01/27/22

1. HW #2 : due tomorrow (move Q1 to HW# 3)
2. Midterm 1 : everything upto today's lecture.

STAT 206B
Chapter 3: From Prior Information to Prior Distributions

Winter 2022

- A fundamental basis of Bayesian decision theory is that statistical inference should start with the rigorous determination of three factors.

  $\star\star$ the distribution family for the observations (sampling distribution), $f(x \mid \theta)$ for $x \in \mathcal{X}$

  $\star\star$ the prior distribution for the parameter $\pi(\theta)$, $\theta \in \Theta$

  $\star\star$ the loss association with the decisions, $L(\theta, \delta) \in [0, +\infty)$.

- In this chapter, we will discuss prior distributions – CR Chapter 3 and JB Chapters 3 & 4.

† Priors!

- Priors are carriers of external knowledge (outside the data being modeled and analyzed) that is coherently incorporated via Bayes theorem to the inference.

- Parameters (θ) are unobservable.

  $\Rightarrow$ Prior specification is **subjective** in nature.

  $x \in \{0, 1, 2, \ldots\}$

  $\begin{cases} \text{Poi} \\ \text{NB} \end{cases}$

- There is no unique way of choosing a prior distribution.

  $\Rightarrow$ There is no such a thing as *the* prior distribution.

- The choice of the prior distribution has an influence on the resulting inference.

  $\Rightarrow$ Ungrounded prior distributions produce unjustified posterior inference.

† *Is using a prior a problem?*

- The elicitation of a model (likelihood) and loss function is highly subjective, and Bayesians merely divide the necessary subjectivity to two sources - that from the model and from the prior.
- Vast amount scientific information coming from theoretical and physical models is guiding specification of priors and merging such information with the data for better inference.
- Being subjective $\neq$ Being nonscientific

- If complete information is given, an exact prior can be elicited. **However**, it is very rare!

- How to specify priors?

  ⋆⋆ Subjective determination and approximations (Sec 3.2)

  ⋆⋆ Conjugate priors (Sec 3.3)

  ⋆⋆ Noninformative prior distributions (Sec 3.5): *have* ***little influence*** *on the posterior distribution*

- *criticism:* Bayesian inference is overly sensitive to the choice of a prior.

  ⇒ the development of non-informative and robust priors (so change in the prior distribution does not change the posterior inference much)

† Subjective Determination (Sec 3.2) $\theta \in \Theta$

- Subjective prior distributions exist as a consequence of an ordering of relative likelihoods.

- Approximations to the prior distribution. e.g.

  ⋆⋆ When the parameter space Θ is finite, obtain a subjective evaluation of the probabilities of the different values of θ.

  ⋆⋆ When Θ is noncountable (e.g. an interval of the real line), may use the histogram approach.

  - Divide Θ into intervals
  - Determine the subjective probability of each interval
  - Plot a probability histogram
  - If needed, a smooth density $\pi(\theta)$ can be sketched.

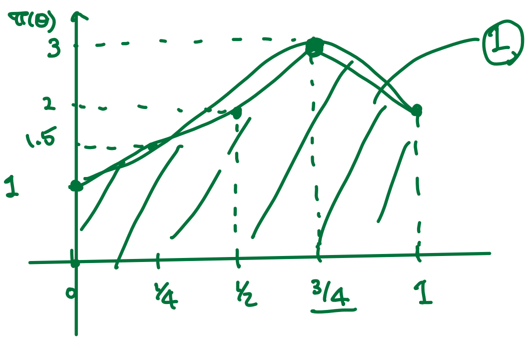
- Approximations to the prior distribution. (contd)

JB Example 1 Assume that $\Theta = [0,1]$. Suppose that

⋆⋆ the parameter point $\theta = 3/4$ is felt to be the most likely, while $\theta = 0$ is the least likely.

⋆⋆ $3/4$ is estimated to be three times as likely to be the true value of θ as is 0.

⋆⋆ $\theta = 1/2$ and $\theta = 1$ are twice likely as $\theta = 0$ while $\theta = 1/4$ is 1.5 times as likely as $\theta = 0$.



$$\pi(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

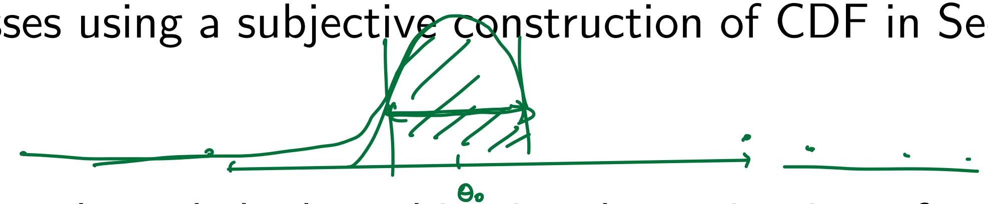
α, β

$\alpha,$

- Approximations to the prior distribution. (contd)

  ⋆⋆ So far we have seen "histogram approach" and "relative likelihood approach".

  ⋆⋆ JB discusses using a subjective construction of CDF in Section 3.2.



- When Θ is not bounded, the subjective determination of π is complicated due to the difficulty of subjectively evaluating the probabilities of the extreme regions of Θ (will see this from Example 3.2.6).

- Using marginal distribution to determine the prior (JB 3.5)

- Parametric Approximations
  - ⋆⋆ *How?* Assume that $\pi(\theta)$ is of a given functional form and then choose the density of this given form which most closely matches prior beliefs (through the *moments*, the *quantiles*, etc).
  - ⋆⋆ Most used (and misused)
  - ⋆⋆ Very useful when a density of a standard functional form gives a good match to the prior information.
  - ⋆⋆ Also useful when only vague prior information is available.
  - ⋆⋆ Considerably different functional forms can often be chosen for the prior density (as will be seen in Example 3.2.6).
  - ⋆⋆ *drawback:* The choice of the parameterized family is often based on ease in the mathematical treatment. The resulting posterior inference is affected by the choice.

- **Ex 3.2.5** Let $X_i \sim \text{Bin}(n_i, p_i)$ be the number of passing students in a freshman calculus course of n_i students. Over the previous years, the average of the p_i is 0.70, with variance 0.1. If we assume that the p_i's are all generated according to the same beta distribution, $\text{Be}(\alpha, \beta)$, then we choose the values of α and β which most closely matches the prior beliefs. That is, set

$\alpha = (0.7)(\alpha+\beta)$

$$\mu = \frac{\alpha}{\alpha+\beta} = 0.7 \quad \text{and} \quad \tau^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, = 0.1$$

and solve for α and β.

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

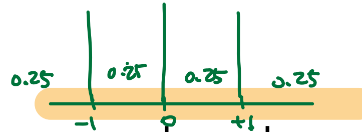
$$= \left(\frac{\alpha}{\alpha+\beta}\right)\left(1 - \frac{\alpha}{\alpha+\beta}\right) \cdot \frac{1}{(\alpha+\beta+1)}$$

$$= 0.7 \times (1-0.7) \cdot \frac{1}{\left(\frac{\alpha}{0.7}+1\right)} = 0.1$$

$\alpha = 0.77$ & $\beta = 0.33$

$\Rightarrow \quad p_i \sim \text{Be}(0.77, 0.33)$

$\theta \in \mathbb{R} = \textcircled{H}$



- **Example 3.2.6** Let $x \sim N(\theta, 1)$. Assume that the prior median of θ is 0, the first quartile is -1, and the third quartile is +1. Use the quadratic loss function.

$m(x)$ is $N(\mu, \sigma^2 + \tau^2)$ $\quad \mu = 0$ $\quad = 1 + 2.19 = 3.19$

★★ Case 1: Assume $\theta \sim N(\mu, \tau^2)$ and set $\mu = 0$ and $\tau^2 = 2.19$. $\sqrt{3.19} = 1.786$

$$\Rightarrow \delta_1^\pi(x) = x - \frac{x}{3.19}$$

$\mu = ?$, $\tau^2 = ?$ $\quad \Rightarrow \quad \mu = 0$

$Pr(\theta < -1) = 0.25$

CDF of $N(0,1)$

$$Pr\Big(\underbrace{\frac{\theta - 0}{\tau}}_{=N(0,1)} < \frac{-1-0}{\tau}\Big) = \Phi\left(\frac{-1}{\tau}\right) = 0.25$$

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}}$$

$$\Phi^{-1}(0.25) = -\frac{1}{\tau}$$

inverse of CDF

$$\Rightarrow \tau = -\frac{1}{\Phi^{-1}(0.25)}$$

$$= \sqrt{2.19}$$

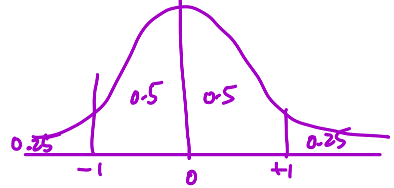
$$\Rightarrow \tau^2 = 2.19$$

$\theta \sim N(0, 2.19)$

$$\Rightarrow \theta | x \sim N\left(\left(\frac{1}{1} + \frac{1}{2.19}\right)^{-1}\left(\frac{x}{1} + \frac{0}{2.19}\right), \left(\frac{1}{1} + \frac{1}{2.19}\right)^{-1}\right)$$

$$\Rightarrow \delta_1^\pi(x) = \left(1 + \frac{1}{2.19}\right)^{-1} x$$

- **Example 3.2.6** (contd)



  ⋆⋆ Case 2: Assume θ has a Cauchy distribution and set $\theta \sim$ Cauchy$(0, 1)$.

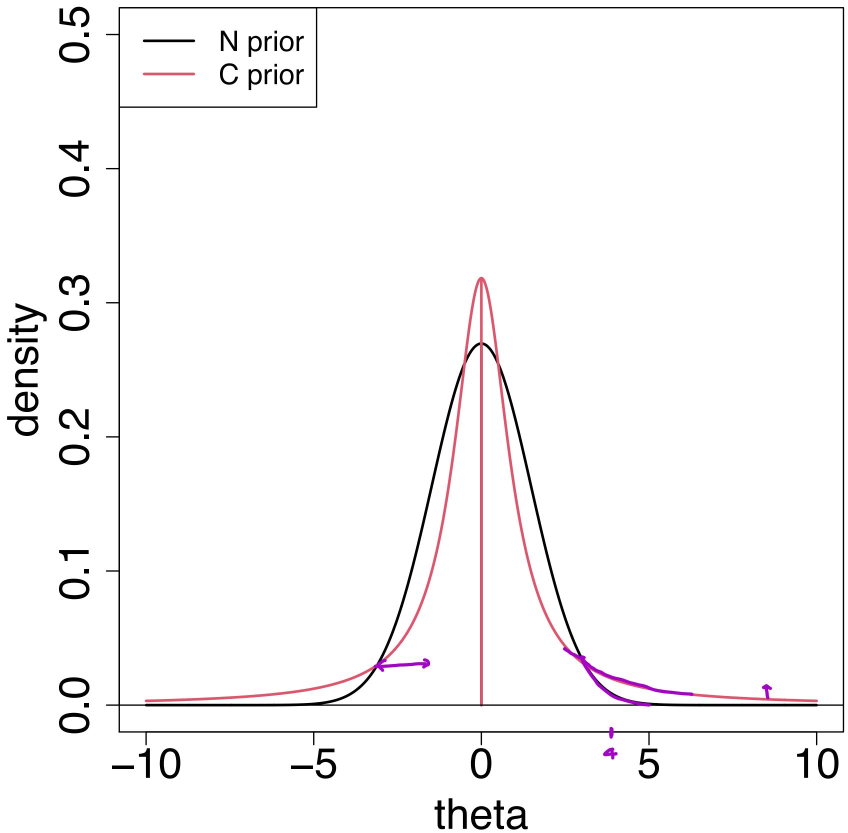
$$\Rightarrow \delta_2^{\pi}(x) \approx x - \frac{x}{1+x^2} \text{ for } |x| \geq 4$$

$$\pi(\theta) = \frac{1}{\pi \cdot (1+\theta^2)}, \quad \theta \in \mathbb{R} = \Theta$$

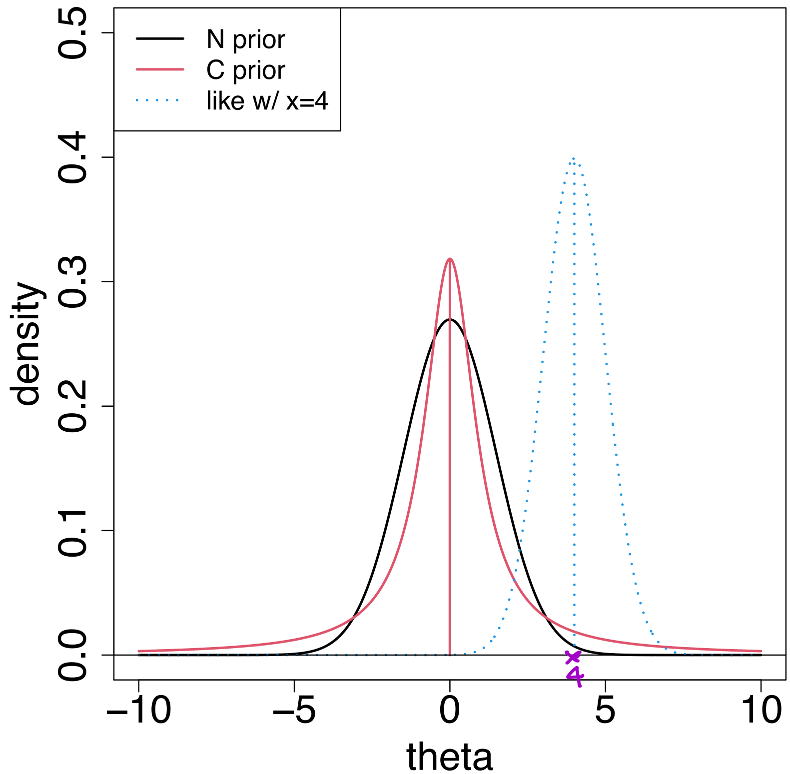
$$\pi(\theta \mid x) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} \cdot \frac{1}{\pi(1+\theta^2)}}{\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} \cdot \frac{1}{\pi(1+\theta^2)} \, d\theta}, \quad \theta \in \mathbb{R} = \Theta$$

  ⋆⋆ For $x = 4$, we have $\delta_2^{\pi}(x) = 3.76$.

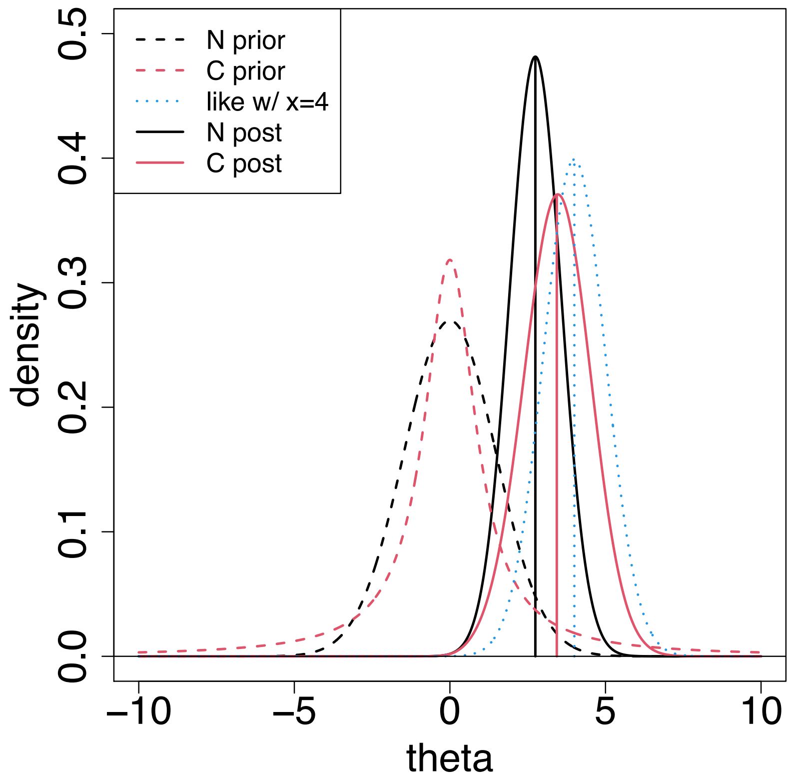
- **Example 3.2.6** (contd)



- **Example 3.2.6** (contd) Suppose $x = 4$ is observed. N(



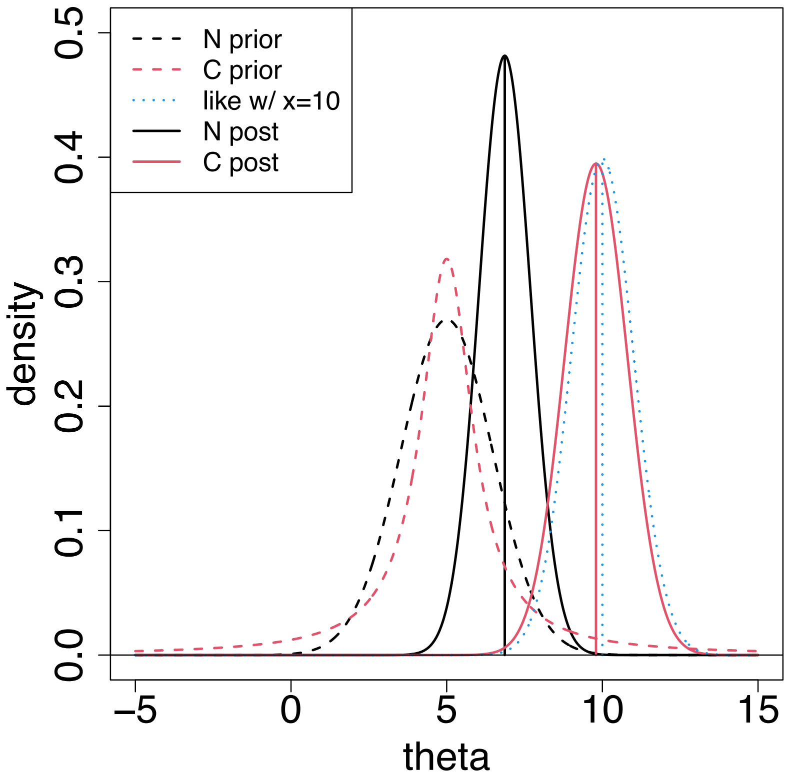
• **Example 3.2.6** (contd) Case 1: $\delta_1^\pi(x) = \underline{2.75}$ vs Case 2: $\delta_2^\pi(x) = \underline{3.76}$.



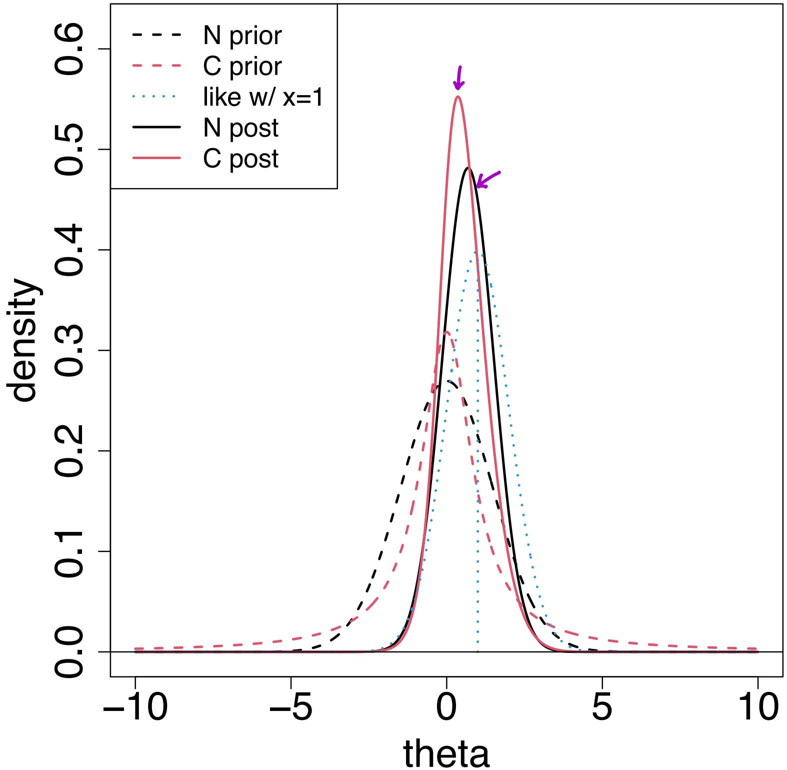
- **Example 3.2.6**(contd) If $x = 10$ is observed,

$\delta_1^\pi(x) = 6.86$

$\delta_2^\pi(x) = 9.90$



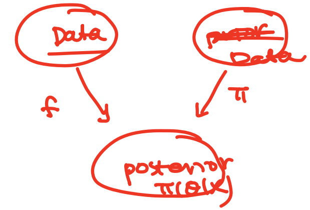
- **Example 3.2.6**(contd) If $x = 1$ is observed,



- **Example 3.2.6** (contd) Take-home message;
  - ⋆⋆ The selection of the parameterized family greatly affects the inference about θ, especially due to the tail of the chosen prior where prior information is scarce.
  - ⋆⋆ These posterior discrepancies call for some tests on the validity (or robustness) of the selected priors.

$x_i|\theta_i \sim N(\theta_i, \sigma^2)$, $i=1,\dots,p$

$\theta_i \overset{iid}{\sim} N(\mu, \tau^2)$



† *Empirical Bayes*

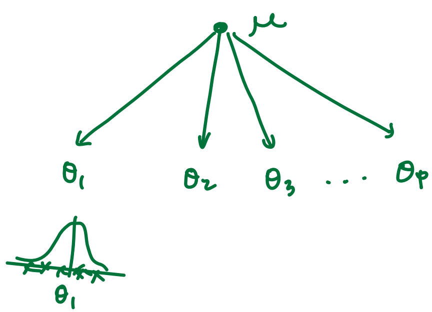
- **<u>Use data</u>** to estimate some features of the prior distribution
- Choose a *prior* distribution a posteriori! ⇒ It does not belong to the Bayesian paradigm.
- Parametric empirical Bays:

  ⋆⋆ Assume that the prior distribution of θ is in some parametric class with unknown parameters.

  ⋆⋆ Use data to specify the unknown parameters.

JB in Section 4.5.2 Assume that $X_i \mid \theta_i \overset{indep}{\sim} \mathrm{N}(\theta_i, \sigma^2)$ with known σ^2, $i = 1, \ldots, p$ and θ_i are from a common prior distribution. Specify the prior distribution for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ using data. Assume $\theta_i \overset{iid}{\sim} \mathrm{N}(\mu, \tau^2)$. The hyperparameters μ and τ^2 are unknown.

X_i is the test score of individual i, random about his/her true ability θ_i with known "reliability" σ^2. True abilities θ_i, $i = 1, \ldots, p$ are from an unknown normal population.



JB 4.5.2 (contd) How do we specify values for μ and τ^2?

- ⋆⋆ We use the data to estimate μ and τ^2.
- ⋆⋆ One way is to consider $m(\boldsymbol{x} \mid \pi)$ as a likelihood function for π as follows;
- ⋆⋆ Intuition $m(\boldsymbol{x} \mid \underline{\pi})$ is the density according to which X will actually occur.

  If X_i is a test score of individual i which was normally distributed about "true ability" θ_i, and the true ability in the population varied according to a normal distribution with mean μ and τ^2, then $m(x_i)$ would be the actual distribution of observed test scores.

  ∗∗ Recall we called $m(x \mid \pi)$ the predictive distribution for x.

$X_i \mid \theta_i \overset{\text{indep}}{\sim} N(\theta_i, \sigma^2)$, σ^2 known, $i = 1, \dots, p$

$\theta_i \overset{\text{iid}}{\sim} N(\mu, \tau^2)$

$$\underline{m(x \mid \mu, \tau^2)} = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x_i-\theta_i)^2}{2\sigma^2}\right) \cdot \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\theta_i-\mu)^2}{2\tau^2}\right) d\theta_1 \cdots d\theta_p$$

marginal distr. of x

or (prior predictive distr. of x)

$$= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} \exp\left(-\frac{(x_i-\mu)^2}{2(\sigma^2+\tau^2)}\right)$$

$x_i \mid \mu, \tau^2 \sim N(\mu, \sigma^2+\tau^2)$

$\Rightarrow$ Find the values of (μ, τ^2) that maximize

$m(x \mid \mu, \tau^2)$

$$\left.\begin{aligned} \frac{\partial \log m(x \mid \mu, \tau^2)}{\partial \mu} &= 0 \\ \frac{\partial \log m(x \mid \mu, \tau^2)}{\partial \tau^2} &= 0 \end{aligned}\right\} \longrightarrow \text{solve for } \mu, \tau^2$$

JB 4.5.2 (contd)

- ⋆⋆ Seek to maximize $m(\boldsymbol{x} \mid \pi)$ over the hyperparameters μ and τ^2 by maximum likelihood.

  Intuition If $m(x \mid \pi_1) > m(x \mid \pi_2)$, we can conclude that the data provides more support for π_1 than for π_2.

- ⋆⋆ Recall that

$$
\begin{aligned}
m(\boldsymbol{x} \mid \mu, \tau^2) &= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} \exp\left\{-\frac{(x_i-\mu)^2}{2(\sigma^2+\tau^2)}\right\} \\
&= \{2\pi(\sigma^2+\tau^2)\}^{-p/2} \exp\left\{-\frac{s^2}{2(\sigma^2+\tau^2)}\right\} \exp\left\{-\frac{p(\bar{x}-\mu)^2}{2(\sigma^2+\tau^2)}\right\},
\end{aligned}
$$

where $\bar{x} = \sum_{i=1}^{p} x_i/p$ and $s^2 = \sum_{i=1}^{p}(x_i-\bar{x})^2$.

JB 4.5.2 (contd)

⋆⋆ We find the MLEs

$$\hat{\mu} = \bar{x} \quad \text{and} \quad \hat{\tau}^2 = \max\left\{0, \frac{1}{p}s^2 - \sigma^2\right\}.$$

⋆⋆ We can pretend that the θ_i are iid from $N(\hat{\mu}, \hat{\tau}^2)$ and proceed with a Bayesian analysis.

⋆⋆ **Or** we can use the moment method by matching the first two moments, $\hat{\mu} = \bar{x}$ and $\hat{\tau}^2 = \sum_{i=1}^{p}(x_i - \bar{x})^2/(p-1) - \sigma^2$.