Published in final edited form as:

Front Ecol Evol. 2019 November ; 7: . doi:10.3389/fevo.2019.00413.

Model Projections in Model Space: A Geometric Interpretation of the AIC Allows Estimating the Distance Between Truth and Approximating Models

José Miguel Ponciano[1,*], **Mark L. Taper**[1,2]

[1]Biology Department, University of Florida, Gainesville, FL, United States,

[2]Department of Ecology, Montana State University, Bozeman, MT, United States

Abstract

Information criteria have had a profound impact on modern ecological science. They allow researchers to estimate which probabilistic approximating models are closest to the generating process. Unfortunately, information criterion comparison does not tell how good the best model is. In this work, we show that this shortcoming can be resolved by extending the geometric interpretation of Hirotugu Akaike's original work. Standard information criterion analysis considers only the divergences of each model from the generating process. It is ignored that there are also estimable divergence relationships amongst all of the approximating models. We then show that using both sets of divergences and an estimator of the negative self entropy, a model space can be constructed that includes an estimated location for the generating process. Thus, not only can an analyst determine which model is closest to the generating process, she/he can also determine how close to the generating process the best approximating model is. Properties of the generating process estimated from these projections are more accurate than those estimated by model averaging. We illustrate in detail our findings and our methods with two ecological examples for which we use and test two different neg-selfentropy estimators. The applications of our proposed model projection in model space extend to all areas of science where model selection through information criteria is done.

[*]**Correspondence**: José Miguel Ponciano, josemi@ufl.edu.

AUTHOR CONTRIBUTIONS

JP and MT contributed equally to the conceptualization, development, programing, analysis, writing, and figure creation.

DATA AVAILABILITY STATEMENT

This paper's code is available at https://github.com/jmponciano/ModelsProjection.

Specialty section: This article was submitted to Environmental Informatics, a section of the journal Frontiers in Ecology and Evolution

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Keywords

error rates in model selection; Kullback-Leibler divergence; model projections; model averaging; Akaike’s Information Criterion

1. INTRODUCTION

Recent decades have witnessed a remarkable growth of statistical ecology as a discipline, and today, stochastic models of complex ecological processes are the hallmark of the most salient publications in ecology (e.g., Leibold et al., 2004; Gravel et al., 2016; Zeng and Rodrigo, 2018). Entropy and the Kullback-Liebler divergence as instruments of scientific inquiry are now at the forefront of the toolbox of quantitative ecologists, and many exciting new opportunities for their use are constantly being proposed (e.g., Casquilho and Rego, 2017; Fan et al., 2017; Kuricheva et al., 2017; Milne and Gupta, 2017; Roach et al., 2017; Cushman, 2018). One of the most important, but under explored, applications of the Kullback-Liebler divergence remains the study or characterization of the error rates incurred while making model selection according to information criteria (Taper and Ponciano, 2016b). This research is particularly relevant when, as it almost always happens in science, none of the candidate models exactly corresponds to the chance mechanism generating the data.

Understanding the impact of misspecification of statistical models constitutes a key knowledge gap in statistical ecology, and many other areas of biological research for that matter (e.g., Yang and Zhu, 2018). Research by us and many others (see citations in Taper and Ponciano, 2016b and in Dennis et al., 2019) has led to detailed characterizations of how the probability of making the wrong model choice using any given information criterion, not only may depend on the amount of information (i.e., sample size) available, but also on the degree of model misspecification.

Consequently, in order to estimate the error rates of model selection according to any information criterion, practitioners are left with the apparent paradox (“catch-22”) of being able to estimate how likely it is to erroneously deem as best that model which is furthest apart from the generating model, only after having accomplished the unsolved task of estimating the location of the candidate models relative to the generating process and to each other.

In this paper, we propose a solution to this problem. Our solution was motivated by the conceptualization of models as objects in a multi-dimensional space as well as an extension of the geometrical thinking that Akaike used so brilliantly in his 1973 paper introducing the AIC. Starting from Akaike’s geometry, we show how to construct a model space that includes not only the set of candidate models but also an estimated location for the generating process. Now, not only can an analyst determine which model is closest to the generating process, she/he can also determine the (hyper)spatial relationships of all models and how close to the generating process the best model is.

In 1973, Hirotugu Akaike wrote a truly seminal paper presenting what came to be known as the AIC. Akaike initially called the statistic “An Information Criterion,” but soon after its publication it came to be known as “Akaike’s Information Criterion.” Various technical accounts deriving the AIC exist (e.g., Burnham and Anderson, 2004, Chapter 7), but few explain in detail every single step of the mathematics of Akaike’s derivation (but see De Leeuw, 1992). Although focusing on the measure-theoretic details, deLeeuw’s account makes it clear that Akaike’s paper was a paper about ideas, more than a paper about a particular technique. Years of research on this project has led us to understand that only after articulating Akaike’s ideas, the direction of a natural extension of his work is easily revealed and understood. Although thinking of models and the generating mechanism as objects with a specific location in space is mathematically challenging, this exercise may also prove to be of use to study the adequacy of another common statistical practice in multi-model inference: model averaging.

Intuitively, if one thinks of the candidate models as a cloud of points in a Euclidean space, then it would only make sense to “average” the model predictions if the best approximation of the generating chance mechanism in that space is located somewhere inside the cloud of models. If however the generating model is located outside such cloud, then performing model average will only at best, worsen the predictions of the closest models to the generating mechanism. The question then is, can this idea of thinking about models as points in a given space be mathematically formalized? Can the structure and location of the candidate models and the generating mechanism be somehow estimated and placed in a space? If so, then the answer to both questions above (i.e., the error rates of multi-model selection under misspecification and when should an analyst perform model averaging) could be readily explored. These questions are the main motivation behind the work presented here.

2. THE AIC AND A NATURAL GEOMETRIC EXTENSION: MODEL PROJECTIONS IN MODEL SPACE

In his introduction to Akaike (1973)’s original paper, De Leeuw (1992) insisted on making sure it was understood that Akaike’s contribution was much more valuable for its ideas than for its technical mathematical developments: “…This is an ‘ideas’ paper,’ promoting a new approach to statistics, not a mathematics paper concerned with the detailed properties of a particular technique…” After this explanation, De Leeuw undertakes the difficult labor of teasing Akaike’s thought process from the measure-theoretic techniques. In so doing, the author manages to present a clear and concise account clarifying both, Akaike’s mathematical approach and his ideas. De Leeuw was keenly aware of the difficulty of trying to separate the ideas from the mathematical aspects of the paper: in introducing the key section in Akaike’s paper, he describes it as “a section not particularly easy to read, that does not have the usual proof/theorem format, expansions are given without precise regularity conditions, exact and asymptotic identities are freely mixed, stochastic and deterministic expressions are not clearly distinguished and there are some unfortunate notational… typesetting choices” (De Leeuw, 1992). To us, however, the importance of De Leeuw’s account stems from the fact that it truly brings home the crucial point that at the very heart of Akaike's derivation there was a geometrical use of Pythagoras' theorem (see Equation 1, page 604 in De Leeuw, 1992). The modern literature has been able to reduce Akaike's derivation to just a few lines (see Davison, 2003). However, such condensed proofs conceal the original geometric underpinnings of Akaike's thinking, which De Leeuw exposed. Our contribution for this special issue consists of taking Akaike's derivation one step further by using Pythagoras' theorem again to attain not a relative, but an absolute measure of how close each model in a model set is from the generating process.

Akaike's (1973) paper is difficult and technical but at the same time, it is a delightful reading because he managed to present his information criterion as the natural consequence of a logical narrative. That logical narrative consisted of six key insights that we strung together to arrive at what we believe is a second natural consequence of Akaike's foundational thoughts: our model projections proposal. After introducing our notation following Akaike's, we summarize those six key insights. We stress that these insights and the accompanying key figure we present below are none other than a simple geometric representation of De Leeuw's measure-theoretic re-writing of Akaike's proof. We encourage readers with a strong probability background to read De Leeuw's account. We then present our main model projections proposal and contribution and support it with a fully illustrated example.

2.1. Theoretical Insights From Akaike (1973)

Akaike's quest was motivated by a central goal of modern scientific practice: obtaining a comparison measure between many approximating models and the data-generating process. Akaike began thinking about how to characterize the discrepancy between any given approximating model and the generating process. He denoted the probability densities of the generating process and of the approximating model as $f(x, \theta_0)$ and $f(x, \theta)$, respectively, where θ_0 denoted the column vector of dimension L of true parameter values. Although he started by characterizing the discrepancy between the true model and the approximating model, his objective was to come up with an estimate of such discrepancy that somehow was free of the need of knowing either the dimension or the model form of $f(x, \theta_0)$. The fact that he was able to come up with an answer to such problem is not only outstanding, but the reason why the usage of the AIC has become ubiquitous in science. Akaike's series of arguments arriving to the AIC can be summarized by stringing together these six key insights:

2.1.1. Insight 1: Discrepancy From the Generating Process (Truth) Can Be Measured by the Average of Some Function of the Likelihood Ratio—

Akaike's first important insight follows from two observations. First, under the parametric setting defined above, a direct comparison between an approximating model and the true, generating stochastic process can be achieved *via* the likelihood ratio, or some function of the likelihood ratio. Second, because the data X are random, the expected discrepancy (average over all possible realizations of the data) would be written as

$$\begin{aligned}\mathscr{D}(\theta,\theta_0;\boldsymbol{\Phi}) &= \int f(x;\theta_0)\boldsymbol{\Phi}(\tau(x,\theta,\theta_0))dx \\ &= \mathbb{E}_X[\boldsymbol{\Phi}(\tau(X,\theta,\theta_0))],\end{aligned}$$

where the expectation is, of course, taken with respect to the generating stochastic process X. We denote the likelihood ratio as $\tau(x,\theta,\theta_0) = \frac{f(x;\theta)}{f(x;\theta_0)}$ and a twice differentiable function of it as $\Phi(\tau(x, \theta, \theta_0))$.

Akaike then proposed to study under a general framework how sensitive this average discrepancy would be to the deviation of θ from the truth, θ_0.

2.1.2. Insight 2: $\mathscr{D}(\theta,\theta_0;\boldsymbol{\Phi})$ Is Scaled by Fisher's Information Matrix—Akaike thought of expanding the average discrepancy $\mathscr{D}(\theta,\theta_0;\boldsymbol{\Phi})$ using a second order series approximation around θ_0. Akaike's second insight then consisted of noting the strong link between such approximation and the theory of Maximum Likelihood (ML).

For a univariate θ, the Taylor series approximation of the average function Φ of the likelihood ratio is written as

$$\begin{aligned}&\mathscr{D}(\theta,\theta_0;\boldsymbol{\Phi}) \approx \mathscr{D}(\theta_0,\theta_0;\boldsymbol{\Phi}) + (\theta-\theta_0)\frac{\partial\mathscr{D}(\theta,\theta_0;\boldsymbol{\Phi})}{\partial\theta}\bigg|_{\theta=\theta_0} \\ &+\frac{(\theta-\theta_0)^2}{2!}\frac{\partial^2\mathscr{D}(\theta,\theta_0;\boldsymbol{\Phi})}{\partial\theta^2}\bigg|_{\theta=\theta_0} + \ldots\end{aligned} \quad (1)$$

To find an interpretable form of this approximation, just like Akaike did following Kullback and Leibler (Kullback and Leibler, 1951; Akaike, 1973), we use two facts: first, by definition $\tau(x, \theta, \theta_0)|_{\theta=\theta_0} = 1$ and second, that $\int f(x; \theta)dx = 1$ because f is a probability density function. Together with the well-known regularity conditions used in mathematical statistics that allow differentiation under the integral sign (Pawitan, 2001), these two facts give us the following: first, $\int\frac{\partial f(x;\theta)}{\partial\theta}dx = \int\frac{\partial^2 f(x;\theta)}{\partial\theta^2}dx = 0$. Hence, $\frac{\partial\mathscr{D}(\theta,\theta_0;\boldsymbol{\Phi})}{\partial\theta}\bigg|_{\theta=\theta_0} = 0$. This result then allows writing the second derivative of the approximation as

$$\begin{aligned}\left.\frac{\partial^2 \mathscr{D}(\theta,\theta_0;\boldsymbol{\Phi})}{\partial\theta^2}\right|_{\theta=\theta_0} &= \left.\int \frac{\partial}{\partial\theta}\left(\frac{\partial\boldsymbol{\Phi}(\tau)}{\partial\tau}\frac{\partial\tau}{\partial\theta}\right) f(x;\theta_0)dx\right|_{\theta=\theta_0} \\ &= \left.\int \frac{\partial^2\boldsymbol{\Phi}(\tau)}{\partial\tau^2}\left(\frac{\partial\tau}{\partial\theta}\right)^2 f(x;\theta_0)dx\right|_{\theta=\theta_0} \\ &+ \left.\int \frac{\partial^2\tau}{\partial\theta^2}\frac{\partial\boldsymbol{\Phi}(\tau)}{\partial\tau} f(x;\theta_0)dx\right|_{\theta=\theta_0} \\ &= \boldsymbol{\Phi}''(1)\int\left(\frac{1}{f(x;\theta_0)}\frac{\partial f(x;\theta)}{\partial\theta}\right)^2 f(x;\theta_0)dx \mid S_{\theta} = \theta_0 \\ &= \boldsymbol{\Phi}''(1)\int\left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)^2 f(x;\theta)dx\bigg|_{\theta=\theta_0} \\ &= \boldsymbol{\Phi}''(1)\mathscr{I}(\theta_0),\end{aligned}$$

where $\mathscr{I}(\theta_0)$ is Fisher's information. To move from the first line of the above calculation to the second line we used a combination of the product rule and of the chain rule. To go from the second to the third line, note that because the first derivative is equal to 0 as shown immediately above of this equation, the integral in the right hand is null.

Hence, in this univariate case, the second order approximation is given by $\mathscr{D}(\theta,\theta_0) \approx \boldsymbol{\Phi}(1) + \frac{1}{2}\boldsymbol{\Phi}''(1)(\theta-\theta_0)^2\mathscr{I}(\theta_0)$, where $\mathscr{I}(\theta_0)$ is Fisher's information. Thus, the average discrepancy between an approximating and a generating model is scaled by the inverse of the theoretical variance of the Maximum Likelihood estimator, regardless of the form of the function Φ().

2.1.3. Insight 3: Setting Φ(*t*) = −2 log *t* Connects $\mathscr{D}(\theta,\theta_0;\Phi)$ With Entropy and Information Theory—

Akaike proceeded to arbitrarily set the function Φ(*t*) to Φ(*t*) = −2 log *t*. Using this function not only furthered the connection with ML theory, but also introduced the connection of his thinking with Information Theory. By using this arbitrary function, the average discrepancy becomes a divergence because $\mathscr{D}(\theta_0,\theta_0) = \boldsymbol{\Phi}(1) = 0$ and the approximation of the average discrepancy, heretofore denoted as $\mathscr{W}(\theta,\theta_0)$, is modulated by Fisher's information, the variance of the Maximum Likelihood estimator: $\mathscr{D}(\theta,\theta_0) \approx \mathscr{W}(\theta,\theta_0) = (\theta-\theta_0)^2\mathscr{I}(\theta_0)$. For a multivariate θ_0 we get then that $\mathscr{W}(\theta,\theta_0) = (\theta-\theta_0)'\mathscr{I}(\theta_0)(\theta-\theta_0)$ where $\mathscr{I}(\theta_0)$ is Fisher's Information matrix (Pawitan, 2001). Conveniently then, the arbitrary factor of 2 gave his general average discrepancy function the familiar "neg-entropy" or Kullback-Leibler (KL) divergence form

$$\begin{aligned}\mathscr{D}(\theta,\theta_0) &= -2\int f(x;\theta_0)\log\left(\frac{f(x;\theta)}{f(x;\theta_0)}\right)dx \\ &= -2\mathbb{E}_X\left[\log\frac{f(X;\theta)}{f(X;\theta_0)}\right] \\ &= -2[\mathbb{E}_X(\log f(X;\theta)) - \mathbb{E}_X(\log f(X;\theta_0))] \\ &= 2\mathbb{E}_X(\log f(X;\theta_0)) - 2\mathbb{E}_X(\log f(X;\theta)) \\ &= 2KL(\theta,\theta_0)\end{aligned} \tag{2}$$

thus bringing together concepts in ML estimation with a wealth of results in Information Theory. The two expectations (integrals) in the last line of the above equation were often succinctly denoted by Akaike as *Sgg* and *Sgf*, respectively: these are the neg-selfentropy and the neg-crossentropy terms. Thus, he would write that last line as $2KL(\theta, \theta_0) = 2[Sgg - Sgf]$. Note that for consistency with Akaike (1973) we have retained his notation and in particular, the order of arguments in the KL function, as opposed to the notation we use in Dennis et al. (2019).

2.1.4. Insight 4: $\mathscr{D}(\theta, \theta_0)$ Is Minimized at the ML Estimate of θ—Aikaike's fourth critical insight was to note that a Law of Large Numbers (LLN) approximation of the Kullback-Leibler divergence between the true, generating stochastic process and a statistical model is minimized by evaluating the candidate model at its maximum likelihood estimates. Such conclusion can be arrived at even if the generating stochastic model is not known. Indeed, given a sample of size n, $X_1, X_2, \ldots, X_n$ from the generating model, from the LLN we have that

$$\widehat{\mathscr{D}}_n(\hat{\theta}, \theta_0) = -2 \times \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i; \hat{\theta})}{f(x_i; \theta_0)},$$

which is minimized at the ML estimate $\hat{\theta}$. Akaike actually thought that this observation could be used as a *justification for the maximum likelihood principle*: "Though it has been said that the maximum likelihood principle is not based on any clearly defined optimum consideration, our present observation has made it clear that it is essentially designed to keep minimum the estimated loss function which is very naturally defined as the mean information for discrimination between the estimated and the true distributions" Akaike (1973).

2.1.5. Insight 5: Minimizing $\mathscr{D}(\theta, \theta_0)$ Is an Average Approximation Problem—Akaike's fifth insight was to recognize the need to account for the randomness in the ML estimator. Because multiple realizations of a sample $X_1, X_2, \ldots, X_n$ each results in different estimates of θ, the average discrepancy should be considered a random variable. The randomness hence, is with respect to distribution of the maximum likelihood estimator $\hat{\theta}$. Let $\mathscr{R}(\theta_0) = \mathbb{E}_{\hat{\theta}}\left[\mathscr{D}\left(\hat{\theta}, \theta_0\right)\right]$ denote our target average over the distribution of $\hat{\theta}$. Then, the problem of minimizing the Kullback Leibler divergence can be conceived as an approximation problem where the target is the average:

$$\begin{aligned}
\mathscr{R}(\theta_0) = \mathbb{E}_{\hat{\theta}} \mathscr{D}\left(\hat{\theta}, \theta_0\right) &= 2\mathbb{E}_{\hat{\theta}}\left[\mathbb{E}_X(\log f(X; \theta_0)) \right. \\
&\left. -\mathbb{E}_X(\log f(X; \hat{\theta}) \mid \hat{\theta})\right] \\
&= 2\mathbb{E}_X(\log f(X; \theta_0)) \\
&- 2\mathbb{E}_{\hat{\theta}}\left[\mathbb{E}_X(\log f(X; \hat{\theta}) \mid \hat{\theta})\right].
\end{aligned}$$

In the final expression of the equation above, the first term is an unknown constant. The second term on the other hand, is the expected value of a conditional expectation.

2.1.6. Insight 6: $\mathscr{D}(\theta, \theta_0)$ Can Be Approximated Geometrically Using Pythagoras' Theorem—Instead of estimating the expectations above, Akaike thought of substituting the probabilistic entropy $\mathscr{D}\left(\hat{\theta}, \theta_0\right)$ with its Taylor Series approximation $\mathscr{W}\left(\hat{\theta}, \theta_0\right) = \left(\hat{\theta} - \theta_0\right)' \mathscr{I}(\theta_0)\left(\hat{\theta} - \theta_0\right)$, which can then be interpreted as a squared statistical distance. This approximation is indeed the square of a statistical distance wherein the divergence between any two points $\hat{\theta}$ and θ_0 is weighted by their dispersion in multivariate space, measured by the eigenvalues of the positive definite matrix $\mathscr{I}(\theta_0)$. This sixth insight led him straight into the path to learning about the KL divergence between a generating process and a set of proposed probabilistic mechanisms/models. By viewing this quadratic form as a statistical distance, Akaike was able to use a battery of clear measure-theoretic arguments relying on various convergence proofs to derive the AIC.

Interestingly, and although he doesn't explicitly mentions it in his paper, his entire argument can be phrased geometrically: if the average discrepancy that he was after could be approximated with the square of a statistical distance, its decomposition using Pythagoras theorem was the natural thing to do. By doing such decomposition, one can immediately visualize the ideas in his proof with a simple sketch. We present such sketch in Figure 1. In that figure, the key triangle with a right angle has as vertices the truth θ_0 of unknown dimension L, the ML estimator $\hat{\theta}$ of dimension $k \leq L$, denoted $\hat{\theta}_k$ and finally, θ_{0k}. This quantity represents the orthogonal projection of the truth in the plane where all estimators of dimension k lie, which is in turn denoted as Θ_k (Figure 1A). Figure 1B shows a fourth crucial point in this geometrical interpretation: it is the estimator of θ_0 from the data using a model with the same model form as the generating model, but with parameters estimated from the data. To distinguish it from $\hat{\theta}_k$ we denote this estimator $\hat{\theta}_0$. Because it has the same dimensions than the generating model, $\hat{\theta}_0$ can be thought of as being located in the same model surface as the generating model θ_0. Akaike's LLN approximation of the KL divergence as an average of log-likelihood ratios $\widehat{\mathscr{D}}_n\left(\hat{\theta}, \theta_0\right) = -2 \times \frac{1}{n}\sum_{i=1}^{n} \log \frac{f\left(x_i; \hat{\theta}\right)}{f(x_i; \theta_0)}$ comes to play in this geometric derivation as the edge labeled e^2 in Figure 1B that traces the link between $\hat{\theta}_0$ and the ML estimator $\hat{\theta}_k$. Following Akaike's derivation then, the ML estimator $\hat{\theta}_k$ can be thought as the orthogonal projection of $\hat{\theta}_0$ onto the plane Θ_k.

Before continuing with our geometric interpretation, we alert the reader that in Figure 1 all the edges are labeled with a lowercase letter with the purpose of facilitating this geometric visualization. The necessary calculations to understand Akaike's results are presented as simplified algebraic calculations but the reader however, is warned that these edges or lower case letters denote for the most part random variables. We leave these simple letters here because in Akaike's original derivations, the technical measure-theoretic operations may end up distracting the reader from a natural geometric understanding of the AIC.

In simple terms then, the objective of this geometric representation is to see that obtaining an estimate of the discrepancy between the approximating model and the generating process amounts to solving for the square of the edge length b, which is in fact the KL divergence quadratic form approximation. That is, $b^2 = \mathscr{W}\left(\hat{\theta}, \theta_0\right)$. Proceeding with our geometric interpretation, note that the angle ϕ between edges h and c in Figure 1B is not by necessity a right angle, and that the generalized Pythagoras Theorem to find the edge length d applies. Akaike then noted that provided that the approximating model is in the vicinity of the generating mechanism, the third term of the generalized Pythagoras form of the squared distance $d^2 = c^2 + h^2 - 2ch\cos\phi$ was negligible when compared with c^2 and h^2 [see Akaike, 1973, his Equation (4.15) and his comment about that term in the paragraph above his Equation (4.19). See also De Leeuw 1992, text under his Equation (4)], and so he proceeded to simply use only the first two terms, c^2 and h^2 (see Figure 1C). The immense success of the AIC in a wide array of scientific settings to date shows that this approximation, as rough as it may seem, is in fact quite reliable. This approximation allowed him to write the squared distance d^2 in two different ways: as $d^2 \approx c^2 + h^2$ and as $d^2 = a^2 + e^2$. Because by construction, we have that $b^2 = h^2 + a^2$, one can immediately write the difference $b^2 - e^2$ as

$$\begin{aligned} b^2 - e^2 &= h^2 + a^2 - d^2 + a^2 \\ &= h^2 + a^2 - c^2 - h^2 + a^2, \end{aligned}$$

and then solve for b^2 (see Figure 1D):

$$b^2 = e^2 + 2a^2 - c^2\,. \tag{3}$$

Using asymptotic expansions of these squared terms, the observed Fisher's information and using known convergence in probability results, Akaike showed when multiplied by the sample size n, the difference of squares $c^2 - a^2$ was approximately chi-squared distributed with degrees of freedom $L - k$ and that $na^2 \sim \chi^2_k$. Then, multiplying equation (3) by n gives

$$nb^2 = n\mathscr{W}\left(\hat{\theta}_k, \theta_0\right) \approx \underbrace{n\mathscr{D}_n\left(\hat{\theta}_k, \hat{\theta}_0\right)}_{= 2 \times \text{log-likelihood ratio}} + \underbrace{na^2}_{\sim\chi^2_k} - \underbrace{n\left(c^2 - a^2\right)}_{\sim\chi^2_{L-k}}.$$

Finally, one may arrive at the original expected value of the conditional expectation shown above by replacing the chi-squares with their expected values, which are given by their degrees of freedom. Hence,

$$\begin{aligned} n\mathbb{E}_{\hat{\theta}_k}\left[\mathscr{W}\left(\hat{\theta}_k, \theta_0\right)\right] &\approx n\mathscr{D}_n\left(\hat{\theta}_k, \hat{\theta}_0\right) + 2k - L, \text{ or} \\ \mathbb{E}_{\hat{\theta}_k}\left[\mathscr{W}\left(\hat{\theta}_k, \theta_0\right)\right] &\approx \frac{-2}{n}\sum_{i=1}^{n} \log f\left(x_i; \hat{\theta}_k\right) + \frac{2k}{n} - \frac{L}{n} \\ &\quad + \frac{2}{n}\sum_{i=1}^{n} \log f\left(x_i; \hat{\theta}_0\right). \end{aligned} \tag{4}$$

The first two terms in the above expression, $-2\sum_{i=1}^{n} \log f\left(x_i; \hat{\theta}_k\right) + 2k$, constitute what came to be known as the AIC. These terms correspond respectively to twice the negative log-likelihood evaluated at the MLE and twice the number of parameters estimated in the approximating model. To achieve multi-model comparison (see Figure 2), Akaike swiftly pointed out that in fact, only these first two terms are needed because the true model dimension L and the term $\sum_{i=1}^{n} \log f\left(x_i; \hat{\theta}_0\right)$ both terms (1) remain the same across models, as long as the same data set is used and (2) *cannot be known* because they refer to the true model dimension. Akaike rightly noted that if one were to compute Equation (4) for a suite of approximating models, these two terms would remain the same across all models and hence, could in practice be ignored for comparison purposes: these unknowns then act as constants of proportionality that are invariant to model choice. Therefore, in order to compare the value of this estimated average discrepancy across a suite of models, the user only needs to calculate the AIC score $-2\sum_{i=1}^{n} \log f\left(x_i; \hat{\theta}_k\right) + 2k$ for each model and deem as best that model for which the outcome of this calculation is the smallest. The logic embedded in Akaike's reasoning is represented graphically in Figure 2 (redrawn from Burnham et al., 2011). This reasoning kickstarted the practice, still followed in science 46 years later, to disavow the absolute truth in favor of a careful examination of multiple, if not many, models.

Finally, the reader should recall that what Equation (4) is in fact approximating is

$$\begin{aligned}\mathscr{R}(\theta_0) = \mathbb{E}_{\hat{\theta}}\mathscr{D}\left(\hat{\theta}_k, \theta_0\right) = & -2\mathbb{E}_{\hat{\theta}}\left[\mathbb{E}_X\left(\log f\left(X; \hat{\theta}_k\right) \mid \hat{\theta}_k\right)\right] \\ & + 2\mathbb{E}_X(\log f(X; \theta_0)) .\end{aligned} \tag{5}$$

and that this last expression is in fact the expectation with respect to $\hat{\theta}_k$ of

$$\begin{aligned}-2\int f(x; \theta_0)\log\frac{f\left(x; \hat{\theta}_k\right)}{f(x; \theta_0)}dx = & -2\int f(x; \theta_0)\log f\left(x; \hat{\theta}_k\right)dx \\ & + 2\int f(x; \theta_0)\log f(x; \theta_0)dx .\end{aligned} \tag{6}$$

Later, Akaike (1974) referred to the integral $\int f(x; \theta_0)\log f\left(x; \hat{\theta}_k\right)dx$ as *Sgf* and to $\int f(x; \theta_0)$ log $f(x; \theta_0)dx$ as *Sgg*, which are names easy to remember because it's almost as if the *S* in *Sgf* and *Sgg* represent the integral sign and *g* and *f* are a short hand representation of the probability density function of the generating stochastic process and of the approximating model, respectively.

One of our central motivations to write this paper is the following: by essentially ignoring the remainder terms in Equation (4), since 1973 practitioners have been almost invariably selecting the "least worst" model among a set of models (but see Spanos, 2010). In other words, we as a scientific community, have largely disregarded the question of how far, *in absolute terms not relative*, is the generating process from the best approximating model. Suppose the generating model is in fact very far from all the models in a set of models currently being examined. Then, the last term in Equation (4) will be very large with respect to the first two terms for all the models in a model set that is being examined, and essentially any differences between the terms $-2\sum_{i=1}^{n} \log f\left(x_i; \hat{\theta}_k\right) + 2k$ for every model will be meaningless.

2.2. The Problem of Multiple Models

Akaike's realization that "truth" did not need to be known in order to select from a suite of models which one was closest to truth shaped the following four and a half decades of scientific undertaking of model-centered science. Scientists were then naturally pushed toward the confrontation of not one or two, but multiple models with their experimental and observational data. Such approach soon led to the realization that basing the totality of the inferences on the single best model was not adequate because it was often the case that a small set of models would appear indistinguishable from each other when compared (Taper and Ponciano, 2016b).

Model averaging is by far, the most common approach used today to make inferences and predictions following an evaluation of multiple models *via* the AIC. Multiple options to do model averaging exist but in all cases, this procedure is an implicit Bayesian methodology that results in a set of posterior probabilities for each model. These posterior probabilities are called the "Akaike weights." For the i^{th} model in a set of candidate models, this weight is computed as

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

In this expression, Δ_i is the i^{th} difference between the AIC value and the best (i.e., the lowest) AIC score in the set of R candidate models. Although this definition is very well-known, cited and used (Taper and Ponciano, 2016b), it is seldom acknowledged that because these weights are in fact posterior probabilities, they must result from adopting a specific set of subjective model priors. Burnham et al. (2011) actually show that the weights shown above result from adopting the following subjective priors q_i:

$$q_i = C \cdot \exp\left(\frac{1}{2}k_i \log(n) - k_i\right), \tag{7}$$

where C is a normalization constant, k_i is the model dimension (the estimated number of parameters) of model i and n denotes the total sample size. Note that with sample sizes above 7, those weights increase with the number of parameters, thus favoring parameter rich models. The use of these priors makes model averaging a confirmation approach (Bandyopadhyay et al., 2016).

For someone using evidential statistics, adopting the model averaging practice outline above presents two important problems: first, the weights are based on prior beliefs that favor more parameter rich models and are not based on actual evidence (data). Second, and much more practically, model averaging appears to artificially favor redundancy of model specification: the more models that are developed in any given region of model space, the stronger this particular region gets weighted during the model averaging process. To counter these two problems, here we propose alternatively to estimate (1) the properties of a hyperplane containing the model set, (2) the location in such plane of the best projection of the generating process and (3) an overall general discrepancy between each of the models in the model set and the generating process or truth. We achieve these goals by using the estimated

KL divergences amongst all estimated models, that is, the estimated Sf_if_j for all models i and j in the candidate set. This is information that is typically ignored. Here again, we use Akaike's mnemonic notation where g denotes the generating model and f the approximating model. Then the so called neg-crossentropy and neg-selfentropy are written as

$$\begin{aligned} Sgf &= \int f(x;\theta_0)\log f\left(x;\hat{\theta}_k\right)dx \text{ and} \\ Sgg &= \int f(x;\theta_0)\log f(x;\theta_0)dx, \text{ respectively.} \end{aligned}$$

In his 1974 paper, Akaike observed that the neg-crossentropy could be estimated with

$$\widehat{Sgf} = \frac{1}{n}\sum_{i=1}^{n}\log f\left(x_i;\hat{\theta}_k\right) - \frac{k}{n} = -\frac{AIC}{2n}. \tag{8}$$

We wish to point out that in the "popular" statistical literature within the Wildlife Ecology sciences (e.g., Burnham and Anderson, 2004; Burnham et al., 2011), it is often repeated that an estimator of $\mathbb{E}_{\hat{\theta}}\left[\mathbb{E}_X\left(\log f\left(X;\hat{\theta}_k\right) \mid \hat{\theta}_k\right)\right]$ is given by $-AIC/2$. In fact, Akaike (1974) shows that the correct estimator is given by Equation (8). This distinction, albeit subtle, marks a difference when the analyst wishes to compare not only which model best approximates the generating process, but also the strength of the evidence for one or the other model choice.

In what follows, we extend Akaike's geometric derivation to make inferences regarding the spatial configuration of the ensemble of models being considered as approximations to the generating process. As we show with an ecological example, unlike model averaging this natural geometric extension of the AIC is fairly robust to the specification of models around the same region of model space and is actually aided, not hampered, by proposing a large set of candidate models.

2.3. A Geometrical Extension of Akaike's Extension to the Principle of Maximum Likelihood

As modelers, scientists are naturally drawn to visualize a suite of candidate models as entities in a (hyper)plane. By so doing, the geometric proximities between these entities are then intuitively understood as similarities amongst models. The key questions we answer in this paper are whether it is possible to estimate the architecture of such model space, locate a suite of approximating models within such space as well as estimating the location of the projection of truth onto that plane. All of this while not having to formulate an explicit model for the generating model. The estimation of the location of the truth projection in that plane would open the door to a formulation of an overall goodness of fit measure qualifying every single one of the AIC scores computed for a set of candidate models. Additionally, answering these questions automatically provides valuable insights to intuitively understand why or why not model averaging may be an appropriate course of action. As we show below, these questions are answerable precisely because any given set of models has a set of relationships which are typically ignored but that can be translated directly to a set of geometrical relationships that carry all the needed information and evidence.

One of the key observations of this contribution is the fact that while at the time of Akaike's publication his approach could not be extended due to mathematical intractabilities, nowadays computer intensive methods allow the design of a straightforward algorithm to solve the model projection problem outlined above. These computational tools basically involve two methodologies: first, a numerical estimation of Kullback-Leibler (KL) divergences between arbitrary distributions and second, parallel processing to carry a Non-Metric Multidimensional (NMDS) space scaling algorithm. With the help of a NMDS, a matrix of amongst-candidate models estimated KL divergences can be transformed into an approximated Euclidean representation of models in a (hyper)plane. The coordinates of each model in that plane, that we heretofore denote (y_1, y_2, …) are used to solve the model projection problem. The algorithm presented here is not necessarily restricted to a two-dimensional representation of model space, but for the sake of visualization we present our development in $\mathscr{R}^2$.

Consider the sketch in Figure 3. There, to begin with we have drawn only two approximating models f_2 and f_3 on a Euclidean space, along with a depiction of the location of the generating process g outside that plane. Such representation immediately leads to the definition of a point m in that plane that correspond to the orthogonal projection of the generating process onto the plane. The location of such point is denoted as $(y_1^\star, y_2^\star)$. The length h in that sketch represents the deviation of the generating process from the plane of approximating models as a line from g to the plane that crosses such plane perpendicularly. Note also that every one of the approximating models f_i in that plane is situated at a distance $d(f_i, m)$ from the orthogonal projection m. In reality, both the edges as well as the points in this plane are random variables associated with a sampling error. But we ask the reader's indulgence for the sake of the argument, just as we did above when we explained Akaike's results, and think of these simply as points and fixed lengths. Doing so, one may also indulge, as Akaike did, in using the right-angle, simple version of the Pythagoras theorem, and assume that all the amongst-models KL divergences have a corresponding squared Euclidean distance in that representation. Then, the following equations hold

$$\begin{cases} KL(g, f_1) = d(f_1, m)^2 + h_1^2 \\ KL(g, f_2) = d(f_2, m)^2 + h_2^2 \\ \quad\vdots \end{cases}$$

where necessarily $h_1 = h_2 = h_i = \ldots = h$. Recalling Equation (8) we note that every one of the divergences between the approximating models and g can be expressed as a sum of an estimable term and a fixed, unknown term. These terms are Sgf_i and Sgg, respectively. Writing such decomposition of the KL divergences for all the equations above, and explicitly incorporating the coordinates of m then results in this system of equations

$$\begin{cases} Sgg - \widehat{Sgf_1} - d\left(f_1, m\left(y_1^\star, y_2^\star\right)\right)^2 = h_1^2, \\ Sgg - \widehat{Sgf_2} - d\left(f_2, m\left(y_1^\star, y_2^\star\right)\right)^2 = h_2^2, \\ \qquad\qquad\vdots \qquad\qquad\qquad \vdots \quad \vdots \end{cases} \tag{9}$$

which can be solved and optimized computationally by constructing an objective function that, for any given set of values of sgg, $y_1^\star$, $y_2^\star$ in the left hand of these equations returns the sum of squared differences between all the h_i. Because by necessity (see Figure 3) $h^2 = h_i^2$ for all i, a routine minimization of this sum of squared differences can be used as the target to obtain optimal values of the unknown quantities of interest and obtain the model-projection representation shown in Figure 5. Although previously unrecognized by Taper and Ponciano (2016a), in these equations the terms Sgg and h^2 appear always as a difference, and hence are not separable. Fortunately, a non-parametric, multivariate estimate of Sgg can be readily computed. We use the estimator proposed by Berrett et al. (2019), a multivariate extension of the well-known univariate estimator by Kozachenko and Leonenko (1987). Other non-parametric entropy estimators could be used if they prove to be more appropriate. For instance, the Berrett et al. (2019) estimator assumes that the data are iid. This restricts the class of problems for which we are able to separate Sgg and h^2. An estimator for Sgg for dependent data would expand the class.

3. EXAMPLES

In what follows we illustrate our ideas and methodology with two ecological examples. The first example is an animal behavior study aiming to understand the mechanism shaping patterns of animal aggregations. The second one is an ecosystems ecology example, where the aim was to try to understand the biotic and abiotic factors that shape the species diversity and composition of a shrubland ecosystem in California.

3.1. An Application in Animal Behavior

The phenomenon of animal aggregations has long been the focus of interest for evolutionary biologists studying behavior (Brockmann, 1990). In some animal species, males form groups surrounding females, seeking breeding opportunities. Often, these mating groups vary substantially in size, even during the same breeding season and breeding occasion. This is particularly true in some species with external fertilization where females spawn the eggs and one or more males may fertilize them. The females of the American horseshoe crab, *Limulus polyphemus* leave sea “en masse” to spawn at the beach during high tide, 1–4 times a year. As females enter the beach and find a place to spawn, males land in groups and begin to surround the females. Nesting typically occur in pairs, but some females attract additional males, called satellites, and spawn in groups. As a result, when surveys of the mating group size are done, one may encounter horseshoe crab pairs with 0, 1, 2, 3, … satellite males. That variation in the number of satellite males is at the root of the difficulty in characterizing the exact make-up of the crab population. Hence, for years during spawning events, Brockmann (1990) focused on recording not only the total number of spawning females in a beach in Seahorse Key (an island along Florida’s northern west coast) but also the number of satellite males surrounding each encountered pair. Those data have long been the focus of attempts at a probabilistic description of the distribution of the number of satellite males surrounding a pair of horseshoe crabs using standard distribution models (e.g., Poisson, zero inflated Poisson, negative binomial, zero inflated negative binomial, hurdle-negative binomial distributions).

When one of us (JMP) met H. J. Brockmann in 2010, she asked the following: “how will fitting different discrete probability distributions to my data help me understand the biological mechanisms underlying group formation in this species?” After years of occasional one-on one meetings and back and forth discussions, we put together a detailed study (Brockmann et al., 2018) in which we compared the observed distribution of the number of satellites surrounding a female to the same distribution resulting from a complex, individual-based model simulation program. Importantly, this individual-based model allowed us to translate different hypotheses regarding the influence of different factors, like female density or male density around a female, into the decision by a new satellite male of joining a mating group or continuing the search.

The comparison between the real data and the simulated data *via* discrete probability distributions then allowed these authors to identify the biological settings that resulted in *in silico* distributions of satellites that most resembled the real, observed distributions of satellite males. To do that comparison, Brockmann et al. (2018) first fitted a handful of discrete probability models to the counts of the number of satellites surrounding each pair from each one of $N = 339$ tides, and proceeded to find the standard probability model that best described the data. These authors then fitted the same models to the simulated data sets under different biological scenarios and found the simulation setting that yielded the highest resemblance between the real data and the digital data. Finally Brockmann et al. (2018) discuss the implications of the results.

One of the most relevant conclusions of these authors was that their comparative approach was useful as a hypothesis generator. Indeed, by finding via trial and error which biological processes gave rise in the individual-based simulations to distributions of satellites that most resembled the real distributions, the researchers basically came up with a system to elicit viable biological explanations for the mechanisms shaping the distribution of the number of satellite males surrounding a pair. This approach was an attempt to answer Brockmann’s initial question to JMP.

Here, we used the simulation setting of Brockmann et al. (2018) to exemplify how our Model Projections in Model Space (MPMS) approach can further our understanding of what are the model attributes that make a model a good model to better understand the underlying mechanisms generating the data. By having a complex simulation program, we can describe exactly the probability distribution of the data-generating process and we can validate our MPMS approach.

In what follows we first explain how we fitted our proposed models to the tides’ count data, and then how we compute the quantities needed to generate an approximate representation of models in model space that includes the estimated projection of the true, data-generating process.

3.1.1. Likelihood Function for the Satellites Count Data—

A handful of discrete probability models can be fit conveniently to the male satellites counts data using the same general likelihood functions by means of a reduced-parameter multinomial distribution model parameterization. As we will see below, this reduced-parameter multinomial likelihood formulation is instrumental to compute analytically the KL divergences between each one of the models as well as the neg-selfentropy. Many modern biological models, like phylogenetic Markov models, use this reduced-parameter formulation (Yang, 2000), and the example presented here can be readily used in many other settings in ecology and evolution (e.g., Rice, 1995).

In this example we adopt the following notation: the probability mass function of each discrete probability model i (i = 1, 2, … r where r is the number of models in the model set) is denoted as $f_i(x)$. Following Brockmann et al. (2018), we use $f_1(x)$ to denote the Poisson distribution (Poisson), $f_2(x)$ the negative binomial distribution (NegBin), $f_3(x)$ the zero inflated Poisson distribution (ZIP), $f_4(x)$ the zero inflated negative binomial distribution (ZINegBi), $f_5(x)$ a hurdle negative binomial distribution (HurdNBi), $f_6(x)$ a Poisson-negative binomial mixture (PoiNB), $f_7(x)$ a negative-binomial-Poisson mixture (NBPois), $f_8(x)$ a one-inflated Poisson distribution (OIPoiss), and $f_9(x)$ a one inflated negative-binomial distribution (OINegBi). In this example, $r = 9$.

We begin with the likelihood function for the counts for one tide, and extend it to the ensemble of counts for N tides Because for each tide j, j = 1, 2, …, N the data consisted of the number of 0’s, 1’s, etc…, the data can be represented as a multinomial sample with k categories and probabilities $\pi_1, \pi_2, \ldots, \pi_k$: Let Y_1 be the number of pairs with no satellites found at the beach in one tide, Y_2 the number of pairs with 1 satellite male in one tide, Y_3 the number of pairs with 2 satellite males in one tide, …, Y_{k-1} the number of pairs with $k -$ 2 satellites in one tide and Y_k the number of pairs with $k - 1$ or more satellites in one tide. Suppose for instance that we are to fit the Poisson distribution model with parameter λ to the counts of one tide. Then, the reduced parameter multinomial distribution arranged to fit the Poisson model would be parameterized using the following probabilities for each category:

$$\begin{aligned}
\pi_1 &= P(X=0) &&= f_1(0) &&= e^{-\lambda}, \\
\pi_2 &= P(X=1) &&= f_1(1) &&= \lambda e^{-\lambda}, \\
\pi_3 &= P(X=2) &&= f_1(2) &&= \frac{\lambda^2 e^{-\lambda}}{2!}, \\
&\vdots \\
\pi_{k-1} &= P(X=k-2) &&= f_1(k-2) &&= \frac{\lambda^{k-2} e^{-\lambda}}{(k-2)!} \\
\pi_k &= P(X \ge k-1) &&= 1 - \sum_{s=0}^{k-2} f_1(s) &&= 1 - \sum_{s=0}^{k-2} \frac{\lambda^s e^{-\lambda}}{(s)!}.
\end{aligned} \tag{10}$$

It follows that if in a given tide j a total of n_j pairs are counted and $y_{j,1}$ is the number of females with no satellites, $y_{j,2}$ is the number with one satellites, etc., such that $\sum_{i=1}^{k} y_{j,k} = n_j$, the likelihood function needed to fit the Poisson probability model to the data of one tide is simply written as:

$$
\begin{aligned}
L_j(\lambda) &= P\left(Y_{j,1} = y_{j,1}, Y_{j,2} = y_{j,2}, \ldots, Y_{j,k-1} = y_{j,k-1}, Y_{j,k} = y_{j,k}\right) \\
&= \frac{n!}{y_{j,1}!y_{j,2}!y_{j,3}!\ldots y_{j,k}!}\pi_1^{y_{j,1}}\pi_1^{y_{j,2}}\ldots\pi_k^{y_{j,k}},
\end{aligned}
$$

and the overall likelihood function for the N tides is simply

$$L(\lambda) = \prod_{i=1}^{N} L_j(\lambda).$$

Finally, note that for this reduced parameter multinomial model, the ML expected frequencies would simply be computed as $n_j\hat{\pi}_1$. For example, under the Poisson model, the expected number of 0's in a sample would be computed as $n_j\hat{\pi}_1 = \widehat{P(X=0)} = e^{-\hat{\lambda}}$, where $\hat{\lambda}$ denotes the ML estimate of λ.

The likelihood function and each of the predicted probabilities for every model were computed using the programs in the files CrabsExampleTools.R and AbundanceToolkit2.0.R downloadable from our github webpage, which works as follows. Suppose that for a single tide, the counts of the number of pairs with 0, 1, 2, 3, 4, and 5 or more satellites are 112, 96, 101, 48, 22, 16, respectively. Then, the program abund.fit (found in the set of functions AbundanceToolkit2.0.R) takes those counts and returns, for every model in a pre-specified model set, the expected frequencies (from which the probabilities of every category in the reduced-parameter multinomial are retrievable), the ML estimates of each set of model parameters, the maximized log-likelihood and other statistics.

The processes of simulating any given number of tide counts according to Brockmann et al. (2018) and computing the ML estimates and other statistics for every model and every tide in a pre-specified model set are packaged within our function short.sim() whose output is (1) a matrix of simulated counts, with one row per tide. In each row the data for a single tide is displayed from left to right, showing the number of pairs with 0, 1, 2, 3, 4, and 5 or more satellites. (2) a list with the statistics (ML estimates, maximized log-likelihood, predicted counts, etc…) for every model and every tide. (3) A matrix of information criteria values for every tide (row) and every model (column) in the set of tested models.

3.1.2. Calculation of Quantities Needed to Generate a MPMS—

The generation of the MPMS necessitates solving the system of Equation (9). To solve that system of equations for any given dat set we need

1. A non-parametric estimate of the neg-selfentropy Sgg, $\widehat{Sgg}$. Berrett et al. (2019) recently proposed such an estimator. Their estimator is in essence a weighted (Kozachenko and Leonenko, 1987) estimator, and uses k nearest neighbors of each observation as follows:

$$H_n^w = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} w_j \log \xi_{(j),i},$$

where $\xi_{(j),i} = (n - 1)e^{-\psi(j)}\, V_q \| X_{(j),i} - X_i \|^q$ with $X_{(j),i}$ indicating the j-th nearest neighbor from the i-th observation X_i. Also, in these equation n indicates the number of observations, $\psi(j)$ is the digamma function and $V_q = \pi^{q/2}/\Gamma(1 + q/2)$ is the volume of the unit q-dimensional ball and q is the dimension of the multivariate observations.

The focus of Berrett et al. (2019) was writing a complete theoretical proof of the statistical properties of their estimator. Practical guidance as to how to find these weights is however lacking in their paper, but through personal communication with T. Berrett we learned that their weights w_j must only satisfy the constraints (see their Equation 2):

$$\sum_{j=1}^{k} w_j = 1 \quad \text{and} \quad \sum_{j=1}^{k} w_j \Gamma(j + 2l/q)/\Gamma(j) = 0 \quad \text{for} \quad l = 1, \ldots, q/4,$$

where k is the number of observations that define a local neighborhood of observations around any given observation. Berrett (personal communication) recommends arbitrarily choosing k as the sample size to the power of a third. The other restrictions on Berrett et al. (2019)' theorem about the support of these weights were needed only for technical convenience for the proof. Berrett et al. (2019) also mentioned that for small sample sizes, the unweighted estimator may be preferable. For larger problems he recommended solving the above restrictions with a non-linear optimizator. We wrote such non-linear optimization routine to compute the weights w_j's and tested it extensively via simulations and embedded it into a function whose only argument is the data itself. Through extensive simulations we have verified that this routine works well for dimensions at least up to $q = 15$. We coded our optimization in R and is now part of a package of functions accompanying this paper. The function is found in the file MPcalctools.R and was named Hse.wKL. Finally, note that a typical data set for our crabs example is of dimension 6, so our routine is more than enough for a typical set of counts similar to the ones in this example. For instance, one set of counts of pairs with 0 satellites, 1 satellite, 2 satellites, …, 5 or more satellite males for one tide is $y_1 = 112$, $y_2 = 96$, $y_3 = 101$, $y_4 = 48$, $y_5 = 22$, $y_6 = 16$.

2. A matrix of KL divergences between all models estimated in the model set being considered. If a total of r models are being considered, then the elements of this matrix are $\{KL(f_i, f_j)\}_{i,j}$, $i, j = 1, 2, \ldots, r$. Computing these divergences may seem like a daunting task, especially because these quantities are, in fact, different expectations (i.e., infinite sums) evaluated at the ML estimates for each model in the model set. However, those calculations are enormously simplified by adopting the general reduced-parameter likelihood approach because the neg-crossentropy $H(f_r, f_s)$ between two multinomial models f_r and f_s with a total sample size n can be computed exactly:

$$
\begin{aligned}
H(f_r, f_s) \;=\; & \sum\nolimits_{(y_1, y_2, \ldots, y_k) \geq 0, (\sum_k y_k) = n} \frac{n!}{y_1! \ldots y_k!} \pi_{1,r}^{y_1} \cdots \\
& \pi_{k,r}^{y_k} \log\left[\frac{n!}{y_1! \ldots y_k!} \pi_{1,s}^{y_1} \ldots \pi_{k,s}^{y_k}\right] \\
& = \log n! + n \sum i = 1^k \pi_{i,r} \log \pi_{i,s} \\
& - \sum_{i=1}^{k} \sum_{y_i = 0}^{n} \binom{n}{y_i} \pi_{i,r}^{y_i} (1 - \pi_{i,r})^{(n - y_i)} \log y_i! \,.
\end{aligned}
\tag{11}
$$

Note that when $s = r$, then $H(f_r, f_s)$ becomes the neg-selfentropy. Because the KL divergence is the sum of a neg-selfentropy and a crossentropy, in practice, to compute the KL divergence between two count models for a single vector of counts for one tide we only needed to compute the probabilities in Equation (10) for every model using the ML estimates for each data set and use Equation (11) above. The function in R used to compute either the neg-crossentropies or the neg-selfentropies is named H.multinom.loop() and found in the file MPcalctools.R. Following simple rules of expected values, the overall KL divergence between two count models for a set of N vectors of tide counts, each drawn from the same true generating process (the individual-based model simulator program), was just computed as the sum of the divergences between the two models for each vector of counts. Note that the same simplification in Equation (11) applies to the computation of the neg-selfentropy for a multinomial distribution, a fact that we used to compute the true Sgg for our simulator algorithm, given that the individual-based model simulator of Brockmann et al. (2018) could be used to the estimate numerically true probabilities for 0,1,2,… satellites.

3. The estimates of the neg-crossentropies $\widehat{Sgf_i}$ and of $\widehat{Sf_ig}$ for $i = 1, 2, \ldots, r$. Although the first set of divergences, the $\widehat{Sgf_i}$, can be estimated either using the AIC and Equation (8), by definition of the KL divergence, the estimates $\widehat{Sf_ig}$ are in general not equal to the estimates $\widehat{Sgf_i}$ and cannot be computed using the AIC and Equation (8). If however, h^2 is very small, then using the approximation $\widehat{Sgf_i} \approx \widehat{Sf_ig}$ works quite well as we show in example 3.2 and in Taper and Ponciano (2016a). Fortunately, using this approximation is not always necessary and does not have to be used for a large class of statistical problems. Indeed, for the example at hand where we are fitting multiple count models and for any other case where the likelihood function may be written by means of a reduced-parameter multinomial model (like the likelihood function for most phylogenetics models, for instance), both the $\widehat{Sgf_i}$ and the $\widehat{Sf_ig}$ can be computed using Equation (11) by using the ML estimates of the multinomial $\boldsymbol{\pi}$'s for each model and the ML estimates of the $\boldsymbol{\pi}$'s for the fully parameterized (i.e., the empirical model) in lieu of the $\boldsymbol{\pi}$ parameter values for g. We will denote these empirical estimates (i.e., the sample proportions) as $\bar{\pi}_i$. These estimates and Equation (11) can be used to compute $\widehat{Sgg}$. For a set of models and a data set including one or more tides, the estimates $\widehat{Sgf_i}$, $\widehat{Sf_ig}$, $\widehat{Sf_if_j}$ and $\widehat{Sgg}$ are computed using the function entropies.matcalc() found in the file MPcalctools.R.

For a simulated example where the data consisted of counts for 300 tides for which the first 5 tides were

```
> simdat [1:5,]
        0    1    2   3   4   5
[1,]  112   96  101  48  22  16
[2,]  135  125  108  44  19  12
[3,]  141  108   91  55  23  16
[4,]  119  117   99  60  18  10
[5,]  139  120  117  37  26  11
```

The estimated matrix of neg-crossentropies for these N = 300 tides was

```
$Sfifjs.hat
            Poisson     NegBin     ZIPoiss    ZINegBi    HurdNBi    PoisNB     NBPois     OIPoiss
Poisson    −4693.860  −6788.198  −7144.127  −7261.358  −7276.347  −7240.670  −7268.412  −4694.360
NegBin     −7140.595  −4801.609  −5748.393  −5402.616  −5393.795  −5141.644  −5417.371  −7142.133
ZIPoiss    −7269.760  −5688.618  −4778.731  −4958.789  −4991.042  −5157.351  −4973.420  −7271.568
ZINegBi    −7483.025  −5374.228  −4980.700  −4792.072  −4868.308  −5004.579  −4952.840  −7484.694
HurdNBi    −7504.330  −5366.989  −5015.873  −4870.053  −4792.890  −4980.233  −4974.241  −7503.976
PoisNB     −7476.723  −5131.165  −5178.562  −5008.269  −4982.275  −4794.920  −4975.992  −7477.044
NBPois     −7453.540  −5389.715  −4984.999  −4949.951  −4969.751  −4970.038  −4790.289  −7455.437
OIPoiss    −4694.328  −6789.643  −7145.910  −7262.925  −7275.909  −7240.883  −7270.251  −4693.826
OINegBi    −5606.894  −6051.992  −6648.368  −6590.061  −6559.321  −6453.790  −6596.554  −5606.393
```

The true generating process neg-selfentropy, Sgg was −16.01199 and the estimated neg-selfentropy $\widehat{Sgg}$ was −15.96137. The real neg-crossentropies between the generating process and each of the models Sgf_i's and Sf_ig's were:

```
$Sgfis
  Poisson     NegBin     ZIPoiss    ZINegBi    HurdNBi    PoisNB     NBPois     OIPoiss    OINegBi
−7601.606  −5603.842  −5485.607  −5432.819  −5457.657  −5450.109  −5463.949  −7606.407  −6832.673

$Sfisg
  Poisson     NegBin     ZIPoiss    ZINegBi    HurdNBi    PoisNB     NBPois     OIPoiss    OINegBi
−7276.506  −5615.999  −5431.557  −5414.428  −5439.071  −5436.292  −5435.686  −7281.250  −6653.690
```

whereas the estimated neg-crossentropies were

```
$Sgfis.hat
  Poisson     NegBin     ZIPoiss    ZINegBi    HurdNBi    PoisNB     NBPois     OIPoiss    OINegBi
−7891.890  −5652.049  −5403.773  −5200.907  −5179.331  −5323.294  −5253.094  −7891.705  −7051.753

$Sfisg.hat
  Poisson     NegBin     ZIPoiss    ZINegBi    HurdNBi    PoisNB     NBPois     OIPoiss    OINegBi
```

−7638.034	−5682.275	−5396.418	−5213.190	−5193.804	−5339.449	−5264.447	−7637.716	−6906.558

4. The coordinates of every model in an NMDS space. Multidimensional scaling, MDS, is an established method (Borg et al., 2018) for representing the information in the $s \times s$ matrix D of distances/divergences among s objects as a set of coordinates for the objects in a k-dimensional euclidian space ($k \leq s$). If $k < s$, there may be some loss of information. MDS has two major varieties, metric multidimensional scaling, MMDS, in which D is assumed to be comprised of Euclidean distances, and non-metric multidimensional scaling, NMDS, in which D can be made up of divergences only monotonically related to distances. The MMDS projection can be made analytically, while the NMDS projection can only be found algorithmically by iteratively adjusting the configuration to minimize a statistic known as "Stress," which is a weighted average squared deviation of the distances between points (models in our case) calculated from the proposed configuration and the distances given in D.

The matrix D required by NMDS should be symmetric. KL divergences are not, however, symmetric. The KL divergence can be reasonably symmetrized in a number of ways (Seghouane and Amari, 2007). We symmetrize using the arithmetic average of $KL(\theta_i, \theta_j)$ and $KL(\theta_j, \theta_i)$. As mentioned above in this problem we can directly calculate the symmetric KL. For other applications the symmetric KL can be estimated (up to the constant Sgg) using the KIC and its small sample version the KiCc (Cavanaugh, 1999, 2004). We follow Akaike in considering the KL divergence as a squared distance, and thus construct the matrix D from the square roots of the symmetrized KL divergence. We use the function smacofSym (De Leeuw and Mair, 2009) from the R package smacof (version 2.0, Mair et al., 2019) to calculate the NMDS. For the purposes of this paper we chose $k = 2$ so that we could have a graphical representation after augmenting the dimension to 3 to show the orthogonal distance from the generating process to its orthogonal projection M in the estimated plane of models. Nevertheless, the Stress of 0.029 indicates an excellent fit. Except for the very important aspect of visualization, dimension reduction is not an essential aspect of our method. Finally, the tight pseudo-confidence ellipses (95%) illustrated in Figure 4, based on Stress derivatives (Mair et al., 2019) indicate that this NMDS is quite stable.

Once all these components are computed, the system of Equation (9) can be solved with non-linear optimization. We coded such solution in the R function MP.coords found in the file *MPcalctools.R*. This function takes as input the estimated neg-crossentropies between all models, an estimate of Sgg or the neg-selfentropy of the generating process, and the vectors of estimated neg-crossentropies $\widehat{Sgf_i}$ and $\widehat{Sf_ig}$ to output the matrix of dimension $(r + 1) \times (r + 1)$ of symmetrized KL divergences, and the results of the NMDS with the coordinates of every model in a two-dimensional space, the estimated location of the orthogonal projection of g in such plane, M, and the estimate of h. Notably, this function works for any example for which these estimated quantities are available. Its output is taken by our function plot.MP to produce the three-dimensional representation of the Model Projection in Model

Space shown in Figure 5. For this example, the estimated distances in the model projection space between all models, *g* and its projection *M* were

```
          Poisson  NegBin   ZIPoiss  ZINegBi  HurdNBi  PoisNB   NBPois   OIPoiss  OINegBi        M
NegBin    4.46711
ZIPoiss   5.09358  1.51616
ZINegBi   5.21422  1.26069  0.42784
HurdNBi   5.22444  1.19699  0.52347  0.09773
PoisNB    5.10858  0.89561  0.81309  0.42542  0.33690
NBPois    5.19883  1.24348  0.43208  0.01798  0.09139  0.41228
OIPoiss   0.00181  4.46859  5.09530  5.21588  5.22609  5.11018  5.20050
OINegBi   1.69500  2.85332  3.73875  3.76818  3.75926  3.58746  3.75125  1.69618
M         4.51235  2.33280  1.26566  1.67451  1.76140  1.97444  1.67387  4.51416  3.50306
g         4.51235  2.33280  1.26566  1.67451  1.76140  1.97444  1.67387  4.51416  3.50306  0.00032
```

whereas the real distances (because we knew what the simulation setting was) were

```
> dist(true.MP$XYs.mat)
          Poisson  NegBin   ZIPoiss  ZINegBi  HurdNBi  PoisNB   NBPois   OIPoiss  OINegBi        M
NegBin    4.46711
ZIPoiss   5.09358  1.51616
ZINegBi   5.21422  1.26069  0.42784
HurdNBi   5.22444  1.19699  0.52347  0.09773
PoisNB    5.10858  0.89561  0.81309  0.42542  0.33690
NBPois    5.19883  1.24348  0.43208  0.01798  0.09139  0.41228
OIPoiss   0.00181  4.46859  5.09530  5.21588  5.22609  5.11018  5.20050
OINegBi   1.69500  2.85332  3.73875  3.76818  3.75926  3.58746  3.75125  1.69618
M         4.14688  2.14942  1.34587  1.70959  1.78455  1.93970  1.70493  4.14868  3.12059
g         4.14688  2.14942  1.34587  1.70959  1.78455  1.93970  1.70493  4.14868  3.12059  0.00037
```

From these matrices, it is readily seen that the real value of *h* in the model projection space was 0.000372 whereas its corresponding estimated *h* value is 0.000323. A quick calculation yields the distances between the true location of the orthogonal projection *M*, its estimate, the true location of *g* and its estimate:

```
          hat.m     hat.g     true.m
hat.g    0.000323
true.m   0.383074  0.383074
true.g   0.383074  0.383074  0.000372
```

As expected, variation in the quality of these estimates and the difference with the true locations changes from simulated dat set to simulated data set. Two questions are a direct consequence of this observation: first, the MPMS data representation in Figure 5 could be more accurately depicted via bootstrap and confidence clouds or spheres for the location of each model in model space could be drawn. Such task would however involve entertaining the problem of the representation of multiple bootstrap NMDS runs in a single space, using the same rotation.

Classically, variation among NMDS object has been estimated only after Procrustes rotation has oriented the various coordinate systems for maximal similarity among the NMDS objects (see Mardia et al., 1979). A long series of articles involving authors, such as T. M. Cole, S. R. Lele, C. McCulloch, and J. Richtsmeir demonstrates that this approach is deeply flawed. This work is summarized in the monograph by Lele and Richtsmeier (2001). The problem is that the apparent variability among equivalent points in the multiple objects depends on distance from the center of rotation. Lele and Richtsmeir argue that inference is better made regarding variation in estimated distances between points than on the coordinates of points. A mean distance matrix can be estimated from a set of bootstrapped replicates, and it is almost certain that the mean distance matrix will be the most informative matrix both for inference and for graphical purposes as this mean corresponds to the expectation with respect to $\hat{\theta}$ in Akaike's 5th insight (see section 2.1.5). Further, variation and covariation in all estimated distances and contrasts of distances can be invariantly calculated and used for inference. Finally, extending our MPMS methodology to include confidence bounds for our estimates is a topic of current research in our collaboration and will be treated in a future manuscript because it necessitates the same degree of care used to generate confidence intervals for Model-Average inferences (see for instance Turek, 2013).

The second question has to do with how would our estimate of the location in model space of the orthogonal projection of the generating process compare to the location of the model-average. For our example at hand, the AIC values as well as the ΔAICs were:

```
> AICs
   Poisson     NegBin    ZIPoiss   ZINegBi   HurdNBi     PoisNB    NBPois     OIPoiss    OINegBi
14301.321   9823.639   9327.086  8923.355  8880.203   9168.129  9027.727   14302.951  12625.047

> delta.is
    Poisson      NegBin      ZIPoiss    ZINegBi    HurdNBi      PoisNB      NBPois      OIPoiss      OINegBi
5421.11814   943.43554   446.88328   43.15146    0.00000   287.92594   147.52444   5422.74769   3744.84389
```

To compare the estimated location of the model average with our estimated our model projection, we plotted both panels in Figure 6 into a single, two-dimensional figure with: the location of every estimated model, the location of the model averaged coordinates using the AIC weights, the location of the estimated orthogonal projection of *g*, and the location of the true location of the orthogonal projection *g*. Such figure is presented in Figure 6. In this figure, the distance between the real projection *M* of *g* and our estimated projection is 0.383074 whereas the distance between the model-average and the real projection of *g* is 1.784555. A quick inspection of Figure 6 shows that this case in fact, is a real-life illustration of the point brought up by Figure 3B. When the geometry of the model space is as in Figures 3B, 5, 6, model averaging may not be a suitable enterprise.

3.2. An Ecosystems Ecology Application

Here we discuss a worked example highlighting the strengths of the model projections approach to multi-model inference. This example was originally presented in Taper and Ponciano (2016a) which is freely downloadable from: https://link.springer.com/book/10.1007/978-3-319-27772-1 This example is an analysis of data simulated from a structural equation model (SEM) based on a study by Grace and Keeley (2006). Simulation from a known model in necessary to understand how well our methods capture information about the generating process, while basing that model on published research guarantees that our test-bed is not a toy, but is a problem of scientific interest. SEM is a flexible statistical method that allows scientists to analyze the causal relationship among variables and even general theoretical constructs (Grace and Bollen, 2006, 2008; Grace, 2008; Grace et al., 2010). Grace and Keeley (2006) analyzed the development of plant diversity in California shrublands after natural fires. Structural equations models were used to make inferences as to the causal mechanisms influencing changes in diversity. Plant composition at 90 sites was followed for 5 years. The Grace and Keely final model is displayed in Figure 7. To summarize the causal influences, species richness is directly affected by heterogeneity, local abiotic conditions, and plant cover. Heterogeneity and local abiotic conditions are in turn affected by landscape position, but total cover is only directly affected by burn intensity. Burn intensity is in turn only affected by stand age, which itself depend on landscape position. Affects and their direction are shown as arrows in the figure. The strength of affects (i.e., the path coefficients) are shown both as numbers on the figure and as the thickness of the arrows).

Forty-one models were fit to our generated data. The models ranged from underfitted to overfitted approximations of the generating process. The actual generating model was not included in this model set. Using this set of fitted models, we estimated a 2-d Non-Metric Dimensional Scaling model space as discussed above. The calculated stress was tiny (0.006%) indicating almost all higher dimensional structure is captured by an $\mathscr{R}^2$ plane. A mapping of the estimated space analogous to our Figure 6 is shown in their Figure 6 Taper and Ponciano (2016a). ΔAIC values are indicated by color. As in Figure 6 of this paper, on this map of model space we also indicated: (1) The estimated projection (location) of the generating process to the 2-d NMDS space, (2) The Akaike weighted model averaged location and 3) The actual projection of the true generating process l onto the 2-d manifold (in this worked example this can be done because we have simulated from a known model).

Two important observations can be made based on the graph in Figure 6 (both in this manuscript and in Taper and Ponciano, 2016a) : First while there is a rough agreement between proximity to the generating process and ΔAIC values, this relationship is not as tight as one might naively expect. The inter-model KL distances do have substantial impact on the map. Second, using our methods and just like in example 3.1 above, the estimated projection of the generating process is somewhat nearer to the actual projection of the generating process than the location produced by model-averaging (Figure 6 in this manuscript).

Figure 8 demonstrates the sensitivities of both the estimated projection and model average of eliminating fitted models from the estimation of the NMDS space. We repeatedly eliminate the left-most model in the model set and reestimate the space after each cycle. With each model elimination, the model-averaged location moves toward the right. On the other hand, the estimated projection stays near its original location, even after all fitted models in that side of the map have been eliminated. Conversely, eliminating from the right, the model average shifts to the left as anticipated. Under right-side model elimination, the model projection is somewhat more variable than under elimination from the other direction.

This model elimination example illuminates differences in the two kinds of estimates the generating process location. These differences follow directly from the geometric development of the AIC by Akaike, and from the mathematics of model averaging. (1) The model average must fall inside of the bounds of the fitted models. changing the model set will, except in contrived cases, change the model average. (2) Because it is a projection, our method's estimate of the generating process' location can fall outside the bounds of the model set. And (3), because of the nature of projection geometry, farther models can inform the estimated situation of the generating process in the NMDS map. Point (3) is demonstrated in the discrepancy in the stability of the model projection location under model elimination from the left and model elimination from the right. There are several models with high influence that are deleted quickly under model elimination from the right that stay in the model set much longer under elimination from the left.

Our approach calculates two important diagnostic statistics not even thought of in model averaging. The first is measure of the dispersion of the generating process. This is the neg-selfentropy or Sgg. In this example it is calculated to be −9.881, very close to the known magnitude of −9.877. The second statistic is an estimate of the perpendicular distance of the generating process to the NMDS manifold (h in Equation 9). This diagnostic is critical for proper interpretation of your model set. If the generating process is far from NMDS manifold, then any statistic based on models in the model set is likely to be inaccurate. Using our approach we calculate h to be 0.0002. The known h is $6e-08$.

3.3. Testing the Non-parametric Estimation of *Sgg*

To exemplify the independent estimation of Sgg with a data set we simulated samples from a seven-dimensional multivariate normal distribution and compared the true value of Sgg with its non-parametric estimate according to Berrett et al. (2019). We chose to simulate data from a multivariate normal distribution because its Sgg value is known analytically. When the dimension of a multivariate normal distribution is p and is variance-covariance matrix is Σ, then

$$Sgg = -\frac{1}{2}\ln\left\{(2\pi e)^{p}\det(\Sigma)\right\}. \tag{12}$$

To carry our test, we chose five testing sample sizes 10, 25, 50, 75, 150, and for each sample size we simulated 2,000 data sets according to a multivariate normal distribution with $p = 7$ and $\Sigma = I$, and computed each time Berrett et al.'s non-parametric estimate. The resulting estimates, divided by the true value of 9.93257 are plotted as boxplots in Figure 9.

4. DISCUSSION

We have constructed a novel approach to multi-model inference. Standard multi-model selection analyses only estimate the relative, not overall divergences of each model from the generating process. Typically, divergence relationships amongst all of the approximating models are also estimable (dashed lines in Figure 5). We have shown that using both sets of divergences, a model space can be constructed that includes an estimated location for the generating process (the point *g* in Figure 5). The construction of such model space stems directly from a geometrical interpretation of Akaike's original work.

The approach laid out here has clear and substantial advantages over standard model identification and Bayesian based model averaging. A heuristic approach aiding the development of novel models is now possible by simply being able to visualize a set of candidate models in an Euclidean space. Now the overall architecture of model space vis-a-vis the generating process is statistically estimable. Such architecture is composed of a critical set of quantities and relationships. Among these objects, we now include the estimated coordinates of the closest orthogonal generating model projection onto the manifold of candidate models (the point *M* in Figure 5). Second, the estimated magnitude of the total divergence between the truth and its orthogonal projection onto the manifold of models can give the analyst an indication of whether important model attributes have been overlooked.

In the information criterion literature and all scientific application, the neg-selfentropy *Sgg* of the generating process is simply treated as an unknown quantity. In fact, it can be estimated quite precisely as our example shows. *Sgg* is itself of great interest because with it the overall discrepancy to the generating process becomes estimable. Because this quantity is estimable, now the analyst can discern the overall quality and proximity of the model set under scrutiny. Thus, our approach solves a difficulty that has long been recognized (Spanos, 2010) but yet treated as an open problem.

Studying the model space architecture gives the information to correct for misleading evidence (the probability of observing data that fails to support the best model), accommodation (over-fitting), and cooking your models (Dennis et al., 2019). The scaffolding from which to project the location of the generating process is estimated can be rendered more robust simply by considering more models. This is an interesting result that we expect will later contribute to the discussion of data dredging. On the other hand, non-identifiability and weak estimability (Ponciano et al., 2012) are, of course, still a problem, but at least the model space approach will clearly indicate the difficulties.

As conceived here, model projection is an evidential alternative (Taper and Ponciano, 2016b) to model averaging using Akaike weights (or other Bayesian alternatives) because it incorporates the available information estimated by many models without the redundancy inherent in model averaging. Through model projection the analyst can use more of the information available but usually ignored. Furthermore, our methodology provides new important diagnostic statistics previously not considered by model averaging: *Sgg* and *h*. As we showed in our results, model projection is not as sensitive as model average to the composition of the set of candidate models being investigated. Model averaging appears to artificially favor redundancy of model specification: the more models are developed in any given region of model space, the stronger this particular region gets weighted during the model averaging process. Finally, an emergent pattern in the analysis is that the optimization problem of our model projection methodology can be used to project outside the bounds of the available model set whereas the model averaging methodology, by definition, cannot.

As well as proposing solutions to existing problems, any new method also raises a variety of technical problems that need to be solved. This is certainly the case with the model projection approach presented here.

Our methodology bears a near-model limitation that, although important, is shared with the usage of Akaike's Information Criterion. Our exposition makes it clear that near model requirement is due to the imperfect yet useful approximation employed by Akaike while setting $\phi \approx \pi/2$ (see Figure 1). It was only thanks to this approximation that Akaike was able to solve for the estimable divergence contrasts between all approximating models and the generating process. This approximation breaks down in curved model spaces as the divergence from the generating process increases. Indeed, as the KL distance between approximating models and the generating model increases, $-AIC/2n$ becomes an increasingly biased and variable estimate of the Sgf component of the KL distance between the approximating model and the generating model. This effect is strong enough that sometimes very bad models can have very low ΔAIC scores, sometimes even as low as the minimum score. The TIC (Takeuchi, 1976) and the EIC2 (Konishi and Kitagawa, 2008; Kitagawa and Konishi, 2010) are model identification criteria designed to be robust to model misspecification. Substituting one of these information criteria for the AIC in constructing the matrix of inter-model divergences should allow the use of models more distant from truth than is acceptable using the AIC.

Our methodology focuses on estimation of the model space geometry but uncertainties around such estimation are not fully worked out as of yet. Work in progress by Taper, Lele, Ponciano and Dennis, the estimation of the uncertainties associated with doing inference with evidence functions, such as ΔSIC scores, can be assessed *via* non-parametric bootstrap techniques. We expect bootstrap to be also useful to reduce the variance of information criterion's bias correction (Kitagawa and Konishi, 2010).

We think that this model projection methodology should be the starting point to do a careful, science-based inquiry of what are the model attributes that make a model a good model. Knowing the location of the projected best model is an essential component of our multi-model development strategy because a response surface analysis can reveal what model attributes tend to be included near the location of the projected best model thus aiding in the construction of a model closer to the best projection.

ACKNOWLEDGMENTS

The authors appreciate the feedback and insights from the audience at the International Symposium What is a good model?: Evidential statistics, information criterion and model evaluation held January 2016 at the Institute of Statistical Mathematics in Tokyo, Japan (the home institute of Hirotugu Akaike). We also appreciate the insightful comments from the audiences from Virginia Tech's Statistics Department Seminar and University of Idaho's Applied Statistics Seminar, specially to Drs. Leah Thompson and Bill Price.

FUNDING

Both authors were partially funded via a grant from JSPS KAKENHI Grant Number 25340110 (to Shimatani Kenichiro) and from the Risk Analysis Center in the Institute of Statistical Mathematics. JP was partially supported by two grants from the US National Institute of Health grant number to University of Florida NIH R01 GM103604 to JP and NIH 1R01GM117617 to Profs. Jason K. Blackburn and JP.

REFERENCES

Akaike H (1973). "Information theory as an extension of the maximum likelihood principle," in Second International Symposium on Information Theory, eds Petrov B, and Csaki F (Budapest: Akademiai Kiado), 267–281.

Akaike H (1974). A new look at statistical-model identification. IEEE Trans. Autom. Control 19, 716–723.

Bandyopadhyay PS, Brittan G Jr., and Taper ML (2016). Belief, Evidence, and Uncertainty Problems of Epistemic Inference. Basel: Springer International Publisher.

Berrett TB, Samworth RJ, Yuan M (2019). Efficient multivariate entropy estimation via *k*-nearest neighbour distances. Ann. Stat 47, 288–318. doi: 10.1214/18-AOS1688

Borg I, Groenen PJ, and Mair P (2018). Applied Multidimensional Scaling and Unfolding. Cham: Springer.

Brockmann HJ (1990). Mating behavior of horseshoe crabs, limulus polyphemus. Behaviour 114, 206–220.

Brockmann HJ, St Mary CM, and Ponciano JM (2018). Discovering structural complexity and its causes: breeding aggregations in horseshoe crabs. Anim. Behav 143, 177–191. doi: 10.1016/j.anbehav.2017.10.020

Burnham KP, and Anderson DR (2004). Multimodel inference: understanding aic and bic in model selection. Sociol. Method Res 33, 261–304. doi: 10.1007/b97636

Burnham KP, Anderson DR, and Huyvaert KP (2011). Aic model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behav. Ecol. Sociobiol 65, 23–35. doi: 10.1007/s00265-010-1029-6

Casquilho JP, and Rego FC (2017). Discussing landscape compositional scenarios generated with maximization of non-expected utility decision models based on weighted entropies. Entropy 19:66. doi: 10.3390/e19020066

Cavanaugh JE (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. Stat. Probab. Lett 42, 333–343.

Cavanaugh JE (2004). Criteria for linear model selection based on kullback's symmetric divergence. Austr. N. Zeal. J. Stat 46, 257–274. doi: 10.1111/j.1467-842X.2004.00328.x

Cushman SA (2018). Calculation of configurational entropy in complex landscapes. Entropy 20:298. doi: 10.3390/e20040298

Davison AC (2003). Statistical Models, Vol. 11. Cambridge, UK: Cambridge University Press.

De Leeuw J (1992). "Introduction to akaike (1973) information theory and an extension of the maximum likelihood principle," in Breakthroughs in Statistics, eds Kotz S, and Johnson NL (London: Springer), 599–609.

De Leeuw J, and Mair P (2009). Multidimensional scaling using majorization: Smacof in R. J. Stat. Softw 31, 1–30. doi: 10.18637/jss.v031.i03

Dennis B, Ponciano J, Taper M, and Lele S (2019). Errors in statistical inference under model misspecification: evidence, hypothesis testing, and AIC. Front. Ecol. Evol 7:372. doi: 10.3389/fevo.2019.00372

Fan Y, Yu G, He Z, Yu H, Bai R, Yang L, et al. (2017). Entropies of the chinese land use/cover change from 1990 to 2010 at a county level. Entropy 19:51. doi: 10.3390/e19020051

Grace JB (2008). Structural equation modeling for observational studies. J. Wildl. Manage 72, 14–22. doi: 10.2193/2007-307

Grace JB, Anderson TM, Olff H, and Scheiner SM (2010). On the specification of structural equation models for ecological systems. Ecol. Monogr 80, 67–87. doi: 10.1890/09-0464.1

Grace JB, and Bollen KA (2006). The Interface Between Theory and Data in Structural Equation Models. Reston, VA: US Geological Survey.

Grace JB, and Bollen KA (2008). Representing general theoretical concepts in structural equation models: the role of composite variables. Environ. Ecol. Stat 15, 191–213. doi: 10.1007/s10651-007-0047-7

Grace JB, and Keeley JE (2006). A structural equation model analysis of postfire plant diversity in California shrublands. Ecol. Appl 16, 503–514. doi: 10.1890/1051-0761(2006)016[0503:ASEMAO]2.0.CO;2 [PubMed: 16711040]

Gravel D, Massol F, and Leibold MA (2016). Stability and complexity in model meta-ecosystems. Nat. Commun 7:12457. doi: 10.1038/ncomms12457 [PubMed: 27555100]

Kitagawa G, and Konishi S (2010). Bias and variance reduction techniques for bootstrap information criteria. Ann. Stat. Math 62:209. doi: 10.1007/s10463-009-0237-1

Konishi S, and Kitagawa G (2008). Information Criteria and Statistical Modeling. New York, NY: Springer Science & Business Media.

Kozachenko L, and Leonenko NN (1987). Sample estimate of the entropy of a random vector. Probl. Pered. Inform 23, 9–16.

Kullback S, and Leibler RA (1951). On information and sufficiency. Ann. Math. Stat 22, 79–86.

Kuricheva O, Mamkin V, Sandlersky R, Puzachenko J, Varlagin A, and Kurbatova J (2017). Radiative entropy production along the paludification gradient in the southern taiga. Entropy 19:43. doi: 10.3390/e19010043

Leibold MA, Holyoak M, Mouquet N, Amarasekare P, Chase JM, Hoopes MF, et al. (2004). The metacommunity concept: a framework for multi-scale community ecology. Ecol. Lett 7, 601–613. doi: 10.1111/j.1461-0248.2004.00608.x

Lele SR, and Richtsmeier JT (2001). An Invariant Approach to Statistical Analysis of Shapes. Boca Raton, FL: Chapman and Hall/CRC.

Mair P, Groenen P, and De Leeuw J (2019). More on multidimensional scaling and unfolding in R: smacof version 2. J. Stat. Softw [Epub ahead of print].

Mardia K, Kent J, and Bibby J (1979). Multivariate Statistics. San Diego, CA: Academic Press.

Milne BT, and Gupta VK (2017). Horton ratios link self-similarity with maximum entropy of eco-geomorphological properties in stream networks. Entropy 19:249. doi: 10.3390/e19060249

Pawitan Y (2001). In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford: Oxford University Press.

Ponciano JM, Burleigh JG, Braun EL, and Taper ML (2012). Assessing parameter identifiability in phylogenetic models using data cloning. Syst. Biol 61, 955–972. doi: 10.1093/sysbio/sys055 [PubMed: 22649181]

Rice J (1995). Mathematical Statistics and Data Analysis. Duxbury advanced series. Belmont, CA: Duxbury Press.

Roach TN, Nulton J, Sibani P, Rohwer F, and Salamon P (2017). Entropy in the tangled nature model of evolution. Entropy 19:192. doi: 10.3390/e19050192

Seghouane A-K, and Amari S-I (2007). The aic criterion and symmetrizing the Kullback-Leibler divergence. IEEE Trans. Neural Netw 18, 97–106. doi: 10.1109/TNN.2006.882813 [PubMed: 17278464]

Spanos A (2010). Akaike-type criteria and the reliability of inference: model selection versus statistical model specification. J. Econometr 158, 204–220. doi: 10.1016/j.jeconom.2010.01.011

Takeuchi K (1976). The distribution of information statistics and the criterion of goodness of fit of models. Math. Sci 153, 12–18.

Taper ML, and Ponciano J (2016a). “Book appendix. projections in model space: multi-model inference beyond model averaging,” in Belief, Evidence, and Uncertainty: Problems of Epistemic Inference, eds Bandyopadhyay PS, Brittan GG, and Taper ML (Springer), 157–173.

Taper ML, and Ponciano JM (2016b). Evidential statistics as a statistical modern synthesis to support 21st century science. Popul. Ecol 58, 9–29. doi: 10.1007/s10144-015-0533-y

Turek D (2013). Frequentist model-averaged confidence intervals (Ph.D. thesis), University of Otago, Dunedin, New Zealand.

Yang Z (2000). Complexity of the simplest phylogenetic estimation problem. Proc. R. Soc. Lond. B Biol. Sci 267, 109–116. doi: 10.1098/rspb.2000.0974

Yang Z, and Zhu T (2018). Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. Proc. Natl. Acad. Sci. U.S.A 115, 1854–1859. doi: 10.1073/pnas.1712673115 [PubMed: 29432193]

Zeng Q, and Rodrigo A (2018). Neutral models of short-term microbiome dynamics with host subpopulation structure and migration limitation. Microbiome 6:80. doi: 10.1186/s40168-018-0464-x [PubMed: 29703247]

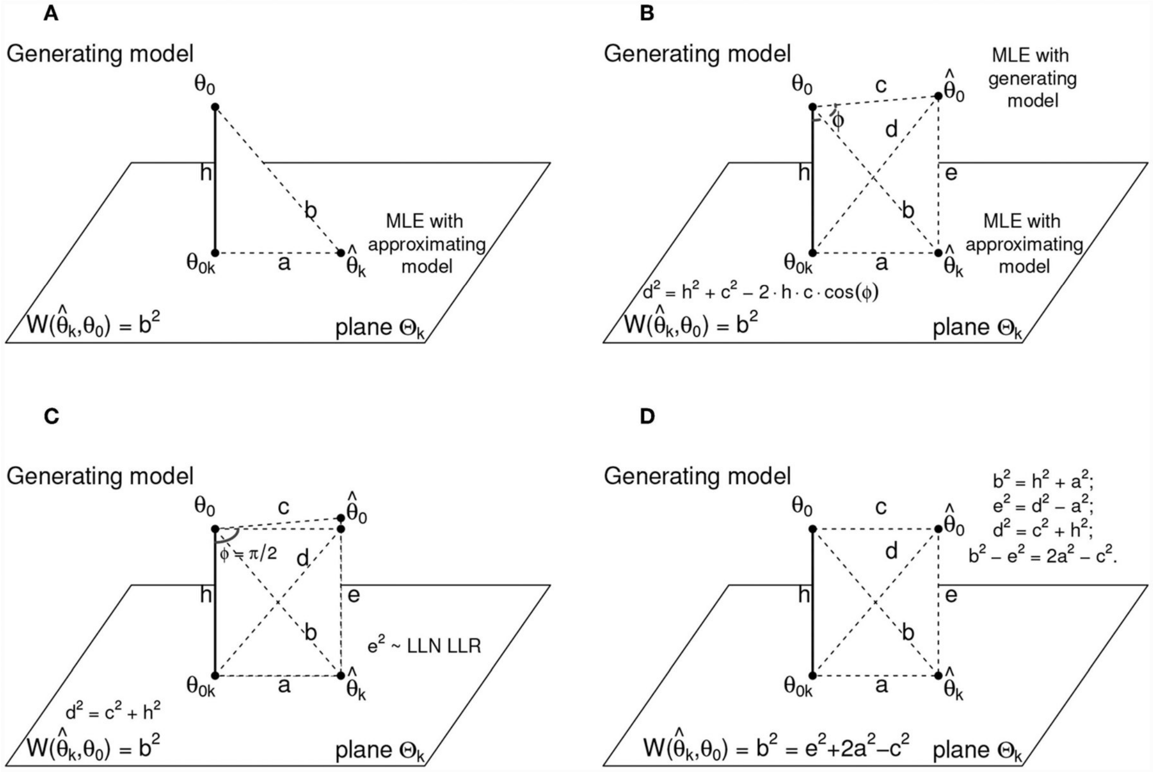

FIGURE 1 |.

The geometry of Akaike’s Information Criterion. **(A)** Shows θ_0, which is the generating model and θ_{0k} which is the orthogonal projection of the generating model into the space Θ_k of dimension k. $\hat{\theta}_k$ is the ML estimate (MLE) of an approximating model of dimension k given a data set of size n. Akaike’s objective was to solve for b^2, which represents in this geometry $\mathscr{W}\left(\hat{\theta}, \theta_0\right)$, the quadratic form approximation of the divergence between the generating and the approximating models. Akaike showed that $\hat{\theta}_k$ can be thought of as the orthogonal projection of the MLE of $\hat{\theta}_0$ **(B)**. This last quantity $\hat{\theta}_0$ represents the MLE of θ_0 with a finite sample of size n and assuming that the correct model form is known. The angle ϕ is not necessarily a right angle, but Akaike used $\phi \approx \pi/2$ so that the generalized Pythagoras theorem [equation on the lower left side of **(B)**] could be approximated with the simple version of Pythagoras [equation on the lower left side of **(C)**] when the edge h is not too long. When implemented, this Pythagoras equation can be used in conjunction with the other Pythagorean triangles in the geometry to solve for the squared edge b. The equations leading to such solution are shown in **(D)**.

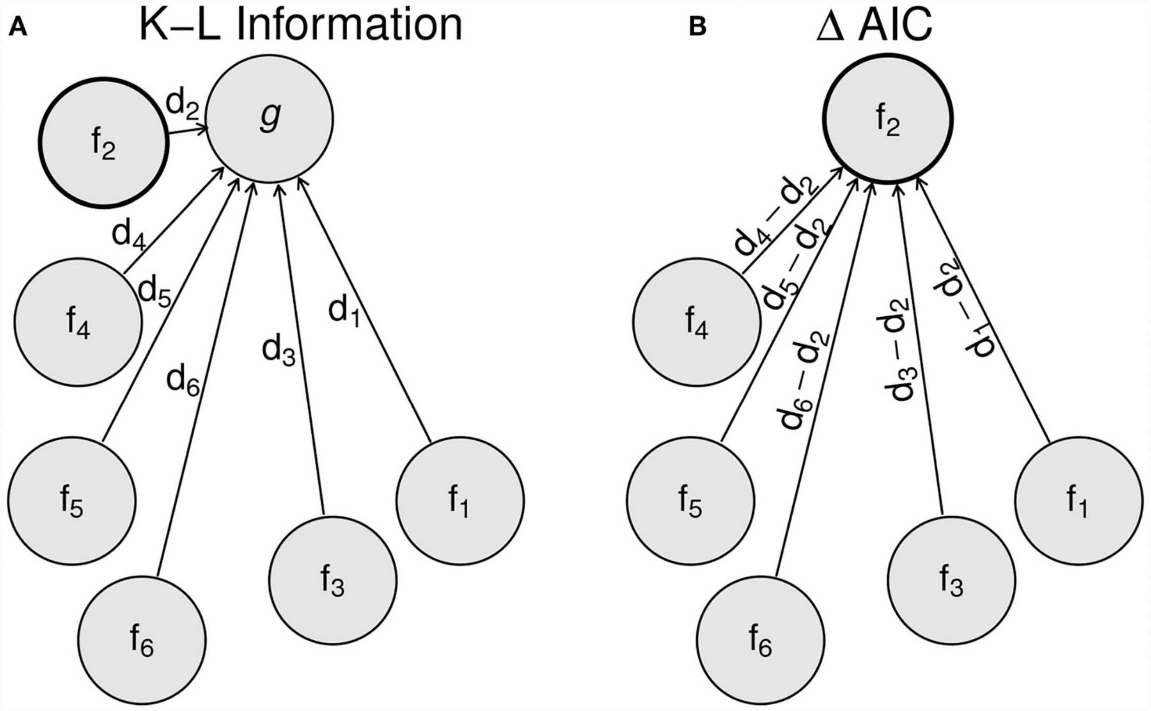

FIGURE 2 |.

Schematic representation of the logic of multi-model selection using the AIC. g represents the generating model and f_i the i^{th} approximating model. The Kullback-Leibler information discrepancies (d_i) are shown on the left **(A)** as the distance between approximating models and the generating model. The ΔAICs shown on the right **(B)** measures the distance from approximating models to the best approximating model. All distances are on the information scale.

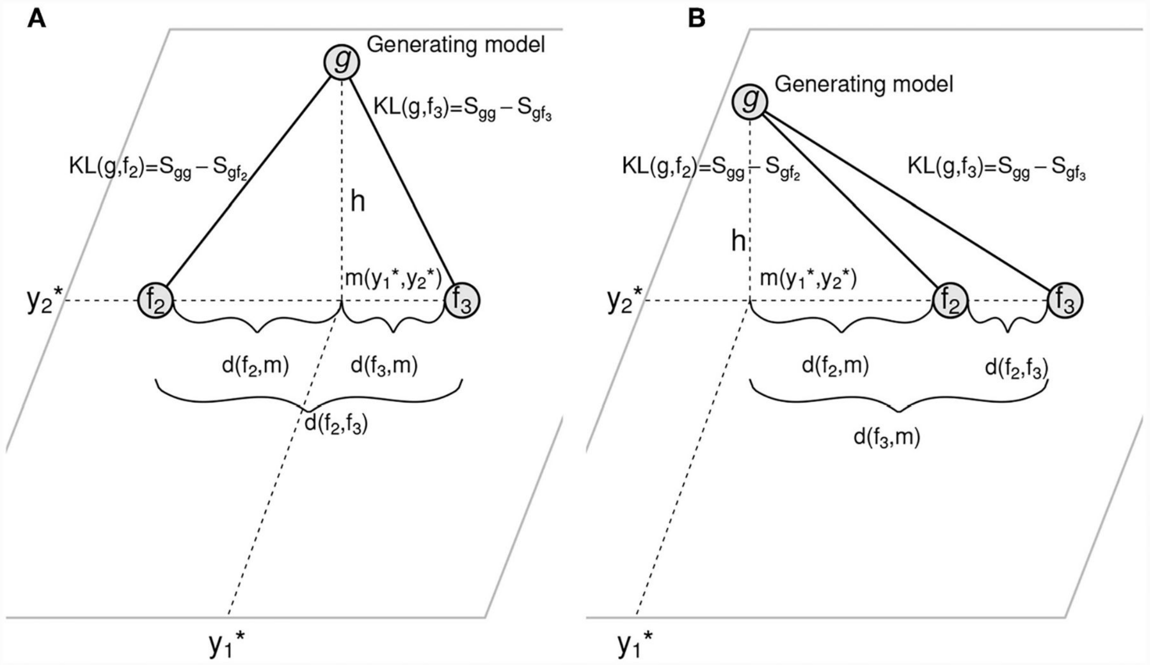

FIGURE 3 |.

The geometry of model space. In this figure, f_2 and f_3 are approximating models residing in a (hyper)plane. g is the generating model. m is the projection of g onto the (hyper)plane. $d(;)$ $\cdot$ are distances between models in the plane. $d(f_2, f_3) \approx KL(f_2, f_3)$ with deviations due to the dimension reduction in NMDS and non-Euclidian behavior of KL divergences. As KL divergences decrease, they become increasingly Euclidian. **(A)** Shows a projection when m is within the convex hull of the approximating models, and **(B)** shows a projection when m is outside of the convex hull. Prasanta S. Bandyopadhyay, Gordon Brittan Jr., Mark L. Taper, Belief, Evidence, and Uncertainty. Problems of Epistemic Inference, published 2016 Springer International Publisher, reproduced with permission of Springer Nature Customer Service Center.

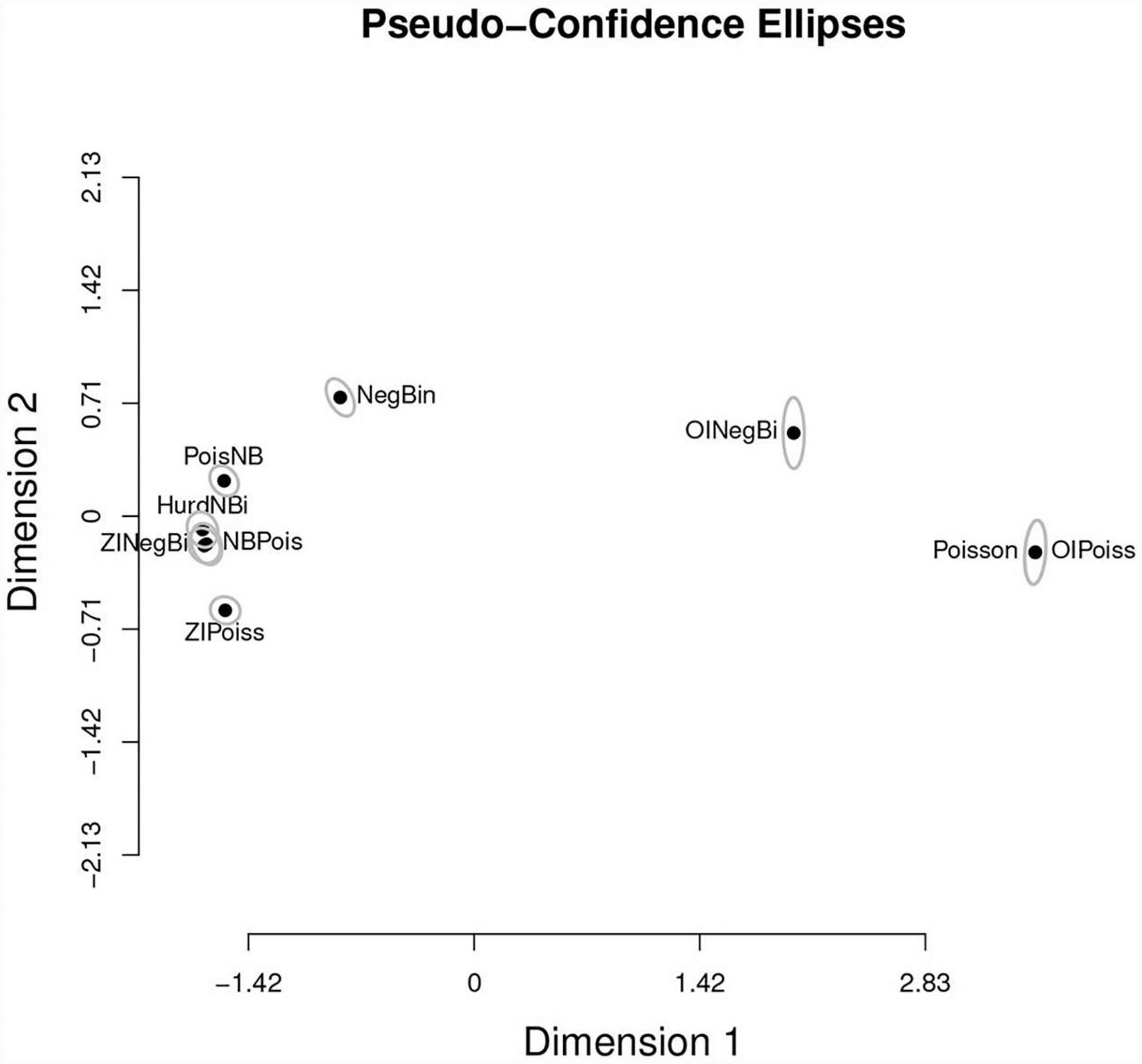

FIGURE 4 |.

Count models for the horseshoe crab example (section 3.1) in NMDS space, along with pseudo-confidence ellipses (95%). These ellipses are based on Stress derivatives (Mair et al., 2019) and indicate in this case that the NMDS is quite stable (overall stress is 0.029).

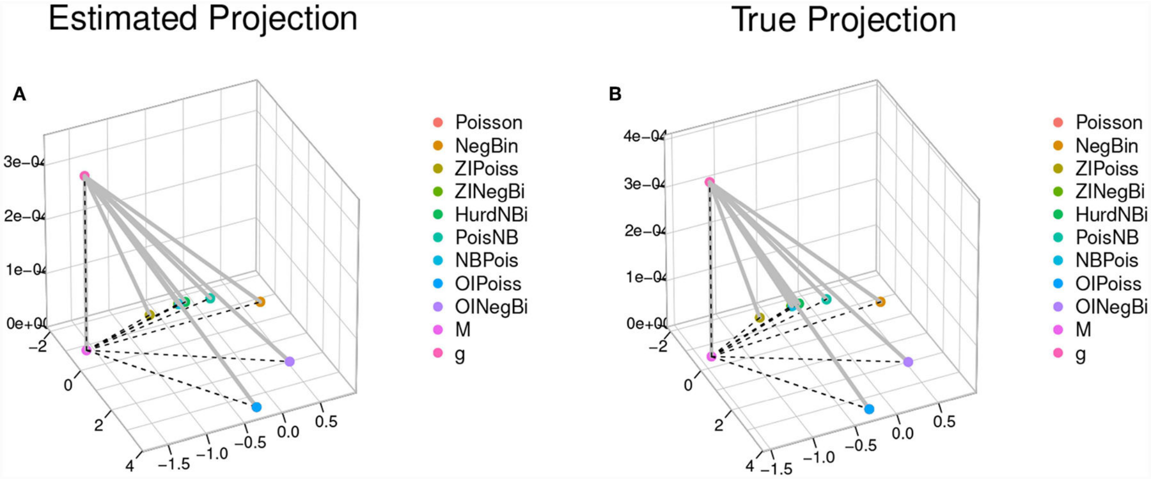

FIGURE 5 |.

The models of Figure 2 visualized by our new methodology, and applied to our Horseshoe crab example (section 3.1). As before, g is the generating model and models $f_1, \ldots, f_9$, are the approximating models and named in the legend of each panel. **(A)** Shows the estimated model projection “M” and the estimated location of the true generating process whereas **(B)** shows the location of the true model projection “M” and of the true generating process. The dashed lines are KL distances between approximating models, which were calculated according to Equation 2. The solid gray lines are the KL distances from approximating models to the generating model. The vertical dotted line shows h, the discrepancy between the generating model and its best approximation in the NMDS plane, whereas all the other dotted lines mark the discrepancy between the approximating models and the model projection “M.” A 2-dimensional representation of only the plane of models, the estimated g model projection and the true model projection of g onto that plane is shown in Figure 6.

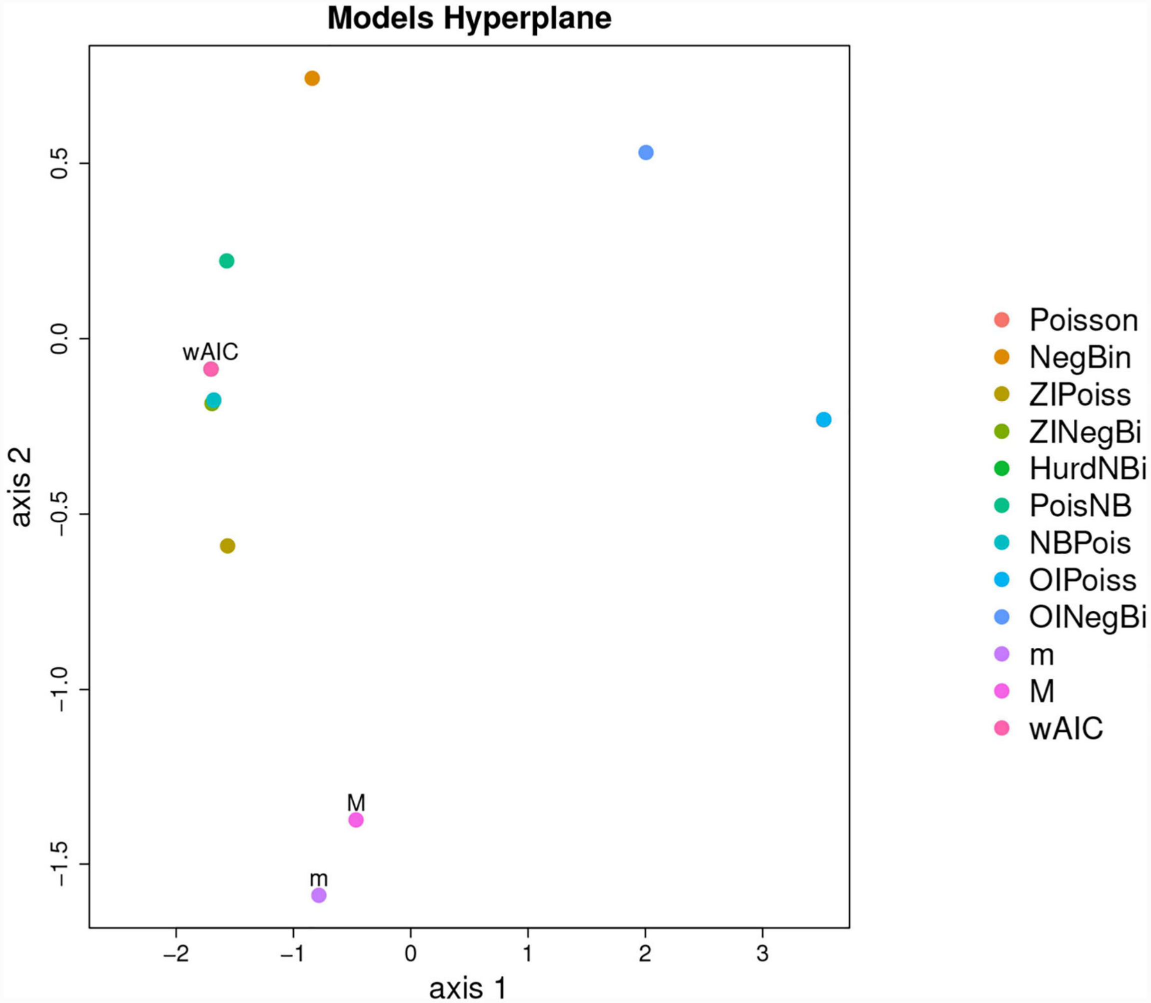

FIGURE 6 |.

NMDS space of nine models for the Horseshoe crab example (section 3.1). The true projection, *M*, of the generating model onto the NMDS plane is shown, along with the location of the estimated location of such projection, *m*, and of the model average, *wAIC*.

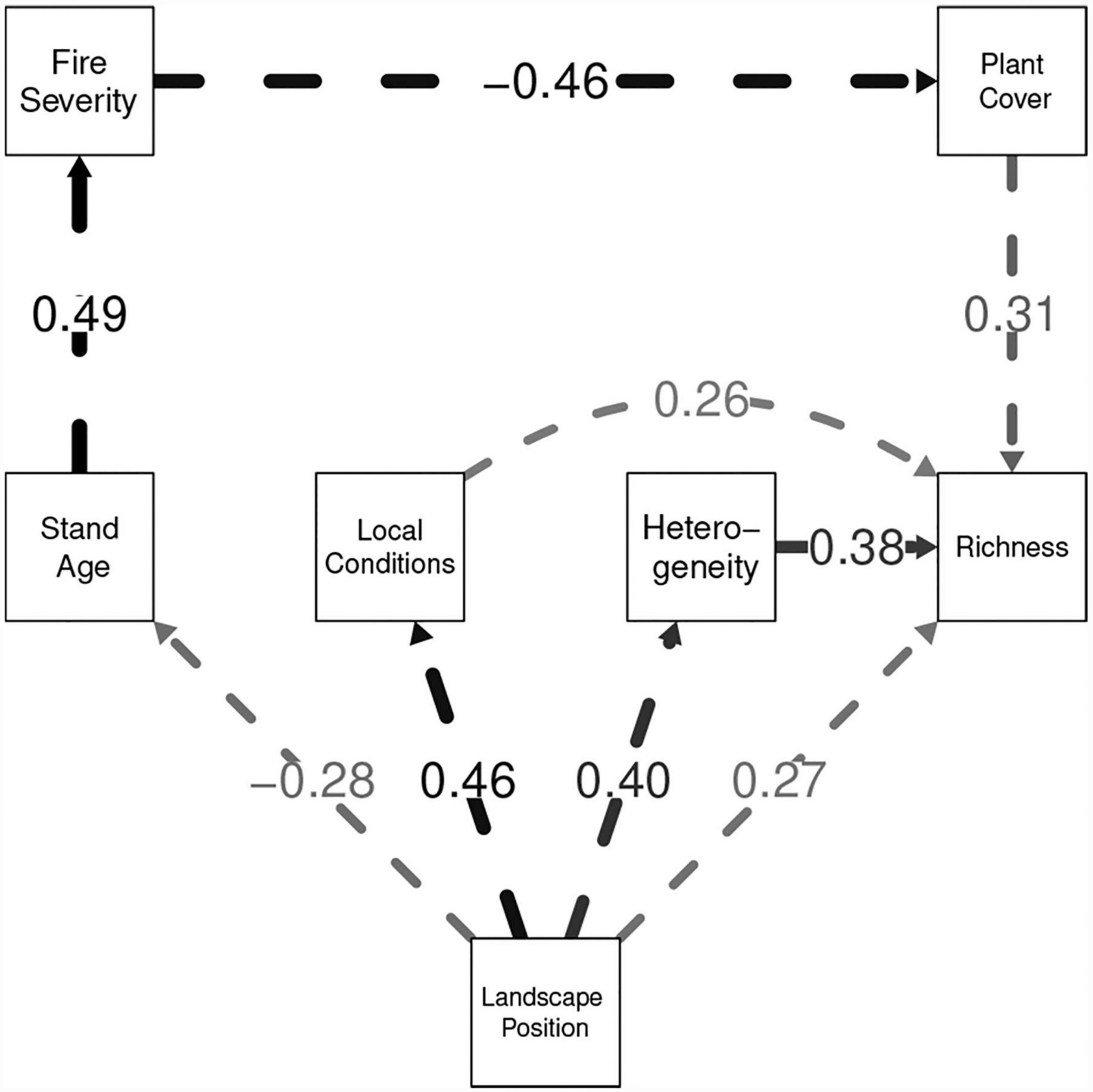

FIGURE 7 |.

The final, simplified model explaining plant diversity from Grace and Keeley (2006). Arrows indicate causal influences. The standardized coefficients are indicated by path labels and widths. See section 3.2 for details. Prasanta S. Bandyopadhyay, Gordon Brittan Jr., Mark L. Taper, Belief, Evidence, and Uncertainty. Problems of Epistemic Inference, published 2016 Springer International Publisher, reproduced with permission of Springer Nature Customer Service Center.

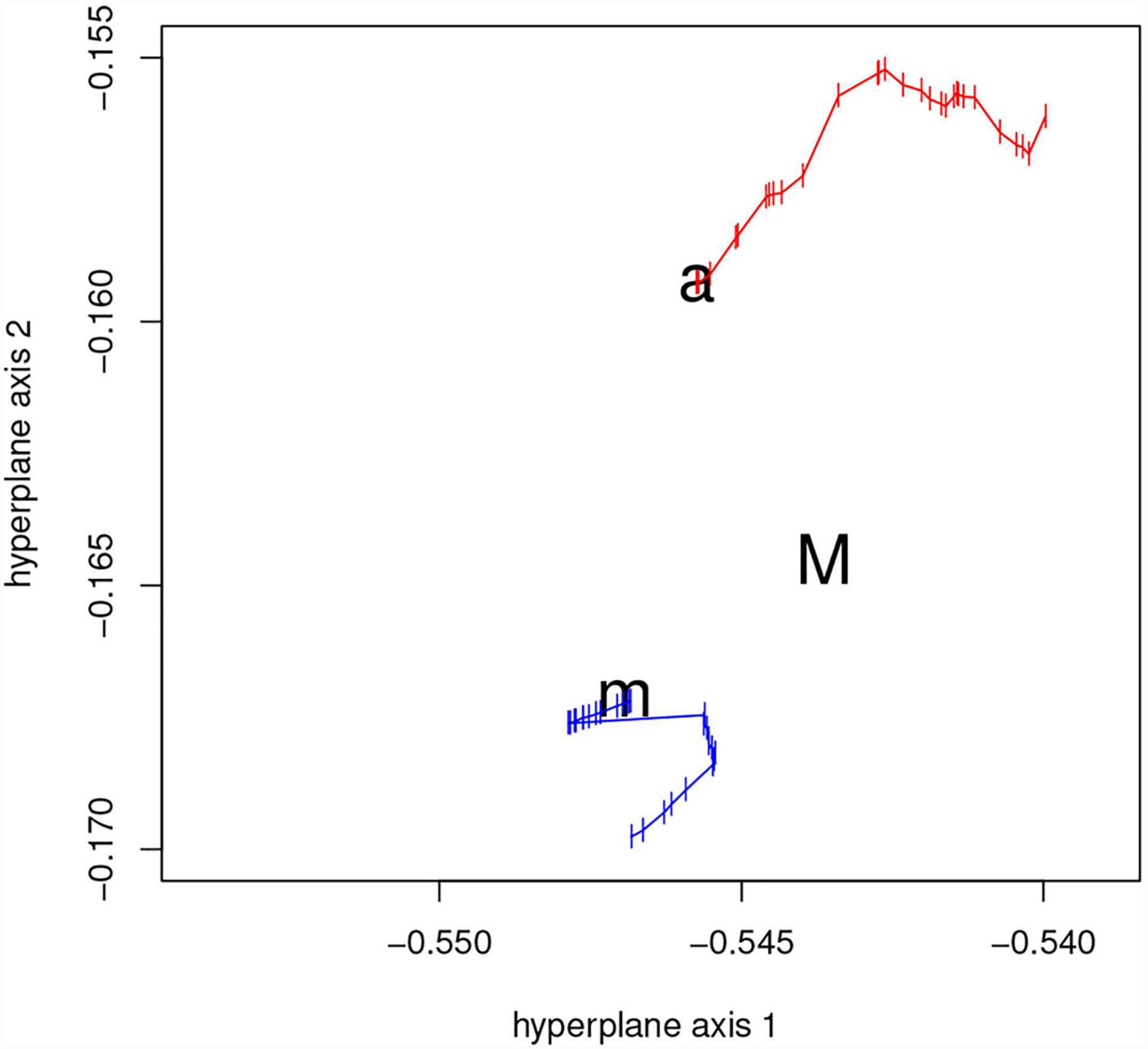

FIGURE 8 |.

Stability test of the displacement (trajectories) of the model prediction (in blue) and the model average (in red) under deletion of 1 – 30 models. *M* denotes the true location of the orthogonal projection of the generating model in the hyperplane. *m* and *a* mark the location of the model projection and the model average, respectively, when the 30 models are used. In both cases, as models are removed one by one from the candidate model set, the location of both *m* and *a* changes (little vertical lines). Note how the model projection estimate is more stable to changes in the model set than the model average. Prasanta S. Bandyopadhyay, Gordon Brittan Jr., Mark L. Taper, Belief, Evidence, and Uncertainty. Problems of Epistemic Inference, published 2016 Springer International Publisher, reproduced with permission of Springer Nature Customer Service Center.

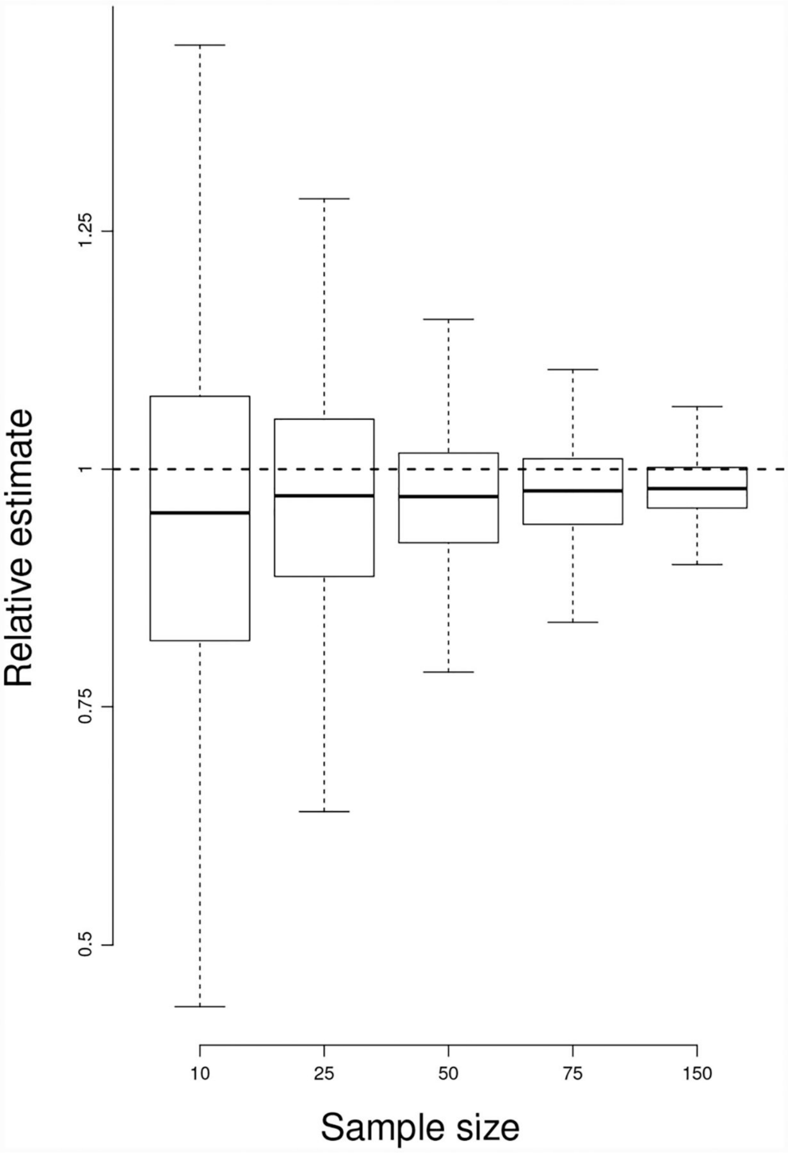

FIGURE 9 |.

Boxplots of sets of 2,000 non-parametric estimates of *Sgg* (from Berrett et al., 2019) relative to the true *Sgg* value of 9.93257, for different sample sizes. The simulated data comes from a seven-dimensional Multivariate Normal distribution with means equal to 10 and the identity matrix as a variance-covariance matrix. The dashed, horizontal line at 1 shows the zero-bias mark.