Philosophy of Science (2023), **90**, 648–668
doi:10.1017/psa.2022.32

PSA PHILOSOPHY OF SCIENCE ASSOCIATION

PHILOSOPHY OF SCIENCE

ARTICLE

Simplicity and the Sub-Family Problem for Model Selection

Alireza Fatollahi[1]* and Kasra Alishahi[2]*

[1]Department of Philosophy, Princeton University, Princeton, USA and [2]Department of Mathematical Sciences, Sharif University of Technology, Tehran, Iran
*Corresponding author. Emails: alireza.fatollahi@gmail.com, alishahikasra@gmail.com

(Received 30 September 2021; revised 05 April 2022; accepted 07 April 2022; first published online 25 May 2022)

Abstract

Forster and Sober (1994) introduced the "sub-family problem" for model selection criteria that recommend balancing goodness-of-fit against simplicity. This problem arises when a maximally simple model (family of hypotheses) is artificially constructed to have excellent fit with the data. We argue that the problem arises because of a violation of the general maxim that balancing goodness-of-fit against simplicity leads to desirable inferences only if one is comparing models for the consideration of which one has a positive reason independently of the current data.

1. Introduction

Malcolm Forster and Elliott Sober's highly influential paper (1994) introduced to the philosophical literature some of the major developments of the latter half of the twentieth century statistics on the model selection problem. They argued that Akaike's results (1973) on how to correct for over-fitting the data sheds light on a number of topics in philosophy of science, especially the problem of explaining why simpler and less ad hoc theories have better predictions. However, Forster and Sober also posed a potential problem, which they call the "sub-family problem," that would arise if one uses model selection criteria in an ad hoc way. This is a problem for "any proposal that measures simplicity by the paucity of adjustable parameters," including the Akaikean one. They then offer a solution by showing how such ad hoc use of the Akaikean criterion is disallowed in the broader Akaikean framework because of a "meta-theorem" about the error in employing Akaike's results.

Although we find the sub-family problem interesting in itself, we think it can be of deeper philosophical value by illuminating some of the conditions under which respecting simplicity-favoring considerations results in desirable inferences. Our goal in this essay is to establish the following claim.

> **Independent Motivation Requirement (IMR).** Weighing considerations of simplicity against those of goodness-of-fit, as it is recommended by Akaike

> Information Criterion (AIC) or Bayesian Information Criterion (BIC), results in reliable inferences only if one's domain of options (i.e., the set of candidate models) consists of models for the inclusion of which one has a positive reason independently of the current data. Optimizing the balance between simplicity and goodness-of-fit can lead one astray, if the models are constructed post hoc or if one liberally adds to the number of models.

IMR is violated in two ways. First, when one allows knowledge of the extant data to play a role in the design of the model itself (post hoc model construction). Second, when the number of models is unduly large because one doesn't have a positive reason for taking some of the candidate models into consideration. In both events, simplicity-favoring considerations that are appealed to in model selection criteria cannot amend for systematic over-fitting.

We begin, in section 2, by introducing the Akaikean framework for model selection, AIC, and the sub-family problem. The problem equally applies to structurally similar model selection criteria, such as BIC. We discuss this issue and a solution for the problem in the BIC-based framework in section 3. This solution is a manifestation of our thesis that violating IMR results in untoward inferential practices. Sections 4 and 5 discuss two solutions to the problem in the Akaikean framework. In section 4, we examine Forster and Sober's solution and argue that, although much of what they say is true, their solution isn't *fully* satisfactory. In particular, it *appears* to follow from their solution that the AIC scores of simple models with excellent goodness-of-fit are "unreliable" (epistemically biased) estimates, even if there are reasons independent of the current data for considering those models (i.e., even if considering them as candidate models does not violate IMR). We argue that this idea is false. In section 5, we offer our own solution, according to which the problem arises because of a violation of IMR. In section 6, we talk about a case in which one can re-introduce essentially the same error involved in the sub-family problem by considering too many candidate models—and thereby violating IMR but not through post hoc construction. We will then explain why such a practice is problematic. In section 7, we conclude our argument for IMR, presenting the main ideas less technically. We end the paper by pointing out an important practical consequence of IMR. The Akaikean framework for model selection tells us why fudged hypotheses that fit the data well (i.e., best-fitting members of complex models) have poor predictions.[1] We will argue that IMR gives us a clear criterion for determining whether fixing certain parameters in an n-parameter model results in *fudged models* with poor predictions and must be avoided.[2]

2. The Setup of the Problem

In model selection, one is concerned with the comparison of families of hypotheses (hereafter "models"). We will reserve the term "hypothesis" for individual members of models. For example, "$y = 2x+3+N(0,1)$" is a hypothesis belonging to the model

[1] See Forster (2006) for how the Akaikean framework helps distinguish between meritorious and fudged fit.

[2] André Kukla (1995) suggests another problem similar to the sub-family problem in which models are constructed by consulting the data. See also Forster (1995). As we mention in footnote 25, IMR disallows Kukla's problem as well.

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

{y=ax+b+N(0,1)}. Typically, one's background theory specifies a finite set of candidate models prior to consulting the data and one is interested in comparing them based on the data. For example, background theory may restrict the set of plausible models to polynomials of degrees no more than 5. One's goal can then be to find some weighted ordering of such models (here, degrees of polynomials) with respect to various desirable features, such as predictive accuracy or posterior probability.

Suppose you have a suitably large set of observational data consisting of ordered pairs of values of two variables, *X* and *Y*, generated by an unknown "true" function, *T*. The data might, for example, record the length of a metallic bar in different temperatures. Your background theory tells you that *X*s and *Y*s are linearly related (though because of the existence of error, the observed values might not exactly fit a line). You want to find the particular linear function that best fits the data. This is a rather straightforward statistical problem and doesn't involve model selection, because you are considering only one model (you know *X* and *Y* are linearly related). The standard solution on which there is consensus among statisticians is to find the line with maximum likelihood relative to the data, where the likelihood of a hypothesis is defined as the probability (or probability density) of the data conditional on the hypothesis.

A much more difficult question arises if the inference problem concerns the choice between models, say between parabolic functions and linear functions. Here, one cannot simply maximize likelihood, because to do so would almost always lead one to choose a parabolic function (or generally, a member of the largest model). More complicated models have more freedom to fit the data. Thus, the likeliest members of those models are more likely to fit the noise, as opposed to the main pattern, in the data. This is called "over-fitting." There are various model selection techniques for how to avoid over-fitting. Most of them recommend balancing considerations of goodness-of-fit with data against considerations of simplicity, though the "optimal" balance is naturally different for different techniques, since they either pursue different goals or make different assumptions about the inference problem or both.

Before we proceed, some notational conventions must be mentioned. In order to avoid confusion, we refer to random variables by capital letters, single pieces of data by lower-case letters and data sets by bolded lower-case letters. In order to distinguish between specific data sets and data sets considered as random variables, we show the former by subscripted, bolded lower-case letters, like $\mathbf{y_0}$, and the latter by non-subscripted, bolded lower-case letters, like, $\mathbf{y}$. Consider a model, $F(\theta)$, defined over a parameter space Θ. For example, if F is the family of linear functions with a normally-distributed error with mean 0 and variance 1, it can be characterized as F: {y = ax+b+N(0,1); (a,b) ∈ R^2}. Here the parameter space of F is R^2.[3] We can associate a likelihood function, $\mathscr{L}(\theta) = \mathscr{L}(\theta,\mathbf{z}) = \mathrm{g}(\mathbf{z}|\theta)$, to F, where $\mathrm{g}(\mathbf{z}\,|\theta)$ is the probability density of obtaining the n-tuple data set $\mathbf{z} = \{\mathrm{z}_1,\mathrm{z}_2,\ldots,\mathrm{z}_n\}$ [where z_i is the ith observed datum (x_i,y_i)] conditional on θ being the true parametric value. The value

[3] We don't need to assume that the variance of error is known or that it is normally distributed. It is sufficient if we can characterize the distribution of error by a general parametric equation, in which case the unknown parameters of the error distribution are part of the parameters of the model. For example, if F is the family of linear functions with a normally distributed error of unknown variance (F:{y=ax+b+N(0,σ)}), the parameter space of F is three-dimensional.

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

of θ for which $\mathscr{L}(\theta)$ is maximized is called the Maximum Likelihood Estimate (MLE) of *F*, and we denote it by $\hat{\theta}$. We denote the member of *F* obtained by taking $\theta = \hat{\theta}$ by L(F).[4] Note that both $\hat{\theta}$ and L(*F*) are functions of data; that is, $\hat{\theta} = \hat{\theta}(\mathbf{y})$ and L(*F*) = L(*F*,**y**), where **y** is the data.

In the Akaikean framework, the goal is to find a plausible estimate of the predictive accuracies of various models. The predictive accuracy of a model, *M*, is a measure of how close, on average, the predictions of the best fitting member of *M*, with respect to an initial data set, are to subsequent data generated by the same generating function. Suppose you obtain a data set, **y**, generated by *T* and use it to determine L(*M*,**y**). Now obtain a new data set, **x**, and calculate the logarithm of its likelihood (hereafter, "log-likelihood") of L(*M*,**y**) with respect to **x**. The average value—with respect to both **x** and **y**—of this log-likelihood is called the predictive accuracy of *M*. The following defines A(*M*), the predictive accuracy of model *M*.[5]

$$\mathrm{A}(M) =^{\mathrm{df}} \mathrm{E}_{\mathbf{y}}\mathrm{E}_{\mathbf{x}}[\log(\mathrm{g}(\mathrm{x}|\hat{\theta}(\mathrm{y})))] = E_{\hat{\theta}}\mathrm{E}_{\mathbf{x}}[\log(\mathrm{g}(\mathrm{x}|\hat{\theta}))] \tag{1}$$

$\mathrm{E}_{\mathbf{y}}(.)$ and $\mathrm{E}_{\mathbf{x}}(.)$ are both expectations with respect to *T*. The last term on the right-hand side better captures the nature of predictive accuracy, by emphasizing that it averages over the MLE.[6]

Akaike showed that the AIC value of a model can be used to provide an estimate of its predictive accuracy. AIC of model *F* is defined thus.

$$\mathrm{AIC}(F) =^{\mathrm{df}} -2\log\mathscr{L}(\mathrm{L}(F)) + 2k \tag{2}$$

$\log\mathscr{L}(\mathrm{L}(F))$ is the log-likelihood of L(*F*). If error is normally distributed, this becomes the familiar sum of squares of error terms of L(*F*). *k* is the dimension of the parameter space of *F*, which is usually equal to its number of adjustable parameters.

If one is interested in finding the model with the highest predictive accuracy, the Akaikean framework[7] recommends choosing the model with minimum AIC. However,

[4] Note that "L(*F*)" refers to the best fitting *member* of *F*, while $\mathscr{L}(\theta)$ is the likelihood function over the parameter space of *F*. We use this notation in order to maximally stick to the notation used by Forster and Sober.

[5] Forster and Sober define predictive accuracy as the average *per datum* of this double expectation, "so that the accuracy of a hypothesis does not change when we consider the prediction of data sets of different sizes" (Forster and Sober 1994, 10). We find this choice sensible. However, since such minutia will not affect the soundness of our argument, we will use the simpler definition in equation (1).

[6] More precisely, the goal is to obtain a reliable estimate of the relative Kullback-Leibler (KL) distance from the truth of different candidate models. The KL distance of probability density *g* from *f* is defined by: $\mathrm{D}_{\mathrm{KL}}(f\|g) =^{\mathrm{df}} \mathrm{E}_f[\log(\frac{f}{g})]$, where $\mathrm{E}_f(.)$ is the expectation with respect to *f*. This equation does not define the KL distance of a model from *f*, since a model does not have a single probability density. However, a plausible way to understand the KL distance of model *M* from truth, *f*, is the average KL distance from *f* of L(*M*,**y**) (average with respect to **y**). Turns out this quantity is equal to the predictive accuracy of *f* minus the predictive accuracy of *M*. Since the predictive accuracy of *f* is an unknown constant depending only on *f*, the KL distance of a model from truth is equal to a constant minus its predictive accuracy—the more the accuracy, the less the distance. Thus, to maximize predictive accuracy and to minimize KL distance from the truth are effectively the same goal.

[7] A referee suggested that we clarify what we mean by the "Akaikean framework." Here is a characterization largely borrowed from Forster and Sober (2011). The *goal* of this framework is to estimate predictive accuracy. The *criterion* by which one approaches this goal is AIC. This practice is justified by Akaike's result that AIC is an unbiased estimator of predictive accuracy. Note that not every model selection practice that tries to find a balance between goodness-of-fit and simplicity belongs to this

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

as Forster and Sober observe, blindly minimizing AIC is problematic. Suppose one wishes to find the predictively most accurate hypothesis among polynomials of degrees 3 or less. One must find the family with minimum AIC and then select the likeliest member of that family. Let M_i be the family of polynomials of degree *i-1* with *i* adjustable parameters. Suppose M_3 turns out to have the lowest AIC value. Since the family of parabolic functions is embedded in the family of cubic functions, the likeliest member of the cubic family ($L(M_4)$) has a better (or equally good) fit with the data than the likeliest member of the parabolic family ($L(M_3)$). Now construct an ad hoc family, $\{L(M_4)\}$, whose only member is $L(M_4)$. The number of adjustable parameters in $\{L(M_4)\}$ is 0 (because it is a singleton family) and therefore, its AIC value is equal to $-2\log\mathscr{L}(L(M_4))$. We have, $AIC(\{L(M_4)\}) = -2\log\mathscr{L}(L(M_4)) \leq -2\log\mathscr{L}(L(M_3)) < AIC(M_3)$, where $\log\mathscr{L}(L(M_4))$ is the log-likelihood of the best fitting member of M_4. Indeed, $\{L(M_4)\}$ has the lowest possible AIC value (relative to the extant data) among all families the members of which are restricted to polynomials of degrees 3 or less. Thus, if we blindly minimize AIC scores, we must choose $\{L(M_4)\}$ (hereafter, "the sub-family model") as the predictively most accurate model, which is tantamount to choosing the likeliest hypothesis at our disposal and giving no weight to simplicity. This is what Forster and Sober have called the sub-family problem.[8]

In order to show the importance of IMR, we will contrast the sub-family problem with another inference problem, which is very similar but differs in only one salient way: in that problem, IMR *is* respected. Suppose you have reasons independently of the extant data (e.g., theoretic reasons) for including the singleton model $\{5x^3+6x^2+2x+7+N(0,1)\}$ among your candidate models. *Then* you obtain a data set, $\mathbf{y_0}$, and you observe that $L(M_4,\mathbf{y_0})=5x^3+6x^2+2x+7+N(0,1)$.[9] This inference problem, which we will call the singleton family problem,[10] is very similar to the sub-family problem. The set of candidate models ($M_1,M_2,M_3,M_4,\{L(M_4,\mathbf{y_0})\}$) and the data ($\mathbf{y_0}$) are the same. However, they have an important difference: here IMR is respected because you had reasons to include $\{5x^3+6x^2+2x+7+N(0,1)\}$ (which happens to be identical with $\{L(M_4,\mathbf{y_0})\}$) among your candidate models. As we shall see, this makes a big difference.

framework. For example, the BIC-based framework has a different goal and a different criterion (BIC), but tries to strike a balance between goodness-of-fit and simplicity.

[8] A referee objected that the sub-family problem might be a "pseudo-problem," because the way in which $\{L(M_4)\}$ was constructed "is a way of adjusting 4 parameters and it should be penalized as such." Notice that the complexity of a model is the dimension of its parameter space and the parameter space of $\{L(M_4)\}$ is zero-dimensional. Thus, the AIC score of the sub-family model does not reflect the fact that $\{L(M_4)\}$ was constructed by adjusting 4 parameters. And that is a problem. A proper solution must account for this fact by telling us why the AIC score of $\{L(M_4)\}$ is not a good estimate in this case, despite Akaike's result that AIC is an unbiased estimator.

[9] The exact equality in this condition is unlikely to happen even if $5x^3+6x^2+2x+7+N(0,1)$ is the truth. This is an idealization, which we have made for its theoretic simplicity. However, our argument is applicable, mutatis mutandis, to the more realistic case in which $L(M_4,\mathbf{y_0}) \approx 5x^3+6x^2+2x+7+N(0,1)$. We talk about the more realistic case in the context of a numerical example in section 4.

[10] As we shall argue, both BIC and AIC treat this problem very differently from the sub-family problem. We call it a "problem" only because it involves an inference problem, not because there is anything problematic in it.

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

3. BIC

In the BIC-based framework, the average likelihood of a model[11] is estimated by −½ times exponential of its BIC defined as follows,

$$\mathrm{BIC}(F) =^{\mathrm{df}} -2\log\mathscr{L}(\mathrm{L}(F)) + k\log(n), \tag{3}$$

where n is the number of data points. Lower BIC scores are better. Since this is a Bayesian framework, its ultimate goal is to determine the posterior probabilities of candidate models. It is customary—though by no means necessary—to assign equal prior probabilities to all candidate models. Thus, the most probable model is often the one with the lowest BIC value. Now, the sub-family model has the lowest possible BIC value, since it has no adjustable parameters and its only member has maximum log-likelihood. Thus, we are faced with the sub-family problem: the sub-family model appears to be the most probable model.

A referee suggests that some of the general problems for the BIC-based framework might complicate our solution to the sub-family problem. Thus, before offering our solution, we will briefly mention one of those problems, which is most relevant. If the models are nested (as in our example of polynomial models), then larger models entail smaller ones. (The set of polynomials of degree n contains the set of polynomials of degree n-1.) It follows that $\mathrm{P}(M_n) \geq \mathrm{P}(M_{n-1})$, no matter what the data is. If so, it is difficult, in this framework, to make sense of the fact that scientists sometimes prefer smaller models to larger ones. To the best of our knowledge, there has been no fully satisfactory response to this problem. Forster and Sober (1994) discuss the following way to address this difficulty. Instead of M_2, construct $M_2^* = M_2 - M_1$ (and so on for larger models), so that no model entails any other. Then compare those newly-formulated models. Forster and Sober find this maneuver unsatisfactory, because it changes the subject. The question was why scientists prefer M_1 to M_2, not M_1 to M_2^*. This is a fair objection, a proper reply to which (if it exists) goes beyond the scope of this essay. Here we use this *mathematical maneuver* to offer a solution for the sub-family problem, but we don't claim to offer a solution for the above problem or any other general problem for BIC. All we wish to establish is that if BIC can be rescued from the general difficulties it faces, it won't be *further* subject to the sub-family problem.

Construct the non-nested models (M_1-M_4^*) as described above. Make sure that the prior probability functions over the members of your non-nested models do not have probability masses. That is, if $\pi_4(\theta)$ is the prior probability function over the members of M_4^*, then make sure for no single value of θ, $\pi_4(\theta) > 0$. (If this is not the case, BIC is not a good approximation of average likelihood. This is another limitation of BIC, again, independently of the sub-family problem.) How can you make sure this is the case? If for θ_1, $\pi_4(\theta_1) > 0$, construct the singleton model $\{\theta_1\}$, where $\mathrm{P}(\{\theta_1\}) = \pi_4(\theta_1)$. Then redefine M_4^* in the following way, $M_{4\,new}^* = M_{4\,old}^* - \{\theta_1\}$, with $\mathrm{P}(M_{4\,new}^*) = \mathrm{P}(M_{4\,old}^*) - \mathrm{P}(\{\theta_1\})$. Once this is done, BIC can be used as an estimate of average likelihood of the non-nested models. Now suppose we obtain a data set $\mathbf{y_0}$ and

[11] The likelihood of a model, unlike that of a hypothesis, isn't determined solely by the content of the data and the model. We have $\mathscr{L}(F(\theta),\mathbf{y}) = \mathrm{P}(\mathbf{y}|F(\theta)) = [\mathrm{P}(y|\theta)]\pi(\theta)d\theta$, where the integral is over the parameter space of F and $\pi(\theta)$ is the prior probability function over the members of F. Therefore, to talk about the likelihood of a model is to talk about its *average* likelihood.

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

$\{L(M_4,\mathbf{y_0})\} = \{5x^3+6x^2+2x+7+N(0,1)\}$. There are two possibilities. Either $P(\{5x^3+6x^2+2x+7+N(0,1)\}) > 0$, or $P(\{5x^3+6x^2+2x+7+N(0,1)\}) = 0$, where P(.) is the prior probability function.

If $P(\{5x^3+6x^2+2x+7+N(0,1)\}) > 0$, then you had a positive reason to include $\{5x^3+6x^2+2x+7+N(0,1)\}$ among your candidate models independently of the data[12] and this is an instance of the singleton family problem—considering $(\{5x^3+6x^2+2x+7+N(0,1)\})$ does not violate IMR. The likelihood of a model is proportional to the *exponential* of −½BIC and since $\{5x^3+6x^2+2x+7+N(0,1)\}$ has an exceptionally low BIC, its likelihood will be massively higher than other models. Thus, unless its prior probability is extremely lower than other models, it will be by far the most probable model among the non-nested ones. This is a welcome result. If one had theoretic reasons that a single parameter value has positive probability, and one subsequently learns that this single hypothesis fits data excellently well, the data provides very powerful evidence for that hypothesis and ought to make one significantly more confident of its truth. Again, this doesn't solve the above-mentioned general difficulty about BIC. In the singleton family problem, $\{5x^3+6x^2+2x+7+N(0,1)\}$ might end up having a higher posterior probability than M_4^* but it can never have a higher probability than M_4.

The sub-family problem corresponds to the case where $P(\{5x^3+6x^2+2x+7+N(0,1)\}) = 0$. Here you have no reason to include $\{5x^3+6x^2+2x+7+N(0,1)\}$ among the candidate models independently of the data. Thus, to choose $\{5x^3+6x^2+2x+7+N(0,1)\}$ because of its excellent BIC score involves violating IMR. The BIC-based framework disallows choosing $\{5x^3+6x^2+2x+7+N(0,1)\}$, because if $P(\{5x^3+6x^2+2x+7+N(0,1)\})=0$, no matter how likely $\{5x^3+6x^2+2x+7+N(0,1)\}$ is (no matter how good its BIC score is), its posterior probability will remain zero. As we shall see, although the Akaikean framework doesn't appeal to model priors, it equally presupposes IMR.

4. Forster and Sober's Solution

For large data sets, AIC is an approximately unbiased estimator of predictive accuracy. An estimator is statistically unbiased if its expected value equals the value it estimates. The following equation expresses this fact.

$$E_{\mathbf{y}}\left[-\tfrac{1}{2}\mathrm{AIC}(M,\mathbf{y})\right] = A(M), \tag{4}$$

where $\mathbf{y}$ is a random data set. In this paper, we are not concerned with the approximate nature of this equation. So we will talk about the "unbiasedness" of AIC for convenience.

Forster and Sober argue that the AIC of the sub-family model is an unbiased estimator of its predictive accuracy, but statistical unbiasedness is not the only criterion by which to judge an estimate. They introduce another such criterion called "epistemic unbiasedness" by the following example. Consider a simple measurement of an object's mass with a kitchen scale. Normally, the measured value is an unbiased estimator of the actual value, because it is just as likely to over-measure the mass by a given amount as it is to under-measure it by that same amount.

[12] Notice that typically no single hypothesis has a positive prior probability. So when it does, you must have a reason for thinking that the hypothesis is particularly good.

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

> But now suppose that we modify this estimate by adding +10 or -10 depending on whether a fair coin lands heads or tails, respectively. Suppose that the measured value of mass was 7 kg, and the fair coin lands heads. Then the new estimate is 17 kg. Surprisingly, this new estimate is also a statistically unbiased estimate of the true mass! The reason is that in an imagined series of repeated instances, the +10 will be subtracted as often as it is added, so that the value of the average value of the modified estimate will still be equal to the true mass value. However, we know that the modified estimate is an overestimate in this instance, because we know that the coin landed heads. If the coin had landed tails, then the estimate would have been -3 kg, and would have been known to be an underestimate. In either case, we say that the modified estimate is epistemically biased. (Forster and Sober 1994, 19)

It is helpful to make a distinction between an "estimator" and an "estimate" here. An *estimator* is a function of data that yields individual *estimates*. Statistical unbiasedness is a feature of an estimator, while epistemic bias is a feature of an individual estimate. Forster and Sober argue that AIC of the sub-family model is statistically unbiased (qua estimator) but epistemically biased (qua estimate).[13] They argue that the AIC of the sub-family model is not a good estimate because it is epistemically biased. To show this, they appeal to the following "meta-theorem" about AIC.

Error[Estimated(A(F))] $=^{df}$ A(F) – (−½)AIC(F) = Residual Fitting Error + Common Error + Sub-family Error (Forster and Sober 1994, 19).

This theorem concerns the error involved in taking −½AIC as an estimate of predictive accuracy. Forster and Sober argue that the first two terms on the right-hand side are both statistically and epistemically unbiased, but the third term, although statistically unbiased, is sometimes epistemically biased, which (given the epistemic unbiasedness of the other two terms) makes the total error sometimes epistemically biased. They further argue that an important occasion in which this happens is in the sub-family problem. Thus, a fuller understanding of the Akaikean framework, which includes this meta-theorem, dissolves the sub-family problem.

Here is why Forster and Sober believe the sub-family error is epistemically biased for the sub-family model. Suppose we embed the parameter spaces of all our models in a larger parameter space (call it K) that contains the truth, T. Forster and Sober state that this space can be considered as a vector space in such a way that i) the closer a point in this space is to truth, the higher its predictive accuracy; and ii) the sub-family error is equal to the scalar product of the following two vectors in this space: the vector that goes from T to the likeliest hypothesis in K, L(K), (i.e., $\overrightarrow{T.\mathrm{L}(K)}$) and the vector $\overrightarrow{T.\theta_0}$ that goes from T to the (unknown) predictively most accurate member of M, which we denote by θ_0.[14] These vectors are shown in figure 1 below.[15]

[13] Forster and Sober don't offer an exact definition for epistemic bias. A referee expressed skepticism about the usefulness and clarity of the concept. Although we disagree with this assessment, notice that the idea that epistemic bias is hopelessly unclear, if true, adds to the motivation for our solution, because our solution only uses the notion of statistical bias. All discussion of epistemic bias in this paper is for the purposes of examining Forster and Sober's solution.

[14] θ_0 is generally different from $\hat{\theta}$. The former is the member of the model that is predictively most accurate and doesn't depend on data. The latter is the likeliest member given the extant data.

[15] Figure 1 essentially represents the same idea represented by figure 4 in Forster and Sober (1994), although our notations are slightly different.

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

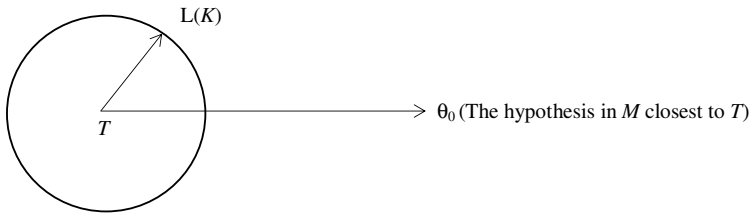

Figure 1. *T* is the truth. L(*K*) is the likeliest hypothesis in *K*, which contains *T*. For a Gaussian distribution of error, L(*K*) is equally likely to fall on any point on the circle. *T* and L(*K*) are independent of the model. θ_0 is the predictively most accurate hypothesis in the model (i.e., the closest member of the model to *T*).

All three points are in general unknown. However, the scalar product of $\overrightarrow{T.\mathrm{L}(K)}$ and $\overrightarrow{T.\theta_0}$ is equal to the product of their lengths multiplied by the cosine of the angle between them. Forster and Sober argue that for the sub-family model, the tips of the two vectors tend to be close. Here is their argument.

> The Akaike estimate for a low dimensional family whose best fitting case is close to the data (and such families are the dangerous "pretenders," for they "unfairly" combine high log-likelihoods with small penalties for complexity) exhibits an epistemic bias, as we now explain. The most predictively accurate hypothesis in such small families will also be close to the data, and therefore close to *L(K)*. The danger is that the tips of the two vectors will be close together. Then the cosine factor is close to +1 and the subfamily error is large *and positive*.[16] (Forster and Sober 1994, 20–21)

On Forster and Sober's view, we shouldn't follow the sub-family policy, because if we do, the AIC values of the models we construct, although statistically unbiased, tend to be epistemically biased (over)estimates of the predictive accuracies of these models.

Before discussing what we find potentially problematic about this argument, we would like to give a summary of what we will claim, in order to avoid confusion. When Forster and Sober say that the AIC of the sub-family model is statistically unbiased, they are referring to the fact that the estimate AIC($\{\mathrm{L}(M_4,\mathbf{y_0})\},\mathbf{y_0}$) is an instantiation of the estimator AIC($\{\mathrm{L}(M_4,\mathbf{y_0})\},\mathbf{y}$), which is—we agree—an unbiased estimator. We also agree that in the sub-family problem, AIC($\{\mathrm{L}(M_4,\mathbf{y_0})\},\mathbf{y_0}$) is epistemically biased. However, we suggest that a better estimator (than AIC($\{\mathrm{L}(M_4,\mathbf{y_0})\},\mathbf{y}$)) for determining the (de)merits of the estimate AIC($\{\mathrm{L}(M_4,\mathbf{y_0})\},\mathbf{y_0}$) *in the sub-family problem* is AIC($\{\mathrm{L}(M_4,\mathbf{y})\},\mathbf{y}$), which is statistically biased. This is because in the sub-family problem, AIC($\{\mathrm{L}(M_4,\mathbf{y_0})\},\mathbf{y_0}$) is more similar to other instantiations of AIC($\{\mathrm{L}(M_4,\mathbf{y})\},\mathbf{y}$) than to other instantiations of AIC($\{\mathrm{L}(M_4,\mathbf{y_0})\},\mathbf{y}$). The exact opposite situation holds for the singleton family problem. *There* the statistical bias of AIC($\{\mathrm{L}(M_4,\mathbf{y_0})\},\mathbf{y}$) provides more relevant information (than that of AIC($\{\mathrm{L}(M_4,\mathbf{y})\},\mathbf{y}$)) for deciding how good the estimate AIC($\{\mathrm{L}(M_4,\mathbf{y_0})\},\mathbf{y_0}$) is. Therefore, we will argue that the Akaikean framework treats the sub-family problem and the singleton family problem differently.

[16] Forster and Sober define AIC as the negative value of our equation (2); thus, a positive error means overestimation of predictive accuracy.

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

Now we will unpack this. Consider the singleton family problem first. That is, suppose we had reasons independently of $\mathbf{y_0}$ for considering the model $\{5x^3+6x^2+2x+7+N(0,1)\}$. *Then* we obtained $\mathbf{y_0}$ and observed that $L(M_4,\mathbf{y_0})=5x^3+6x^2+2x+7+N(0,1)$. Thus, including $\{5x^3+6x^2+2x+7+N(0,1)\}$ among the models doesn't violate IMR. Is $AIC(\{5x^3+6x^2+2x+7+N(0,1)\},\mathbf{y_0})$ an epistemically biased estimate in this case, according to Forster and Sober? We don't know for sure because they don't discuss this problem. However, if we are to infer from the *literal* meaning of their argument, they would have to say "yes," because in every relevant respect to their solution, the two problems are identical. In their argument, the relevant factors are the low-dimensionality of the model and its closeness to data, which are shared in the two problems. *If they are committed to this idea*, then this is the only point of disagreement between our account and theirs. We believe that in the singleton family problem, the exceptionally low AIC score of $\{5x^3+6x^2+2x+7+N(0,1)\}$ *is* exceptionally good reason that $\{5x^3+6x^2+2x+7+N(0,1)\}$ has high predictive accuracy. There is nothing "unfair" in the AIC score of a low-dimensional model with good fit per se. The AIC scores of such models are "too good to be true" only if the model has these features *because* it was designed in an ad hoc fashion to have a low AIC score.

For this to be true, something must be missing in Forster and Sober's account of the sub-family error *as applied to the singleton family problem*. Here is what is missing. They offer a consideration (hereafter consideration$_1$) that θ_0 tends to be close to $L(K)$, which leads to the sub-family error for $\{5x^3+6x^2+2x+7+N(0,1)\}$ being large and positive. However, there is a *competing* consideration (hereafter, consideration$_2$) that mitigates the effect of consideration$_1$: for models with low AIC scores, θ_0 tends to be close to T. (A low AIC score means high predictive accuracy, which means closeness to truth.) Consideration$_2$ is *a* reason for sub-family error to be small, because the error is equal to the product of the lengths of the two vectors $\overrightarrow{T.L(K)}$ and $\overrightarrow{T.\theta_0}$ times the cosine of the angle between them. Consideration$_2$ is *a* reason that the length of $\overrightarrow{T.\theta_0}$ is small. In both the sub-family problem and the singleton family problem, $AIC(\{5x^3+6x^2+2x+7+N(0,1)\},\mathbf{y_0})$ is exceptionally low, which is a reason to think that $5x^3+6x^2+2x+7+N(0,1)$ is close to truth. This is the case, unless one *otherwise* knows that $AIC(\{5x^3+6x^2+2x+7+N(0,1)\})$ is *not* a good estimate of the predictive accuracy of $\{5x^3+6x^2+2x+7+N(0,1)\}$. We will argue in the next section that in the sub-family problem, one knows this independently. Thus, consideration$_2$ is irrelevant to the sub-family problem and $AIC(\{5x^3+6x^2+2x+7+N(0,1)\},\mathbf{y_0})$ is epistemically biased in that problem. However, our argument in the next section doesn't apply to the singleton family problem. In that problem, we are left with two competing considerations bearing on how good an estimate $AIC(\{5x^3+6x^2+2x+7+N(0,1)\},\mathbf{y_0})$ is, and in general we have no way of comparing the relative strengths of these considerations. Notice that consideration$_2$ doesn't tell us anything about the sign of the sub-family error (whether it is negative or positive). Thus, we must expect to have a positive epistemic error but smaller in absolute value relative to the sub-family problem.

The *amount* of bias matters a lot. In both the sub-family problem and the singleton family problem, $AIC(\{5x^3+6x^2+2x+7+N(0,1)\},\mathbf{y_0})$ is *significantly* better than the AIC of other models. For example, $AIC(\{5x^3+6x^2+2x+7+N(0,1)\},\mathbf{y_0}) = AIC(M_4,\mathbf{y_0}) - 4$,[17]

[17] Since $AIC(M_3) \geq AIC(M_4) - 1$, $AIC(\{5x^3+6x^2+2x+7+N(0,1)\},\mathbf{y_0})$ is at least 3 units better than $AIC(M_3,\mathbf{y_0})$ and so on for other models.

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

and 4 units of AIC difference is usually a big difference. (To have an intuitive idea for why this is the case, consider the fact that if the data is *perfectly* linear, M_2 will have only a 2 units AIC advantage relative to M_4.) In the sub-family problem, the average sub-family error is exactly 4, which means that the *entire* difference in AIC score (between $\{5x^3+6x^2+2x+7+N(0,1)\}$ and M_4) is due to error. However, in the singleton family problem, little of the AIC difference between $\{5x^3+6x^2+2x+7+N(0,1)\}$ and M_4 (which is 4 units) is due to epistemic error. Therefore, $\{5x^3+6x^2+2x+7+N(0,1)\}$ is in *great* shape relative to M_4. This is a welcome result. If you have a singleton model among the candidate models for reasons independently of the extant data, and it fits the data as precisely as the family of cubic functions, you ought to be very confident of its predictive accuracy *even if its AIC score is slightly biased*.

The following numerical example can help illustrate our point. Here all the data is generated by the function $T= 0.2x^5-0.2x^4-3x^3+x^2-1+N(0,1)$. Figure 2a depicts the sub-family error of the model $\{L(M_4,\mathbf{y})\}$ for 10^7 values of $\mathbf{y}$ each consisting of 100 data points.[18] That is, for each data set, $\mathbf{y}$, $L(M_4,\mathbf{y})$ is determined separately and then its corresponding sub-family error is calculated.[19] This is essentially a 10^7 repetition of the sub-family problem. Evidently, the sub-family error of $\{L(M_4,\mathbf{y})\}$ tends to be positive and large, as Forster and Sober rightly argue. Figure 2b depicts the sub-family error of the fixed singleton model M_0:$\{-3x^3+x^2-1+N(0,1)\}$, which is chosen because it is close to T. The sub-family error of this fixed model is statistically unbiased, again as observed by Forster and Sober.

The difficult part is how to simulate the singleton family problem, because in that problem you must have a singleton family for reasons independently of the data and then something amazing happens: the only member of that model turns out to be equal to the best fitting member of your largest model. (Even if your singleton family contains the truth, this is unlikely to happen, unless the data is huge.) However, we can approximate this situation. Figure 3a is the histogram of the distance between M_0 and $\{L(M_4,\mathbf{y})\}$ for the 10^7 data sets. Instead of looking at cases where $\{-3x^3+x^2-1+N(0,1)\}=\{L(M_4,\mathbf{y})\}$, we first looked at cases in which M_0 is "close" to $\{L(M_4,\mathbf{y})\}$. This corresponds to cases where the member of the singleton family is not exactly identical with $L(M_4,\mathbf{y})$ but is close to it (both in the parameter space and in terms of log-likelihood). In 3b, the sub-family error for M_0 is depicted only for those cases in which $||M_0\text{-}L(M_4,\mathbf{y})||< 0.3$ (cases that are left of the red line in 3a). The choice of 0.3 is arbitrary. We chose this value so that we can still have a large number of cases. (Smaller values result in larger errors because for them consideration$_1$ becomes stronger, but due to the small number of cases, the histograms become very jagged.)

The average sub-family error for cases depicted in figure 3b is 0.4286. When we took this condition to the limit (i.e., $||M_0\text{-}L(M_4,\mathbf{y})||\to 0$), the error approached a value slightly less than 1. A comparison between 2a and 3b shows the mitigating effect of consideration$_2$. In 2a, the average $L(M_4,\mathbf{y})$ is not particularly close to T; thus, the only relevant consideration is consideration$_1$. The average sub-family error was

[18] The values of X were first generated randomly from the uniform distribution on [-1,1]. Then those values were fixed throughout the data generating process.

[19] The sub-family error is the second item in the right-hand side of equation 4.55 in Sakamoto et al. (1986, 77). We have used -1 times the value in that equation in order to simplify things, so that a positive error means (as Forster and Sober also mean it) an overestimation of predictive accuracy.

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

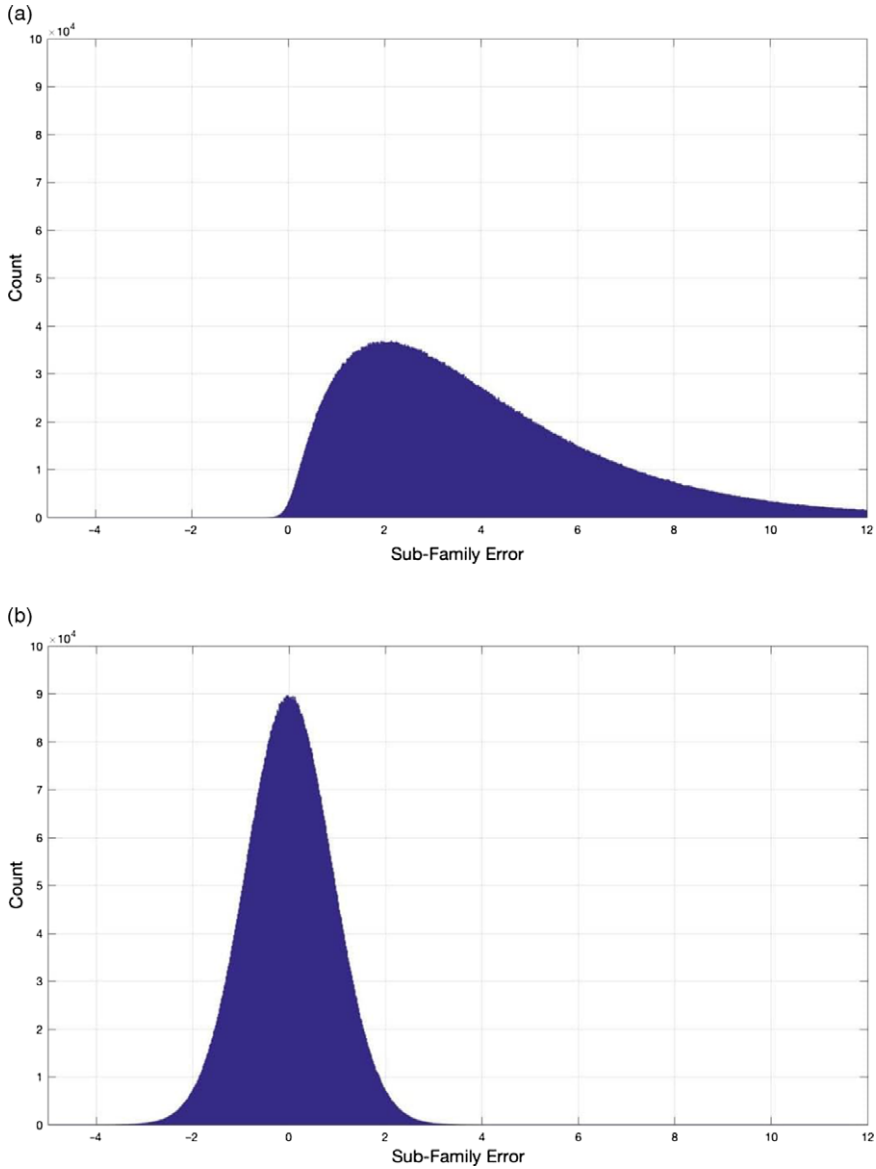

Figure 2. Histogram of the Sub-family Error of {L(M_4,**y**)}. Histogram of the Sub-family Error for M_0:{$-3x^3+x^2-1+N(0,1)$}.

3.9991 in 2a.[20] In 3b, the model is still singleton and close to the data (thus consideration$_1$ is still pertinent), but because M_0 is close to T, the average sub-family error was 0.4286, quite smaller than 3.9991. The importance of this fact can be best understood

[20] See equation (9) for why this average for large data sets is equal to 4.

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

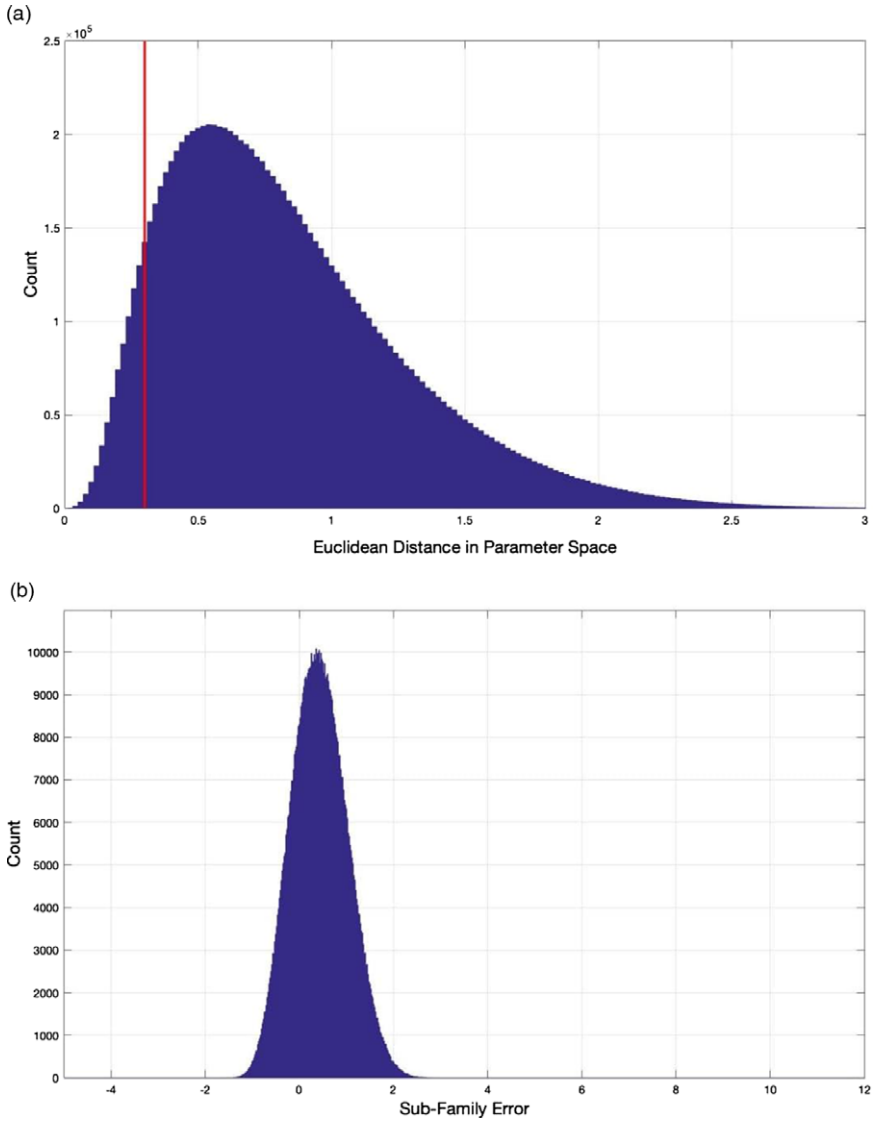

Figure 3. Histogram of the distance between M_0 and L(M_4,**y**). Histogram of the sub-family error for M_0 only for those data sets for which $||M_0\text{-L}(M_4,\mathbf{y})|| < 0.3$ (left of the red line in 3a).

with the help of figure 4, which depicts the histogram of AIC(M_4)-AIC(M_0) for those cases depicted in 3b. For $||M_0\text{-L}(M_4,\mathbf{y})|| < 0.3$, the average AIC difference was 3.2197.

The important point is that as we focus on data sets for which M_0 is closer and closer to L(M_4,**y**), AIC(M_0) becomes smaller and smaller (better) but at the same time the sub-family error becomes larger and larger. For $||M_0\text{-L}(M_4,\mathbf{y})|| \to 0$, (AIC($M_0$,**y**)–AIC($M_4$,**y**))→ -4 and the sub-family error approached 1 in this particular case. That is, *if you correct for the sub-family error*, AIC(M_0,**y**) is still 3 units better than that of

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

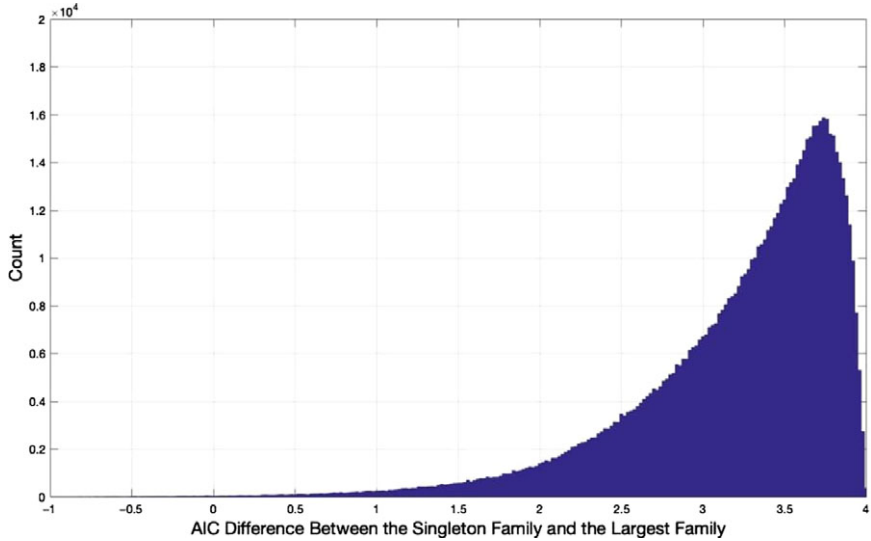

Figure 4. Histogram of the AIC difference between M_0 and M_4 for data sets for which $||M_0\text{-L}(M_4,\mathbf{y})|| < 0.3$ (left of the red line in 3a).

AIC(M_4,**y**). For the average case that satisfied $||M_0\text{-L}(M_4,\mathbf{y})|| < 0.3$, *if you correct for the sub-family error* (i.e., subtract the average sub-family error, 0.4286), you still have a better AIC value of 2.7911. This illustrates the main point of our argument. In the singleton family problem, our model is singleton and very close to the data. However, its excellent AIC score (although slightly biased) *is* excellent evidence that it is predictively accurate.

Objection: the difference between the sub-family problem and the singleton family problem is only a historical fact about how the model was constructed. So it cannot affect how good an estimate AIC($\{5x^3+6x^2+2x+7+N(0,1)\}$,$\mathbf{y_0}$) is.

Answer: certain facts about AIC($\{5x^3+6x^2+2x+7+N(0,1)\}$,$\mathbf{y_0}$) are not affected by this historical fact, including the definition of A($\{5x^3+6x^2+2x+7+N(0,1)\}$), the value of AIC($\{5x^3+6x^2+2x+7+N(0,1)\}$,$\mathbf{y_0}$) itself and the statistical bias of the estimator AIC($\{5x^3+6x^2+2x+7+N(0,1)\}$,**y**). However, this list does not exhaust all the relevant information bearing on how good an estimate AIC($\{5x^3+6x^2+2x+7+N(0,1)\}$,$\mathbf{y_0}$) is. Indeed, Forster and Sober introduced the notion of epistemic bias in order to be able to account for the intuitive idea that in the sub-family problem there is something wrong with AIC($\{5x^3+6x^2+2x+7+N(0,1)\}$,$\mathbf{y_0}$) as an estimate, which cannot be captured by the items in the above list. We think what is wrong with this estimate is that in the sub-family problem, our model is tailored to $\mathbf{y_0}$[21] in an ad hoc fashion in order to have an optimal AIC score. However, this is not the case in the singleton family problem. In that problem, $\{5x^3+6x^2+2x+7+N(0,1)\}$ is a contender model for independent reasons. And that such a simple model fits $\mathbf{y_0}$ so well (without us having

[21] The term "our model" can refer rigidly to $\{5x^3+6x^2+2x+7+N(0,1)\}$ and non-rigidly to whatever model we construct. That our model is tailored to data in the sub-family problem is true only if "our model" is understood non-rigidly, which is what we intend in this paragraph.

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

chosen it because it fits $\mathbf{y_0}$ well) is excellent reason that it is highly predictively accurate. In other words, *the historical fact that constitutes the difference between the two problems bears important information about the ad hocness of AIC($\{5x^3+6x^2+2x+7+N(0,1)\}$,$\mathbf{y_0}$) as an estimate.* Of course, this intuitive idea needs a technical explanation. That is what we will offer in the next section.

5. Our Solution

We think a better solution for the sub-family problem can be given by studying another estimator namely AIC($\{L(M_4,\mathbf{y})\},\mathbf{y}$) (instead of AIC($\{L(M_4,\mathbf{y_0})\},\mathbf{y}$)) one of the instantiations of which is the estimate AIC($\{L(M_4,\mathbf{y_0})\},\mathbf{y_0}$). But how can one decide which estimator provides better information about the merits/demerits of the individual estimate? In order to answer this question, consider why statisticians study the statistical biases of estimators in the first place. Suppose a_1 is an estimate of the quantity, a. For example, a_1 might be the value one reads on a kitchen scale when one weighs an apple. Naturally, one might wish to know how good an estimate a_1 is, but usually one cannot directly talk about how good or bad a single estimate is. However, sometimes one can talk about how good, in general, the estimates obtained by the "same" *procedure* tend to be. That is, one can talk about various features of an estimator by considering *repeated* estimates obtained by the "same" procedure. Statistical bias is one such feature. An estimator whose expected value equals the value it estimates is statistically unbiased. But what exactly does this tell us about the individual estimate? *Insofar as the individual estimate is similar to the other instantiations of the estimator*, the expected error of the estimator (its statistical bias) contains information about how good the estimate is. Importantly, the inherent vagueness in what it means for the procedure to be the "same" makes it the case that a particular estimate can be an instantiation of more than one estimator.[22] However, it doesn't follow that the statistical biases of those estimators bear equally valuable information on the merits of the individual estimate. If the estimate a_1 is an instantiation of two estimators A and A^*, and if other instantiations of A better resemble a_1 than other instantiations of A^* in ways that affect the value of the estimate, then the statistical bias of A bears more pertinent information than the statistical bias of A^* on the merits/demerits of a_1.

In ordinary inference problems, there is only one estimator one would naturally associate with an individual AIC score (qua estimate). Thus, these considerations are usually unimportant. However in the sub-family problem, things are different because not only the individual AIC score is a function of the data, but the model itself is designed on the basis of the data too. Thus, the AIC score one calculates - AIC($\{L(M_4,\mathbf{y_0})\},\mathbf{y_0}$) - is an instantiation of two estimators: AIC($\{L(M_4,\mathbf{y_0})\},\mathbf{y}$) and AIC($\{L(M_4,\mathbf{y})\},\mathbf{y}$). Notice that both of these estimators can be understood as the AIC of "the sub-family model." If you understand "the sub-family model" *rigidly* to refer

[22] There is a well-known criticism of frequentist statistics that the same individual estimate can be produced by many different estimation procedures. (Though see Hájek [2007] for why this is everyone's problem.) However, this criticism doesn't sanction an "anything goes" kind of attitude towards the estimate-estimator relation. We believe the notion of statistical bias is a helpful notion to the extent that one chooses estimators the other instantiations of which are not different from the estimate in question in intuitively salient ways. Otherwise, the average error of the estimator bears no or little information on how good the particular estimate is.

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

to the model constructed after obtaining $\mathbf{y_0}$, you'll have AIC({L(M_4,$\mathbf{y_0}$)},$\mathbf{y}$); and if you understand it *nonrigidly* to refer to whichever model one constructs on the basis of data, you'll have AIC({L(M_4,$\mathbf{y}$)},$\mathbf{y}$). The former is an unbiased estimator, as Forster and Sober correctly claim. The latter is a biased estimator, as we shall argue later. But which one gives us better information about the merits/demerits of the individual estimate AIC({L(M_4,$\mathbf{y_0}$)},$\mathbf{y_0}$)? In order to answer this question, consider another instantiation of each estimator, say for data set $\mathbf{y_1}$. The estimator AIC({L(M_4,$\mathbf{y_0}$)},$\mathbf{y}$) yields AIC({L(M_4,$\mathbf{y_0}$)},$\mathbf{y_1}$). This keeps the model {$5x^3+6x^2+2x+7+N(0,1)$} and calculates the AIC score of *that fixed* model with respect to $\mathbf{y_1}$. However AIC({L(M_4,$\mathbf{y}$)},$\mathbf{y}$) yields AIC({L(M_4,$\mathbf{y_1}$)},$\mathbf{y_1}$). Here one constructs the model {L(M_4,$\mathbf{y_1}$)} and calculates AIC({L(M_4,$\mathbf{y_1}$)}) with respect to $\mathbf{y_1}$. AIC({L(M_4,$\mathbf{y_0}$)},$\mathbf{y_1}$) is dissimilar to AIC({L(M_4,$\mathbf{y_0}$)},$\mathbf{y_0}$) in the sub-family problem in an important respect. In AIC({L(M_4,$\mathbf{y_0}$)},$\mathbf{y_1}$) the model is chosen because it fits one data set ($\mathbf{y_0}$) well but its AIC score is calculated with respect to another data set ($\mathbf{y_1}$). Whereas in AIC({L(M_4,$\mathbf{y_0}$)},$\mathbf{y_0}$), the model is chosen because it has excellent fit with $\mathbf{y_0}$ and its AIC score is calculated with respect to that *same* data set. Obviously, AIC({L(M_4,$\mathbf{y_0}$)},$\mathbf{y_0}$) and AIC({L(M_4,$\mathbf{y_1}$)},$\mathbf{y_1}$) are similar in this respect. Another way one can see this point is by considering the following question: What would have been the AIC estimate if instead of $\mathbf{y_0}$ one had obtained $\mathbf{y_1}$ in the sub-family problem? Clearly the answer is AIC({L(M_4,$\mathbf{y_1}$)},$\mathbf{y_1}$). This is the reason we believe the statistical bias of the estimator AIC({L(M_4,$\mathbf{y}$)},$\mathbf{y}$) provides better information about the estimate AIC({L(M_4,$\mathbf{y_0}$)},$\mathbf{y_0}$) in the sub-family problem.[23]

What about the singleton family problem? Things are quite different in that problem. What would have been the AIC estimate if instead of $\mathbf{y_0}$ one had obtained $\mathbf{y_1}$ in the singleton family problem? Clearly AIC({$5x^3+6x^2+2x+7+N(0,1)$},$\mathbf{y_1}$), which happens to be equal to AIC({L(M_4,$\mathbf{y_0}$)},$\mathbf{y_1}$), because here {$5x^3+6x^2+2x+7+N(0,1)$} is a contender model regardless of the fact that $5x^3+6x^2+2x+7+N(0,1) = L(M_4,\mathbf{y_0})$. Therefore, for this problem the estimator AIC({L(M_4,$\mathbf{y}$)},$\mathbf{y}$) is irrelevant, because we care about the singleton family {$5x^3+6x^2+2x+7+N(0,1)$}. Here the statistical bias of AIC({L(M_4,$\mathbf{y_0}$)},$\mathbf{y}$) gives one the relevant information on how good an estimate AIC({$5x^3+6x^2+2x+7+N(0,1)$},$\mathbf{y_0}$) (or incidentally, AIC({L(M_4,$\mathbf{y_0}$)},$\mathbf{y_0}$)) is.

Now we would like to show that AIC({L(M_4,$\mathbf{y}$)},$\mathbf{y}$) is in fact a biased estimator. This might appear to contradict Akaike's results, but it doesn't. Those results presuppose IMR according to which the set of candidate models must be determined independently of the current data. In fact, Akaike's results about models that respect IMR help us prove that AIC({L(M_4,$\mathbf{y}$)},$\mathbf{y}$) is biased.

In order to show this, we first show that the expectation of −½(AIC({L(M_4,$\mathbf{y}$)},$\mathbf{y}$) with respect to $\mathbf{y}$ is larger than the predictive accuracy of *any* singleton family whose only member is a member of M_4. Suppose f is a variable that ranges over the members of M. For each value of f, we can define the singleton family {f}. Also suppose m is an arbitrary member of M and {m} is the singleton family whose only member is m. By definition of L(M_4,$\mathbf{y}$) we have,

[23] It is worth mentioning that this is not a criticism of Forster and Sober's argument. They introduce the notion of epistemic bias in order to express the same idea (in a different way than we do) that the statistical unbiasedness of the estimator AIC({L(M_4,$\mathbf{y_0}$)},$\mathbf{y}$) does not exhaust all the relevant information on the demerits of the estimate AIC({L(M_4,$\mathbf{y_0}$)},$\mathbf{y_0}$) in the sub-family problem.

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

$$-\tfrac{1}{2}\mathrm{AIC}\big(\{\mathrm{L}(M_4,\mathbf{y})\},\mathbf{y}\big) = \max[-\tfrac{1}{2}(\mathrm{AIC}(\{\mathrm{f}\},\mathbf{y}),\mathrm{f}\in M_4] \geq -\tfrac{1}{2}\mathrm{AIC}(\{\mathrm{m}\},\mathbf{y}) \quad (5)$$

If we take expectation with respect to $\mathbf{y}$ we have:

$$\mathrm{E}_{\mathbf{y}}[-\tfrac{1}{2}\mathrm{AIC}((\{\mathrm{L}(M_4,\mathbf{y})\},\mathbf{y})] = \mathrm{E}_{\mathbf{y}}[\max[-\tfrac{1}{2}\mathrm{AIC}(\{\mathrm{f}\},\mathbf{y}),\mathrm{f}\in M_4]] \geq \mathrm{E}_{\mathbf{y}}\big[-\tfrac{1}{2}\mathrm{AIC}(\{\mathrm{m}\},\mathbf{y})\big] = \mathrm{A}(\{\mathrm{m}\}) \quad (6)$$

The last equality obtains because of (4). In (6), only if $\mathrm{m} = \mathrm{L}(M_4)$ the equality holds. However, $\mathrm{L}(M_4)$ is a function of data. For any m, such that $\mathrm{m} = \mathrm{L}(M_4,\mathbf{y_i})$, there is (almost always) a data set $\mathbf{y_j}$ generated by the same generating function such that m is not the likeliest member of M_4 with respect to $\mathbf{y_j}$. Thus, in fact we have a stronger result than (6):

$$\mathrm{E}_{\mathbf{y}}[-\tfrac{1}{2}\mathrm{AIC}\big(\{\mathrm{L}(M_4,\mathbf{y})\},\mathbf{y}\big)] = \mathrm{E}_{\mathbf{y}}[\max[-\tfrac{1}{2}\mathrm{AIC}(\{\mathrm{f}\},\mathbf{y}),\mathrm{f}\in M_4]] > \mathrm{E}_{\mathbf{y}}\big[-\tfrac{1}{2}\mathrm{AIC}(\{\mathrm{m}\},\mathbf{y})\big] = A(\{\mathrm{m}\}) \quad (7)$$

It follows from (7) that the average value of $-\tfrac{1}{2}\mathrm{AIC}(\{\mathrm{L}(M_4,\mathbf{y}),\mathbf{y}\})$ is strictly larger than the predictive accuracy of any singleton family one can construct from the members of M_4 (since m is an arbitrary member of M_4)—including, of course, $\{\mathrm{L}(M_4,\mathbf{y_0})\}$.

Since $\{\mathrm{L}(M_4,\mathbf{y})\}$ is a random variable and a function of data, $\mathrm{A}(\{\mathrm{L}(M_4,\mathbf{y})\})$ is a random variable too and will vary for different data sets. This is unlike the usual application of the Akaikean framework, where the model is fixed and its predictive accuracy is a fixed number. Indeed if $\{\mathrm{L}(M_4)\}$ was not a function of $\mathbf{y}$, then by (4), $-\tfrac{1}{2}\mathrm{AIC}(\{\mathrm{L}(M_4)\})$ would have been an unbiased estimator of $\mathrm{A}(\{\mathrm{L}(M_4)\})$. However, regardless of the data at hand, since $\mathrm{L}(M_4)$ is a member of M_4, we have $\mathrm{A}(\{\mathrm{L}(M_4,\mathbf{y})\}) \leq \max[(\mathrm{A}(\{\mathrm{f}\}), \mathrm{f}\in M_4]$. And since (7) is true for all m, $\mathrm{m}\in M_4$, then we have,

$$\mathrm{E}_{\mathbf{y}}\big[-\tfrac{1}{2}\mathrm{AIC}\big(\{\mathrm{L}(M_4,\mathbf{y})\},\mathbf{y}\big)\big] > \mathrm{A}\big(\{\mathrm{L}(M_4,\mathbf{y})\}\big) \quad (8)$$

That is, $\mathrm{AIC}(\{\mathrm{L}(M_4,\mathbf{y})\},\mathbf{y})$ is statistically biased. A comparison between equations (4) and (8) shows the difference between $\mathrm{AIC}(\{\mathrm{L}(M_4)\})$ and AIC of 'normal' models that are constructed independently of the data. In the same way that equation (4) motivates using AIC scores of 'normal' models as an estimate of their predictive accuracies, equation (8) shows why $-\tfrac{1}{2}$ times $\mathrm{AIC}(\{\mathrm{L}(M_4,\mathbf{y_0})\},\mathbf{y_0})$ is not a good estimate of $\mathrm{A}(\{\mathrm{L}(M_4,\mathbf{y_0})\})$.

There is another way of looking at $\mathrm{AIC}(\{\mathrm{L}(M_4,\mathbf{y})\},\mathbf{y})$ that makes it obvious why it is a biased estimator. Since $\{\mathrm{L}(M_4,\mathbf{y})\}$is singleton,

$$\mathrm{AIC}\big(\{\mathrm{L}(M_4,\mathbf{y})\},\mathbf{y}\big) = -2\log\mathscr{L}\big(\{\mathrm{L}(M_4,\mathbf{y})\},\mathbf{y}\big) = \mathrm{AIC}\big(M_4,\mathbf{y}\big) - 4, \quad (9)$$

The last equality obtains because M_4 has 4 adjustable parameters. But by the definition of predictive accuracy (equation (1)) for model M_4, the average predictive accuracy of $\{\mathrm{L}(M_4,\mathbf{y})\}$ (i.e., $\mathrm{E}_{\mathbf{y}}[\mathrm{A}(\{\mathrm{L}(M_4)\})]$) must equal $\mathrm{A}(M_4)$. It follows that $\mathrm{AIC}(\{\mathrm{L}(M_4,\mathbf{y})\},\mathbf{y})$ is on average 4 units lower than $\mathrm{A}(\{\mathrm{L}(M_4,\mathbf{y})\})$. The fact that the estimator $\mathrm{AIC}(\{\mathrm{L}(M_4,\mathbf{y})\},\mathbf{y})$ is biased gives us information about why the individual estimate $\mathrm{AIC}(\{\mathrm{L}(M_4,\mathbf{y_0})\},\mathbf{y_0})$ is not a good estimate of predictive accuracy. This concludes our technical treatment of the sub-family problem. In the next section, we will talk about another inference problem in which IMR is violated.

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

7. Too Many Models

Post hoc model construction is not the only way one can violate IMR. Another way is by adding models to one's set of candidate models without any reason for them to be considered. This simply increases the probability that a model with low predictive accuracy will have a good AIC score because it fits the current data well. An extreme version of this strategy is to add every member of the largest candidate model as a singleton model. In our working example of polynomial models, this would involve adding a singleton family $\{M_{a,b,c,d}: y = ax^3+bx^2+cx+d+N(0,1)\}$ for all $(a,b,c,d) \in R^4$. The number of candidate models will be infinite, but the one with the lowest AIC score is obviously $\{L(M_4,\mathbf{y})\}$. Therefore, in this case one would always choose the same model as one does in the sub-family problem. No doubt this is a problematic practice, but the question is why?

In order to see why, first consider the rather common practice of computing AIC scores for all candidate models, choosing the model with minimum AIC as the winner and taking its AIC value as an estimate of its predictive accuracy. Call this policy the minimizing policy. We can define the error involved in this policy as follows.

$$\text{Error}(\text{minimizing}) =^{\text{df}} -\tfrac{1}{2}\text{AIC}(M_{min}(\mathbf{y})) - \text{A}(M_{min}(\mathbf{y})), \tag{10}$$

where $M_{min}(\mathbf{y})$ is the model with the minimum AIC score given the data $\mathbf{y}$. The minimizing policy seems quite unproblematic, but here is an interesting fact: the expectation of Error(minimizing) is positive; that is, the AIC score of $M_{min}(\mathbf{y})$ is a *biased* estimator of its predictive accuracy. Generally, if $\hat{r}$ and $\hat{s}$ are unbiased estimators of r and s, $\max(\hat{r}, \hat{s})$ is usually a biased estimator (in fact over-estimator) of max(r,s). Suppose r is in fact bigger than s. Obviously, $\max(\hat{r}, \hat{s}) \geq \hat{r}$, and after taking expectation, $E(\max(\hat{r}, \hat{s})) \geq E(\hat{r}) = r = \max(r,s)$. However, if $P(\hat{r} < \hat{s}) > 0$, $\max(\hat{r}, \hat{s}) > \hat{r}$ with positive probability and therefore, $E(\max(\hat{r}, \hat{s})) > \max(r, s)$. But unless r and s are massively different or $\hat{r}$ and $\hat{s}$ are estimators with extremely small variances, $P(\hat{r} < \hat{s}) > 0$. In fact, in the case of AIC scores, even if model *A* is significantly predictively more accurate than model *B*, there is still a positive (however small) probability that one acquires a data set for which $\text{AIC}(B) < \text{AIC}(A)$.

The expected value of Error(minimizing) is normally small. However, its size grows with an increase in the number of models. Without going into the technical details, we will only gesture towards an explanation for why this happens.[24] Recall that the AIC score of a model, unlike its predictive accuracy, is a random variable; it varies with different data sets. Equation (4) states that the average value of AIC equals predictive accuracy, but it doesn't say anything else about the distribution of AIC. In the most straightforward cases, −½AIC has an approximately chi-squared distribution plus a constant. Thus, in an inference problem with N models, $M_1, M_2, \ldots, M_N$, we can think of −½AICs as N approximately chi-squared distributed random variables (plus a constant) with means equal to predictive accuracies, $\text{A}(M_1), \text{A}(M_2), \ldots, \text{A}(M_N)$. For simplicity, suppose that predictive accuracies are not very different. (Neither this assumption nor the chi-square distribution is necessary for the validity of our argument, but they make things easier for our explanatory purposes.) It is unlikely for

[24] Giraud (2015, 33–38) contains a rather technical treatment.

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

each such variable to be significantly smaller than the mean value, but if N is large, it is *very likely* that at least one of the variables (and thereby the minimum of all of them) is significantly smaller than the mean. Therefore, the more candidate models you have, the more biased the AIC value of M_{min} will be.

The fact that AIC is an unbiased estimator of predictive accuracy can mislead one into thinking that one can add to one's candidate models at will, in the hope that if any model is not plausible it will have a poor AIC score and will thus be discarded. This is a mistaken idea. Choosing the model with minimum AIC score can be a hopelessly misguided practice if one isn't stingy about which models to consider in the first place.

8. Concluding Remarks

We discussed Forster and Sober's solution to the sub-family problem. Although we agree with much of what they say, we disagree about a *potential* implication of their argument concerning the singleton family problem. We offered our own solution for the sub-family problem, which makes the difference between the two problems salient.

Although we find the sub-family problem interesting in itself, we believe a much more important lesson about simplicity-favoring considerations can be learned from our solution to the problem. Here we would like briefly to discuss what is going on beyond the technicalities. The fundamental difference between the sub-family problem and the singleton family problem is that the sub-family model is itself a random variable dependent on and tailored to the data. (Hence the difference between equations [4] and [8].) Simple models that are *designed* to have excellent goodness-of-fit with the extant data tend to perform poorly in predicting future data.[25] Here is a non-technical explanation for this. Consider the very idea behind the Akaikean framework. Why is it a bad idea to use the goodness-of-fit of the best fitting member of a model as an estimate of the model's predictive accuracy? Because to do so would essentially amount to using the current data twice: both in determining the best fitting member of the model and in determining how close the model is to the data, which is measured in terms of the fit of that *same* best fitting member. If the data was not used to pick out the representative (best fitting) member of the model (as is the case in singleton models), the fit of the model with the extant data was a good (unbiased) estimator of its predictive accuracy. Thus, the source of the problem with this proposal is essentially the double-use of the data.[26] But the more complex the model is, the more effective such double-use will be, because the data will have more power in selecting among the members of the model. That is why the bias in taking the goodness-of-fit of the model as an estimator of its predictive accuracy increases when the complexity of

[25] This also explains why other post hoc model construction policies (such as the one described in Kukla 1995) are likewise disallowed by the Akaikean framework.

[26] Note that this is a weaker claim than what is sometimes called the "no double-counting rule" in the debate over the thesis of predictionism (see Steele and Werndl 2018). According to that idea, no piece of evidence that was used in the construction of a theory can provide support for it. Whereas the idea discussed here simply asserts that if a piece of data is used in the construction of a theory, the support it provides for the theory is diminished (not necessarily nullified).

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

the model increases. The beauty of Akaike's results is in offering a way to calculate this bias. Now, when one *designs* one's model to be simple and to have an excellent degree of fit with the current data, one re-introduces that bias into one's estimation of predictive accuracy. In such an event, AIC is no longer an unbiased estimator, because the bias introduced by the double-use of the data is not relevant only to the number of adjustable parameters (which AIC corrects for) but also to the construction of the model itself (which AIC does not correct for).

We also discussed another problematic practice, which involves considering too many candidate models. We showed that one can effectively re-introduce the same error involved in the sub-family problem by engaging in an extreme version of this practice. Although post hoc model construction and comparing too many candidate models are problematic for two different reasons, there is a unified solution for both, namely, to respect IMR.

Before ending the paper, we would like to mention an important practical consequence of IMR. Respecting IMR gives rise to a clear criterion for determining whether a given model is gerrymandered or not. In the highly artificial examples in which the sub-family problem or similar problems are usually formulated (such as in Kukla 1995), it is crystal clear which models are fudged or gerrymandered (e.g., the sub-family model). However, in more realistic cases, it is sometimes not so clear. Thus, Douglas and Magnus write: "it would be perverse to do this arbitrarily, but in the general case of n-parameter models it may be possible to motivate specific values for some of the parameters. There is no formal rule for when this is or is not legitimate" (Douglas and Magnus 2013, 583). A comparison between the sub-family problem and the singleton family problem suggests exactly such a rule: models in which certain parameters are held fixed are not fudged just in case there are grounds independently of the current data for holding them so fixed.

Acknowledgments. We are grateful to Adam Elga, Elliott Sober, Erfan Salavati, David Schroeren, two anonymous referees, and an audience at Institute for Research in Fundamental Sciences in Tehran for their helpful feedback.

References

Akaike, Hirotugu. 1973. "Information Theory as an Extension of the Maximum Likelihood Principle." In *Second International Symposium on Information Theory*, edited by B. N. Petrov and F. Csaki, 267–81. Budapest: Akademiai Kiado.

Douglas, Heather, and P.D. Magnus. 2013. "State of the Field: Why Novel Prediction Matters." *Studies in the History and Philosophy of Science* 44 (4):580–89.

Forster, Malcolm R. 1995. "The Golfer's Dilemma: A Reply to Kukla on Curve-Fitting." *British Journal for the Philosophy of Science* 46 (3):348–60.

Forster, Malcolm R. 2006. "A Philosopher's Guide to Empirical Success." *Philosophy of Science* 74 (5): 588–600.

Forster, Malcolm R., and Elliott Sober. 1994. "How to Tell when Simpler, More United, or Less Ad Hoc Theories Will Provide More Accurate Predictions." *British Journal for the Philosophy of Science* 45 (1):1–35.

Forster Malcolm R., and Elliott Sober. 2011. "AIC Scores as Evidence – a Bayesian Interpretation." In *Philosophy of Statistics*, edited by Prasanta S. Bandyopadhyay and Malcolm R. Forster, 535–49. Amsterdam: Elsevier.

Giraud, Christophe. 2015. *Introduction to High-Dimensional Statistics*. Boca Raton, FL: CRC Press.

Hájek, Alan. 2007. "The Reference Class Problem Is Your Problem Too." *Synthese* 156 (3):563–85.

https://doi.org/10.1017/psa.2022.32 Published online by Cambridge University Press

Kukla, André. 1995. "Forster and Sober on the Curve-Fitting Problem." *British Journal for the Philosophy of Science* 46 (2):248–52.
Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1986. *Akaike Information Criterion Statistics*. Tokyo: KTK Scientific Publishers.
Steele, Katie, and Charlotte Werndl. 2018. "Model-Selection Theory: The Need for a More Nuanced Picture of Use-Novelty and Double-Counting." *British Journal for the Philosophy of Science* 69 (2):351–75.

Cite this article: Fatollahi, Alireza and Kasra Alishahi. 2023. "Simplicity and the Sub-Family Problem for Model Selection." *Philosophy of Science* 90 (3):648–668. https://doi.org/10.1017/psa.2022.32