Lecture Stat 461-561
Bayesian Statistics

AD

March 2008

Introduction

- The following problem is taken from *Pillow Problems* by Lewis Carroll Consider a bag containing a ball of unknown colour, which may be either black or white. A white ball is added to the bag, then a ball is drawn at random from it. The drawn ball happens to be white. What is the probability that the remaining ball is white? The answer is 2/3 (under assumptions).
- Let C_I be the colour of the ball initially in the bag. We set $P(C_I = W) = P(C_I = B) = \frac{1}{2}$.
- Let D the ball that is been drawn, whether the initial one (I) or the white ball that has been added (A). We have $P(D = I) = P(D = A) = \frac{1}{2}$.
- Let C_R, C_D the respective colours of the remaining ball and the drawn ball. Then

$$P(C_R = W | C_D = W) = \frac{P(C_R = W, C_D = W)}{P(C_D = W)} = \frac{1/2}{3/4} = \frac{2}{3}.$$

- Implicitly this problems imposes to accept some concepts that deserve discussion
- Prior uncertainty/prior beliefs (Of which colour is the initial ball?) can be expressed in terms of probabilities

$$P(C_I = W) = P(C_I = B) = \frac{1}{2}.$$

- Taking into account any new information that arises from the experiment (the drawn ball is white) can be done by writing conditional probabilities

$$P(C_R = W | C_D = W) = \frac{2}{3}$$

- The fact that we accept almost without noticing them these concepts prove they are natural and convenient. Therefore we are all Bayesian... sometimes.

- Let A and B be two events then we have

$$\textbf{Bayes' rule}: \; P\left(A\middle|B\right) = \frac{P\left(B\middle|A\right)P\left(A\right)}{P\left(B\right)}$$

where

$$P\left(B\right) = P\left(B\middle|A\right)P\left(A\right) + P\left(B\middle|\overline{A}\right)P\left(\overline{A}\right).$$

- *Example*: You feel sick and your GP thinks you might have contracted a rare disease (0.01% of the population has the disease).
- A test is available but not perfect.
  - If a tested patient has the disease, 100% of the time the test will be positive.
  - If a tested patient does not have the diseases, 95% of the time the test will be negative (5% false positive).
- Your test is positive, should you really care?

- Let A be the event that the patient has the disease and B be the event that the test returns a positive result

$$P(A|B) = \frac{1 \times 0.0001}{1 \times 0.0001 + 0.05 \times 0.9999} \approx 0.002$$

- Such a test would be a complete waste of money for you or the National Health System.
- A similar question was asked to 60 students and staff at Harvard Medical School: 18% got the right answer, the modal response was 95%!

Bayes Formula

- Now consider the case where you model your observations with the *likelihood*

$$X \sim f\left(x|\,\theta\right).$$

- In a Bayesian approach, the unknown parameter θ is assumed **random** and we set a *prior* distribution on it

$$\theta \sim \pi\left(\theta\right).$$

  This distribution expresses our belief about θ before having seen any data.
- In this context, we have

$$\pi\left(\theta|\,x\right) = \frac{\pi\left(\theta\right) f\left(x|\,\theta\right)}{\pi\left(x\right)}$$

  where $\pi\left(x\right)$ is the *marginal likelihood* also called evidence

$$\pi\left(x\right) = \int \pi\left(\theta\right) f\left(x|\,\theta\right) d\theta.$$

Frequentist vs Bayes

So far we have followed a frequentist approach where

- Probabilities refer to limiting relative frequencies. They are (supposed to be) objective properties of the real world.
- Parameter are fixed unknown constants. Because they are not random, we cannot make any probability statements about parameters.
- Statistical procedures should have well-defined long-run properties. For example, a 95% confidence interval should include the true value of the parameter with limiting frequency at least 95%.

Bayesian inference takes a very different stance.

- Probability describes degrees of subjective belief, not limiting frequency. Thus we can make probability statements about things other than data that can recur from some source; e.g. the probability that they will be no snow in Whistler during the 2010 Olympics.
- We can make probability statements about parameters, even though they are fixed constants.
- We make inference about a parameter by producing a probability distribution for it. Point estimates and interval estimates may then be extracted from this distribution.

More on Bayes Formula

- Usually, we simply write

$$\pi(\theta|x) \propto \pi(\theta) f(x|\theta)$$

  where '$\propto$' means 'proportional to'. The proportionality constant can be obtained by normalization.
- **Example**. In Paris, $n - x$ =241,945 girls and x =251,527 boys were born from 1745 to 1770. Let θ be the probability of a male birth. What is the probability that $\theta \geq \frac{1}{2}$.
- We model $X \sim Bin(n, \theta)$, that is

$$\Pr(X = x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

- We set

$$\pi(\theta) = 1_{[0,1]}(\theta).$$

- We have $\Pr\left(\theta \geq \frac{1}{2} \middle| x\right) = \int_{1/2}^{1} \pi\left(\theta \middle| x\right) d\theta$ where

$$\pi\left(\theta \middle| x\right) = \frac{\binom{n}{x} \theta^{x} \left(1-\theta\right)^{n-x} 1_{[0,1]}\left(\theta\right)}{\int_{0}^{1} \binom{n}{x} \theta^{x} \left(1-\theta\right)^{n-x} 1_{[0,1]}\left(\theta\right) d\theta}$$

- Let us introduce the class of Beta densities defined for $\alpha, \beta > 0$

$$\mathcal{B}e\left(\theta; \alpha, \beta\right) = \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} \theta^{\alpha-1} \left(1-\theta\right)^{\beta-1} 1_{[0,1]}\left(\theta\right)$$

where $\Gamma\left(u\right) = \int_{0}^{\infty} t^{u-1} e^{-t} dt$ then $\pi\left(\theta \middle| x\right) = \mathcal{B}e\left(1+x, n+1-x\right)$ and

$$\Pr\left(\theta \geq \frac{1}{2} \middle| x = 251,527\right) = 1 - 1.15 \times 10^{-42} \approx 1.$$

- Consider now that $\pi(\theta) = \mathcal{B}e(\theta; \alpha, \beta)$ which implies that

$$\mathbb{E}(\theta) = \frac{\alpha}{\alpha+\beta}, \ \mathbb{V}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

- Be careful! (α, β) are *fixed* quantities. To distinguish them from θ, we call them *hyperparameters*.
- Then we have for $X \sim Bin(n, \theta)$

$$\pi(\theta|x) = \mathcal{B}e(\theta; \alpha + x, \beta + n - x).$$

- In this simple case, $\pi(\theta)$ and $\pi(\theta|x)$ are in the same parametric family (Beta), albeit with different coefficients.
- The prior on θ can be conveniently reinterpreted as an imaginary initial sample of size $(\alpha + \beta - 2)$ with $\alpha - 1$ observations in favour of the 'yes' answer. Provided that $(\alpha + \beta)$ is small with respect to n, the information carried by the data is prominent.

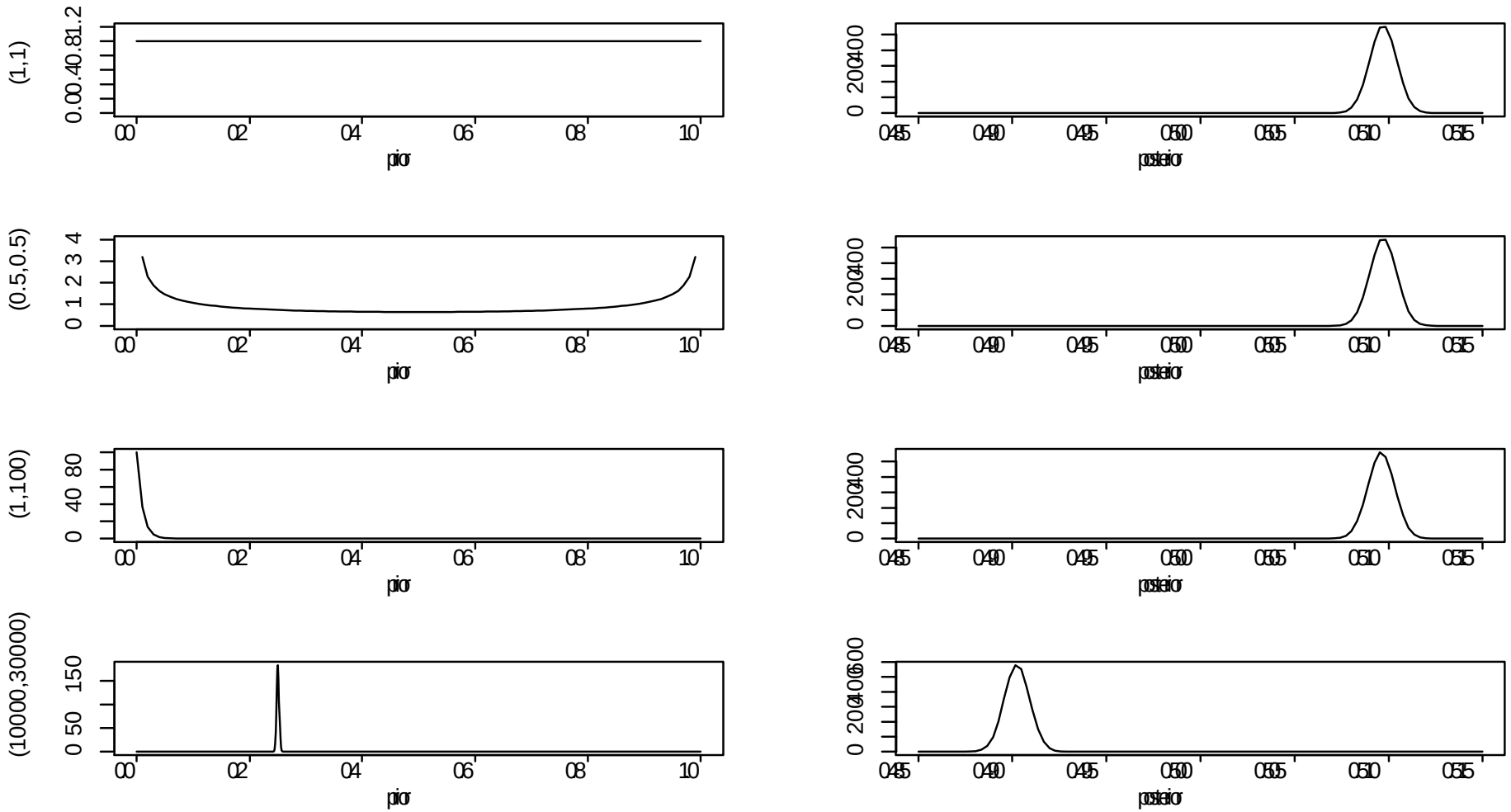

Figure: Prior (left) and posterior distributions (right) of θ for Laplace's example, with different choices for hyperparameters (α, β)

Comparison with the classical/frequentist approach

- In a classical framework, we propose a point estimate of θ which hopefully has nice asymptotic properties. In the Binomial model, we could pick the MLE

$$\widehat{\theta} = \frac{x}{n}, \ \widehat{V} = \frac{x\left(1 - x/n\right)}{n^2}$$

  the latter being an estimate of the variance of $\widehat{\theta}$.
- **Problem.** In a political survey, $n = 1000$ individual are asked if they will vote for candidate W. They all answer 'no'. What can we say about the proportion of people ready to vote for W in the entire population?
- Classical answer: $\widehat{\theta} = \widehat{V} = 0$ versus Bayesian answer (for $\alpha = \beta = 1$)

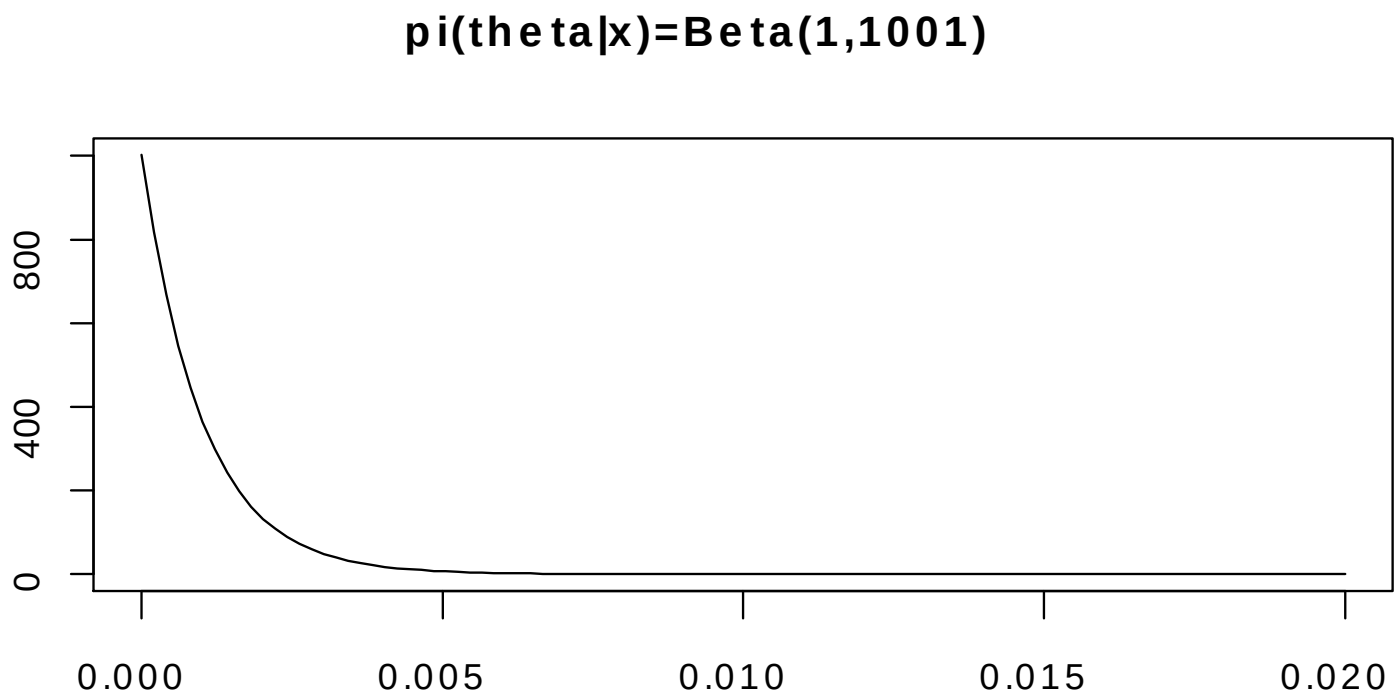

Figure: Posterior distribution $\pi(\theta|x=0)$ for $\alpha=\beta=1$

A more realistic application

(Gelman et al., 2003, p. 43) Placenta previa is an unusual condition of pregnancy in which the placenta is implanted very low in the uterus, obstructing the foetus from a normal delivery. An early study concerning the sex of placenta previa found that for a total of 980 births, 437 were female. We want to compute the posterior probability that a placenta previa birth is a female?

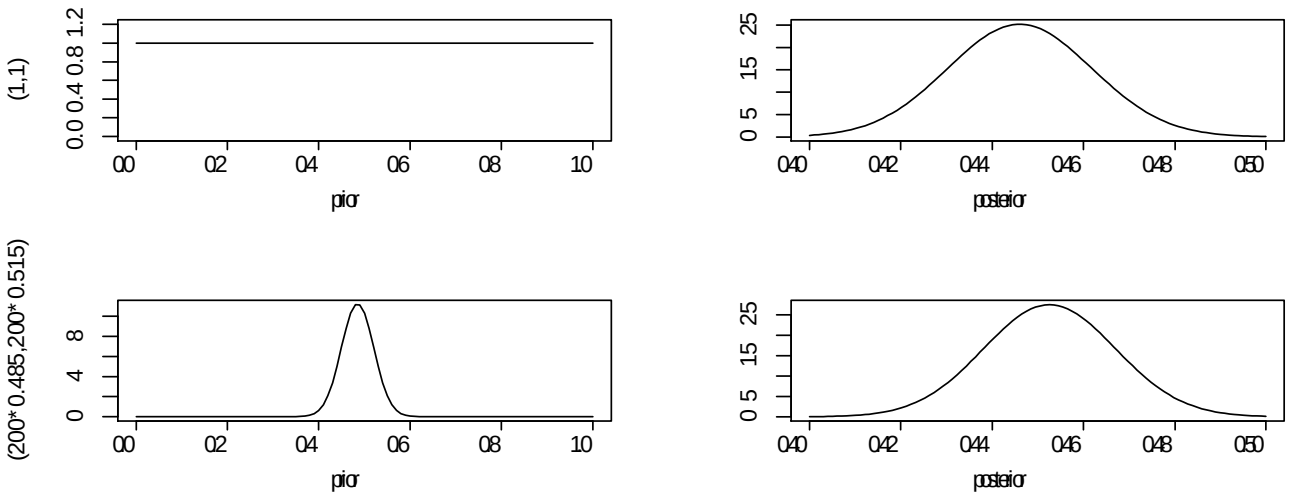

Figure: Prior (left) and posterior (right) for the placenta previa data

- For the posterior distribution $\pi(\theta|x)$, we have

$$\begin{aligned}\mathbb{E}(\theta|x) &= \frac{\alpha+x}{\alpha+\beta+m},\\ \mathbb{V}(\theta|x) &= \frac{(\alpha+x)(\beta+n-x)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}.\end{aligned}$$

- The posterior means behave asymptotically like x/n (the 'frequentist' estimator) and converge to θ_0, the 'true' value of θ; i.e. the posterior mean is an estimator of θ in the classical sense.
- The posterior variance decreases to zero as $n \longrightarrow \infty$, at rate n^{-1}: the information you get on θ gets more and more precise.
- For n large enough, the prior is washed out by the data. For a small n, the impact can have a huge impact. However, even with little or no prior information, the Bayesian analysis can lead to much more sensible answers in ill-behaved cases (e.g. political survey).

Prior Specification

- The construction of the prior distribution may be the most delicate part of Bayesian analysis. We will look at two approaches here.
- *Subjective Approach*: Prior information is available (expert knowledge, previous experiments, common sense, etc.). Expressing this prior information into a prior distribution is known as the problem of prior elicitation.
- *Objective Approach*: Prior information is not available, or too sparse to be taken into account. One must find a way to express this lack of information (for instance through an uniform distribution like in Laplace's example).

Subjective Approach: conjugate prior distributions

- **Definition**. A parametric family $\mathcal{F}$ of prior distributions is said to be conjugate for a given model $f(x|\theta)$ if and only if any prior distribution in $\mathcal{F}$ yields a posterior distribution for this model that is still in $\mathcal{F}$,

$$\pi(\theta) \in \mathcal{F} \Rightarrow \pi(\theta|x) \in \mathcal{F}.$$

- Some examples

Model	Prior	Posterior
$\mathcal{B}\text{in}(n,\theta)$	$\mathcal{B}\text{eta}(\alpha,\beta)$	$\mathcal{B}\text{eta}(\alpha+x,\beta+n-x)$
$\mathcal{G}\text{amma}(v,\theta)$	$\mathcal{G}\text{amma}(\alpha,\beta)$	$\mathcal{G}\text{amma}(\alpha+v,\beta+x)$
Gaussian	....	see next slides.

- In the Beta-Binomial model, the appeals of conjugacy where the mathematical tractability and the possibility of interpreting the prior information as one carried by an imaginary initial sample.

- Gamma distributions

$$X \sim \mathcal{G}(\alpha, \beta) \text{ with } \alpha, \beta > 0$$

$$f_{\alpha,\beta}(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) 1_{(0,\infty)}(x)$$

$$\mathbb{E}(\theta) = \frac{\alpha}{\beta}, \ \mathbb{V}(\theta) = \frac{\alpha}{\beta^2}.$$

- Particular cases are the exponential distribution $\mathcal{G}(1, \beta)$ and the chi-squared distribution χ^2_v given by $\mathcal{G}(v/2, 1/2)$.
- Clearly we have

$$\begin{aligned} \mathcal{G}(x; v, \theta)\, \mathcal{G}(\theta; \alpha, \beta) &\propto \frac{\theta^{v}}{\Gamma(v)} x^{v-1} \exp(-\theta x) . \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta) \\ &\propto \theta^{v} \exp(-\theta x) . \theta^{\alpha-1} \exp(-\beta\theta) \\ &\propto \mathcal{G}(\theta; \alpha + v, \beta + x). \end{aligned}$$

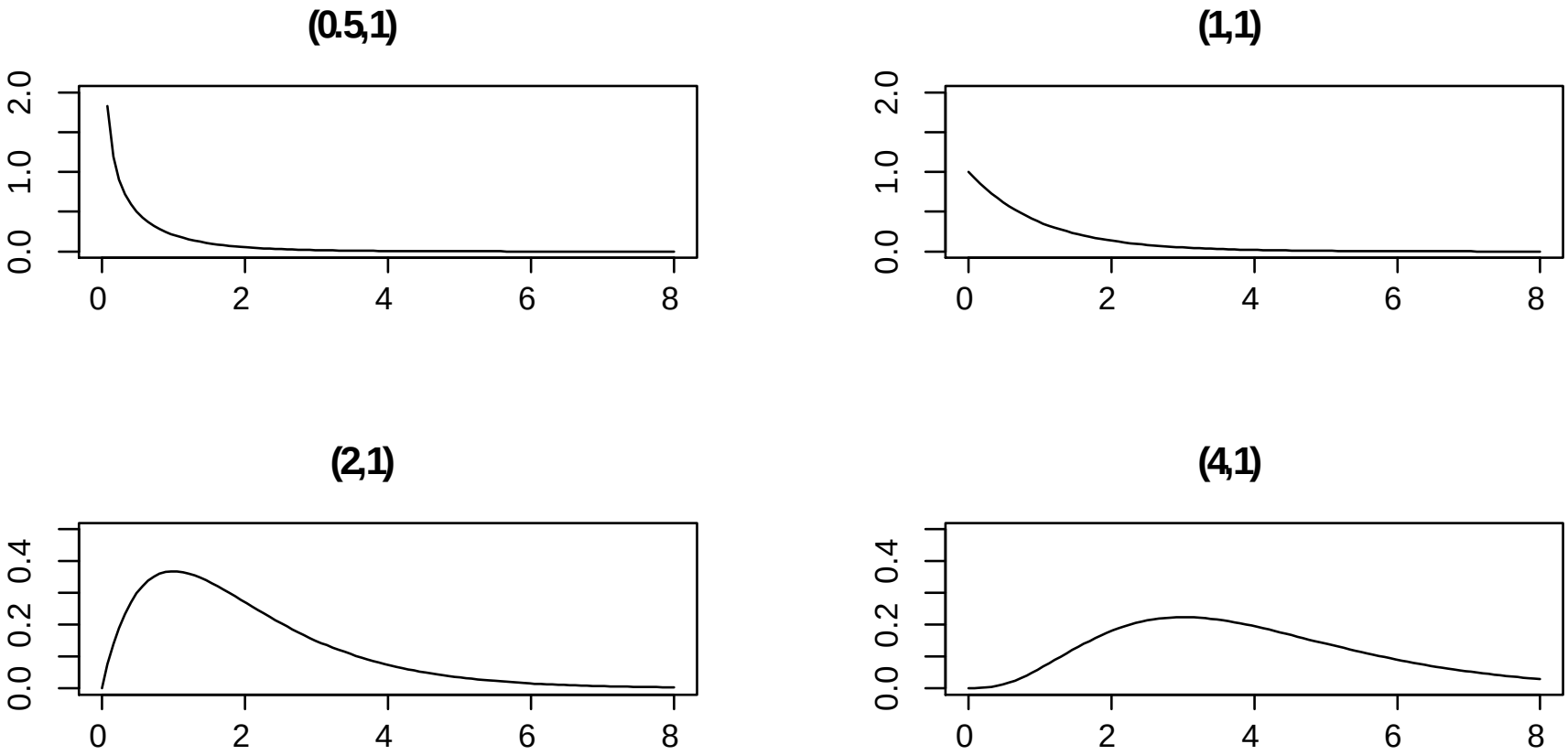

Figure: Gamma densities for various parameter settings.

Poisson Gamma model

- Assume you have some counting observations

$$X_i|\,\theta \sim \mathcal{P}\left(\theta\right), \text{ i.e. } \Pr\left(X_i = k|\,\theta\right) = \exp\left(-\theta\right)\frac{\theta^k}{k!}$$

- Assume the following prior distribution on θ

$$\theta \sim \mathcal{G}\left(\alpha, \beta\right).$$

- The posterior is given by

$$\theta|\,x_1, ..., x_n \sim \mathcal{G}\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n\right).$$

- Once more we can think of the conjugate prior as an imaginary sample of size β with α counts.

Gaussian model with unknown mean and known variance

- Assume you have
$$X_i | \theta \sim \mathcal{N}\left(\theta, \sigma^2\right)$$
- We assume the following prior distribution on θ
$$\theta \sim \mathcal{N}\left(\mu, \tau^2\right)$$
- We have
$$\theta | x_1, ..., x_n \sim \mathcal{N}\left(\mu_n, \sigma_n^2\right)$$
where
$$\sigma_n^2 = \frac{\left(\sigma^2/n\right)\tau^2}{\sigma^2/n + \tau^2}, \ \mu_n = \frac{\sigma^2\mu/n + \tau^2\overline{x}}{\sigma^2/n + \tau^2}.$$
- The posterior mean is a weighted average of μ and $\overline{x}$.
- The prior information can be interpreted as an imaginary initial sample of size σ^2/τ^2.
- When n is large, we have
$$\mu_n \approx \overline{x}, \ \sigma_n^2 \approx \sigma^2/n.$$

Gaussian model with known mean and unknown variance

- In this case we have

$$X_i | \, \theta \sim \mathcal{N}\left(\mu, 1/\theta\right), \quad \theta \sim \mathcal{G}\left(\alpha, \beta\right)$$

- The posterior is given by

$$\theta | \, x_1, ..., x_n \sim \mathcal{G}\left(\alpha_n, \beta_n\right)$$

with

$$\alpha_n = \alpha + n/2, \;\; \beta_n + \sum_{i=1}^{n} \left(x_i - \mu\right)^2 /2.$$

- When n is large, we have

$$\mathbb{E}\left[\theta | \, x_1, ..., x_n\right] = \frac{\alpha_n}{\beta_n} \approx \frac{n}{\sum_{i=1}^{n} \left(x_i - \mu\right)^2}.$$

Gaussian model with unknown mean and variance

- In this case, $\theta = \left(\mu, \sigma^2\right)$ and

$$\begin{aligned} \pi\left(\theta\right) &= \pi\left(\sigma^2\right)\pi\left(\mu\middle|\,\sigma^2\right) \\ &= \mathcal{IG}\left(\sigma^2;\alpha,\beta\right)\mathcal{N}\left(\mu;\mu_0,\frac{\sigma^2}{\kappa_0}\right) \end{aligned}$$

  where

$$\mathcal{IG}\left(x;\alpha,\beta\right) = \frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}x^{-\alpha-1}\exp\left(-\beta/x\right)1_{(0,\infty)}\left(x\right).$$

- The posterior is given by $\pi\left(\theta\middle|\,x_{1:n}\right) = \pi\left(\sigma^2\middle|\,x_{1:n}\right)\pi\left(\mu\middle|\,x_{1:n},\sigma^2\right)$ where

$$\sigma^2\middle|\,x_{1:n} \sim \mathcal{IG}\left(\sigma^2;\alpha+n/2,\beta+\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2/2+\frac{n\kappa_0}{n+\kappa_0}\left(\overline{x}-\mu_0\right)^2\right),$$

$$\mu\middle|\,x_{1:n},\sigma^2 \sim \mathcal{N}\left(\mu;\frac{\kappa_0\mu_0}{\kappa_0+n}+\frac{n\overline{x}}{\kappa_0+n},\frac{\sigma^2}{\kappa_0+n}\right)$$

- Bayesian analysis work exactly the same in the multivariate case. Now if μ is a nuisance parameter and you are just interested in σ^2, then integrate out μ and consider

Conjugate prior for the exponential family

- Many likelihood do not admit conjugate distributions BUT it is feasible when the likelihood is in the exponential family

$$f\left(x|\,\theta\right)=h\left(x\right)\exp\left(\theta^{\mathsf{T}}x-\Psi\left(\theta\right)\right)$$

- In this case the conjugate distribution is (for the hyperparameters μ,λ)

$$\pi\left(\theta\right)=K\left(\mu,\lambda\right)\exp\left(\theta^{\mathsf{T}}\mu-\lambda\Psi\left(\theta\right)\right).$$

- It follows that

$$\pi\left(\theta|\,x\right)=K\left(\mu+x,\lambda+1\right)\exp\left(\theta^{\mathsf{T}}\left(\mu+x\right)-\left(\lambda+1\right)\Psi\left(\theta\right)\right).$$

Mixture of Conjugate Priors

- If you have a prior distribution $\pi(\theta)$ which is a mixture of conjugate distributions, then the posterior is in closed form and is a mixture of conjugate distributions; i.e. with

$$\pi(\theta) = \sum_{i=1}^{K} w_i \pi_i(\theta)$$

then

$$\pi(\theta|x) = \frac{\sum_{i=1}^{K} w_i \pi_i(\theta) f(x|\theta)}{\sum_{i=1}^{K} w_i \int \pi_i(\theta) f(x|\theta) d\theta} = \sum_{i=1}^{K} w_i' \pi_i(\theta|x)$$

where

$$w_i' \propto w_i \int \pi_i(\theta) f(x|\theta) d\theta, \quad \sum_{i=1}^{K} w_i' = 1.$$

- **Theorem** (Brown, 1986): It is possible to approximate arbitrary closely any prior distribution by a mixture of conjugate distributions.

- The conjugate prior can have a strange shape or be difficult to handle.
- Consider the logistic regression model

$$\Pr\left(y=1\middle|\,\theta,x\right)=\frac{\exp\left(\theta^{\mathsf{T}}x\right)}{1+\exp\left(\theta^{\mathsf{T}}x\right)}$$

then the likelihood for n observations is conditional upon x_i's of the form

$$f\left(y_1,\ldots,y_n\middle|\,x_1,\ldots,x_n,\theta\right)=\exp\left(\theta^{\mathsf{T}}\sum_{i=1}^{n}y_ix_i\right)\prod_{i=1}^{n}\left(1+\exp\left(\theta^{\mathsf{T}}x_i\right)\right)^{-1}$$

- The conjugate prior is thus given by

$$\pi\left(\theta\right)\propto\exp\left(\theta^{\mathsf{T}}\mu\right)\prod_{i=1}^{n}\left(1+\exp\left(\theta^{\mathsf{T}}x_i\right)\right)^{-\lambda}$$

- Subjective prior distributions are often built from conjugate prior distributions. In this way, prior elicitation reduces to tune the hyperparameters according to the available prior information.
- This approach has several drawbacks
  - Outside simple models, conjugate prior distributions are rarely available.
  - Mathematical convenience does not mean practical relevance.
  - Approximation by mixtures feasible but very tiedous and almost never used in practice.

Objective Approach

- In the probability of male birth problem, Laplace used the uniform distribution on $[0, 1]$. Two questions arise from this particular choice.
- How can we generalize the uniform distribution to cases where the parameter is defined on an infinite interval?
- Is the uniform distribution the best choice to represent the absence of knowledge about θ? For instance, let $\theta' = \theta^2$ and reexpress the model in terms of θ'

$$\Pr\left(X = x \middle| \theta'\right) = \binom{n}{x} \theta'^{x/2} \left(1 - \theta'^{1/2}\right)^{n-x}.$$

Clearly the model for X is unchanged, and we have no more prior information on θ' than θ. Yet if we assign to θ' an uniform distribution, we will get different results! This is the reparameterisation issue.

Improper Prior Distributions

- **Definition**: A prior density $\pi(\theta)$ is said to be improper if and only if it does not integrate to a finite value, that is

$$\int \pi(\theta)\, d\theta = +\infty$$

- A Bayesian model including an improper prior density is considered as valid provided that the corresponding posterior is proper, that is

$$\int f(x|\theta)\, \pi(\theta)\, d\theta < +\infty.$$

- **Example**: Consider the Gaussian model with unknown mean and known variance $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$ then the following improper prior density

$$\pi(\theta) \propto 1,$$

yields a proper density

$$\pi(\theta|x_{1:n}) = \mathcal{N}(\theta; \overline{x}, \sigma^2/n).$$

This simply corresponds to the case where $\tau^2 \to \infty$ in a conjugate

Reparametrisation issues: location parameter

- If the likelihood is of the form

$$f\left(x|\,\theta\right) = f\left(x-\theta\right)$$

  then θ is a *location parameter*, as such a model is translation invariant.
- A natural requirement is that the prior distribution enjoys the same property, that is π

$$\pi\left(\theta\right) = \pi\left(\theta-\theta_0\right)$$

  for every θ_0. The solution is an improper prior

$$\pi\left(\theta\right) \propto 1.$$

- **Example**: Gaussian distribution with known variance and unknown mean.

Reparametrisation issues: scale parameter

- If the likelihood is of the form

$$f\left(x|\,\theta\right) = 1/\theta f\left(x/\theta\right)$$

  then θ is a *scale parameter*. Such a model is scale invariant.
- A natural requirement is that the prior distribution enjoys the same property, that is the prior distribution should satisfy

$$\pi\left(\theta\right) = \frac{1}{c}\pi\left(\frac{\theta}{c}\right)$$

  for any $c > 0$. The solution is an improper prior

$$\pi\left(\theta\right) \propto 1/\theta.$$

- **Example**: Gaussian distribution with unknown variance and known mean.

Jeffreys' Prior

- **Definition**: For a given parametric model $f\left(x|\,\theta\right)$, the Jeffreys prior is defined by the density

$$\pi_J\left(\theta\right) \propto \left\{\det\left[I\left(\theta\right)\right]\right\}^{1/2}$$

where $I\left(\theta\right)$ is the Fisher information matrix,

$$I\left(\theta\right) = \mathbb{E}_\theta\left[\frac{\partial \log f\left(X|\,\theta\right)}{\partial\theta}\frac{\partial \log f\left(X|\,\theta\right)}{\partial\theta^{\mathsf{T}}}\right].$$

- **Property**: Jeffreys prior is invariant by reparametrisation.

- *Proof*: Let $\eta = h(\theta)$ where h is a one-to-one mapping, and assume $\Theta \subset \mathbb{R}$, then

$$\pi(\eta) = \pi_J\left(h^{-1}(\eta)\right)\left|\frac{dh^{-1}(\eta)}{d\eta}\right| = \pi_J(\theta)\left|\frac{d\theta}{d\eta}\right| \propto |I(\eta)|^{1/2} = \pi_J(\eta)$$

as

$$\begin{aligned} I(\eta) &= -\mathbb{E}_\eta\left[\frac{\partial^2 \log f(X|\eta)}{\partial\eta^2}\right] = -\mathbb{E}_\theta\left[\frac{\partial^2 \log f(X|\theta)}{\partial\theta^2}.\left|\frac{d\theta}{d\eta}\right|^2\right] \\ &= I(\theta)\left|\frac{d\theta}{d\eta}\right|^2. \end{aligned}$$

- *Example*: For $X_i \sim \mathcal{N}(\theta, \sigma^2)$ then

$$f(x_{1:n}|\theta) \propto \exp\left(-\frac{n(\overline{x}-\theta)^2}{2\sigma^2}\right) \Rightarrow \frac{\partial^2 \log f(x_{1:n}|\theta)}{\partial\theta^2} = -\frac{n}{\sigma^2}$$

thus

$$\pi_J(\theta) \propto 1.$$

- The location invariant prior is the Jeffreys' prior for any location parameter. The scale invariant prior is the Jeffreys' prior for any scale parameter.
- Jeffreys' prior is appealing in that it solves completely the issue of reparametrisation. It is invariant through any transformation of the parameter.
- The fact that it is improper in most cases supports the idea that Jeffreys's prior is noninformative.
- Unfortunately, Jeffreys' prior can lead to incoherences and paradoxes in multivariate problems and does not satisfy the likelihood principle.
- **Example**. Jeffreys' prior for the Gaussian model with unknown mean and variance is

$$\pi_J(\mu, \sigma) \propto \sigma^{-2}.$$

For such a model, it seems more natural to have

$$\pi_J(\mu, \sigma) \propto \sigma^{-1}$$

as μ is a location parameter and σ is a scale parameter.

More on Bayesian Inference

- Bayesian inference satisfies the likelihood principle if you don't build your prior using $f(x|\theta)$.
- If $f_2(x|\theta) = cf_1(x|\theta)$ then

$$\frac{f_2(x|\theta)\pi(\theta)}{\int f_2(x|\theta)\pi(\theta)\,d\theta} = \frac{f_1(x|\theta)\pi(\theta)}{\int f_1(x|\theta)\pi(\theta)\,d\theta}.$$

- If $f(x|\theta) = h(x)\,g(T(x)|\theta)$ then

$$\frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\,d\theta} = \frac{g(T(x)|\theta)\pi(\theta)}{\int g(T(x)|\theta)\pi(\theta)\,d\theta}.$$

Variance Decomposition

- By using the information x, you might expect that $\mathbb{V}(\theta|x) \leq \mathbb{V}(\theta)$ but this wrong, this is only true on average over the distribution of X.
- You have

$$\mathbb{E}(\mathbb{V}(\theta|X)) = \mathbb{V}(\theta) - \mathbb{V}(\mathbb{E}[\theta|X]) \leq \mathbb{V}(\theta).$$

- *Proof.* We have

$$\begin{aligned}
\mathbb{V}(\theta) &= \mathbb{E}(\theta^2) - \mathbb{E}^2(\theta) \\
&= \mathbb{E}(\mathbb{E}(\theta^2|X)) - \mathbb{E}(\mathbb{E}(\theta|X))^2 \\
&= \mathbb{E}(\mathbb{E}(\theta^2|X)) - \mathbb{E}(\mathbb{E}^2(\theta|X)) \\
&\quad + \mathbb{E}(\mathbb{E}^2(\theta|X)) - \mathbb{E}(\mathbb{E}(\theta|X))^2.
\end{aligned}$$

Predictive Distribution

- Assume you have a Bayesian model $\pi(\theta)$, $f(x|\theta)$. Having observed $x_1, ..., x_m$, you want to predict $Y|\theta \sim g(y|\theta)$.
- A plug-in estimate consists of picking $\widehat{\theta}_{MLE}$ and estimating the distribution of Y through

$$g\left(y|\widehat{\theta}_{MLE}\right).$$

This estimate of the distribution of Y is over-confident, especially if we do not have access to many data $x_{1:n}$.

- In a Bayesian approach, the predictive distribution is simply given by

$$\begin{aligned} g(y|x) &= \int \pi(y,\theta|x)\,d\theta = \int \pi(y|x,\theta)\,\pi(\theta|x)\,d\theta \\ &= \int g(y|\theta)\,\pi(\theta|x)\,d\theta. \end{aligned}$$

- As $n \to \infty$ then $\pi(\theta|x) \to \delta_{\widehat{\theta}_{MLE}}(\theta)$ and $g(y|x) \to g\left(y|\widehat{\theta}_{MLE}\right)$.

- *Example*. Beta-Binomial model with $f\left(x\middle|\theta\right) = \mathcal{B}\text{in}(x; n, \theta)$ and $\pi\left(\theta\right) = \mathcal{B}e\left(\alpha, \beta\right)$, we have $\widehat{\theta}_{MLE} = \frac{x}{n}$ and

$$\pi\left(\theta\middle|x\right) = \mathcal{B}e\left(\alpha + x, \beta + n - x\right).$$

- If we are interested in the predictive distribution of $Y\middle|\theta \sim \textit{Bernoulli}\left(\theta\right)$, then

$$g\left(y = 1\middle|\widehat{\theta}_{MLE}\right) = \widehat{\theta}_{MLE} = \frac{x}{n},$$

whereas

$$\begin{aligned} g\left(y = 1\middle|x\right) &= \int \theta \mathcal{B}e\left(\theta; \alpha + x, \beta + n - x\right) d\theta \\ &= \frac{\alpha + x}{\alpha + \beta + n}. \end{aligned}$$

- *Example*. Consider $X_1 | \theta \sim \mathcal{N}\left(\theta, \sigma^2\right)$ and $\theta \sim \mathcal{N}\left(m_0, \sigma_0^2\right)$ then

$$\theta | x_1 \sim \mathcal{N}\left(m_1, \sigma_1^2\right)$$

with

$$\begin{aligned} \frac{1}{\sigma_1^2} &= \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2} \Rightarrow \sigma_1^2 = \frac{\sigma_0^2 \sigma^2}{\sigma_0^2 + \sigma^2}, \\ m_1 &= \sigma_1^2 \left(\frac{x_1}{\sigma^2} + \frac{m}{\sigma_0^2} \right). \end{aligned}$$

- To predict the distribution of a new observation $X | \theta \sim \mathcal{N}\left(\theta, \sigma^2\right)$ in light of x_1 we use the predictive distribution

$$f\left(x | x_1\right) = \int f\left(x | \theta\right) \pi\left(\theta | x_1\right) d\theta$$

- We can do direct calculations or alternatively use the fact that $f\left(x | x_1\right)$ is Gaussian so is characterized by its mean and variance

$$\begin{aligned} \mathbb{E}\left[X | x_1\right] &= \mathbb{E}\left[\theta + V | x_1\right] = \mathbb{E}\left[\theta | x_1\right] = m_1, \\ \mathbb{V}\left[X | x_1\right] &= \mathbb{V}\left[\theta + V | x_1\right] = \mathbb{V}\left[\theta | x_1\right] + \mathbb{V}\left[V\right] = \sigma_1^2 + \sigma^2. \end{aligned}$$

Sequential Bayesian Estimation

- Now assume that you observe a realization x_2 of $X_2|\theta \sim \mathcal{N}\left(\theta, \sigma^2\right)$. Then you are interested now in

$$\begin{aligned}\pi\left(\theta|x_1, x_2\right) &\propto f\left(x_2|\theta\right) f\left(x_1|\theta\right) \pi\left(\theta\right) \\ &\propto f\left(x_2|\theta\right) \pi\left(\theta|x_1\right) \\ &\propto f\left(x_1|\theta\right) \pi\left(\theta|x_2\right).\end{aligned}$$

- Updating the prior one observation at a time, or all observations together, does not matter.
- The sequential approach can be useful for massive dataset. In this case at time n,

$$\pi\left(\theta|x_1, ..., x_n\right) \propto f\left(x_n|\theta\right) \pi\left(\theta|x_1, ..., x_{n-1}\right);$$

i.e. 'the prior at time n is the posterior at time $n-1$'.

Hypothesis Testing

- Consider (again!) the Beta-Binomial model $f\left(x\middle|\theta\right)=\mathcal{B}\text{in}(x;n,\theta)$ and $\pi\left(\theta\right)=\mathcal{B}e\left(\alpha,\beta\right)$.
- Assume we want to test $H_0:\theta\geq\frac{1}{2}$ vs $H_1:\theta<\frac{1}{2}$.
- In a Bayesian approach, you can simply compute

$$\pi\left(H_0\middle|x\right)=1-\pi\left(H_1\middle|x\right)=\int_{1/2}^{1}\pi\left(\theta\middle|x\right)d\theta.$$

- Contrary to frequentists, your test is not based on observations you do not observe.

- In general, we want to compare two hypothesis: $H_0 : \theta \sim \pi_0$ versus $H_1 : \theta \sim \pi_1$. You can think of it as putting a prior given by

$$\begin{aligned} \pi(\theta) &= \pi(H_0)\,\pi(\theta|\,H_0) + \pi(H_1)\,\pi(\theta|\,H_1) \\ &= \pi(H_0)\,\pi_0(\theta) + \pi(H_1)\,\pi_1(\theta) \end{aligned}$$

  where $\pi(H_0) + \pi(H_1) = 1$.
- In the previous example, $\pi_0(\theta) = \mathcal{U}\left(\theta; \left[\frac{1}{2}, 1\right]\right)$ and $\pi_1(\theta) = \mathcal{U}\left(\theta; \left[0, \frac{1}{2}\right)\right)$ and $\pi(H_0) = \pi(H_1) = \frac{1}{2}$.
- To compare H_0 versus H_1, we can compute the posterior probabilities of H_0 and H_1 which are of the form

$$\begin{aligned} \pi(H_i|\,x) &= \frac{\pi(x|\,H_i)\,\pi(H_i)}{\pi(x)} \\ &= \frac{\pi(x|\,H_i)\,\pi(H_i)}{\pi(x|\,H_0)\,\pi(H_0) + \pi(x|\,H_1)\,\pi(H_1)}. \end{aligned}$$

- Clearly this approach can be extended straightforwardly to the multi-hypothesis case.

- To compare H_0 versus H_1, we typically compute the *Bayes factor* which partially eliminated the influence of the prior modelling (i.e. $\pi(H_i)$!)

$$\begin{aligned} B_{10}^{\pi} &= \frac{\pi(x|H_1)}{\pi(x|H_0)} = \frac{\pi(H_1|x)}{\pi(H_0|x)} \frac{\pi(H_0)}{\pi(H_1)} \\ &= \frac{\int f(x|\theta)\,\pi(\theta|H_1)\,d\theta}{\int f(x|\theta)\,\pi(\theta|H_0)\,d\theta} \\ &= \frac{\int f(x|\theta)\,\pi_1(\theta)\,d\theta}{\int f(x|\theta)\,\pi_0(\theta)\,d\theta}. \end{aligned}$$

- For realistic models, these integrals can be very difficult to compute.

- Jeffreys' scale of evidence says that

  - if $\log_{10}\left(B_{10}^{\pi}\right)$ varies between 0 and 0.5, the evidence against H_0 is poor,
  - if it is between 0.5 and 1, it is substantial,
  - if it is between 1 and 2, it is strong, and
  - if it is above 2, it is decisive.

- Bayes factor tell you where one should prefer H_0 to H_1: it does NOT tell you whether these models are sensible!

- Note that Bayes procedures can be directly used to test point null hypothesis; i.e. $H_0 : \theta = \theta_0$ (that is $\pi_0(\theta) = \delta_{\theta_0}(\theta)$) versus $H_1 : \theta \sim \pi_1$ where the prior is then defined as

$$\pi(\theta) = \pi(H_0)\,\delta_{\theta_0}(\theta) + \pi(H_1)\,\pi_1(\theta)$$

- The associated Bayes factor is simply

$$B_{10}^{\pi} = \frac{\pi(x|H_1)}{\pi(x|H_0)} = \frac{\int f(x|\theta)\,\pi_1(\theta)\,d\theta}{f(x|\theta_0)}.$$

- *Example*: Assume you have an coin, you toss it n times and gets x heads. Is it biased?
- Let θ be the proba of having a head then we can test $H_0 : \theta = \frac{1}{2}$ versus $H_1 : \theta \sim \mathcal{U}\left(\frac{1}{2}, 1\right]$ using

$$B_{10}^{\pi} = \frac{2\int_{\frac{1}{2}}^{1} \theta^x (1-\theta)^{n-x} \, d\theta}{\left(\frac{1}{2}\right)^x \left(1-\frac{1}{2}\right)^{n-x}}$$

or $H_0 : \theta = \frac{1}{2}$ versus $H_1 : \theta \sim \mathcal{U}[0, 1]$ using

$$B_{10}^{\pi} = \frac{\int_0^1 \theta^x (1-\theta)^{n-x} \, d\theta}{\left(\frac{1}{2}\right)^x \left(1-\frac{1}{2}\right)^{n-x}} = \frac{2^n \Gamma(x+1)\Gamma(n+1-x)}{\Gamma(n+2)}.$$

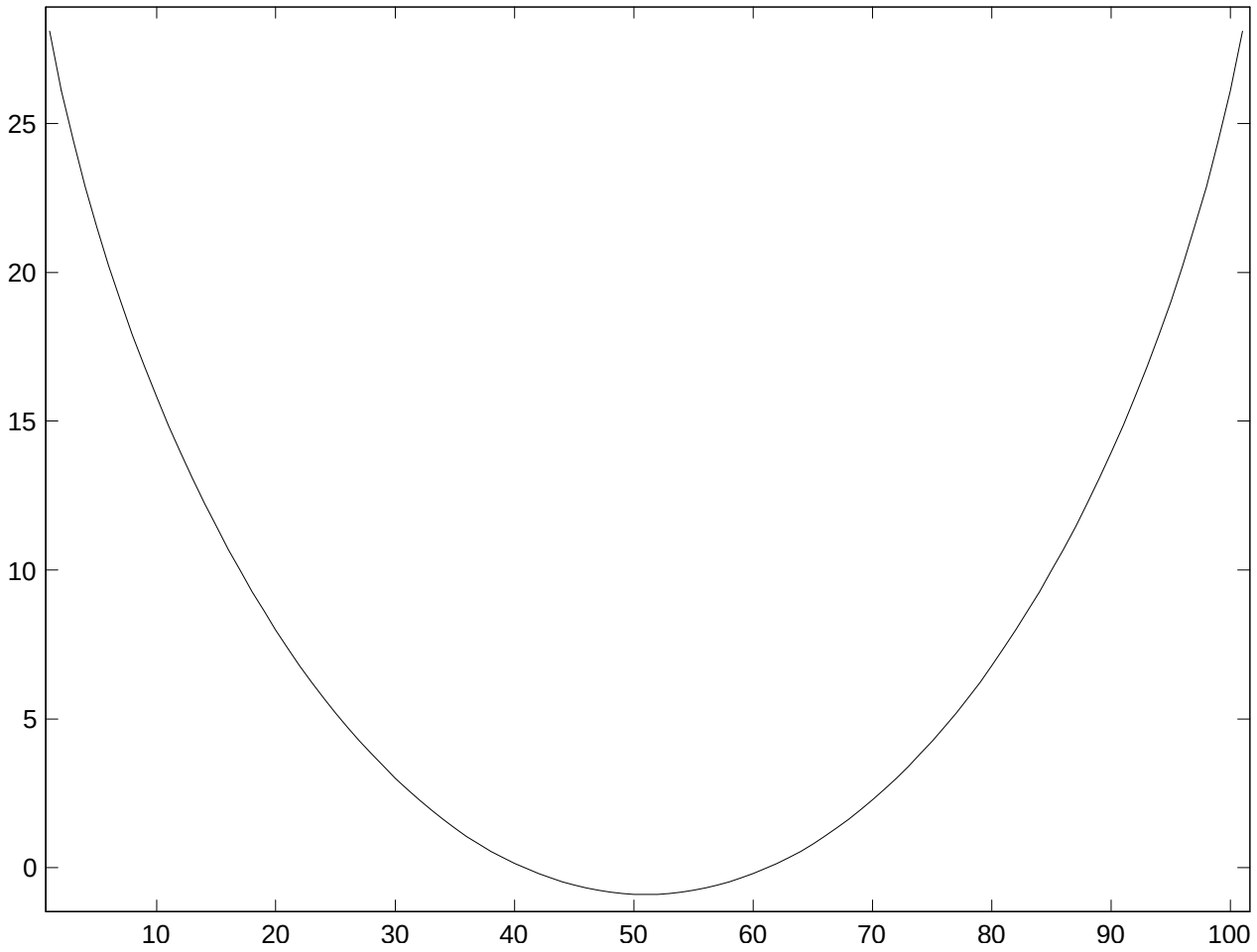

Figure: $\log_{10}\left(B_{10}^{\pi}\right)$ for $n = 100$ and $x = 0, ..., n$.

- *Example*: Gaussian model

$$X_i|\,\theta \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 1),$$
$$\pi_0(\theta) = \delta_0(\theta),\ \pi_1(\theta) = \mathcal{N}(0, \tau^2)$$

- In this case we have

$$B_{01}^{\pi} = \frac{\pi(x|\,H_0)}{\pi(x|\,H_1)} = (1+\tau^2)^{n/2} \exp\left(-\frac{\tau^2 \left(\sum_{i=1}^n x_i\right)^2}{2(1+\tau^2)}\right).$$

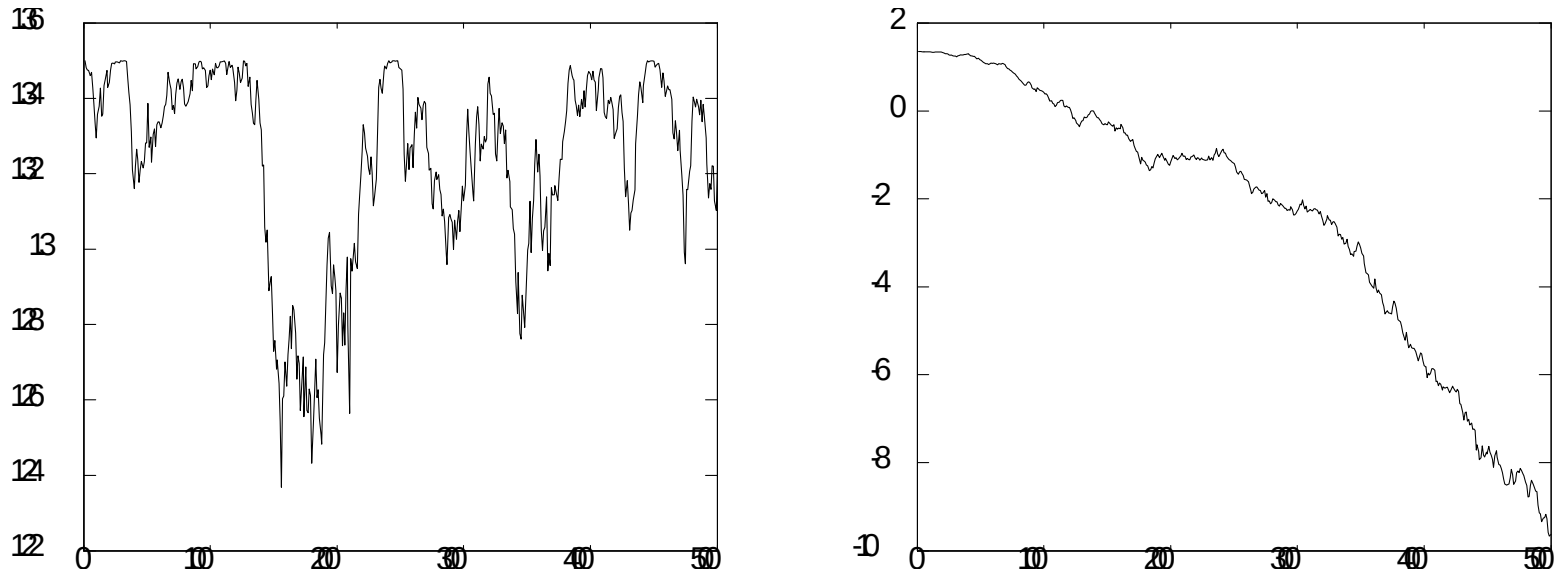

Figure: $\log_{10}(B_{01})$ for data generated from $\mathcal{N}(0,1)$ (left) and $\mathcal{N}(0.25,1)$ (right)

- Bayes factors are not limited to the comparison of models with the same parameter space.
- Assume you have some data and two statistical models.
- Under H_0, $\theta_0 \in \Theta_0$, the prior is $\pi_0(\theta_0)$ and the likelihood is $f_0(x|\theta_0)$. Under H_1, $\theta_1 \in \Theta_1$, the prior is $\pi_1(\theta_1)$ and the likelihood is $f_1(x|\theta_1)$ then

$$B_{10}^{\pi} = \frac{\pi(x|H_1)}{\pi(x|H_0)} = \frac{\int f_1(x|\theta_1)\,\pi_1(\theta_1)\,d\theta_1}{\int f_0(x|\theta_0)\,\pi_0(\theta_0)\,d\theta_0}$$

- One can have $\Theta_0 = \mathbb{R}$ and $\Theta_1 = \mathbb{R}^{1000}$.
- In this case you have a parameter space $\Theta = \Theta_0 \cup \Theta_1$ which is the union two subspaces of different dimensions.

- We can straightforwardly extend this hypothesis approach to model selection.
- Assume you have a countable collection of hypothesis/models $\{H_i\}$.
- For each model $\mathcal{M}_i$, you have a prior $\pi_i(\theta_i)$ on Θ_i and a likelihood function $f_i(x|\theta_i)$.
- You attribute a prior probability $\pi(i)$ to each hypothesis/model H_i.
- The parameter space is $\bigcup_i \{i\} \times \Theta_i$ and the prior on this space is

$$\pi(i, \theta) = \pi(i)\, \pi_i(\theta_i).$$

Lindleys' paradox

- Testing hypothesis in a Bayesian way is attractive.... but be careful to vague priors!!!
- Assume you have $X|\left(\mu,\sigma^2\right)\sim\mathcal{N}\left(\mu,\sigma^2\right)$ where σ^2 is assumed known but μ (the parameter θ) is unknown. We want to test $H_0:\mu=0$ vs $H_1:\mu\sim\mathcal{N}\left(0,\tau^2\right)$ then

$$\begin{aligned} B_{10}^{\pi}\left(x\right) &= \frac{\pi\left(x|\,H_1\right)}{\pi\left(x|\,H_0\right)}=\frac{\int\mathcal{N}\left(x;\mu,\sigma^2\right)\mathcal{N}\left(\mu;0,\tau^2\right)d\mu}{f\left(x|\,0\right)} \\ &= \frac{\sigma}{\sqrt{\sigma^2+\tau^2}}\exp\left(\frac{\tau^2x^2}{2\sigma^2\left(\sigma^2+\tau^2\right)}\right)\underset{\tau^2\to\infty}{\to}0 \end{aligned}$$

- Vague priors should be banned for Bayesian hypothesis testing/model selection.
- Alternative apporaches have been proposed to overcome this problem. For example, intrinsic Bayes factors consist in replacing the improper prior by a proper posterior distribution computed using a small number of data points.

Confidence Regions

- We are interested in deriving *confidence regions*, that is regions that contain the 'most likely values' for the parameter.
- In classical statistics, and for an univariate problem, the confidence interval at level α is of the form

$$\left[\widehat{\theta} - z_{\alpha/2}\widehat{\sigma}, \widehat{\theta} + z_{\alpha/2}\widehat{\sigma}\right]$$

  where $\widehat{\theta}$ is the classical estimator (say MLE) and $\widehat{\sigma}$ is an estimate of its standard deviation.
- In this frequentist perspective, the true value of the parameter is fixed, and the confidence interval is random, having a probability of $(1-\alpha)$ to actually contain this true value (when we repeat the same experiment a great number of times).
- It is not possible to interpret $(1-\alpha)$ as the probability that the parameter lies in the confidence interval for the considered experiment.

Credibility Regions

- The Bayesian viewpoint allows for a conceptually simpler determination of 'confidence regions'.
- **Definition**: A subset C of the parameter space is a $100\,(1-\alpha)\,\%$ credible region for θ if and only if

$$\pi\,(\theta \in C|\,x) = 1-\alpha.$$

- There is usually obviously an infinity of credible regions for a given level.
- One may restrict ourselves to the credible interval centred at a given Bayesian point estimate (e.g. posterior mean, median, etc.) but this is arbitrary and does not make much sense when the posterior is not symmetric around this particular value. Such intervals may even not exist.
- *Example*: Assume that the posterior distribution is $\mathcal{B}e\,(1,30)$ then the posterior mean is $1/31$, and the posterior probability of being above $1/31$ is approx. 0.37, therefore credible intervals centred at $1/31$ exists only if $1-\alpha$ is smaller than 0.74.

Highest posterior density regions

- A more satisfactory approach is to restrict our attention to the credible set that countains the 'most likely values'/
- **Definition**: The subset $C_\alpha(x)$ of the parameter space is a highest posterior density (HPD) region at level α if and only if it is of the form

$$C_\alpha(x) = \{\theta \in \Theta : \pi(\theta|x) > \gamma\}$$

where γ is chosen so that $\pi(\theta \in C_\alpha(x)|x) = 1 - \alpha$.

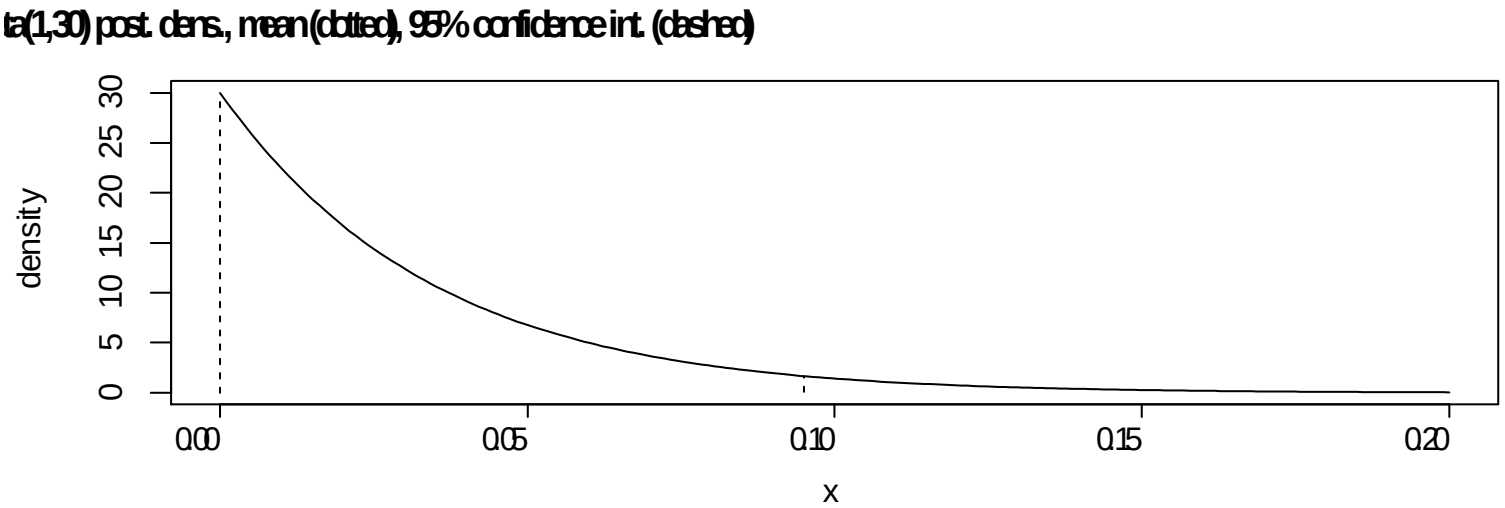

Figure: HPD for the Beta$(1, 30)$ and $\alpha = 0.05$.

- In well-behaved cases, there exists exactly one HPD region for a given confidence level.
- One can show that the HPD region at level α is the credible interval of minimal width (or minimal surface/volume in a bivariate/multivariate setting).
- The issue of setting α remains, with the obvious trade-off that small values give large regions, while large values lead to restrictive regions.

The Regression Problem

- Regression problem: Determining the relationship between some response variable Y and a set of predictor variables $\mathbf{X} = (X_1, ..., X_p)$.
- The most common form of structural assumption is that the responses are assumed to be related through some deterministic function f and some additive random error component ϵ so that

$$Y = f(\mathbf{X}) + \epsilon$$

  where ϵ is a zero-mean error distribution.
- Typically X is observed and we have

$$\mathbb{E}[Y | \mathbf{X} = \mathbf{x}] = f(\mathbf{x}).$$

- We are interested in determining f over some range of plausible predictor values.

- We have no way of determining its analytic form exactly, even if one actually exists.
- We must content ourselves with finding approximations which are close to the truth.
- To do this we must make use of the observed dataset $D = \{y_i, \mathbf{x}_i\}_{i=1}^{n}$.
- A simple and much used approximation of f consists of using

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

- More generall, we shall make use of the more general basis function models

$$g(\mathbf{x}) = \sum_{i=1}^{k} \beta_i B_i(\mathbf{x});$$

the linear model being just a special case where $k = p + 1$, $B_1(\mathbf{x}) = 1$ and $B_i(\mathbf{x}) = x_{i-1}$.

Bayesian Linear Model

- We have approximated f by g so we get the model

$$y_i = \sum_{j=1}^{k} \beta_j B_j(\mathbf{x}_i) + \epsilon_i, \quad i = 1, ..., n$$

- In matrix notation

$$\mathbf{Y} = \mathbf{B}\beta + \boldsymbol{\epsilon}$$

where $\mathbf{Y} = (y_1, ..., y_n)^{\mathsf{T}}$, $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^{\mathsf{T}}$ and the design matrix

$$\mathbf{B} = \begin{pmatrix} B_1(\mathbf{x}_1) & \cdots & B_k(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ B_1(\mathbf{x}_n) & \cdots & B_k(\mathbf{x}_n) \end{pmatrix}$$

- We make the assumption that

$$\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$$

This defines the likelihood

$$f\left(D|\,\beta, \sigma^2\right) = \mathcal{N}\left(\mathbf{Y}; \mathbf{B}\beta, \sigma^2 I_n\right).$$

- The MLE of $\left(\beta, \sigma^2\right)$ is given by

$$\begin{aligned}
\widehat{\beta} &= \left(\mathbf{B}^{\mathsf{T}}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathsf{T}}\mathbf{Y}, \\
\widehat{\sigma}^2 &= \frac{1}{n}\left(\mathbf{Y} - \mathbf{B}\widehat{\beta}\right)^{\mathsf{T}}\left(\mathbf{Y} - \mathbf{B}\widehat{\beta}\right).
\end{aligned}$$

The MLE of β can be extremely unstable as $\mathbf{B}^{\mathsf{T}}\mathbf{B}$ is becoming singular when covariates are co-linear.

- We consider the following prior distribution

$$\pi\left(\beta,\sigma^2\right)=\pi\left(\beta|\,\sigma^2\right)\pi\left(\sigma^2\right)$$

where

$$\begin{aligned}\pi\left(\sigma^2\right) &= \mathcal{IG}\left(\sigma^2;a,b\right),\\ \pi\left(\beta|\,\sigma^2\right) &= \mathcal{N}\left(\beta;m,\sigma^2V\right).\end{aligned}$$

- We can show that the posterior satisfies

$$\pi\left(\beta,\sigma^2\middle|\,D\right)=\mathcal{N}\left(\beta;m^*,\sigma^2V^*\right)\mathcal{IG}\left(\sigma^2;a^*,b^*\right)$$

where

$$\begin{aligned}V^* &= \left(V^{-1}+\mathbf{B}^{\mathsf{T}}\mathbf{B}\right)^{-1}\\ m^* &= V^*\left(V^{-1}m_0+\mathbf{B}^{\mathsf{T}}\mathbf{Y}\right)^{-1},\\ a^* &= a+n/2,\\ b^* &= b+\left(m^{\mathsf{T}}V^{-1}m+\mathbf{Y}^{\mathsf{T}}\mathbf{Y}-m^{*\mathsf{T}}V^{*-1}m^*\right)/2.\end{aligned}$$

- Clearly for vague priors, we have

$$m^* \longrightarrow \widehat{\beta}$$

- Also $\mathbf{Y}^{\mathsf{T}}\mathbf{Y} - m^{*\mathsf{T}} V^{*-1} m^*$ is equal to the residual sum of squares when $a, b \longrightarrow 0$.
- Assuming we are interested in predicting y at a new design point $\mathbf{x}$, we have

$$\begin{aligned} p\left(y \middle| D, x\right) &= \int \mathcal{N}\left(y; B^{\mathsf{T}}\left(\mathbf{x}\right)\beta, \sigma^2\right) \pi\left(\beta, \sigma^2 \middle| D\right) d\beta d\sigma^2 \\ &= St\left(y; B^{\mathsf{T}}\left(\mathbf{x}\right) m^*, b^*\left(I + B^{\mathsf{T}}\left(\mathbf{x}\right) V^* B\left(\mathbf{x}\right)\right), a^*\right) \end{aligned}$$

where for $Y \sim St\left(\mu, v, c\right)$

$$p\left(y\right) = \frac{\Gamma\left(\frac{1}{2}\left(c+1\right)\right)}{\Gamma\left(\frac{1}{2}c\right)\sqrt{v\pi}} \left\{1 + \frac{\left(y-\mu\right)^2}{v}\right\}^{-(c+1)/2}$$

which has mean μ and variance $v/\left(c-2\right)$ (for $c > 2$), c being known at the degrees of freedom.

- The Bayesian linear model with uncertainty in the regression variance leads to Student predictive distributions.
- Student distributions have thicker tails than Gaussian for the same location and scale parameters, hence we can expect greater robustness.
- For reasonable samples size, $n > 100$ then $a^* > 100$ and the Student and Gaussian are almost similar.

Bayesian Model Selection

- It is often the case that we have a competing number of Bayesian linear models/hypothesis say $\mathcal{M}_1, ..., \mathcal{M}_M$ where

$$\mathcal{M}_i : \mathbf{Y} = \mathbf{B}_i \beta_i + \boldsymbol{\epsilon}$$

with $\pi_i\left(\beta, \sigma^2\right) = \mathcal{N}\left(\beta; m_i, \sigma^2 V_i\right) \mathcal{IG}\left(\sigma^2; a, b\right)$ where m_i is of dimension k_i

- We can put a prior distribution on $\biguplus_{i=1}^{m} \{\mathcal{M}_i\} \times \Theta_i$ of the form

$$\pi\left(\mathcal{M}, \theta\right) = \pi\left(\mathcal{M}\right) \pi_{\mathcal{M}}\left(\theta\right)$$

to establish the expression of

$$\pi\left(\mathcal{M}, \theta \middle| D\right) = \frac{f\left(D \middle| \mathcal{M}, \theta\right) \pi_{\mathcal{M}}\left(\theta\right) \pi\left(\mathcal{M}\right)}{\sum_{i=1}^{M} \int_{\Theta_i} f\left(D \middle| \mathcal{M}_i, \theta\right) \pi_{\mathcal{M}_i}\left(\theta\right) d\theta . \pi\left(\mathcal{M}_i\right)}$$

from which we can deduce

$$\pi\left(\mathcal{M} \middle| D\right) = \frac{\pi\left(D \middle| \mathcal{M}\right) \pi\left(\mathcal{M}\right)}{\sum_{i=1}^{M} \pi\left(D \middle| \mathcal{M}_i\right) \pi\left(\mathcal{M}_i\right)}.$$

- In the Bayesian linear model case, we have

$$\begin{aligned}\pi\left(D|\,\mathcal{M}_i\right) &= \int f\left(D|\,\beta,\sigma^2,\mathcal{M}_i\right)\pi\left(\beta,\sigma^2\middle|\,\mathcal{M}_i\right)d\beta d\sigma^2 \\ &= \frac{b^a}{\pi^{n/2}\Gamma\left(a\right)}\frac{\left|V_i^*\right|^{1/2}\Gamma\left(a_i^*\right)}{\left|V_i\right|^{1/2}\left(b_i^*\right)^{a_i^*}}.\end{aligned}$$

- When computing the Bayes factor we obtain

$$\frac{\pi\left(D|\,\mathcal{M}_i\right)}{\pi\left(D|\,\mathcal{M}_j\right)} = \frac{\left|V_j\right|^{1/2}\left|V_i^*\right|^{1/2}\Gamma\left(a_i^*\right)\left(b_j^*\right)^{a_j^*}}{\left|V_i\right|^{1/2}\left|V_j^*\right|^{1/2}\Gamma\left(a_j^*\right)\left(b_i^*\right)^{a_i^*}}.$$

- Essentially Bayesian model selection can be performed analytically in the linear model case... when not too many models have to be assessed.

- Example: Consider the following polynomial regression problem where for any $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$.

$$\mathcal{M}_M : y_i = \sum_{k=0}^{M} \beta_k x_i^k + \epsilon_i$$

- If j is too large then there will be overfitting if we use MLE.

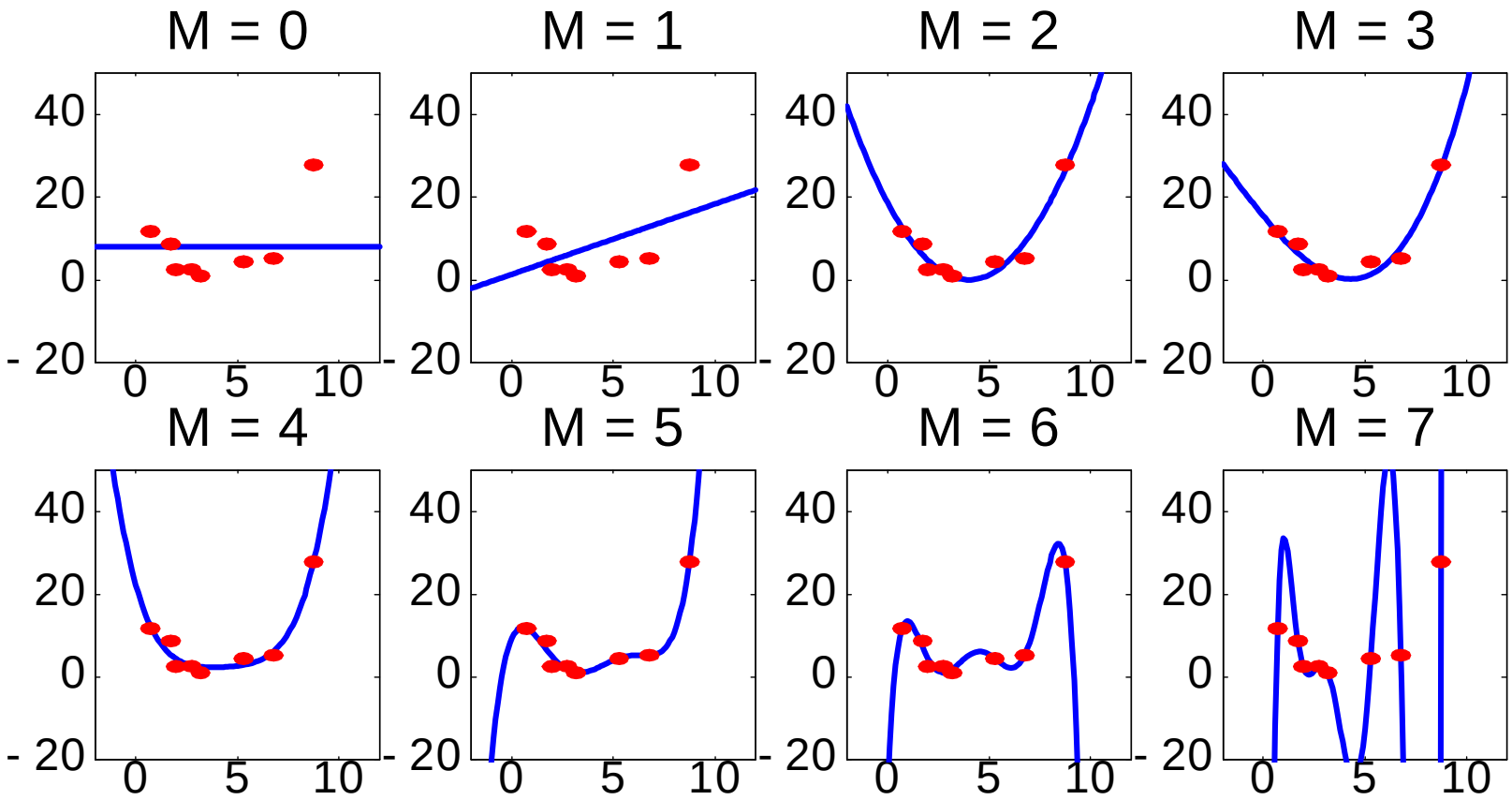

Figure: As M increases, the model overfits in a MLE framework.

- We select $V_M = \delta^2 I_{M+1}$ where $\delta^2 = 10$, $a = b = 1$.
- In this case, we have $\Theta_M = \mathbb{R}^{M+1} \times \mathbb{R}^+$.
- For $M \in \{0, ..., M_{\max}\}$, we can define $\pi(\mathcal{M}_M) = \frac{1}{M_{\max}+1}$.

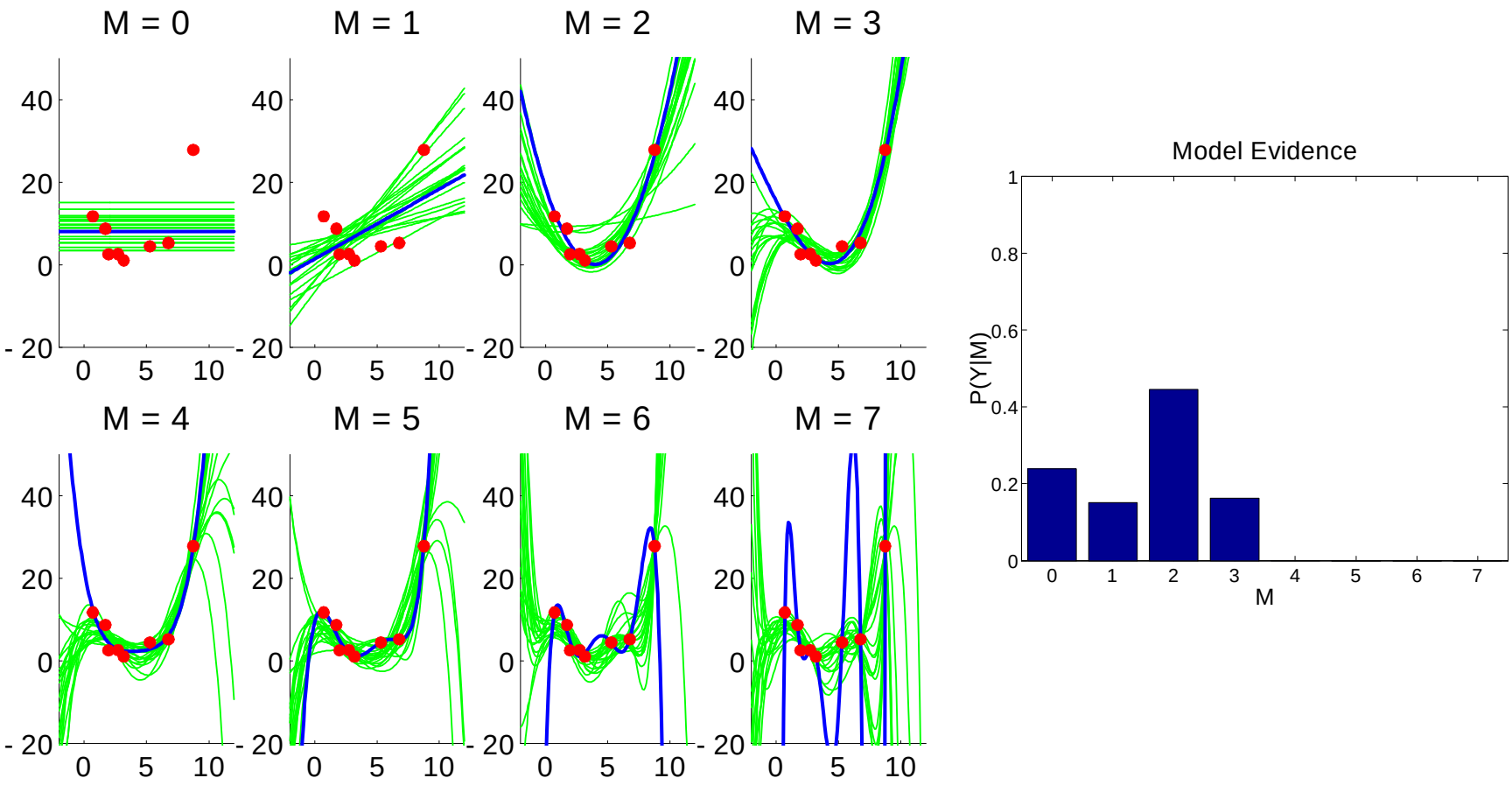

Figure: Marginal likelihood $\pi\left(D\middle|\mathcal{M}_i\right)$ and $f\left(\mathbf{x}\right)$ for random draws from $\pi\left(\beta\middle|\mathcal{M}_i\right)$.

- We have assumed here that δ^2 was fixed and set to $\delta^2 = 1$.
- As $\delta^2 \to \infty$, the prior on β is getting vague but then

$$\lim_{\delta^2 \to \infty} \pi\left(\mathcal{M}_0 | D\right) = 1$$

as for $M \geq 1$

$$\frac{\pi\left(D | \mathcal{M}_0\right)}{\pi\left(D | \mathcal{M}_M\right)} = \frac{\left|\delta^2 I_{M+1}\right|^{1/2} \left|\delta^{-2} I_{M+1} + \mathbf{B}_M^{\mathsf{T}} \mathbf{B}_M\right|^{-1/2} \Gamma\left(a_0^*\right) \left(b_M^*\right)^{a_M^*}}{\delta \left|\delta^{-2} + \mathbf{B}_0^{\mathsf{T}} \mathbf{B}_0\right|^{1/2} \Gamma\left(a_M^*\right) \left(b_0^*\right)^{a_0^*}} \underset{\delta^2 \to \infty}{\longrightarrow} \infty$$

- **Do not use vague priors for model selection**!!!
- For a robust model, select a random δ^2 and estimate it from the data. However, numerical methods are then necessary.

Bayesian Information Criterion

- Bayesian model selection is a flexible and principled approach but can be difficult to implement if the marginal likelihoods do not admit a closed form expression.
- When a large number of data is available, it is possible to approximate these marginal likelihood terms.
- This simple but powerful approach was proposed by Schwarz (1978) and is very popular in the literature.

- Let us consider a Bayesian model where $X_i \overset{\text{i.i.d.}}{\sim} f\left(x \middle| \theta\right)$ and we consider the prior $\pi\left(\theta\right)$ on $\Theta \subseteq \mathbb{R}^d$.
- The marginal likelihood of n observations is given by

$$\pi\left(x_{1:n}\right) = \int f\left(x_{1:n} \middle| \theta\right) \pi\left(\theta\right) d\theta.$$

and we use the following notation

$$g\left(\theta\right) = \log\left(f\left(x_{1:n} \middle| \theta\right) \pi\left(\theta\right)\right).$$

- We perform a Taylor expansion about $\widehat{\theta}$ the posterior mode; i.e. the value of θ maximizing $g\left(\theta\right)$.

$$\begin{aligned} g\left(\theta\right) &= g\left(\widehat{\theta}\right) + \left(\theta - \widehat{\theta}\right)^{\mathsf{T}} g'\left(\widehat{\theta}\right) + \frac{1}{2}\left(\theta - \widehat{\theta}\right)^{\mathsf{T}} g''\left(\widehat{\theta}\right)\left(\theta - \widehat{\theta}\right) \\ &\quad + o\left(\left\|\theta - \widehat{\theta}\right\|^2\right) \end{aligned}$$

where

$$g'\left(\theta\right) = \left(\frac{\partial g\left(\theta\right)}{\partial \theta_1}, ..., \frac{\partial g\left(\theta\right)}{\partial \theta_d}\right)^{\mathsf{T}}, \quad \left[g''\left(\theta\right)\right]_{i,j} = \frac{\partial^2 g\left(\theta\right)}{\partial \theta_i \partial \theta_j}.$$

- Because $g'\left(\widehat{\theta}\right) = 0$ then

$$g(\theta) \approx g\left(\widehat{\theta}\right) + \frac{1}{2}\left(\theta - \widehat{\theta}\right)^{\mathsf{T}} g''\left(\widehat{\theta}\right)\left(\theta - \widehat{\theta}\right).$$

- This approximation will not be good unless θ is close to $\widehat{\theta}$. However, when n is large, $g(\theta)$ is concentrated around its maximum and declines fast as one moves away from $\widehat{\theta}$, so that only values from θ close to $\widehat{\theta}$ will contribute much to the marginal likelihood.
- It follows that

$$\begin{aligned}
\pi(x_{1:n}) &= \int \exp(g(\theta))\, d\theta \\
&\approx \exp\left(g\left(\widehat{\theta}\right)\right) \int \exp\left(\frac{1}{2}\left(\theta - \widehat{\theta}\right)^{\mathsf{T}} g''\left(\widehat{\theta}\right)\left(\theta - \widehat{\theta}\right)\right) d\theta \\
&= \exp\left(g\left(\widehat{\theta}\right)\right) (2\pi)^{d/2} |A|^{-1/2}
\end{aligned}$$

where $A = -g''\left(\widehat{\theta}\right)$. This is the so-called Laplace approximation.

- Under regularity assumptions, the error is of order $O\left(n^{-1}\right)$ and we have

$$\begin{aligned}\log \pi\left(x_{1:n}\right) &= \log f\left(x_{1:n}|\widehat{\theta}\right)+\log \pi\left(\widehat{\theta}\right) \\ &\quad +\frac{d}{2}\log\left(2\pi\right)-\frac{1}{2}\log\left(|A|\right)+O\left(n^{-1}\right).\end{aligned}$$

- Remember that $O\left(n^{-\alpha}\right)$ represents any quantity such that $n^{\alpha}O\left(n^{-\alpha}\right)\leq Cst$ as $n\longrightarrow\infty$.

- For large samples, we have $\widehat{\theta} \approx \theta_{MLE}$ and

$$[A]_{i,j} = -\frac{\partial^2 g(\theta)}{\partial\theta_i \partial\theta_j} = -\frac{\partial^2 \log \pi(\theta)}{\partial\theta_i \partial\theta_j} - \sum_{i=1}^{n} \frac{\partial^2 \log f(X_i|\theta)}{\partial\theta_i \partial\theta_j}$$

so

$$A \approx nI$$

where I is

$$[I]_{i,j} = -\mathbb{E}_X \left[\frac{\partial^2 \log f(X|\theta)}{\partial\theta_i \partial\theta_j} \right] \Bigg|_{\theta_{MLE}}.$$

- I is the Fisher information matrix if the model is correctly specified.
- Thus it follows that

$$|A| \approx n^d |I|$$

- These two approximations introduce an $O\left(n^{-1/2}\right)$ error and

$$\begin{aligned} \log \pi(x_{1:n}) &= \log f(x_{1:n}|\theta_{MLE}) + \log \pi(\theta_{MLE}) \\ &\quad + \frac{d}{2}\log(2\pi) - \frac{d}{2}\log n - \frac{1}{2}\log(|I|) + O\left(n^{-1/2}\right). \end{aligned}$$

- We have $\log f\left(x_{1:n}|\,\theta_{MLE}\right)$ of order $O\left(n\right)$, $\frac{d}{2}\log n$ of order $O\left(\log n\right)$ whereas the other terms are of order $O\left(1\right)$ or less.
- Hence we can conclude that

$$\log \pi\left(x_{1:n}\right) = \log f\left(x_{1:n}|\,\theta_{MLE}\right) - \frac{d}{2}\log n + O\left(1\right).$$

- This equation means that in general the approximation error does not vanish even with an infinite number of data. This is not crucial as the other terms will go to infinity as $n \longrightarrow \infty$ and will dominate the $O\left(1\right)$ term; i.e. the error will tend toward zero as a proportion of $\log \pi\left(x_{1:n}\right)$, ensuring that the error will not affect the conclusion reached.
- Moreover empirical experience has found that in many scenarios of interest the error is of a much smaller order of magnitude.

- Consider the case where

$$\pi(\theta) = \mathcal{N}\left(\theta; \theta_{MLE}, I^{-1}\right)$$

so that, roughly speaking, the prior distribution contains the same amout of information as would, on average, a single observation.

- In this case we have

$$\log \pi(\theta_{MLE}) = -\frac{d}{2}\log(2\pi) + \frac{1}{2}\log|I|$$

so

$$\log \pi(x_{1:n}) = \log f(x_{1:n}|\theta_{MLE}) - \frac{d}{2}\log n + O\left(n^{-1/2}\right).$$

- In this case, the approximation error is of order $O\left(n^{-1/2}\right)$.

- The approximation of $\log \pi\left(x_{1:n}\right)$ can be used to approximate the Bayes factor

$$B_{21} = \frac{\pi\left(x_{1:n} \middle| M_2\right)}{\pi\left(x_{1:n} \middle| M_1\right)} = \frac{\int f_2\left(x_{1:n} \middle| \theta_2\right) \pi_2\left(\theta_2\right) d\theta_2}{\int f_1\left(x_{1:n} \middle| \theta_1\right) \pi_1\left(\theta_1\right) d\theta_1}$$

  where $\theta_1 \in \mathbb{R}^{d_1}$ and $\theta_2 \in \mathbb{R}^{d_2}$.
- We have

$$\begin{aligned} 2 \log B_{21} &\approx 2\left(\log f\left(x_{1:n} \middle| \theta_{2,MLE}\right) - \log f\left(x_{1:n} \middle| \theta_{1,MLE}\right)\right) \\ &\quad - \left(d_2 - d_1\right) \log n. \end{aligned}$$

- We can see that BIC penalized the number of parameters, this prevents overfitting.
- We can also compute the approximate posterior probabilities

$$\begin{aligned} \pi\left(M_k \middle| x_{1:n}\right) &\propto \pi\left(M_k\right) \pi\left(x_{1:n} \middle| M_k\right) \\ &\propto \pi\left(M_k\right) \exp\left(\log f\left(x_{1:n} \middle| \theta_{k,MLE}\right) - \frac{d_k}{2} \log n\right) \\ &\propto \pi\left(M_k\right) n^{-d_k/2} f\left(x_{1:n} \middle| \theta_{k,MLE}\right) \end{aligned}$$

- **Example**. Consider the following autoregressive (AR) time series model

$$M_p: \; X_n = \sum_{i=1}^{p} a_i X_{n-i} + \sigma \varepsilon_n$$

where $V_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. The AR $(a_1, ..., a_p)$ coefficients and σ^2 are unknown.

- Clearly the larger p the better the fit to the data. We want to determine the model order $p \in \{0, ..., p_{\max}\}$.

- For the data $x_{p_{\max}:T}$, we can rewrite the model in the matrix-vector form

$$M_p : \ \mathbf{X} = \mathbf{B}_p \mathbf{a} + \boldsymbol{\epsilon}$$

where $\mathbf{X} = (x_{p_{\max}}, ..., x_T)^{\mathsf{T}}$, $\mathbf{a} = (a_1, ..., a_p)^{\mathsf{T}}$, $\boldsymbol{\epsilon} = (\epsilon_{p_{\max}}, ..., \epsilon_T)^{\mathsf{T}}$ and the design matrix

$$\mathbf{B}_p = \begin{pmatrix} x_{p_{\max}-1} & \cdots & x_{p_{\max}-p} \\ \vdots & & \vdots \\ x_{n-1} & \cdots & x_{n-p} \\ \vdots & & \vdots \\ x_{T-1} & \cdots & x_{T-p} \end{pmatrix}$$

- For each candidate model M_p, we can easily come up with the MLE $\theta_{MLE} = \left(\mathbf{a}_{\text{MLE}}, \sigma^2_{\text{MLE}}\right)$ and hence obtain

$$\pi\left(x_{p_{\max}:T} \middle| M_p\right) \approx \log f\left(x_{p_{\max}:T} \middle| \theta_{MLE}\right) - \frac{p+1}{2} \log\left(T - p_{\max} + 1\right).$$

- Alternatively, we can do a conjugate analysis by setting a normal-inverse Gamma prior on $\left(\mathbf{a}, \sigma^2\right)$. Hence, we can compute analytically $\pi\left(\mathbf{a}, \sigma^2 \middle| x_{p_{\max}:T}\right)$ and $\pi\left(x_{p_{\max}:T} \middle| M_p\right)$ for such a prior.
- We select $p_{\max} = 20$ and compared BIC to MMAP when $\delta^2 = 10$ on 100 realizations of a 3rd order AR process (roots at 0.9 and 0.5 $\pm j$ 0.85π and $\sigma^2 = 10$).

  Number of occurrences where the true model order $k = 3$ is selected

T	35	50	75	100	200	300
AIC	20	31	49	59	74	76
BIC	19	30	46	57	76	94
MMAP	23	33	49	64	78	95

Akaike Information Criterion

- An alternative non-Bayesian approach to BIC is the celebrated AIC.
- In this framework, we select the model minimizing

$$AIC = -2 \log f\left(x_{1:n} \middle| \theta_{MLE}\right) + 2d$$

in contrast with BIC where we minimize

$$-2 \log f\left(x_{1:n} \middle| \theta_{MLE}\right) + d \log n.$$

- Contrary to BIC, AIC is typically not asymptotically consistent... although this does not mean very much!

- We outline here the derivation of AIC which is based on the KL

$$I(g;f) = \int g(x) \log \frac{g(x)}{f(x)} dx$$

which is such that $I(g;f) \geq 0$ and $I(g;f) = 0$ iff $g(x) = f(x)$.

- Assuming we have a family of candidate models $f_m(x) = f(x|\theta_m)$, $m = 1, ..., M$, to model the true distribution g of the observations then we want to select the model m minimizing $I(g;f_m)$ or equivalently maximizing

$$\int g(x) \log f_m(x) dx.$$

- Now assume we have n observations $X_i \overset{\text{i.i.d.}}{\sim} g(x)$. We have

$$\int g(x) \log f(x|\theta_m) dx \approx \frac{1}{n} \sum_{i=1}^{n} \log f(x_i|\theta_m).$$

- However, in almost all the applications, θ_m is unknown and needs to be estimated from the data.
- It is natural to estimate θ_m using ML and then plug $\widehat{\theta}_m$ to obtain the estimate of the KL

$$\frac{1}{n}\sum_{i=1}^{n}\log f\left(x_i|\widehat{\theta}_m\right).$$

- However, we have

$$\mathbb{E}_X\left[\frac{1}{n}\sum_{i=1}^{n}\log f\left(X_i|\widehat{\theta}_m\right)\right] \neq \mathbb{E}_X\left(\mathbb{E}_Y \log f\left(Y|\widehat{\theta}_m\right)\right)$$

  as the same dataset was used twice for the estimation of the parameters and for the estimation of the expected log-likelihood.
- AIC is an approximate correction for this bias.

- To simplify notation, we suppress further on the model index m.
- We introduce the bias

$$
\begin{aligned}
B &= \mathbb{E}_X \left(\mathbb{E}_Y \log f\left(Y | \widehat{\theta} \right) - \frac{1}{n} \sum_{i=1}^{n} \log f\left(X_i | \widehat{\theta} \right) \right) \\
&= \underbrace{\mathbb{E}_X \left(\mathbb{E}_Y \log f\left(Y | \widehat{\theta} \right) - \mathbb{E}_Y \log f\left(Y | \theta_0 \right) \right)}_{B_1} \\
&\quad \underbrace{+ \mathbb{E}_X \left(\mathbb{E}_Y \log f\left(Y | \theta_0 \right) - \frac{1}{n} \sum_{i=1}^{n} \log f\left(X_i | \theta_0 \right) \right)}_{B_2} \\
&\quad \underbrace{+ \mathbb{E}_X \left(\frac{1}{n} \sum_{i=1}^{n} \log f\left(X_i | \theta_0 \right) - \frac{1}{n} \sum_{i=1}^{n} \log f\left(X_i | \widehat{\theta} \right) \right)}_{B_3}
\end{aligned}
$$

where θ_0 maximizes $\mathbb{E}_Y \log f\left(Y | \theta \right)$.

- We have

$$\begin{aligned}
&\mathbb{E}_Y \log f\left(Y|\widehat{\theta}\right) - \mathbb{E}_Y \log f\left(Y|\theta_0\right) \\
\approx\ & \left(\frac{\partial}{\partial\theta}\mathbb{E}_Y \log f\left(Y|\theta_0\right)\right)^{\mathsf{T}}\left(\widehat{\theta}-\theta_0\right) \\
&+\frac{1}{2}\left(\widehat{\theta}-\theta_0\right)^{\mathsf{T}}\frac{\partial^2}{\partial\theta\partial\theta^{\mathsf{T}}}\mathbb{E}_Y \log f\left(Y|\theta_0\right)\left(\widehat{\theta}-\theta_0\right)
\end{aligned}$$

  where $J = -\frac{\partial^2}{\partial\theta\partial\theta^{\mathsf{T}}}\mathbb{E}_Y\left(\log f\left(Y|\theta_0\right)\right)$.
- We have seen in class before that

$$\sqrt{N}\left(\widehat{\theta}-\theta_0\right) \Rightarrow \mathcal{N}\left(0, J^{-1}IJ\right),$$
$$I = \mathbb{E}_Y\left[\frac{\partial \log f\left(Y|\theta_0\right)}{\partial\theta}\left(\frac{\partial \log f\left(Y|\theta_0\right)}{\partial\theta}\right)^{\mathsf{T}}\right]$$

  so

$$\mathbb{E}_X\left[\left(\widehat{\theta}-\theta_0\right)^{T} J\left(\widehat{\theta}-\theta_0\right)\right] = \frac{1}{n}\text{trace}\left(IJ^{-1}\right).$$

- Finally we have

$$B_1 \approx -\frac{1}{2n}\text{trace}\left(IJ^{-1}\right)$$

- We have

$$B_2 = \mathbb{E}_X \left(\mathbb{E}_Y \log f\left(Y | \theta_0 \right) - \frac{1}{n} \sum_{i=1}^{n} \log f\left(X_i | \theta_0 \right) \right) = 0$$

- Now

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^{n} \log f\left(X_i | \theta_0 \right) - \frac{1}{n} \sum_{i=1}^{n} \log f\left(X_i | \widehat{\theta} \right) \\ \approx & \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log f\left(X_i | \widehat{\theta} \right)}{\partial \theta} \right)^{\mathsf{T}} \left(\theta_0 - \widehat{\theta} \right) \\ & + \frac{1}{2} \left(\theta_0 - \widehat{\theta} \right)^{\mathsf{T}} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f\left(X_i | \widehat{\theta} \right)}{\partial \theta \partial \theta^{\mathsf{T}}} \right) \left(\theta_0 - \widehat{\theta} \right). \end{aligned}$$

- The law of large numbers yields

$$\left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f\left(X_i | \widehat{\theta} \right)}{\partial \theta \partial \theta^{\mathsf{T}}} \right) \rightarrow -J$$

- Now taking the expectation with respect to $\{X_i\}$

$$\begin{aligned} B_3 &= \mathbb{E}_X \left(\frac{1}{n} \sum_{i=1}^{n} \log f\left(X_i | \theta_0\right) - \mathbb{E}_Y \log f\left(Y | \widehat{\theta}\right) \right) \\ &\approx -\frac{1}{2} \mathbb{E}_X \left(\left(\theta_0 - \widehat{\theta}\right)^{\mathsf{T}} J \left(\theta_0 - \widehat{\theta}\right) \right) \\ &\approx -\frac{1}{2n} \text{trace}\left(IJ^{-1}\right). \end{aligned}$$

- It follows that

$$\begin{aligned} B &= B_1 + B_2 + B_3 \\ &\approx -\frac{1}{n} \text{trace}\left(IJ^{-1}\right). \end{aligned}$$

- In the special situation where the true distribution belongs to the family of models (!!), then we have $I = J$ and

$$B = -\frac{1}{n} \text{trace}\left(IJ^{-1}\right) = -\frac{1}{n} \text{trace}\left(\text{Id}\right) = -\frac{d}{n}$$

where Id is the identity matrix of size $d \times d$ where $\theta \in \mathbb{R}^d$.

- To conclude, we have

$$\begin{aligned}\mathbb{E}_X\left(\mathbb{E}_Y \log f\left(Y|\widehat{\theta}_m\right)\right) &= B + \mathbb{E}_X\left[\frac{1}{n}\sum_{i=1}^{n}\log f\left(X_i|\widehat{\theta}_m\right)\right] \\ &\approx \frac{1}{n}\left(\log f\left(x_{1:n}|\widehat{\theta}_m\right) - d\right)\end{aligned}$$

- Maximizing this quantity over m is equivalent to minimizing

$$AIC = -2\log f\left(x_{1:n}|\widehat{\theta}_m\right) + 2d.$$

- Since its introduction in the mid-70s, many variations have been proposed.