Bayesian Statistics

Agenda

Part I - Generative Models

Part II - Bayesian Analysis

Part I - Generative Models

$$p(y = c|\mathbf{x}, \boldsymbol{\theta}) \propto p(\mathbf{x}|y = c, \boldsymbol{\theta})p(y = c|\boldsymbol{\theta})$$

posterior

likelihood

prior

The numbers game - human estimates

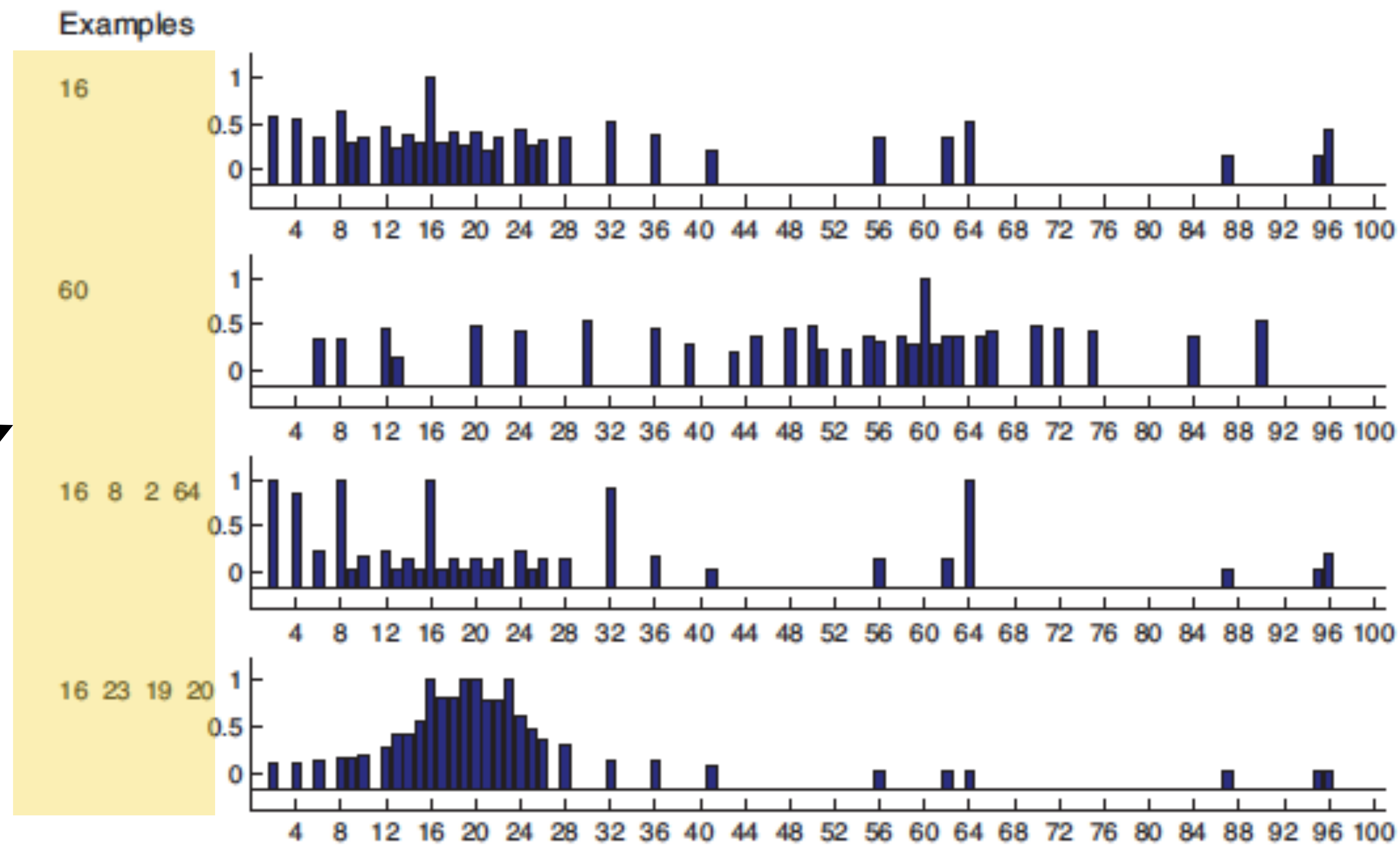
The game proceeds as follows.

I choose some simple arithmetical concept C , such as “prime number” or “a number between 1 and 10”.

I then give you a series of randomly chosen positive examples D = {x1, . . . , xN} drawn from C , and ask you whether some new test case ˜x belongs to C

i.e., I ask you to classify ˜x

The numbers game - human estimates



data seen

Say $D = \{ 16, 8, 2, 64 \}$

Am I taking C = “powers of 2”, or C = “even numbers”?

(both are consistent with the data)

Say D = { 16, 8, 2, 64 }

Am I taking C = “powers of 2”, or C = “even numbers”?

(both are consistent with the data)

Occam’s razor

$$p(\mathcal{D}|h) = \left[\frac{1}{\text{size}(h)}\right]^N = \left[\frac{1}{|h|}\right]^N$$

likelihood

prior

h = “powers of 2”   h’ = “powers of 2, except 32”

What now?

prior

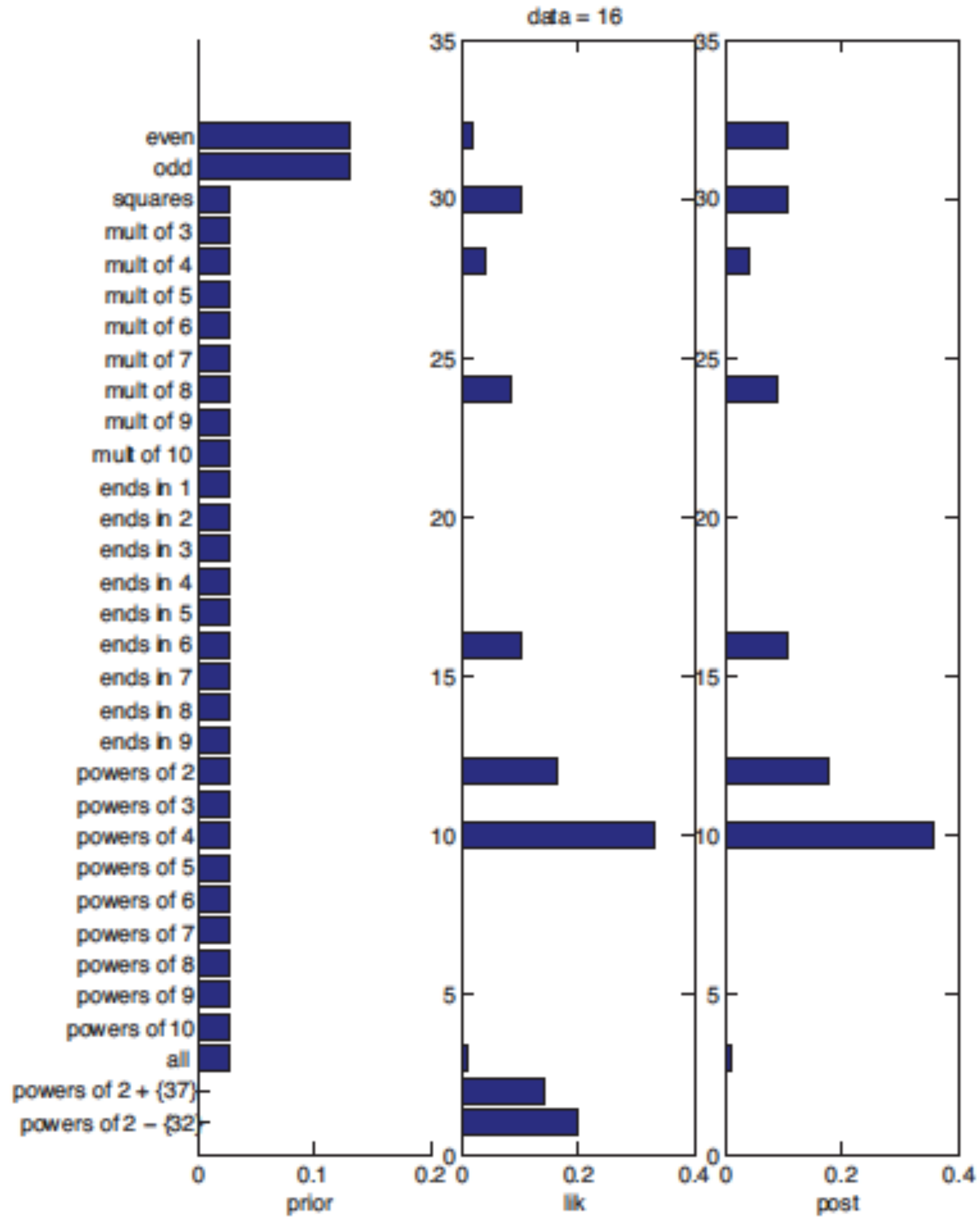
h = "powers of 2"   h' = "powers of 2, except 32"

What now?

the prior $p(h)$ is the **subjective** part of Bayesian reasoning

$$p(h \mid D) = p(D \mid h) \cdot p(h) / \text{normalisation}$$

posterior   likelihood   prior



data = 16
even
odd
squares
mult of 3
mult of 4
mult of 5
mult of 6
mult of 7
mult of 8
mult of 9
mult of 10
ends in 1
ends in 2
ends in 3
ends in 4
ends in 5
ends in 6
ends in 7
ends in 8
ends in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
all
powers of 2 + {37}
powers of 2 − {32}
0
0.1
0.2
prior
0
0.2
0.4
lik
0
0.2
0.4
post
0
5
10
15
20
25
30
35

Recall Naïve Bayes

We want p(y=class | D=words)

What is likelihood p(D | y=c) ?

What is prior p(y = c)?

e.g. Mary had a little lamb, little lamb, little lamb,
Mary had a little lamb, its fleece as white as snow

Beta-Binomial

Counting dice head, unknown bias

likelihood $\mathrm{Bin}(k|n,\theta) \triangleq \binom{n}{k}\theta^k(1-\theta)^{n-k}$

prior $\mathrm{Beta}(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$

posterior
$$p(\theta|\mathcal{D}) \propto \mathrm{Bin}(N_1|\theta, N_0+N_1)\mathrm{Beta}(\theta|a,b)$$
$$\propto \mathrm{Beta}(\theta|N_1+a, N_0+b)$$

Bayesian inference for online learning

$$
\begin{aligned}
p(\theta|\mathcal{D}_a, \mathcal{D}_b) &\propto p(\mathcal{D}_b|\theta)p(\theta|\mathcal{D}_a) \\
&\propto \text{Bin}(N_1^b|\theta, N_1^b + N_0^b)\text{Beta}(\theta|N_1^a + a, N_0^a + b) \\
&\propto \text{Beta}(\theta|\ N_1^a + N_1^b + a, N_0^a + N_0^b + b)
\end{aligned}
$$

Part II - Bayesian Analysis

Frequentist approach:

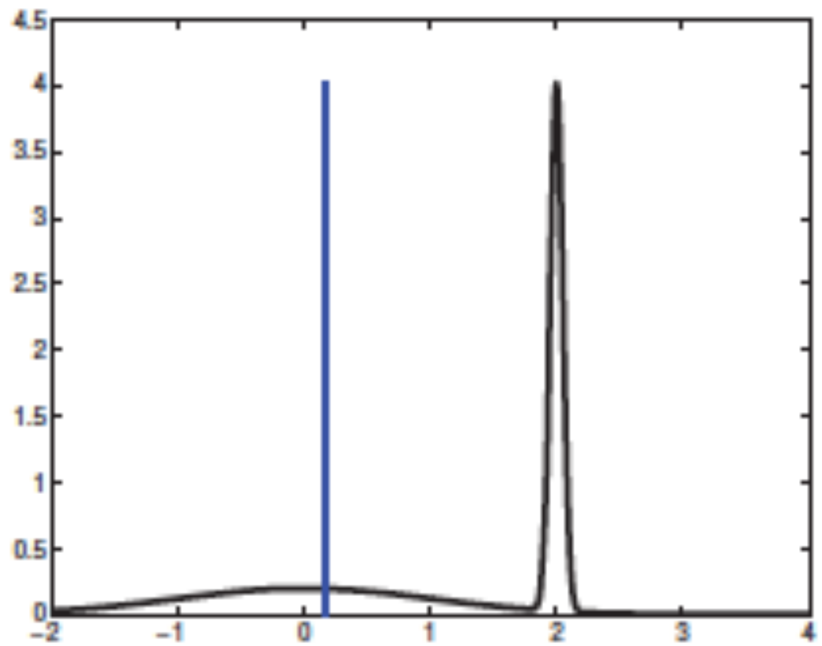
$$\hat{\theta}_{MLE} = \arg\max_{\theta} p(D|\theta)$$

Bayesian approach:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(D|\theta)p(\theta)$$

we care about the prior

problems with a point estimate



the mode is an untypical point

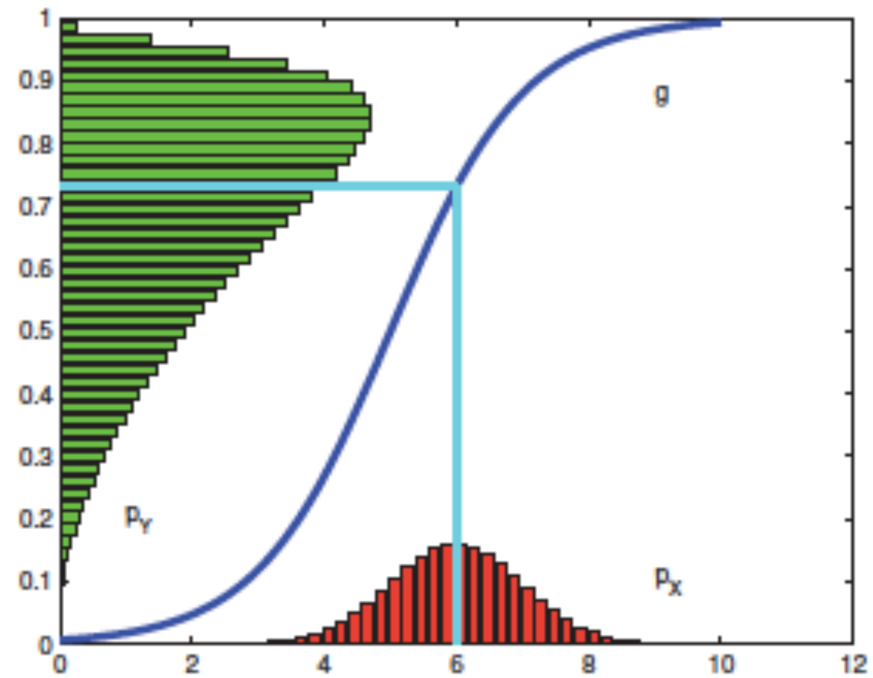


Figure 5.2 Example of the transformation of a density under a nonlinear transform. Note how the mode of the transformed distribution is not the transform of the original mode. Based on Exercise 1.4 of (Bishop 2006b). Figure generated by `bayesChangeOfVar`.

credible interval (CI) vs HPD

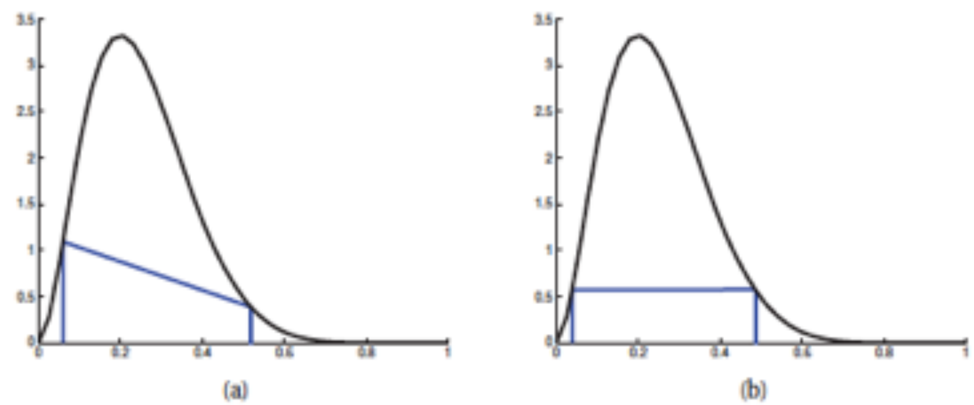


Figure 5.3 (a) Central interval and (b) HPD region for a Beta(3,9) posterior. The CI is (0.06, 0.52) and the HPD is (0.04, 0.48). Based on Figure 3.6 of (Hoff 2009). Figure generated by `betaHPD`.

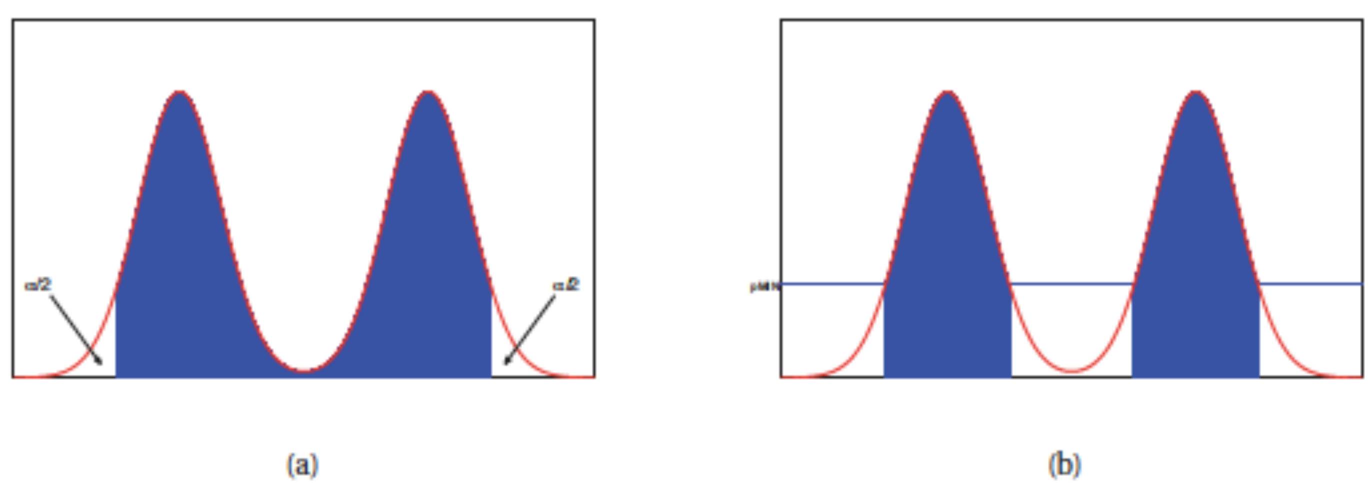


Figure 5.4 (a) Central interval and (b) HPD region for a hypothetical multimodal posterior. Based on Figure 2.2 of (Gelman et al. 2004). Figure generated by `postDensityIntervals`.

Going further than MAP

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