Is the number of parameters the best measure of model complexity?

Christopher Fogelberg and Vasile Palade

February 8, 2008

Abstract

This paper considers an argument which has been interpreted as showing that Bayesian reasoning automatically embodies and justifies Occam's razor, because it automatically balances goodness of fit with the number of parameters. However, careful analysis shows that this interpretation is incorrect. Through theoretical considerations and by analysing some examples we see the kinds of situations in which Bayesian reasoning embodies Occam's razor and the kinds of situations it does not. We also explore the directions our analysis suggests for research into Bayesian model scoring.

1 Introduction

This papers analyses a common argument (see e.g. [18, 19, 23, 12]) which has been interpreted as showing that Bayesian reasoning automatically embodies and justifies *Occam's razor* (OR). This paper considers the problems with this interpretation and why it is incorrect.

1.1 Occam's Razor Defined

Occam's razor is a rule of thumb that we often use without consideration. It was first made popular by William of Ockham in the 14th century when he said that "plurality ought never be posed without necessity". As with Murphy's law, this phrasing has given rise to an enormous number of variants: That the simplest explanation is the right one, that the epiphenomenal[1] parts of a theory are untrue, that a theory should be as simple as possible but no simpler[4], that simplicity is the ultimate sophistication[2] and so forth.

[1] Loosely, "undetectable".

[2] Leonardo da Vinci

In this paper we are only considering the use of Occam's razor for model comparison. In this situation, OR can be defined descriptively as "the simplest of two or more models that fit the data is usually the right model". This is similar to Einstein's definition, and also accords with the common, modern, prescriptive definition: "accept the simplest explanation that fits the data"[19, ch. 28].

1.2 Organisation of the Paper

The remainder of this section, subsection 1.3, describes several historical and contempary justifications of Occam's razor. Section 2 presents and explains the easily-misinterpreted Bayesian argument in depth. Section 3 discusses the theoretical problems with this argument and section 4 relates examples used to validate the argument and shows how further analysis undermines them. Finally, section 5 summarises the problems and the new areas of research for model scoring measures that the problems suggest.

1.3 Justifications for Occam's Razor

OR has been justified aesthetically, epistemically and by its past success. Aquinas (1225–1274) argued that a more beautiful theory is more likely to be true and that this is reflected in natural processes. Historically, it has also often been the case that the simpler explanation which fits the data turns out to be true. Do the planets and the sun (epicyclically) orbit the earth, or do Earth and the other planets orbit the sun?

More recently, philosophers have argued[21, 25] that simpler theories are more informative. However, these justifications for OR and the debate surrounding them (e.g. [26, 24, 2]) are not the focus of this paper. It is important to note though that Occam's razor makes a claim about *objective* truth. For this reason, justifying it is very difficult.

It has also been claimed that coherent Bayesian inference automatically embodies OR[19, p. 344]. Jerrold Katz's deductive argument[13] is very similar to the Bayesian argument described in section 2, although we have not considered the applicability of our analysis to Katz's argument.

However the Bayesian argument for justification and embodiment very subtly begs the question. The problems it faces also highlight important consequences for model scoring measures. These consequences are similar to what has been discussed in the philosophy of model science and population biology[16] and they suggest new areas of research in model scoring measures. Beyond this, it is also important that the non-justification of OR is made

clear in the context of machine learning as well, as past research (e.g. [14, 3]) has considered the argument without deeper critical inspection.

2 The Bayesian Argument for Occam's Razor?

Given the data and using Bayes's law we can calculate the relative posterior probability of one model as shown in equation 1. The posterior probability relative to some other model is determined by the likelihood of the data given the model and the prior probability of that model.

$$p(M|D) \propto p(M)p(D|M) = prior \times likelihood \tag{1}$$

Given equation 1, one could explicitly assume OR by giving simpler models proportionally greater prior probability. However, if the prior over model complexity is uniform then it appears that Bayesian reasoning automatically justifies OR. This is because of the impact of the size of the parameter space on the evidence factor. If we arrange the models into families, such as linear models, quadratic models, cubic models etc., then some families of models will have more parameters than others. The number of parameters is often used as a measure of complexity.

2.1 From Bayes's Law to Occam's Razor

At this point the argument becomes quite subtle. If two models in different families are equally accurate and if the prior distribution over families is uniform (as specified above) then a model in a more complex family has a lower posterior than a model in a simpler family. This is a consequence of the sum law of probability and because the parameter space of the more complex model is larger.

We will now show this formally and in more detail. Consider two families of models, M_c and M_s, where M_c is a family of complex models and M_s is a family of simple models. Denote a particular model in a family M_m as M_m^i. Because a model in M_c has more parameters, there are more models possible in the family M_c than in M_s, i.e. $|M_c| > |M_s|$.

Assume that the best fitting model in the family M_m is denoted by M_m^{top}. If M_c^{top} and M_s^{top} fit the data equally well then the argument in table 1 shows that $p(M_s^{top}|D) > p(M_c^{top}|D)$. Bayesian reasoning appears to automatically include an Occam's bias for simpler models. Further analysis[18] can show how much more accurate a more complex model must be to be more probable

than a simpler model (the *Occam factor*), but this extension does not concern us here.

Table 1: How Bayesian reasoning appears to entail Occam's razor.

	$p(M_c) = p(M_s)$	*prior assumption*
	$p(M_m) = \sum_{i \in M_m} p(M_m^i)$	*sum law*
	$\lvert M_c \rvert > \lvert M_s \rvert$	*by definition*
$\therefore$	$p(M_s^i) > p(M_c^j)$ for all $i \in M_s$, $j \in M_c$	
	$p(D \lvert M_s^{top}) = p(D \lvert M_c^{top})$	*by definition*
	$p(M_m^i \lvert D) = p(D \lvert M_m^i) p(M_m^i)$	
$\therefore$	$p(M_s^{top} \lvert D) > p(M_c^{top} \lvert D)$	*substitution*

2.2 A More Intuitive Phrasing

The more complex family is a larger set of models. Thus it spreads its share of the prior distribution more thinly. As a result, any particular model in that family is less likely *a priori*. Therefore if a simple model and complex model predict equally well then the simpler model is more probable *a posteriori*.

An immediate objection to this argument is that it does not show that Bayesian reasoning jusifies OR, only that Bayesian reasoning assumes it for certain priors and in certain situations. This is as discussed by Sivia[23, ch. 4]. There he notes that, while this argument is deductively correct, the size of the parameter space is not always sufficient or appropriate. We will explore this objection in the next two sections.

3 Theoretical Problems with the Bayesian Argument

As noted, section 2 could be interpreted as only showing that Bayesian reasoning implicitly assumes OR if the prior is uniform over model complexity, and to not make that assumption we must adjust our prior or otherwise normalise for complexity.

3.1 Representational Dependence

More importantly, there is no way to choose the representation we use without making assumptions. This is important because a model's simplicity de-

pends entirely on its representation[17]. We will explore the practical impact of this in 4.2. For now, consider that Bayesian reasoning assumes the *principle of indifference*[11], and that this assumption can lead to contradictory and paradoxical outcomes when we try to compare different representations with continuous variables.

We will show this with reference to a modified version of the predictive problem in [26, p. 303]. Imagine a factory that makes cubic boxes with side length between 0 and 1 metres. We are interested in both the side length of the boxes (perhaps because we need to pack them) and also their surface area (perhaps we need to paint them). For the sake of argument, we will assume that knowledge of the side length and surface area are equally important to us.

Over time we will use the sequence of boxes that we see the factory produce to infer the distribution over side length and surface area. To maximise the informativeness of this data we will use a uniform prior, $\theta = U(0, 1)$. Thus our probability distributions for length and surface area are as shown in equations 2 and 3.

$$p(len|D, \theta_{len}) \tag{2}$$

$$p(area|D, \theta_{area}) \tag{3}$$

This is the setup of our Bayesian predictive problem. Now imagine that we want to predict the posterior probability $p(len > 0.5|D, \theta_{len})$. To simplify the calculation and so that we do not need to assume an evidence distribution in our analysis we specify that we are asked to answer this question before we have seen the factory produce any boxes. This means that $D = \{\}$ and therefore:

$$\begin{aligned} p(len > 0.5|D, \theta_{len}) &= p(len > 0.5|\{\}, \theta_{len}) \\ &= p(len > 0.5|\theta_{len}) \\ &= \int_{0.5}^{1} U(0, 1) \\ &= 0.5 \end{aligned} \tag{4}$$

However if we ask an equivalent question of the surface area, where $area = 6 \times len^2$, we see that:

$$
\begin{aligned}
p(area > 1.5|D, \theta_{area}) &= p(area > 1.5|\{\}, \theta_{area}) \\
&= p(area > 1.5|\theta_{area}) \\
&= \int_{0.25}^{1} U(0,1) \\
&= 0.75
\end{aligned} \tag{5}
$$

This is a different and inconsistent (contradictory) answer to the same question. At first this might not seem problematic, because the side length and surface area are different quantities, and we explicitly specified a different prior for each. If the variables were unrelated, the inconsistency would not even exist. However, the distribution over the surface area of the boxes can be calculated from the distribution over the side length of the boxes and vice versa. Furthermore, we chose our priors in such a way as to maximise the informativeness of the data.

As a result of these choices, all of which can be strongly argued for *a priori*, we now have an *inconsistent triad*. Either we cannot maximise the informativeness of the data for the surface area, or we cannot maximise the informativeness of the data for the side length, or we will get inconsistent results.

Symmetrically, there are three ways this triad could be resolved. We can prioritise the surface area over the side length, the side length over the surface area, or accept our inconsistent predictions. But by the specification of the problem and through a desire for consistent predictions none of these is acceptable. Because there are no other possibilities we nonetheless must choose one of them. This is Bertrand's paradox[10].

Importantly, there is no way of making a principled or objective choice of one over the other, which we choose for this problem will depend on our personal inclinations. The intrinsic and principled subjectivity at the heart of Bayesian reasoning makes it very powerful. It's inability to privilege one representation over another also precludes it from justifying Occam's razor.

4 Some Examples

This section relates the two examples presented in [19]. It shows that each example can be both an example and a counter-example for the Bayesian argument in section 2.

4.1 Sequence Prediction

The problem of sequence prediction is a special case of the more general problem of curve fitting. Therefore almost all of what we say applies also to curve fitting problems. However, for the sake of simplicity and also to stay focused on examples used to support justification-interpretation of the Bayesian argument we will only consider sequence prediction.

4.1.1 Automatic Embodiment

Sequence prediction involves predicting the next value in some sequence, given a finite number of past values. If we consider the same sequence as [19] (-1, 3, 7, 11, ...) we can immediately see that there is an infinite range of possible relationships between the data points. These relationships fall into families. One family is *lin*: $x \rightarrow cx + d$. Another family is *cub*: $x \rightarrow ax^3 + bx^2 + cx + d$. Because *lin* has fewer parameters (2) than *cub* (4) and by the argument in 2.1, any member of *lin* which is as accurate as a member of *cub* will be *a posteriori* more probable than it. This is an example of automatic embodiment.

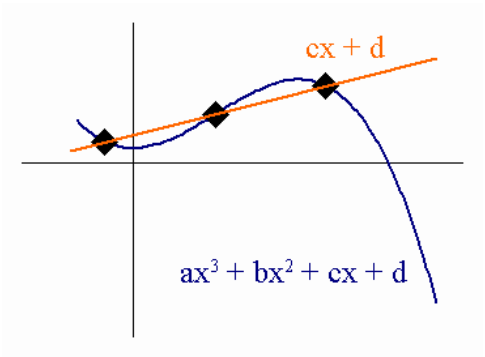

Figure 1: Two possible relationships explaining the data.

4.1.2 *Ad Hoc* Families

However, consider that $lin \subset cub$, as all models in *lin* are also models in *cub* with $a = 0, b = 0$. Are *lin* and *cub* really different families? If they are not, i.e. if *lin* is really just a short hand representation of some members of *cub*, then any inference which depends on this difference is invalid.

Alternately, if we still claim that *lin* and *cub* are different families then the sub-family problem[5] becomes unavoidable. Consider, *lin-sub*, the sub-

family of linear models which has b fixed at its maximum likelihood value B. The best member of this family will (by definition) fit the data as well as the best member of *lin* does. By the logic of our claim that *lin* and *cub* are different families then *lin-sub* and *lin* must be as well. Therefore, if simplicity is our guide, we should prefer a model from the *ad hoc* family *lin-sub*. And similarly with the family *lin-sub-sub*, fixing $c = C$. In this family there is only one model. Occam's razor and Bayesian reasoning has forced us to over-fit our model to the data.

Further, despite our earlier denial, we must accept that *lin* is itself a sub-family of *cub* with $a = 0, b = 0$, i.e.: $x \rightarrow 0x^3 + 0x^2 + bx + c$. Just as *lin-sub-sub* can be obtained by fixing parameters in *lin*, so *lin* can be obtained by fixing parameters in *cub*. *Ad hoc* restrictions on what constitutes valid inference can be used to avoid this problem, but such restrictions undermine the principled way in which Bayesian reasoning appears to automatically embody OR.

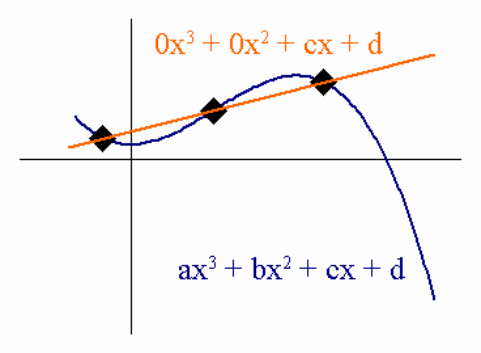

Figure 2: Two possible relationships explaining the data. Is *lin* really simpler than *cub*?

In summary, it appears that just using the size of the parameter space as a definition of complexity is insufficient. Section 5 will speculate about alternatives and extensions which may address this.

4.2 Selecting a World Model

Is there one box behind the tree in figure 3, two, or more? Occam's razor would suggest just one. So far we have focused on showing that Bayesian reasoning does not justify OR. Now we will show that coherent Bayesian inference does not even necessarily embody it. Our analysis relies on another

example of representational dependence. Section 3 also discussed representational dependence.

This subsection presents two representations. For the sake of a simpler exposition we will contrast just the one box world, M_1, and the two box world, M_2. Implicit in our discussion of this problem is the assumption that one box is simpler. What we are investigating is whether or not it is also automatically more probable.

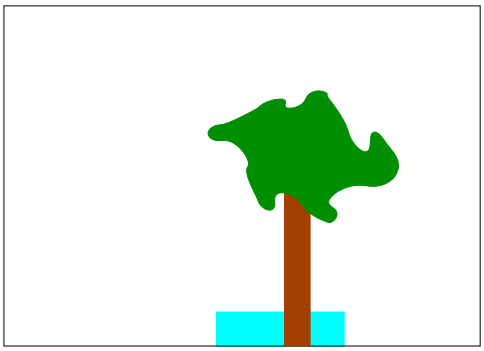

Figure 3: Is there one box behind the tree, two, or more?

4.2.1 World Specification

The first and most important thing to do is to make sure that the world is explicitly specified in every way that it needs to be. If it is not any conclusions that we reach could be an invisible artifact of our world. For comparative purposes our world specification will be as Mackay's was, to the extent that we can determine it.

Firstly, boxes can overlap[19, p. 351]. Despite this, the figure appears to be a warped photograph which doesn't display distance stereoscopically and which only shows the front-side of the box(es) and the tree. For this reason we will not consider distance except to allow overlaps.

The distribution over box side length and height is uniform and ranges over 20 discrete values, although the extent of these is not given[3]. For that reason we will assume a uniform distribution over all side lengths (and/or side area as appropriate) which fit in the image. We do not need to give a

[3]Pixels? But if pixels are used then the relative dimensions of the given maximum box and tree trunk width do not match in the original image. In any case, difference in this area is unimportant. Just as a rough estimate was sufficient in [19], a rough comparison will make our point clearly.

prior over the number of boxes, as we assume this *a priori* in the two models we are comparing.

We will not consider a parameterisation which would allow for falling boxes. This is as with Mackay and because such a parameterisation only has a multiplicative effect on the relative probability of M_1 and M_2 when using either of the representations we will shortly describe. Such a requirement does more to constrain how the image appears than affect the relative probability of M_1 and M_2.

In addition, the boxes can also be parameterised by their colour. As with falling boxes, this only has a multiplicative effect on the relative likelihood of M_1 and M_2 and so we will not consider it in our detailed analysis. The world is clearly abstract. This means that other constraints could be invisibly implicit, but the box world is now sufficiently specified for the two representations to be compared.

4.2.2 A First Representation

The first representation we will consider was described (along with the world, in 4.2.1) in [19]. Each box is represented by three parameters[4]: its position, its width and its height. Therefore $|M_1| = 3$ and $|M_2| = 6$. Because $|M_1| < |M_2|$ and through the argument in table 1 we find that $p(M_1|img) > p(M_2|img)$.

4.2.3 Another Representation

A priori there is no reason to prefer the parameterisation summarised in 4.2.2. We will now describe another parameterisation and consider its implications. Take the following variables:

- B, the visible surface area of boxes in figure 3
- B', the actual surface area of all boxes in figure 3
- B_h, the height of the box(es) visible in figure 3
- A, the area of figure 3 (i.e. tree + sky + boxes)
- T, the area of the tree trunk in figure 3
- T_w, the width of the tree trunk in figure 3

[4]Plus one for colour; we do not consider colour here, for reasons as described in 4.2.1

With these variables we can express $p(M_1|img)$ and $p(M_2|img)$ as distributions in terms of the surface area of boxes in the image. This is shown in equations 6 and 7. According to the world specification the priors over M_1 and M_2 is uniform. Thus we do not need to include the priors in the equations (with a uniform prior the *maximum a posteriori* probability equals the *maximum likelihood*).

$$p(M_1|img) = p(B' = B + B_h \times T_w \pm \frac{\delta}{2}|img) \quad (6)$$

$$p(M_2|img) = p(B \leq B' \leq 2(B + B_h \times T_w)|img) \quad (7)$$

Because the observed surface area of boxes in the image, B, is less than the surface area in M_1 (i.e. $B + B_h \times T_w \pm \frac{\delta}{2}$) and because $B + B_h \times T_w \pm \frac{\delta}{2}$ is a very small subset of the possible surface area of boxes in the image for M_2 (i.e. $[B, 2(B + B_h \times T_w)]$), we can immediately conclude that $p(M_2|img) > p(M_1|img)$. This means two boxes are more likely than one! Coherent Bayesian reasoning does not even always embody Occam's razor.

Note that $B' > A$ is possible, if unlikely, as B' refers to the total surface area of all boxes. If there were actually ≈ 40 boxes all lined up one behind the other in figure 3 then this would be the case. This stacking factor introduces a further slight bias against M_1 and in favour of $M_{i,i>1}$, but as we are just comparing M_1 and M_2 we will not consider this fact.

Finally we return to the issue of colour. The fact that the two box-parts visible are the same colour does introduce a further bias in favour of M_1 (if we also assume that each box is mono-coloured). However it should be clear that the relative width of the tree trunk and boxes will affect the relative probabilities.

If forced to include colour then an image with a thicker tree trunk or taller and narrower visible box area would still order M_2 before M_1. A similar argument can be made for falling boxes and boxes of different height. As claimed above and explained here, using this representation we would always be able to construct an image in which some $M_{i,i>1}$ was more probable than M_1, even though intuitively it might appear that there was only one box. Bayesian reasoning does not necessarily automatically embody OR.

4.2.4 Hyper-parameterisation?

An immediate objection to the representation discussed in 4.2.3 is that it isn't as good (for some definition of "good") as that used in 4.2.2. In making such a claim one must be extremely careful. It is very easy to implicitly assume (or automatically embody) Occam's razor. Certainly if such an assumption is

made one cannot use the result as a justification for OR or general automatic embodiment, else one begs the question. Embodiment and justification are, of course, different things. Simplicity's dependence on the representation being used is discussed in greater depth in [17].

An apparently more robust objection would be that we might make some principled choice over representations by hyper-parameterising them. In a sense this is true and we could. However it does not help. This is because such a hyper-parameterisation depends on our assumptions. If we implicitly or explicitly assumed OR then we would choose the representation which gave us simplicity, if we did not then we would not. We do not rule out some form of bootstrapping which justifies Occam's razor but it is important to remember just how difficult this task is. There is a vast gulf between objective and subjective knowledge.

5 Conclusion and Future Work

This section summarises the results of the investigation and explores the consequences for model scoring measures. The analysis (section 3, 4.1 and 4.2) makes clear that the argument does not justify OR. Instead, it very subtly begs the question. The claim that coherent inference using Bayesian reasoning automatically embodies it fails for the same reasons. As Mackay acknowledges, "there is no such thing as inference or prediction without assumptions"[19, p. 345]. Occam's razor is one of those assumptions.

5.1 Defining Simplicity

The analysis suggests that we need a more comprehensive definition of simplicity than the size of the parameter space. It is not clear that there is any simple and general definition. One possibility might involve considering both the number and also the *robustness* of the parameters. By robustness we mean the level of precision which the parameters need to be specified with and how much the predictive accuracy of the model varies as the parameters are varied. A model with 3 parameters which was approximately as accurate for small variations in any of the parameters could be considered simpler than a model with only 1 parameter which needed it to be specified very exactly.

Another factor which could also be considered is the extent to which each parameter contributes. Consider the family of predictive models represented by $x \rightarrow asin(bx)$. Because most values of a can cause the predicted sequence to vary over a much greater range than b we might conclude that a contributes

more to the complexity than b and therefore that it is more important for a to be robust than b. This question bears further investigation.

Similarly, properties of the graph itself (such as its density, girth, connectedness and so forth) could also be incorporated into the model scoring measure. Investigating these other measures of complexity and their utility for Bayesian network inference is an area of research that the authors hope to pursue.

A *Minimum description length*[6, 7, 15] encoding which adaptively considered the precision of the parameters as well as the number could provide a more comprehensive representation of OR. Note that such a representation would not be a proof of Occam's razor; it would only be another representation which could sometimes automatically embody it.

In addition, it could be argued that such a representation assumes the parameter encoding (e.g. IEEE) and that any encoding will arbitrarily represent some numbers more compactly. For example, although a simple base-10 encoding of numbers represent tenths very efficiently it also encodes fractions of 9 inefficiently and with limited accuracy. An MDL algorithm using such a representation will optimise goodness-of-fit versus closeness-to-fraction-often, and not goodness-of-fit versus simplicity.

5.2 Model Scoring Measures

When considering the problem of choosing a model scoring measure, the analysis in this paper has several implications. Firstly, it stresses that the *Bayesian information criterion*[22], the *Akaike information criterion*[1, 5], the *Bayesian Nonparametric heteroscedastic Regression Criteria*[8, 9] and other model scoring methods are approximations. The tradeoffs they make for simplicity while models are being searched over is not a principled Bayesian trade offs and are heuristic in nature. The nature of these tradeoffs for model based science have been discussed in more detail in, e.g., [16] and [20]. Which model scoring measure is best is dependent on the problem, and it is always possible to create a pathological counter example for each of these heuristic measures.

Furthermore, the BIC, AIC and BNRC all trade off model accuracy with the size of the parameter space. As just discussed in 5.1, the number of parameters is not always the best measure of complexity. Development of new model scoring measures which consider other dimensions holds promise as an area of future research.

6 Acknowledgements

This research is supported by the CSCUK. The authors gratefully thank John Matthewson (ANU) for his philosophical insights, Tony Vignaux (VUW) and Yuichi Hiroshe (VUW) for their statistical assistance and Peter Andreae (VUW) for his advice. We would also like to thank people in Elvis and the Programming Language Theory research groups at VUW for their comments and feedback.

References

[1] Hirogu Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *2nd International Symposium on Information Theory*, pages 267–281, 1973.

[2] David Armstrong. Reply to van Fraassen. *Australasian Journal of Philosophy*, 66:224–229, 1988.

[3] James Berger. The case for objective bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006.

[4] Albert Einstein. *The World As I See It*. Citadel Press, 1998.

[5] Malcolm Forster and Elliot Sober. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, 45(1):1–35, 1994.

[6] Peter Grünwald. The minimum description length principle and non-deductive inference. In Peter Flach, editor, *Proceedings of the IJCAI Workshop on Abduction and Induction in AI, Japan*, 1997.

[7] Peter Grünwald. A tutorial introduction to the minimum description length principle. *CoRR*, math.ST/0406077, 2004. informal publication.

[8] S. Imoto, T. Goto, and S. Miyano. Estimation of genetic networks and functional structures between genes by using bayesian networks and nonparametric regression. In *Pacific Symposium on Biocomputing*, volume 7, pages 175–186, 2002.

[9] Seiya Imoto, SunYong Kim, Takao Goto, Sachiyo Aburatani, Kousuke Tashiro, Satoru Kuhara, and Satoru Miyano. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *Journal of Bioinformatics and Computational Biology*, 1(2):231–252, 2003.

[10] Edwin Thompson Jaynes. The well posed problem. *Foundations of Physics*, 3:477–493, 1973.

[11] Edwin Thompson Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[12] William H. Jefferys and James O. Berger. Sharpening ockham's razor on a bayesian strop. Technical Report 91–44C, Department of Statistics, Purdue University, August 1991.

[13] Jerrold J. Katz. *Realistic Rationalism*. MIT Press, 1998.

[14] Harri Lähdesmäki, Ilya Shmulevich, and Olli Yli-Harja. On learning gene regulatory networks under the boolean network model. *Machine Learning*, 52(1–2):147–167, 2003.

[15] Wai Lam and Fahiem Bacchus. Learning Bayesian Belief Networks: An Approach Based on the MDL Principle. In *Computational Intelligence*, volume 10, pages 269–293, 1994.

[16] Richard Levins. The strategy of model building in population biology. *American Scientist*, 54(4):421–431, 1966.

[17] Aidan Lyon. Gruesome simplicity: A guide to truth. In *ANU-Sydney-Kyoto Probability Workshop*, June 2007.

[18] David J. C. MacKay. Bayesian interpolation. In C.R. Smith, G.J. Erickson, and P.O. Neudorfer, editors, *Maximum Entropy and Bayesian Methods, Seattle 1991*, pages 39–66, Dordrecht, 1992. Kluwer.

[19] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003. Available from `http://www.inference.phy.cam.ac.uk/mackay/itila/`.

[20] Steven Hecht Orzack and Elliot Sober. A critical assement of Levin's The strategy of model building in population biology (1966). *The Quarterly Review of Biology*, 68(4):533–546, December 1993.

[21] Karl R. Popper. *The Logic of Scientific Discovery*. Hutchinson, London, 1968.

[22] Gideo Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[23] D. S. Sivia. *Data Analysis: A Bayesian Tutorial*. Clarendon Press, Oxford, 1996.

[24] J. J. C. Smart. Laws of nature and cosmic coincidence. *Philosophical Quarterly*, 35:272–280, 1985.

[25] Elliot Sober. *Simplicity*. Clarendon Press, Oxford, 1975.

[26] Bas C. van Fraassen. *Laws and Symmetry*. Oxford University Press, 1989.