Occam's Razor and Bayesian Complexity Penalisation

David Barber

University College London

Occam's Razor

- William of Occam: "It is futile to do with more things that which can be done with fewer"
- One of the strengths of the Bayesian framework is that presents a conceptually clear framework for comparing competing models:
  For a model M with parameters $\boldsymbol{\theta}$

  $$p(M|\mathcal{D}) = \frac{p(\mathcal{D}|M)p(M)}{p(\mathcal{D})}$$

  where we obtain $p(\mathcal{D}|M)$ by integrating over the parameter space

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

- Bayesian model inference automatically penalises models which are over complex.
- How this effect occurs can perhaps be best explained using some simple assumptions.

Bayesian Modelling

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

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

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

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

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

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

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

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

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

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

Simple versus Complex model

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

$$p(\mathcal{D}|M_{\text{simple}}) = L^*_{\text{simple}} \frac{V^{\epsilon}_{\text{simple}}}{V_{\text{simple}}}$$

$$p(\mathcal{D}|M_{\text{complex}}) = L^*_{\text{complex}} \frac{V^{\epsilon}_{\text{complex}}}{V_{\text{complex}}}$$

- Which model 'best' fits the data? One solution is to use $p(M|\mathcal{D}) \propto p(\mathcal{D}|M)p(M)$.

- Complex 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.
- Simple Hence for a simple model M_{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 $p(\mathcal{D}|M_{\text{simple}}) > p(\mathcal{D}|M_{\text{complex}})$.

If we have no prior preference for either model $p(M_{\text{simple}}) = p(M_{\text{complex}})$ the Bayes factor is given by

$$\frac{p(M_{\text{simple}}|\mathcal{D})}{p(M_{\text{complex}}|\mathcal{D})} = \frac{p(\mathcal{D}|M_{\text{simple}})}{p(\mathcal{D}|M_{\text{complex}})} \tag{1}$$

and the Bayes factor will typically prefer the simpler of two competing models with similar maximum likelihood values.

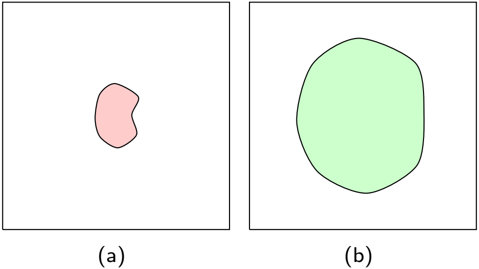


Figure : **(a)**: The likelihood $p(\mathcal{D}|\boldsymbol{\theta}, M_{\text{complex}})$ has a higher maximum value than then simpler model, but the likelihood drops quickly as we move away from regions of high likelihood. **(b)**: The likelihood $p(\mathcal{D}|\boldsymbol{\theta}, M_{\text{simple}})$ 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 regression example

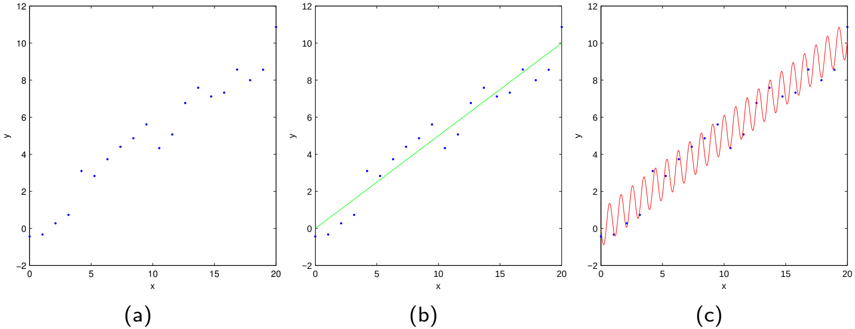


Figure : **(a)**: Data for which we wish to fit a regression model. **(b)**: The best ‘simple’ model fit $y = ax$ has maximum likelihood 1.65×10^{-10}. **(c)**: The best ‘complex’ model fit $y = ax + \cos(bx)$ has maximum likelihood 3.36×10^{-10}. Whilst the complex model has a higher likelihood, it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

A regression example

Consider the following two models of the ‘clean’ underlying regression function:

$$\begin{aligned} M_{\text{simple}} &: \; y_0 = ax \\ M_{\text{complex}} &: \; y_0 = ax + \cos(bx) \end{aligned}$$

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

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

The likelihood of independent observations $\mathcal{Y}$ given inputs $\mathcal{X}$ is

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

Similarly,

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

and

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

- For this data, the maximum likelihoods are:

$$\max_a p(\mathcal{Y}|\mathcal{X}, a, M_{\textsf{simple}}) = 1.65 \times 10^{-10}$$

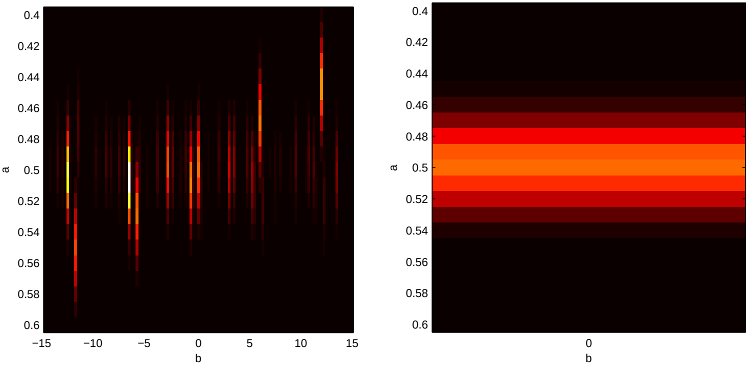
$$\max_{a,b} p(\mathcal{Y}|\mathcal{X}, a, b, M_{\textsf{complex}}) = 3.36 \times 10^{-10}$$

  so that the more complex model has a higher maximum likelihood.
- Using a flat prior for the parameter spaces of both models, we obtain

$$p(\mathcal{Y}|\mathcal{X}, M_{\textsf{simple}}) = 3.88 \times 10^{-11}$$

$$p(\mathcal{Y}|\mathcal{X}, M_{\textsf{complex}}) = 5.28 \times 10^{-12}$$

  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.



(a)

(b)

Figure : **(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 $p(\mathcal{Y}|\mathcal{X}, M_{\text{simple}}) > p(\mathcal{Y}|\mathcal{X}, M_{\text{complex}})$.

Fitting models to continuous data

A model class

Consider an additive set of periodic functions

$$y^0 = w_0 + w_1 \cos(x) + w_2 \cos(2x) + \ldots + w_K \cos(Kx)$$

This can be conveniently written in vector form

$$y^0 = \mathbf{w}^\mathsf{T} \boldsymbol{\phi}(x)$$

where

$$\boldsymbol{\phi}(x) = (1, \cos(x), \cos(2x), \ldots, \cos(Kx))^\mathsf{T}$$

Data

We are given a set of data $\mathcal{D} = \{(x^n, y^n), n = 1, \ldots, N\}$ drawn from this distribution, where y is the clean $y^0(x)$ corrupted with additive zero mean Gaussian noise with variance σ^2,

$$y^n = y^0(x^n) + \epsilon^n, \quad \epsilon^n \sim \mathcal{N}\left(\epsilon^n | 0, \sigma^2\right)$$

The task

How many components K should we use?

Fitting models to continuous data

The posterior

Assuming i.i.d. data,

$$\begin{aligned} p(K|\mathcal{D}) &= \frac{p(\mathcal{D}|K)p(K)}{p(\mathcal{D})} \\ &= \frac{p(K)\prod_n p(x^n)}{p(\mathcal{D})} p(y^1,\ldots,y^N|x^1,\ldots,x^N,K) \end{aligned}$$

We will assume $p(K) = \text{const.}$.

The likelihood

$$p(y^1,\ldots,y^N|x^1,\ldots,x^N,K) = \int_{\mathbf{w}} p(\mathbf{w}|K) \prod_{n=1}^{N} p(y^n|x^n,\mathbf{w},K)$$

Evaluating the likelihood

$$p(y^1, \ldots, y^N | x^1, \ldots, x^N, K) = \int_{\mathbf{w}} p(\mathbf{w}|K) \prod_{n=1}^{N} p(y^n | x^n, \mathbf{w}, K)$$

For $p(\mathbf{w}|K) = \mathcal{N}\left(\mathbf{w}|0, \mathbf{I}_K/\alpha\right)$, the integrand is a Gaussian in $\mathbf{w}$ for which it is straightforward to evaluate the integral,

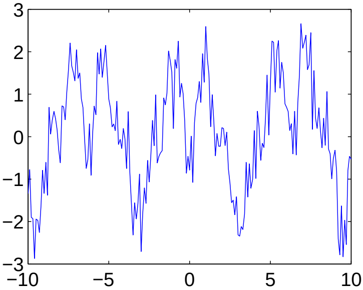
$$N \log\left(2\pi\sigma^2\right) - \sum_{n=1}^{N} \frac{\left(y^n\right)^2}{\sigma^2} + \mathbf{b}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{b} - \log\det\left(2\pi\mathbf{A}\right) + K\log\left(2\pi\alpha\right)$$

where

$$\mathbf{A} \equiv \alpha\mathbf{I} + \frac{1}{\sigma^2}\sum_{n=1}^{N} \boldsymbol{\phi}(x^n)\boldsymbol{\phi}^{\mathsf{T}}(x^n), \qquad \mathbf{b} \equiv \frac{1}{\sigma^2}\sum_{n=1}^{N} y^n \boldsymbol{\phi}(x^n)$$

Example

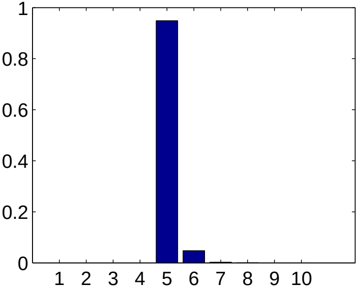
The data



Setting $\alpha = 1$ and $\sigma = 0.5$, we sampled some data from a model with $K = 5$ components.

Example

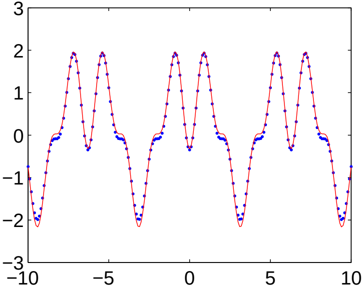
The posterior $p(K|\mathcal{D})$



We assume that we know the correct noise level σ. This is sharply peaked at $K = 5$, which is the 'correct' value used to generate the data.

Example

The reconstruction



The clean reconstructions for $K = 5$ are plotted. The reconstruction of the data using $\langle \mathbf{w} \rangle^{\mathsf{T}} \boldsymbol{\phi}(x)$ where $\langle \mathbf{w} \rangle$ is the mean posterior vector of the optimal dimensional model $p(\mathbf{w}|\mathcal{D}, K = 5)$. Plotted in red is the mean reconstruction. Plotted in dots is the true underlying clean data.

Approximating the Model Likelihood

The model likelihood

For a model with continuous parameter vector $\boldsymbol{\theta}$, $\dim(\boldsymbol{\theta}) = K$ and data $\mathcal{D}$, the model likelihood is

$$p(\mathcal{D}|M) = \int p(\mathcal{D}|\boldsymbol{\theta}, M) p(\boldsymbol{\theta}|M) d\boldsymbol{\theta}$$

Need for approximations

For a generic expression

$$p(\mathcal{D}|\boldsymbol{\theta}, M) p(\boldsymbol{\theta}|M) = e^{-f(\boldsymbol{\theta})}$$

unless f is of a particularly simple form (quadratic in $\boldsymbol{\theta}$ for example), one cannot compute the integral and approximations are required.

Laplace's method

A simple approximation is given by Laplace's method,

$$\log p(\mathcal{D}|M) \approx -f(\boldsymbol{\theta}^*) + \frac{1}{2}\log\det\left(2\pi\mathbf{H}^{-1}\right)$$

where $\boldsymbol{\theta}^*$ is the MAP solution

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ p(\mathcal{D}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)$$

and $\mathbf{H}$ is the Hessian of $f(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$.

---

i.i.d. data

For i.i.d. data $\mathcal{D} = \left\{x^1, \ldots, x^N\right\}$

$$p(\mathcal{D}|M) = \int p(\boldsymbol{\theta}|M)\prod_{n=1}^{N} p(x^n|\boldsymbol{\theta}, M)d\boldsymbol{\theta}$$

In this case Laplace's method computes the optimum of the function

$$-f(\boldsymbol{\theta}) = \log p(\boldsymbol{\theta}|M) + \sum_{n=1}^{N}\log p(x^n|\boldsymbol{\theta}, M)$$

Bayes Information Criterion

For i.i.d. data the Hessian scales with the number of training examples, N, and a crude approximation is to set $\mathbf{H} \approx N\mathbf{I}_K$ where $K = \dim \boldsymbol{\theta}$. In this case one may take as a model comparison procedure the function

$$\log p(\mathcal{D}|M) \approx \log p(\mathcal{D}|\boldsymbol{\theta}^*, M) + \log p(\boldsymbol{\theta}^*|M) + \frac{K}{2}\log 2\pi - \frac{K}{2}\log N$$

For a simple prior that penalises the length of the parameter vector, $p(\boldsymbol{\theta}|M) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \mathbf{I})$, the above reduces to

$$\log p(\mathcal{D}|M) \approx \log p(\mathcal{D}|\boldsymbol{\theta}^*, M) - \frac{1}{2}(\boldsymbol{\theta}^*)^{\mathsf{T}}\boldsymbol{\theta}^* - \frac{K}{2}\log N$$

The BIC approximates ignores the penalty term, giving

$$BIC = \log p(\mathcal{D}|\boldsymbol{\theta}^*, M) - \frac{K}{2}\log N$$

The term $-\frac{K}{2}\log N$ penalises model complexity. In general, the Laplace approximation is to be preferred to the BIC criterion since it more correctly accounts for the uncertainty in the posterior parameter estimate.