Bayesian Statistics

Rob Weston

Applied Artificial Inteligence Lab

March 18, 2021

Overview

1. Introduction
   - Bayesian Inference
   - Why Bother?
   - Bayesian vs a Frequentist Approach
   - Challenges
     - Calculating the Marginal Evidence
     - Subjective Priors
2. The Model
   - The Prior
   - The Likelihood
   - Hyper-parameters
   - Examples
3. Priors
   - The Exponential Family
   - Conjugate Priors
   - Non Informative Priors
     - Invariant Priors
     - Jefferey Priors
     - Reference Priors
4. Approximate Inference
   - Collapsing the posterior to a point
   - Laplace Approximation
   - Variational Inference
   - Sampling Approaches
5. Bayesian Model Comparison
   - Bayesian Model Comparison
   - Occams Razor
   - Approximating the Marginal Likelihood
   - Alternatives to the Marginal Likelihood

Overview

1. Introduction
   - Bayesian Inference
   - Why Bother?
   - Bayesian vs a Frequentist Approach
   - Challenges
     - Calculating the Marginal Evidence
     - Subjective Priors
2. The Model
   - The Prior
   - The Likelihood
   - Hyper-parameters
   - Examples
3. Priors
   - The Exponential Family
   - Conjugate Priors
   - Non Informative Priors
     - Invariant Priors
     - Jefferey Priors
     - Reference Priors
4. Approximate Inference
   - Collapsing the posterior to a point
   - Laplace Approximation
   - Variational Inference
   - Sampling Approaches
5. Bayesian Model Comparison
   - Bayesian Model Comparison
   - Occams Razor
   - Approximating the Marginal Likelihood
   - Alternatives to the Marginal Likelihood

Bayesian Inference

- **Aim** Estimate an *unknown* variable θ given known *observations* $\mathcal{D}$
  - θ is assumed random. It might correspond to the outcome of a *repeatable* (eg. a coin toss) or *non-repeatable* event (eg. film preference)
  - θ might be a single variable (eg. the probability of a coin toss) or a set of variables (eg. the mean and variance of a Gaussian distribution)
  - θ is normally *discrete* or *continuous* but can be something else
  - $\mathcal{D}$ in general corresponds to the set of all observed data points (eg. $\mathcal{D} = \{\boldsymbol{x}_n : \boldsymbol{x}_n \in \mathbb{R}^D\}_{n=1}^N$)
- **Approach**
  1. Assume a *prior* over the unknown quantity $p(\theta)$
  2. Choose a *likelihood* function $p(\mathcal{D}|\theta)$
  3. Infer the posterior distribution

$$\overbrace{p(\theta|\mathcal{D})}^{\text{Posterior}} = \frac{\overbrace{p(\mathcal{D}|\theta)}^{\text{Likelihood}}\overbrace{p(\theta)}^{\text{Prior}}}{\underbrace{p(\mathcal{D})}_{\text{Marginal Evidence}}} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D})p(\mathcal{D}|\theta)d\theta} = \frac{p(\mathcal{D},\theta)}{\int \underbrace{p(\mathcal{D},\theta)}_{\text{Joint}}\, d\theta} \tag{1}$$

Why Bother?

Using the posterior $p(\theta|\mathcal{D})$:

- **Uncertainty Prediction** "quantify the uncertainty in θ"

$$C_\alpha^{CI}(\mathcal{D}) = (\theta_l, \theta_u) \text{ given } P(\theta_l \leq \theta \leq \theta_u|\mathcal{D}) = 1 - \alpha$$

$$C_\alpha^{HDI}(\mathcal{D}) = \{\theta : p(\theta|\mathcal{D}) \geq p^*\} \text{ given } 1 - \alpha = \int_{\theta: p(\theta|\mathcal{D}) > p^*} p(\theta|\mathcal{D}) d\theta$$

- **Predictive Posterior** "predict a new data point"

$$p(\mathcal{D}^*|\mathcal{D}) = \int p(\mathcal{D}^*|\theta) p(\theta|\mathcal{D}) d\theta$$

- **Posterior Expectations** "make optimal decisions under uncertainty"

$$a^* = \arg\max_a \mathbb{E}_{p(\theta|\mathcal{D})}[\ell(a; \theta)]$$

Using the marginal likelihood $p(\mathcal{D})$:

- **Bayesian Model selection** "choose the best model amongst possible candidates"

$$\mathcal{M}^* = \arg\max_{\mathcal{M}}[p(\mathcal{D}|\mathcal{M})] \quad \text{or} \quad \phi^* = \arg\max_\phi [p(\mathcal{D}|\phi)]$$

Bayesian vs a Frequentist Approach

1. **Probability**
   - *A Frequentist view*
     - "The probability of an event is equal to the *frequency* with which that event occurs"
     - Most applicable when talking about repeatable experiments (eg. the probability of a coin toss)
   - *A Bayesian view*
     - "The probability of an event quantifies our *belief* that an event is likely to happen"
     - It does not necessarily correspond to the *frequency* with which that event occurs (although it can do)
     - It can be used for non-repeatable experiments (eg. Netflix film matches)
2. **Inference**
   - *A Frequentist view*
     - Assume that θ is *fixed* but *unknown* and that the data is uncertain $\mathcal{D} \sim p(\mathcal{D}|\theta)$
     - Maximum likelihood estimate: $\theta_{MLE} = \max_\theta[\log p(\mathcal{D}|\theta)]$
     - Uncertainty in θ_{MLE} is generated by uncertainty in the observed data $\mathcal{D}$ (eg. using p-values)
   - *A Bayesian view*
     - Assume that θ is *uncertain* and *unknown* $\theta \sim p(\theta)$
     - Update our belief in light of observation $\mathcal{D}$ using *bayes* rule $p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$
     - Uncertainty in θ is quantified directly from $p(\theta|\mathcal{D})$ (eg. using credible intervals)

Bayesian vs a Frequentist Approach

Pros

1. A consistent and logical framework for reasoning about *both* repeatable and non-repeatable random events
2. Allows us to assume that θ is inherently random (not possible in a frequentist approach)
3. Assuming that θ is unknown and that the data is fixed is often more natural (as the data *is* what we have and θ is unknown).
4. Can lead to more intuitive characterisations of uncertainty (eg. p-values do *not* give you the probability that θ lies in a particular interval).
5. When reasoning under uncertainty the sum rule and the product rule (and so Bayes rule) emerge as a natural consequence of common sense assumptions. Bayesian inference can therefore be seen as a natural extension of Boolean logic under uncertainty.
6. Bayesian inference naturally favours the simplest model - built in Occam's razor
7. Models of different dimensionality can be easily compared using the marginal likelihood

Cons

1. In many cases calculating the posterior distribution is intractable
2. Sensitivity of results on the prior $p(\theta)$

Challenges

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D})p(\mathcal{D}|\theta)d\theta} \tag{2}$$

Intractable Inference

Calculating the posterior distribution $p(\theta|\mathcal{D})$ involves calculating the *marginal evidence* $p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta$ which depending on our choice of model is often intractable.

In this case several options are available:

- Choose a model such that calculating $p(\theta|\mathcal{D})$ is tractable $\implies$ **Conjugate Priors**
- Approximate the posterior distribution
  - As a single point $p(\theta|\mathcal{D}) \approx \delta(\theta - \hat{\theta})$ $\implies$ **MAP estimation**
  - As a Gaussian distribution $p(\theta|\mathcal{D}) \approx \mathcal{N}(\theta|\mu, \Sigma)$ $\implies$ **Laplace Approximation**
  - As a General approximate distribution $p(\theta|\mathcal{D}) \approx q(\theta)$ $\implies$ **Variational Inference**
  - As a set of samples $p(\theta|\mathcal{D}) = \sum_i \omega_i \delta(\theta - \theta_i)$ $\implies$ **Monte Carlo**, **Histogram Approaches**

Challenges

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D})p(\mathcal{D}|\theta)d\theta} \tag{3}$$

Subjective Priors

The prior $p(\theta)$ encodes information about the likely value / properties of θ. But what do we do if we know nothing about θ?

Several solutions have been developed to overcome this:

- **Jeffreys Priors** Assume that the prior is invariant under some transformation $\phi = h(\theta)$
- **Reference Priors** Assume a prior which is maximally far from all posteriors given the data
- **Bayesian Model Selection** Try different priors and choose the one that maximises the marginal evidence $p(\mathcal{D})$

Overview

1. Introduction
   - Bayesian Inference
   - Why Bother?
   - Bayesian vs a Frequentist Approach
   - Challenges
     - Calculating the Marginal Evidence
     - Subjective Priors
2. The Model
   - The Prior
   - The Likelihood
   - Hyper-parameters
   - Examples
3. Priors
   - The Exponential Family
   - Conjugate Priors
   - Non Informative Priors
     - Invariant Priors
     - Jefferey Priors
     - Reference Priors
4. Approximate Inference
   - Collapsing the posterior to a point
   - Laplace Approximation
   - Variational Inference
   - Sampling Approaches
5. Bayesian Model Comparison
   - Bayesian Model Comparison
   - Occams Razor
   - Approximating the Marginal Likelihood
   - Alternatives to the Marginal Likelihood

The Model

The prior $p(\theta)$:

- Typically assumed to be some parametric distribution $p(\theta) = p(\theta|\phi)$ with *hyper-parameters* ϕ
- *Can* be used to encode any prior information we might have about the likely value of θ
  - Domain Knowledge - "we expect the person to be around 3 ± 0.5 metres away" or "a coin is likely to be heads with a 50 % chance"
  - Known Properties - "the probability of a person being a certain distance away from us should be independent of whether we use inches or metres and / or our choice of origin" (Jeffrey Priors / Reference Priors)
- *Might* be chosen to make the posterior easy to compute (Conjugate Priors)

The Model

The Likelihood $p(\mathcal{D}|\theta, \phi)$

- Describes how the data $\mathcal{D}$ is generated given θ
- Typically constructed assuming *i.i.d* observations $p(\mathcal{D}|\theta) = \prod_i p(\mathcal{D}_i|\theta_i)$
- Might correspond to a model we already have available to us
- Typically will be a parametric model with hyper-parameters ϕ

Choosing Hyper-parameters

How do we choose the hyper-parameters ϕ?

- **Domain Specific Knowledge** In some cases we have domain specific knowledge available or we might be able to measure the hyper-parameters (eg. Kalman Filter)
- **A fully Bayesian treatment** When no domain knowledge is available, in a fully Bayesian treatment we could add the hyper-parameters into θ, assuming a prior $p(\theta)$. Whilst, this *may* lead to better results this comes at the expense of increased model complexity.
- **Type II Maximum Likelihood / Empirical Bayes** As a compromise we could think about setting the hyper-parameters to point estimates estimated from the data (eg. MLE or MAP estimation). Over-fitting becomes less of an issue the higher up the chain we go.

Method	Definition
Maximum likelihood	$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D} \mid \boldsymbol{\theta})$
MAP estimation	$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \phi)$
ML-II (Empirical Bayes)	$\hat{\phi} = \text{argmax}_{\phi} \int p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \phi) d\boldsymbol{\theta}$
MAP-II	$\hat{\phi} = \text{argmax}_{\phi} \int p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \phi) p(\phi) d\boldsymbol{\theta}$
Full Bayes	$p(\boldsymbol{\theta}, \phi \mid \mathcal{D}) \propto p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \phi) p(\phi)$

Examples

Example

Linear Regresssion For latents $\theta = \{w\}$, observations $\mathcal{D} = \{X, Y\}$, and hyper-parameters $\phi = \{\tau, \kappa\}$ we have

$$p(\mathcal{D}|\theta, \phi) = \text{Nor}(Y|Xw, \kappa^{-1}I) \quad \text{and} \quad p(\theta|\phi) = \text{Nor}(w|0, \tau^{-1}) \tag{4}$$

Example

PCA $\theta = \{X\}$, $\mathcal{D} = \{Y\}$, $\phi = \{\mu, \sigma^2, W\}$

$$p(\mathcal{D}|\theta, \phi) = \text{Nor}(Y|XW + \mu, \sigma^2 I) \quad \text{and} \quad p(\theta|\phi) = \text{Nor}(X|0, I) \tag{5}$$

Examples

Example

Gaussian Mixture Model $\mathcal{D} = \boldsymbol{x}_{1:N}$, $\theta = \{z_{1:N}, \boldsymbol{\Lambda}_{1:K}, \boldsymbol{\mu}_{1:K}\}$, $\phi = \{\gamma, \boldsymbol{\Lambda}_0, \nu, \beta\}$

$$p\left(\mathcal{D}|\theta, \phi\right) = \left[\prod_{n=1}^{N} \mathrm{Nor}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_{z_n}, \boldsymbol{\Lambda}_{z_n}\right)\right] \tag{6}$$

$$p\left(\theta \mid \phi\right) = \mathrm{Dir}(\boldsymbol{\pi} \mid \gamma) \left[\prod_{n=1}^{N} \mathrm{Cat}\left(z_n \mid \boldsymbol{\pi}\right)\right] \left[\prod_{k=1}^{K} \mathrm{Wis}\left(\boldsymbol{\Lambda}_k \mid \boldsymbol{\Lambda}_0, \nu\right) \mathrm{Nor}\left(\boldsymbol{\mu}_k \mid \boldsymbol{\Lambda}_k, \beta\right)\right] \tag{7}$$

Examples

Example

Kalman Filter for observations $\mathcal{D} = x_{1:T}$, latents $\theta = z_{0:T}$, hyper-parameters $\phi = \{A, a, Q, B, b, b_0, R, R_0\}$ the kalman filter assumes a model

$$p(\mathcal{D}, \theta|\phi) = \overbrace{\left[\prod_{t=1}^{T} \text{Nor}(x_t|Az_t + a, Q)\right]}^{p(\mathcal{D}|\theta,\phi)} \overbrace{\left[\prod_{t=1}^{T} \text{Nor}(z_t|Bz_{t-1} + b, R)\right] N(z_0|b_0, R_0)}^{p(\theta|\phi)}$$

Overview

1. Introduction
   - Bayesian Inference
   - Why Bother?
   - Bayesian vs a Frequentist Approach
   - Challenges
     - Calculating the Marginal Evidence
     - Subjective Priors
2. The Model
   - The Prior
   - The Likelihood
   - Hyper-parameters
   - Examples
3. Priors
   - The Exponential Family
   - Conjugate Priors
   - Non Informative Priors
     - Invariant Priors
     - Jefferey Priors
     - Reference Priors
4. Approximate Inference
   - Collapsing the posterior to a point
   - Laplace Approximation
   - Variational Inference
   - Sampling Approaches
5. Bayesian Model Comparison
   - Bayesian Model Comparison
   - Occams Razor
   - Approximating the Marginal Likelihood
   - Alternatives to the Marginal Likelihood

The Exponential Family

The exponential family includes a wide range of common likelihoods:

- **Continuous Variables**: Normal, Multivariate Normal, Uniform, Pareto, Log-Normal, Gamma, Inverse Gamma, Weibull, Exponential, Beta, Chi-Squared, Wishart, Inverse Wishart, Normal-gamma
- **Discrete Variables**: Bernoulli, Binomial, Poisson, Categorical, Mutlinomial, Hypergeometric, Geometric, Dirichilet

Theorem (The Exponential Family)

Every member of the exponential family has a likelihood $p(x|\eta)$ of the form

$$p(x|\eta) = h(x)g(\eta)e^{\eta^\top u(x)} \tag{8}$$

*where η are the **natural parameters**, $h(x) \in \mathbb{R}$ is the **base measure**, $g(\eta) \in \mathbb{R}$ is referred to as the **partition function** and $u(x)$ are the sufficient statistics of the distribution and $h(x)$, $g(\eta)$ and $u(x)$ are all known functions.*

Conjugate Priors

Theorem (Conjugate Priors)

Every member of the exponential family $p(x|\eta)$ has a conjugate prior $p(\eta)$ given as

$$p(\eta|\chi,\nu) = f(\chi,\nu)g(\eta)^{\nu}e^{\nu\eta^{\top}\chi} \tag{9}$$

where $f(\chi,\nu)$ is a partition function and $g(\eta)$ is partition function of $p(x|\eta)$. The hyper-parameter $\nu \in \mathbb{R}$ can be thought of as the effective number of psuedo-observations whilst χ their psuedo-average. The posterior distribution in this case will take the same form as the prior:

$$p(\eta|X,\chi,\nu) \propto g(\eta)^{\nu+\eta}e^{\eta^{\top}\left(\sum_n u(x_n)+\nu\mathcal{X}\right)} \tag{10}$$

Conjugate Priors

Observation

Inferring $p(\theta|\mathcal{D})$ will *always* be tractable for any likelihood $p(\mathcal{D}|\theta)$ in the exponential family if we use the *conjugate* prior $p(\theta)$. In this case $p(\theta|\mathcal{D})$ will take the same form as $p(\theta)$.

Examples

- Beta-Binomial
- Categorical-Dirichlet (latent dirichilet allocation)
- Gaussian-Gaussian
  1. Unknown Mean (Bayesian linear regression, Kalman Filter)
  2. Unknown Co-variance
  3. Unknown Mean + Unknown Co-variance

The Beta-Bernoulli Model

Example

Beta-Bernoulli Model $\theta = \{p\}$, $\mathcal{D} = \{y_n\}_{n=1}^N$, $\phi = \{a, b\}$

- Likelihood: $p(\mathcal{D}|\theta, \phi) = \prod_{n=1}^N \text{Bern}(y_n|p) = \prod_{n=1}^N p^{y_n}(1-p)^{1-y_n} = p^{N_0}(1-p)^{N_1}$
- Prior: $p(\theta|\phi) = \text{Beta}(\theta|a, b) = \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1}$
- Posterior: $p(\theta|\mathcal{D}, \phi) = \text{Beta}(p|a + N_0, b + N_1)$
- Posterior-Predictive: $p(y = 1|\mathcal{D}, \phi) = \frac{a+N_0}{(a+N_0)(b+N_1)}$
- Marginal Likelihood: $p(\mathcal{D}|\phi) = \frac{B(a+N_0, b+N_1)}{B(a,b)}$

where N_0 and N_1 are the number of observations belonging to class 0 and 1 respectively.

The Dirichlet-Categorical Model

Example

Dirichlet-Categorical $\theta = \{\boldsymbol{\theta}\}$, $\mathcal{D} = \{\boldsymbol{y}_n\}_{n=1}^N$, $\phi = \{\boldsymbol{\alpha}\}$

- Likelihood: $p(\mathcal{D}|\theta, \phi) = \prod_{n=1}^N \text{Cat}(y_n|\boldsymbol{\theta}) = \prod_{c=1}^C \theta_k^{N_k}$
- Prior: $p(\theta|\phi) = \text{Dirichlet}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^K \theta_k^{\alpha_k - 1}$
- Posterior: $\text{Dirichlet}(\boldsymbol{\theta}|\alpha_1 + N_1, ..., \alpha_K + N_k)$
- Posterior-Predictive: $p(y = 1|\mathcal{D}, \phi) = \frac{\alpha_k + N_k}{\sum_j \alpha_j + N_j}$
- Marginal Likelihood: $p(\mathcal{D}|\phi) = \frac{B(\boldsymbol{\alpha}+\boldsymbol{N})}{B(\boldsymbol{\alpha})}$

where N_k are the number of observations belonging to class k.

Multi-Variate Normal with Unknown Mean and Co-variance

Example

Multi-Variate Normal $\theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, $\mathcal{D} = \{\boldsymbol{y}_n\}_{n=1}^N$, $\phi = \{\check{\boldsymbol{m}}, \check{\kappa}, \check{\boldsymbol{S}}, \check{\nu}\}$

- Likelihood: $p(\mathcal{D}|\theta, \phi) = \prod_{n=1}^N \text{Norm}(\boldsymbol{y}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Prior: $p(\theta|\phi) = \text{Norm}(\boldsymbol{\mu}|\check{\boldsymbol{m}}, \check{\kappa}^{-1}\boldsymbol{\Sigma})\text{InvWas}(\boldsymbol{\Sigma}|\check{\boldsymbol{S}}, \check{\nu})$
- Posterior: $p(\theta|\mathcal{D}, \phi) = \text{Norm}(\boldsymbol{\mu}|\hat{\boldsymbol{m}}, \hat{\kappa}^{-1}\boldsymbol{\Sigma})\text{InvWas}(\boldsymbol{\Sigma}|\hat{\boldsymbol{S}}, \hat{\nu})$
- $\boldsymbol{\mu}$-posterior: $p(\boldsymbol{\mu}|\mathcal{D}|\phi) = \text{Student}(\boldsymbol{\mu}|\hat{\boldsymbol{\mu}}, (\hat{\kappa}\hat{\nu})^{-1}\hat{\boldsymbol{S}}, \hat{\nu})$
- $\boldsymbol{\Sigma}$-posterior: $p(\boldsymbol{\Sigma}|\mathcal{D}, \phi) = \text{InvWas}(\boldsymbol{\Sigma}|\hat{\boldsymbol{S}}, \hat{\nu})$
- Posterior Predictive: $p(\boldsymbol{y}|\mathcal{D}, \phi) = \text{Student}(\boldsymbol{\mu}|\hat{\boldsymbol{\mu}}, \frac{(\hat{\kappa}+1)\hat{\boldsymbol{S}}}{\hat{\kappa}(\hat{\nu}-D+1)}, \hat{\nu} - D + 1)$

Where...

1. *Empirical Mean* $\bar{\boldsymbol{y}} = \frac{1}{N}\sum_n \boldsymbol{y}_n$
2. *Empirical Scatter Matrix* $\boldsymbol{S}_{\bar{\boldsymbol{y}}} = \sum_n \boldsymbol{y}_n\boldsymbol{y}_n^\top - \frac{1}{N}(\sum_n \boldsymbol{y}_n)(\sum_n \boldsymbol{y}_n)^\top$
3. *Posterior Mean* $\hat{\boldsymbol{m}} = \frac{\check{\kappa}}{\check{\kappa}+N}\hat{\boldsymbol{m}} + \frac{N}{\check{\kappa}+N}\bar{\boldsymbol{y}}$
   - 3.1 Convex combination of prior mean and empirical mean (MLE)
4. *Posterior Confidences* $\hat{\kappa} = \check{\kappa} + N$ and $\hat{\nu} = \check{\nu} + N$
   - 4.1 Confidence increased by the number of observations
5. *Posterior Scatter Matrix* $= \hat{\boldsymbol{S}} = \check{\boldsymbol{S}} + \boldsymbol{S}_{\bar{\boldsymbol{y}}} + \frac{\check{\kappa}N}{\check{\kappa}+N}(\bar{\boldsymbol{y}} - \hat{\boldsymbol{m}})(\bar{\boldsymbol{y}} - \hat{\boldsymbol{m}})^\top$
   - 5.1 Prior scatter matrix + Empirical scatter matrix + Uncertainty from unknown mean

(Less)Non-informative Priors

- In many situations we have domain specific knowledge that we can use to construct a prior $\implies$ by introducing biases we can reach the answer much more quickly (eg. we expect a coin flip to be heads 50% of the time)
- In situations where no-domain knowledge is available how do we choose a prior which encodes minimal assumptions?
  - **Invariant Priors** assume that the prior should preserve some symmetry (eg. invariant to co-ordinate origin / choice of units).
  - **Jeffrey Priors** choose a prior encoding minimal information such that scientists working with different but valid parameterisations of θ are still guaranteed to reach the same conclusions.
  - **Reference Priors** Assume a prior which is maximally far from all posteriors given the data
- But...
  - can result in improper priors (priors that do not normalise to one) $\implies$ Bad for Bayesian model comparison
  - inference many no longer be tractable

Non-Informative Priors

Non-Informative priors *are* still informative - *every* prior encodes some prior assumption

Invariant Priors - A Motivating Example

Example

Permutation Invariance You arrive into a room to find three cups turned upside down on the table. You are told, by your closest friend (who never lies) that under one of the cups is a ball. You will most likely assume a uniform prior over each cup (eg. $p = 1/3$) but why? To see why we reached this conclusion, note that our choice of label is arbitrary; we could label the cups A, B, C or B, C, A or C, A, B. Each is valid. The conclusion of the experiment should therefore be the same independent of our choice of label. A uniform prior is the *only* prior which respects this (permutation) *invariance*.

Translation Invariant Priors

Example

You need to measure how far away you are from a laser. You know the laser has an accuracy of $\pm\sigma$m standard deviations but no nothing about your current location. What is a sensible choice of prior in this case?

The likelihood of the laser measurement is given as $p(x|\mu) = \text{Norm}(x|\mu, \sigma^2)$ where μ is your true position. Whilst you have no knowledge about your current position you expect the accuracy of the laser will be the same no matter how far you are from it; if you move to a new location $\hat{\mu} = \mu + c$ the likelihood of the new laser observation will be given as $p(\hat{x}|\hat{\mu}) = \text{Norm}(\hat{x}|\hat{\mu}, \sigma^2)$. A sensible choice of prior will therefore satisfy this symmetry...

Translation Invariant Priors

Theorem

Translational Invariant Priors A likelihood of the form $p(x|\mu) = f(x - \mu)$ respects translational invariance if $p(\hat{x}|\hat{\mu}) = f(\hat{x} - \hat{\mu})$ where $\hat{x} = x + c$ and $\hat{\mu} = \mu + c$. To choose a prior that respects this invariance, we need a prior that gives equal probability mass to every possible μ in the domain such that $p(\mu) = constant$.

Proof The integral of the probability mass over an interval $A \leq \mu \leq B$ and the shifted interval $A - c \leq \mu \leq B - c$ should be the same. This implies

$$\int_A^B p(\mu)d\mu = \int_{A-c}^{B-c} p(\mu)d\mu = \int_A^B p(\mu - c)d\mu \implies p(\mu - c) = p(\mu) \implies p(\mu) = \text{constant}$$

Scale Invariant Priors

Example

You need to measure your location relative to a laser. You know you are most likely at a distance μ from the laser but know nothing about the its accuracy. What is a sensible choice of prior in this case?

The likelihood of the laser measurement is given as $p(x|\sigma) = \text{Norm}(x|\mu, \sigma^2)$ where μ is your expected position. Whilst you have no knowledge about the accuracy of the laser, you would not expect your choice of scale to effect the outcome of the experiment; if you choose a new unit such that $\hat{\sigma} = c\sigma$ the likelihood of the laser observation in the new units will be given as $p(\hat{x}|\hat{\sigma}) = \text{Norm}(\hat{x}|\hat{\sigma}^2)$. A sensible choice of prior will therefore satisfy this symmetry...

Scale Invariant Priors

Theorem

Scale Invariant Priors A likelihood of the form $p(x|\mu) = \frac{1}{\sigma} f(\frac{x}{\sigma})$ *respects scale invariance if* $p(\hat{x}|\hat{\mu}) = \frac{1}{\hat{\sigma}} f(\frac{x}{\hat{\sigma}})$ *where* $\hat{x} = cx$ *and* $\hat{\mu} = c\mu$. *A prior that respects this symmetry will be given as* $p(\sigma) \propto \frac{1}{\sigma}$

Proof The integral of the probability mass over an interval $A \leq \sigma \leq B$ and the shifted interval $A/c \leq \sigma \leq B/c$ should be the same. This implies

$$\int_A^B p(\sigma)d\sigma = \int_{A/c}^{B/c} p(\sigma)d\sigma = \int_A^B \frac{1}{c} p(\frac{\sigma}{c})d\sigma \implies p(\sigma) = \frac{1}{c} p(\frac{\sigma}{c}) \implies p(\sigma) \propto \frac{1}{\sigma}$$

Jefferey Priors

- In the case where we know nothing about θ we might assume that all possible θ are equally likely (eg. uniform distribution)
- For example a scientist trying to infer the probability that a coin comes up heads θ might assume that θ is uniformly distributed over the interval $[0, 1]$.
- In this case the odds of success $\tau = \frac{\theta}{1-\theta} = f(\theta)$ is distributed as

$$p(\tau) = \left| \frac{df^{-1}}{d\theta} \right| \text{Uniform}(f^{-1}(\tau)) = \frac{1}{(1+\tau^2)}$$

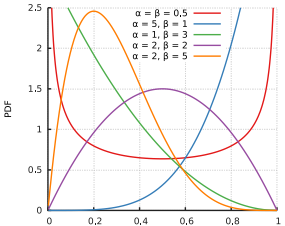
- Another scientist assuming that τ is uniformally distributed will therefore reach a completely different conclusion! And yet there is no reason to believe that the first scientist is any more right than the second.
- Jeffreys priors ensure that two statisticians will reach the same conclusion independent of whether they use τ or θ.

Jefferey Priors

Theorem

For any likelihood $p(\mathcal{D}|\theta)$, Jeffreys prior is given as $p_J(\theta) \propto \sqrt{F(\theta)}$ where $F(\theta)$ is the Fischer information $F(\theta) = \mathbb{E}_{p(\mathcal{D}|\theta)}\left[\frac{\partial}{\partial\theta}\log p(\mathcal{D}|\theta)\right]$. In this case the prior will be invariant to all parameterisations $\tau = f(\theta)$ such that $P_J(\theta) = P_J(\tau)\left|\frac{d\tau}{d\theta}\right|$

Example Using Jeffreys prior for the example above we find that $p(\theta) = \text{Beta}(\theta|\frac{1}{2}, \frac{1}{2})$ which is a horse shoe shape with most of the mass concentrated on $\theta = 0$ and $\theta = 1$.



Jeffrey Priors

- **Examples**
  - Bernoulli Distribution: $p(\theta) = \text{Beta}(\theta|\frac{1}{2}, \frac{1}{2})$
  - Categorical Distribution: $p(\theta) = \text{Dir}(\theta|\frac{1}{2}, ..., \frac{1}{2})$
  - Mean of a Gaussian: $p(\theta) \propto 1$ or $p(\theta) = \text{Norm}(0, \infty)$
  - Standard Deviation of a Gaussian: $p(\theta) \propto 1/\sigma$ or $p(\theta) = \text{InvGamma}(\sigma|0, 0)$
- **Improper Priors** in many instances jeffrey priors are *improper* (i.e do not integrate to 1). This is not generally a problem if they lead to a *proper* posterior (i.e that does integrate to 1). Improper priors can lead to difficulties with Model Selection
- **Relation to Invariant Priors** the jeffrey prior for the mean and the variance of a Gaussian distribution are the same as the invariant shift and scale priors derived earlier

Reference Priors

- Jeffreys priors are also applicable to higher dimensions: $p_J(\boldsymbol{\theta}) \propto \sqrt{\det(\boldsymbol{F}(\boldsymbol{\theta}))}$ where $\boldsymbol{F}(\boldsymbol{\theta})$ is the fisher information matrix. However, this can be problematic...
- Reference priors are a superior alternative: "choose a prior that is maximally far from all posterior distributions when averaged across the data":

$$p^*(\boldsymbol{\theta}) = \arg\max_{p(\boldsymbol{\theta})} \left(\mathbb{E}_{p(\mathcal{D})}[\mathbb{KL}(p(\boldsymbol{\theta}|\mathcal{D})\|p(\boldsymbol{\theta}))]\right) \tag{11}$$

$$= \arg\max_{p(\boldsymbol{\theta})} \left(\mathbb{E}_{p(\boldsymbol{\theta})}[\mathbb{KL}(p(\mathcal{D}|\boldsymbol{\theta})||p(\mathcal{D}))]\right) \tag{12}$$

- In 1d reference priors are equivalent to Jeffrey priors
- In higher dimensions reference priors can be computed for one parameter at a time using the chain rule. Can become computationally intractable. In this case variational approximations may be used instead.

Overview

1. Introduction
   - Bayesian Inference
   - Why Bother?
   - Bayesian vs a Frequentist Approach
   - Challenges
     - Calculating the Marginal Evidence
     - Subjective Priors
2. The Model
   - The Prior
   - The Likelihood
   - Hyper-parameters
   - Examples
3. Priors
   - The Exponential Family
   - Conjugate Priors
   - Non Informative Priors
     - Invariant Priors
     - Jefferey Priors
     - Reference Priors
4. Approximate Inference
   - Collapsing the posterior to a point
   - Laplace Approximation
   - Variational Inference
   - Sampling Approaches
5. Bayesian Model Comparison
   - Bayesian Model Comparison
   - Occams Razor
   - Approximating the Marginal Likelihood
   - Alternatives to the Marginal Likelihood

Approximate Inference

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D})p(\mathcal{D}|\theta)d\theta} \tag{13}$$

Intractable Inference

Calculating the posterior distribution $p(\theta|\mathcal{D})$ involves calculating the *marginal evidence* $p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta$ which depending on our choice of model is often intractable.

In this case several options are available:

- Choose a model such that calculating $p(\theta|\mathcal{D})$ is tractable $\implies$ **Conjugate Priors**
- Approximate the posterior distribution
  - As a single point $p(\theta|\mathcal{D}) \approx \delta(\theta - \hat{\theta})$ $\implies$ **MAP estimation**
  - As a Gaussian distribution $p(\theta|\mathcal{D}) \approx \mathcal{N}(\theta|\mu, \Sigma)$ $\implies$ **Laplace Approximation**
  - As a General approximate distribution $p(\theta|\mathcal{D}) \approx q(\theta)$ $\implies$ **Variational Inference**
  - As a set of samples $p(\theta|\mathcal{D}) = \sum_i \omega_i \delta(\theta - \theta_i)$ $\implies$ **Histogram Filter**, **MCMC**

Collapsing the posterior to a point

- Perhaps the simplest option is to collapse the posterior to a point $p(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$
- The optimal point estimation is dependent on our choice of regret $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \int p(\boldsymbol{\theta}|\mathcal{D})\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})d\boldsymbol{\theta}$
  - $\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2 \implies \hat{\boldsymbol{\theta}} = \overline{\boldsymbol{\theta}} = \mathbb{E}[\boldsymbol{\theta} \mid \mathcal{D}]$
  - $\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = |\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|_1 \implies \hat{\boldsymbol{\theta}} = \text{median}(p(\boldsymbol{\theta}|\mathcal{D}))$
  - $\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \mathbb{I}(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}) \implies \hat{\boldsymbol{\theta}}_{MAP} = \text{argmax}_{\boldsymbol{\theta}}\, p(\boldsymbol{\theta} \mid \mathcal{D}) = \text{argmax}_{\boldsymbol{\theta}}\, p(\boldsymbol{\theta}, \mathcal{D})$
- In practice the easiest point estimate to compute is $\hat{\boldsymbol{\theta}}_{MAP}$ at it only involves optimisation and not integration.
- Both $\hat{\boldsymbol{\theta}} = \overline{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}} = \text{median}(p(\boldsymbol{\theta}|\mathcal{D}))$ can be estimated using Monte Carlo integration

The Laplace Approximation

Theorem (The Laplace Approximation)

Any distribution $p(\boldsymbol{z}) = \frac{1}{Z} f(\boldsymbol{z})$ can be approximated as a Gaussian distribution $p(\boldsymbol{z}) \approx Norm(\boldsymbol{z}|\hat{\boldsymbol{z}}, \boldsymbol{A}^{-1})$ where $\hat{\boldsymbol{z}}$ is the mode of $p(\boldsymbol{z})$ and $\boldsymbol{A} = -\nabla\nabla \log f(\boldsymbol{z})|_{\boldsymbol{z}=\hat{\boldsymbol{z}}}$ is the negative log hessian evaluated at $\hat{\boldsymbol{z}}$.

Proof Considering $p(\boldsymbol{z}) = \frac{1}{Z} f(\boldsymbol{z})$ and performing a taylor expansion of $\log f(\boldsymbol{z})$ around the mode $\hat{\boldsymbol{\theta}}$ of $p(\boldsymbol{z})$ gives,

$$\log f(\boldsymbol{z}) \approx \log f(\hat{\boldsymbol{z}}) - \frac{1}{2}(\boldsymbol{z} - \hat{\boldsymbol{z}})^\top \boldsymbol{A}(\boldsymbol{z} - \hat{\boldsymbol{z}}) \implies p(\boldsymbol{z}) \propto e^{-\frac{1}{2}(\boldsymbol{z}-\hat{\boldsymbol{z}})^\top \boldsymbol{A}(\boldsymbol{z}-\hat{\boldsymbol{z}})} \tag{14}$$

$$\implies p(\boldsymbol{z}) = \text{Norm}(\boldsymbol{z}|\hat{\boldsymbol{z}}, \boldsymbol{A}^{-1}) \tag{15}$$

where we have used the fact that $\nabla \log p(\boldsymbol{z}) = 0$ when $\boldsymbol{z}$ is equal to the mode $\hat{\boldsymbol{z}}$.

The Laplace Approximation

Observation use the Laplace approximation to approximate $p(\boldsymbol{\theta}|\mathcal{D}) \approx \text{Norm}(\boldsymbol{\theta}|\boldsymbol{\theta}_{MAP}, \boldsymbol{A}^{-1})$...

Laplace Approximation to the Posterior Distribution

The posterior $p(\boldsymbol{\theta}|\mathcal{D})$ can be approximated as $p(\boldsymbol{\theta}|\mathcal{D}) \approx \text{Norm}(\boldsymbol{\theta}|\boldsymbol{\theta}_{MAP}, \boldsymbol{A}^{-1})$ where $\boldsymbol{\theta}_{MAP}$ is the *MAP* estimate for $\boldsymbol{\theta}$ and $\boldsymbol{A} = -\nabla\nabla \log p(\boldsymbol{\theta}, \mathcal{D})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{MAP}}$.

Proof Follows almost immediately from the previous definition of the laplace approximation setting $\boldsymbol{z} = \boldsymbol{\theta}$, $p(\boldsymbol{z}) = p(\boldsymbol{\theta}|\mathcal{D})$ and $f(\boldsymbol{z}) = p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}, \mathcal{D})$ whilst noting that the mode of $p(\boldsymbol{\theta}|\mathcal{D})$ is given by $\hat{\boldsymbol{\theta}}_{MAP}$.

Variational Inference

- Laplace Approximation = Approximate $p(\theta|\mathcal{D})$ with a normal distribution
- Variational Inference = Approximate $p(\theta|\mathcal{D})$ with a *general* distribution $q(\theta) \in \mathcal{Q}$ where

$$q = \text{argmax}_{q \in \mathcal{Q}} d[q(\theta), p(\theta|\mathcal{D})]$$

- Any divergence between $q(\theta)$, $p(\theta|\mathcal{D})$ can be used - in practice KL divergence is a good choice (arises naturally from information considerations)

$$\begin{aligned} q &= \arg\min_{q \in \mathcal{Q}} \mathbb{KL}[q(\theta) \| p(\theta|\mathcal{D})] && (16) \\ &= \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{q(\theta)}[\log q(\theta) - \log(p(\theta|\mathcal{D})] && (17) \\ &= \arg\min_{q \in \mathcal{Q}} \mathbb{ELBO}[q|\mathcal{D}] + \log p(\mathcal{D}) && (18) \\ &= \arg\min_{q \in \mathcal{Q}} \mathbb{ELBO}[q|\mathcal{D}] && (19) \end{aligned}$$

where $\mathbb{ELBO}[q|\mathcal{D}] = \mathbb{E}_{q(\theta)}[\log p(\mathcal{D}|\theta)p(\theta) - \log q(\theta)]$.

Sampling Approaches

Variational Inference

Variational approaches are fast *but* will only be accurate if the true posterior $p(\theta|\mathcal{D})$ belongs to $\mathcal{Q}$. In all other cases $q(\theta)$ will be **sub-optimal**.

- Sampling approaches approximate $q(\theta) = \sum_s \omega_s \delta(\theta - \theta_s)$ where $\sum_s \omega_s = 1$ and guarantee in the limit as $N \to \infty$ that $q(\theta) \to p(\theta|\mathcal{D})$
- The histogram filter
  - Convert continuous probability space $\theta \in \mathbb{R}^N$ to a finite set of possibilities $\Theta = \{\theta_1, ..., \theta_K\}$
  - Approximate $p(\theta|\mathcal{D}) = \sum_s \omega_s \delta(\theta - \theta_s)$ with $\omega_s = \frac{p(\mathcal{D}|\theta)p(\theta)}{\sum_k p(\mathcal{D}|\theta_k)p(\theta_k)}$
  - Other normalisation schemes are possible (eg. using the trapezoid rule)
- **Curse of Dimensionality**: The number of samples needed to evenly cover a space $\theta \in \mathbb{R}^N$ is exponential in N.

Markov Chain Monte Carlo

- **Solution**: Using samples where $p(\theta|\mathcal{D}) \approx 0$ is inefficient $\implies$ use $p(\theta|\mathcal{D})$ to guide the sampling procedure
- But evaluating $p(\theta|\mathcal{D}) = p(\theta, \mathcal{D})p(\mathcal{D})$ exactly is infeasible as $p(\mathcal{D}) = \int p(\theta|\mathcal{D})p(\mathcal{D})d\theta$ (the problem we are trying to solve). How do we achieve a more efficient sampling approach without evaluating $p(\mathcal{D})$?
- **Metropolis Hasting (MH) Algorithm**: start at a random point in parameter space and perform a random walk sampling new θ_k from a *proposal distribution* $q(\theta'|\theta)$ which is chosen such that (eventually) the fraction of time we spend visiting each θ is proportional to the posterior distribution. To decide whether to stay at the current state or to move to a new θ' all we need to do is evaluate the *unormalised* density ratio $\frac{p(\theta|\mathcal{D})}{p(\theta'|\mathcal{D})} = \frac{p(\mathcal{D},\theta)}{p(\mathcal{D}',\theta)}$
- Proposal distribution $q(\theta'|\theta)$ in MH is typically chosen to be a Gaussian distribution $\rightarrow$ *Inefficient due to blind random walk behaviour*
- **Gibbs Sampling**: improve efficiency of sampling approach using conditional relations between latents in θ
- **Hamiltonian Monte Carlo**: guide the walk using gradient information when $\theta \in \mathbb{R}^D$

Overview

1. Introduction
   - Bayesian Inference
   - Why Bother?
   - Bayesian vs a Frequentist Approach
   - Challenges
     - Calculating the Marginal Evidence
     - Subjective Priors
2. The Model
   - The Prior
   - The Likelihood
   - Hyper-parameters
   - Examples
3. Priors
   - The Exponential Family
   - Conjugate Priors
   - Non Informative Priors
     - Invariant Priors
     - Jefferey Priors
     - Reference Priors
4. Approximate Inference
   - Collapsing the posterior to a point
   - Laplace Approximation
   - Variational Inference
   - Sampling Approaches
5. Bayesian Model Comparison
   - Bayesian Model Comparison
   - Occams Razor
   - Approximating the Marginal Likelihood
   - Alternatives to the Marginal Likelihood

Bayesian Model Comparison

The *marginal evidence* $p(\mathcal{D}|m) = \int p(\mathcal{D}|\theta, m)p(\theta|m)d\theta$ can be used to compare different models $m \in \mathcal{M}$

- **Bayesian Model Selection**:

$$\hat{m} = \arg\max_m p(m|\mathcal{D}) \quad \text{where} \quad \frac{p(\mathcal{D}|m)p(m))}{\sum_{m\in\mathcal{M}} p(\mathcal{D}|m)p(m))} \tag{20}$$

$$= \arg\max_{m\in\mathcal{M}} p(\mathcal{D}|m) \quad \text{if} \quad p(m) = \text{Uniform}(|\mathcal{M}|^{-1}) \tag{21}$$

- **Bayesian Model Averaging**:

$$p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{D}) = \sum_m p(\boldsymbol{y}|\boldsymbol{x}, m)p(m|\mathcal{D}) \tag{22}$$

$$\propto \sum_m p(\boldsymbol{y}|\boldsymbol{x}, m)p(\mathcal{D}|m) \quad \text{if} \quad p(m) = \text{Uniform}(|\mathcal{M}|^{-1}) \tag{23}$$

Improper Priors

Improper priors should be avoided with model selection.

Bayesian Model Comparison

- **Bayesian Hypothesis Testing**:

$$BF = \frac{p(\mathcal{D}|m_1)}{p(\mathcal{D}|m_0)} = \frac{p(m_1|\mathcal{D})}{p(m_0|\mathcal{D})} \frac{p(m_0)}{p(m_1)} \tag{24}$$

Bayes factor $BF(1,0)$	Interpretation
$BF < \frac{1}{100}$	Decisive evidence for M_0
$BF < \frac{1}{10}$	Strong evidence for M_0
$\frac{1}{10} < BF < \frac{1}{3}$	Moderate evidence for M_0
$\frac{1}{3} < BF < 1$	Weak evidence for M_0
$1 < BF < 3$	Weak evidence for M_1
$3 < BF < 10$	Moderate evidence for M_1
$BF > 10$	Strong evidence for M_1
$BF > 100$	Decisive evidence for M_1

Occams Razor

Occams Razor

For two models with the same performance prefer the simpler model.

Bayesian Occams Razor

The marginal likelihood $p(\mathcal{D}|m)$ naturally favours the simplest model. A more complex model...

- $\Longrightarrow$ has larger parameter space
- $\Longrightarrow$ has less mass $p(\theta|m)$ per each parameter (must normalise to 1)
- $\Longrightarrow$ has smaller average $p(\mathcal{D}|m) = \int p(\mathcal{D}|\theta, m)p(\theta|m)d\theta$

Approximating the Marginal Likelihood

- The marginal likelihood can be difficult to compute $\rightarrow$ requires marginalising over the entire parameter space.
- **Bayesian Information Criteria**
  - Approximating $p(\boldsymbol{\theta}|\mathcal{D}) \approx \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\theta}_{MAP}, \boldsymbol{A})$ using the laplace approximation we can estimate $p(\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\theta|\mathcal{D})$ and so

$$\log p(\mathcal{D}) \approx \log p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) + \log p(\boldsymbol{\theta}_{MAP}) + \frac{M}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{A}|$$

  - If we assume the prior distribution is broad then $\boldsymbol{\theta}_{MAP} \rightarrow \boldsymbol{\theta}_{MLE}$ and the hessain matrix has *full rank* then

$$logp(\mathcal{D}) \approx \log p(\mathcal{D}|\boldsymbol{\theta}_{MLE}) - \frac{1}{2}D\log N = BIC[m]$$

    where $\boldsymbol{\theta}_{MLE} \in \mathbb{R}^D$ and N is the number of data points.

Alternatives to the Marginal Likelihood

- **Akaike Information Criteria**: Frequentist equivalent replacing $\frac{1}{2} D \log N$ with a complexity factor of $C[m] = dof(m)$. Penalises complex models more heavily then BIC.
- **Widely Applicable Information Criteria**:

$$WAIC[m] = -LLPD[m] + C[m]$$

  where $LLPD[m] = \sum_n \log \mathbb{E}[p(y_n|\mathcal{D}, m)]$ and $c[m] = \sum_n \log \mathbb{V}[p(y_n|\mathcal{D}, m)]$
- **Minimum description Length**:
  - Model complexity is equal to the number of bits needed to communicate the data to the reciever
  - The sender specifies the model taking $C[m] = -\log p(m)$ bits.
  - The receiver fits the model by computing $\hat{\boldsymbol{\theta}}_m$ approximately reconstructing the data.
  - The sender then sends the residual errors that cannot be explained by the model taking $-L(m) = -\sum_n \log p(y_n|\theta_m, m)$
  - The minimum description length $J[m] = -L[m] + C[m]$