# A note on Occam's razor

David Barber
Department of Computer Science
University College London

May 29, 2013

## 1 Occam's Razor

This is a note to help explain why Bayesian methods automatically penalise over-complex models. A similar description appeared in draft versions of [1].
For a model $M$ with parameters $\boldsymbol{\theta}$, the likelihood of generating the data is given by

$$p(\mathcal{D}|M) = \int_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M) \tag{1}$$

To simplify the argument, we place flat priors on the parameter space,

$$p(\boldsymbol{\theta}|M) = 1/V \tag{2}$$

where $V$ is the volume (number of states in the discrete case) of the parameter space. Then

$$p(\mathcal{D}|M) = \frac{\int_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta}, M)}{V} \tag{3}$$

We can approximate the likelihood $p(\mathcal{D}|\boldsymbol{\theta}, M)$ by thresholding it at a value $\epsilon$:

$$p(\mathcal{D}|\boldsymbol{\theta}, M) \approx \begin{cases} L^* & p(\mathcal{D}|\boldsymbol{\theta}, M) \geq \epsilon \\ 0 & p(\mathcal{D}|\boldsymbol{\theta}, M) < \epsilon \end{cases} \tag{4}$$

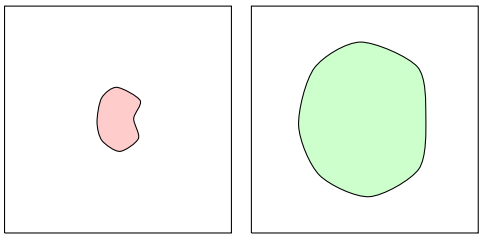
That is, when the likelihood is appreciable (bigger than $\epsilon$) we give it value $L^*$, otherwise we give the value 0. Then

$$p(\mathcal{D}|M) = L^* \frac{V^\epsilon}{V}, \qquad V^\epsilon \equiv \int_{\boldsymbol{\theta}: p(\mathcal{D}|\boldsymbol{\theta}, M) \geq \epsilon} 1 \tag{5}$$

One can then interpret the model likelihood $p(\mathcal{D}|M)$ as approximately the high likelihood value $L^*$ multiplied by the fraction of the parameter volume for which the likelihood is high.

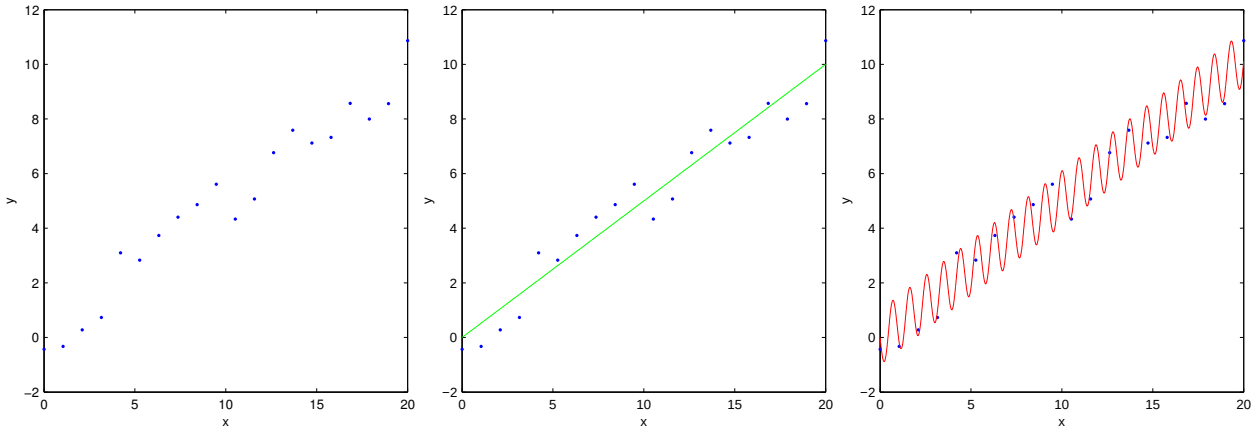
Consider two models $M_{\text{simple}}$ and $M_{\text{complex}}$ with corresponding parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Then, for flat parameter priors, we can approximate

$$p(\mathcal{D}|M_{\text{simple}}) = L^*_{\text{simple}} \frac{V^\epsilon_{\text{simple}}}{V_{\text{simple}}}, \qquad p(\mathcal{D}|M_{\text{complex}}) = L^*_{\text{complex}} \frac{V^\epsilon_{\text{complex}}}{V_{\text{complex}}} \tag{6}$$



(a) (b)

Figure 1: **(a)**: The likelihood $p(\mathcal{D}|\boldsymbol{\theta}, M_{\text{complex}})$ has a higher maximum value than then simpler model, but the likelihood drops quickly as we move away from regions of high likelihood. **(b)**: The likelihood $p(\mathcal{D}|\boldsymbol{\theta}, M_{\text{simple}})$ has a lower maximum value than then complex model, but the likelihood changes less quickly as we move away from regions of high likelihood. The corresponding fraction of parameter space in which the model fits well can then be higher for the simpler model.




(a) (b) (c)

Figure 2: **(a)**: Data for which we wish to fit a regression model. **(b)**: The best 'simple' model fit $y = ax$ has maximum likelihood $1.65 \times 10^{-10}$. **(c)**: The best 'complex' model fit $y = ax + \cos(bx)$ has maximum likelihood $3.36 \times 10^{-10}$. Whilst the complex model has a higher likelihood, it is actually the less suitable model according to the Bayes factor (see text), and the simpler model is to be preferred.

At this point it is useful to reflect on what constitutes a 'complex' model. This can be characterised by a relative flexibility in the data generating process, compared to a 'simple' model. Consequently, this means that as we move in the space of the complex model, we will generate very different datasets compared to our given dataset $\mathcal{D}$. This parameter sensitivity means that the likelihood of generating the observed data $\mathcal{D}$ will typically drop dramatically as we move away from regions of parameter space in which the model fits well. Hence for a simple model $M_{\text{simple}}$ that has a similar maximum likelihood to a complex model, $L^*_{\text{simple}} \approx L^*_{\text{complex}}$, typically the fraction of the parameter space in which the likelihood is appreciable will be smaller for the complex model than the simple model, meaning that $p(\mathcal{D}|M_{\text{simple}}) > p(\mathcal{D}|M_{\text{complex}})$, see fig(1). If we have no prior preference for either model $p(M_{\text{simple}}) = p(M_{\text{complex}})$ the Bayes factor is given by

$$\frac{p(M_{\text{simple}}|\mathcal{D})}{p(M_{\text{complex}}|\mathcal{D})} = \frac{p(\mathcal{D}|M_{\text{simple}})}{p(\mathcal{D}|M_{\text{complex}})} \tag{7}$$

and the Bayes factor will typically prefer the simpler of two competing models with similar maximum likelihood values.
This demonstrates the *Occam's razor* effect of Bayesian model inference which penalises models which are over complex.

## 1.1 A regression example

As an example of this effect, consider the regression problem in fig(2) for which we consider the following two models of the 'clean' underlying regression function:

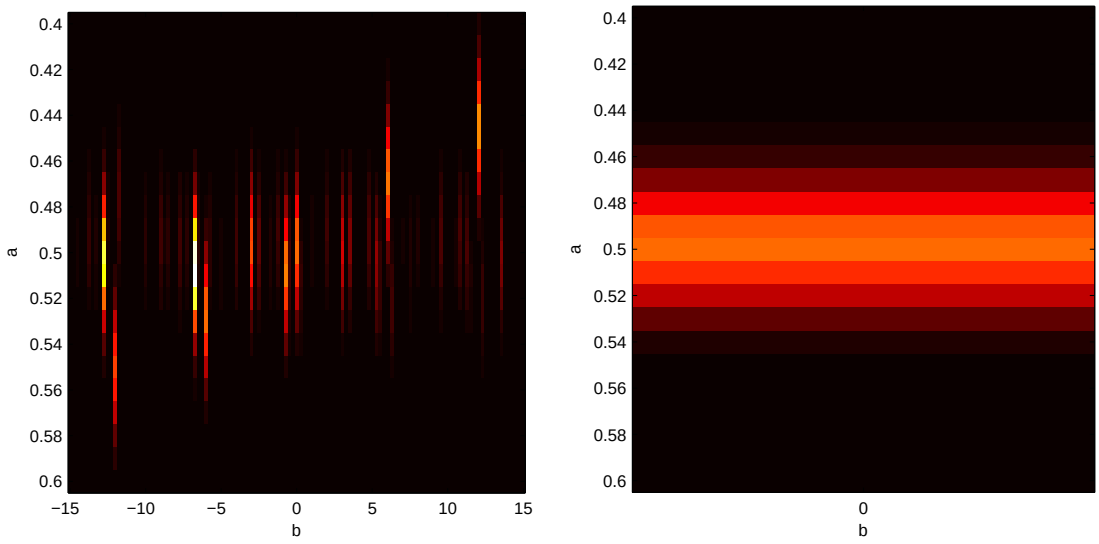
$$M_{\text{simple}} : \; y_0 = ax \tag{8}$$
$$M_{\text{complex}} : \; y_0 = ax + \cos(bx) \tag{9}$$

To account for noise in the observations, $y = y_0 + \epsilon$, $\epsilon \sim \mathcal{N}\left(\epsilon|0, \sigma^2\right)$, we use the model

$$p(y|x, a, M_{\text{simple}}) = \mathcal{N}\left(y|ax, \sigma^2\right) \tag{10}$$

The likelihood of a collection $\mathcal{Y}$ of independent observations for a set of inputs $\mathcal{X}$ is then

$$p(\mathcal{Y}|\mathcal{X}, M_{\text{simple}}) = \int_a p(a|M_{\text{simple}}) \prod_{n=1}^{N} p(y^n|x^n, a, M_{\text{simple}}) \tag{11}$$




(a) (b)

Figure 3: Likelihood plots for the problem in fig(2). **(a)**: The likelihood $p(\mathcal{Y}|\mathcal{X}, a, b, M_{\text{complex}})$ **(b)**: The likelihood $p(\mathcal{Y}|\mathcal{X}, a, M_{\text{simple}})$ plotted on the same scale as (a). This displays the characteristic of a 'complex' model in the likelihood drops dramatically as we move only a small distance in parameter space from a point with high likelihood. Whilst the maximum likelihood of the complex model is higher than the simpler model, the volume of parameter space in which the complex model fits the data well is smaller than for the simpler model, giving rise to $p(\mathcal{Y}|\mathcal{X}, M_{\text{simple}}) > p(\mathcal{Y}|\mathcal{X}, M_{\text{complex}})$.

Similarly,

$$p(y|x, a, b, M_{\text{complex}}) = \mathcal{N}\left(y|ax + \cos(bx), \sigma^2\right) \tag{12}$$

and

$$p(\mathcal{Y}|\mathcal{X}, M_{\text{complex}}) = \int_{a,b} p(a, b|M_{\text{complex}}) \prod_{n=1}^{N} p(y^n|x^n, a, b, M_{\text{complex}}) \tag{13}$$

Using a discrete set of 21 values for $a$, evenly spaced from 0.4 to 0.6, and 121 discrete values for $b$, evenly spaced from -15 to 15, we can compute the corresponding likelihoods $p(\mathcal{Y}|\mathcal{X}, a, M_{\text{simple}})$ and $p(\mathcal{Y}|\mathcal{X}, a, b, M_{\text{complex}})$. For this data, the maximum likelihoods are

$$\max_a p(\mathcal{Y}|\mathcal{X}, a, M_{\text{simple}}) = 1.65 \times 10^{-10}, \qquad \max_{a,b} p(\mathcal{Y}|\mathcal{X}, a, b, M_{\text{complex}}) = 3.36 \times 10^{-10} \tag{14}$$

so that the more complex model has a higher maximum likelihood. However, as we can see in fig(3), the fraction of the parameter space in which the complex model fits the data well is relatively small. Using a flat prior for the parameter spaces of both models, we obtain

$$p(\mathcal{Y}|\mathcal{X}, M_{\text{simple}}) = 3.88 \times 10^{-11}, \qquad p(\mathcal{Y}|\mathcal{X}, M_{\text{complex}}) = 5.28 \times 10^{-12} \tag{15}$$

Whilst the more complex model has a higher maximum likelihood value by a factor 2, it is roughly 7 times less likely to be the correct model, compared to the simpler model.

## References

[1] D. Barber. *Bayesian Reasoning and Machine Learning.* Cambridge University Press, 2012.