Model comparison

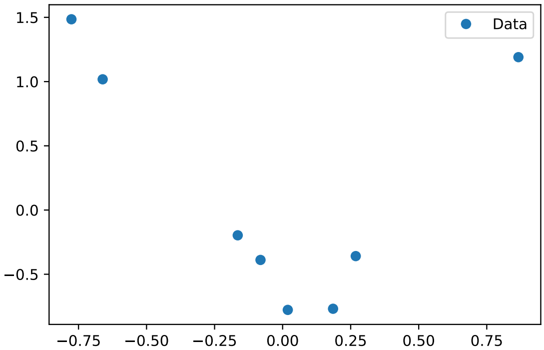
Brooks Paige

COMP0171

Model selection and comparison

One thing we haven't talked about too much is model selection.

A prototypical example, which we mentioned briefly during the first week: what degree polynomial should I fit to this data?



The evidence

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

The evidence

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

The evidence

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} = \frac{p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})}$$

The evidence

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} = \frac{p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})}$$

We use the **evidence**, $p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}, \boldsymbol{\theta}|\mathcal{M})\mathrm{d}\boldsymbol{\theta}$, for model comparison.

- Also known as the **marginal likelihood** as it describes the probability of the data with parameters marginalized out
- Usually the normalizing constant of a Bayesian model

The evidence

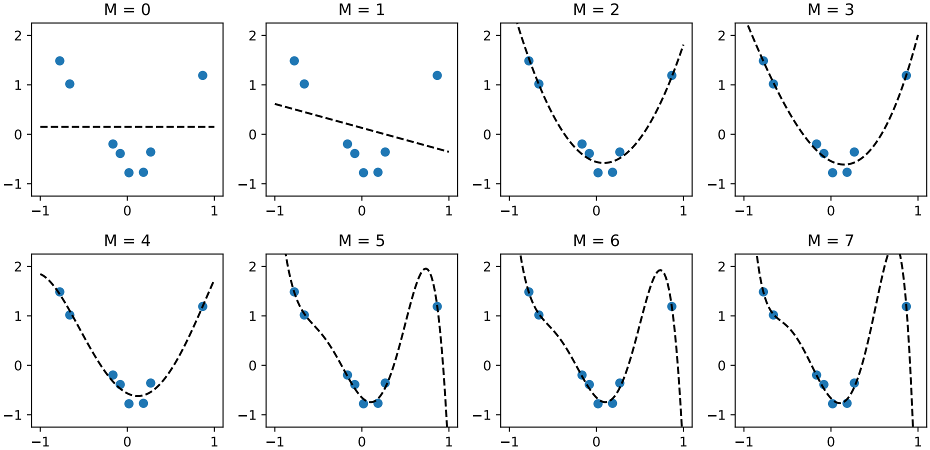
$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} = \frac{p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})}$$

We use the **evidence**, $p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}, \boldsymbol{\theta}|\mathcal{M})\mathrm{d}\boldsymbol{\theta}$, for model comparison.

- Also known as the **marginal likelihood** as it describes the probability of the data with parameters marginalized out
- Usually the normalizing constant of a Bayesian model

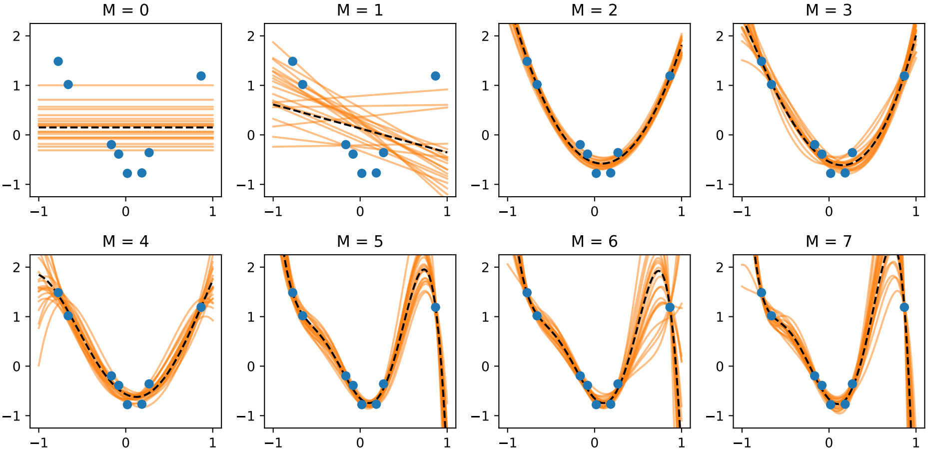
Given two possible model families, e.g. $\mathcal{M}_0$ and $\mathcal{M}_1$, we can compare them using the marginal likelihood.

MAP estimation



More complex models: lower training error

Bayesian estimation

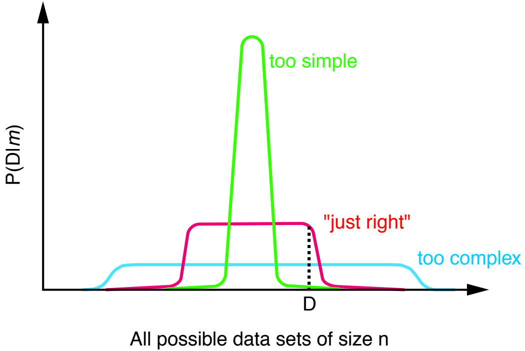


More complex models: more possible “explanations” for the data

Bayesian Occam's Razor

William of Occam, **Occam's Razor**: "Entities should not be multiplied beyond necessity"

In general, we would prefer simple models over complex models, when either would suffice.



Models that are **too simple** are unlikely to have generated the dataset.

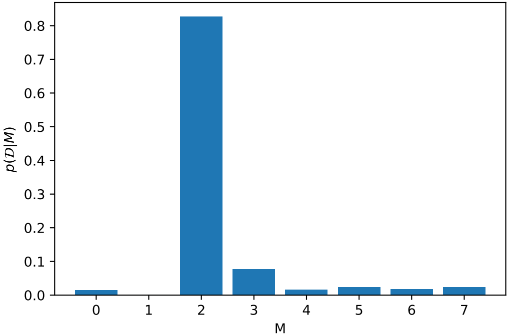
Models that are **too complex** could have generated many possible datasets, so producing this particular one at random is unlikely as well.

Occam figure: Zoubin Ghahramani

Model evidence

For each model M, we can compute $p(\mathcal{D}|M)$.

- A quadratic model looks promising!
- A cubic model is plausible
- Higher-order than that is very unlikely



Model comparison as inference

Given two possible model families, e.g. $\mathcal{M}_0$ and $\mathcal{M}_1$, we can compare them using the marginal likelihood.

Bayes' rule over "models" is just Bayes' rule:

$$p(\mathcal{M}_0|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_0)p(\mathcal{M}_0)}{\sum_{i=0}^{1} p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)}$$

$$p(\mathcal{M}_1|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_1)p(\mathcal{M}_1)}{\sum_{i=0}^{1} p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)}$$

Approximating the model evidence

Unfortunately, it's hard!

- This was a linear Gaussian regression model, so it has a closed form
- Next slide: estimating it with the Laplace approximation

Approximating the model evidence

Unfortunately, it's hard!

- This was a linear Gaussian regression model, so it has a closed form
- Next slide: estimating it with the Laplace approximation
- Q: could we use the ELBO as a proxy? Maybe! (it depends on how tight the lower bound is. . .)

Approximating the model evidence

Unfortunately, it's hard!

- This was a linear Gaussian regression model, so it has a closed form
- Next slide: estimating it with the Laplace approximation
- Q: could we use the ELBO as a proxy? Maybe! (it depends on how tight the lower bound is. . .)
- No easy / foolproof way to do it from MCMC samples

Approximating the model evidence

Unfortunately, it's hard!

- This was a linear Gaussian regression model, so it has a closed form
- Next slide: estimating it with the Laplace approximation
- Q: could we use the ELBO as a proxy? Maybe! (it depends on how tight the lower bound is...)
- No easy / foolproof way to do it from MCMC samples
- Note it can be estimated using importance sampling (not covered here)

Laplace approximation for marginal likelihoods

The **Laplace approximation** we covered last week approximates posterior distributions with Gaussians. (Fortunately, we know how to normalize Gaussians!)

To estimate the normalizing constant instead of the posterior, start by taking the same Taylor approximation at the mode $\boldsymbol{\theta}^\star \in \mathbb{R}^D$,

$$\log p(\mathcal{D}, \boldsymbol{\theta}) \approx \log p(\mathcal{D}, \boldsymbol{\theta}^\star) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)$$

Laplace approximation for marginal likelihoods

The **Laplace approximation** we covered last week approximates posterior distributions with Gaussians. (Fortunately, we know how to normalize Gaussians!)

To estimate the normalizing constant instead of the posterior, start by taking the same Taylor approximation at the mode $\boldsymbol{\theta}^\star \in \mathbb{R}^D$,

$$\log p(\mathcal{D}, \boldsymbol{\theta}) \approx \log p(\mathcal{D}, \boldsymbol{\theta}^\star) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)$$

$$p(\mathcal{D}, \boldsymbol{\theta}) \approx p(\mathcal{D}|\boldsymbol{\theta}^\star)p(\boldsymbol{\theta}^\star) \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \boldsymbol{\Lambda}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\right\}$$

where $\boldsymbol{\Lambda} = -\mathbf{H} = -\nabla\nabla \log p(\mathcal{D}, \boldsymbol{\theta}) = -\nabla\nabla \log p(\boldsymbol{\theta}|\mathcal{D})$.

Laplace approximation for marginal likelihoods

The **Laplace approximation** we covered last week approximates posterior distributions with Gaussians. (Fortunately, we know how to normalize Gaussians!)

To estimate the normalizing constant instead of the posterior, start by taking the same Taylor approximation at the mode $\boldsymbol{\theta}^\star \in \mathbb{R}^D$,

$$\log p(\mathcal{D}, \boldsymbol{\theta}) \approx \log p(\mathcal{D}, \boldsymbol{\theta}^\star) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)$$

$$p(\mathcal{D}, \boldsymbol{\theta}) \approx p(\mathcal{D}|\boldsymbol{\theta}^\star)p(\boldsymbol{\theta}^\star) \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \boldsymbol{\Lambda}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\right\}$$

where $\boldsymbol{\Lambda} = -\mathbf{H} = -\nabla\nabla \log p(\mathcal{D}, \boldsymbol{\theta}) = -\nabla\nabla \log p(\boldsymbol{\theta}|\mathcal{D})$.

This suggests we can (approximately) normalize the distribution by normalizing the approximation, as

$$p(\mathcal{D}) = \int p(\mathcal{D}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.$$

Laplace approximation for marginal likelihoods

First, note that for $\boldsymbol{\theta} \in \mathbb{R}^D$

$$\int \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \boldsymbol{\Lambda}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\right\} \mathrm{d}\boldsymbol{\theta} = \frac{(2\pi)^{D/2}}{|\boldsymbol{\Lambda}|^{1/2}}.$$

(This is exactly the normalization constant of a multivariate Gaussian.)

Laplace approximation for marginal likelihoods

First, note that for $\boldsymbol{\theta} \in \mathbb{R}^D$

$$\int \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \boldsymbol{\Lambda}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\right\} \mathrm{d}\boldsymbol{\theta} = \frac{(2\pi)^{D/2}}{|\boldsymbol{\Lambda}|^{1/2}}.$$

(This is exactly the normalization constant of a multivariate Gaussian.)

$$\int p(\mathcal{D}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \approx \int p(\mathcal{D}|\boldsymbol{\theta}^\star) p(\boldsymbol{\theta}^\star) \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \boldsymbol{\Lambda}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\right\} \mathrm{d}\boldsymbol{\theta}$$

Laplace approximation for marginal likelihoods

First, note that for $\boldsymbol{\theta} \in \mathbb{R}^D$

$$\int \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)^\top \boldsymbol{\Lambda}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)\right\} \mathrm{d}\boldsymbol{\theta} = \frac{(2\pi)^{D/2}}{|\boldsymbol{\Lambda}|^{1/2}}.$$

(This is exactly the normalization constant of a multivariate Gaussian.)

$$\begin{aligned}\int p(\mathcal{D}, \boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} &\approx \int p(\mathcal{D}|\boldsymbol{\theta}^\star)p(\boldsymbol{\theta}^\star) \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)^\top \boldsymbol{\Lambda}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)\right\} \mathrm{d}\boldsymbol{\theta} \\ &= p(\mathcal{D}|\boldsymbol{\theta}^\star)p(\boldsymbol{\theta}^\star) \int \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)^\top \boldsymbol{\Lambda}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)\right\} \mathrm{d}\boldsymbol{\theta}\end{aligned}$$

Laplace approximation for marginal likelihoods

First, note that for $\boldsymbol{\theta} \in \mathbb{R}^D$

$$\int \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \boldsymbol{\Lambda}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\right\} \mathrm{d}\boldsymbol{\theta} = \frac{(2\pi)^{D/2}}{|\boldsymbol{\Lambda}|^{1/2}}.$$

(This is exactly the normalization constant of a multivariate Gaussian.)

$$\begin{aligned}
\int p(\mathcal{D}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} &\approx \int p(\mathcal{D}|\boldsymbol{\theta}^\star) p(\boldsymbol{\theta}^\star) \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \boldsymbol{\Lambda}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\right\} \mathrm{d}\boldsymbol{\theta} \\
&= p(\mathcal{D}|\boldsymbol{\theta}^\star) p(\boldsymbol{\theta}^\star) \int \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \boldsymbol{\Lambda}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\right\} \mathrm{d}\boldsymbol{\theta} \\
&= p(\mathcal{D}|\boldsymbol{\theta}^\star) p(\boldsymbol{\theta}^\star) \frac{(2\pi)^{D/2}}{|\boldsymbol{\Lambda}|^{1/2}}.
\end{aligned}$$

A "regularizer" for model complexity

Taking logarithms of both sides, we have the approximation

$$\log p(\mathcal{D}) \approx \log p(\mathcal{D}|\boldsymbol{\theta}^\star) + \log p(\boldsymbol{\theta}^\star) + \frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Lambda}|$$

A “regularizer” for model complexity

Taking logarithms of both sides, we have the approximation

$$\log p(\mathcal{D}) \approx \log p(\mathcal{D}|\boldsymbol{\theta}^\star) + \underbrace{\log p(\boldsymbol{\theta}^\star) + \frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Lambda}|}_{\text{“Occam factor”}}$$

A "regularizer" for model complexity

Taking logarithms of both sides, we have the approximation

$$\log p(\mathcal{D}) \approx \log p(\mathcal{D}|\boldsymbol{\theta}^\star) + \underbrace{\log p(\boldsymbol{\theta}^\star) + \frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Lambda}|}_{\text{“Occam factor”}}$$

The **Occam factor** here is a general means for penalizing model complexity. Of particular note:

- Partially, this depends on the prior. But it is not just the prior!
- It also depends on $\boldsymbol{\Lambda} = -\nabla\nabla \log p(\boldsymbol{\theta}|\mathcal{D})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star}$, the Hessian at the MAP estimate $\boldsymbol{\theta}^\star$.

A "regularizer" for model complexity

Taking logarithms of both sides, we have the approximation

$$\log p(\mathcal{D}) \approx \log p(\mathcal{D}|\boldsymbol{\theta}^\star) + \underbrace{\log p(\boldsymbol{\theta}^\star) + \frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Lambda}|}_{\text{"Occam factor"}}$$

The **Occam factor** here is a general means for penalizing model complexity. Of particular note:

- Partially, this depends on the prior. But it is not just the prior!
- It also depends on $\boldsymbol{\Lambda} = -\nabla\nabla \log p(\boldsymbol{\theta}|\mathcal{D})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star}$, the Hessian at the MAP estimate $\boldsymbol{\theta}^\star$.

Intuitively, this will prefer "broader" rather than "peaky" posteriors.