Appendix

Projections in Model Space: Multi-model Inference Beyond Model Averaging

Mark L. Taper and José M. Ponciano

A reviewer of the manuscript challenged us to do two things. First, to move beyond simple likelihood ratio examples and show how evidential ideas are used in the practice of science. And, second to solve some deep problem in ecology using this framework. To answer the first challenge, we discuss information criteria differences as a natural extension of the likelihood ratio that overcomes many of the complexities of real data analysis. To answer the second more substantive challenge, we then extend the information criterion model comparison framework to much more effectively utilize the information in multiple models, and contrast this approach with model averaging, the currently dominant method of incorporating information from multiple models (Burnham and Anderson 2002). Model averaging is a confirmation-based-approach.

Because of limitations in both time and allowable word count, this will be a sketch of a solution.¹ We deeply appreciate the reviewer’s challenge because the work it has forced us to do has been very rewarding.

Comparing Models Forms that are not Fully Specified

The evidential theory developed so far in this monograph has concerned fully specified models, that is, models with given parameter values. As such, they completely describe the probability distribution for data consistent with those models. Scientists, on the other hand, generally do not have that much foreknowledge. Instead, they commonly propose model forms, that is, models known up to the functional form of their elements, but without specific parameter values. Parameter values are estimated by fitting the model forms to data, and comparisons of these now fully-specified models made using the same data. As both estimated models do have likelihoods, it would seem straightforward to compare them with likelihood ratios, the measure which has been invoked in the body of this monograph to capture the evidential strength for one model against its rival.

¹Technical statistical details will follow in other fora, beginning with a symposium at the Japanese Institute of Mathematical Statistical Mathematics in January of 2016.
Unfortunately, important problems do arise. The estimation of parameters creates a bias in the calculation of the probability of the data under the model and therefore it also creates a bias in the estimated likelihood of the model. This bias is related to the number of parameters in the model; the more parameters a model has, the greater the (positive) bias there is in the estimated likelihood. Carried to the extreme, if a model has as many parameters as there are data points, it can fit the data exactly, and its likelihood will become infinite. This despite having no predictive power.

Classically this problem has been handled (at least in the context of nested models) with the use of likelihood ratio tests. But, what about non-nested models? Constraining your models to be nested imposes a severe limitation on the ability of scientists to explore nature through models. Further, the likelihood ratio test is an error-statistical approach where the bias due to parameterization is accounted for in the critical value of the test. This book (Chap. 6) and Taper and Ponciano (2015) argue for the greater utility of an evidential approach compared to an error-statistical approach.

In 1973, Hirotogu Akaike wrote a brilliant paper developing the AIC, and thereby solving, in large part, the estimation bias problem of the likelihood ratio, allowing for a huge expansion in the scope of evidential analysis. Although Akaike referred to the AIC as “an Information Criterion,” the AIC is universally termed the “Akaike Information Criterion.”

**Akaike with Tears**

Technical accounts deriving Akaike’s Information Criterion (AIC) exist in the literature (see for instance the general derivation of Burnham and Anderson 2002, Chap. 7), but few have attempted to clarify Akaike’s (1973) paper, step by step. A notable exception is deLeeuw’s (1992) introduction to Akaike (1973) Information Theory, which made it clear that more than a technical mathematical statistics paper, Akaike’s seminal contribution was a paper about ideas: “…This is an ‘ideas’ paper, promoting a new approach to statistics, not a mathematics paper concerned with the detailed properties of a particular technique…” deLeeuw then takes on the task of expunging the ideas from the technical probabilistic details and coming up with a unified account clarifying both the math and the ideas involved. His account is important because it makes evident that at the very heart of the derivation Akaike was using Pythagoras’ theorem. It will be seen later that our contribution is to take this derivation one step further by using Pythagoras’ theorem

If this is their initial encounter with information criteria, we suggest that readers first familiarize themselves with a gentler introduction such as Malcolm Forster and Elliott Sober’s “How to Tell When Simple, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions,” which explains, develops, and applies Akaike’s ideas to central problems in the philosophy of science. *British Journal for the Philosophy of Science* 45: 1–35, 1994. A comparison between different standard information criteria can be found in Bandyopadhyay and Brittan (2002).
again. In what follows we will set the stage to explain our contribution using Akaike (1973, 1974) and deLeeuw (1992). Given this monograph’s context, our account will focus more on the ideas than on the technical, measure theoretic details for the sake of readability and also because this approach will allow us to shift directly to the core of our contribution.

A central objective in scientific practice is trying to come up with some measure of comparison between an approximating model and the generating model. Following Akaike, we shall be concerned for the time being with the parametric situation where the probability densities are specified by a set of parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_L)'$ in the form $f(x; \theta)$. The true, generating model will be specified by setting $\theta = \theta_0$ in the density $f$. Setting aside the fact that truth is unknown, under this setting the comparison between a general model and the true model can be done, as in the rest of the monograph, via the likelihood ratio $s(x, \theta; \theta_0) = \frac{f(x; \theta)}{f(x; \theta_0)}$ without loss of efficiency. This well-known statistical fact suggests using some discrimination function $U(s(x, \theta; \theta_0))$ of the likelihood ratio between $\theta$ and the true model $\theta_0$. The data, $x$, are random and so the average discrimination over all possible data would better represent the distance between a model and the truth. Such an average would then be written as

$$D(\theta, \theta_0; \Phi) = \int f(x; \theta_0) \Phi(s(x, \theta; \theta_0)) dx = \mathbb{E}_x[\Phi(s(x, \theta; \theta_0))]$$

where the expectation is over the sampled stochastic process of interest $X$. Akaike then suggested study of the sensitivity of this quantity to the deviation of $\theta$ from $\theta_0$. Two questions of interest immediately arise: can such an average discrimination be minimized and if so, can its minimization be estimated from realized observations of the process?

To get at this quantity, Akaike thought of expanding it via a Taylor series around $\theta_0$ and keeping a second order approximation, which we write here for a univariate $\theta$:

$$D(\theta, \theta_0; \Phi) \approx D(\theta_0, \theta_0; \Phi) + (\theta - \theta_0) \frac{\partial D(\theta, \theta_0; \Phi)}{\partial \theta} \bigg|_{\theta = \theta_0} + \frac{(\theta - \theta_0)^2}{2!} \frac{\partial^2 D(\theta, \theta_0; \Phi)}{\partial \theta^2} \bigg|_{\theta = \theta_0} + \cdots$$

To write this approximation explicitly, note that $s(x, \theta; \theta_0)|_{\theta = \theta_0} = 1$. Also, note that since $f$ is a probability density $\int f(x; \theta) dx = 1$, which together with the regularity conditions that allow differentiation under the integral sign results in

$$\int \frac{\partial f(x; \theta)}{\partial \theta} dx = \int \frac{\partial^2 f(x; \theta)}{\partial \theta^2} dx = 0.$$
\[
\frac{\partial^2 \mathcal{D}(\theta, \theta_0; \Phi)}{\partial \theta^2} \bigg|_{\theta = \theta_0} = \int \frac{\partial}{\partial \theta} \left( \frac{\partial \Phi(\tau)}{\partial \tau} \right) f(x; \theta_0) \, dx \bigg|_{\theta = \theta_0} \\
= \int \frac{\partial^2 \Phi(\tau)}{\partial \tau^2} \left( \frac{\partial \tau}{\partial \theta} \right)^2 f(x; \theta_0) \, dx \bigg|_{\theta = \theta_0} + \int \frac{\partial^2 \tau}{\partial \theta^2} \frac{\partial \Phi(\tau)}{\partial \tau} f(x; \theta_0) \, dx \bigg|_{\theta = \theta_0} \\
= \Phi''(1) \int \left( \frac{1}{f(x; \theta_0)} \frac{\partial f(x; \theta)}{\partial \theta} \right)^2 f(x; \theta_0) \, dx \bigg|_{\theta = \theta_0} \\
= \Phi''(1) I(\theta_0),
\]
where \( I(\theta_0) \) is Fisher’s information. In going from the first line to the second line, a combination of the product rule and chain rule is employed. In going from the second line to the third, it is noted that the results given immediately above indicate that the right hand integral is 0. This preliminary result is non-trivial because it demonstrates that the resulting approximation of the average discrimination function

\[
\mathcal{D}(\theta, \theta_0) \approx \Phi(1) + \frac{1}{2} \Phi''(1)(\theta - \theta_0)^2 I(\theta_0)
\]

is directly scaled by the theoretical variance of the Maximum Likelihood Estimator (MLE). Next, Akaike proposed using the functional form \( \Phi(t) = -2 \log(t) \) so that \( \mathcal{D} \) behaves like a distance, i.e., is always non-negative and \( \mathcal{D}(\theta_0, \theta_0) = \Phi(1) = 0 \). The factor of 2 is conventional. Conveniently, the approximation, from here on denoted as \( \mathcal{W}(\theta, \theta_0) \) then becomes \( \mathcal{D}(\theta, \theta_0) \approx \mathcal{W}(\theta, \theta_0) = (\theta - \theta_0)^2 I(\theta_0) \). It is straightforward to show that in the multivariate case, the approximation is written as the quadratic form \( \mathcal{W}(\theta, \theta_0) = (\theta - \theta_0)^T \mathcal{I}(\theta_0)(\theta - \theta_0) \), where \( \mathcal{I}(\theta_0) \) is Fisher’s information matrix. On the other hand, inserting Akaike’s choice of a functional form into the original definition of the average discrepancy gives

\[
\mathcal{D}(\theta, \theta_0) = -2 \int f(x; \theta_0) \log \left( \frac{f(x; \theta)}{f(x; \theta_0)} \right) \, dx \\
= -2 \mathbb{E}_x \left[ \log \frac{f(X; \theta)}{f(X; \theta_0)} \right] \\
= -2 \mathbb{E}_x [\log f(X; \theta) - \mathbb{E}_x (\log f(X; \theta_0))] \\
= 2 \mathbb{E}_x (\log f(X; \theta_0)) - 2 \mathbb{E}_x (\log f(X; \theta)).
\]

This form of the average discrimination function is known as the negentropy, or the Kullback–Leibler (KL) divergence. So from the start, Akaike was able to make two crucial connections between his choice measure of discrepancy between the true generating model and an approximating model. One, directly bringing the theory of ML estimation into the scaling of such discrepancy, and the other, linking these
concepts with a wealth of results in Information Theory. Thus, it was natural for Akaike to call $D(\hat{\theta}, \theta_0)$ the probabilistic entropy. If $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ observations from the process $X$ are available, then using the law of large numbers the KL divergence (or probabilistic entropy) could be estimated consistently with the average likelihood ratio

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

where $\hat{\theta}$ is the MLE. In reality, one cannot compute this likelihood ratio because the true model $\theta_0$ in the denominator is unknown. However, because for every data point $x_i$ the denominator in this average log-likelihood ratio is the same constant, Akaike pointed out that even if truth is unknown, we do know that maximizing the (log-) likelihood also minimizes the KL divergence between the estimated density and the true density. This is why Akaike called his contribution an “extension of the principle of maximum likelihood”. Not content with this result, and in a remarkable display of the reaches of frequentist thinking, Akaike pointed out that because multiple realizations of the array of data points $X_1, X_2, \ldots, X_n$ yield multiple estimates of $\theta_0$, one should in fact think of the average discrepancy as a random variable, where the randomness is with respect to the probability distribution of the MLE $\hat{\theta}$. Therefore one may in fact be able to minimize the KL divergence between the true generating model and the approximating model by minimizing the average of $D(\hat{\theta}, \theta_0)$—averaged over the distribution of $\hat{\theta}$. The problem of minimization of the KL divergence then becomes a problem of approximation of an average, something that statisticians are (supposedly) good at. Let $R(\theta_0) = \mathbb{E}_\theta[D(\hat{\theta}, \theta_0)]$ denote our target average. Substituting the probabilistic entropy by its definition using the expectations over the process we get

$$R(\theta_0) = \mathbb{E}_\theta D(\hat{\theta}, \theta_0) = 2\mathbb{E}_\hat{\theta} \left[ \mathbb{E}_X \left( \log f(X; \theta_0) \right) - \mathbb{E}_X \left( \log f(X; \hat{\theta})|\hat{\theta} \right) \right]$$

$$= 2\mathbb{E}_X \left( \log f(X; \theta_0) \right) - 2\mathbb{E}_\hat{\theta} \left[ \mathbb{E}_X \left( \log f(X; \hat{\theta}) \right) \right].$$

The first term in this expression is an unknown constant whereas the second term is a double expectation. Instead of working directly with these expectations, Akaike thought of substituting for the probabilistic entropy $D(\hat{\theta}, \theta_0)$ by its quadratic approximation $W(\hat{\theta}, \theta_0)$ via a Taylor series expansion and a very creative and useful way to re-write this expression. Akaike noted that the quadratic form

$$W(\theta, \theta_0) = (\hat{\theta} - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta} - \theta_0)$$

used to approximate $\hat{D}_n(\hat{\theta}, \theta_0)$ can be seen (as any quadratic form involving a positive definite matrix and a fixed point) as the squared statistical distance between $\hat{\theta}$ and $\theta_0$. This is the square of a statistical distance because proximity between
points is weighted by the dispersion of the points in the multivariate space, which is in turn proportional to the eigenvalues of the positive definite matrix $I(\theta_0)$ (see plots explaining the geometric interpretation of quadratic forms in Johnson and Wichern 2002, Chaps. 1 and 2). Expressing the average discrepancy as the squared of a distance was a crucial step in Akaike’s derivation because it opened the door for its decomposition using Pythagoras’ theorem. By doing such decomposition, one can immediately visualize through a simple sketch the ideas in his proof (see Fig. 1). The first vertex of Akaike’s right-angle’s triangle is the truth $\theta_0$ of dimension $L$ (unknown). The second vertex is the estimator $\hat{\theta}$ of dimension $k \leq L$, as it comes from an approximating model. We will denote this estimator as $\hat{\theta}_k$ to emphasize its smaller dimension. The third vertex is $\theta_{0k}$ which is the orthogonal projection of the truth in the plane where all estimators of dimension $k$ lie, that we will denote $\Theta_k$. A fourth point crucial to derive the AIC is given by the estimator of $\theta_0$ from the data using an approximating model with the same dimension as $\theta_0$. To distinguish it from $\hat{\theta}_k$ we will denote it as $\hat{\theta}_0$. This estimator can be thought of lying in the same plane as $\theta_0$. In the sketch in Fig. 1, we have labeled all the edges with a lowercase letter. To make the derivation as simple as possible, we will do the algebraic manipulations with these letters. In so doing, we run the unavoidable risk of trivializing one of the greatest findings of modern statistical science, all for the sake of transmitting the main idea behind the proof. The reader, however, should be well aware that these edges (lower case letters) denote, by necessity, random
variables and that in the real derivation, more complex arguments including limits in probability and fundamental probability facts are needed.

Given the law of large numbers approximation of the average discrepancy using the average log-likelihood ratio, Akaike’s initial idea was to use, as estimate of $W(\hat{\theta}, \theta_0)$, $\hat{D}_n(\hat{\theta}_k, \hat{\theta}_0) = -2 \times \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i; \hat{\theta}_k)}{f(x_i; \theta_0)}$ (=$e^2$ in the sketch below) and let $n \to \infty$. However, the Pythagorean decomposition illustrated below shows that the estimated discrepancy $\hat{D}_n(\hat{\theta}_k, \hat{\theta}_0)$ will be a biased estimate of the target discrepancy because of the substitution of the ML estimators. One point in favor of the usage of $\hat{D}_n(\hat{\theta}_k, \hat{\theta}_0)$, however, is that ML theory tells us that $n\hat{D}_n(\hat{\theta}_k, \hat{\theta}_0)$ is chi-square distributed with degrees of freedom $L - k$. With this result at hand, and using simple geometry, Akaike sought to re-write the Pythagorean decomposition using $\hat{D}_n(\hat{\theta}_k, \hat{\theta}_0)$. The last piece of the puzzle needed to be able to do that was to demonstrate via convergence in probability calculations that the edge $a = (\hat{\theta}_k, \theta_0)$ was the stochastic projection of the edge $c = (\hat{\theta}_0, \theta_0)$ in the $\Theta_k$ plane. Below is the sketch aforementioned:

In simple terms, the objective is to solve for the edge length $b$ using what we can estimate ($e^2$ through the log-likelihood ratio $n\hat{D}_n(\hat{\theta}_k, \hat{\theta}_0)$). Using Pythagoras’ theorem we get that

$$b^2 = h^2 + a^2.$$  

Note also that $d^2 = e^2 + a^2$ so that

$$e^2 = d^2 - a^2.$$

However, arguing that the third term of the squared distance $d^2 = c^2 + h^2 - 2c.h \cos \phi$ remained insignificant compared with the other squared terms, Akaike re-wrote it as $d^2 \approx c^2 + h^2$, which upon substituting into (2) gives

$$e^2 = h^2 + c^2 - a^2.$$  

Now, doing (1–3) gives $b^2 - e^2 = h^2 + a^2 - h^2 - c^2 + a^2$. Hence, it follows that

$$b^2 = e^2 + 2a^2 - c^2.$$  

Expressing the square distances in Eqs. (1–3), expanding them using Taylor Series expansions, estimating Fisher’s Information in each case with the observed information and using convergence in probability results, Akaike was able to show that

$$nc^2 - na^2 \sim \chi^2_{L-k}.$$
and that
\[ na^2 \sim \chi_k^2, \]
so that Eq. (4) multiplied by \( n \) can be re-written as
\[ nb^2 = nW(\hat{\theta}_k, \theta_0) \approx \frac{nD_n(\hat{\theta}_k, \hat{\theta}_0)}{\text{log-likelihood ratio}} + \frac{na^2}{\sim \chi_k^2} - n(c^2 - a^2). \]

The double expectation from the original average discrepancy definition is then implemented by simply replacing the chi-squares by their expectations, which immediately gives
\[ nE_{\hat{\theta}_k}[W(\hat{\theta}_k, \theta_0)] \approx nD_n(\hat{\theta}_k, \hat{\theta}_0) + 2k - L, \text{ or} \]
\[ E_{\hat{\theta}_k}[W(\hat{\theta}_k, \theta_0)] \approx -\frac{2}{n} \sum_{i=1}^{n} \log f(x_i; \hat{\theta}_k) + \frac{2k}{n} - \frac{L}{n} + \frac{2}{n} \sum_{i=1}^{n} \log f(x_i; \hat{\theta}_0). \] (5)

Recall that what Eq. (5) is approximating is in fact
\[ R(\theta_0) = E_{\hat{\theta}_k}D(\hat{\theta}_k, \theta_0) = -2E_{\hat{\theta}} \left[ E_X \left( \log f(X; \hat{\theta}_k) \mid \hat{\theta}_k \right) \right] + 2E_X(\log f(X; \theta_0)), \] (6)
which is the expected value (with respect to \( \hat{\theta}_k \)) of
\[ -2 \int f(x; \theta_0) \frac{\log f(x; \hat{\theta}_k)}{f(x; \theta_0)} dx = -2 \int f(x; \theta_0) \log f(x; \hat{\theta}_k) dx + 2 \int f(x; \theta_0) \log f(x; \theta) dx. \] (7)

Using Eq. (5) where the first two terms are known and the next two terms include the unknown dimension, and the law of large numbers approximation of the first integral, Akaike concluded that an unbiased estimation of the expected value over the distribution of \( \hat{\theta}_k \) of the first integral would be given by the average of the first two terms in Eq. (5).

The first term in Eq. (5) is \((-2/n)\) times the log likelihood with the approximating model. The last two terms cannot be known, but because upon comparing various models they will remain the same can be ignored in practice. Because \( n \) also remains the same across models, in order to compare an array of models one only has to compute \( AIC = -2 \sum_{i=1}^{n} \log f(x_i; \hat{\theta}_k) + 2k \) and choose the model with the lowest score as the one with the smallest discrepancy to the generating model. The logic can be graphically represented by Fig. 2 (drawn from Burnham et al. 2011).

In the popular literature (e.g. Burnham and Anderson 2002, p. 61, or Burnham et al. 2011) it is often asserted that the \(-AIC/2\) is an estimator of \( E_{\hat{\theta}} \left[ E_X \left( \log f(X; \hat{\theta}_k) \mid \hat{\theta}_k \right) \right] \).
It is not, as Akaike (1974) states, the estimator of this quantity is \(-\frac{\text{AIC}}{2n}\). For the qualitative comparison of models, this distinction makes no difference, but factoring the sample size \((n)\) into the AIC allows a comparer of models to assess not only which model appears best, but what is the strength of evidence for that statement.

The Problem of Multiple Models

A model-centric view of science coupled with a disavowal of the absolute truth of any model pushes the scientist to the use of many models. Once this stance is taken, the question of how to use multiple models in inference naturally arises. Inference by the best model is not adequate as many models may be indistinguishable on the basis of pairwise comparison (see Chap. 2).

Currently, the dominant method for incorporating information from multiple models is model averaging. This comes in several flavors. In all cases model averaging is inherently, and generally explicitly, a Bayesian approach. Most common in ecology is averaging model parameter estimates or model predictions using Akaike weights. The Akaike weight for the \(i\)th model is given as:

\[
W_i = \frac{e^{-\Delta_i/2}}{\sum_{r=1}^{R} e^{-\Delta_r/2}},
\]

where \(\Delta_i\) is the difference between a model’s AIC value and the lowest AIC value from the model set of \(R\) models indexed by \(r\). Although it is not always pointed out, \(w_i\) is a posterior probability based on subjective priors of the form.
where \( q_i \) is the prior for model \( i \), \( C \) is a normalization constant, \( k_i \) is the number of parameters in the model, and \( n \) is the number of observations. The use of this prior makes model averaging a confirmation approach.

Two difficulties with model averaging for an evidentialist are: (1) the weights are based on beliefs, and are thus counter to an evidential approach. And (2) as a practical matter, model averaging does not take into account model redundancy. The more effort put into building models in a region of model space, the more heavily that region gets weighted in the average. We propose the alternative of estimating the properties of the best projection of truth, or a generating model, to the hyper-plane containing the model set. This mathematical development extends Akaike’s insight by using the known KL distances among models as a scaffolding to aid in the estimation of the location of the generating model.

For convenience, we follow Akaike’s (1974) notation and denote \( S_{gf} = \int f(x; \theta_0) \log f(x; \hat{\theta}_k) dx \) and \( S_{gg} = \int f(x; \theta_0) \log f(x; \theta) dx \), where the \( g \) refers to the ‘generating’ model and the \( f \) to the approximating model. Akaike’s observation is then written as:

\[
\tilde{S}_{gf} = \frac{1}{n} \sum_{i=1}^{n} \log f(x_i; \hat{\theta}_k) - \frac{k}{n} = -\frac{\text{AIC}}{2n}. 
\]

Accordingly, the KL divergence between a generating model \( g \) and an approximating model \( f \) can simply be written as \( KL(g, f) = S_{gg} - S_{gf} \). From now on we will stick to this short-hand notation. One last detail that we have not mentioned so far is the fact that Akaike’s approximation works provided \( \hat{\theta}_k \) is close to \( \theta_0 \) for any \( k \).

In fact, this is precisely why the Pythagorean decomposition works. The staggering and successful use of the AIC in the scientific literature shows that such approximation is in many cases reliable. Under the same considerations, we now extend these ideas to the case where we want to draw inferences from the spatial configuration of \( f_1, f_2, \ldots \) approximating models to the generating model \( g \).

The fundamental idea of our contribution is to use the architecture of model space to try to estimate the projection of truth onto a (hyper)plane where all the approximating models lie. Having estimated the location of truth, even without estimating it per se would anchor the AIC statistics in a measure of overall goodness of fit, as well as provide invaluable insights into the appropriateness of model averaging. The intuition of the feasibility of such a task comes from the realization that approximating models have similarities and dissimilarities. A modeler is drawn naturally to speak of the space of models. All that remains is to realize that that language is not metaphor, but fact. KL divergences can be calculated between any distributions and are not restricted to between generating processes and approximating models. A set of models has an internal geometrical
relationship which constrains it and therefore has information about the relationship of approximating models and the generating process.

Computational advances have rendered straightforward algorithmic steps that while conceptually feasible would have been computationally intractable at the time that Akaike was developing the AIC. First, it is now easy to calculate the KL divergence between any two models. For instance, for the Normal distribution, the KL discrepancy can be computed exactly using the package gaussDiff in the statistical software R. Other packages will estimate the KL divergences of arbitrary distributions. Thus for a large set of approximating models, a matrix of estimated KL divergences among the set of models can be constructed. Second, parallel processing has tamed the computer intensive Non-Metric Multidimensional (NMDS) scaling algorithm which can take an estimated matrix of KL divergences and estimate the best Euclidean representation of model space in a (hyper)plane with coordinates \((y_1, y_2, \ldots)\). Nothing in our development restricts model space to be restricted to \(\mathbb{R}^2\). To emphasize this we speak of a (hyper)plane, but to have any hope of visualizing we stay in \(\mathbb{R}^2\) for this paper.

Suppose then that one can place the approximating models \(f_1, f_2, \ldots\), on a Euclidean plane, as in the sketch below. For simplicity we have placed only two models in the sketch. Our derivation is not constrained to their particular configuration in the plane, relative to the generating model (truth), as the Fig. 3a, b show. Define \(m\) with coordinates \((y_{1*}, y_{2*})\) as the projection of the generating model (truth) in the Euclidean plane of models. This projection is separated by the length \(h\) to the generating model. Define \(d(f_i, m)\) as the distance in the hyperplane of model \(i\) from \(m\). Of course, the edges and nodes in this plane are random variables, associated with a sampling error. But, for the sake of simplicity and just as we did above to explain Akaike’s derivation of the AIC, we conceive them for the time being as simple fixed nodes and edges.

Fig. 3 The geometry of model space. \(f_2\) 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. \(d(\ldots)\) 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. Panel a shows a projection when \(m\) is within the convex hull of the approximating models, and Panel b shows a projection when \(m\) is outside of the convex hull.
Then, using Pythagoras and thinking of the KL divergences as squared distances, the following equations have to hold simultaneously:

\[
\begin{aligned}
KL(g,f_1) &= d(f_1, m)^2 + h_1^2 \\
KL(g,f_2) &= d(f_2, m)^2 + h_2^2 \\
&\vdots
\end{aligned}
\]

where necessarily \( h_1 = h_2 = h_i = \cdots = h \). In practice, one can decompose the KL divergence into an estimable component, \( S_{gf_i} \), and a fixed unknown component \( S_{gg} \). Given that the \( S_{gf_i} \) are estimable as in Eq. (8), one can re-write the above system of equations including the unknown constants \( S_{gg}, y_1^*, y_2^* \) as follows:

\[
\begin{aligned}
S_{gg} - S_{gf_1} - d(f_1, m(y_1^*, y_2^*))^2 &= h_1^2, \\
S_{gg} - S_{gf_2} - d(f_2, m(y_1^*, y_2^*))^2 &= h_2^2, \\
&\vdots
\end{aligned}
\]  

(9)

Then, operationally, in order to estimate the location of the orthogonal projection of the generating model in the plane of approximating models, one can easily program the system of Eq. (9) into an objective function that, for a given set of values of the unknown parameters \( S_{gg}, y_1^*, y_2^* \), computes the left hand sides of Eq. (9) and returns the sum of the squared differences between all the \( h_i^2 \). Then, a simple minimization of this sum of squared differences leads to an optimization of the unknown quantities (Fig. 4).

**Fig. 4** The models of Fig. 2 visualized by our new methodology. As before, \( g \) is the generating model and \{\( f_1, \ldots, f_5 \)\}, are the approximating models. The dashed lines are KL distances between approximating models, which can be calculated. The solid black lines are the KL distances from approximating models to the generating model, which now can be estimated. The model labeled \( m \) is the projection of the generating model to the plane of the approximating models. The solid gray line shows \( h \), the discrepancy between the generating model and its best approximation in the NMDS plane.
We demonstrate this approach with a simulation based on the published ecological work of Grace and Keely (2006). Analyzing community composition data at 90 sites over 5 years, they studied the generation of plant community diversity after wildfire using structural equation models. Structural equation modeling is a powerful suite of methods facilitating the incorporation of causal hypotheses and general theoretical constructs directly into a formal statistical analysis (Grace and Bollen 2006, 2008; Grace 2008; Grace et al. 2010). The final model that Grace and Keely arrived at is shown in Fig. 5.

The figure should be read to mean that species richness is directly influenced by heterogeneity, local abiotic conditions, and plant cover. Heterogeneity and local abiotic conditions are themselves both directly influenced by landscape position, while plant cover is influenced by fire severity, which is influenced by stand age, which is itself influenced by landscape position. Numbers on the arrows are path coefficients and represent the strength of influence.

Our purpose in presenting this model is not to critique it or the model identification process by which it was found, but to use it as a reasonably realistic biological scenario from which to simulate. In short, we play god using this as a known true generating process. We consider in this analysis 41 models of varying complexity fitted to the simulated data. They cover a spectrum from underfitted to overfitted.
We calculated the NMDS 2-dimensional model space as described above. The stress for this NMDS is extremely low (0.006 %) indicating the model space fits almost perfectly into an $\mathbb{R}^2$ plane. We have plotted the fitted models in this space, grey-scale coded by the AIC categories. We have also plotted in Fig. 6 the location of our methods estimated projection of the generating model to the NMDS plane, the model averaged location using Akaike weights, and the true projection of the generating model to the NMDS plane (we know this because we are acting as God). We can see in Fig. 6 that the estimated projection is slightly closer to the true projection than is the model-averaged location.

In Fig. 7 we plot the effect on the estimated projection and model average of deleting models from consideration. We sequentially delete the left-most model remaining in the set, recalculating locations with each deletion. We see that the model-averaged location shifts systematically rightward with deletion, and that the location of the estimated projection is in this example more stable than the model averaged location. It remains in the vicinity of its original estimate even after all models in the vicinity of that location have been removed from consideration. If we delete from the right, the model average moves systematically leftward. The model projection location is, in this sequence, less stable than under deletion from the left. These deletion exercises highlight several interesting facts about the two types of location estimates that are implicit in the mathematics, but easily overlooked. First, the model average is constrained to lie within the convex hull\(^3\) of the approximating model set. If you shift the model set, you will shift the average. Second, the

---

\(^3\)The convex hull of a set of points in a plane is easily visualized as the space contained by a rubber band placed around all of the points.
estimated generating model projection as a projection can lie outside of the convex hull. Third, because of the geometrical nature of the projection estimate, distant models can contribute information to the location of the best projection. This is the difference between rightward and leftward deletion. There are several models with high influence on the right hand side of the plot which are retained until the end in rightward deletion, but removed early in leftward deletion.

Unlike model averaging, the model projection methodology also produces estimates of two more quantities. The $S_{gg}$, the neg-selfentropy of the generating process, is estimated as $-9.881$. As God, we know that the true value is $-9.877$. These two agree to three significant figures. Also estimated is the distance of the generating process from the (hyper)plane of the NMDS model space. This is very important, because if the generating process is far from the (hyper)plane then any property estimate based on information from the model set should be suspect. The estimate for this discrepancy is $0.00018$, indicating that is very close the (hyper)-plane. The true discrepancy is $5.8 \times 10^{-8}$.

**Discussion**

This is just a brief sketch of an approach. Much statistical development, extension and validation are needed. These will be reported in other venues. Topics for development include:

Response surface methods to predict the properties of models near the generating model. The model average does weight models with low AIC more heavily than models with higher AIC, but does not take into consideration the rate of change of properties across the space. Thus a response surface approach should yield more accurate estimates.
Extensions beyond the near model constraint. As the KL distance between approximating models and the generating model increases, \(-\text{AIC}/2n\) 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 low delta AIC values, even sometimes appearing as the best model. It seems reasonable to think that using heteroskedastic nonlinear regressions such as described by Carroll and Ruppert (1988) and Carroll et al. (1995) will allow for incorporating information from much more distant models into the estimated projection. If this does not prove effective, at least the region in which the projection of the generating model resides can be found by plotting the density of AIC good models in the NMDS space. The projection methodology can be applied in the high density region.

As described above, one of the expectations taken in calculating the AIC is over parameter estimates. Estimation of the location and properties of the estimated projection can likely be improved using the reduced variance bias corrected bootstrap information criterion of Kitagawa and Konishi (2010). A benefit of this is that confidence intervals on the estimated projection can be simultaneously calculated. These intervals are based in frequentist probability and can be expressed either as error statistical confidence intervals or as evidential support intervals. This contrasts with intervals produced by model averaging, which despite their sometime presentation as error statistics are actually posterior probability intervals (under a cryptic assumption that the posterior distribution is normal).

Despite the substantial work that still needs to be done, the approach laid out here already has shown a number of clear and substantial advantages over the Bayesian-based model averaging. First, heuristically, the simple ability of being able to visualize model space will aid in the development of new models. Second, the estimated generating model projection is less constrained by the configuration of the model set than is the model average. Third, \(Sgg\), the neg-selfentropy of the generating process itself is estimable. It has long been assumed that this is unknowable, but it can be estimated and is estimated as part of our procedure. In the example it is estimated quite precisely. \(Sgg\) as a measure of the dispersion of the generating process itself is of great interest. Fourth, the distance of the generating process from the (hyper)plane of the estimated model space can be estimated. It has long been a complaint of scientific practitioners of model identification through information criteria that they can tell which of their models is closest to the generating process, but they can’t tell if any of the models in their model set are any good. Now discrepancy is statistically estimable. Fifth, the strain between a priori and post hoc inference is vacated. The study of the structure of model space corrects for misleading evidence (chance good results), accommodation (over-fitting), and cooking the models. Theoretically, the more models are considered the more robust the scaffolding from which to project the location of the generating process. Currently, the model set for the model projection approach is limited by the near model requirement common to all information criteria analysis. However, as indicated above, non-linear modeling should allow the analyst to bootstrap
(Glymour sense) the structure of model space to validly include in the projection information from models more distant than is valid for model averaging.

Model projection is an evidential alternative to Bayesian model averaging for incorporating information from many models in a single analysis. Model projection, because it more fully utilizes the information in the structure of model space, is able to estimate several very important quantities that are not estimated by model averaging.

References


We began the monograph with an African Proverb for working together to go farther. It applies in particular to deepening our understanding of scientific methodology. We end with a quote from Stephen Hawking that highlights our approach in an especially significant way: “[N]ot only does God play dice, but...he sometimes throws them where they cannot be seen.” At the level of practice, we believe scientists are bent on finding those hidden dice, whether broken, biased, or unfair. They do so with better and better models to approximate reality, although in the nature of the case those models are strictly false. Methodology is in this general sense probabilistic. But it is more specifically probabilistic as well. Even physics, that traditional paradigm of exactitude, should be construed stochastically with parameters to estimate making room for errors in measurements. The possibility of misleading evidence captures our approach to those model-building at still another level. Statisticians who “get to play in everybody’s backyard” help build those models often aimed at estimating parameters as part of probing natural phenomena. At a meta-level, philosophers also have distinctive roles to play. For example, to make scientists and scientifically-minded philosophers aware of their assumptions “where they cannot be seen” especially in their model-building endeavors in which conflating “evidence” and “confirmation” is likely to occur. Developing a credible philosophy of science that escapes this conflation is indispensable. The position we have advanced in this Monograph is based on a multi-disciplinary approach. In our view it is required for a present day multi-disciplinary audience as it has become the “new normal” in both the practice and theory of science.
**Index**

A
Absolute and incremental confirmation, 26–28
Accumulation of evidence, 69, 115–116
AIC evidential analysis, 158
Akaike, 158–165
Argument from dreaming, 143–144
Assessing models without specific parameter values, 157

B
Base-rate fallacy, 43–45
Bayesianism, 29, 32
Belief, 3–5, 8, 11
Bootstrap circularity, 102–103
Bootstrapping, 98, 99

C
Certain data, 55–57
Composite hypotheses, 23, 32, 33
Confirmation, 16–21, 26–29, 32
Confounding, 100–102
Curve-fitting problem, 153

D
Degree-of-belief, 37–38
Diagnostic testing, 39
Dodo Bird verdict, 10

E
Ecological model estimation analysis, 165
Error-probabilities, 76, 77
Error-statistics, 73–80, 82–88
Estimated likelihoods, 158
Evidence, 3–11, 16–18, 22–28, 32–34

G
Global theories, 74, 80–82, 85, 87, 88
Global warming hypothesis, 40
Good reasons for believing, 111–116
Grue paradox, 130–135

I
Inferential uncertainty, 5–6
Interpretations of Probability, 22, 28–32

L
Likelihood ratio, 37, 45
Likelihoods, 16, 20, 21, 22, 24, 25, 28–33
Lottery paradox, 27

M
Misleading evidence, 113, 116–120
Model projection, 171–173
Multi-model inference, 157–173
Multiple models, 84, 85, 89
Multiple statistical models, 165

O
Old evidence paradox, 135–140

P
Paradox of subjective evidence, 71
Piecemeal testing, 80–82
Plausibility arguments, 147–149
Poisson spot, 12
Positive instance account of confirmation, 126, 127, 130
Probabilistic measures, 51–53
Probability of misleading evidence, 118–120

R
Raven paradox, 127–130

S
Selective confirmation, 94–95, 97, 101
Severe tests, 75–76, 80–82, 85

© The Author(s) 2016
P.S. Bandyopadhyay et al., *Belief, Evidence, and Uncertainty*, Philosophy of Science, DOI 10.1007/978-3-319-27772-1
The “Simple Rule”, 54, 55, 59
S measure of confirmation, 63–65
Subjective Bayesianism, 64, 67–70

T
Theoretical constants, 104–106
Theory acceptance, 49–51
Thought-experiments, 148–150
True-model assumption, 75, 82–86, 155, 156

U
Uncertain inference, 46–47
Under-determination, 150

V
Variety of evidence, 104–105
Veridical evidence, 110–111