a discrete prior distribution defined on the set of candidate models must also be specified (Hoeting *et al.*, 1999). Three distinct sets of Bayesian parameter and model prior distributions were studied, to provide an understanding of their effects on Bayesian model-averaged interval performance. These three sets of prior distributions were:

Standard priors: Using a vague, flat prior distribution for each model parameter, and assigning an equal prior probability to each candidate model.

Jeffreys priors: Using the transformation-invariant Jeffreys prior distribution for each model parameter, and assigning an equal prior probability to each candidate model.

Kullback-Leibler priors: Using a vague, flat prior distribution for each model parameter, and assigning the data-dependent Kullback-Leibler prior probability to each candidate model.

In each simulation study presented in Chapter 4, a consistent pattern was observed in the relative performances of model-averaged Bayesian intervals constructed using these three sets of prior distributions. We consider the use of standard priors to construct the MAB interval as a benchmark for performance, since this selection may be thought of as "uninformative" (Gelman, 2006; Martín and Pérez, 2009).

Relative to the use of standard priors, use of the Jeffreys prior distribution for parameters (Jeffreys, 1946) consistently resulted in error rates further from the nominal level. When using Jeffreys prior distributions, the coverage rates of the MAB_J interval were degraded in both the normal and non-normal simulation settings. This effect was particularly noticeable in the lognormal setting, and for small sample sizes. The use of Jeffreys prior distribution for parameters does not appear to be beneficial in the context of Bayesian model averaging.

Perhaps more interesting, the use of the KL prior probability for each model yielded a consistent improvement in the error rates of the MAB_{KL} interval. This improvement was relative to the use of standard priors, which included equal model prior probabilities. The effect was noticeable in both the normal linear and the lognormal simulation settings. Model-averaged Bayesian intervals constructed using KL prior probabilities yielded coverage rates nearer to the nominal level by 1% or more, across a wide range of sample sizes of interest.

The KL prior probability is suggested by Burnham and Anderson (2002, p.302-305), and is defined to generate posterior model probabilities asymptotically equal to AIC model weights. The derivation of the KL prior is based upon the fact that BIC model weights asymptotically approximate the model posterior probabilities resulting from equal model prior probabilities (Link and Barker, 2006). These relationships were observed through simulation, as seen in Figure 5.1.

In the normal linear setting, the performance of the MATA-Wald interval constructed using AIC weights was nearly identical to that of the MAB_{KL} interval constructed using KL prior probabilities. The methodology of the MATA interval can be motivated by analogy with model-averaged Bayesian intervals (Fletcher and Turek, 2012), and through appropriate choice of priors these intervals result in asymptotically equal model weights and model posterior probabilities. This suggests an equivalence

between the frequentist and Bayesian methods, in the normal linear setting. However, this equivalence was not observed in the lognormal setting, in which the MATA interval construction relies upon the asymptotic distributions of single-model estimators.

The use of KL prior probabilities requires defining one's "prior" beliefs based upon the size of the observed data set. This may seem to contradict the underlying intent of a Bayesian prior, as representing one's beliefs *prior* to data collection (*e.g.,* King *et al.*, 2009, p.7). On the other hand, it could be defensible to allot additional "prior belief" to more complex models when a larger data set is observed, since larger data sets provide the capacity to support increased model complexity (Buckland *et al.*, 1997). The KL prior does exactly that: assigns increasingly large prior probabilities to models of higher complexity, as the sample size increases. Philosophies aside, the use of KL prior probabilities for models was observed to be beneficial for the construction of Bayesian model-averaged credible intervals.

5.2 Assumptions

Throughout, we have made the assumption that the largest model in the candidate set represents truth. Accordingly, in the simulation studies presented, data has been generated under the full (or largest) model under consideration. Use of this assumption does not undermine the applicability of the results, when considering complex, natural systems. In nature, the "true" data-generating mechanism would typically be a highly complex, infinite-dimensional system, which itself may be impossible to fully specify. It could involve individual heterogeneity, as well as depend upon any number of unobservable covariates. If the "true" model could be fully specified, it is certain to be more complex than any of the models in the candidate set. Model selection merely seeks to identify the best approximating model, from within the candidate set. The

process of model-averaging accounts for uncertainty in the question of *which* candidate model provides the best approximation to the conceptual "true" model (Burnham and Anderson, 2002, p.20-23, 437-439).

Generally in the simulations presented, data has been generated under the largest model. However, due to the limited nature of an observed sample, the best approximating model is typically identified to be one of the simpler models from among the candidate set. This mimics reality, where the generating process is highly complex, but the finite nature of any observed sample requires the use an approximating model, which is selected from among those under consideration.

5.2.1 Largest Model is Not Truth

In addition to those already presented, a variety of simulations were also performed in which the generating model was *not* the most complex model in the candidate set. One example occurred as a specific case of the lognormal simulation described in Section 2.5. The simulation parameter β_2 represented the difference between the lognormal population means, and β_2 was varied by increments of 0.1 on the interval $[0, 3]$. In the case where $\beta_2 = 0$, data was in fact generated under M_1, the simpler of the two candidate models. The results of this case appear in Figure 2.2.

When $\beta_2 = 0$, in addition to all non-zero values of β_2, the MATA-PL interval was seen to outperform the traditional MAW_2 and MAW_2^* confidence intervals in terms of error rates. In addition, for all values of β_2 we observed a consistent pattern in the trade-off between coverage rate and interval width (Fletcher and Turek, 2012). Hence, the broad results of this simulation study are robust to the choice of the data generating model.

Two simulation studies were presented in Section 3.4, in the framework of normal linear regression. In the second of these studies, a random model was selected among the candidate set for data generation at the onset of each simulation. The purpose of this was to assess the performance of confidence intervals in situations where truth was not the largest model. The results of this simulation appear in Figure 3.2, where for all values of sample size, the MATA-Wald interval outperformed the traditional MAW intervals in terms of error rates. Likewise, the MATA-Wald interval was not significantly wider than the MAW_1 or MAW_2 constructions (Turek and Fletcher, 2012). When simulation results were averaged across data generation under all candidate models, the superiority of the MATA-Wald interval construction persisted; once again, the conclusions from this simulation study were robust to the choice of the generating model.

A broad comparison between frequentist MATA interval constructions and Bayesian model-averaged credible intervals was presented in Chapter 4. In Section 4.6 we presented simulation results for normal and lognormal data generated under M_2, which was the more complex of the two candidate models. In addition to these simulations, analogous sets of simulations were also performed in which data were generated under the simpler model, M_1. Again, this included both the normal linear and lognormal settings. With data generation occurring under M_1, the underlying means of each sample were equal ($\beta_1 = \beta_2 = 0$). The remaining simulation details were identical to those described in Section 4.5. The results from this alternate study appear in Figure 5.3, for the normal linear setting with data generation under M_1.

With data generation occurring under the simpler model, M_1, no method of model-averaged interval construction is clearly superior to the others. The near equivalence between the MATA-Wald interval and the MAB_{KL} interval is still present, as was also observed in the normal linear simulation with data generation under M_2. In addition,

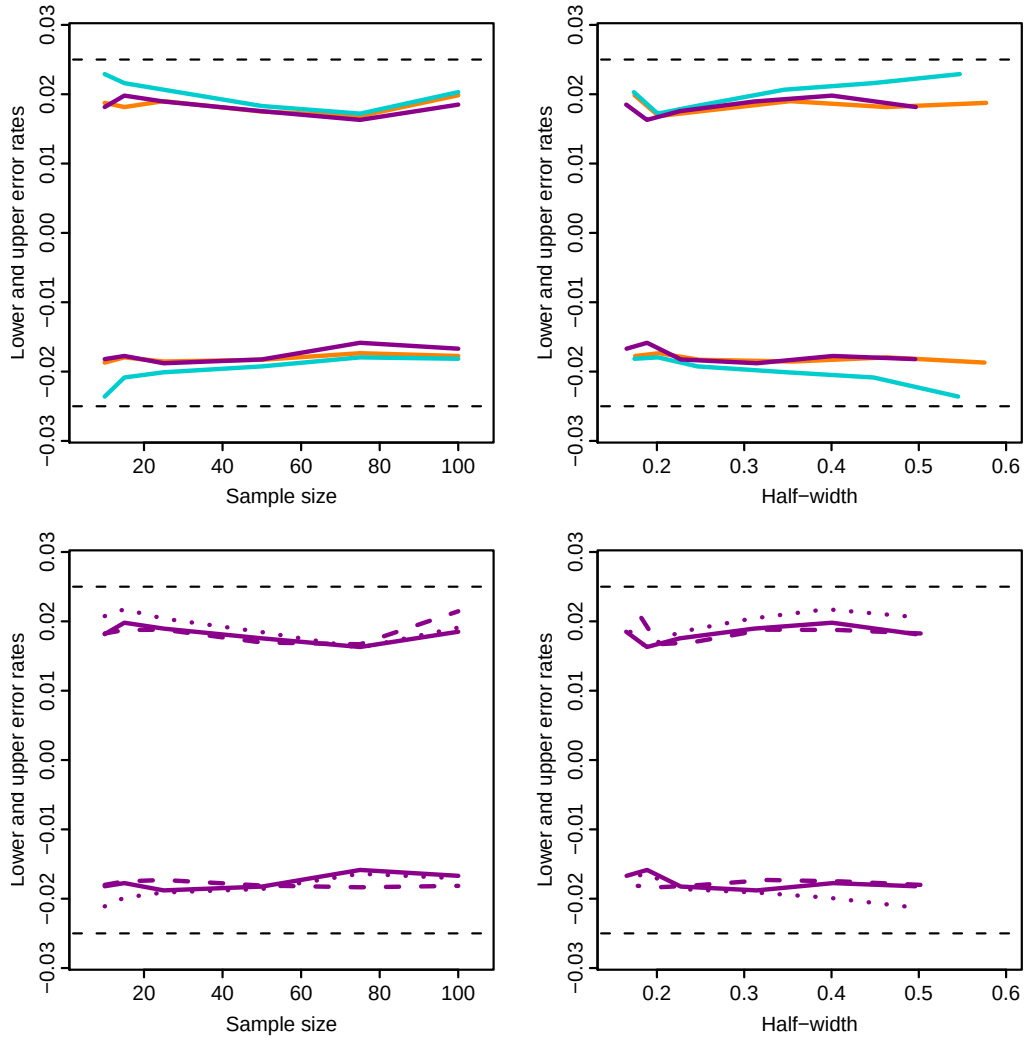

Figure 5.3: Confidence interval performance for prediction of the mean, θ_2, in the normal linear simulation described in Section 4.5, when data is generated under the simpler model (M_1). Top row: MAB (purple), MATA-Wald (orange), and MATA-PL (blue) intervals. Bottom row: MAB (solid), MAB_J (dotted), and MAB_KL (dashed) Bayesian intervals. Nominal error rates are shown as dashed lines, and lower error rates are plotted as negative values.

the use of the KL prior probabilities for the construction of a MAB_{KL} interval no longer produces a noticeable improvement; any choice of Bayesian prior distributions yields a reasonably comparable result.

It is worth noting that in this simulation, *both* models M_1 and M_2 are correct. Model averaging occurs over a set of two *correct* candidate models. The only noteworthy difference is that M_2 requires estimation of an additional, redundant parameter. Model-fitting under M_2 will therefore produce an inflated estimate of the variance, and result in a wider confidence interval. This propagates through the model averaging process to produce a wider model-averaged interval. Hence, the resulting intervals produce upper and lower error rates less than 2.5%. This suggests that when data generation occurs under a simpler model, no particular methodology for the construction of model-averaged intervals is clearly favorable. Appendix B.1 contains additional results of this nature, from the normal linear regression simulation presented in Chapter 3.

5.2.2 Largest Model is Truth: Consequences

Following the assumption that the largest model represents truth may in part explain the consistent benefits which resulted from the use of frequentist AIC weights. Similarly, results favoring the use of Bayesian KL model prior probabilities may be a consequence of this assumption. AIC is known to favor model complexity (Shibata, 1976; Katz, 1981), hence AIC weights will favor the *true* model more often than other information criteria, when data is generated under the largest model. This may, at least in part, explain the observed improvements resulting from frequentist AIC weights and Bayesian KL prior probabilities, in the performance of model-averaged intervals.

This preference towards AIC weights does not affect a broader result: the superior performance of MATA interval constructions, relative to traditional MAW methodolo-

gies for frequentist model averaging. Nor does it have any bearing on the relationships between frequentist and Bayesian model averaging techniques, in various simulation settings. Specifically, the agreement between the MATA-Wald interval and MAB_{KL} interval in the normal setting, and the general preference towards Bayesian model-averaging for non-normal data distributions. However, if one subscribes to the philosophy that natural processes are generally complex, our results *do* suggest that AIC weights and Bayesian KL priors are beneficial, for the construction of model-averaged confidence intervals.

5.3 Future Work

A variety of results have been observed through this thesis, relating to the construction and performance of the newly developed frequentist MATA confidence intervals. Specifically, the performance of MATA intervals in various simulation settings, including a comparison to Bayesian model averaging techniques. Throughout the course of our research, a number of open questions were observed, and new areas of potential future research were realized. We now discuss several of these possibilities for future work.

5.3.1 Profile Likelihood Modifications

An alternate MATA-PL interval could be constructed using a variation of the profile likelihood function. A modification of this function, the *modified* profile likelihood, is described in Davison (2003, p.680-691). This modification aims to alleviate the dependence upon the large-sample asymptotic distribution of the likelihood ratio statistic, which is the theoretical foundation of profile likelihood confidence intervals. Davison

(2003) suggests this modification may produce more accurate inferences in small-sample problems, where we have observed the MATA-PL interval to have poor coverage properties in several situations.

One can envision a *modified* MATA-PL interval, constructed by model-averaging the error rates of single-model *modified* profile likelihood intervals. It is difficult to predict exactly how such an interval would perform; however, if we expect each modified single-model profile likelihood interval to improve upon each standard profile likelihood interval, then it stands to reason that the *modified* MATA-PL interval would experience improvement as well.

5.3.2 Probability Matching Priors

Several different options for Bayesian prior distributions were considered, to gauge their effects on model-averaged Bayesian credible intervals. In addition to generally accepted "flat" non-informative priors (Gelman *et al.*, 2004, p.61-65), both Jeffreys prior distribution for parameters, and the informative, data-dependent KL prior probability for models were studied. Since we are comparing results for frequentist and Bayesian intervals, the study of probability matching priors (Mukerjee and Ghosh, 1997; Datta *et al.*, 2000; Datta and Mukerjee, 2004) could provide a fruitful avenue of future research.

Probability matching priors (for parameters) are defined to generate approximate agreement in coverage probabilities between single-model frequentist and Bayesian intervals. Their construction is such that the coverage probability of a Bayesian credible interval using probability matching priors is asymptotically equivalent to that of a frequentist interval, up to a certain order of approximation (Datta *et al.*, 2000). In the normal linear setting, an approximate equivalence between frequentist and Bayesian techniques was observed, when using KL prior probabilities for models. Probability

matching priors for parameters could be used in the context of model averaging, possibly in combination with KL prior probabilities for models. This combination could provide additional insight toward the relationships between frequentist and Bayesian model-averaged intervals, particularly in the non-linear setting.

5.3.3 Efficient Bayesian Model-Averaged Intervals

Herein, the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm (Green, 1995) has been used for performing Bayesian model averaging. Under the RJMCMC algorithm, an additional discrete-valued "model parameter" is introduced, which indexes the current model state. Using this additional parameter, a Gibbs sampler simultaneously traverses the parameter space within each model, as well as the "model space," or the discrete set of candidate models. The posterior distribution of this model parameter defines the Bayesian posterior model probabilities.

In some situations, the RJMCMC algorithm has been known to result in poor mixing between candidate models. The prior probability mass function for the model parameter may be difficult to tune, in order to achieve the desired mixing between model states (Green *et al.*, 2003, p.179-198). When confronted with a large set of candidate models, the reversible jump algorithm may not be able to achieve the desired mixing between the large number of model states. Furthermore, even when adequate mixing among the model states is achieved, a Gibbs sampler spends only a small fraction of the total computational time in model states with low posterior probability. For such states, the total computational time requirement to produce accurate estimates for parameter posterior distributions may be infeasible.

The KL prior probability for models has been seen to generate posterior model probabilities approximately equal to frequentist AIC model weights. Therefore, AIC

model weights can provide a reasonable approximation to the Bayesian posterior model probabilities resulting from KL prior probabilities. This suggests a method of approximating the results of Bayesian model averaging, without performing the time-intensive RJMCMC algorithm. Single-model Bayesian analyses could be performed under each model, which are each comparatively fast, and provide accurate approximations of posterior distributions within each model. As suggested, AIC weights may then be used to approximate the posterior model probabilities, without directly performing Bayesian model averaging.

A model-averaged posterior distribution may then be constructed, by weighting and combining the single-model posteriors under each model. Assuming good mixing within each model and an accurate approximation to posterior model probabilities, this would provide a justifiable approximation to a model-averaged posterior distribution produced using the RJMCMC algorithm. A full study of the time requirements and accuracy of this approximation has not been performed, but may provide a quick approximation suitable for constructing Bayesian model-averaged credible intervals.

5.4 Conclusions

We have presented a novel methodology for the construction of confidence intervals in the presence of model uncertainty, called the model-averaged tail area (MATA) interval. This approach can be applied to any single-model frequentist interval construction, to produce a model-averaged MATA interval. The methodology underlying the MATA interval is similar to that of Bayesian model averaging, although a MATA interval is entirely frequentist by nature.

Through simulation studies, we have compared MATA intervals against the preex-

isting approaches to constructing frequentist model-averaged intervals, and also against Bayesian model averaging. Intervals were compared in terms of coverage rates, and interval width. We observed the MATA-Wald interval to produce the most favorable coverage rates in the normal linear setting, without incurring an increased width. Similar coverage rates could also be produced using Bayesian methods and the data-dependent KL prior probability for models, which is designed to approximate frequentist AIC model weights. However, the computational requirements of the MATA-Wald interval are substantially less than the Bayesian approach, which can suffer from convergence issues in the multimodel framework.

In the non-normal setting, Bayesian model averaging produced the most favorable coverage properties. The improvement in coverage rate came at the expense of a substantial increase in interval width, compared to the frequentist constructions. However, this merely indicates that the Bayesian methodology better incorporates model uncertainty into the resultant credible intervals, which achieved very near to the nominal coverage rates. The use of KL prior probabilities for models provided a further improvement in the Bayesian coverage rate, however this may partially result from assumptions underlying the simulation design.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, *16*(1), 3–14.

Akaike, H. (1983). Information measures and model selection. *Bulletin of the International Statistical Institute*, *50*(1), 277–291.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*(3), 317–332.

Anderson, D., Burnham, K., and White, G. (1994). AIC model selection in overdispersed capture-recapture data. *Ecology*, *75*(6), 1780–1793.

Biondini, R. (1976). Cloud motion and rainfall statistics. *Journal of Applied Meteorology*, *15*(3), 205–224.

Boltzmann, L. (1896). *Vorlesungen über gastheorie*. JA Barth, Leipzig. [In German].

Box, G. and Tiao, G. (1973). *Bayesian inference in statistical analysis*. Addison-Wesley, Reading.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.

Brazzale, A., Davison, A., and Reid, N. (2007). *Applied asymptotics: case studies in small-sample statistics*. University Press, Cambridge.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: x-fixed prediction error. *Journal of the American Statistical Association*, *87*(419), 738–754.

Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Brown, L., Cai, T., and DasGupta, A. (2003). Interval estimation in exponential families. *Statistica Sinica*, *13*(1), 19–50.

Buckland, S., Burnham, K., and Augustin, N. (1997). Model selection: an integral part of inference. *Biometrics*, *53*(2), 603–618.

Burnham, K. and Anderson, D. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, New York.

Burnham, K. and Anderson, D. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research*, *33*(2), 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. (2003). The focused information criterion. *Journal of the American Statistical Association*, *98*(464), 900–916.

Claeskens, G. and Hjort, N. (2008). *Model selection and model averaging*. University Press, Cambridge.

Cover, T. and Thomas, J. (1991). *Elements of information theory*. John Wiley & Sons, New York.

Cox, D. and Hinkley, D. (1974). *Theoretical statistics*. Chapman & Hall, London.

Cox, D. and Snell, E. (1989). *Analysis of binary data*. Chapman & Hall, London.

Datta, G. and Mukerjee, R. (2004). *Probability matching priors: higher order asymptotics*. Springer-Verlag, New York.

Datta, G., Mukerjee, R., Ghosh, M., and Sweeting, T. (2000). Bayesian prediction with approximate frequentist validity. *Annals of Statistics*, *28*(5), 1414–1426.

Davison, A. (2003). *Statistical models*. University Press, Cambridge.

De Veaux, R., Velleman, P., and Bock, D. (2008). *Stats: data and models*. Pearson Addison-Wesley, New York.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B (Methodological)*, *57*(1), 45–97.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26.

Fletcher, D. and Dillingham, P. (2011). Model-averaged confidence intervals for factorial experiments. *Computational Statistics & Data Analysis*, *55*(11), 3041–3048.

Fletcher, D., MacKenzie, D., and Villouta, E. (2005). Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. *Environmental and Ecological Statistics*, *12*(1), 45–54.

Fletcher, D. and Turek, D. (2012). Model-averaged profile likelihood intervals. *Journal of Agricultural, Biological, and Environmental Statistics*, *17*(1), 38–51.

Freedman, D. (1983). A note on screening regression equations. *The American Statistician*, *37*(2), 152–155.

Freedman, L. and Pee, D. (1989). Return to a note on screening regression equations. *The American Statistician*, *43*(4), 279–282.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, *70*(350), 320–328.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian data analysis*. Chapman & Hall, Boca Raton.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711–732.

Green, P., Hjort, N., and Richardson, S. (2003). *Highly structured stochastic systems*. University Press, Oxford.

Hjort, N. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, *98*(464), 879–899.

Hjort, N. and Claeskens, G. (2006). Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association*, *101*(476), 1449–1464.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, *14*(4), 382–401.

Holländer, N., Augustin, N., and Sauerbrei, W. (2006). Investigation on the improvement of prediction by bootstrap model averaging. *Methods of Information in Medicine*, *45*(1), 44–50.

Hurvich, C. and Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307.

Hurvich, C. and Tsai, C. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, *44*(3), 214–217.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, Series A (Mathematical and Physical Sciences)*, *186*(1007), 453–461.

Kabaila, P. (1995). The effect of model selection on confidence regions and prediction regions. *Econometric Theory*, *11*(3), 537–549.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Katz, R. (1981). On some criteria for estimating the order of a Markov chain. *Technometrics*, *23*(3), 243–249.

King, R., Morgan, B., Gimenez, O., and Brooks, S. (2009). *Bayesian analysis for population ecology*. Chapman & Hall, Boca Raton.

Kullback, S. (1959). *Information theory and statistics*. John Wiley & Sons, New York.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Lebreton, J., Burnham, K., Clobert, J., and Anderson, D. (1992). Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monographs*, *62*(1), 67–118.

Lee, M. and Hugall, A. (2006). Model type, implicit data weighting, and model averaging in phylogenetics. *Molecular Phylogenetics and Evolution*, *38*(3), 848–857.

Link, W. and Barker, R. (2006). Model weights and the foundations of multimodel inference. *Ecology*, *87*(10), 2626–2635.

Lukacs, P., Burnham, K., and Anderson, D. (2010). Model selection bias and Freedman's paradox. *Annals of the Institute of Statistical Mathematics*, *62*(1), 117–125.

Martín, J. and Pérez, C. (2009). Bayesian analysis of a generalized lognormal distribution. *Computational Statistics & Data Analysis*, *53*(4), 1377–1387.

Miller, A. (1984). Selection of subsets of regression variables. *Journal of the Royal Statistical Society, Series A (General)*, *147*(3), 389–425.

Miller, A. (2002). *Subset selection in regression*. Chapman & Hall, Boca Raton.

Mooney, C. and Duval, R. (1993). *Bootstrapping: a nonparametric approach to statistical inference*. Sage Publications, Newbury Park.

Mukerjee, R. and Ghosh, M. (1997). Second-order probability matching priors. *Biometrika*, *84*(4), 970–975.

Nakagawa, S. and Freckleton, R. (2011). Model averaging, missing data and multiple imputation: a case study for behavioural ecology. *Behavioral Ecology and Sociobiology*, *65*(1), 103–116.

Raftery, A., Madigan, D., and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*(437), 179–191.

Rencher, A. and Pun, F. (1980). Inflation of R^2 in best subset regression. *Technometrics*, *22*(1), 49–53.

Ronchetti, E., Field, C., and Blanchard, W. (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association*, *92*(439), 1017–1023.

Rosenfeld, D. and Woodley, W. (1993). Effects of cloud seeding in west Texas: additional results and new insights. *Journal of Applied Meteorology*, *32*(12), 1848–1866.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Severini, T. (1991). On the relationship between Bayesian and non-Bayesian interval estimates. *Journal of the Royal Statistical Society, Series B (Methodological)*, *53*(3), 611–618.

Shannon, C. (1948). A mathematical theory of communications. *The Bell System Technical Journal*, *27*(1), 379–423, 623–656.

Shannon, C. and Weaver, W. (1949). *The mathematical theory of information*. University of Illinois Press, Urbana.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, *88*(422), 486–494.

Shao, J. (1996). Bootstrap model selection. *Journal of the American Statistical Association*, *91*(434), 655–665.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, *63*(1), 117–126.

Simpson, J. (1972). Use of the gamma distribution in single-cloud rainfall analysis. *Monthly Weather Review*, *100*(4), 309–312.

Simpson, J., Woodley, W., Miller, A., and Cotton, G. (1971). Precipitation results of two randomized pyrotechnic cumulus seeding experiments. *Journal of Applied Meteorology*, *10*(3), 526–544.

Spiegelhalter, D., Best, N., Carlin, B., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *64*(4), 583–639.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B (Methodological)*, *36*(2), 111–147.

Sugiura, N. (1978). Further analysts of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics: Theory and Methods*, *7*(1), 13–26.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)*, *153*(1), 12–18. [In Japanese].

Turek, D. and Fletcher, D. (2012). Model-averaged Wald confidence intervals. *Computational Statistics & Data Analysis*, *56*(9), 2809–2815.

Volinsky, C., Madigan, D., Raftery, A., and Kronmal, R. (1997). Bayesian model averaging in proportional hazard models: assessing the risk of a stroke. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *46*(4), 433–448.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, *61*(3), 439–447.

Weiss, N., Holmes, P., and Hardy, M. (2005). *A course in probability*. Pearson Addison-Wesley, New York.

Westfall, P. and Young, S. (1993). *Resampling-based multiple testing: examples and methods for p-value adjustment*. Wiley, New York.

Wintle, B., McCarthy, M., Volinsky, C., and Kavanagh, R. (2003). The use of Bayesian model averaging to better represent uncertainty in ecological models. *Conservation Biology*, *17*(6), 1579–1590.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, *92*(4), 937–950.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, *93*(441), 120–131.

Appendix A

R Programming Code for Model-Averaged Intervals

A.1 Ecklonia Abundance Example, Negative Binomial Simulation (Chapter 2)

```
### Model 3 log-likelihood functions
ll3.mu = function(much.b1.b2.b12.logk, xy) {
   mu.ch=much.b1.b2.b12.logk[1]
   b1=much.b1.b2.b12.logk[2]
   b2=much.b1.b2.b12.logk[3]
   b12=much.b1.b2.b12.logk[4]
   k=exp(much.b1.b2.b12.logk[5])
   mu = mu.ch * exp(b1*(xy$x1-xy$x1.ch)+b2*(xy$x2-xy$x2.ch)+b12*(xy$x1*xy
       $x2-xy$x1.ch*xy$x2.ch))
   sum(dnbinom(xy$y, size=k, mu=mu, log=T))}


ll3.logmu = function(logmuch.b1.b2.b12.logk, xy) {
   mu.ch=exp(logmuch.b1.b2.b12.logk[1])
   b1=logmuch.b1.b2.b12.logk[2]
   b2=logmuch.b1.b2.b12.logk[3]
   b12=logmuch.b1.b2.b12.logk[4]
   k=exp(logmuch.b1.b2.b12.logk[5])
   mu = mu.ch * exp(b1*(xy$x1-xy$x1.ch)+b2*(xy$x2-xy$x2.ch)+b12*(xy$x1*xy
       $x2-xy$x1.ch*xy$x2.ch))
   sum(dnbinom(xy$y, size=k, mu=mu, log=T))}


### Model 2 log-likelihood functions
ll2.mu = function(much.b1.b2.logk, xy) {
   mu.ch=much.b1.b2.logk[1]
   b1=much.b1.b2.logk[2]
   b2=much.b1.b2.logk[3]
   k=exp(much.b1.b2.logk[4])
   mu = mu.ch * exp(b1*(xy$x1-xy$x1.ch)+b2*(xy$x2-xy$x2.ch))
   sum(dnbinom(xy$y, size=k, mu=mu, log=T))}


ll2.logmu = function(logmuch.b1.b2.logk, xy) {
   mu.ch=exp(logmuch.b1.b2.logk[1])
   b1=logmuch.b1.b2.logk[2]
   b2=logmuch.b1.b2.logk[3]
   k=exp(logmuch.b1.b2.logk[4])
   mu = mu.ch * exp(b1*(xy$x1-xy$x1.ch)+b2*(xy$x2-xy$x2.ch))
   sum(dnbinom(xy$y, size=k, mu=mu, log=T))}
```

```
### Model 1 log-likelihood functions
ll1.mu = function(much.b1.logk, xy) {
   mu.ch=much.b1.logk[1]
   b1=much.b1.logk[2]
   k=exp(much.b1.logk[3])
   mu = mu.ch * exp(b1*(xy$x1-xy$x1.ch))
   sum(dnbinom(xy$y, size=k, mu=mu, log=T))}


ll1.logmu = function(logmuch.b1.logk, xy) {
   mu.ch=exp(logmuch.b1.logk[1])
   b1=logmuch.b1.logk[2]
   k=exp(logmuch.b1.logk[3])
   mu = mu.ch * exp(b1*(xy$x1-xy$x1.ch))
   sum(dnbinom(xy$y, size=k, mu=mu, log=T))}


### profile log-likelihood functions
ll3.givenmu = function(b1.b2.b12.logk, much.par, xy) {
   b1=b1.b2.b12.logk[1]
   b2=b1.b2.b12.logk[2]
   b12=b1.b2.b12.logk[3]
   k=exp(b1.b2.b12.logk[4])
   mu = much.par * exp(b1*(xy$x1-xy$x1.ch)+b2*(xy$x2-xy$x2.ch)+b12*(xy$x1
       *xy$x2-xy$x1.ch*xy$x2.ch))
   sum(dnbinom(xy$y, size=k, mu=mu, log=T))}


ll2.givenmu = function(b1.b2.logk, much.par, xy) {
   b1=b1.b2.logk[1]
   b2=b1.b2.logk[2]
   k=exp(b1.b2.logk[3])
   mu = much.par * exp(b1*(xy$x1-xy$x1.ch)+b2*(xy$x2-xy$x2.ch))
   sum(dnbinom(xy$y, size=k, mu=mu, log=T))}


ll1.givenmu = function(b1.logk, much.par, xy) {
   b1=b1.logk[1]
   k=exp(b1.logk[2])
   mu = much.par * exp(b1*(xy$x1-xy$x1.ch))
   sum(dnbinom(xy$y, size=k, mu=mu, log=T))}



### signed likelihood ratio statistic functions
r3 = function(much.par, maxll, muhat, ctlist, xy) {
    pll3 = optim(c(.1,.1,.1,0), ll3.givenmu, much.par=much.par, xy=xy,
        control=ctlist)$value
    ifelse(maxll[3]-pll3<0, 0, sign(muhat[3]-much.par) * sqrt(2*(maxll[3]-
        pll3)))}
```

```
r2 = function(much.par, maxll, muhat, ctlist, xy) {
    pll2 = optim(c(.1,.1,0), ll2.givenmu, much.par=much.par, xy=xy,
        control=ctlist)$value
    ifelse(maxll[2]-pll2<0, 0, sign(muhat[2]-much.par) * sqrt(2*(maxll[2]-
        pll2)))}


r1 = function(much.par, maxll, muhat, ctlist, xy) {
    pll1 = optim(c(.1,0), ll1.givenmu, much.par=much.par, xy=xy, control=
        ctlist)$value
    ifelse(maxll[1]-pll1<0, 0, sign(muhat[1]-much.par) * sqrt(2*(maxll[1]-
        pll1)))}



### model-averaged functions for upper & lower MPI limits
r.ma.upp = function(mu.par, waic, maxll, muhat, ctlist, xy) {
   sum(waic*pnorm(c(r1(mu.par, maxll, muhat, ctlist, xy), r2(mu.par,
       maxll, muhat, ctlist, xy), r3(mu.par, maxll, muhat, ctlist, xy))))
       - 0.025}


r.ma.low = function(mu.par, waic, maxll, muhat, ctlist, xy) {
   sum(waic*pnorm(c(r1(mu.par, maxll, muhat, ctlist, xy), r2(mu.par,
       maxll, muhat, ctlist, xy), r3(mu.par, maxll, muhat, ctlist, xy))))
       - 0.975}


### extract standard error from Hessian matrix
getSE.hess = function(hess, position=1) {
    hessinv = solve(hess)
    sqrt(-hessinv[position,][position])}
```

```
### generate parameters & data
a   =  2.451904
b1  =  0.031744
b2  =  0.151169
b12 = -0.006778
k   =  0.402237
n   =  103

x1 = c(0, 0, 0, 3, 5, 5, 6, 11, 13, 18, 18, 19, 19, 27, 27, 29, 31, 32,
    33, 34, 38, 39, 50, 78, 110, 133, 220, 15, 15, 23, 0, 15, 16, 19, 34,
    35, 14, 115, 22, 15, 53, 64, 28, 52, 14, 0, 11, 11, 14, 20, 7, 5, 0,
    12, 20, 11, 11, 20, 13, 19, 29, 52, 24, 25, 0, 19, 0, 35, 0, 0, 28, 4,
     0, 29, 2, 11, 15, 0, 28, 0, 9, 0, 1, 0, 0, 0, 5, 2, 3, 8, 0, 0, 25,
    2, 28, 2, 4, 29, 13, 0, 4, 5, 50)

x2 = c(7, 7, 12, 11, 10, 11, 10, 9, 11, 9, 9, 8, 7, 7, 7, 7, 8, 6, 10,
    11, 13, 5, 6, 6, 5, 7, 11, 10, 6, 6, 4, 6, 11, 11, 11, 12, 11, 9, 5,
    7, 7, 7, 11, 9, 11, 11, 11, 6, 8, 10, 11, 7, 12, 11, 5, 9, 13, 6, 14,
    11, 8, 5, 6, 7, 13, 11, 13, 4, 9, 12, 4, 13, 9, 8, 12, 7, 11, 11, 7,
    14, 8, 12, 12, 14, 12, 9, 9, 10, 11, 11, 10, 10, 12, 12, 12, 11, 7,
    11, 9, 11, 12, 12, 3)


mu    = exp(a + b1*x1    + b2*x2    + b12*x1   *x2   )
mu.ch = exp(a + b1*x1.ch + b2*x2.ch + b12*x1.ch*x2.ch)

y = rnbinom(n, size=k, mu=mu)
xy = list(y=y, x1=x1, x1.ch=x1.ch, x2=x2, x2.ch=x2.ch)


### muhat, logmuhat, SE values
optim3.mu    = optim(c(    mu.ch   ,b1,b2,b12,log(k)), ll3.mu,    xy=xy)
optim3.logmu = optim(c(log(mu.ch)  ,b1,b2,b12,log(k)), ll3.logmu, xy=xy)
optim2.mu    = optim(c(    mu.ch   ,b1,b2,    log(k)), ll2.mu,    xy=xy)
optim2.logmu = optim(c(log(mu.ch)  ,b1,b2,    log(k)), ll2.logmu, xy=xy)
optim1.mu    = optim(c(    mean(y) ,b1,       log(k)), ll1.mu,    xy=xy)
optim1.logmu = optim(c(log(mean(y)),b1,       log(k)), ll1.logmu, xy=xy)


muhat = c(optim1.mu$par[1], optim2.mu$par[1], optim3.mu$par[1])

logmuhat = c(optim1.logmu$par[1],
             optim2.logmu$par[1],
             optim3.logmu$par[1])

semuhat = c(getSE.hess(hessian(ll1.mu, optim1.mu$par, xy=xy)),
            getSE.hess(hessian(ll2.mu, optim2.mu$par, xy=xy)),
            getSE.hess(hessian(ll3.mu, optim3.mu$par, xy=xy)))

selogmuhat = c(getSE.hess(hessian(ll1.logmu, optim1.logmu$par, xy=xy)),
               getSE.hess(hessian(ll2.logmu, optim2.logmu$par, xy=xy)),
               getSE.hess(hessian(ll3.logmu, optim3.logmu$par, xy=xy)))
```

```
### Full Model intervals: FWI, FWIT, FPI
z = qnorm(0.975)
fwi  =        muhat[3] + z*semuhat[3]    *c(-1,1)
fwit = exp(logmuhat[3] + z*selogmuhat[3]*c(-1,1))


fpi = c(
   uniroot(r.ma.low,  c(1e-7, 1e6), waic=c(0,0,1), maxll=maxll, muhat=
      muhat, ctlist=ctlist, xy=xy, tol=1e-10)$root,
   uniroot(r.ma.upp,  c(1e-5, 1e8), waic=c(0,0,1), maxll=maxll, muhat=
      muhat, ctlist=ctlist, xy=xy, tol=1e-10)$root)


### Best Model (AIC sense) intervals: BWI, BWIT, BPI
best.ind = which(waic==max(waic))[1]
best.aic = c(0,0,0)
best.aic[best.ind] = 1
bwi  =        muhat[best.ind] + z*semuhat    [best.ind]*c(-1,1)
bwit = exp(logmuhat[best.ind] + z*selogmuhat[best.ind]*c(-1,1))


bpi = c(
   uniroot(r.ma.low,  c(1e-7, 1e6), waic=best.aic, maxll=maxll, muhat=
      muhat, ctlist=ctlist, xy=xy, tol=1e-10)$root,
   uniroot(r.ma.upp,  c(1e-5, 1e8), waic=best.aic, maxll=maxll, muhat=
      muhat, ctlist=ctlist, xy=xy, tol=1e-10)$root)


### Model-Averaged intervals: MWI, MWIT, MPI
ind = which(waic>0.001)
mubar = sum(waic[ind]*muhat[ind])
varmubar = sum(waic[ind]*((semuhat[ind])^2+(muhat[ind]-mubar)^2))
mwi = mubar + z*sqrt(varmubar)*c(-1,1)

logmubar = sum(waic[ind]*logmuhat[ind])
varlogmubar = sum(waic[ind]*((selogmuhat[ind])^2+(logmuhat[ind]-logmubar)
   ^2))
mwit = exp(logmubar + z*sqrt(varlogmubar)*c(-1,1))


mpi = c(
   uniroot(r.ma.low,  c(1e-7, 1e6), waic=waic, maxll=maxll, muhat=muhat,
      ctlist=ctlist, xy=xy, tol=1e-10)$root,
   uniroot(r.ma.upp,  c(1e-5, 1e8), waic=waic, maxll=maxll, muhat=muhat,
      ctlist=ctlist, xy=xy, tol=1e-10)$root)
```

A.2 Lognormal Simulation (Chapter 2)

```
y1 = rlnorm(n, theta1, sqrt(sigsq))
y2 = rlnorm(n, theta2, sqrt(sigsq))
y = c(y1, y2)
logy  = log(y)
logy1 = log(y1)
logy2 = log(y2)

mu1 = exp(theta1 + sigsq/2)
mu2 = exp(theta2 + sigsq/2)


### Hessians, SEs, log-likelihood values
fv.ll.theta = function(thetasigsq, y) {
   sum(dlnorm(y,meanlog=thetasigsq[1],sdlog=sqrt(thetasigsq[2]),log=T))}

fv.ll.mu = function(musigsq, y) {
   sum(dlnorm(y, meanlog=log(musigsq[1])-musigsq[2]/2, sdlog=sqrt(
       musigsq[2]), log=T))}

fv.ll.logmu = function(logmusigsq, y) {
   sum(dlnorm(y, meanlog=logmusigsq[1]-logmusigsq[2]/2, sdlog=sqrt(
       logmusigsq[2]), log=T))}

fv.ll.mu12 = function(mu12sigsq, y1, y2) {
   ll1 = dlnorm(y1, meanlog=log(mu12sigsq[1])-mu12sigsq[3]/2, sdlog=sqrt(
       mu12sigsq[3]), log=T)
   ll2 = dlnorm(y2, meanlog=log(mu12sigsq[2])-mu12sigsq[3]/2, sdlog=sqrt(
       mu12sigsq[3]), log=T)
   sum(c(ll1, ll2))}

fv.ll.logmu12 = function(logmu12sigsq, y1, y2) {
   ll1 = dlnorm(y1, meanlog=logmu12sigsq[1] -logmu12sigsq[3]/2, sdlog=
       sqrt(logmu12sigsq[3]), log=T)
   ll2 = dlnorm(y2, meanlog=logmu12sigsq[2] -logmu12sigsq[3]/2, sdlog=
       sqrt(logmu12sigsq[3]), log=T)
   sum(c(ll1, ll2))}


### calculate SE(logmu)
optiml.m1   = optim(c(0,1),   fv.ll.logmu,   y=y,           hessian=T)
optiml12.m2 = optim(c(0,0,1), fv.ll.logmu12, y1=y1, y2=y2, hessian=T)
hessianl.m1 =   optiml.m1$hessian
hessianl12.m2 = optiml12.m2$hessian
hesslinv.m1   = solve(hessianl.m1)
hessl12inv.m2 = solve(hessianl12.m2)
selogmuhat.m1 =  sqrt(-1*hesslinv.m1[1][1])
logmu1row = hessl12inv.m2[1,]
logmu2row = hessl12inv.m2[2,]
selogmu1hat.m2 = sqrt(-1*logmu1row[1])
selogmu2hat.m2 = sqrt(-1*logmu2row[2])
```

```
### calculate SE(mu)
optim.m1    = optim(c(1,1),    fv.ll.mu,    y=y,              hessian=T)
optim12.m2 = optim(c(1,1,1), fv.ll.mu12, y1=y1, y2=y2, hessian=T)
hessian.m1    = optim.m1$hessian
hessian12.m2 = optim12.m2$hessian
hessinv.m1  = solve(hessian.m1)
hess12inv.m2 = solve(hessian12.m2)
semuhat.m1 =  sqrt(-1*hessinv.m1[1][1])
mu1row = hess12inv.m2[1,]
mu2row = hess12inv.m2[2,]
semu1hat.m2 = sqrt(-1*mu1row[1])
semu2hat.m2 = sqrt(-1*mu2row[2])



### maximized log-likelihood values
maxll.m1  = optim.m1$value
maxll.m2  = optim12.m2$value



### FWI
mu1fwi.low = mu1hat.m2 - z*semu1hat.m2
mu1fwi.upp = mu1hat.m2 + z*semu1hat.m2
mu2fwi.low = mu2hat.m2 - z*semu2hat.m2
mu2fwi.upp = mu2hat.m2 + z*semu2hat.m2


### BWI
mu1hat.best    = ifelse(w.m1>w.m2,    muhat.m1,    mu1hat.m2)
mu2hat.best    = ifelse(w.m1>w.m2,    muhat.m1,    mu2hat.m2)
semu1hat.best = ifelse(w.m1>w.m2, semuhat.m1, semu1hat.m2)
semu2hat.best = ifelse(w.m1>w.m2, semuhat.m1, semu2hat.m2)
mu1bwi.low = mu1hat.best - z*semu1hat.best
mu1bwi.upp = mu1hat.best + z*semu1hat.best
mu2bwi.low = mu2hat.best - z*semu2hat.best
mu2bwi.upp = mu2hat.best + z*semu2hat.best


### MWI
mu1bar = w.m1*muhat.m1 + w.m2*mu1hat.m2
mu2bar = w.m1*muhat.m1 + w.m2*mu2hat.m2
varmu1bar = w.m1*(semuhat.m1^2+(muhat.m1-mu1bar)^2) + w.m2*(semu1hat.m2
    ^2+(mu1hat.m2-mu1bar)^2)
varmu2bar = w.m1*(semuhat.m1^2+(muhat.m1-mu2bar)^2) + w.m2*(semu2hat.m2
    ^2+(mu2hat.m2-mu2bar)^2)
mu1mwi.low = mu1bar - z*sqrt(varmu1bar)
mu1mwi.upp = mu1bar + z*sqrt(varmu1bar)
mu2mwi.low = mu2bar - z*sqrt(varmu2bar)
mu2mwi.upp = mu2bar + z*sqrt(varmu2bar)
```

```
### FWIT
mu1fwit.low = exp(logmu1hat.m2 - z*selogmu1hat.m2)
mu1fwit.upp = exp(logmu1hat.m2 + z*selogmu1hat.m2)
mu2fwit.low = exp(logmu2hat.m2 - z*selogmu2hat.m2)
mu2fwit.upp = exp(logmu2hat.m2 + z*selogmu2hat.m2)


### BWIT
logmu1hat.best    = ifelse(w.m1>w.m2,    logmuhat.m1,    logmu1hat.m2)
logmu2hat.best    = ifelse(w.m1>w.m2,    logmuhat.m1,    logmu2hat.m2)
selogmu1hat.best = ifelse(w.m1>w.m2, selogmuhat.m1, selogmu1hat.m2)
selogmu2hat.best = ifelse(w.m1>w.m2, selogmuhat.m1, selogmu2hat.m2)
mu1bwit.low = exp(logmu1hat.best - z*selogmu1hat.best)
mu1bwit.upp = exp(logmu1hat.best + z*selogmu1hat.best)
mu2bwit.low = exp(logmu2hat.best - z*selogmu2hat.best)
mu2bwit.upp = exp(logmu2hat.best + z*selogmu2hat.best)


### MWIT
logmu1bar = w.m1*logmuhat.m1 + w.m2*logmu1hat.m2
logmu2bar = w.m1*logmuhat.m1 + w.m2*logmu2hat.m2
varlogmu1bar = w.m1*(selogmuhat.m1^2+(logmuhat.m1-logmu1bar)^2) + w.m2*(
    selogmu1hat.m2^2+(logmu1hat.m2-logmu1bar)^2)
varlogmu2bar = w.m1*(selogmuhat.m1^2+(logmuhat.m1-logmu2bar)^2) + w.m2*(
    selogmu2hat.m2^2+(logmu2hat.m2-logmu2bar)^2)
mu1mwit.low = exp(logmu1bar - z*sqrt(varlogmu1bar))
mu1mwit.upp = exp(logmu1bar + z*sqrt(varlogmu1bar))
mu2mwit.low = exp(logmu2bar - z*sqrt(varlogmu2bar))
mu2mwit.upp = exp(logmu2bar + z*sqrt(varlogmu2bar))


### Functions for Profile-Likelihood intervals
f1.ll.sig = function(sigsq, mu, y) {
   sum(dlnorm(y, meanlog=log(mu)-sigsq/2, sdlog=sqrt(sigsq), log=T))}

fv.ll.mu1 = function(mu2sigsq, mu1, y1, y2) {
    ll1 = sum(dlnorm(y1, meanlog=log(mu1)          -mu2sigsq[2]/2, sdlog=
        sqrt(mu2sigsq[2]), log=T))
    ll2 = sum(dlnorm(y2, meanlog=log(mu2sigsq[1])-mu2sigsq[2]/2, sdlog=
        sqrt(mu2sigsq[2]), log=T))
    ll1+ll2}

r.m1 = function(mu.par) {
    pll = optimize(f1.ll.sig, c(0, 10^6), mu=mu.par, y=y)$objective
    sign(muhat.m1-mu.par)*sqrt(2*abs(maxll.m1-pll))}

r1.m2 = function(mu1.par) {
    pll = optim(c(1,1), fv.ll.mu1, mu1=mu1.par, y1=y1, y2=y2)$value
    sign(mu1hat.m2-mu1.par)*sqrt(2*(maxll.m2-pll))}

r2.m2 = function(mu2.par) {
    pll = optim(c(1,1), fv.ll.mu1, mu1=mu2.par, y1=y2, y2=y1)$value
    sign(mu2hat.m2-mu2.par)*sqrt(2*(maxll.m2-pll))}
```

```
### FPI
r1u.full = function(mu1.par) {pnorm(r1.m2(mu1.par))-0.025}
r1l.full = function(mu1.par) {pnorm(r1.m2(mu1.par))-0.975}
r2u.full = function(mu2.par) {pnorm(r2.m2(mu2.par))-0.025}
r2l.full = function(mu2.par) {pnorm(r2.m2(mu2.par))-0.975}

mu1fpi.low = uniroot(r1l.full, c(0.0001,10^9))$root
mu1fpi.upp = uniroot(r1u.full, c(0.0001,10^9))$root
mu2fpi.low = uniroot(r2l.full, c(0.0001,10^9))$root
mu2fpi.upp = uniroot(r2u.full, c(0.0001,10^9))$root



### BPI
w1.best = ifelse(w.m1>w.m2, 1, 0)
w2.best = ifelse(w.m1>w.m2, 0, 1)

r1u.best = function(mu1.par) {
   w1.best*pnorm(r.m1(mu1.par))+w2.best*pnorm(r1.m2(mu1.par))-0.025}

r1l.best = function(mu1.par) {
   w1.best*pnorm(r.m1(mu1.par))+w2.best*pnorm(r1.m2(mu1.par))-0.975}

r2u.best = function(mu2.par) {
   w1.best*pnorm(r.m1(mu2.par))+w2.best*pnorm(r2.m2(mu2.par))-0.025}

r2l.best = function(mu2.par) {
   w1.best*pnorm(r.m1(mu2.par))+w2.best*pnorm(r2.m2(mu2.par))-0.975}


mu1bpi.low = uniroot(r1l.best, c(0.0001,10^9))$root
mu1bpi.upp = uniroot(r1u.best, c(0.0001,10^9))$root
mu2bpi.low = uniroot(r2l.best, c(0.0001,10^9))$root
mu2bpi.upp = uniroot(r2u.best, c(0.0001,10^9))$root


### MPI
r1u.ma = function(mu1.par) {
   w.m1*pnorm(r.m1(mu1.par))+w.m2*pnorm(r1.m2(mu1.par))-0.025}

r1l.ma = function(mu1.par) {
   w.m1*pnorm(r.m1(mu1.par))+w.m2*pnorm(r1.m2(mu1.par))-0.975}

r2u.ma = function(mu2.par) {
   w.m1*pnorm(r.m1(mu2.par))+w.m2*pnorm(r2.m2(mu2.par))-0.025}

r2l.ma = function(mu2.par) {
   w.m1*pnorm(r.m1(mu2.par))+w.m2*pnorm(r2.m2(mu2.par))-0.975}

mu1mpi.low = uniroot(r1l.ma, c(0.0001,10^9))$root
mu1mpi.upp = uniroot(r1u.ma, c(0.0001,10^9))$root
mu2mpi.low = uniroot(r2l.ma, c(0.0001,10^9))$root
mu2mpi.upp = uniroot(r2u.ma, c(0.0001,10^9))$root
```

A.3 Normal Linear Regression Simulation (Chapter 3)

```
wald.tailarea.norm = function(mu.par, muhats, semuhats, weights, alpha) {
   sum(weights*pnorm(mu.par, mean=muhats, sd=semuhats)) - alpha}


wald.tailarea.t = function(mu.par,muhats,semuhats,dfs,weights,alpha) {
   mu.quantiles = (mu.par-muhats)/semuhats
   sum(weights*pt(mu.quantiles, df=dfs)) - alpha}



### generate data, fit models
x2 = rnorm (n, mean =params$x2.mean, sd=params$x2.sd)
x3 = rgamma(n, shape=params$x3.shape, rate=params$x3.rate)
x23 = x2*x3



### probabilistic data generation
model.choice = rmultinom(1,1,pis)

if(model.choice[1,1]==1) {mu = b[1]; mu.ch = b[1]}
if(model.choice[2,1]==1) {mu = b[1] + b[2]*x2; mu.ch = b[1] + b[2]*x2.ch}
if(model.choice[3,1]==1) {mu = b[1] + b[3]*x3; mu.ch = b[1] + b[3]*x3.ch}
if(model.choice[4,1]==1) {mu = b[1] + b[2]*x2 + b[3]*x3; mu.ch = b[1] + b
    [2]*x2.ch + b[3]*x3.ch}
if(model.choice[5,1]==1) {mu = b[1] + b[2]*x2 + b[3]*x3 + b[4]*x23; mu.ch
     = b[1] + b[2]*x2.ch + b[3]*x3.ch + b[4]*x23.ch}

y = rnorm(n, mean=mu, sd=sigma)


m1 = glm(y ~ 1)
m2 = glm(y ~ 1+x2)
m3 = glm(y ~ 1   +x3)
m4 = glm(y ~ 1+x2+x3)
m5 = glm(y ~ 1+x2+x3+x23)


p1 = predict(m1, se=T, newdata=data.frame(1))
p2 = predict(m2, se=T, newdata=data.frame(x2=x2.ch))
p3 = predict(m3, se=T, newdata=data.frame(x3=x3.ch))
p4 = predict(m4, se=T, newdata=data.frame(x2=x2.ch,x3=x3.ch))
p5 = predict(m5, se=T, newdata=data.frame(x2=x2.ch,x3=x3.ch,x23=x23.ch))
```

```
z = qnorm(1-alpha)
t.dfs = n - p
ts = qt(1-alpha,t.dfs)


### Best Model Wald interval: b
best.ind = which(waic==max(waic))[1]
b.l = muhat[best.ind,] - ts[best.ind]*semuhat[best.ind,]
b.u = muhat[best.ind,] + ts[best.ind]*semuhat[best.ind,]


### True Model Wald interval: t
true.ind = which(model.choice==1)[1]
t.l = muhat[true.ind,] - ts[true.ind]*semuhat[true.ind,]
t.u = muhat[true.ind,] + ts[true.ind]*semuhat[true.ind,]


### Model-Averaged Wald intervals: ma1, ma1tz, ma2, ma2tz
mubar = t(waic) %*% muhat

semubar1   = apply((waic%*%rep(1,num.ch)) * sqrt((semuhat^2 + (muhat-
    matrix(mubar, nrow=5, ncol=num.ch, byrow=T))^2)),2,sum)

semubar1tz = apply((waic%*%rep(1,num.ch))*sqrt(((diag(ts/z)%*%semuhat)^2
    + (muhat-matrix(mubar, nrow=5, ncol=num.ch, byrow=T))^2)),2,sum)

semubar2   = sqrt(apply((waic%*%rep(1,num.ch)) * (semuhat^2 + (muhat-
    matrix(mubar, nrow=5, ncol=num.ch, byrow=T))^2)  ,2,sum))

semubar2tz = sqrt(apply((waic%*%rep(1,num.ch))*((diag(ts/z)%*%semuhat)^2
    + (muhat-matrix(mubar, nrow=5, ncol=num.ch, byrow=T))^2)  ,2,sum))


ma1.l   = as.numeric(mubar - z*semubar1)
ma1.u   = as.numeric(mubar + z*semubar1)
ma1tz.l = as.numeric(mubar - z*semubar1tz)
ma1tz.u = as.numeric(mubar + z*semubar1tz)
ma2.l   = as.numeric(mubar - z*semubar2)
ma2.u   = as.numeric(mubar + z*semubar2)
ma2tz.l = as.numeric(mubar - z*semubar2tz)
ma2tz.u = as.numeric(mubar + z*semubar2tz)



### Model-Averaged Wald Tail Areas: mata
mata.l = rep(0,num.ch)
mata.u = rep(0,num.ch)

for (i in 1:num.ch) {
   mata.l[i] = uniroot(wald.tailarea.t, c(-1e7, 1e7), muhats=muhat[,i],
       semuhats=semuhat[,i], dfs=t.dfs, weights=waic, alpha=alpha)  $root
   mata.u[i] = uniroot(wald.tailarea.t, c(-1e7, 1e7), muhats=muhat[,i],
       semuhats=semuhat[,i], dfs=t.dfs, weights=waic, alpha=1-alpha)$root}
```

A.4 Cloud Seeding Example, Bayesian vs. Frequentist Simulation (Chapter 4)

```
llbayes.n = function(x, y, b, sigma) {
   sum(dnorm(y, mean=(x%*%b), sd=sigma, log=T))}


llbayes.logn = function(x, y, b, sigma) {
   sum(dlnorm(exp(y), meanlog=(x%*%b)-sigma^2/2, sdlog=sigma, log=T))}




### Bayesian MCMC and RJMCMC functions
gibbs = function(bayespar) { with(bayespar, {

   mod = inits$model

   if(mod==1) {param = c(inits$sigma, inits$mu,  inits$mu )}
   if(mod==2) {param = c(inits$sigma, inits$mu1, inits$mu2)}

   ll = llbayes(x=x, y=y, b=param[2:maxpar], sigma=param[1])
   itns = array(0, dim=c(nit, maxpar))
   output = rep(0, (1+maxpar))

   for (t in 1:nit) {
      output = updateparam(mod=mod, param=param, ll=ll)
      ll = output[1]
      param = output[2:(1+maxpar)]

      if(rj) {
         output = updatemodel(mod=mod, param=param, ll=ll)
         ll     = output[1]
         param = output[2:(1+maxpar)]
         mod = ifelse(param[2]==param[3], 1, 2)
      }

      itns[t,] = param
   }

   itns = itns[(nburn+1):nit,]
   itns
)}
```

```
updateparam = function(mod, param, ll, bayespar) { with(bayespar, {
   for (i.par in 1:modnpar[mod]) {
      oldparam = param[i.par]
      if(i.par==1) {
         param[i.par] = max(oldparam + propose.sigma(), 0.000001)
         newll = llbayes(x=x, y=y, b=param[2:maxpar], sigma=param[1])
         num = newll + prior.sigma(param[i.par])
         den = ll    + prior.sigma(oldparam)
      }
      if(i.par>1) {
         param[i.par] = oldparam + propose.b()
         if(mod==1) {param[3] = param[2]}
         newll = llbayes(x=x, y=y, b=param[2:maxpar], sigma=param[1])
         num = newll + prior.b(param[i.par])
         den = ll    + prior.b(oldparam)
      }
      A = min(1, exp(num-den))
      u = runif(1)
      if(u<A) {ll = newll}
      if(u>A) {
         param[i.par] = oldparam
         if(mod==1) {param[3] = param[2]}
   } }
   output = c(ll, param)
   output}
)}


updatemodel = function(mod, param, ll, bayespar) { with(bayespar, {
   oldparams = param[2:maxpar]
   if(mod==1) {
      u.prop = propose.rjb()
      param[2:maxpar] = oldparams + u.prop*c(-1,1)
      newll = llbayes(x=x, y=y, b=param[2:maxpar], sigma=param[1])
      num = newll + log(mpriors[2]) + prior.b(param[2])      + prior.b(
          param[3])
      den = ll    + log(mpriors[1]) + prior.b(oldparams[1]) + propose.rjb
          .dens(u.prop)
   }
   if(mod==2) {
      u.infer = 1/2 * (oldparams[1]-oldparams[2])
      param[2:maxpar] = mean(oldparams)
      newll = llbayes(x=x, y=y, b=param[2:maxpar], sigma=param[1])
      num = newll + log(mpriors[1]) + prior.b(param[2])      + propose.rjb
          .dens(u.infer)
      den = ll    + log(mpriors[2]) + prior.b(oldparams[1]) + prior.b(
          oldparams[2])
   }
   A = min(1, exp(num-den))
   u = runif(1)
   if(u<A) {ll = newll}
   if(u>A) {param[2:maxpar] = oldparams}
   output = c(ll, param)
   output})
}
```

```
### data generation
if(logn) {
   y1 = log(rlnorm(n, meanlog = theta1, sdlog = sigma))
   y2 = log(rlnorm(n, meanlog = theta2, sdlog = sigma))
} else   {
   y1 =      rnorm(n, mean    = theta1, sd    = sigma)
   y2 =      rnorm(n, mean    = theta2, sd    = sigma)
}


### generate: muhat, semuhat, and maxll
m1 = glm(y ~ -1+x1)
m2 = glm(y ~ -1+x21+x22)
p1 = predict(m1, se=T, newdata=data.frame(x1=c(1,1)))
p2 = predict(m2, se=T, newdata=data.frame(x21=c(1,0), x22=c(0,1)))


if(logn) {
   s1 = p1$residual.scale
   s2 = p2$residual.scale
   muhat = rbind(p1$fit+(s1^2)/2, p2$fit+(s2^2)/2)
   semuhat = matrix(c(sqrt(s1^2/(2*n) + s1^4/(2*(2*n-1))), sqrt(s2^2/(n)
      + s2^4/(2*(2*n-2)))), nrow=2, ncol=2)
   maxll = c(sum(dlnorm(exp(y), m1$coef, sqrt(m1$dev/(2*n)), log=T)), sum
      (dlnorm(exp(y1), m2$coef[1], sqrt(m2$dev/(2*n)), log=T))+sum(dlnorm
      (exp(y2), m2$coef[2], sqrt(m2$dev/(2*n)), log=T)))

} else   {
   muhat = matrix(c(p1$fit, p2$fit), byrow=T, nrow=2)
   semuhat = matrix(c(p1$se.fit, p2$se.fit), byrow=T, nrow=2)
   maxll = c(logLik(m1)[1], logLik(m2)[1])
}


### MATA interval
for (i.mu in 1:2) {     for (i.wt in 1:3) {

   mata = uniroot(wald.tailarea.t, c(-1e6, 1e6), muhats=muhat[,i.mu],
      semuhats=semuhat[,i.mu], dfs=t.dfs, weights=wts[,i.wt], alpha
      =0.025, tol=tol)$root

   mata = uniroot(wald.tailarea.t, c(-1e6, 1e6), muhats=muhat[,i.mu],
      semuhats=semuhat[,i.mu], dfs=t.dfs, weights=wts[,i.wt], alpha
      =0.975, tol=tol)$root}}


### Profile-Likelihood: log-likelihoods and transformation:
if(logn) {
   f1=ll1.logn
   f2=ll2.logn
   g=function(x) {exp(x)}
} else   {
   f1=ll1.n
   f2=ll2.n
   g=function(x) {x}}
```

```
### MPI interval

mpi.mu1 =  c(
  g(uniroot(proma.mu1, c(-1e6, 1e6), wts=wts[,i.wt], maxll=maxll, muhat=
      muhat, ctlist=ctlist, y=y, quant=0.975, f1=f1, f2=f2)$root),
  g(uniroot(proma.mu1, c(-1e6, 1e6), wts=wts[,i.wt], maxll=maxll, muhat=
      muhat, ctlist=ctlist, y=y, quant=0.025, f1=f1, f2=f2)$root))

mpi.mu2 = c(
  g(uniroot(proma.mu2, c(-1e6, 1e6), wts=wts[,i.wt], maxll=maxll, muhat=
      muhat, ctlist=ctlist, y=y, quant=0.975, f1=f1, f2=f2)$root),
  g(uniroot(proma.mu2, c(-1e6, 1e6), wts=wts[,i.wt], maxll=maxll, muhat=
      muhat, ctlist=ctlist, y=y, quant=0.025, f1=f1, f2=f2)$root))


### single-model PL intervals: pl1, pl2
pl1.mu1 = c(
  g(uniroot(pro1.mu,  c(-1e6, 1e6), maxll=maxll, muhat=muhat, ctlist=
      ctlist, y=y, quant=0.975, f1=f1, f2=f2, tol=tol)$root),
  g(uniroot(pro1.mu,  c(-1e6, 1e6), maxll=maxll, muhat=muhat, ctlist=
      ctlist, y=y, quant=0.025, f1=f1, f2=f2, tol=tol)$root))

pl1.mu2 = c(
  g(uniroot(pro1.mu,  c(-1e6, 1e6), maxll=maxll, muhat=muhat, ctlist=
      ctlist, y=y, quant=0.975, f1=f1, f2=f2, tol=tol)$root),
  g(uniroot(pro1.mu,  c(-1e6, 1e6), maxll=maxll, muhat=muhat, ctlist=
      ctlist, y=y, quant=0.025, f1=f1, f2=f2, tol=tol)$root))

pl2.mu1 = c(
  g(uniroot(pro2.mu1, c(-1e6, 1e6), maxll=maxll, muhat=muhat, ctlist=
      ctlist, y=y, quant=0.975, f1=f1, f2=f2, tol=tol)$root),
  g(uniroot(pro2.mu1, c(-1e6, 1e6), maxll=maxll, muhat=muhat, ctlist=
      ctlist, y=y, quant=0.025, f1=f1, f2=f2, tol=tol)$root))

pl2.mu12= c(
  g(uniroot(pro2.mu2, c(-1e6, 1e6), maxll=maxll, muhat=muhat, ctlist=
      ctlist, y=y, quant=0.975, f1=f1, f2=f2, tol=tol)$root),
  g(uniroot(pro2.mu2, c(-1e6, 1e6), maxll=maxll, muhat=muhat, ctlist=
      ctlist, y=y, quant=0.025, f1=f1, f2=f2, tol=tol)$root))
```

Appendix B

Supplementary Results

B.1 Normal Linear Regression Simulation: M_2 and M_3 Data Generation (Chapter 3)

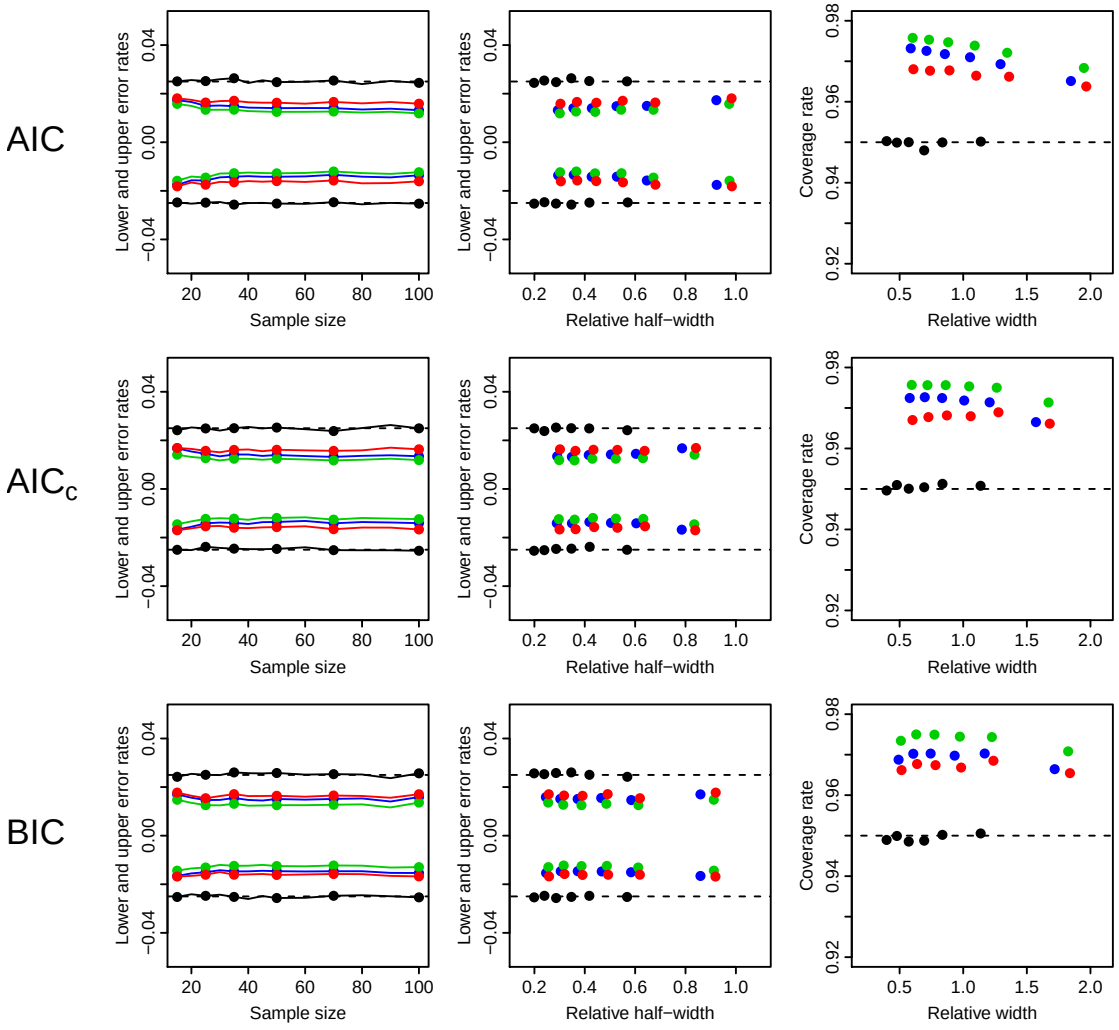

Figure B.1: Performance of the MATA-Wald (red), MAW_1 (blue), and MAW_2 (green) confidence intervals for prediction of the mean, at the 50% and 90% quantiles of the x_1 and x_2 generating distributions, respectively. The first, second, and third rows use AIC, AIC_c, and BIC weights. Nominal rates are shown as dashed lines. The data are generated under M_2, and the black points show the performance of the Wald interval based on this model.

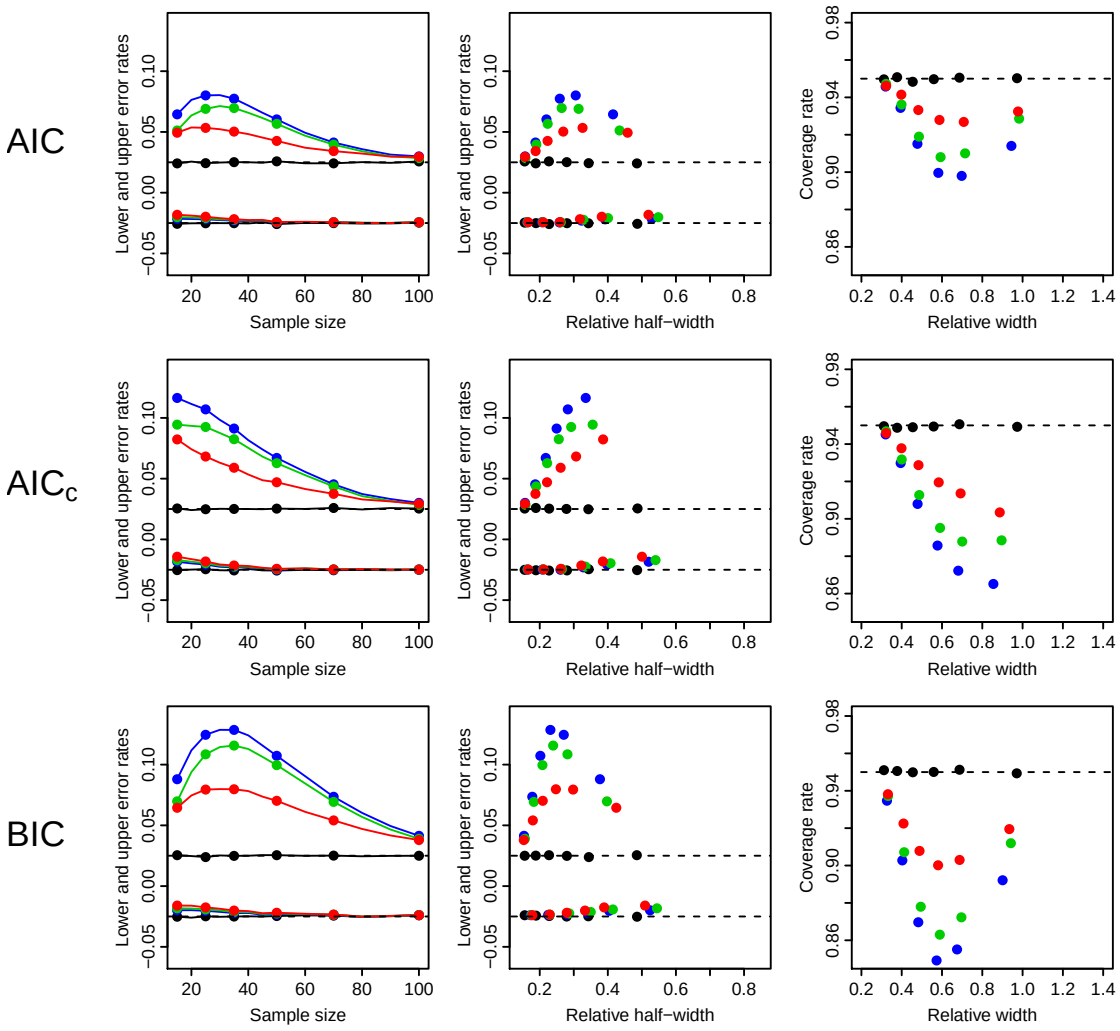

Figure B.2: Performance of the MATA-Wald (red), MAW_1 (blue), and MAW_2 (green) confidence intervals for prediction of the mean, at the 50% and 90% quantiles of the x_1 and x_2 generating distributions, respectively. The first, second, and third rows use AIC, AIC_c, and BIC weights. Nominal rates are shown as dashed lines. The data are generated under M_3, and the black points show the performance of the Wald interval based on this model.

B.2 Bayesian vs. Frequentist Simulation: AIC_c Weights (Chapter 4)

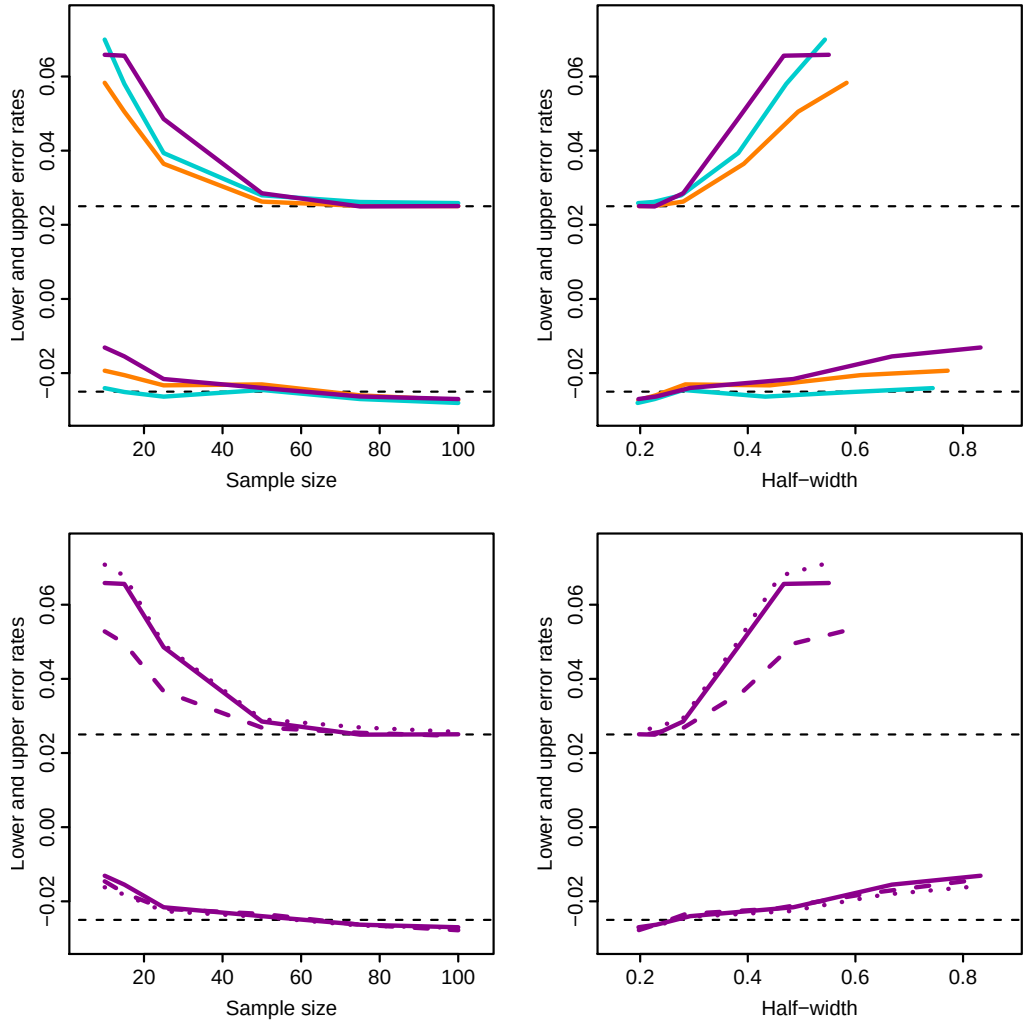

Figure B.3: Confidence interval performance for prediction of the mean, θ_2, in the normal linear simulation. Top row: MAB (purple), MATA-Wald (orange), and MATA-PL (blue) intervals. Bottom row: MAB (solid), MAB_J (dotted), and MAB_{KL} (dashed) Bayesian intervals. Nominal error rates are shown as dashed lines, and lower error rates are plotted as negative values. MATA-Wald and MATA-PL intervals are constructed using AIC_c weights.

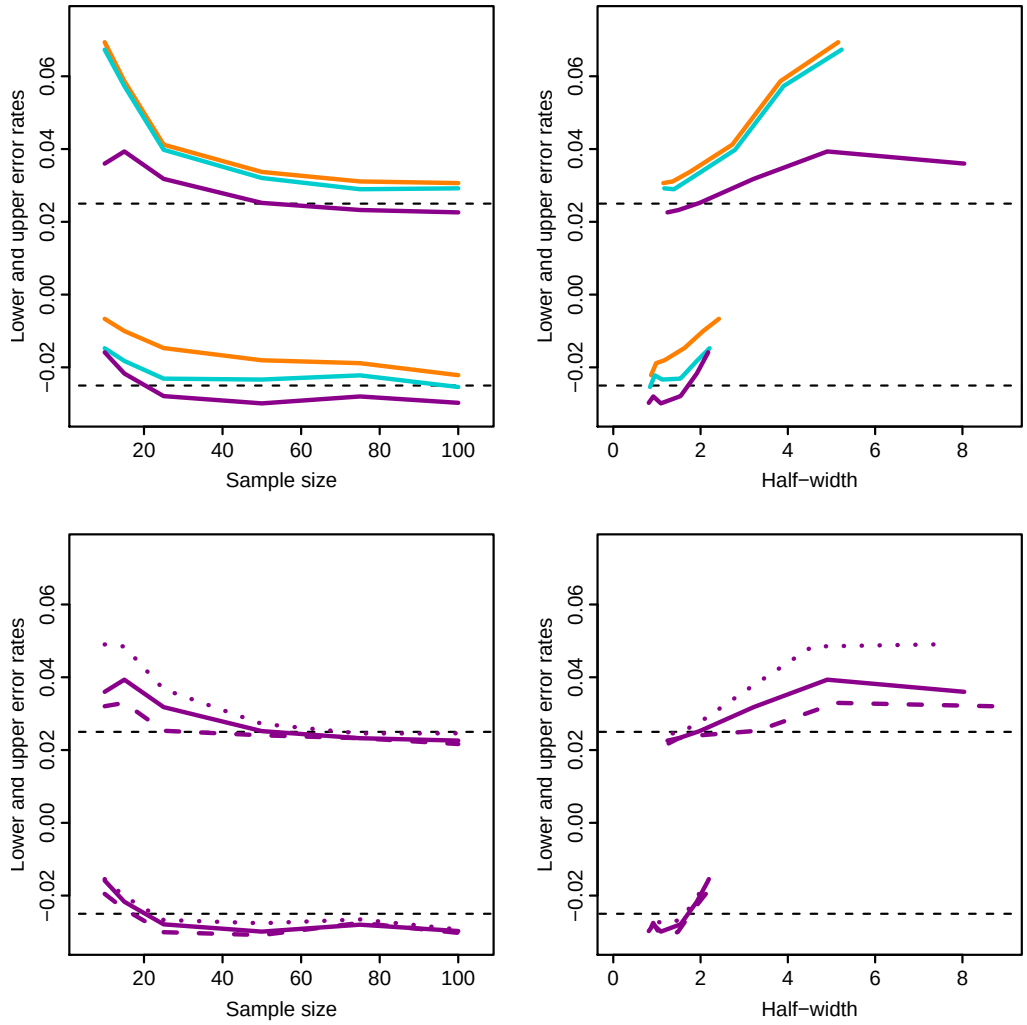

Figure B.4: Confidence interval performance for prediction of the mean, θ_2, in the lognormal simulation. Top row: MAB (purple), MATA-Wald (orange), and MATA-PL (blue) intervals. Bottom row: MAB (solid), MAB_J (dotted), and MAB_KL (dashed) Bayesian intervals. Nominal error rates are shown as dashed lines, and lower error rates are plotted as negative values. MATA-Wald and MATA-PL intervals are constructed using AIC_c weights.

B.3 Bayesian vs. Frequentist Simulation: BIC Weights (Chapter 4)

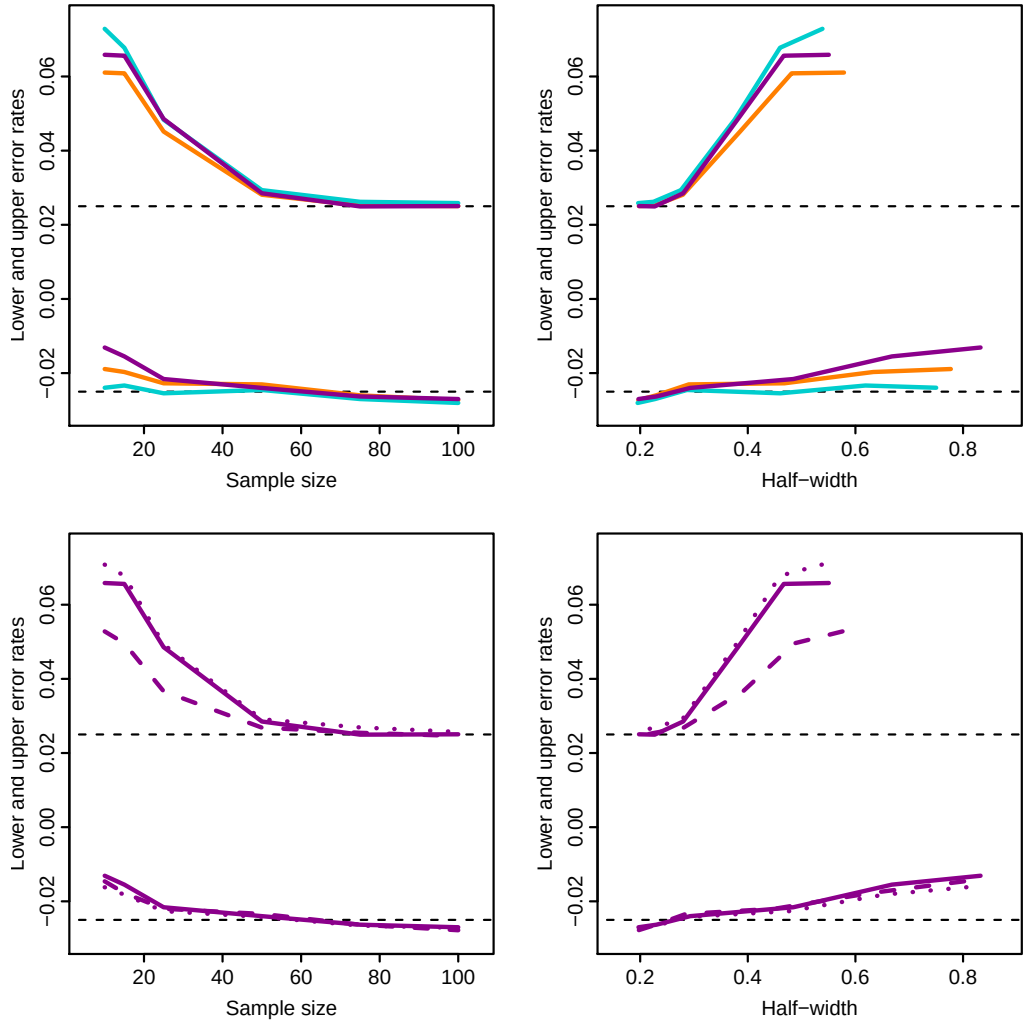

Figure B.5: Confidence interval performance for prediction of the mean, θ_2, in the normal linear simulation. Top row: MAB (purple), MATA-Wald (orange), and MATA-PL (blue) intervals. Bottom row: MAB (solid), MAB_J (dotted), and MAB_KL (dashed) Bayesian intervals. Nominal error rates are shown as dashed lines, and lower error rates are plotted as negative values. MATA-Wald and MATA-PL intervals are constructed using BIC weights.

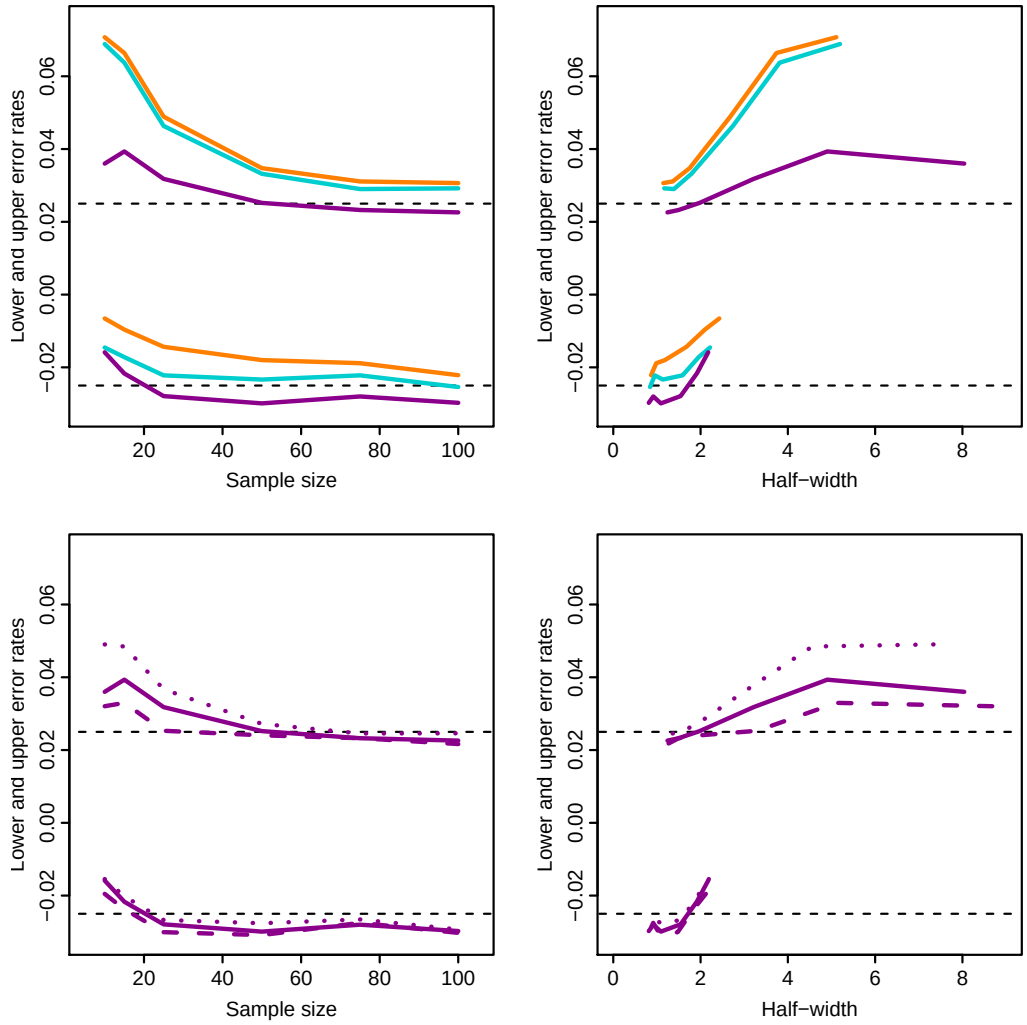

Figure B.6: Confidence interval performance for prediction of the mean, θ_2, in the lognormal simulation. Top row: MAB (purple), MATA-Wald (orange), and MATA-PL (blue) intervals. Bottom row: MAB (solid), MAB_J (dotted), and MAB_KL (dashed) Bayesian intervals. Nominal error rates are shown as dashed lines, and lower error rates are plotted as negative values. MATA-Wald and MATA-PL intervals are constructed using BIC weights.