5.Bayesian statistics

Wenxiang Luo

- Basic concepts
- Summarizing posterior distributions
- **Bayesian model selection**

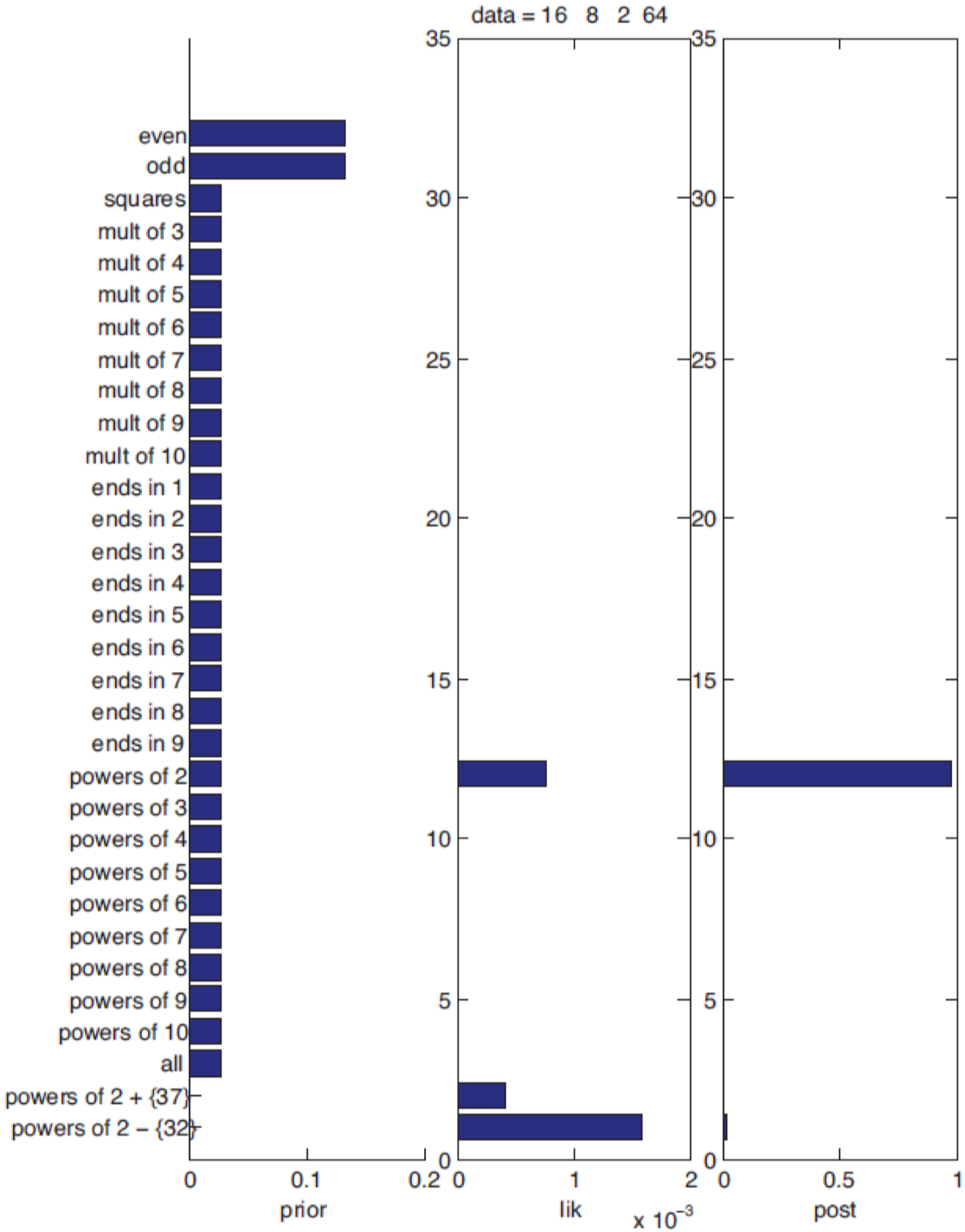
Basic concepts

- Bayes rule

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{p(X = x)p(Y = y|X = x)}{\sum_{x'} p(X = x')p(Y = y|X = x')}$$

Basic concepts

- Another example —— number game
  - To classify x given $\mathcal{D} = \{x_1, \ldots, x_N\}$ drawn from C(a kind of arithmetical concept)
  - $\mathcal{D} = \{16, 8, 2, 64\}$, “even numbers”or “powers of two”or even“powers of two except for 32”
- Likelihood
  - Suppose $p(D|h) = \left[\frac{1}{\text{size}(h)}\right]^N = \left[\frac{1}{|h|}\right]^N$ then $p(\mathcal{D} \mid h_{\text{even}}) = (1/50)^4 = 1.6 \times 10^{-7}$, $p(\mathcal{D} \mid h_{\text{two}}) = (1/6)^4 = 7.7 \times 10^{-4}$
  - Quantifies our intuition
- Prior
  - Why not“powers of two except for 32”?——seems“conceptually **unnatural**”
  - Subjective
- Posterior
  - Likelihood times the prior, and normalized



$$\hat{h}^{MAP} = \underset{h}{\operatorname{argmax}}\, p(\mathcal{D}|h)p(h) = \underset{h}{\operatorname{argmax}}\, [\log p(\mathcal{D}|h) + \log p(h)]$$

$$\hat{h}^{mle} \triangleq \underset{h}{\operatorname{argmax}}\, p(\mathcal{D}|h) = \underset{h}{\operatorname{argmax}} \log p(\mathcal{D}|h)$$

basic concepts

- Posterior predictive distribution
  - $p(\tilde{x} \in C|\mathcal{D}) = \sum_h p(y=1|\tilde{x}, h)p(h|\mathcal{D})$
  - Called Bayes model averaging
  - Plug-in approximation

    $p(\tilde{x} \in C|\mathcal{D}) = \sum_h p(\tilde{x}|h)\delta_{\hat{h}}(h) = p(\tilde{x}|\hat{h})$

- Conjugate prior
  - When the prior and the posterior have the **same form**, we say that the prior is a **conjugate prior** for the corresponding **likelihood**.

- Conjugate prior
  - Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$$

$$log\, p(x) \propto ax^2 + bx \qquad a = -\frac{1}{2\sigma^2} \qquad b = \frac{\mu}{\sigma^2}$$

$$p(x|\theta)p(\theta) \propto exp\{-\frac{(x-\theta)^2}{2\sigma_1^2}\} exp\{-\frac{(\theta-\mu_2)^2}{2\sigma_2^2}\} \propto exp\{a\theta^2 + b\theta\}$$

$$a = -(\frac{n}{2\sigma_1^2} + \frac{1}{2\sigma_2^2}) \quad b = \frac{\Sigma x_i}{\sigma_1^2} + \frac{\mu}{\sigma_2^2}$$

$$\mu_0 = (\frac{n}{\sigma_1^2} + \frac{1}{\sigma_2^2})^{-1}(\frac{\Sigma x_i}{\sigma_1^2} + \frac{\mu}{\sigma_2^2}) \qquad \sigma_0^2 = (\frac{n}{\sigma_1^2} + \frac{1}{\sigma_2^2})^{-1}$$

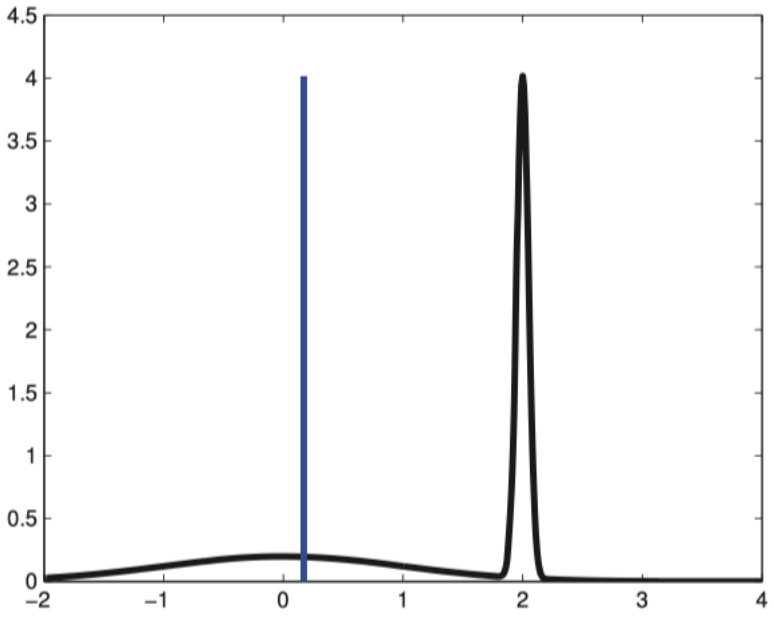
Summarizing posterior distributions

MAP(Maximum A Posteriori)

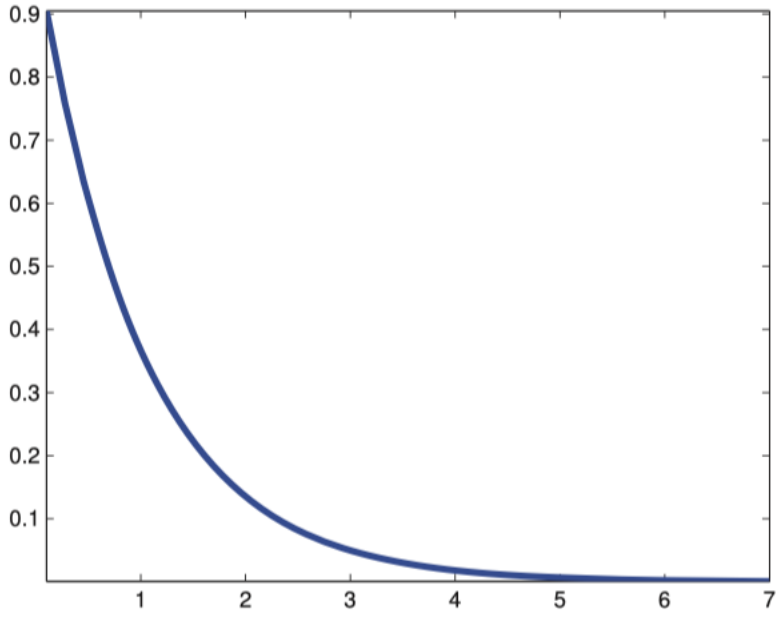
- Basic introduction
  - A point estimate —— the posterior mode
  - Reduces to an optimization problem
  - Can be interpreted in non-Bayesian terms

- Drawbacks
  - No measure of uncertainty
  - The mode is an untypical point
  - MAP estimation is not invariant to reparameterization

The mode is an untypical point



(a)



(b)

MAP estimation is **not invariant** to reparameterization

- $p_y(y) = p_x(x)\left|\frac{dx}{dy}\right|$
  - suppose $p_\mu(\mu) = 1\ \mathbb{I}(0 \leq \mu \leq 1)$

First let $\theta = \sqrt{\mu}$ so $\mu = \theta^2$. The new prior is

$$p_\theta(\theta) = p_\mu(\mu)\left|\frac{d\mu}{d\theta}\right| = 2\theta$$

for $\theta \in [0, 1]$ so the new mode is

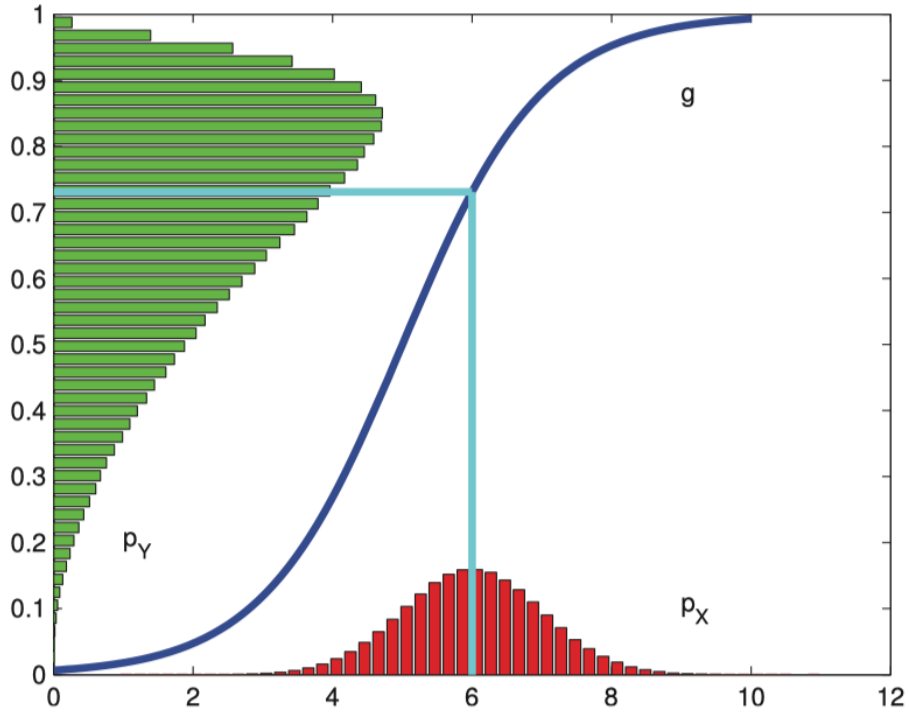
$$\hat{\theta}_{MAP} = \arg\max_{\theta \in [0,1]} 2\theta = 1$$

Now let $\phi = 1 - \sqrt{1-\mu}$. The new prior is

$$p_\phi(\phi) = p_\mu(\mu)\left|\frac{d\mu}{d\phi}\right| = 2(1-\phi)$$

for $\phi \in [0, 1]$, so the new mode is

$$\hat{\phi}_{MAP} = \arg\max_{\phi \in [0,1]} 2 - 2\phi = 0$$



The MLE is a **function**, not a probability density.

Bayesian inference **integrating over the parameter space**.

Summarizing posterior distributions

Credible intervals

http://jakevdp.github.io/blog/2014/06/12/frequentism-and-bayesianism-3-confidence-credibility/

https://bayes.wustl.edu/etj/articles/confidence.pdf

- $C_\alpha(\mathcal{D}) = (\ell, u) : P(\ell \leq \theta \leq u|\mathcal{D}) = 1 - \alpha$
- central interval
  - there is α/2 mass in each
- highest density interval $1 - \alpha = \int_{\theta:p(\theta|\mathcal{D})>p^*} p(\theta|\mathcal{D})d\theta$
  - For a unimodal distribution, the HDI will be the narrowest interval

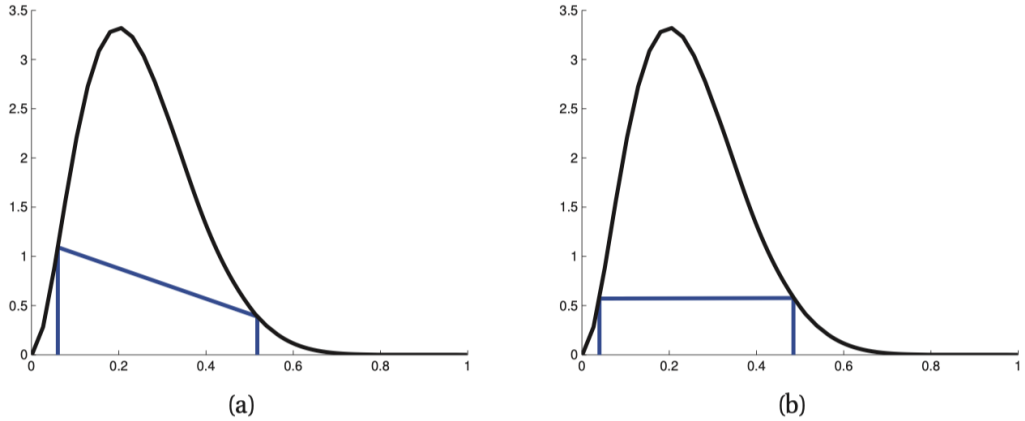


Figure 5.3 (a) Central interval and (b) HPD region for a Beta(3,9) posterior. The CI is (0.06, 0.52) and the HPD is (0.04, 0.48). Based on Figure 3.6 of (Hoff 2009). Figure generated by `betaHPD`.

Bayesian model selection

- Compute the **posterior over models**

$$p(m \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid m)p(m)}{\sum_{m \in \mathcal{M}} p(m, \mathcal{D})} \propto p(\mathcal{D} \mid m)p(m)$$

- **Evidence**
  - Use a uniform prior over models $p(m) \propto 1$, this amounts to picking the model which maximizes $p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$
  - Called the **marginal likelihood**, the **integrated likelihood**, or the **evidence** for model m

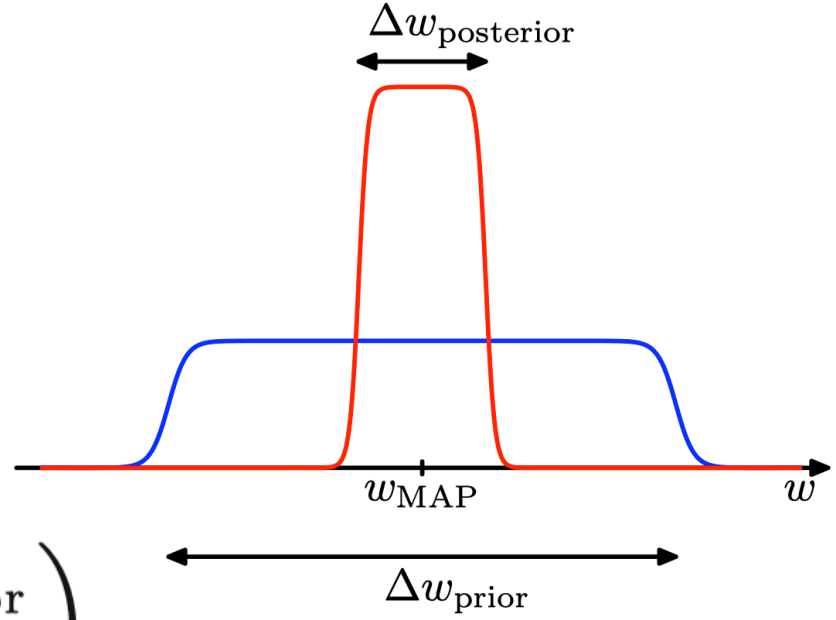
Bayesian model selection

- Evidence
  - assume that the posterior distribution is sharply peaked around the most probable value w_{MAP} , with width $\Delta w_{\mathrm{posterior}}$
  - further assume that the prior is flat with widt $\Delta w_{\mathrm{prior}}$ so that $p(w) = 1/\Delta w_{\mathrm{prior}}$, then we have

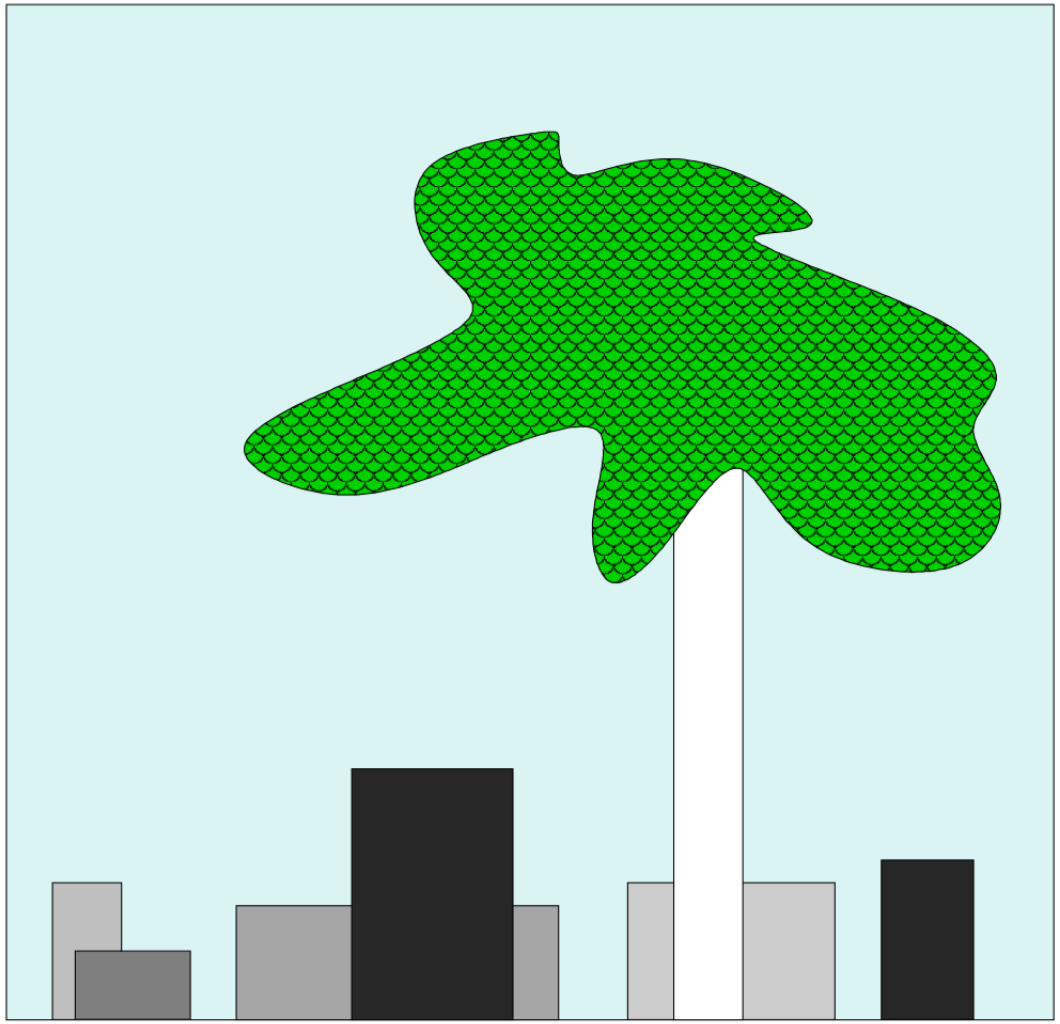
$$p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)\,\mathrm{d}w \simeq p(\mathcal{D}|w_{\mathrm{MAP}})\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}$$

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\mathrm{MAP}}) + \ln\left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}\right)$$

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\mathbf{w}_{\mathrm{MAP}}) + M\ln\left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}\right)$$



Bayesian Occam's razor



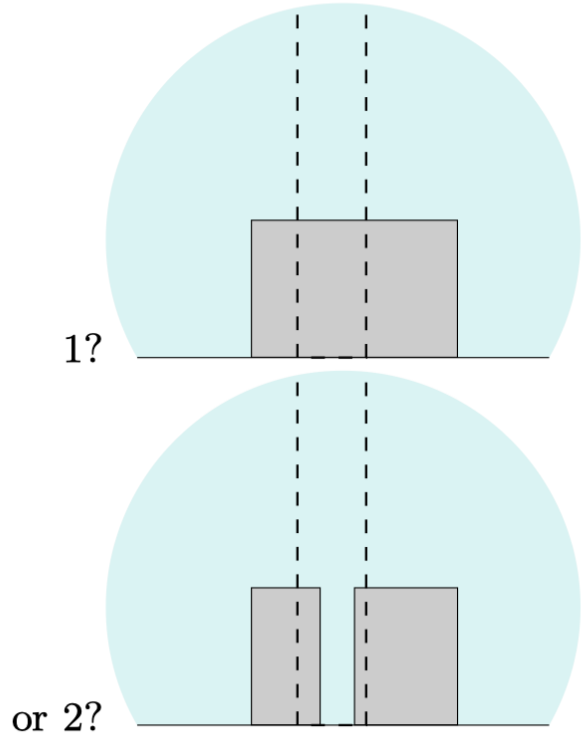
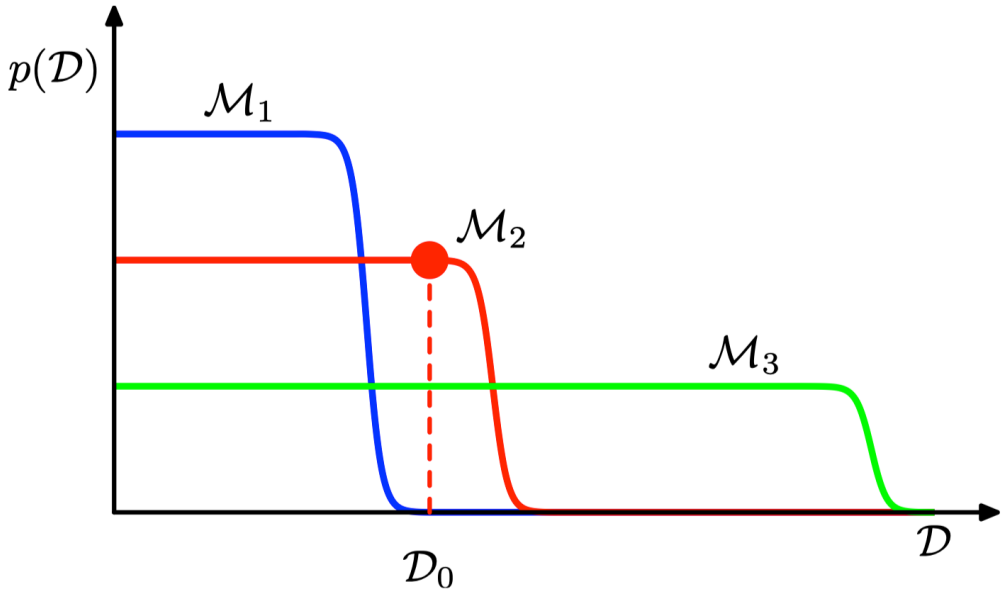


Figure 28.2. How many boxes are behind the tree?

Bayesian model selection

- Bayesian Occam's razor
  - $p(\mathcal{D}) = p(y_1)p(y_2|y_1)p(y_3|y_{1:2}) \ldots p(y_N|y_{1:N-1})$
  - conservation of probability mass $\sum_{\mathcal{D}'} p(\mathcal{D}'|m) = 1$
    - Complex models-spread mass thinly



Bayesian Occam's razor

- A sequence $-1,\ 3,\ 7,\ 11$
  - $\mathcal{H}_a$: add 4 or $\mathcal{H}_c$:$-x^3/11 + 9/11x^2 + 23/11$

$\mathcal{H}_a$ – the sequence is an *arithmetic* progression, 'add n', where n is an integer.

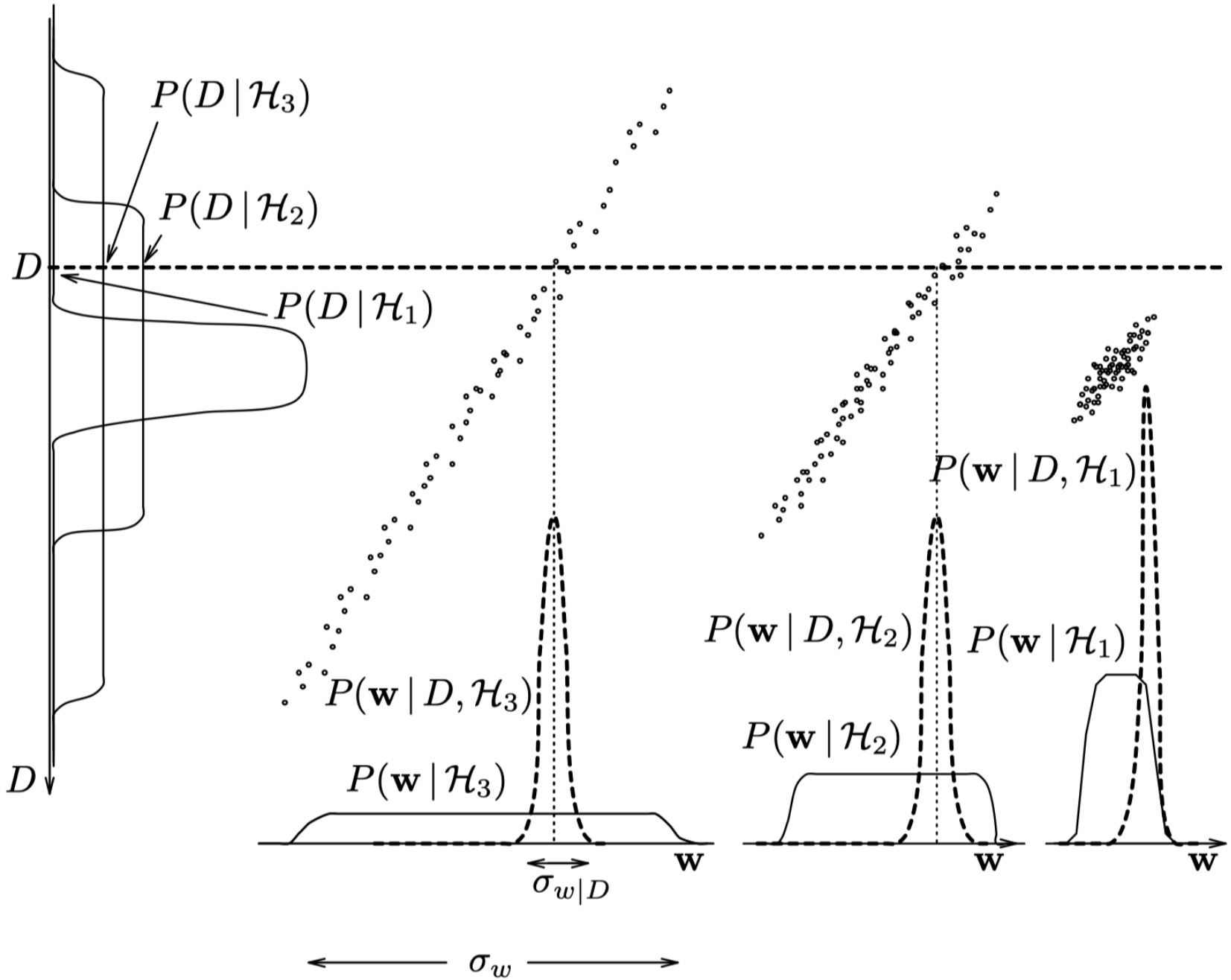
$\mathcal{H}_c$ – the sequence is generated by a *cubic* function of the form $x \to cx^3 + dx^2 + e$, where c, d and e are fractions.

$$\mathcal{H}_a : \{n = 4,\ \text{first number } = -1\}$$

$$P(D\,|\,\mathcal{H}_a) = \frac{1}{101}\frac{1}{101} = 0.00010$$

$$P(D|\mathcal{H}_c) = \left(\frac{1}{101}\right)\left(\frac{4}{101}\frac{1}{50}\right)\left(\frac{4}{101}\frac{1}{50}\right)\left(\frac{2}{101}\frac{1}{50}\right)$$

$$= 0.0000000000025 = 2.5 \times 10^{-12}.$$


$P(D\,|\,\mathcal{H}_3)$
$P(D\,|\,\mathcal{H}_2)$
D
$P(D\,|\,\mathcal{H}_1)$
D
$P(\mathbf{w}\,|\,D,\mathcal{H}_3)$
$P(\mathbf{w}\,|\,\mathcal{H}_3)$
$\mathbf{w}$
$\sigma_{w|D}$
σ_w
$P(\mathbf{w}\,|\,D,\mathcal{H}_1)$
$P(\mathbf{w}\,|\,D,\mathcal{H}_2)$
$P(\mathbf{w}\,|\,\mathcal{H}_2)$
$P(\mathbf{w}\,|\,\mathcal{H}_1)$
$\mathbf{w}$
$\mathbf{w}$

Bayesian model selection

- Computing the marginal likelihood (evidence)
  - BIC approximation
    - Computing the integral $p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$ can be quite difficult
    - One popular approximation

$$\mathrm{BIC} \triangleq \log p(\mathcal{D} \mid \hat{\boldsymbol{\theta}}, m) - \frac{\mathrm{dof}(\hat{\boldsymbol{\theta}})}{2} \log N \approx \log p(\mathcal{D} \mid m)$$

Derivation of the BIC

- Laplace approximation

$$\text{Suppose } \boldsymbol{\theta} \in \mathbb{R}^D \quad p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z} e^{-E(\boldsymbol{\theta})} \quad E(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}, \mathcal{D} \mid m), \text{ with } Z = p(\mathcal{D} \mid m)$$

$$E(\boldsymbol{\theta}) \approx E(\boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{g} + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{H} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)$$

$$\mathbf{g} \triangleq \nabla E(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}^*}, \; \mathbf{H} \triangleq \frac{\partial^2 E(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \big|_{\boldsymbol{\theta}^*}$$

$$\hat{p}(\boldsymbol{\theta}|\mathcal{D}) \quad \approx \quad \frac{1}{Z} e^{-E(\boldsymbol{\theta}^*)} \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{H} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right]$$

$$\exp\left[-\tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{H} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right] = \frac{(2\pi)^{D/2}}{|\mathbf{H}|^{\frac{1}{2}}} \mathcal{N}\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*, \mathbf{H}^{-1}\right)$$

$$Z = p(\mathcal{D} \mid m) \approx \int \hat{p}(\boldsymbol{\theta}, \mathcal{D} \mid m) d\boldsymbol{\theta} = e^{-E(\boldsymbol{\theta}^*)} (2\pi)^{D/2} |\mathbf{H}|^{-\frac{1}{2}}$$

- Dropping irrelevant constants

$$\log p(\mathcal{D} \mid m) \approx \log p(\mathcal{D} \mid \boldsymbol{\theta}^*) + \log p(\boldsymbol{\theta}^*) - \frac{1}{2}\log |\mathbf{H}|$$

- Assume uniform prior $p(\boldsymbol{\theta}) \propto 1$, and replace $\boldsymbol{\theta}^*$ with MLE, $\hat{\boldsymbol{\theta}}$

$$\mathbf{H} = \sum_{i=1}^{N} \mathbf{H}_i \quad \mathbf{H}_i = \nabla\nabla \log p(\mathcal{D}_i \mid \boldsymbol{\theta})$$

- Approximate each $\mathbf{H}_i$ by a fixed matrix $\hat{\mathbf{H}}$

$$\log |\mathbf{H}| = \log |N\hat{\mathbf{H}}| = \log(N^d|\hat{\mathbf{H}}|) = D \log N + \log |\hat{\mathbf{H}}|$$

- Assumed $\mathbf{H}$ is full rank

$$\log p(\mathcal{D} \mid m) \approx \log p(\mathcal{D} \mid \hat{\boldsymbol{\theta}}) - \frac{D}{2}\log N$$

Bayesian model selection

- Computing the marginal likelihood (evidence)
  - Effect of the prior
    - Sometimes it is not clear how to set the prior
    - If the prior is unknown, the correct Bayesian procedure is to **put a prior on the prior**.

$$p(\mathcal{D}|m) = \int \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w}|\alpha, m)p(\alpha|m)d\mathbf{w}d\alpha$$

    - Make the hyper-prior **uninformative**.
    - EB

$$p(\mathcal{D}|m) \approx \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w}|\hat{\alpha}, m)d\mathbf{w}$$

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}}\, p(\mathcal{D}|\alpha, m) = \underset{\alpha}{\operatorname{argmax}} \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w}|\alpha, m)d\mathbf{w}$$

Bayesian model selection

- Bayes factors
  - Suppose our prior on models is uniform $p(m) \propto 1$. Then model selection is equivalent to picking the model with the highest marginal likelihood.
  - the Bayes factor

$$BF_{1,0} \triangleq \frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_0)} = \frac{p(M_1|\mathcal{D})}{p(M_0|\mathcal{D})} / \frac{p(M_1)}{p(M_0)}$$

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

Table 5.1 Jeffreys' scale of evidence for interpreting Bayes factors.

Bayesian model selection

- Bayes factors
  - an example —— testing if a coin is fair

$$M_0 : \theta = 0.5 \quad M_1 : \theta \in [0, 1]$$

$$p(\mathcal{D}|M_0) = \left(\frac{1}{2}\right)^N \quad p(\mathcal{D}|M_1) = \int p(\mathcal{D}|\theta)p(\theta)d\theta = \frac{B(\alpha_1 + N_1, \alpha_0 + N_0)}{B(\alpha_1, \alpha_0)}$$

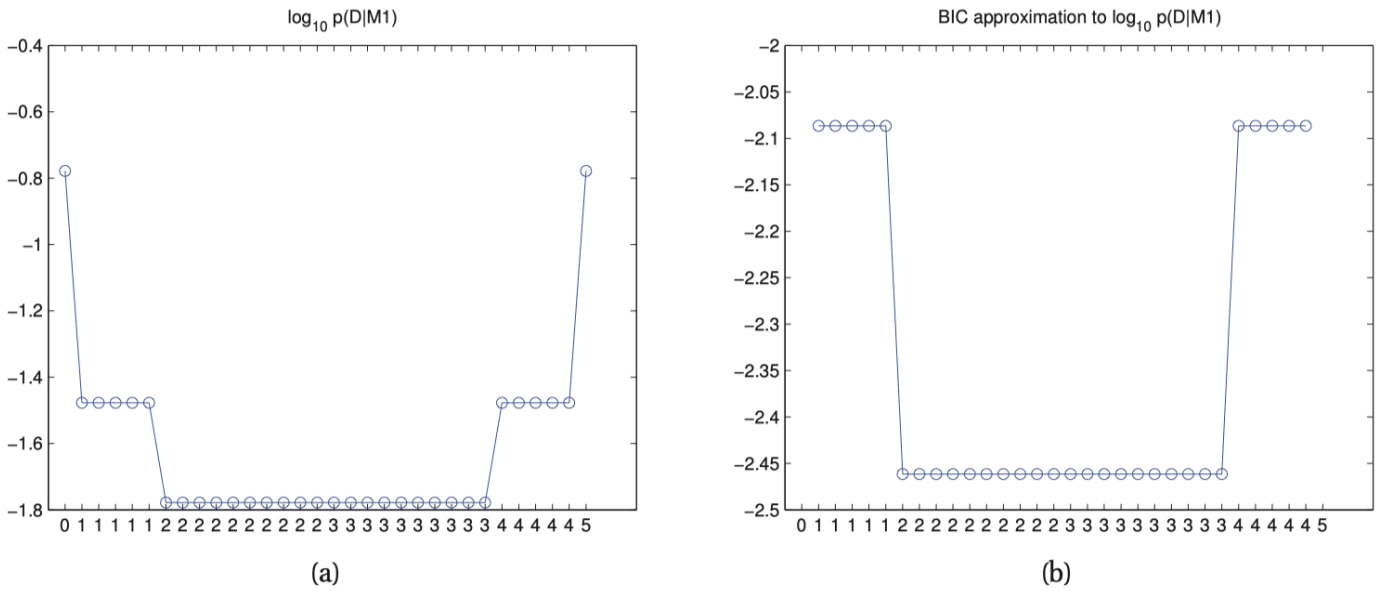


Figure 5.9 (a) Log marginal likelihood for the coins example. (b) BIC approximation. Figure generated by `coinsModelSelDemo`.

5.Bayesian statistics (5.4-5.7)

Liyuan Hu

5.4 Priors

- Uninformative priors
  - Uniform distribution $Beta(1,1)$
  - Haldane prior $Beta(0,0)$
  - Jeffreys priors
- Robust priors
- Mixtures of conjugate priors

Uninformative Priors

- If we don't have strong beliefs about what θ should be, it is common to use an uninformative or non-informative prior, and to let the data speak for itself.
- Designing uninformative priors is tricky.

Uninformative Priors for the Bernoulli

- Consider the Bernoulli parameter $P(x|\theta) = \theta^x(1-\theta)^{n-x}$, and in the coin flipping experiment, we have observed N_1 heads and N_0 tails.
- The uniform distribution Uniform(0, 1)? (equivalent to $Beta(1,1)$ on θ)
  - We can predict the MLE is $N_1/(N_1+N_0)$
  - Posterior: $Beta(N_1+1, N_0+1)$; posterior mean: $E[\theta|X] = (N_1+1)/(N_1+N_0+2)$. Prior isn't completely uninformative!

Uninformative Priors for the Bernoulli

- By decreasing the magnitude of the pseudo counts, we can lessen the impact of the prior.

  Haldane prior: $Beta(0, 0)$

  - An improper prior: does not integrate to 1. $1/\mathrm{B}(a, b) \int_0^1 \theta^{-1}(1-\theta)^{-1} d\theta$
  - Results in the posterior Beta(x, n − x) which will be proper as long as $n - x! = 0$ and $x! = 0$.
- We will see that the "right" uninformative prior is $Beta(\ 1/2\ ,1/2\ )$.

Jeffreys Prior

- A uninformative prior should be invariant to the parametrization used. if $p(\theta)$ is uninformative, then any reparameterization of the prior, such as $\theta = h(\varphi)$ for some function h, should also be uninformative.
  - $h(\phi)$ should be a smooth bijection.
  - Jeffreys prior : $p(\theta) \propto \sqrt{\det I(\theta)}$, where $I(\theta)$ is the Fisher information for θ, and is invariant under reparameterization of the parameter vector θ.
  - $p(\theta) \propto \sqrt{\det I(\theta)} \Leftrightarrow p(\phi) \propto \sqrt{\det I(\phi)}$

Jeffreys Prior

- For an alternative parametrization φ we can derive $p(\phi) \propto \sqrt{\det I(\phi)}$ from $p(\theta) \propto \sqrt{\det I(\theta)}$, using the change of variables theorem and the definition of Fisher information:

$$p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| \propto \sqrt{I(\theta)\left(\frac{d\theta}{d\phi}\right)^2} = \sqrt{E\left[\left(\frac{d\ln L}{d\theta}\right)^2\right]\left(\frac{d\theta}{d\phi}\right)^2}$$

$$= \sqrt{E\left[\left(\frac{d\ln L}{d\theta}\frac{d\theta}{d\phi}\right)^2\right]} = \sqrt{E\left[\left(\frac{d\ln L}{d\phi}\right)^2\right]} = \sqrt{I(\phi)}$$

Jeffreys Prior for the Bernoulli and Multinoulli

- Suppose X ~ Ber(θ). The log-likelihood for a single sample is log p(X|θ) = X

$$\log p(X|\theta) = X\ log\theta + (1 - X)\log(1 - \theta)$$

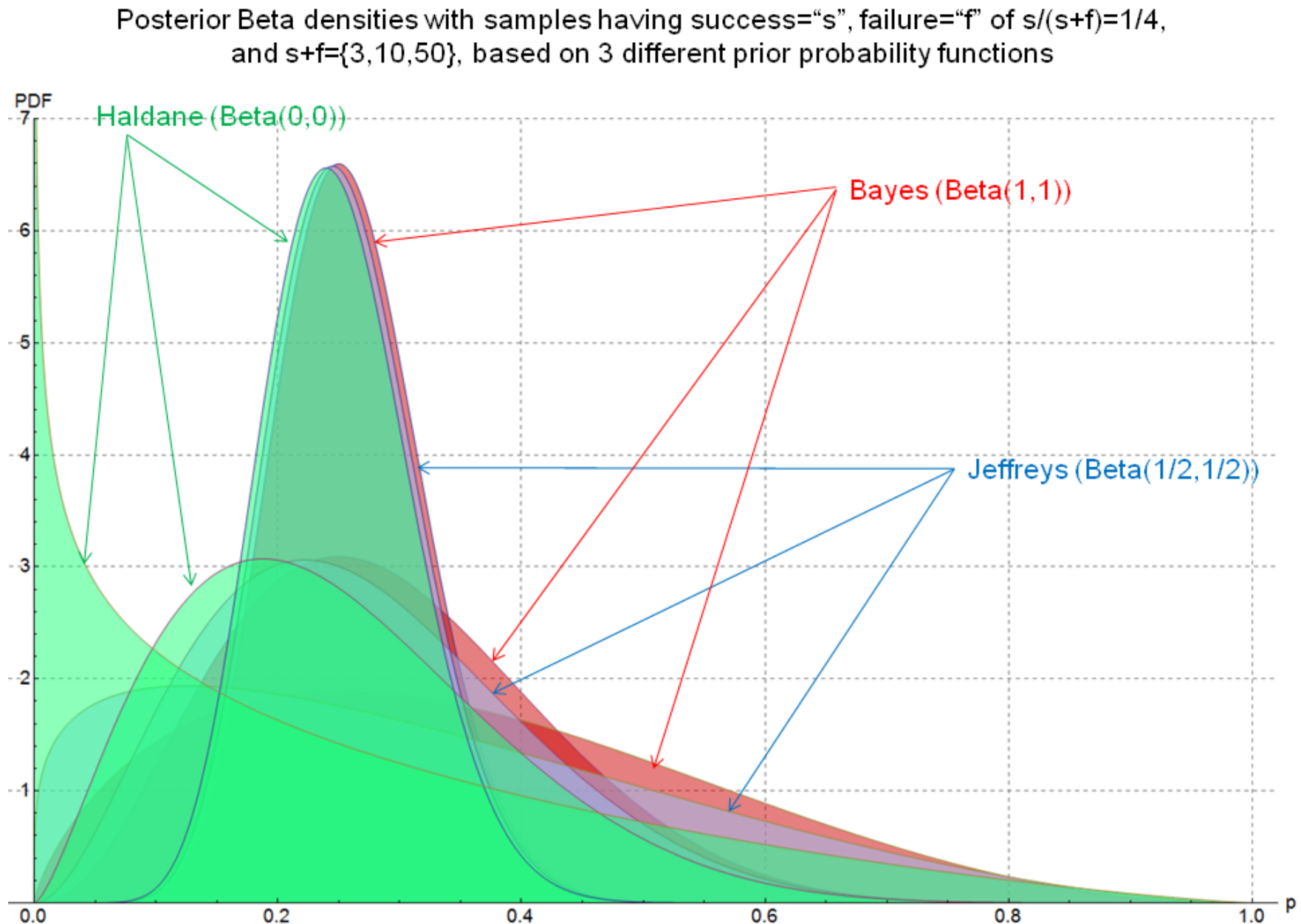
- The Fisher information is $\frac{1}{\theta(1-\theta)}$
- Hence Jeffreys prior is

$$p(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \propto \text{Beta}(\frac{1}{2}, \frac{1}{2}).$$

- Jeffreys prior for a Multinoulli random variable with K states is

$$p(\theta) \propto Dir(\frac{1}{2}, \ldots, \frac{1}{2})$$

Beta(1,1), Beta(0,0), and Beta(½, ½)



Jeffreys Prior for Location Parameters

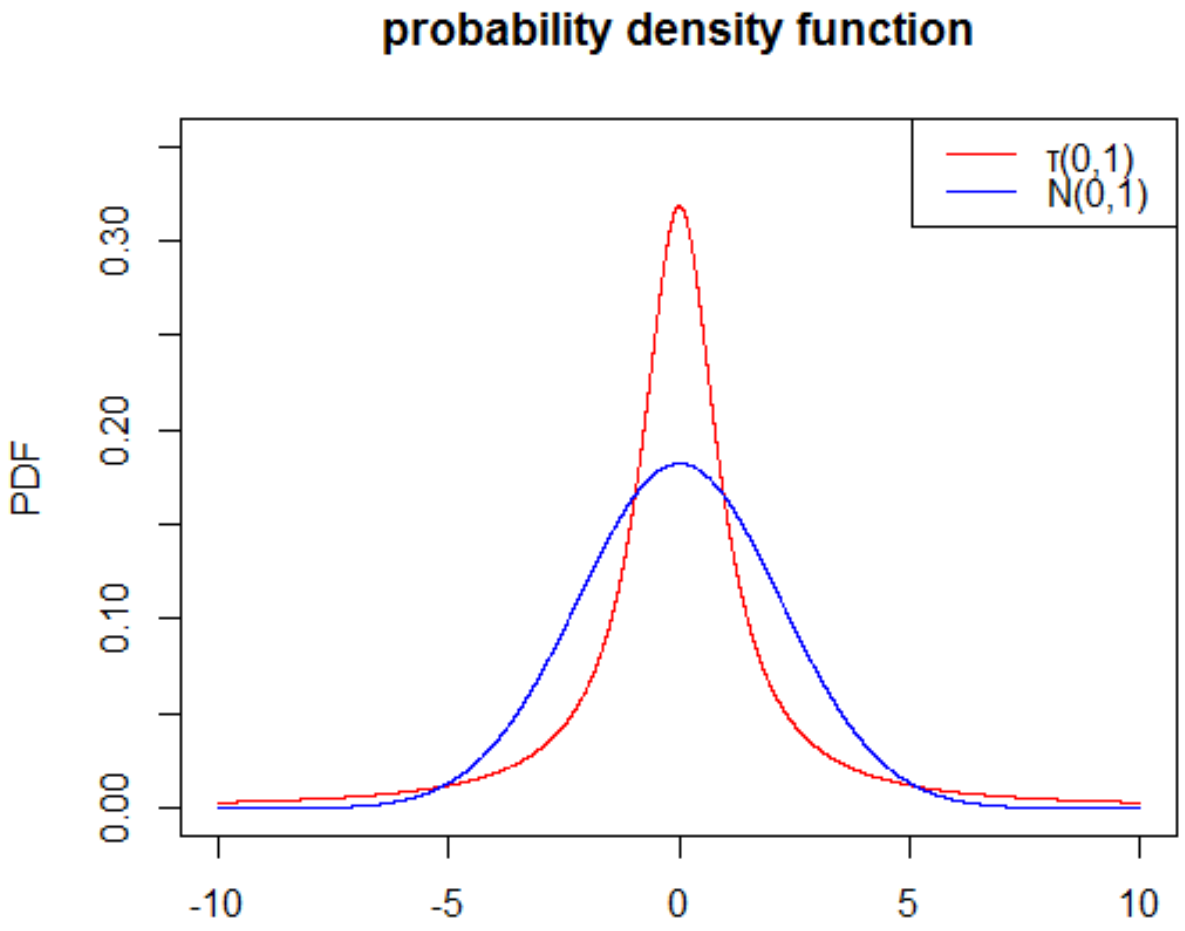
- Translation invariant prior: The probability mass assigned to any interval, $[A, B]$ is the same as that assigned to any other shifted interval of the same width, such as $[A - C, B - C]$
- Jeffreys prior for a location parameters (such as the Gaussian mean), is $p(\mu) \propto 1$.
  - The score function is the gradient of the log-likelihood: $s(\mu) = (x - \mu)/\sigma$
  - The observed information is the second derivation of the log-likelihood: $-1/\sigma$
  - The fisher information: $I(\mu) = 1/\sigma$
- $\int_{A-C}^{B-C} p(\mu)d\mu = (B - C) - (A - C) = (B - A) = \int_{A}^{B} p(\mu)d\mu$
- Can be approximate by using a $p(\mu) = N(\mu|0, \infty)$. This is an improper prior, but we can "nail down" the location as soon as we have seen a single data point.

Jeffreys Prior for Scale Parameters

- Scale invariant prior: The probability mass assigned to any interval $[\mathrm{A}, \mathrm{B}]$ is the same as that assigned to any other interval $[\mathrm{A}/\mathrm{C}, \mathrm{B}/\mathrm{C}]$.
- Jeffreys prior for a scale parameter (such as the Gaussian variance), is $p(\sigma^2) \propto 1/\sigma^2$.
- $\int_{A/C}^{B/C} p(s)ds = [\log s]_{A/C}^{B/C} = \log\left(\frac{B}{C}\right) - \log\left(\frac{A}{C}\right) = \int_{A}^{B} p(s)ds$
- Can be approximate by using a degenerate Gamma distribution $p(s) = \mathrm{Ga}(s|0, 0)$. The prior $p(s) \propto 1/s$ is also improper, but the posterior is proper as soon as we have seen $N \geq 2$ data points (since we need at least two data points to estimate a variance).

Robust Priors

- Not very confident about our prior, want to make sure it does not have an undue influence on the result.
- Typically have heavy tails: avoids forcing things to be too close to the prior mean.



Mixtures of Conjugate Priors

- Robust priors are useful but computationally expensive.
- Conjugate priors simplify the computation, but often not robust, and not flexible
- A mixture of conjugate priors is also conjugate. such priors provide a good compromise between computational convenience and flexibility.
- For example, suppose we are modelling coin tosses, and we think the coin is either fair, or is biased towards heads. This cannot be represented by a Beta distribution. However, we can model it using a mixture of two Beta distributions. We, might use:

$$p(\theta) = 0.5\, Beta(\theta|20, 20) + 0.5\, Beta(\theta|30, 10)$$

Mixtures of Conjugate Priors

- The prior has the form

$$p(\theta) = \sum_k P(Z = k) p(\theta | Z = k)$$

where $Z = k$ means that θ comes from mixture component k, $P(Z = k)$ are called the prior mixing weights, and each $p(\theta | Z = k)$ is conjugate.

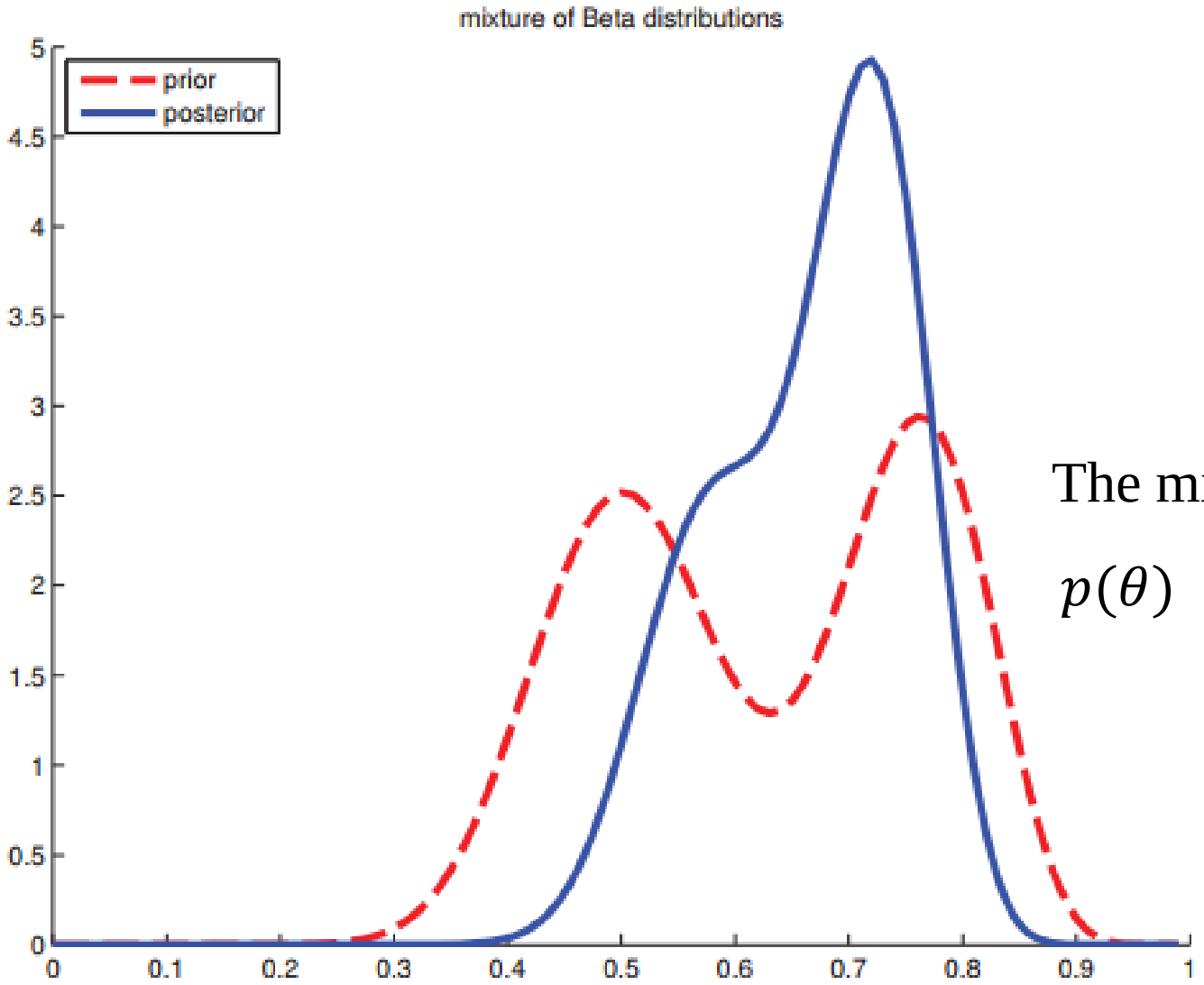
- Posterior can also be written as a mixture of conjugate distributions as follows:

$$p(\theta | X) = \sum_k P(Z = k | X) p(\theta | X, Z = k)$$

where $P(Z = k | X)$ are the posterior mixing weights given by

$$p(Z = k | D) = \frac{P(Z=k) p(D | Z=k)}{\sum_{k'} P(Z=k') p(D | Z=k')}$$

Mixtures of Conjugate Priors - Example



The mixture prior:

$$p(\theta) = 0.5\, Beta(\theta|20,20) + 0.5\, Beta(\theta|10,10)$$

5.5 Hierarchical Bayes

- Prior: $p(\theta|\eta)$. η is called a hyper-parameters.
- If we don't know how to set η
  - Use uninformative priors
  - Put a prior on the priors
- Two important concepts in deriving the posterior distribution
  - Hyper-parameters: parameters of the prior distribution, i.e. the η in $p(\theta|\eta)$
  - Hyperpriors: distributions of hyper-parameters η

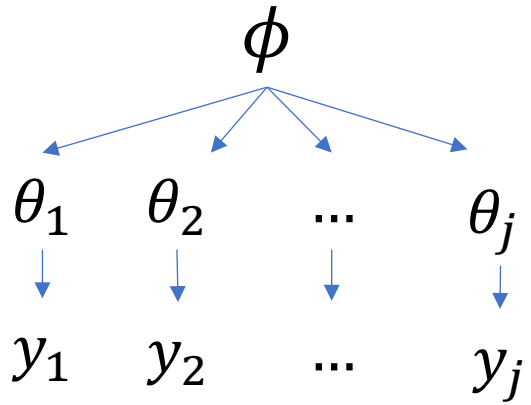
Framework of Hierarchical Bayes

- Let y_j be an observation and θ_j a parameter governing the data generating process for y_j. Assume further that the parameters $\theta_1, \theta_2, \ldots, \theta_j$ are generated exchangeably from a common population, with distribution governed by a hyperparameter ϕ.
- The Bayesian hierarchical model contains the following stages:

  Stage I: $y_j|\theta_j, \phi \sim p(y_j|\theta_j, \phi)$

  Stage II: $\theta_j|\phi \sim p(\theta_j|\phi)$

  Stage III: $\phi \sim p(\phi)$



- The posterior distribution is proportional to:

$$p(\phi, \theta_j|y) \propto p(y_j|\theta_j, \phi) \text{ [using Bayes' Theorem]}$$

$$\propto p(y_j|\theta_j)p(\theta_j|\phi)p(\phi)$$

*Note that the likelihood depends ϕ only through θ_j, i.e. the likelihood does not depend on ϕ.

Hierarchical Bayes – Computations of $p(\theta|y)$

1. Conditional on ϕ(i.e., by keeping it fixed), compute:
   a. the prior predictive distribution of y:

$$p(y|\phi) = \int p(y,\theta|\phi)d\theta = \int p(y|\theta,\phi)p(\theta|\phi)d\theta$$

   b. the posterior distribution of θ:

$$p(\theta|y,\phi) = \frac{p(y|\theta,\phi)p(\theta|\phi)}{p(y|\phi)}$$

2. By using $p(y|\phi)$ from step 1, compute:
   a. the distribution of y:

$$p(y) = \int p(y,\phi)d\phi = \int p(y|\phi)p(\phi)d\phi$$

   b. the posterior marginal distribution of ϕ:

$$p(\phi|y) = \frac{p(y|\phi)p(\phi)}{p(y)}$$

Hierarchical Bayes – Computations of $p(\theta|y)$

3. Compute the posterior joint distribution of ϕ and θ:

$$p(\phi, \theta|y) = p(\theta|y, \phi)p(\phi|y)$$

4. Compute the posterior marginal distribution of θ:

$$p(\theta|\mathrm{y}) = \int p(\phi, \theta|y)d\phi$$

When we are not able to carry out the integrations required to derive the predictive distributions, or when we cannot compute posteriors with Bayes' rule, then we can use other computational methods.

Hierarchical Bayes - Example

- Consider the problem of predicting cancer rates in various cities. In particular, suppose we measure the number of people in various cities, N_i, and the number of people who died of cancer in these cities, x_i . We assume $x_i \sim Bin(N_i, \theta_i)$, and we want to estimate the cancer rates θ_i.
  - Estimate them all separately (sparse data problem/ black swan paradox, see p77 3.3.4.1) underestimation of the rate.
  - Parameter tying**:** assume all the θ_i are the same (too strong an assumption)
  - Assuming the θ_i are drawn from some common distribution, say $\theta_i \sim$ Beta(a, b).

Hierarchical Bayes - Example

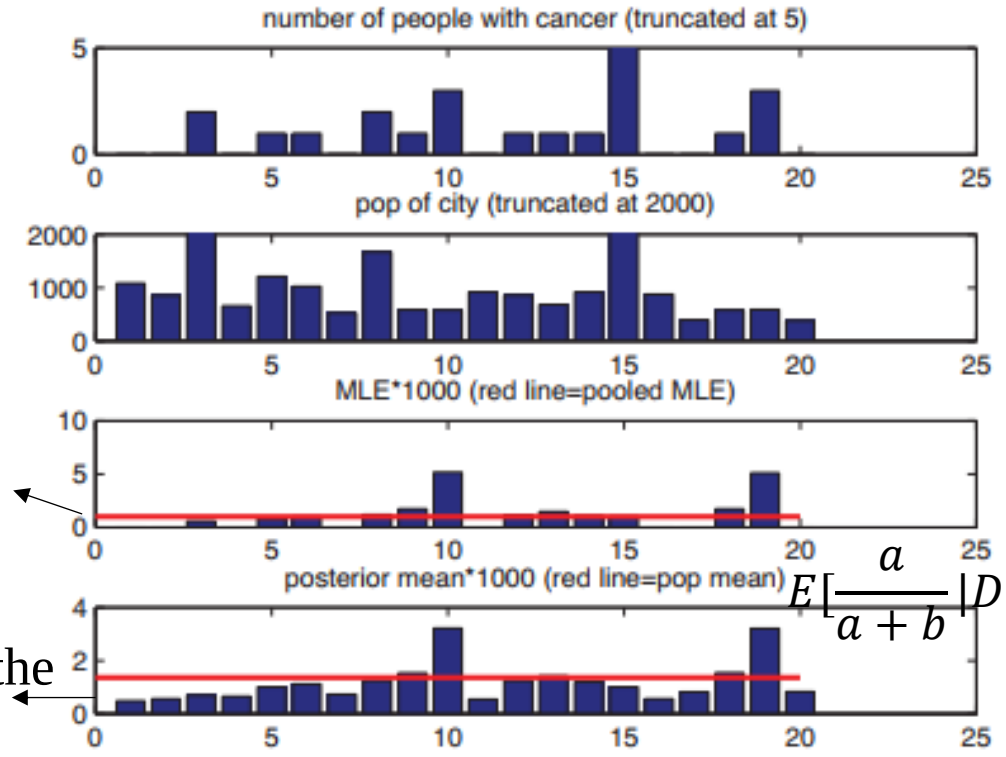
- The full joint distribution:

$$p(D, \boldsymbol{\theta}, \boldsymbol{\phi}|N) = p(\boldsymbol{\phi}) \prod_{i=1}^{N} Bin(x_i|N_i, \theta_i) Beta(\theta_i|\boldsymbol{\phi})$$

Where $\boldsymbol{\phi} = (a, b)$.

- It is crucial that we infer $\boldsymbol{\phi}$ = (a, b) from the data.
  - If we just clamp it to a constant, the θ_i will be conditionally independent.
  - By treating $\boldsymbol{\phi}$ as an unknown (hidden variable), we allow the data-poor cities to borrow statistical strength from data-rich ones.

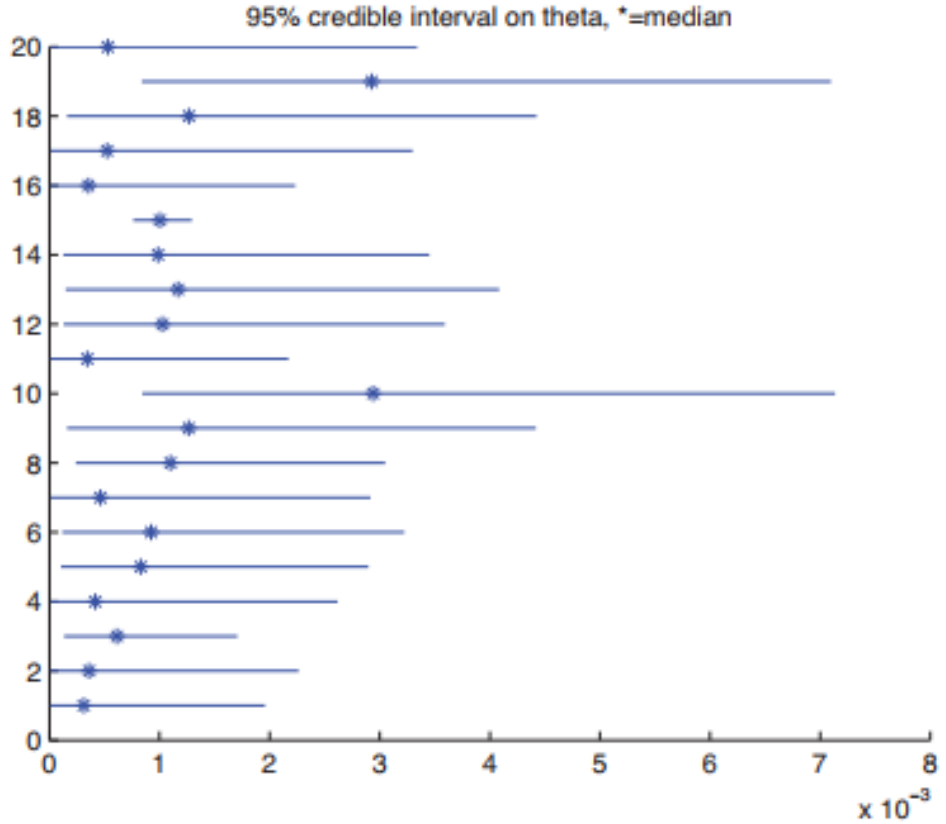
Hierarchical Bayes - Example



Sparse data pro

Shrunk towards the pooled estimate

$E[\frac{a}{a+b}|D]$



- Hierarchical Bayes is a way to tackle the sparse data problem. The posterior mean is shrunk towards the pooled estimate more strongly for cities with small sample sizes N_i

- City 15, which has a very large population, has small posterior uncertainty. And has the largest impact on the posterior estimate of $\boldsymbol{\phi}$.
- Cities 10 and 19, which have the highest MLE, also have the highest posterior uncertainty

5.6 Empirical Bayes

- Can we use the training data to set the hyper-parameters?
- In theory: No!
  - It would not be a “prior”.
  - It’s no longer the right thing to do.
- In practice: Yes!
  - Approach 1: split into training/validation set or use cross-validation as before.
  - Approach 2: optimize the marginal likelihood (“evidence”):

$$\int p(D|\theta)p(\theta|\eta)d\theta$$

- Also called type II maximum likelihood or evidence maximization or empirical Bayes. In machine learning, it is sometimes called the evidence procedure.

Empirical Bayes for Beta-binomial Model

- Let us return to the cancer rates model. The marginal likelihood:

$$p(\mathcal{D}|a,b) = \prod_i \int \text{Bin}(x_i|N_i,\theta_i)\text{Beta}(\theta_i|a,b)d\theta_i$$
$$= \prod_i \frac{B(a+x_i, b+N_i-x_i)}{B(a,b)}$$

- Having estimated *a* and *b*, we can plug in the hyper-parameters to compute the posterior $p(\theta_i|\hat{a},\hat{b},D)$ in the usual way, using conjugate analysis.

Empirical Bayes for Beta-binomial Model

- Local MLE: $\hat{\theta}_{i\,MLE} = \frac{x_i}{N_i}$, prior mean: $m_1 = \frac{a}{a+b}$, posterior mean: $E[\theta_i|D] = \frac{a+x_i}{a+b+N_i} = \lambda_i m_1 + (1-\lambda_i)\hat{\theta}_{i\,MLE}$, where $\lambda_i = \frac{a+b}{N_i+a+b}$. So larger the N_i, the smaller is λ_i, and hence the closer the posterior mean is to the MLE.
  - The posterior mean of each θ_i is a weighted average of its local MLE and the prior means, but since ϕ= (a, b) is estimated based on all the data, each θ_i is influenced by all the data.

5.7 Bayesian Decision Theory

- Bayesian decision theory:
  - Minimizing the posterior expected loss: $\rho(a|\mathbf{x}) \triangleq E\big(L(y,a)\big) = \sum_y L(y,a)p(y|\mathbf{x})$
- The expected loss is a combination of:
  - posterior: $p(y|\mathbf{x})$, $\mathbf{x} \in$ X(observation), $y \in$ Y (label or parameters)
  - decision rule: $a(\mathbf{x})$, $a(\mathbf{x}) \in A$ (action space)
  - loss function: $L(y, a(\mathbf{x}))$ cost of making decision $a(\mathbf{x})$ when true state is y
- The action that minimizes the posterior expected loss is called the Bayes estimator/ Bayes decision rule:

$$\delta(\mathbf{x}) = arg \min_{a \in A} \rho(a|\mathbf{x})$$

(If y is continuous, we should replace the sum with an integral.)

Bayesian Estimator for 0-1 Loss

- The 0-1 loss is defined by

$$L(y, a) = \mathbb{I}(y \neq a) = \begin{cases} 0 & \text{if } a = y \\ 1 & \text{if } a \neq y \end{cases}$$

$a = \hat{y}$ is the estimate.

- The posterior expected loss is

$$\rho(a|\mathbf{x}) = P(a(\mathbf{x}) \neq y|\mathbf{x}) = 1 - P(y = a(\mathbf{x})|\mathbf{x})$$

- Hence the action that minimizes the expected loss is the posterior mode or MAP estimate

$$y^*(\mathbf{x}) = \arg\max_{y \in Y} P(y|\mathbf{x})$$

- Bayes decision is MAP estimator if the loss function penalizes all errors by the same amount.

This is called the Bayes classifier. The LDA model is a Bayes classifier.

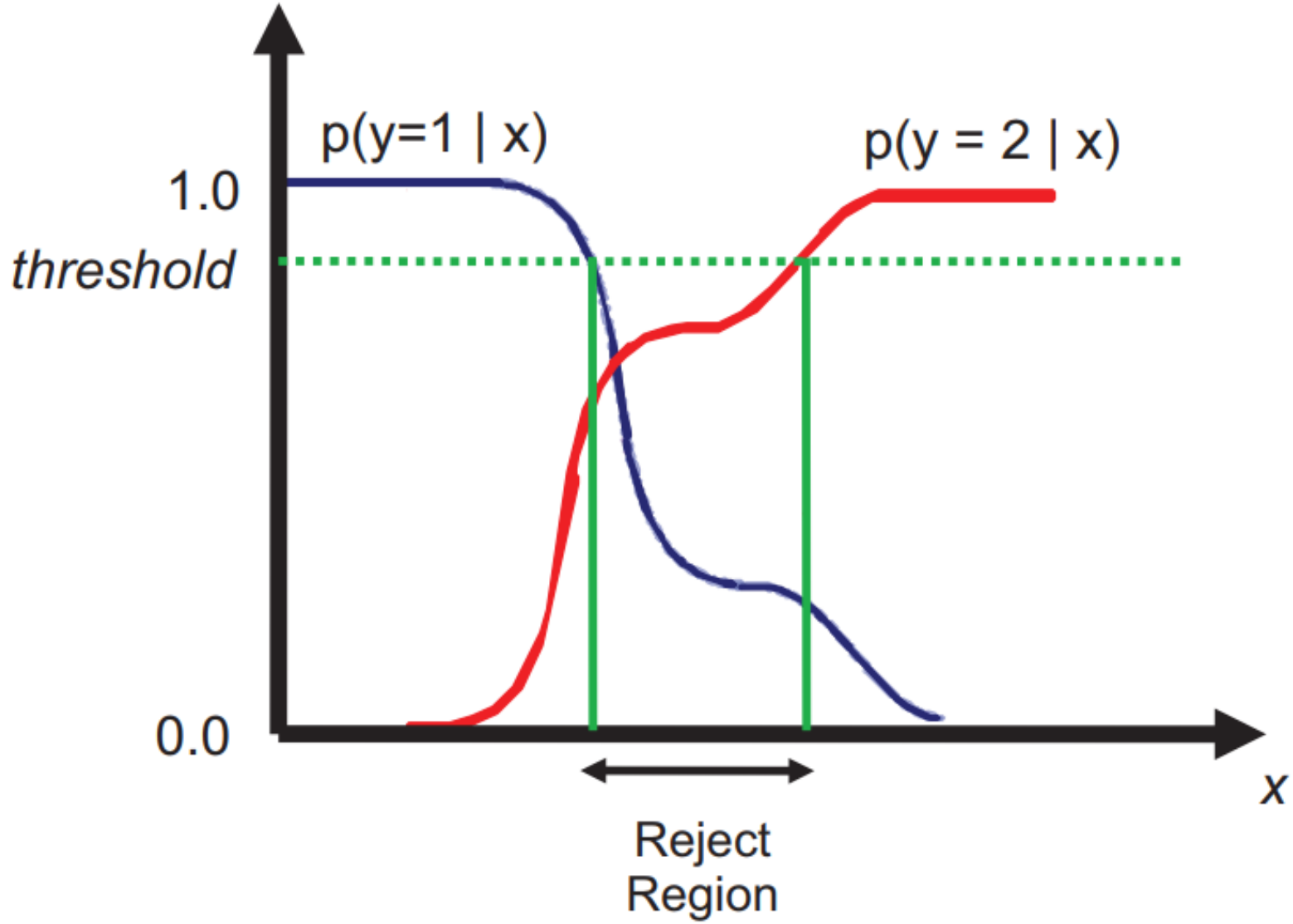
Bayesian Estimator for 0-1 Loss

- Reject option: In classification problems where $p(y|\mathbf{x})$ is very uncertain, we refuse to classify the example.
- Let choosing $a = C + 1$ correspond to picking the reject action, and choosing $a \in \{1, \ldots, C\}$ correspond to picking one of the classes. The loss function:

$$L(y = j, a = i) = \begin{cases} 0 & \text{if } i = j \text{ and } i, j \in \{1, \ldots, C\} \\ \lambda_r & \text{if } i = C + 1 \\ \lambda_s & \text{otherwise} \end{cases}$$

- where λ_r is the cost of the reject action, and λ_s is the cost of a substitution error. The optimal action is to pick the reject action if the most probable class has a probability below 1-λ_r/λ_s

Bayesian Estimator for 0-1 Loss



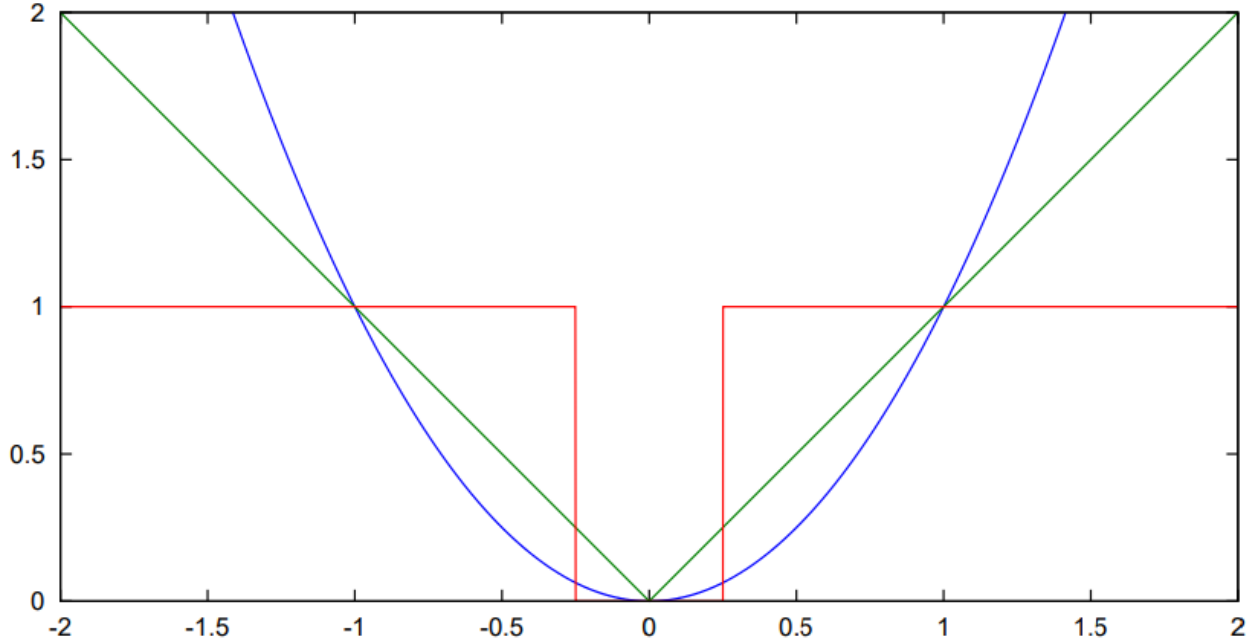
Bayesian Estimator for Continuous Parameters

- Let $\epsilon = \hat{\theta} - \theta$. Many cost assignments can be written as $C_\theta(\hat{\theta})$
  - Squared loss: $C_\theta(\hat{\theta}) = \|\epsilon\|_2^2$
  - Absolute loss: $C_\theta(\hat{\theta}) = \|\epsilon\|_1$.
  - Uniform loss ("hit or miss"):

$$C_\theta(\hat{\theta}) = \begin{cases} 0 & \|\epsilon\|_\infty \leq \frac{\Delta}{2} \\ 1 & \text{otherwise} \end{cases}$$

The hit or miss loss become

$\|\epsilon\|_0 = 1(\theta = a)$. when Δ approaches 0.



Bayesian Estimator for l_2 Loss is the Posterior Mean

- The posterior expected loss is given by

$$\rho(a|\mathbf{x}) = \mathbb{E}\left[(y-a)^2|\mathbf{x}\right] = \mathbb{E}\left[y^2|\mathbf{x}\right] - 2a\mathbb{E}\left[y|\mathbf{x}\right] + a^2$$

- Hence the optimal estimate is the posterior mean:

$$\frac{\partial}{\partial a}\rho(a|\mathbf{x}) = -2\mathbb{E}\left[y|\mathbf{x}\right] + 2a = 0 \;\Rightarrow\; \hat{y} = \mathbb{E}\left[y|\mathbf{x}\right] = \int yp(y|\mathbf{x})dy$$

This is often called the minimum mean squared error estimate or MMSE estimate

- In a linear regression problem, we have $p(y|\boldsymbol{x},\theta) = N\left(y\middle|\mathbf{x}^{\boldsymbol{T}}\boldsymbol{w},\sigma^2\right)$. In this case, the optimal estimate given some training data D is given by $E[y|\mathbf{x},D] = x^T E[\boldsymbol{w}|D]$

Bayesian Estimator for l_1 Loss is the Posterior Median

- The square loss penalizes deviations from the truth quadratically, and thus is sensitive to outliers.
- A more robust alternative is the absolute or 1 loss, $L(y, a) = |y - a|$.
- If we choose the absolute value as the cost function, we have to minimize

$$\arg\min_{\hat{y}} \int |y - \hat{y}|\, p(y|\mathbf{x})\, dy$$

- $\int |y - \hat{y}| p(y|\mathbf{x})\, dy = \int_{-\infty}^{\hat{y}} (\hat{y} - y) p(y|\mathbf{x})\, dy - \int_{\hat{y}}^{\infty} (\hat{y} - y) p(y|\mathbf{x})\, dy$.

Bayesian Estimator for l_1 Loss is the Posterior Median

- $\frac{\partial \int |y-\hat{y}|p(y|\mathbf{x})\,dy}{\partial \hat{y}} = \int_{-\infty}^{\hat{y}} p(y|\mathbf{x})\,dy - \int_{\hat{y}}^{\infty} p(y|\mathbf{x})dy = 0$

  $\Rightarrow \hat{y}$ is median of the posteriori PDF.

- Because when $\hat{y} =$ Median $y|\mathbf{x}$, the posterior loss reaches its minimum 0, $\hat{y}$ is what we want. This is called the Minimum Mean Absolute Error estimate or MMAE estimate.

Bayesian Estimator for Uniform Loss is the Mode

- For the hit-or-miss case, we also need to minimize the inner integral:

$$\arg\min_{\hat{y}} \int C(\hat{y}-y)p(y|\mathbf{x})dy$$

With $C(x) = \begin{cases} 0, |x| < \Delta/2 \\ 1, |x| > \Delta/2 \end{cases}$

- The integral becomes

$$\int C(\hat{y}-y)p(y|\mathbf{x})dy = \int_{-\infty}^{\hat{y}-\frac{\Delta}{2}} p(y|\mathbf{x})dy + \int_{\hat{y}+\frac{\Delta}{2}}^{\infty} p(y|\mathbf{x})dy$$

or in a simplified form $\int C(\hat{y}-y)p(y|\mathbf{x})dy = 1 - \int_{\hat{y}-\frac{\Delta}{2}}^{\hat{y}+\frac{\Delta}{2}} p(y|\mathbf{x})dy$

- This is minimized by maximizing $\int_{\hat{y}-\frac{\Delta}{2}}^{\hat{y}+\frac{\Delta}{2}} p(y|\mathbf{x})dy$

Bayesian Estimator for Uniform Loss is the Mode

- Maximizing $\int_{\hat{y}-\frac{\Delta}{2}}^{\hat{y}+\frac{\Delta}{2}} p(y|\mathbf{x})dy$
- For small Δ and smooth $p(y|\mathbf{x})$ the maximum of the integral occurs at the maximum of $p(y|\mathbf{x})$. Therefore, the estimator is the mode (the highest value) of the posteriori PDF. Thus the name Maximum a Posteriori (MAP) estimator.
- Assume that $\Delta \rightarrow 0^{+}$. Then the limiting case of the above loss is l_0 loss

Bayesian Estimator: Summary of Common Loss

- l_2 loss: MMSE

$$\hat{y}_{mmse} = E[y|\mathbf{x}]$$

- l_1 loss: MMAE

$$\hat{y}_{mmae} = Median[y|\mathbf{x}]$$

- Uniform loss: MAP

$$\hat{y}_{map} = Mode[y|\mathbf{x}]$$