Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2015, Article ID 420483, 9 pages
http://dx.doi.org/10.1155/2015/420483

Research Article

Comparison of the Frequentist MATA Confidence Interval with Bayesian Model-Averaged Confidence Intervals

Daniel Turek

Department of Mathematics and Statistics, University of Otago, Dunedin 9054, New Zealand

Correspondence should be addressed to Daniel Turek; dturek@maths.otago.ac.nz

Received 1 June 2015; Revised 22 July 2015; Accepted 13 September 2015

Academic Editor: Shesh N. Rai

Model averaging is a technique used to account for model uncertainty, in both Bayesian and frequentist multimodel inferences. In this paper, we compare the performance of model-averaged Bayesian credible intervals and frequentist confidence intervals. Frequentist intervals are constructed according to the model-averaged tail area (MATA) methodology. Differences between the Bayesian and frequentist methods are illustrated through an example involving cloud seeding. The coverage performance and interval width of each technique are then studied using simulation. A frequentist MATA interval performs best in the normal linear setting, while Bayesian credible intervals yield the best coverage performance in a lognormal setting. The use of a data-dependent prior probability for models improved the coverage of the model-averaged Bayesian interval, relative to that using uniform model prior probabilities. Data-dependent model prior probabilities are philosophically controversial in Bayesian statistics, and our results suggest that their use is beneficial when model averaging.

1. Introduction

Historically, statistical inference has been based on a single model selected from among a set of predetermined candidate models, with no allowance made for model uncertainty. This process of model selection has been shown to produce biased estimators and result in the incorrect calculation of standard error terms [1–4]. Recently, model averaging has gained popularity as a technique to incorporate model uncertainty into the process of inference [5–7]. The use of model averaging has been studied in a variety of settings (e.g., [8, 9]), where it generally exhibits favorable results relative to traditional model selection.

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. Posterior model probabilities are defined by the posterior distribution of the model parameter, and the posterior distributions for model parameters are not conditional upon a particular model and hence naturally account for model uncertainty [10, 11]. In practice, Bayesian model averaging is achieved by allowing a Gibbs sampler to traverse the augmented parameter space, which generates approximations to the posterior distributions of interest. Facilitated by recent advances in computation, Bayesian model averaging has been widely applied in a variety of application domains (e.g., [12–14]).

In the frequentist setting, a model-averaged estimate $\widehat{\theta}$ is defined as the weighted sum of single-model estimates: $\widehat{\theta} = \sum_{i=1}^{R} w_i \widehat{\theta}_i$, where $\widehat{\theta}_i$ is the estimate under model M_i, model weights w_i are determined from an information criterion such as AIC, and the summation is over the set of R candidate models.

Several approaches to constructing frequentist model-averaged confidence intervals have been suggested. Wald intervals of the form $\widehat{\theta} \pm z_\alpha \widehat{\text{se}}(\widehat{\theta})$, where z_α is the $(1 - \alpha)$ quantile of the standard normal distribution, rely on accurate estimation of $\text{se}(\widehat{\theta})$, the standard error of $\widehat{\theta}$. Estimation of this term is complicated by the fact that the model weights and the single-model estimates are all random quantities. Burnham and Anderson [6] have suggested a variety of forms for $\widehat{\text{se}}(\widehat{\theta})$, which are studied by Claeskens and Hjort [7] and by Turek

and Fletcher [15]. In each of these studies, model-averaged Wald intervals of this form were found to perform poorly in terms of coverage rate.

An alternate methodology for the construction of frequentist model-averaged intervals is proposed by Turek and Fletcher [15]. 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. As this involves averaging the "tail areas" of the sampling distributions of single-model estimates, this new construction is called a model-averaged tail area Wald (MATA-Wald) interval. In a simulation study by Turek and Fletcher [15], the MATA-Wald interval outperformed model-averaged intervals of the form $\widehat{\theta} \pm z_\alpha \widehat{\text{se}}(\widehat{\theta})$. Fletcher and Turek [16] applied the MATA construction to profile likelihood intervals to produce a model-averaged tail area profile likelihood (MATA-PL) interval. Coverage properties of MATA confidence intervals are also studied in Kabaila et al. [17], and a transformed version of the MATA interval was proposed by Yu et al. [18].

In this paper, we compare the performance of model-averaged Bayesian credible intervals and the MATA-Wald and MATA-PL intervals of Turek and Fletcher [15] and Fletcher and Turek [16]. 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 Section 2, we define the Bayesian and frequentist model-averaged intervals. The differences between these intervals are shown in Section 3, through an example involving cloud seeding. We describe the simulation study used to compare these intervals in Section 4 and present the results of this study in Section 5. We conclude with a discussion in Section 6.

2. Model-Averaged Intervals

Assume a set of R candidate models $\{M_i\}$ exists, 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. We now define the Bayesian and frequentist model-averaged intervals for θ.

2.1. Bayesian Interval. The model-averaged posterior distribution for θ is

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

where $p(\theta \mid M_i, y)$ is the posterior distribution of θ under model M_i and $p(M_i \mid y)$ is the posterior probability of M_i [10]. An equal-tailed $(1 - 2\alpha)100\%$ model-averaged Bayesian (MAB) credible interval is defined as the α and $(1 - \alpha)$ quantiles of $p(\theta \mid y)$.

Each posterior distribution $p(\theta \mid M_i, y)$ in (1) may be expressed through integration of the joint posterior, as

$$\begin{aligned} p(\theta \mid M_i, y) &= \int p(\theta, \lambda_i \mid M_i, y)\, d\lambda_i \\ &\propto \int L_i(\theta, \lambda_i)\, p(\theta, \lambda_i \mid M_i)\, d\lambda_i, \end{aligned} \tag{2}$$

following Bayes' theorem, where $p(\theta, \lambda_i \mid M_i)$ is the joint prior distribution for parameters θ and λ_i under M_i. The posterior model probabilities in (1) may be expressed as $p(M_i \mid y) \propto p(y \mid M_i)p(M_i)$, where $p(M_i)$ is the prior probability of model M_i and $p(y \mid M_i)$ is the integrated likelihood under M_i, given by

$$p(y \mid M_i) = \iint L_i(\theta, \lambda_i)\, p(\theta, \lambda_i \mid M_i)\, d\theta\, d\lambda_i. \tag{3}$$

Evaluation of the integrals in (2) and (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 [19].

2.2. Frequentist Interval. The frequentist MATA 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 $\text{AIC} = -2\log\widehat{L} + 2p$ to define model weights, where $\widehat{L}$ 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 $\Delta\text{AIC}_i \equiv \text{AIC}_i - \min_{j=1,\ldots,R}(\text{AIC}_j)$ and AIC_i is the value of the information criterion for model M_i [20]. Other choices of information criteria for defining model weights are addressed in the discussion in Section 6.

2.2.1. MATA-Wald Interval. In the normal linear model, the confidence limits θ_L and θ_U of a single-model $(1 - 2\alpha)100\%$ Wald interval for θ satisfy the equations

$$\begin{aligned} 1 - F_\nu(t_L) &= \alpha, \\ F_\nu(t_U) &= \alpha, \end{aligned} \tag{4}$$

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, $t_L = (\widehat{\theta} - \theta_L)/\widehat{\text{se}}(\widehat{\theta})$, $t_U = (\widehat{\theta} - \theta_U)/\widehat{\text{se}}(\widehat{\theta})$, and $\widehat{\text{se}}(\widehat{\theta})$ is the estimated standard error of $\widehat{\theta}$ [21, 22]. 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

$$\sum_{i=1}^{R} w_i \left(1 - F_{\nu_i}\left(t_{L,i}\right)\right) = \alpha,$$
$$\sum_{i=1}^{R} w_i F_{\nu_i}\left(t_{U,i}\right) = \alpha, \tag{5}$$

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

The MATA-Wald interval may be generalized to non-normal data, assuming that we can specify a transformation $\phi = g(\theta)$ for which the sampling distribution of $\widehat{\phi}_i = g(\widehat{\theta}_i)$ 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 confidence limits θ_L and θ_U are the values satisfying the pair of equations

$$\sum_{i=1}^{R} w_i \left(1 - \Phi\left(z_{L,i}\right)\right) = \alpha,$$
$$\sum_{i=1}^{R} w_i \Phi\left(z_{U,i}\right) = \alpha, \tag{6}$$

where $\Phi(\cdot)$ is the standard normal distribution function, $z_{L,i} = (\widehat{\phi}_i - \phi_L)/\widehat{\text{se}}(\widehat{\phi}_i)$, $z_{U,i} = (\widehat{\phi}_i - \phi_U)/\widehat{\text{se}}(\widehat{\phi}_i)$, $\phi_L = g(\theta_L)$, and $\phi_U = g(\theta_U)$, as set out by Turek and Fletcher [15].

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

$$\Phi\left(r\left(\theta_L\right)\right) = \alpha,$$
$$1 - \Phi\left(r\left(\theta_U\right)\right) = \alpha, \tag{7}$$

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

$$r(\theta) = \text{sign}\left(\widehat{\theta} - \theta\right)\sqrt{2\left(\log L_p\left(\widehat{\theta}\right) - \log L_p(\theta)\right)}, \tag{8}$$

and $L_p(\theta) = \max_\lambda L(\theta, \lambda)$ is the profile likelihood function for θ [23, p. 126–129]. The limits θ_L and θ_U of the MATA-PL interval are defined as the values which satisfy

$$\sum_{i=1}^{R} w_i \Phi\left(r_i\left(\theta_L\right)\right) = \alpha,$$
$$\sum_{i=1}^{R} w_i \left(1 - \Phi\left(r_i\left(\theta_U\right)\right)\right) = \alpha, \tag{9}$$

where $r_i(\theta)$ is defined in terms of the corresponding likelihood function $L_i(\theta, \lambda_i)$, as in (8), and as described by Fletcher and Turek [16].

Table 1: Rain volume data recorded in 1968 and 1970 cloud seeding experiments. All clouds are stationary and are categorized as seeded or unseeded. Rain volume is measured in thousands of cubic meters (10^3 m^3).

Seeded clouds	Unseeded clouds
Rain volume	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	

3. Example

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 [24–26]. 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 [27], and the subset relevant to our analysis is presented in Table 1.

Suppose that we aim to predict the expected rainfall from seeded, stationary clouds. The lognormal distribution can provide a good model for total rain volume [27]. 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, $\theta_S \equiv E[R_S] = \exp(\beta_S + \sigma^2/2)$, and we 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 [28], and an equal prior probability for each model. We ran an MCMC

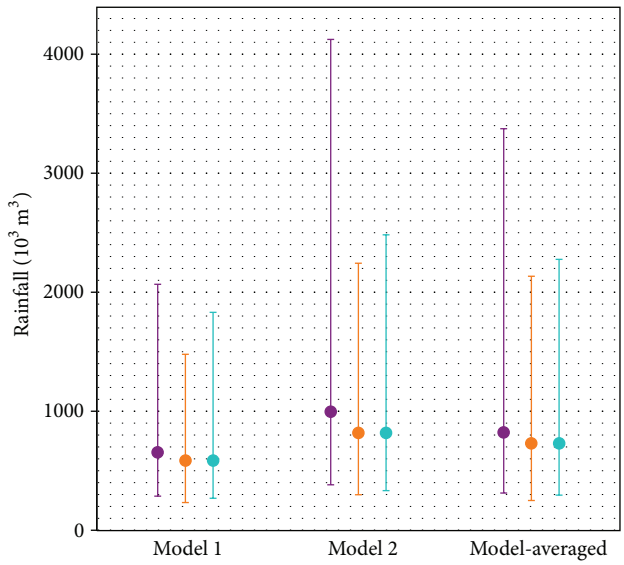

Figure 1: Expected mean rainfall for seeded, stationary clouds, under each model and using model averaging. Vertical bars show 95% intervals for each prediction. Intervals shown: Bayesian and MAB (purple), Wald and MATA-Wald (orange), and profile likelihood and MATA-PL (blue).

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 [29, 30]. This indicated convergence for each model, with all BGR values being 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 $\log\theta_S - \sigma^2/2$ in place of β_S and $\log\theta_U - \sigma^2/2$ in place of β_U. The MATA-Wald interval was constructed according to (6) and the MATA-PL interval following (9), both of which used AIC weights for w_i.

The resulting Bayesian posterior model probabilities were $p(M_1 \mid y) = p(M_2 \mid y) = 0.50$, 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 1 shows the predicted mean rain volume $\widehat{\theta}_S$ 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 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 model-averaged 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. It is important to realize the generalizability of the analysis presented here. The same approach is equally applicable to any data analysis situation in which there is model uncertainty, meaning that the true, underlying data-generating model is unknown.

4. Simulation Study

Based on the example in Section 3, 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 $\theta_i \equiv E[Y_{ij}]$, for $i = 1, 2$. In the lognormal case, $\theta_i = \exp(\beta_i + \sigma^2/2)$, so the likelihood was again reparameterized using $\log\theta_i - \sigma^2/2$ 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 assessed by the actual coverage rate achieved, defined as the proportion of simulations for which $\theta \in [\theta_L, \theta_U]$. We averaged results over 20,000 simulations, ensuring a standard error for the coverage rate less than 0.3%. In addition, we calculated the mean interval width of each method, defined as $\theta_U - \theta_L$. All calculations were performed in R, version 2.13.0 [31].

4.1. Bayesian Implementation. Three sets of prior probabilities were considered, for the construction of three distinct model-averaged Bayesian intervals. The first Bayesian interval (MAB) 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 by Gelman [28]. The second interval (MAB_J) used equal model prior probabilities and improper Jeffreys' prior distributions [32] for the parameters: $p(\beta_i) \propto 1$ and $p(\sigma) \propto 1/\sigma$ (see, e.g., [33]). The third interval (MAB_KL) used flat prior distributions for the parameters and the Kullback-Leibler (KL) prior probability for each model, defined as

$$p(M_i) \propto \exp\left(p_i\left(\frac{1}{2}\log n - 1\right)\right), \tag{10}$$

where p_i is the number of parameters in model M_i [6, 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 burn-in.

4.2. Frequentist Implementation. Frequentist model-averaged intervals were constructed using AIC weights, as defined in Section 2.2. For the normal linear simulation, the MATA-Wald interval was constructed using (5) and the lognormal MATA-Wald interval following (6). The MATA-PL interval was defined according to (9), using the reparameterized likelihood in the lognormal case. Numerical solutions to these equations were found using the R root-finding command *uniroot*.

5. 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 2(a) shows the estimated coverage rate for the MATA-Wald, MATA-PL, and MAB intervals. The MATA-Wald interval performs best, in particular for small sample sizes, followed by the MATA-PL and MAB intervals. All intervals asymptotically approach the nominal coverage rate of 95%. 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, coverage is also plotted against mean interval width. For comparable interval width, the MATA-Wald interval achieves notably improved performance.

Figure 2(b) provides the same comparison for the Bayesian MAB, MAB_J, and MAB_KL intervals. The MAB_KL interval provides a noticeable improvement in coverage performance, 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 coverage rate for small sample sizes. This improvement comes at no noticeable increase in interval 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 3 provides analogous comparisons in the lognormal setting. The MAB interval outperforms the frequentist intervals for small sample sizes, although it requires a substantial increase in interval width. The MAB interval remains roughly within 1% of the nominal coverage rate for all sample sizes, while the frequentist intervals deviate by as much as 3%. The MATA-Wald interval performs better than the MATA-PL interval, with both exhibiting comparable interval widths.

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 an 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.

6. Discussion

The aim of this paper 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 constructing an interval around a model-averaged estimator. This construction is analogous to Bayesian model averaging, and the idea was initially motivated using an analogy to a model-averaged Bayesian interval [16]. The MATA construction was studied further in Turek and Fletcher's work [15], where it is shown that, asymptotically, a MATA 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 fully allows for parameter uncertainty and does not rely on the asymptotic distributions of estimators. 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 nonnormal 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 [6] 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 [34] argue that for large sample sizes the data ought to completely dominate the priors, and the use of prior probabilities which depend upon

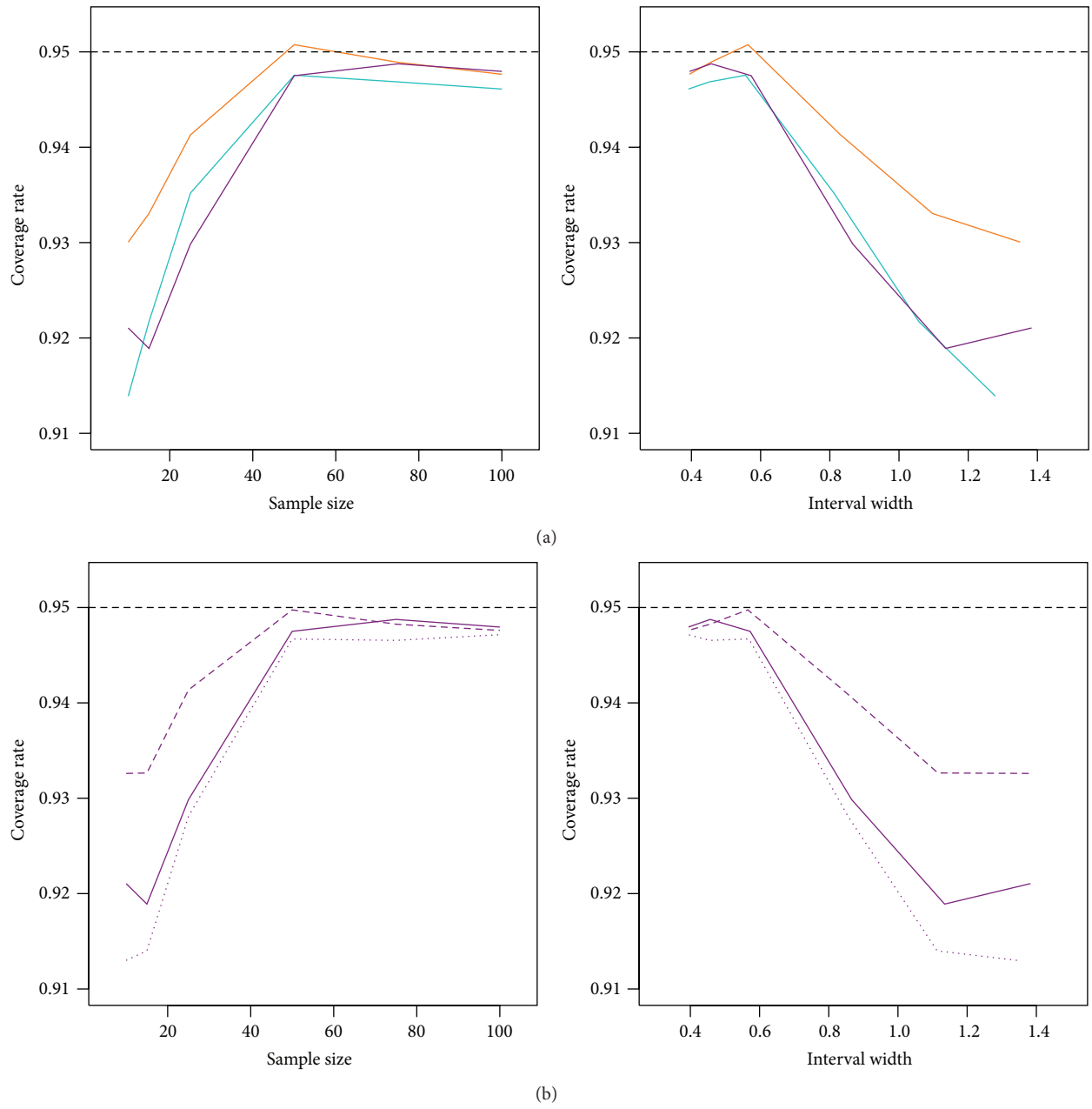

Figure 2: Confidence interval coverage rate performance for prediction of the mean, θ_2, in the normal linear simulation. (a) MAB (purple), MATA-Wald (orange), and MATA-PL (blue) intervals. (b) MAB (solid), MAB_J (dotted), and MAB_KL (dashed) Bayesian intervals. Nominal 95% coverage rate is shown as a dashed line.

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 [33], where a prior 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 MATA intervals constructed using AIC model weights. Two alternate information criteria were also considered: AIC_c [35] and BIC [36]. AIC_c was derived as a small-sample correction to AIC and in certain contexts may be favorable for use in model selection [37]. 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 [34].

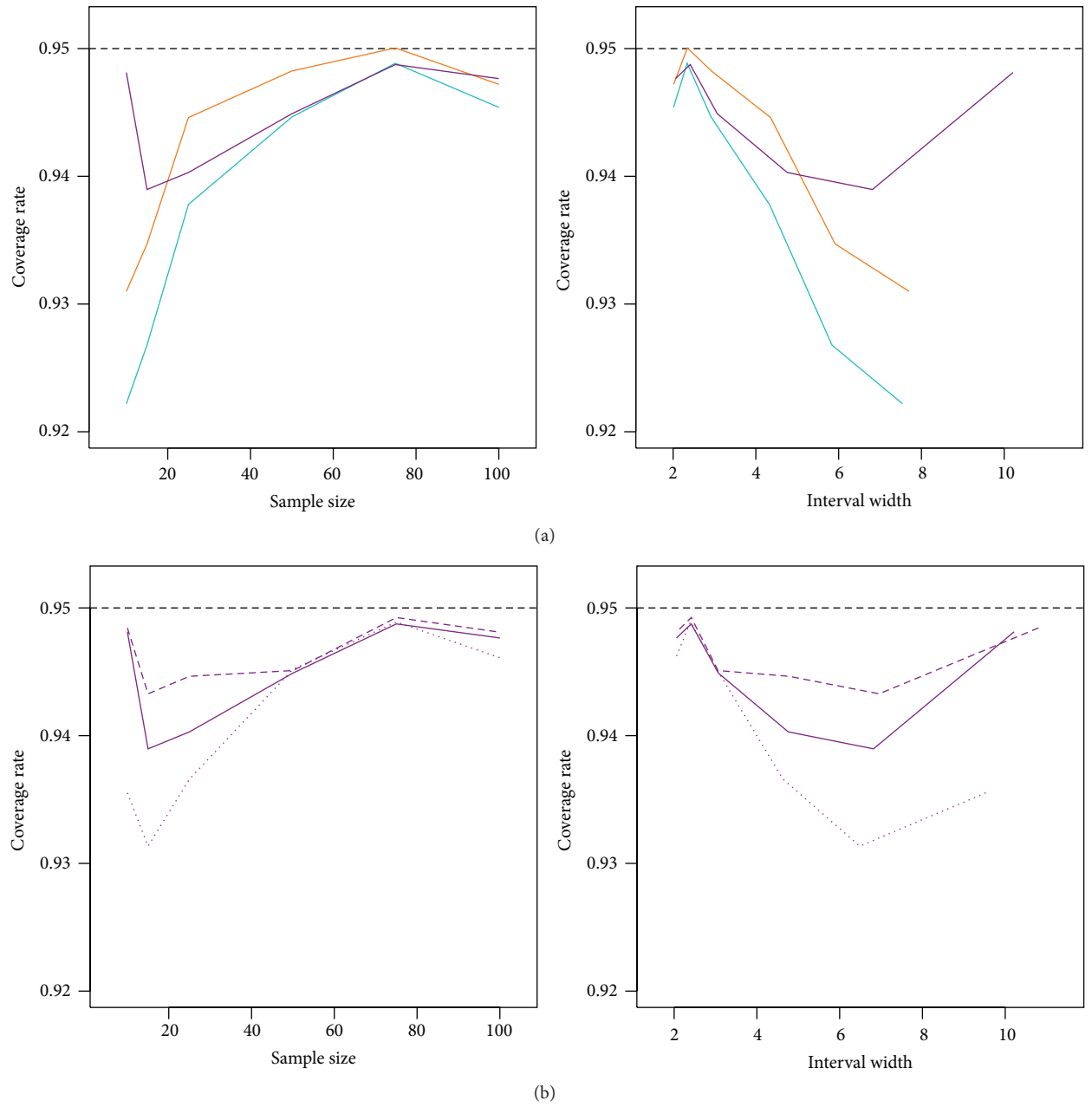

FIGURE 3: Confidence interval coverage rate performance for prediction of the mean, θ_2, in the lognormal simulation. (a) MAB (purple), MATA-Wald (orange), and MATA-PL (blue) intervals. (b) MAB (solid), MAB_J (dotted), and MAB_{KL} (dashed) Bayesian intervals. Nominal 95% coverage rate is shown as a dashed line.

In our study, the MATA 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. This result is consistent with the findings of Fletcher and Dillingham [38], 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 assumption is also used in the derivations of 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 [6] believe that nature is arbitrarily complex, and it is unrealistic to assume that we might fully characterize the underlying process. From this viewpoint, model selection attempts to identify the most parsimonious approximating model to truth, given the data available. This assumption may in part explain the superior performance of AIC model weights, since AIC is known to favor increased model complexity [39]. However we do not consider this an issue, since results from Fletcher and Turek [16] 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. Furthermore, all simulations presented herein were repeated using data generated under the simpler of the two candidate models (M_1). In these simulations all model-averaged constructions performed similar to one another and achieved very near to the nominal coverage rates. This would be expected, since model averaging takes place over two models, both of which now represent truth.

Any simulation study is inherently limited in scope. We have considered both normal and nonnormal data, as well as a wide range of sample sizes, and observed consistent patterns throughout. Bayesian model averaging was better suited for the nonnormal 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 Bayesian priors represent; in particular, whether or not knowledge of the *size* of an observed sample provides grounds to update model prior probabilities.

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

References

[1] C. M. Hurvich and C. Tsai, "The impact of model selection on inference in linear regression," *The American Statistician*, vol. 44, no. 3, pp. 214–217, 1990.

[2] C. Chatfield, "Model uncertainty, data mining and statistical inference," *Journal of the Royal Statistical Society Series A (Statistics in Society)*, vol. 158, no. 3, pp. 419–466, 1995.

[3] D. Draper, "Assessment and propagation of model uncertainty," *Journal of the Royal Statistical Society. Series B. Methodological*, vol. 57, no. 1, pp. 45–97, 1995.

[4] P. Kabaila, "The effect of model selection on confidence regions and prediction regions," *Econometric Theory*, vol. 11, no. 3, pp. 537–549, 1995.

[5] A. E. Raftery, D. Madigan, and J. Hoeting, "Bayesian model averaging for linear regression models," *Journal of the American Statistical Association*, vol. 92, no. 437, pp. 179–191, 1997.

[6] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer, New York, NY, USA, 2002.

[7] G. Claeskens and N. L. Hjort, *Model Selection and Model Averaging*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, UK, 2008.

[8] V. Viallefont, A. E. Raftery, and S. Richardson, "Variable selection and Bayesian model averaging in case-control studies," *Statistics in Medicine*, vol. 20, no. 21, pp. 3215–3230, 2001.

[9] M. Pesaran and P. Zaffaroni, "Model averaging and value-at-risk based evaluation of large multi asset volatility models for risk management," CESifo Working Paper 1358, 2004.

[10] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: a tutorial," *Statistical Science*, vol. 14, no. 4, pp. 382–417, 1999.

[11] L. Wasserman, "Bayesian model selection and model averaging," *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 92–107, 2000.

[12] C. T. Volinsky, D. Madigan, A. E. Raftery, and R. A. Kronmal, "Bayesian model averaging in proportional hazard models: assessing the risk of a stroke," *Journal of the Royal Statistical Society C: Applied Statistics*, vol. 46, no. 4, pp. 433–448, 1997.

[13] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski, "Using Bayesian model averaging to calibrate forecast ensembles," *Monthly Weather Review*, vol. 133, no. 5, pp. 1155–1174, 2005.

[14] Q. Duan, N. K. Ajami, X. Gao, and S. Sorooshian, "Multi-model ensemble hydrologic prediction using Bayesian model averaging," *Advances in Water Resources*, vol. 30, no. 5, pp. 1371–1386, 2007.

[15] D. Turek and D. Fletcher, "Model-averaged Wald confidence intervals," *Computational Statistics & Data Analysis*, vol. 56, no. 9, pp. 2809–2815, 2012.

[16] D. Fletcher and D. Turek, "Model-averaged profile likelihood intervals," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 17, no. 1, pp. 38–51, 2012.

[17] P. Kabaila, A. Welsh, and W. Abeysekera, "Model-averaged confidence intervals," *Scandinavian Journal of Statistics*, 2015.

[18] W. Yu, W. Xu, and L. Zhu, "Transformation-based model averaged tail area inference," *Computational Statistics*, vol. 29, no. 6, pp. 1713–1726, 2014.

[19] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.

[20] S. T. Buckland, K. P. Burnham, and N. H. Augustin, "Model selection: an integral part of inference," *Biometrics*, vol. 53, no. 2, pp. 603–618, 1997.

[21] A. Wald, "The fitting of straight lines if both variables are subject to error," *The Annals of Mathematical Statistics*, vol. 11, no. 3, pp. 285–300, 1940.

[22] C. Stein and A. Wald, "Sequential confidence intervals for the mean of a normal distribution with known variance," *The Annals of Mathematical Statistics*, vol. 18, pp. 427–433, 1947.

[23] A. C. Davison, *Statistical Models*, Cambridge University Press, Cambridge, UK, 2003.

[24] J. Simpson, W. L. Woodley, A. H. Miller, and G. F. Cotton, "Precipitation results of two randomized pyrotechnic cumulus seeding experiments," *Journal of Applied Meteorology*, vol. 10, no. 3, pp. 526–544, 1971.

[25] J. Simpson, "Use of the gamma distribution in single-cloud rainfall analysis," *Monthly Weather Review*, vol. 100, no. 4, pp. 309–312, 1972.

[26] D. Rosenfeld and W. L. Woodley, "Effects of cloud seeding in west Texas: additional results and new insights," *Journal of Applied Meteorology*, vol. 32, no. 12, pp. 1848–1866, 1993.

[27] R. Biondini, "Cloud motion and rainfall statistics," *Journal of Applied Meteorology*, vol. 15, no. 3, pp. 205–224, 1976.

[28] A. Gelman, "Prior distributions for variance parameters in hierarchical models," *Bayesian Analysis*, vol. 1, no. 3, pp. 515–533, 2006.

[29] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical Science*, vol. 7, no. 4, pp. 457–472, 1992.

[30] S. P. Brooks and A. Gelman, "General methods for monitoring convergence of iterative simulations," *Journal of Computational and Graphical Statistics*, vol. 7, no. 4, pp. 434–455, 1998.

[31] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014.

[32] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society. London. Series A. Mathematical, Physical and Engineering Sciences*, vol. 186, pp. 453–461, 1946.

[33] G. Box and G. Tiao, *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Mass, USA, 1973.

[34] W. A. Link and R. J. Barker, "Model weights and the foundations of multimodel inference," *Ecology*, vol. 87, no. 10, pp. 2626–2635, 2006.

[35] N. Sugiura, "Further analysts of the data by akaike's information criterion and the finite corrections," *Communications in Statistics—Theory and Methods*, vol. 7, no. 1, pp. 13–26, 2007.

[36] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[37] C. M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307, 1989.

[38] D. Fletcher and P. W. Dillingham, "Model-averaged confidence intervals for factorial experiments," *Computational Statistics & Data Analysis*, vol. 55, no. 11, pp. 3041–3048, 2011.

[39] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.