Inherent Complexity: a problem for Statistical Model Evaluation

Jan-Willem Romeijn
University of Groningen

Abstract

This paper investigates a problem for statistical model evaluation, in particular for curve fitting: by employing a different family of curves we can fit a scatter plot almost perfectly at apparently minor costs in terms of model complexity. The problem is resolved by an appeal to prior probabilities. This leads to some general lessons about how to approach model evaluation.

1 Introduction

Theories often interface with empirical fact through a statistical model, namely a collection of hypotheses that each determine a probability distribution over possible observations. Most statistical inference is carried out on the basis of a model, for example by getting the data to choose among the hypotheses in it, or by redistributing the probability assignment over the hypotheses in the model.

Curve-fitting is an instance of statistical inference. For example, the yearly number of car accidents with claimable damage follows a Poisson distribution, whose characteristics depend on the total distance covered by the vehicle. What determines the statistical model is the exact functional dependence of frequency

on distance. Since vehicles that do not cover any distance will not incur any damage, the intercept will be zero. One statistical model may be that the dependence is linear, so that the hypotheses in the model differ in the slope of the line that relates distance to expected number of accidents. Another statistical model might postulate a more complicated relation between distance and expected number of accidents, e.g., a quadratic dependence, perhaps with the idea that long-distance drivers have proportionally fewer accidents.

While models are typically chosen at the outset, sometimes they are under scrutiny themselves. For example, we might compare the linear and the quadratic models sketched above. Statistical model evaluation allows us to compare such models on a variety of performance measures. Model evaluation is important for scientists and philosophers of science alike. It allows scientists to submit their modeling assumptions to empirical testing, and thereby address the uncertainty over their theoretical starting points. And it gives philosophers of science a concrete and formally precise handle on a fundamental kind of uncertainty. Examples of model evaluation abound, ranging from climate science and ecology to psychiatry and computational archaeology. If philosophers can motivate and develop norms for dealing with model uncertainty, this will have direct implications for the practice of science.

This paper contributes to our understanding of the norms that drive statistical model evaluation. After an introduction into model evaluation tools in section 2, I present a new problem for them in section 3. I then offer a diagnosis of the problem in section 4. In section 5 I show that the problem can be avoided if we involve prior probability assignments in the model evaluation. Throughout I will mostly avoid mathematical detail, to leave more space for conceptual considerations.

2 Statistical model evaluation

The curve fitting problem sketched in the introduction may not seem statistical. Given a family of curves, we simply choose by minimizing the errors, i.e., the

sum of the discrepancies between curve and point. In the so-called least-squares approach, for example, the error is calculated as the sum of the squares of the vertical distance between point and curve. No model seems to be involved in this.

Underneath such a minimization procedure, however, we do find a statistical inference. One central modeling assumption is that the number of accidents N follows a Poisson distribution. A further assumption is that the mean of this distribution depends on the distance D covered by the vehicle,

$$P_\theta(\langle D, N\rangle) = \frac{(\lambda(D))^N}{N!} e^{-\lambda(D)}, \tag{1}$$

with $\lambda(D) = \theta_1 D + \theta_2 D^2$. Then we choose $\theta_1 > 0 > \theta_2$ for the quadratic model and $\theta_1 > 0 = \theta_2$ for the linear one. Note that the model dictates a distribution over N for all values of D but that it does not determine a probability distribution over the values of D itself. The distance D is an explanatory variable, and we presume that it is randomly sampled from a uniform distribution.

The data consist of m pairs of distances and numbers of accidents, collected in a scatter plot:

$$S_{DN} = \{\langle d_1, n_1\rangle, \langle d_2, n_2\rangle, \ldots, \langle d_m, n_m\rangle\}. \tag{2}$$

For any curve and associated hypothesis we can calculate the probability of a scatter plot, i.e., the likelihood of the hypothesis for the data, by multiplying the probability of all the points,

$$P_\theta(S_{dn}) = \prod_{i=1}^{m} P_\theta(\langle d_i, n_i\rangle). \tag{3}$$

A data point $\langle d, n\rangle$ lying outside the normal range for some hypothesis, e.g., with n high while d is low and θ_2 is too, will be improbable, and hence it will strongly decrease the likelihood of the hypothesis. To fit the curve we look for the value of θ, denoted $\hat{\theta}$, that makes the probability of the scatter plot maximal. Generally speaking, maximizing the likelihood of the curve will correspond to minimizing the distance of points to the curve under some notion of distance. Figure 1 offers an impression of what these curves may look like.

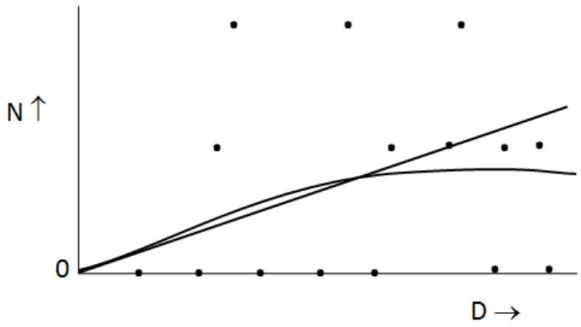

Figure 1: The polynomial curves fitted to the scatter plot.

Against this background it will be clear that evaluating the general shape of the curves, comparing linear and quadratic ones, is indeed part of statistical model evaluation. Note that I use the term "model evaluation", not the more often used "model selection". The selection of a model is a decision, and so involves decision-theoretic as well as inferential aspects. But in what follows I will only consider norms for the comparison of models from an epistemic standpoint.

A very common idea about model evaluation is that, next to the fit with data, it involves the complexity of the model. If a neat fit with the data is achieved by adding many bells and whistles, we are rightly reluctant to put our trust in it. We then say that the model is fitting to noise, or overfitting. In the example, the best fitting curve from the quadratic model will have a higher likelihood than the best curve from the linear model. But this is not to say that the quadratic curve is better. The question is whether the gain in fit weighs up against the cost of a more complicated model.

The extant model evaluation tools, most notably the various information criteria (Claeskens and Hjort, 2008), provide specific formats for this trade-off between simplicity and fit. The two most prominent tools, the Akaike and Bayesian information criteria or AIC and BIC for short, express the simplicity by means of the number of free parameters in the model (cf. Akaike, 1973; Burnham and Anderson, 2002; Raftery, 1995; Schwarz, 1978). The linear model

of the example has one free parameter, and the quadratic model has two. The ICs then differ in how they factor the number of parameters into the trade-off:

$$\mathrm{AIC}(\mathcal{M}_\theta) = 2\log\left(P_{\hat{\theta}}(S)\right) - 2\dim(\mathcal{M}_\theta), \quad (4)$$

$$\mathrm{BIC}(\mathcal{M}_\theta) = 2\log\left(P_{\hat{\theta}}(S)\right) - \log(m)\dim(\mathcal{M}_\theta), \quad (5)$$

in which $\mathcal{M}_\theta$ is the model parameterized by the vector θ, the number of free parameters is given by $\dim(\mathcal{M}_\theta)$, and $\hat{\theta}$ is the hypothesis in the model with maximum likelihood for the data D, so that $P_{\hat{\theta}}(S)$ is the likelihood of the maximum likelihood estimator for the data S. In the BIC the penalty for complexity is scaled according to the sample size of the data m.

The involvement of the complexity of models in their evaluation may seem intuitive on pragmatic or metaphysical grounds. A simpler model is easier to use, or we might think that the world itself is a simple place, perhaps because the Demiurge is an efficient or lazy being. The actual reason for the appearance of the complexity penalty in the ICs is epistemic though. Moreover, the motivation is different for the various information criteria on offer. For example, the AIC factors in the number of parameters as a result of approximating the expected Kullback-Leibler divergence to the true hypothesis. And for the BIC the penalty for complexity drops out of an approximation of the past predictive performance of the model, as measured by the marginal likelihood.

The number of model parameters surfaces repeatedly as a criterion for model evaluation, under a variety of epistemic good-making features of models. Very roughly, the underlying reason is that the predictions and general empirical claims of more complex models will be less robust and reliable. In a more complex model the same number of data points will be used to determine a larger number of parameter values, and so the available information will have to be spread more thinly. For the AIC this shows up in the stumpness of the likelihood function over the model, and in the BIC it appears as the stumpness of the posterior probability distribution within the model. The general idea is that we can always introduce an additional parameter that improves the best

fit in the model, but that we might then lack the data to properly back up a stable value for this additional parameter.

However, this intuition does not cover everything that is salient about complexity in model evaluation. There is another epistemic good-making feature, strongly related to complexity and the number of parameters, that needs to be taken into account when we compare models. This further feature concerns something like model size. It can be expressed by means of the prior probability distribution within the models, as the following model evaluation problem will reveal.

3 Cheap and almost perfect fit

Consider again the example of the scatter plot and the polynomial model. But instead of using the polynomial curves, as detailed above, imagine fitting the data with a model based on trigonometric functions, or sine curves for short. We use the Poisson distributions of Equation (1) but instead of choosing $\lambda(D)$ to be polynomial we choose

$$\lambda(D) = \alpha_1 - \alpha_1 \cos(\alpha_2 D). \tag{6}$$

Figure 2 gives an impression of the fit that may be achieved by the so-called sine model. Importantly, all the points in the scatter plot are given close to maximal probability, because they all end up sitting arbitrarily close to the curve, and hence to the mean for the distribution at the given distance D.

The key observation is that we have achieved this remarkable fit at the expense of only two parameters, α_1 and α_2. It is known that we can obtain a perfect fit to m data points with a polynomial curve of degree $m - 1$. But fitting any number m of points with two parameters seems inexplicably efficient. Clearly, if we were to apply model evaluation criteria like AIC or BIC, or indeed any other method in which complexity is expressed by the number of parameters, the sine model wins out on the quadratic model, and most likely also on the linear model. What is going on?

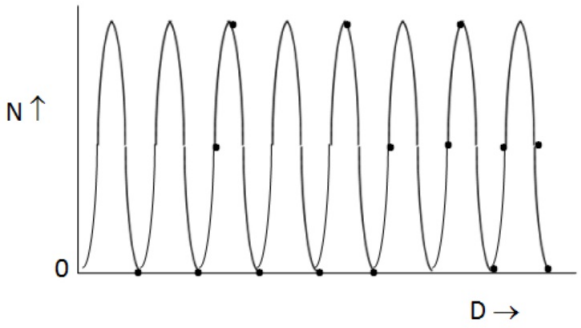

Figure 2: The sine curve that perfectly fits the scatter plot.

Before providing a diagnosis, let me emphasize that the claim that a near-perfect fit is always possible is mathematically non-trivial. In the remainder of this section I will provide more detail to substantiate it. Notably, the fit does not hinge on the assumption of any particular distribution, be it Poisson, normal, or otherwise, or on any particular format of the data, be it real numbers, integers or otherwise. Moreover, given that the scatter plot will manifest on a finite domain $0 < D < L$ we need not even suppose that the parameters are real valued: it is enough to consider sine curves with a period L/t for $t \in \mathbb{N}$, as one does in a Fourier series. Despite all this, it turns out that there are always infinitely many almost perfect fits to a set of points. This abundance of solutions will turn out to be of crucial importance for the resolution of the problem.

Say that we have been given a scatter plot S_{dn} whose farthest points are at $d_i = L$ and $n_j = H$. For convenience we set $\alpha_1 = H/2$, but any α_1 larger than that will work too. Take any specific point $\langle d, n \rangle$ from the scatter plot, consider the curves with $\alpha_2 = L/t$ for increasing t, and ask: for what values of t does the sine curve intersect with the line $D = d$ in very close proximity to the value n? Observe that $d \in [kL/t, (k+1)L/t]$, and that the curve covers the whole of the range $[0, H]$ over this interval of D twice. If we allow for a discrepancy of ϵ between the curve and the value n and assume that d falls within the first

half of the interval, we must require that

$$\frac{L}{\pi t}\cos^{-1}\left(1-{}^{2n-\epsilon}/_{H}\right) \;<\; d-\frac{kL}{t} \;<\; \frac{L}{\pi t}\cos^{-1}\left(1-{}^{2n+\epsilon}/_{H}\right). \tag{7}$$

If d falls in the second half of the interval $\left[{}^{kL}/_{t}, {}^{(k+1)L}/_{t}\right]$, we require an analogous constraint. Because the slope of the cosine is bounded between -1 and 1, we may replace the above inequalities with

$$\frac{L}{\pi t}\cos^{-1}\left(1-{}^{2n}/_{H}\right)-\frac{\epsilon L}{\pi t H} \;<\; d-\frac{kL}{t} \;<\; \frac{L}{\pi t}\cos^{-1}\left(1-{}^{2n}/_{H}\right)+\frac{\epsilon L}{\pi t H}, \tag{8}$$

and similarly for d sitting in the second half of the interval. Consequently, for every t there is a specific region of length ${}^{4\epsilon L}/_{\pi t H}$ within the interval of length ${}^{L}/_{t}$ that includes d, for which the resulting error lies within an ϵ bound. The question merely is, for every separately t, whether d indeed lies within this specific region.

To establish when the latter obtains, we first recall that the d_i's from the scatter plot S_{dn} were randomly sampled from a uniform distribution over $[0, L]$. This means that the individual d from the sample is almost surely, i.e., with probability one, a random number. Consequently, there will be no pattern in how d shows up inside the intervals $\left[{}^{kL}/_{t}, {}^{(k+1)L}/_{t}\right]$ for increasing t. The locations of d are evenly distributed over all parts of this interval. Hence for any $\epsilon > 0$ there will be infinitely many t for which d will fall within the portion of length ${}^{4\epsilon L}/_{\pi t H}$ inside the interval of length ${}^{L}/_{t}$. The relative size of the region in which the curve is sufficiently close to the value n is constant for increasing t at ${}^{4\epsilon}/_{\pi H}$. And so there will be infinitely many sine curves that have an arbitrarily small error in fitting the point $\langle d, n\rangle$.

This suffices as an argument for there being an infinity of curves that fit any finite number of points almost perfectly. For a single point, the fraction of sine curves will tend to ${}^{4\epsilon}/_{\pi H}$. So for a set of m points that are randomly distributed over D, the fraction will tend to $({}^{4\epsilon}/_{\pi H})^{m}$. When making ϵ small and thus maximizing the likelihoods, the fraction of curves with good enough fit will be very small. But there will still be infinitely many fitting ones.

4 Diagnosis of the problem

The fact that there are infinitely many equally well-fitting sine curves incapacitates some of the standard model evaluation tools. The AIC, for example, is not defined for unidentified models. While being silent may be better than positively evaluating the intuitively incorrect sine model, a negative evaluation of the sine model seems preferable. Our discussion revolves around three observations: the sine model is not robust, counting parameters is a nontrivial affair, and the set of best fitting sine curves is not well-behaved. This sets us up for a solution to the problem along Bayesian lines in the next section.

First consider the robustness of the sine model. Imagine that we alter the scatter plot by slightly nudging a single data point. What will be the result if the curve is a polynomial of a given degree? Clearly, any curve that was fitted to the data will change a little as well. But the rough shape of the curve will not change a lot: a small change in data is matched by a similarly small change to the best fitting curve. By contrast, if the curve is a trigonometric function, then nudging a single data point slightly will radically alter the best fitting curve. It will lead to a completely new set of best fits. We might say that the sine model is too versatile, lacking robustness, or skittish: it is oversensitive to the smallest of changes in the scatter plot.

The AIC and BIC do not accommodate this feature of models. MDL-based model evaluation tools and extensions of the AIC and BIC fare slightly better. The Fisher information approximation (FIA) for example includes a so-called geometric complexity term, based on the Fisher information. One might say that this expresses model size in terms of how densely packed the model is with likelihood functions (Grunwald, 2007; Myung et al, 2000; Ly et al, 20XX). The term penalizes skittish models because they will in general cover a larger set of probable data patters: small changes lead to wildly different functions, and in this sense the skittish models are indeed packed densely. Furthermore, developing the AIC and the BIC, as discussed in Bozdogan (1987) and further references therein, we also encounter the Fisher information. So there are nat-

ural extensions of the AIC and the BIC that accommodate something of the skittishness.

However, in all of these refined methods, the contribution of the Fisher information (FI) term is not of the required order of magnitude to resolve the problem of the sine curves. Apart from the original AIC, the FI term is trumped by the term that captures complexity as the number of free parameters, and which grows with $\log(m)$. And the FI term cannot compete with the fit term, which grows with m in all model evaluation tools. For larger data sets the influence of the FI term on the model score therefore dwindles, so that the sines seem preferable after all.

A second observation concerns the deceptively low dimensions of the sine model: it seems to harbor an inherent complexity that is not expressed in the number of parameters. The sine model illustrates that model dimension is a fleeting notion. As a quick illustration, note that statistical parameters are often real numbers. But real numbers are such that we can package any amount of information into them. For example, a sufficiently complicated function will allow us to compact two real numbers in a single one, by constructing the numerical expansion of a number from two such expansions, e.g., $0.135\ldots$ and $0.246\ldots$ yield $0.123456\ldots$, and so on. While this sort of function is of course hopelessly contrived, it illustrates that counting statistical parameters does not give us a fair indication of model dimensions.

This general observation has been made about model evaluation criteria more often, for example in Bozdogan (1987), who proposes to adapt the AIC by involving the sample size, thereby bringing it closer to the BIC. His motivation for adapting the penalty term is, by and large, that the notion of complexity is not adequately captured by the dimension term in the original AIC. Similar sentiments are expressed in Balasubramanian (2005) who develops minimum description length (MDL), and in Romeijn and van de Schoot (2008); Romeijn et al. (2012) who investigates and extends the BIC. The latter two point to a more general notion of model size as a component of complexity. However,

while these proposals are in the right direction, the adapted versions of AIC, BIC, and MDL still give the number of parameters a central role.

A more promising method for dealing with the problem of the sine curves is offered by the so-called Deviance information criterion, or DIC for short (Spiegelhalter et al, 2002). The DIC was originally designed for comparing hierarchical Bayesian models, in which the number of free parameters is not clearly defined. Central to the DIC is the so-called deviance, i.e., the reduction in surprise due to estimation, which can be thought of as a degree of overfitting. The penalty for complexity in the DIC is given by the effective number of parameters, which is based on the notion of deviance. However, in this paper I will not investigate in detail how the DIC responds to the sine model.

A final observation brings us closest to the ultimate reason that trigonometric curves are problematic for the purpose at hand. Note that both polynomial and trigonometric curves can be used as a basis for the space of functions on a finite domain, in the algebraic sense that they parameterize that space: we can write down functions by their Taylor or Fourier series. We can collect the curves that almost perfectly fit some scatter plot into a set within the space of functions. But for the Taylor and Fourier series this set will look very different. In the Taylor parameterization, the set is a well-behaved region sitting somewhere in the linear combination of at least m axes. But the set of well-fitting curves will look much wilder and disjointed in the Fourier parameterization, disjointed and intersecting with distinct axes rather than being lumped together.

The implications of this are best brought out through a variant of the robustness discussed above, namely by considering what happens if we add a point in the scatter plot. The original polynomial curve will not change too radically: the region of well fitting curves shifts slightly. By contrast, the set of best fitting sine curves alters significantly with the addition of a point, not so much by being relocated but rather by being constrained severely. There are infinitely many sine curves that fit the scatter plot, but almost all of those curves will miss the additional point by a stretch, and so be eliminated from the set of well

fitting curves. The solution of the problem hinges on exactly this elimination of hypotheses.

5 Priors to the rescue

This section develops a particular response to the problem of the sine curves. It relies on so-called Bayesian model selection, or Bayesian model evaluation (BME). Following BME the sine model looses against polynomial models because of the specific failure of robustness introduced above.

The message of this section is not that we should embrace BME as the new standard in model selection. I will not make a systematic comparison with other model evaluation criteria and their relation to the salient notion of robustness. Looking at the solution that BME provides and the central role for the so-called marginal likelihood in BME, we might expect that other approaches in which the marginal likelihood is central, e.g., the DIC and MDL-based criteria, will also provide a solution. Because we can only compute something like a marginal likelihood if we adopt some version of a prior within the model, the central point of this section is rather that solutions will have to rely on priors of some kind.

The central idea of BME is to compare models by their posterior probability assignment:

$$\frac{P(\mathcal{M}_1|S)}{P(\mathcal{M}_2|S)} = \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)} \times \frac{P(S|\mathcal{M}_1)}{P(S|\mathcal{M}_2)}. \tag{9}$$

Assuming an equal prior for the models $\mathcal{M}_i$, the posteriors are completely determined by the ratio of the so-called marginal likelihoods,

$$P(S|\mathcal{M}_i) = \int_\Theta P(H_\theta|\mathcal{M}_i) P(S|H_\theta \cap \mathcal{M}_1)\, d\theta, \tag{10}$$

in which Θ is the parameter space. The likelihoods $P(S|H_\theta \cap \mathcal{M}_1)$ are a different notation for the $P_\theta(S)$ of the foregoing. Notice that the prior within the model, $P(H_\theta|\mathcal{M}_i)$, plays a key role in the computation of the marginal likelihood. Many approaches to model evaluation rely on the marginal likelihood of the model, including the BIC, the DIC, and MDL-based approaches. All these approaches must use some notion of a prior.

Now recall how sine curves manage to fit any scatter plots almost perfectly, in particular that there are infinitely many such curves. With the addition of a new point this set of best fitting curves will loose a large number of members, and this will severely impact the marginal likelihood of the sine model. Following Equation (3), we see that the likelihood of sine curves that retain their fit be multiplied by a maximal probability for every new data point. But this only holds for a small fraction $4\epsilon/\pi H$ of the sine curves. The fraction $1 - 4\epsilon/\pi H$ of sine curves will be multiplied by a factor that falls far short of the maximum probability.

By comparison, the likelihoods of the well fitting polynomial curves will pick up a factor that is somewhat lower than the maximal probability for each point, though not falling very far short of the maximum. Importantly, this high but not maximum factor will apply to a set of curves that is more or less stable and that will accumulate more and more probability with the addition of data points. Consequently, the overall factor picked up by the marginal likelihood of the polynomial models will tend to this high but not maximum factor.

To put this in a more mathematical format, say that the average factor picked up by the likelihood of a sine curve outside of the set of best fitting curves is U, that the same factor applies to badly fitting polynomial curves, that the factor for a well fitting polynomial is V, and for a best fitting sine curve W, so that $U < V < W$. For the sine curves we obtain

$$P(\langle d_{m+1}, n_{m+1}\rangle | \mathcal{M}_{\text{Sine}} \cap S_{dn}) \approx \left(1 - \frac{4\epsilon}{\pi H}\right) U + \frac{4\epsilon}{\pi H} W, \tag{11}$$

which is arbitrarily close to U. For the polynomial curves we will have

$$P(\langle d_{m+1}, n_{m+1}\rangle | \mathcal{M}_{\text{Poly}} \cap S_{dn}) \approx (1 - R_m)U + R_m V, \tag{12}$$

in which R_m tends to 1 for increasing m so that the factor tends to V. The result is that the sine model performs less well than the polynomial model on the BME criterion. On BME, therefore, the inherent complexity of the sine curves is adequately factored in.

It will be insightful to return to the observations that the set of well fitting sine curves is skittish. Well fitting polynomial curves of a given degree are concentrated in a particular region within the model, in which posterior probability will accumulate when data size increases: all of them will respond to new data points in roughly the same way. By contrast, with every new data point a small fraction of the well fitting sine curves is multiplied by a high likelihood, while a large portion picks up a low factor. It expresses the skittishness of sine curves that such a large portion of curves is suddenly far off in their prediction.

We can also convert this reasoning to arrive at the observation about model size. Judged from the prior probability distribution within the sine model, the set of well fitting curves is very small indeed: after m points it has decreased to $(4\epsilon/\pi H)^m$. But considering the prior within the polynomial model, the set of well fitting curves retains a reasonable size. What this signals is that the sine model, although it has only two free parameters, has many more different statistical hypotheses packed into it. It is versatile at the cost of a particular kind of robustness. The use of a prior within the model enables us to bring this kind of robustness out.

6 Conclusion

We cannot turn the foregoing into an argument for BME: other model evaluation criteria may also have a response to the problem at stake. But there are several general lessons to take away. One is that we must never mistake the number of parameters in a model for its actual complexity. A related lesson is that we must not forget the deeper motivations for the model evaluation tools, i.e., the good-making features that the tools are based on. Concentrating on those features will guide us to a better understanding of our evaluations.

Another general general lesson ties in with earlier work on the role of size in model evaluation (Romeijn et al., 2012), and indeed with scientific methodology as a whole. In the solution of the problem with the sine model, we can recognize a Popperian theme. Models that allow for fewer possible data patterns are

preferable to those that allow for a very wide range of data patterns. To express some notion of model size in our evaluations, we have to adopt some measure over the space of distributions over data. So we must involve something akin to a prior.

There is, however, a problem with the idea that we can objectively determine how densely distributions are packed together in a model. To say that a set of distributions shows a wide variety in the data patterns that it can adapt to, we need to presuppose a notion of similarity among data patters or, more generally speaking, a metric over sample space. In this paper that metric was adopted implicitly, as part of the way in which we depict and conceptualize the data. This dependence on the metric of the sample space points to a potential subjectivity in adjudicating between statistical models, or at least a reliance on a natural conceptualization of the sample space. This idea deserves to be studied in its own right.

Acknowledgements

The author wishes to thank Elliott Sober and Tom Sterkenburg, as well as audiences in Santiago de Compostella, Padova, Helsinki, Gent, and Groningen.

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *2nd International Symposium on Information Theory*, eds. B. N. Petrov and F. Csaki, Akademiai Kiado, Budapest, pp. 267–281.

Balasubramanian, V. (2005). MDL, Bayesian inference, and the geometry of the space of probability distributions. In *Advances in Minimum Description Length: Theory and Applications*, P. J. Grunwald et al. (eds.), pp. 81–99. MIT Press, Boston.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion. Psychometrika 52(3), 345–370.

Burnham, K.P. and D.R. Anderson (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. New York: Springer.

Claeskens, G. and N. L. Hjort (2008). Model selection and model averaging, Cambridge: Cambridge University Press.

Grunwald, P. (2007). The Minimum Description Length Principle. MIT press, Cambridge (MA).

Ly, A. J. Verhagen, R. Grasman, and E-J. Wagenmakers (20XX). A Tutorial on Fisher Information, unpublished manuscript.

Myung, J. et al. (2000). Counting probability distributions: Differential geometry and model selection. Proceedings of the National Academy of Sciences 97(21), pp. 11170–11175.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25, pp. 111–163.

Romeijn, J. W. and R. van de Schoot (2008). A Philosophical Analysis of Bayesian model selection. In Hoijtink, H., Klugkist, I., and Boelen, P. A. (2008). *Bayesian Evaluation of Informative Hypotheses*, Springer, New York.

Romeijn, J.W., R. van de Schoot, and H. Hoijtink (2012). One size does not fit all: derivation of a prior-adapted BIC. In Dieks, D., W. Gonzales, S. Hartmann, F. Stadler, T. Uebel, and M. Weber (eds.), Probabilities, Laws, and Structures. Berlin: Springer.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6, pp. 461–464.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society B*, 64, pp. 583–639.