Bootstrapped Information-Theoretic Model Selection with Error Control (BITSEC)

by

Michael J. Cullan

A Thesis Presented in Partial Fullfilment
of the Requirements for the Degree
Master of Science

To be approved September 2018 by the
Graduate Supervisory Committee

Beckett Sterner, Chair
John Fricks
Ming-Hung Kao

ARIZONA STATE UNIVERSITY

December 2018

ACKNOWLEDGMENTS

This paper would not have been possible without the support of Dr. Beckett Sterner, an outstanding advisor and mentor whom it has been a true honor to work with.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Statistical model selection using the Akaike Information Criterion (AIC) and similar criteria is a useful tool for comparing multiple and non-nested models without the specification of a null model, which has made it increasingly popular in the natural and social sciences. Despite their common usage, model selection methods are not driven by a notion of statistical confidence, so their results entail an unknown degree of uncertainty. This paper introduces a general framework which extends notions of Type-I and Type-II error to model selection. A theoretical method for controlling Type-I error using Difference of Goodness of Fit (DGOF) distributions is given, along with a bootstrap approach that approximates the procedure. Results are presented for simulated experiments using normal distributions, random walk models, nested linear regression, and nonnested regression including nonlinear models. Tests are performed using an R package developed by the author which will be made publicly available on journal publication of research results.

1 Introduction

The merits of hypothesis testing are many; it is well suited to making inferences about associative, and in certain situations causal, relationships between observable quantities. The use of known distributions in testing allows for analytic solutions for the construction of test rejection regions with desirable, provable properties. A class of tests can be constructed with a specified rate of Type-I error, the probability of obtaining false positives, and among such a class, tests can be often be identified with optimized power, the probability of obtaining true positives. However, the rich and rigorous underpinnings of hypothesis testing come at a cost. Though well-developed and widely effective, these methods are limited in the types of conclusions they can draw and the types of questions they can address (Spanos 2010, Burnham and Anderson 2002).

Hypothesis testing uses known distributions to make binary decisions between rejecting null and alternate hypotheses with a quantifiable degree of certainty. As such, test results are clear and commini-

cable, but their implications are less so. Hypothesis testing cannot prove that a distribution speficied by the alternative hypothesis is true; to do so would be to commit a"fallacy of acceptance" (Spanos 2010). The constructs of null and alternative hypotheses also do not permit comparisons between two disparate, plausible models. This is limiting when investigating phenomena with competing explanations. The ability to evaluate such structural explanations, and indeed to make claims about their validity, is of central importance in fields characterized by quantitative theory and noisy data, such as biology, psychology, and sociology (Ullman and Bentler 2012). These shortcomings have motivated the devlopment and adoption of new statistical paradigms such as information-theoretic model selection.

There are certainly cases in which statisticans have been able to analytically compute likelihood ratio distributions for non-nested models, but this is not possible in general. As such, even these tractable cases provide only a limited and specialized extension to the families and pairs of distributions that characterize most of hypothesis testing. It was not until the development of the Akaike Information Criterion (AIC) that

disparate models could be compared easily and generally (Akaike 1974, Shmueli 2010). Though derived as an estimator for a given model's "distance" from a true data generating process (K-L Divergence), the AIC can be seen as a modified likelihood function that penalizes models for their intrinsic complexity (Burnham and Anderson 2010). The AIC is widely used for selecting and constructing predictive models because its basis in information theory makes it effective at minimizing predictive loss (Shmueli 2010). It can be used to construct weighted averages of models which can outperform their individual components (Burnham and Anderson 2002, Posada and Buckely 2004).

In the context of predicting new data based on past observations, the AIC has considerable justification and proven performance (Akaike 1974). In identifying a true model from a set of candidates, it lacks such justification. It is shown that even as sample size approaches infinity, the the probability that the AIC identifies a true model, assuming the true model is in the set of candidates, does not converge to 1 in general (Bozdogan 1987). Moreover, the AIC cannot be used to make decisions based on allowable degrees of error, the very foundation

of classical hypothesis testing. In spite of these facts, the objections of some statisticians, and indeed the recommendations of those who popularized its use (e.g. Burnham and Anderson 2002), the AIC is often employed with the aim of selecting a single, correct (i.e. "true") model (Spanos 2010, Shmueli 2010, Bozdogan 1989). While alternatives to the AIC have been developed, such as the Bayesian Information Criterion (BIC), which does consistently identify true models (Schwarz 1978), they still do not enable testing at specified error levels.

Information-theoretic methods remedy some of the limitations of classical hypothesis testing at the expense of control over error rates. In light of these discrepancies, this paper introduces a framework for model selection with error control as well as an easy to implement computational method which selects models with approximate error bounds. In doing so, this paper aims to reconcile the evident needs of statistical practitioners with the valid complaints raised against model selection.

2 Background

2.1 Classical Hypothesis Testing

The origins of "classical" testing in use today can be traced back to (Fisher 1924) on "significance testing," and a more generalized formulation by (Neyman and Pearson 1933) on hypothesis testing.

(Fisher 1925) introduced the notion of statistical significance and its use in decision-making. The text does not provide a concise, formal definition of significance, but uses the term widely when describing decision-making processes which use as evidence the probability of results being generated by a null distribution. This formulation of testing is concerned primarily with testing the differences of sample means and thus employs the t and z statistics stemming from the t and normal distributions.

The test statistic is then used to generate p-values, given for example by $p = P(t > t^*|H_0)$, the probability that a random sample from the statistic's associated distribution would be greater than (similarly, less than or with absolute value greater than) the test statistic. These values

are used in two ways in the text. (Fisher 1924) suggests that p-values can be compared to certain benchmarks, and gives as an example the familiar 0.05 "significance level," though it is noted that these values should be selected depending on the context of an experiment. In addition to this type of decision-making, p-values are used by Fisher to report degrees of evidence against the null hypothesis.

Neyman and Pearson (1933) introduced a refinement of this testing procedure. Neyman-Pearson-style testing aimed to formalize Fisher's tests of significance in a more mathematically rigorous manner, expanding the types of tests that can be performed while also restricting the ways in which results are meant to be interpreted. Neyman and Pearson also introduced the "alternative hypothesis," and from the two-hypothesis formulation they introduced the concepts of Type-I and Type-II error as well as statistical power. The fundamentals of Neyman-Pearson testing are encoded in the Neyman Pearson lemma, stated below.

Neyman-Pearson Lemma: Assume X is a random sample $X = x_1, x_2, \ldots, x_n$ generated by a distribution parametrized by Θ. Then, construct a likelihood ratio test with hypotheses

$$\mathrm{H}_0 : \Theta = \Theta_0$$

$$\mathrm{H}_\alpha : \Theta = \Theta_\alpha$$

and rejection region C determined by the critical value k defined by some $\alpha \in [0, 1]$, i.e. k is a constant for which

$$\frac{\mathcal{L}(\theta_0|X)}{\mathcal{L}(\theta_\alpha|X)} \leq k \text{ inside } C \text{ and}$$
$$P(\frac{\mathcal{L}(\theta_0|, X)}{\mathcal{L}(\theta_\alpha|X)} \leq k) = \alpha \text{ for } \theta = \theta_0$$

Then, the Neyman-Pearson Lemma concludes that the test determined by the rejection region C is the most powerful test for this set of hypotheses.

While the Neyman-Pearson lemma applies to a specific test statistic and set of hypotheses, it was the foundation for the goal of constructing tests to attain maximum power for a given α, and more broadly

for emphasizing the relationship between error and power in testing. Based on this emphasis, one can see how simply reporting p-values as quantifiers of evidence can be seen as an abuse of testing (Wagenmakers 2007). If a researcher desired to report results for a test with an α level of 0.01 for instance, they should not, according to Neyman-Pearson principles, simultaneously allow for results using an α value of 0.05. The more rigorous results stemming from a test with $\alpha = 0.01$ depend not on the observed p-value, which is itself a random variable, but on the underlying probability that the test should be accepted or rejected. A p-value below 0.01 has different implications for $\alpha = 0.01$ and $\alpha = 0.05$, because the primary result according to this method of testing should be the decision between rejecting and failing to reject the null hypothesis.

The misuse of test results is one of several complaints raised against hypothesis testing. The process is subject to both the "fallacy of rejection" and "fallacy of acceptance," the former of which states that failure to reject the null hypothesis can be misinterpreted as evidence in support of it (Spanos 2010). Perhaps a more nuanced point is that the

fallacy of acceptance describes how rejecting the null hypothesis can be conflated with accepting the alternative hypothesis as true. This is especially problematic for experiments comparing explanatory models for real-world processes. Comparing explanatory models is common practice in fields such as psychology and sociology, and in these fields is referred to as Structural Equation Modeling (Hox and Bechger 1998, Ullman and Bentler 2012). An NP test is useful for detecting the presence of an effect between treatments, but is not meant for conclusions that the alternative hypothesis defines a model that contains the underlying truth.

Hypothesis tests also limit the types of statistical tests that can be performed and the types of conclusions that can be drawn. Testing between nested null and alternative models means that a practitioner cannot easily compare the relative evidence for two disparate explanatory models. Moreover, there are some cases in which a null model describes a result which would be qualitatively important. One such example in paleobiological studies of evolutionary rates and directions is noted in (Hunt 2015). In this line of research, practitioners are particularly in-

terested in whether fossil data exhibits evidence for evolutionary stasis, or if it resembles a directed or undirected random walk. Hunt notes studies which posit as null models either stasis or undirected walks. In either of these cases, the null models describe results which would be biologically relevant to the field.

The specification of scientifically relevant explanations as null models has led to a misuse of the hypothesis testing framework in biology and psychology (Bausman 2018). Bausman (2018) notes that it is common in these fields to use the more simple of two explanations as a null hypothesis, and upon failing to reject the null hypothesis, concludes that there is support for the null. This practice is justified on the grounds that the more simple explanation should be used until an explanation outperforms it, but such a justification is not rooted in the principles which drive hypothesis testing. Instead, it results in an "epistemological privilege" for the null model, which can lead to unsubstantiated support for the null. This misuse of hypothesis testing points not only to a larger misunderstanding of the role of statistics in scientific study, but to the need for statistical methods which more

appropriately address the types of inquiry that comprise scientific practice.

These limits of hypothesis testing provide substantial motivation for statistical methods that allow for the comparison of multiple and disparate models, and with more interpretable notions of evidence in the contexts of the experiment being performed. One such school of statistical decision-making with these motivations is based on information criteria methods, described in the next section.

2.2 Model Selection

Akaike (1974) formally introduced "an Information Criterion," later to be recognized as the Akaike Information Criterion (AIC), as an estimator of the relative expected KL Divergences among a set of statistical models and an underlying true distribution. The AIC was groundbreaking in relating KL Divergence to Fisher's maximized likelihood to create a simple, tractable measure of relative information loss that could be used to compare multiple models simultaneously, as well as non-nested models with disparate functional forms (Burnham and Anderson 2002).

The formula for the AIC of a model g_i with K parameters, given observed data $\mathbf{x} = x_1, x_2, \ldots, x_n$, is given by

$$AIC(g_i|\mathbf{x}) = -2\mathcal{L}(g_i|\mathbf{x}) + 2K.$$

Because it is a function of the negative likelihood, the best AIC score among a set of models is the minimum score. Then, the quantity $2K$ in the AIC function can be said to penalize a model with more parameters. The AIC is therefore useful for protecting against overfitting models. According to the AIC, the increased likelihood from increasing complexity must outweigh the penalty associated with increasing complexity. In the context of linear regression, for example, including an additional covariate will always decrease the error sum of squares, and thus increase goodness-of-fit. If the additional covariates are fitting observational error, however, predictions made using the more complex model will be inferior to those made by the simple model. information-theoretic methods are commonly used for selecting which variables to use in a linear model and which to leave out.

Making accurate predictions is a key goal of statistics, and was the

orginally intended application of the AIC (Shmueli 2010). However, the AIC is sometimes used in other contexts, for different research goals, and thus has different implications. It is then important to elucidate the differences in types of statistical modeling, and of particular interest is the difference between predictive and explanatory modeling.

Predictive and Explanatory Modeling

The AIC's basis in information theory, and effectiveness minimizing information loss, have made AIC-based practices popular for predictive modeling, both for selecting a best singular model and for creating weighted ensemble models (Shmueli 2010). However, in some scientific fields, predicting future data can be less informative than finding a model with an interpretable functional form (Hox 1998, Wagenmakers 2007).

Models in psychology, for example, are often used to develop theories of the characteristics of human behavior. The intent of such research is not to predict future human behavior, but to fit into a larger context of psychological theory. Similarly, one could imagine a social

scientist using census data to identify demographic factors that lead to homelessness or poverty. Conclusions made by this research would be used to make claims about larger sociological and economic structures. As noted above, these examples would be referred to in these fields as structural equation modeling, used to investigate abstract theoretical constructs which "describe a pheonomenon of theoretical interest" (Hox 1998, Edwards and Bagozzi 2000). Because of their use in communicating real world structures, it is desirable for these models to contain a relatively small number of covariates (Vandekerchkove 2015). Such models are called **parsimonious**, meaning that they describe a phenomenon with as few predictors as possible. The motivation for parsimonious models is well summarized by Occam's razor —given two plausible explanations, the simpler one is usually better.

Explanatory modeling also plays a large role in the biological sciences. A study in cell biology might aim to compare two possible biological mechanisms to attain a better understanding of how cells function (Williams 1970). Candidate models in this case would be approximations of physical or chemical processes. These models would

have functional forms that try to adequately describe real world processes.

Critically, explanatory models offer causal explanations as opposed to associative relationships (Shmueli 2010). Statistical theory would state that many testing contexts, such as observational studies, are not amenable to making conclusions about causal relationships, but this distinction is often ignored in scientific practice (Shmueli 2010).

As stated above, the AIC was not intended for use in explanatory modeling, but its ease of use in comparing multiple and non-nested models has led to its widespread application in explanatory setting (Shmueli 2010, Ullman and Bentler 2012). A naive use of the AIC for this application would be to select and report the model which obtains the best AIC score for given data. However, to do so is to ignore the principle of "model selection uncertainty", and can lead to unpredictable conclusions (Wagenmakers et al. 2004). The AIC is, after all, a modified likelihood function, and thus AIC rankings are a random variable of the data. Thus, the observed best model can vary considerably with sampling fluctuation (Preacher and Merkle 2012). The use

of the AIC and other information-theoretic criteria requires a more nuanced framework than simply reporting the best model. Guidelines for use of AIC, with an emphasis in the natural sciences, are laid out in (Burnham and Anderson 2002) and have been widely used, if at times misused.

Model Selection Methods

Burnham and Anderson (2002) situates model selection not as a replacement to classical statistical testing, but as a supplement to it, especially for observational data and exploratory analysis. According to Burnham and Anderson (B-A) hypothesis tests based on observational data often overlook important assumptions made for those tests, and can often lead to a sort of "false significance." Observational data can contain biases due to sampling methods, and thus are not often representative of the populations they describe. Applying N-P testing to these data is then violating assumptions of the underlying statistical principles (Wagenmakers 2007). Moreover, due to sampling noise, and because of the common interpretation of p-values, exploratory analyses

using N-P testing and p-values can easily overfit a model, for example including erroneous variables in a linear regression analysis.

It is noted in (Burnham and Anderson 2002) that the AIC is not a statistical test and does not entail notions of confidence or error. Thus, the methods they provide are intended for assessing relative evidence for each model, and not for making definitive conclusions regarding models. The foundations of this method are as follows.

For given data X, obtain AIC scores for each model M given X. Then, record the minimum AIC score as $AIC_{\min}$. Using the minimum AIC, one computes the AIC differences Δ_i,

$$\Delta_i = AIC_i - AIC_{min}$$

In the B-A paradigm, score differences allow a practitioner to appraise the level of support for a given model as compared to the best fitting model. The text recommends a loose set of guidelines as a tool to drive further testing, and so do not immediately seek to discount models that do not perform best. As such, AIC differences are compared to a table of benchmark rules of thumb for relative evidence as follows:

Δ_i	Level of Empirical Support of Model i
0-2	Substantial
4-7	Considerably less
$\geq$10	Essentially none.

Table 1: Proposed values for classifying relative support between models based on AIC differences, taken from (Burnham and Anderson 2002)

Burnham and Anderson employ AIC differences as a clear description of relative model fits, or in information-theoretic terms, relative expected information loss. It should be noted that the B-A threshold values are not selected arbitrarily, but have a further justification in terms of each model's likelihoods given the observed data. Burnham and Anderson show that this conditional likelihood, denoted $\mathcal{L}(g_i, x)$ is proportional to the quantity $\exp(-\frac{1}{2}\Delta_i)$. This property is then used to describe model support in two ways: using normalized values which are termed **Akaike weights** and using ratios of $\exp(-\frac{1}{2}\Delta_i)$ for differing models, which are termed **evidence ratios**.

Akaike Weights Given observed data x and models $M_1, M_2, \ldots, M_R$, construct the normalized terms

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^{R} \exp(-\frac{1}{2}\Delta_r)}$$

The interpretation of the w_i is given as follows: "A given w_i is considered as the weight of evidence in favor of model i being the actual K-L best model for the situation at hand given that one of the R models must be the K-L best model of that set of R models" (Burnham and Anderson 2002). Under this interpretation, the Akaike weights can be a useful tool for appraising the degree of support for the best model as compared to the entire model set. Moreover, these weights have been shown to be useful for constructing weighted ensemble models for predictive modeling (Burnham and Anderson 2002,).

Evidence Ratios In addition to the overall weights of conditional likelihoods, define the **evidence ratios** between the conditional likelihoods for two given models M_i, M_j as

$$\frac{\exp(-\frac{1}{2}\Delta_i)}{\exp(-\frac{1}{2}\Delta_j)}.$$

This quantity is of particular interest when model i is the model which obtained the minimum AIC. In this case, because $\exp(-\frac{1}{2}\Delta_i) = 1$ for model i, the evidence ratio is a function of Δ_j. (Burnham and Anderson 2002) gives the following values of this function for some values of Δ_j:

Δ_j	Evidence ratio
2	2.7
4	7.4
8	54.6
10	148.4
15	1808.0
20	22026.5

Table 2: Selected AIC difference values and associated evidence ratios from (Burnham and Anderson 2002). According to the table, if model j has an AIC score 2 points higher than the best fitting model, its likelihood is lower than the best model's by a factor of 2.7.

These evidence ratio values do not give a strict interpretation for decision-making, but are shown in the text to illustrate the nonlinearity between the score differences and the evidence ratios. It is claimed that "this information helps to justify the rough rules of thumb given for judging the evidence for models being the best K-L model in the set" (Burnham and Anderson 2002).That a score difference of 10 corresponds to the a the best model being 148.4 greater in relative likelihood than model j does provide evidence against model j. However, the choice of the specifc cutoff points between the classifications are not derived from any underlying principle.

The AIC was not originally developed for the sake of decisively choosing among statistical models, and neither were the methods described in (Burnham and Anderson 2002), which warns against the equating of their methods to testing:

> "Information-theoretic criteria such as AIC, AIC_c, and $QAIC_c$ are not a 'test' in any sense, and there are no associated concepts such as test power or P-values or $\alpha-$levels. Statistical hypothesis testing represents a very different, and generally inferior, paradigm for the analysis of data in complex settings. It seems best to avoid use of the word 'significant' in reporting research results under an information-theoretic paradigm" (Burnham and Anderson 2002, p. 84).

That these two paradigms are different is clear; that hypothesis test-

ing is "generally inferior," as claimed above, is less so. Crucially, the popular practice of information-theoretic model selection differs from the intended use, much like the combined use of Neyman-Pearson testing and Fisher's p-values. While p-values take on the role of quantifying evidence, as opposed to determining test outcomes, the AIC is used for classification and selection, as opposed to quantifying evidence.

Complaints against AIC selection Insofar as information-theoretic selection procedures are used to report a single "best" model, a practice not entirely consistent with their intended use, it has been shown that these procedures can produce results which are less interpretable and less reliable than results obtained from hypothesis testing. As stated in (Spanos 2010), "Akaike-type procedures are often unreliable because their minimization of a normed-based function is tantamount to comparisons among the models within the prespecified family $M_i(z), i = 1, 2, ...m$, based on Neyman-Pearson (N-P) hypothesis testing with unknown error probabilities." Spanos presents one such example, summarized below:

Example: Consider the following models M_1 and M_2 for observed data $y_1, \dots y_n$ and predictors $x_1, \dots, x_n$, where

$$M_1 : y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i,$$

$$M_2 : y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_t.$$

For the case in which the true model is M_1, Spanos (2010) demonstrates that selecting M_2 based on the best AIC can be described equivalently as a likelihood ratio test. Moreover, it is shown that the error attained by this test varies with n. For $n = 35$, this test has a Type-I error rate of $\alpha = 0.180$.

This example shows a clear incompatibility between the notions of evidence used in hypothesis-testing procedures and in information-theoretic procedures. According to (Spanos 2010), this incompatability is problematic in the context of a larger scientific practice. Despite this incongruence, the benefits of AIC model selection, such as the ability to test between multiple and non-nested models, cannot be overlooked, especially in fields such as biology, which rely on descriptive modeling.

A key aspect of N-P style testing is the availability of solutions to the distribution of test statistics, made possible by the relationships between null- and alternate distributions. Analytical solutions to distribution functions are, in general, not accessible when considering a set of two non-nested models (Wagenmakers et al. 2004). This means, for instance, that the construction of a likelihood ratio test with an analytically determined critical value is generally not possible. Vuong (1989) has made use of asymptotic approximations to likelihood ratio distributions between non-nested models, but which do not allow for multiple model comparisons. Attempts have been made to approximate these distributions computationally, for the purpose of refining selection procedures beween non-nested models (Williams 1970, Wagenmakers et al. 2004, Preacher and Merkle 2012).

Williams (1970a) described a method in which likelihood ratios are generated for synthetic, parametrically bootstrapped data originating from the MLEs of each model with respect to the data. These are used to generate approximate distributions of the LR under each assumption, which are then used to determine critical values for making conclusions

between the two models, given the data. Given that an LR is equivalent to the difference of log-likelihoods, this is cited in (Wagenmakers et al. 2004) as the first use of "Differences in Goodness-of-Fit" (DGOF) distributions under competing assumptions of two models being the true generating process for observed data.

Williams' method is built upon the idea of likelihood ratio tests, but makes a significant departure in including two competing hypotheses as opposed to the classical null and alternative hypotheses. In short, according to this method, a decision can be made to favor either candidate model, or can conclude that the models both perform too poorly or too similarly to choose one.

The method suggested in (Williams 1970a) is developed further and tested in (Williams 1970), using two disparate models of regression. For observed data y_i observed at time t_i, the first model, referred to as the

"Segmented Model," is shown below:

$$y_i = f(\Theta, t_i) + \varepsilon_i, \Theta = (\alpha, \beta_1, \beta_2, \beta_3, T_1, T_2), \text{ where}$$

$$\begin{aligned} f(\Theta, t) &= \alpha + \beta_1 t \text{ for } t \leq T_1 \\ &= \alpha + \beta_1 T_1 + \beta_2(t - T_1) \text{ for } T_1 \leq t \leq T_2 \\ &= \alpha + \beta_1 T_1 + \beta_2(T_2 - T_1) + \beta_(t - T_2) \text{ for } T_2 \leq t, \end{aligned}$$

and $\varepsilon_i \sim N(0, \sigma_f^2)$ and independently distributed.

The fit for this model is compared to a "Smooth Model," given by

$$y_i = g(\Psi, t_i) + \varepsilon_i, \qquad \Psi = (a, b, c),$$

where $g(\Psi, t) = a + b\exp(ct)$ and $\varepsilon_i \sim N(0, \sigma_f^2)$ and independently distributed.

In this experiment, the models represent differing qualitative hypotheses for modeling the patterns of synthesis of various enzymes in bacterial cells. Both regression models describe patterns of synthesis which approximate real world processes, and both models are of distinct interest to the field of research. As such, it would be less informative to

perform a hypothesis test associated with either model than to compare them outright.

The likelihood ratio λ used in this test is given by the ratio of the residual sum of squares for the segmented model to those of the smooth model. (Williams 1970) states that "to use this criterion knowledge is needed of the distributions of λ under the assumption in turn that each of the two models is true." These distributions are dependent on unknown model parameters, which are estimated by the MLEs of each model, given data.

The likelihood difference distributions are approximated using parametric bootstrap resampling (James et al. 2013), as there is no analytical solution for them. As previously stated, the lack of solutions for likelihood ratios of disparate distributions has prevented their use in classical tests, so estimating them with parametric bootstrap is innovative. Despite an innovative sampling method, what follows in the decision process lacks a clear statistical foundation. With regard to the decision regions, Williams (1970) states that "no justification of this rule in terms of misclassification probabilities is claimed, for the form

of the distributions of the conditional likelihood ratios are not known."

Despite its shortcomings, this method approaches a problem in statistical model selection distinct from the problem of intrinsic model complexity. Where the AIC is used to penalize models that are more complex in general, based on numbers of parameters, (Williams 1970) considers models' ability to fit data generated by other models, a property given the term model mimicry (Wagenmakers et al. 2004).

The motivation for decision procedures that take into account model mimicry is clear. Suppose one is comparing fits of two models, A and B, where A is relatively inflexible and unable to fit data from B, but B is able to fit well to data generated by A. Even if A has more parameters, and thus more complexity according to the AIC, this complexity does not lead to increased GOF in the case where B is true.

Wagenmakers et al. (2004) Wagenmakers et al. (2004) also uses a bootstrap procedure to estimate distributions of relative goodness of fit functions for model selection. The text describes these distributions as "Difference in Goodness of Fit" distributions, which use log-likelihoods (or in this case, AIC scores) to convert the likelihood

ratio into a difference. In AIC-based selection between two non-nested models, model mimicry can lead to unequal misclassification probabilities. It is claimed that if these probabilities are mismatched, e.g. if data generated from model A is more likely to be attributed to model B than vice versa, the procedure exhibits a type of bias toward model B.

To combat this phenomenon, Wagenmakers' procedure uses approximate DGOF distributions to move the decision criterion away from zero (i.e. simply choosing the best fitting model) and to a point "that maximizes the probability of a correct binary classification." This procedure is outlined below:

For two models A and B and observed data X, estimated parameters and goodness-of-fit values are obtained for each model based on X. This gives fitted parameter values $\hat{\Theta}_A$ and $\hat{\Theta}_B$ for models A and B, and an observed difference in goodness-of-fit ΔGOF_{AB} given by $GOF_A - GOF_B$.

Under competing assumptions that each model is true, approximate distributions $(\Delta GOF^*_{AB}|A \text{ is true})$ and $(\Delta GOF^*_{AB}|B \text{ is true})$ are com-

puted via bootstrap. The densities of these distributions are denoted by $P_A(x)$ and $P_B(x)$, respectively. Then, an observed ΔGOF_{AB} is more likely to have come from the assumption that model A is true when $P_A(x)/P_B(x) > 1$ for $x = \Delta GOF_{AB}$.

From $P_A(x)$ and $P_B(x)$, (Wagenmakers et al. 2004) constructs a decision criterion which differs from the usual interpretation of the AIC, in which the best model is that which obtains the lowest score. In this procedure, the decision criterion is set to the point denoted (opt) at which $P_A(x)/P_B(x) = 1$. The relationships between the distributions and the decision criterion is best depicted graphically:

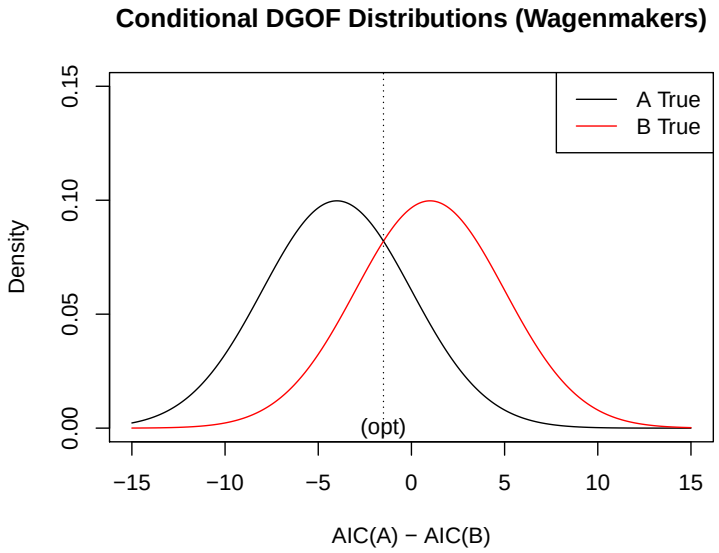

Figure 1: Model Mimicry: example DGOF distributions conditional on the true model. A line has been drawn at -1.5, marking the optimized decision threshold between the two models.

In the plot above, the two density curves intersect at the value $\Delta GOF_{AB} = -1.5$. This represents a case in which model A exhibits greater mimicry than model B. The distribution for $\Delta GOF_{AB}|B$ is true has a mean value closer to zero compared to the mean of the distribution of $\Delta GOF_{AB}|A$ is true. This means model A is observed best more often when B is true than vice versa. Thus, the decision threshold (opt) = -1.5 gives equal classification error, while a threshold of 0 would result in greater error in the case that model A is true.

Wagenmakers' procedure differs from Williams' in two critical ways. Firstly, it relates the decision criterion to misclassification error, which was thought to be inaccessible to Williams. However, the Wagenmakers procedure has only two possible decisions: a conclusion in favor of each model. Thus, Wagenmakers' procedure serves only to optimize decision-making within the current standard of AIC-based model selection. It falls short of addressing complaints leveraged against model selection such as those in (Spanos 2010): that model selection amounts to statistical testing with undefined and unmeasured degrees of error.

This paper aims to adapt and extend ideas from (Williams 1970) and (Wagenmakers et al. 2004) to define a model selection procedure driven by the notion of error control. This will involve defining a general, functional framework for decision-making in model selection that permits notions of Type-I and Type-II error, as well as a theoretical procedure which uses conditional DGOF distributions to bound Type-I error. I will then present a bootstrap method for approximating the necessary functions for the decision procedure along with results on simulated i.i.d. and time series data.

3 A Framework for Model Selection

Toward a Translatable Standard of Evidence Motivated by the gap in evidence between classical hypothesis testing and model selection, as well as the increased generality and flexibility of information-theoretic model selection, we aim to develop a procedure by which model selection can be performed with a standard of evidence comparable to that of hypothesis testing. In particular, we will define a general theoretical framework for model selection which allows us to extend the notions of Type-I and Type-II Errors to information-theoretic procedures. Moreover, we will propose and demonstrate a procedure which uses this framework to perform AIC model selection with specified maximum Type-I error rate α.

Given data $D \in \mathcal{D}$ and candidate model set $\mathcal{M} = \{M_1, M_2, \ldots, M_k\}$, specify

1. A criterion function $f : \mathcal{M} \times \mathcal{D} \to \mathbb{R}$, which assigns a score to each model M_i based on the observed data.

    - Let F denote the vector of criterion values $= \big(f(M_i, D) :$

$i = 1, \ldots, k\big)$.

2. A preference function $g : \mathbb{R}^k \to \mathcal{M}$, which selects a best model from the candidate model set given the model score vector.

    - For observed F as above, let $M_b = g(F)$. In other words, denote the best model according to g by M_b.

3. A decision function $h : \mathcal{M} \times \mathbb{R}^k \to \{0, 1\}$, which is equal to 1 if the specified model is chosen to be correct, and 0 otherwise.

    - Generally, we will take $h(M_i, \cdot) = 0$ when $i \neq b$. This means that in general, only the observed best model can be chosen as correct.

3.1 Example: Simple AIC-based selection

This model selection framework permits a standard use of the AIC for selecting a correct model according to best-fit. For data D and candidate model set $\mathcal{M} = \{M_1, M_2, \ldots, M_k\}$, define

1. $f(M, D) = \text{AIC}(M, D) = 2k - 2\hat{\mathcal{L}}(M; D)$, where k is the number of parameters for M and $\hat{\mathcal{L}}(M; D)$ is log-likelihood evaluated at

the maximum likelihood estimator.

2. $g(F(D)) = \arg\min_M(F(D))$. Here we report the model with the lowest AIC as best.

3. $h(M_b, \cdot) = 1$. The observed best model is automatically reported to be correct.

3.2 Example: Changing the Decision Function

Reporting the best-fitting model as correct can lead to a high rate of misclassification error, a commonly raised complaint against information-theoretic selection methods (Spanos 2010). With this in mind, many have specified rules-of-thumb for quantifying evidence in support of the best model (Burnham and Anderson 2003). In the case of Burnham and Anderson, alternative models are said to attain relatively no support given data compared to the best model when they attain AIC scores at least 10 points above that of the best model.

Along these lines, we can encode a decision function which reports the best model as correct if its score is at least 10 points less than all other model scores. We define the criterion and preference functions as

above. Letting M_b be the model which attains the best AIC score, as above, define

$$\Delta f = F_b - min\{F_i : i \neq b\}$$

but now define the decision function as follows.

$$h(M_b, F(D)) = \begin{cases} 1 \text{ if } \Delta f < -10 \\ \\ 0 \text{ otherwise} \end{cases}$$

4 Error Rates and the Confusion Matrix

In adapting error rates to a general information-theoretic model selection context, we must consider what constitutes a Type I or Type II error. These terms typically refer to conclusions about the null hypothesis in classical hypothesis testing, while our framework doesn't include the specification of a null model. As such, there is no null hypothesis to be rejected, but rather the difference is between selecting or not selecting one of the candidate models as correct.

When performing error controlled model selection, we assume that the model set contains the true generating process for the data. Under this assumption, we can define errors based on reporting a false model as correct or on failing to identify the true model. The latter case is more complex, and to examine it further we consider the possible outcomes of the model selection procedure. Suppose that among the model set $\mathcal{M} = \{M_1, \ldots, M_p\}$ and for some $t \in \{1, \ldots, p\}$, model M_t is the true generating process. Then according to our framework, the possible outcomes for the preference and decision functions are

1. $M_b = M_t$ & $h(M_b, F(D)) = 1$: The true model is observed best, and we conclude that it is correct. This is a "true positive."

2. $M_b \neq M_t$ & $h(M_b, F(D)) = 1$: An incorrect model is observed best, and we conclude that it is correct. This is a "false positive."

3. $M_b = M_t$ & $h(M_b, F(D)) = 0$: The true model is observed best, and we fail to conclude it is correct. This is a "false negative."

4. $M_b \neq M_t$ & $h(M_b, F(D)) = 0$: An incorrect model is observed best, and we fail to conclude it is correct. This is a "true negative."

In separating cases three and four, we recognize case three as a shortcoming of the error control process and case four as a feature. We say that a Type II error occurs under case three, but not case four.

To evaluate and compare decision procedures we consider their Type I error rates as well as their *sensitivity*, given by the ratio of the number of true positives to the number of true positives and false negatives. Thus, a decision function in which the output of the preference function is always reported (e.g. always choosing the model with the best AIC),

the sensitivity is 100%, but there is no control on the error rate. Sensitivity leads to a salient interpretation of the amount of signal lost due to error control, and also could be used to compare decision procedures across multiple score and preference functions.

The Confusion Matrix Suppose that we have p different models. We define the confusion matrix M to be the $p \times p$ matrix where each element M_{ij} corresponds to the probability of reporting model M_i as true when M_j is the true model. The elements of the confusion matrix are given by

$$M_{ij} = p(M|T_j = 1)p(h = 1|B_i = 1, T_j = 1) \tag{1}$$

The term $p(B = i|T_j = 1)$ is the probability that model i will be chosen as best while assuming that model j is the true generating mechanism. The term $p(h = 1|M_j, B = i)$ is the probability of obtaining a decision that model i is correct, assuming that model j is true.

Thus, for $i = j$, i.e. the diagonal elements of M describe the probability of recovering the true model, given the data. The elements where

$i \neq j$ describe the probability of false positives, i.e. for concluding that model i is true when the data was actually generated by model j.

Thus, if we have models $M_1, \ldots, M_p$, with true model M_T, the value M_{TT} gives the power of the test, and the value M_{ij} gives the probability of committing a Type-I error under the condition that M_i is observed best and M_j is the true distribution.

The advantage of situating model selection within our general framework, i.e. using preference and decision functions, is that it gives us a means of comparing model selection procedures similarly to how we compare statistical tests. For instance, since we can define the type I error rate α, we might approach the choice of a model selection procedure similarly to identifying a uniformly most powerful statistical test —by choosing an acceptable error rate and then seeking to maximize statistical power.

This thesis defines a class of decision functions h_τ motivated by the elements of the confusion matrix and a bootstrap-based approach for estimating the tuning parameter τ to bound the error rate at a specified value. In doing so, we hope to work toward facilitating the translation

or comparison between standards of evidence for model selection.

5 A Class of Decision Functions h_τ

Recall that the element M_{ij} of confusion matrix M is given by $M_{ij} = p(B_i = 1|T_j = 1)p(h = 1|B_i = 1, T_j = 1)$.

A decision function with error rate α would output $h = 1$ in such a manner that all $M_{ij} = p(B_i = 1|T_j = 1)p(h = 1|B_i = 1, T_j = 1) = \alpha$ for $i \neq j$, $j = t$. Note that the probability term $p(B_i = 1|T_j = 1)$ does not depend on the choice of h. Suppose that for a given experiment, model M_j is the true model, and model M_i is observed best. Suppose further that we could compute the true value of $p(B_i = 1|T_j = 1)$. Then, suppose that for a given $\tau \in [0, 1]$, we could define a function h_τ such that the event $(h_\tau(\dots) = 1|B_i = 1|T_j = 1)$ occurs with probability τ.

Then, for $\alpha \in [0,1]$, define $\alpha' = \frac{\alpha}{p(B_i=1,T_j=1)}$. Then, let $\tau = \alpha'$ we have

$$\begin{aligned} M_{ij} &= p(B_i = 1|T_j = 1)p(h_{\alpha'} = 1|B_i = 1|T_j = 1) \\ &= p(B_i = 1|T_j = 1)\frac{\alpha}{p(B_i = 1|T_j = 1)} \\ &= \alpha \end{aligned}$$

Toward a tractable form of h_τ We have specified a class of functions of the form h_τ such that given models M_i and M_j, score function f, and preference function g, $P(h_\tau = 1|B_i = 1, T_j = 1) = \tau$. Suppose that f was the AIC and g selected the model with the lowest AIC. Then, to approach an h_τ which relates to the probability that model M_i is chosen as best when model M_j is the true generating mechanism, we consider the statistic

$$\Delta AIC_i = AIC(D, M_i) - \min\{AIC(D, M_j)\}_{i \neq j},$$

which gives the difference between the AIC for the first- and second-ranked models. Recall that this statistic is used by Burnham and Anderson to describe evidence in support of the observed best model based

on rules-of-thumb, e.g. report the best model if $\Delta AIC_i < 10$. Instead of using a preordained value, suppose instead that we could access the conditional distribution of $(\Delta AIC_i | B_i = 1, T_j = 1)$. For $\tau \in [0, 1]$, define q_τ to be the τth percentile of this distribution. Then, define h_τ to be

$$h_\tau = \begin{cases} 1 \text{ if } \Delta AIC_i < q_\tau \\ \\ 0 \text{ otherwise} \end{cases}$$

By definition of percentile, $P(h_\tau = 1 | B_i = 1, T_j = 1) = \tau$. Thus, supposing we can access the distribution of $(\Delta AIC_i | B_i = 1, T_j = 1)$, we can define a function h_τ that attains type-I error rate of $p(B_i = 1 | T_j = 1) p(h = 1 | B_i = 1, T_j = 1) = \alpha$ by setting $\tau = \alpha'$. Note that τ varies with ij, i.e. that each pair of models will have a different decision function with differing thresholds, but indices are suppressed here for convenient notation.

In practice, we do not have access to the threshold values q_τ, because we do not have advance solutions for the probabilities $p(B_i = 1 | T_j = 1)$ or the conditional distributions of $(\Delta AIC_i = 1 | B_i = 1, T_j = 1)$.

We cannot generally expect to find analytic solutions based on the distributions of the models and the functions used in the model selection procedure. This issue is prevalent in model selection - the flexibility afforded in comparing multiple and non-nested models using e.g. the AIC comes at a cost.

In light of the unavailability of analytic solutions, we opt for computational methods to estimate the necessary components to produce approximate functions $\hat{h}_\tau$ for h_τ. However, the parametric bootstrap has been shown to be effective in estimating quantiles (Falk 1989), and my method will make use of this parametric resampling to construct approximate DGOF distributions to estimate quantiles.

6 BITSEC - A Bootstrap Procedure for Approximating the Decision Function

As above, suppose that we have data D, candidate model set $\mathbf{M} = \{M_1, \ldots, M_p\}$, and true model $M_t \in \mathbf{M}$. Moreover, assume we have chosen a score function f and a preference function g. Given a de-

sired Type I error rate α, I introduce a bootstrap method, abbreviated BITSEC (Bootstraped Information-Theoretic Selection with Error Control), which implements a decision function $\hat{h}$ with estimated error rate $\hat{\alpha}$.

Given the data, the model set, and the score and preference functions, we are able to determine the best performing model with respect to the preference function. This gives us $B_k = 1$, i.e. that model M_k is the observed best. At this point, we now seek to estimate α' and use it to estimate $h_{\alpha'}$

Estimating $\mathbf{p(B_i = 1|T_j = 1)}$ We use parametric bootstrap sampling to estimate the probabilities $p(B_i = k|T_j = 1)$ for each $j \neq k$:

Let $\hat{M}_j$ be the maximum likelihood estimated fit of model M_j. For each $j \neq k$, generate N_1 new datasets from $\hat{M}_j$ and record the frequency with which M_k is observed best. This is an estimate of $p(B_i = k|T_j = 1)$ because it assumes that the data was generated by M_j and uses the MLE $\hat{M}_j$ as the estimated true distribution corresponding to model M_j.

Specifying the Decision Function As previously mentioned, we now seek a function h which, conditioning on $B_k = 1$ and $T_j = 1$, is equal to 1 with probability $\frac{\alpha}{P(B_k=1|T_j=1)}$. To do so, we first record the observed score delta for the experiment:

$$\Delta AIC = AIC(M_k) - \min\{AIC(M_j)\}_{j \neq k}$$

To use this value as the basis for our decision function, we use a parametric bootstrap to estimate the conditional distribution $\Delta AIC|T_j = 1$. After fitting the model M_j, we parametrically sample from it to create N new datasets of length n, each of which gives $B_i = 1$. We compute the score differences $\Delta_1^*, \ldots, \Delta_N^*$. Finally, we select the order statistic $\Delta_{(k)}^*$ where k is the largest integer in $1, \ldots, N$ for which $k/N \leq \alpha'$. For example, with $N = 1000$ and $\alpha' = 0.1525$, we select $\Delta_{(152)}^*$ to approximate the α' percentile. Define the quantity $\hat{q}_j = \Delta_{(k)}^*$. We perform this sampling and threshold estimation for each candidate model M_j and obtain q_j for all $j \neq i$.

Finally, we have a set of threshold values $\{q_j\}_{j \neq i}$ corresponding to conditional probabilities of observing various differences in AIC scores.

We use these analogously to critical values in a hypothesis test. For instance, because the threshold $\{q_j\}$ approximates the α percentile of the distribution of $(\Delta AIC|B_i = 1, T_j = 1)$, if the observed score difference $\Delta AIC_{obs} < q_j$, then there is less than an α percent probability that ΔAIC_{obs} could have occurred under the assumption that model M_j was true.

Multiple Decision Thresholds Because the BITSEC procedure has been designed with the preference of obtaining an error rate less than α, we take the minimum of all $\tau_{\alpha'}$ values as the decision threshold that the observed score difference is compared to. An alternative to choosing the minimum threshold might be choosing between the thresholds randomly, but this would result in cases for which too lax of a threshold is used for the decision. This procedure may be more appropriate in cases where data is assumed to come from a mixture of true generating distributions.

As such, we set $h = 1$ when $\Delta AIC_{obs} \leq min\{q_j\}_{j \neq k}$. This means that we only accept the observed score difference ΔAIC_{obs} if it is sufficiently unlikely to have come from any of the alternate assumptions

of models M_j, being true, $j \neq i$. Thus we have approximated a decision function with an estimated maximum Type I error rate $\hat{\alpha}$. Shown below is a visualization of two distributions of $\Delta AIC|T_j = 1$ and the associated $\alpha = 0.05$ decision thresholds. In this example, the model M_3 has been observed best, so thresholds are computed under the assumptions $T_1 = 1$ and $T_2 = 1$. The leftmost threshold value of -5.48 is used for the procedure's decision. Also shown below is a flow chart summarizing the entire BITSEC procedure.

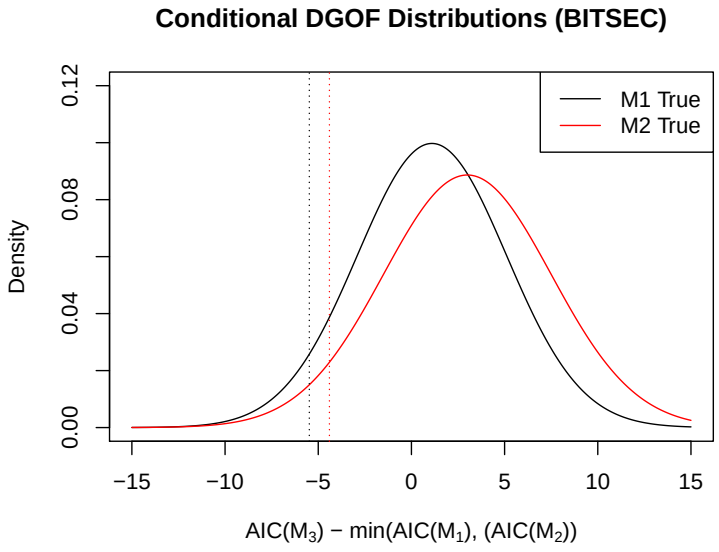

Figure 2: Error Control: Conditional DGOF distributions and α level thresholds. Model M_3 has been observed best, and thresholds are computed for the assumptions that each of the other two models is true.

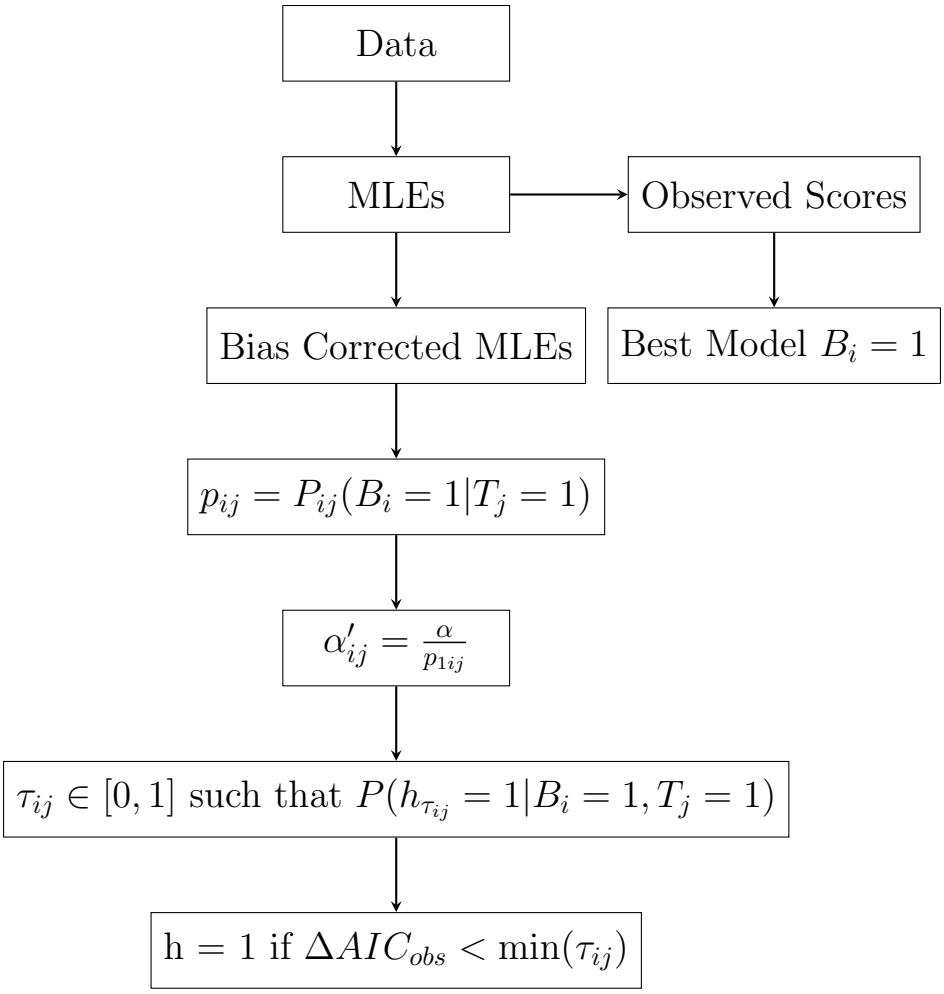

Figure 3: The full BITSEC procedure: First, MLEs are fit to the observed data, resulting in observed scores and observed best model. Then, a bias correction is performed on the parameter estimates, and the new values are used to estimate the probabilities $P_{ij}(B_i = 1|T_j = 1.$ These probabilites are used to determine the proper quantile of the DGOF distribution which will be approximated by the bootstrap to give a decision threshold. After performing this step for all models other than the observed best, the minimum threshold is taken and compared to the observed score differnce. If the observed score is less than the given threshold, the best model is declared to be correct. Otherwise a "null decision" is given.

7 Results

7.1 Bias in MLEs

Our model selection procedure depends on parametrically sampling from the MLE for the observed data to estimate the conditional probability of each model attaining the lowest AIC score as well as the score difference distributions which determine the decision thresholds. For the model $N(\mu, \sigma)$, the MLE given by $\hat{\mu} = \hat{\mu}, \hat{\sigma} = s$ is unbiased overall, but loses this property when conditioning on the observed best model, which the BITSEC procedure cannot avoid.

Example For the case in which the candidate model set is $N_0 = N(0, 1), N_1 = N(\mu, 1), N_2 = N(\mu, \sigma)$, consider the effect of conditioning on N_0 attaining the best AIC score. It is clear to see that $B_0 = 1$ occurs for samples in which $\hat{\mu}$ and $\hat{\sigma}$ are suitably close to 0 and 1 respectively. As such, we can see how the average MLE, conditional on N_0, denoted by $(\hat{\mu}, \hat{\sigma})|B_0 = 1$, will be biased toward $(0, 1)$. Shown below are the results of conditioning on each of the three models, where the MLE's

shown are averaged over 10,000 runs.

	$B_0 = 1$		$B_1 = 1$		$B_2 = 1$	
Parameter	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$
n = 25	0.08169	1.0777	0.4104	1.0603	0.2671	1.2882
n = 50	0.07806	1.0749	0.3258	1.0560	0.2425	1.2342
n = 100	0.0688	1.0659	0.2617	1.0502	0.2225	1.1790

Table 3: Estimation of (μ, σ), true values = $(0.214, 1.124)$ when sampling from $N(\mu, \sigma)$ and conditioning on the observed best model under the AIC. These parameter values have been selected such that each model attains the best AIC with nearly equal probability when n = 50,

It should be noted that as the amount of data increases, the effects of conditioning are different among the different models. We see that the conditional MLEs for (μ, σ) get closer to the true parameter values when $B_1 = 1$ or $B_2 = 1$, but farther away when $B_0 = 1$. This stands to reason because with more data, the parametric models continue to fit the data better while the fixed model stays the same. Moreover, as the amount of data increases, the penalty for complexity plays less of a role in choosing the best model. As such, we observe that the data must give an MLE increasingly close to $(\mu, \sigma) = (0, 1)$ for N_0 to be observed best.

The effects of conditioning on observed best model and of increasing sample size are summarized below. For $N = 1000$ samples at each size of $n = 25, 50, 100$, observed parameter estimates are recorded and coded by the best observed model. The bottom-right plot shows the trends displayed by the mean MLEs after conditioning on each model.

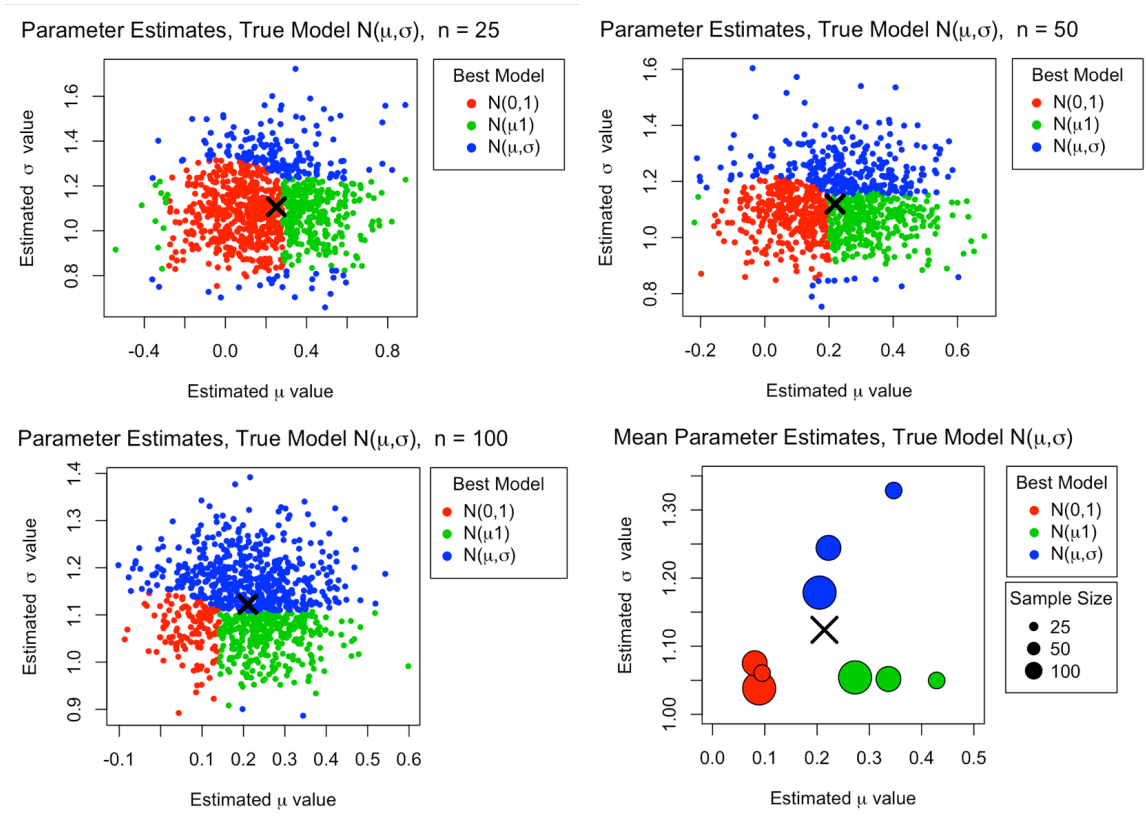

Figure 4: Visualizations of classifiaction regions as a function of sample size, true model N_2 with $(\mu, \sigma) = (0.214, 1.124)$ Bottom left: mean parameter estimates conditional on the observed best model as a function of sample size. Parameter estimates for N_1 and N_2 improve, while those for N_0 worsen.

7.2 Bias Correction

To combat the effect of the bias on the MLEs resulting from conditioning on $B_i = 1$, we estimate the conditioning affect and correct for it. The bias correction procedure proceeds as follows:

- Observe $B_i = 1$ and the MLE $\hat{\Theta}$
- Parametrically sample $\vec{X}_1^*, \ldots \vec{X}_N^*$ from the assumed true model $M_j(\hat{\Theta})$, where $B_i = 1$ for each X_k^*.
- Obtain $\hat{\Theta}_1^*, \ldots, \hat{\Theta}_N^*$ from the simulated data, compute $\hat{\Theta}' = \text{mean } (\overline{\Theta}_1 *, \ldots, \hat{\Theta}_N^*)$.
- Estimate bias B using $\hat{B} = \hat{\Theta} - \overline{\Theta}$
- Corrected estimate of Θ is given by $\hat{\Theta}_C = \overline{\Theta} - \hat{B}$

In general we cannot assume that $\hat{B} = B$, but the bias correction results in estimated parameter values that are closer to the true parameters, which results in better downstream estimates in the model selection procedure.

Example Once again we consider the candidate models N_0, N_1, N_2 with true model $N_2 = N(\mu, \sigma)$. Shown below are corrected estimates of μ, σ when $B_i = 1$.

	$B_0 = 1$		$B_1 = 1$		$B_2 = 1$	
Parameter	$\hat{\mu}$	s	$\hat{\mu}$	s	$\hat{\mu}$	s
n = 25	0.134	1.118	0.323	1.098	0.251	1.200
n = 50	0.117	1.101	0.270	1.086	0.230	1.177
n = 100	0.110	1.093	0.236	1.076	0.219	1.144

Table 4: Bias corrected estimates of (μ, σ), true values = $(0.214, 1.124)$, when sampling from $N(\mu, \sigma)$ and conditioning on the observed best model under the AIC.

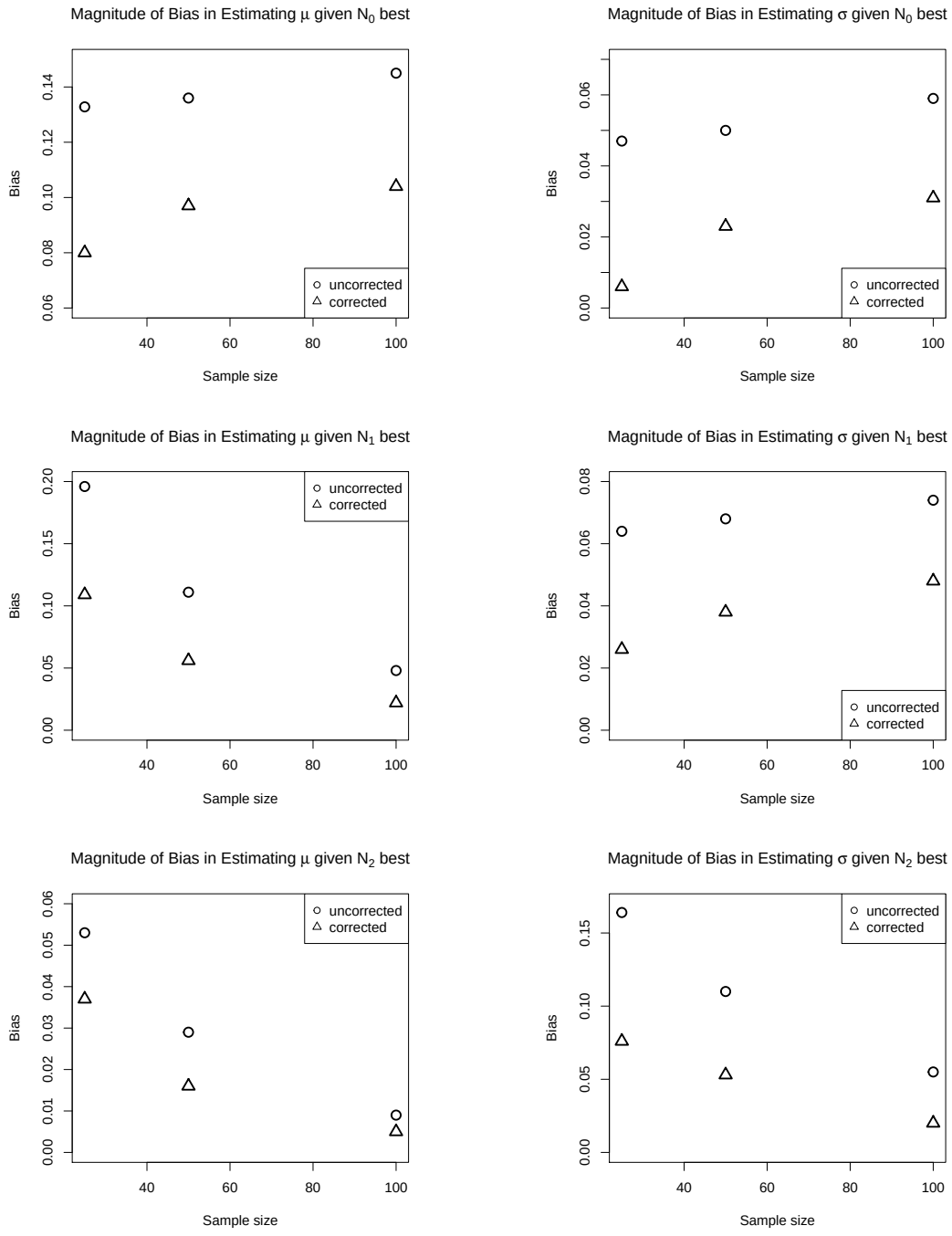

Figure 5: Plots of the absolute value of the bias when estimating (μ, σ), given that each model is true. In every case, the bias correction method reduces the bias.

8 Results

8.1 3 Nested Normal Models

Finally, we test the full selection procedure using N_2 as the true model, with $(\mu, \sigma) = (0.214, 1.124)$, across $\alpha = 0.01, 0.05, 0.10$ and $n = 25, 50, 100$. The values for (μ, σ) have been selected attain nearly equal probability for each model being observed best when $n = 50$. Equal selection probabilities pose a considerable problem in model selection, and provide an opportunity to demonstrate the potential strengths of BITSEC. To attain approximate error rates, we use the proportion of decisions obtained from running the procedure 1000 times on simulated data, and perform this simulation multiple times to obtain an average error rate. We compare these proportions to those obtained by selecting the model with the best AIC or using Burnham and Anderson's decision threshold of 10.

		N_0	N_1	N_2	No Decision	Total Error	Sensitivity
	Best AIC	499	313	188	0	0.812	1
n = 25	$\Delta = -10$ Rule	0	0	3	997	0	0.016
	Error Controlled	0	5	28	967	0.005	0.149
	Best AIC	343	336	321	0	0.679	1
n = 50	$\Delta = -10$ Rule	0	0	10	990	0	0.031
	Error Controlled	1	3	74	922	0.004	0.231
	Best AIC	143	331	526	0	0.474	1
n = 100	$\Delta = -10$ Rule	0	0	29	971	0	0.055
	Error Controlled	1	4	162	833	0.005	0.308

		N_0	N_1	N_2	No Decision	Total Error	Sensitivity
	Best AIC	512	285	203	0	0.797	1
n = 25	$\Delta = -10$ Rule	0	0	2	998	0	0.001
	Error Controlled	25	15	117	843	0.04	0.576
	Best AIC	349	328	323	0	0.677	1
n = 50	$\Delta = -10$ Rule	0	0	9	991	0	.028
	Error Controlled	22	24	200	754	0.046	0.619
	Best AIC	136	308	556	0	0.444	1
n = 100	$\Delta = -10$ Rule	0	0	37	963	0	.067
	Error Controlled	3	13	388	596	.026	0.698

		N_0	N_1	N_2	No Decision	Total Error	Sensitivity
	Best AIC	513	289	198	0	0.802	1
n = 25	$\Delta = -10$ Rule	0	0	1	999	0	0.005
	Error Controlled	51	53	173	717	0.11	0.874
	Best AIC	351	351	298	0	.702	1
n = 50	$\Delta = -10$ Rule	0	0	9	991	0	0.009
	Error Controlled	25	59	258	658	0.084	0.866
	Best AIC	130	342	528	0	0.472	1
n = 100	$\Delta = -10$ Rule	0	0	44	956	0	.009
	Error Controlled	10	49	473	468	.059	0.896

Table 5: Results: Error and Sensitivity, Normal Models

As we predicted above, the observed error rates decrease well below α as n increases due to increased bias in the conditional MLEs, and thus in the estimates of α'. While an improved bias correction to the MLEs could result in an improved estimator for α', we note that increasing n still results in increased sensitivity, from which we conclude that the procedure continues to improve with increased data.

8.2 Random Walks

Earlier in this paper I mentioned the question of comparing rates and modes of evolution in fossil data, a problem described in Hunt (2015). One question of interest is in distinguishing between data depicting stasis and data depicting unbiased or biased random walks. This line of questioning is not suited to hypothesis testing, because the indication of any model as correct would be have a relevant biological interpretation. However, stasis has been commonly used as a null model, and has been used in the manner of accepting the null model when failing to reject it (Hunt 2015, Bausman 2018).

This experiment uses the model set $\mathcal{M} = \{M_1, M_2, M_3\}$, where

M_1 denotes stasis, M_2 denotes an undirected random walk, and M_3 denotes a directed random walk. The stasis model is given by $N(\mu, \sigma)$, and the random walk models are given as follows: for observation y_t and predictor variable x_t taken at time t, y_t is given by

$$y_t = \beta_0 + y_{t-1} + \varepsilon_t,$$

with the ε_ts independently distributed as $N(0, \sigma)$ and $\beta_0 = 0$ for the undirected random walk.

Using data generated by true model M_3 parametrized by $\beta_0 = 0.1, \sigma = 1.1$, chosen to assign the best AIC to M_2 with much greater frequency than M_3, sample sizes $n = 15$ and confidence levels $\alpha = 0.01, 0.05, 0.10$, shown below is a summary of error rates and sensitivities using basic AIC model selection, the Burnham and Anderson threshold of a score difference less than -10, and the BITSEC procedure:

	M_1	M_2	M_3	No Decision	Total Error	Sensitivity
Best AIC	2	737	261	0	0.739	1
$\Delta = -10$ Rule	0	0	0	1000	0	0
BITSEC ($\alpha = 0.01$)	0	12	32	717	0.012	0.123
BITSEC ($\alpha = 0.05$)	0	48	129	717	0.048	0.494
BITSEC ($\alpha = 0.1$)	1	94	169	717	0.095	0.643

Table 6: Simulation results for random walk models with true model M_3 with parametrized by $\beta_0 = 0.1, \sigma = 1.1$

8.3 Nested Linear Regression Models

(Spanos 2010) raises an example for which the simple AIC procedure can be represented as a hypothesis test, one with an implicit Type-I error rate of 0.180. Not only is this error rate relatively high compared to typical benchmarks, but is dependent on the amount of data in the test. In this section we apply the error controlled model selection method to the models used in this example.

Let $\mathcal{M}$ be the candidate model set $\mathcal{M} = M_1, M_2$, with

$$M_1 : y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \beta_3 x_t^3 + \varepsilon_t,$$

$$M_2 : y_t = \alpha_0 + \alpha_1 x_1 + \varepsilon_t$$

and suppose that data $X = x_1, x_2, \ldots, x_{35}$ is generated from M_2, with

parameters

$$a_0 = 167.115$$

$$a_1 = 1.907$$

$$\sigma = 1.77$$

as presented in Spanos (2010) and originally seen in Spanos (2000). We simulate 1000 datasets from the true distribution and apply the error controlled selection method with $\alpha = 0.05$. A table of decisions is shown below:

		M_1	M_2	No Decision	Total Error	Sensitivity
n = 25	Best AIC	166	834	0	0.166	1
	$\Delta = -10$ Rule	0	0	1000	0	0
	Error Controlled	44	82	874	0.044	0.098

Table 7: Simulation results for nested linear regression models with true model M_2 and $\alpha = 0.05$.

Thus, using the error controlled procedure, we conclude the experiment in Spanos (2010) can be performed successfully with an approximate maximum error rate of $\alpha = 0.05$. The sensitivity of this test

is considerably low, which attests to the difficulty in discriminating between these two models.

8.4 Nonlinear Nonnested Regression Models

We compare the AIC model selection methods to simplified versions of the models described in (Williams 1970). These models have been simplified for ease of computation, as the goal of the experiment was to assess how the procedures perform in choosing between the differing functional forms of the models. Thus, the breakpoints in the piecewise model were assumed to be known, as was the term inside the exponential function in the smooth model. This allowed for fitting the piecewise model without considering all breakpoints, and linearized the smooth model. The models are given as follows, where M_1 denotes the segmented model and M_2 denotes the smooth model.

$$\begin{aligned} y_i &= \alpha + \beta_1 t \text{ for } t \leq 10 \\ &= \alpha + 10\beta_1 + \beta_2(t-10) \text{ for } 10 \leq t \leq 20 \\ &= \alpha + 10\beta_1 + 10\beta_2 + \beta_(t-20) \text{ for } 20 \leq t, \end{aligned}$$

and $\varepsilon_i \sim N(0, \sigma_f^2)$ and independently distributed. Then, the smooth model is given by

$$y_i = a + b\exp(0.2t) + \varepsilon_i$$

with $\varepsilon_i \sim N(0, \sigma_f^2)$ and independently distributed. Data of sample size $n = 25$ was sampled from M_2 with parameters chosen such that the correct model was selected by the AIC with approximate probability 0.80. As in previous experiments, a simulation was performed on 1000 samples from the true distribution to obtain approximate decision frequencies for specified error rate $\alpha = 0.05$, summarized below:

		M_1	M_2	No Decision	Total Error	Sensitivity
n = 25	Best AIC	198	802	0	0.198	1
	$\Delta = -10$ Rule	0	0	1000	0	0
	Error Controlled	53	446	501	0.053	0.556

Table 8: Simulation results for nonnested regression models, $\alpha = 0.05$

As expected, the true model is identified by the AIC approximately 80% of the time. The rule of thumb method fails to choose a model in almost every run. The attained error rate for $\alpha = 0.05$ was slightly above

the desired α. This experiment may have benefitted from improved parameter estimation or an increased number of bootstrap samples, but the results still corroborate the notion that error can be controlled in a setting such as the one specified in (Williams 1970).

8.5 Summarized Results

In the experiments presented above, common AIC-based methods provided unpredictable results for explanatory modeling, and lead to conclusions without a clear degree of confidence. In particular, reporting the observed best model as correct leads in all cases to high degrees of Type-I error, and reporting a model based on the rule of thumb from (Burnham and Anderson 2002) leads in all cases to poor performance in terms of sensitivity. Either choice of fixed decision thresholds ignores the fact that AIC differences lack a uniform scale across different experiments and model sets. Thus, along with undesirable performance in terms of error and sensitivity, the standard of evidence employed by these methods is inconsistent among different statistical contexts.

The error controlled method introduced in this paper improves upon

the constant threshold methods in all of the cases shown. Across many types of data and models, the method succeeded in setting approximate error bounds and achieving relatively high degrees of sensitivity. Moreover, sensitivity is seen to increase as the allowable error rate increases and as the amount of data increases. By reducing model selection uncertainty, this method leads to a standard of evidence better suited to descriptive modeling and causal inference. This is especially useful in applications to natural sciences, where model selection is already widely employed.

The analyses given above were performed using an R package developed as part of this paper. The `ICError` package is currently undergoing error handling and documentation and will be released for public use.

9 Software Package

As part of this paper I developed an R package which provides simple functions for implementing the BITSEC procedure. It contains functions for generating data, estimating parameters, and computing AIC

scores for specified models. Familiar models and families of models are supported, including linear regression models, nonlinear regression models using the nonlinear least squares package `nls`, segmented regression using the `segmented` package, random walks, as well as standard distributions. The package will be extended to support users with models that are not included, by letting users specify generating, fitting, and scoring functions as needed.

In developing and improving the BITSEC functions, I took many steps to improve runtime. The package has an option for parallel processing, which can greatly improve speed relative to the number of cores on the user's computer. The package also heavily implements vectorized functions such as `apply` as opposed to iterative loops, which conserve memory and greatly decrease runtime. Moreover, for models whose MLEs or score functions have analytic solutions, direct computations are used whenever possible. For instance, when obtaining scores for linear regression models, sums-of-squares are computed using matrix forms instead of referencing R's `lm` package.

10 Discussion

Despite its original intention of predictive modeling, the AIC is widely used in many scientific fields for explanatory modeling and structural equation modeling (Shmueli 2010). As a result, the validity and interpretation of these results has been contested (Spanos 2010). Even though commonly acknowledged rules of thumb for safeguarding against choosing between highly competitive models, the constant values used in these rules lead to unpredictable results, because pairwise score differene distributions vary greatly based on the the choice of models as well as the true data generating process. This is clear for the rule of adopting the model with the minimum AIC as well as using the (Burnham and Anderson 2002) recommended threshold of -10, as evidenced by (Spanos 2010) and the results of this paper. The use of these constant rules, then, creates a non-uniform standard of evidence across the practice of information-theoretic model selection. Moreover, this non-uniformity, as well as the statistical justifications for model selection as compared to those for hypothesis testing, lead to a discrepancy in the

implications of the results of these methods.

Attempts have been made to improve model selection by computationally estimating distributions of differences of model scores (Williams 1973, Wagenmakers et al. 2004). However, in (Williams 1973), the resulting estimates of difference distributions were not used in a principled manner, leading to an ambiguous interpretation of results of this method. The basis for setting decision thresholds in (Wagenmakers et al. 2004) claims to achieve equal error rates between two models, but it should be noted that the method in this paper actually stems from statistical classification instead of testing. While optimizing a confusion matrix in classification settings is a relevant problem, an underlying assumption in this context is that observed data is drawn from differing sub-populations, i.e. has been generating by differing processes (James, Witten, Hastie, and Tibshirani 2013). The experiment performed in (Wagenmakers et al. 2004) sought to compare competing theories of perception, results which would be generalized to the entire population. Using classification principles in this context has, at best, unclear motivation, and at worst questionable validity and interpreta-

tion.

This paper has presented a general framework for model selection with notions of Type-I and Type-II error, and thus for selecting decision functions based on controlling error. These decision functions may be compared using statistical sensitivity, which factors out the error inherent in the score function in the procedure. Moreover, a bootstrap implementation of a decision function with approximate maximum error rate α) has been introduced and demonstrated on multiple use cases and data types. It was seen in these experiments that the desired error rate was approximately obtained in all cases, and that when fixing α, the decision function has the desirable property of increased sensitivity as sample size increases. A demonstrable, easy to implement method for error controlled model selection has wide application in the natural sciences, and provides an opportunity for reexamining past studies under a new standard of evidence.

Following the results of the error controlled model selection procedure, one is motivated to examine the individual elements of the procedure. The general model selection framework gives the benefit of

distinctly separate steps which can be substituted in a modular fashion.

Alternative Score Criteria While the BITSEC has employed the AIC, there are a number of information criteria available for use that, depending on context, possess desirable properties as compared to the AIC. For instance, there exists a modified AIC intended for small sample sizes (AICc). In addition, one might consider the Bayesian Information Criterion (BIC), which adjusts its complexity penalty with sample size, or the Extended Information Criterion (EIC), which estimates a complexity penalty through the use of the bootstrap. More work is needed for comparing the relative merit of these criteria with respect to the error controlled framework, and they are likely dependent on the models and sample size used in an experiment.

Parameter Estimation The use of parametric bootstrap for resampling data allows for easy application to non i.i.d. data, but relies on estimated model parameters that have been shown to be biased by conditioning on the observed best fitting model. While a method for reducing this bias has been implemented, it was not shown or demonstrated

to remove bias completely. This indicates future work in improved parameter estimation in the model selection context. The problem of parameter estimation is not restricted to this application however. It should be noted that practitioners of model selection often wish to report not only the correct model, but approximate parameters corresponding to a real-world process. Thus, if these estimates are biased by observing a model as best, theoretical conclusions drawn from them can be inaccurate. Even if a score criterion can be shown to asymptotically attain a 100% probability of selecting the true model, bias in finite samples carries strong implications, especially for fields such as paleobiology in which research depends on small populations.

Smoothing the Bootstrap The bootstrap method used by BITSEC gives an empirical distribution as an approximation of the true distribution. Methods have been shown for fitting a curve to this discrete distrubtion, giving a smoothed bootstrap (Efron 1979, Silverman 1987). The relative efficiency of nonsmoothed and smoothed bootstrap estimates of quantile functions is examined in (Falk and Reiss 1989).

Conclusion While improvements could potentially be made to the BITSEC procedure, the results in this paper suggest that achieving error controlled model selection is plausible, a goal which would be applicable to many scientific fields. The principles behind an error controlled framework improve the standard of evidence, and thus strengthen the conclusions drawn from the results of model selection analyses. Because the BITSEC procedure is designed in a flexible, computational manner and can be implemented through an R package, it has an ease-of-use which lends itself to being adopted by scientists and statistical practitioners. Thus, a transition to error controlled model selection for explanatory and structural equation modeling is not only desirable, but plausibly attainable.

References

William Bausman and Marta Halina. Not null enough: pseudo-null hypotheses in community ecology and comparative psychology. *Biology & Philosophy*, 33(3):1–20, 2018.

Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

Kenneth P Burnham. *Model selection and multimodel inference : a practical information-theoretic approach*. Springer, New York, 2nd ed.. edition, 2002.

Chris Donkin, Scott Brown, Andrew Heathcote, and Eric-Jan Wagenmakers. Diffusion versus linear ballistic accumulation: different models but the same conclusions about psychological processes? *Psychonomic bulletin & review*, 18(1):61–69, 2011.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

M Falk and R-D Reiss. Weak convergence of smoothed and nonsmoothed bootstrap quantile estimates. *The Annals of Probability*, pages 362–371, 1989.

Peter Hall, Wolfgang Härdle, and Léopold Simar. On the inconsistency of bootstrap distribution estimators. *Computational statistics & data analysis*, 16(1):11–18, 1993.

Joop J Hox and Timo M Bechger. *An introduction to structural equation modeling*. Structural Equation Modeling, 1998.

Gene Hunt. Fitting and comparing models of phyletic evolution: Random walks and beyond. *Paleobiology*, 32(4):578–601, 2006.

Gene Hunt. The relative importance of directional change, random walks, and stasis in the evolution of fossil lineages. *Proceedings of the National Academy of Sciences*, 104(47):18404–18408, 2007.

Gene Hunt. Evolutionary patterns within fossil lineages: model-based assessment of modes, rates, punctuations and process. *The Paleontological Society Papers*, 14:117–131, 2008.

Gene Hunt. Measuring rates of phenotypic evolution and the inseparability of tempo and modemeasuring rates of phenotypic evolution. *Paleobiology*, 38(3):351, 2012.

Gene Hunt, Melanie J Hopkins, and Scott Lidgard. Simple versus complex models of trait evolution and stasis as a response to environmental change. *Proc Natl Acad Sci U S A*, 112(16):4885–90, Apr 2015.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Sokol Koço and Cécile Capponi. On multi-class classification through the minimization of the confusion matrix norm. In *Asian Conference on Machine Learning*, pages 277–292, 2013.

Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer Science & Business Media, 2008.

Todd D Little, William A Cunningham, Golan Shahar, and Keith F Widaman. To parcel or not to parcel: Exploring the question, weighing the merits. *Structural equation modeling*, 9(2):151–173, 2002.

Gitta H. Lubke, Ian Campbell, Dan Mcartor, Patrick Miller, Justin Luningham, and Stéphanie M. Van Den Berg. Assessing model selection uncertainty using a bootstrap approach: An update. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(2):230–245, 2017.

Deborah Mayo. Experimental practice and an error statistical account of evidence. *Philosophy of Science*, 67(3):S193–S207, 2000.

Deborah G. Mayo. An error-statistical philosophy of evidence. In *The Nature of Scientific Evidence*. University of Chicago Press, 2004.

David Posada and Thomas R Buckley. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology*, 53(5):793–808, 2004.

Kristopher J. Preacher and Edgar C. Merkle. The problem of model selection uncertainty in structural equation modeling. *Psychological Methods*, 17(1):1–14, 2012.

Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

BW Silverman and GA Young. The bootstrap: To smooth or not to smooth? *Biometrika*, 74(3):469–479, 1987.

Aris Spanos. Akaike-type criteria and the reliability of inference: Model selection versus statistical model specification. *Journal of Econometrics*, 158(2):204–220, 2010.

Jodie B Ullman and Peter M Bentler. Structural equation modeling. *Handbook of Psychology, Second Edition*, 2, 2012.

J Vandekerckhove, D Matzke, and EJ Wagenmakers. Model comparison and the principle of parsimony. *Oxford Handbook of Computational and Mathematical Psychology*, 2015.

Quang H Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333, 1989.

Eric-Jan Wagenmakers. A practical solution to the pervasive problems ofp values. *Psychonomic bulletin & review*, 14(5):779–804, 2007.

Eric-Jan Wagenmakers and Simon Farrell. Aic model selection using akaike weights. *Psychonomic bulletin & review*, 11(1):192–196, 2004.

Eric-Jan Wagenmakers, Roger Ratcliff, Pablo Gomez, and Geoffrey J. Iverson. Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, 48(1):28–50, 2004.

D A Williams. Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures. *Biometrics*, 26(1), 1970.