ORIE 4742 - Info Theory and Bayesian ML

Chapter 6: Intro to Bayesian Statistics

March 15, 2021

Sid Banerjee, ORIE, Cornell

Bayesian basics

marginals and conditionals

let X and Y be discrete rvs taking values in $\mathbb{N}$. denote the joint pmf:

$$p_{XY}(x,y) = \mathbb{P}[X = x, Y = y]$$

marginalization: computing individual pmfs from joint pmfs as

$$p_X(x) = \sum_{y\in\mathbb{N}} p_{XY}(x,y) \qquad p_Y(y) = \sum_{x\in\mathbb{N}} p_{XY}(x,y)$$

conditioning: pmf of X given $Y = y$ (with $p_Y(y) > 0$) defined as:

$$\mathbb{P}[X = x|Y = y] \triangleq p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)}$$

more generally, can define $\mathbb{P}[X \in \mathcal{A}|Y \in \mathcal{B}]$ for sets $\mathcal{A}, \mathcal{B} \in \mathbb{N}$

see also this visual demonstration

the basic ‘rules’ of Bayesian inference

let X and Y be discrete rvs taking values in $\mathbb{N}$, with joint pmf $p(x, y)$

product rule

for $x, y \in \mathbb{N}$, we have: $p_{XY}(x, y) = p_X(x)p_{Y|X}(y|x) = p_Y(y)p_{X|Y}(x|y)$

sum rule

for $x \in \mathbb{N}$, we have: $p_X(x) = \sum_{y \in \mathbb{N}} p_{X|Y}(x|y)p_Y(y)$

and most importantly!

Bayes rule

for any $x, y \in \mathbb{N}$, we have:

$$p_{X|Y}(x|y) = \frac{p_X(x)p_{Y|X}(y|x)}{\sum_{x \in \mathbb{N}} p_{Y|X}(y|x)p_X(x)}$$

see also this video for an intuitive take on Bayes rule

fundamental principle of Bayesian statistics

- assume the world arises via an underlying generative model $\mathcal{M}$
- use random variables to model all unknown parameters θ
- incorporate all that is known by conditioning on data D
- use Bayes rule to update prior beliefs into posterior beliefs

$$p(\theta|D, \mathcal{M}) \propto p(\theta|\mathcal{M})p(D|\theta, \mathcal{M})$$

pros and cons

in praise of Bayes

- conceptually simple and easy to interpret
- works well with small sample sizes and overparametrized models
- can handle all questions of interest: no need for different estimators, hypothesis testing, etc.

why isn't everybody Bayesian

- they need priors (subjectivity...)
- they may be more computationally expensive: computing normalization constant and expectations, and updating priors, may be difficult

basics of Bayesian inference

the likelihood principle

given model $\mathcal{M}$ with parameters Θ, and data D, we define:

- the prior $p(\Theta|\mathcal{M})$: what you believe before you see data
- the posterior $p(\Theta|D, \mathcal{M})$: what you believe after you see data
- the marginal likelihood or evidence $p(D|\mathcal{M})$: how probable is the data under our prior and model

  these three are probability distributions; the next is not

- the likelihood: $\mathcal{L}(\Theta) \triangleq p(D|\mathcal{M}, \Theta)$: function of Θ summarizing data

the likelihood principle

given model $\mathcal{M}$, all evidence in data D relevant to parameters Θ is contained in the likelihood function $\mathcal{L}(\Theta)$

this is not without controversy; see Wikipedia article

REMEMBER THIS!!

given model $\mathcal{M}$ with parameters Θ, and data D, we define:

- the prior $p(\Theta|\mathcal{M})$: what you believe before you see data
- the posterior $p(\Theta|D, \mathcal{M})$: what you believe after you see data
- the marginal likelihood or evidence $p(D|\mathcal{M})$: how probable is the data under our prior and model
- the likelihood: $\mathcal{L}(\Theta) \triangleq p(D|\mathcal{M}, \Theta)$: function of Θ summarizing the data

the fundamental formula of Bayesian statistics

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

also see: Sir David Spiegelhalter on Bayes vs. Fisher

example: the mystery Bernoulli rv

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution

fix θ; what is $\mathbb{P}[D|\mathcal{M}]$ for any $i \in [n]$?

let $H = \#$ of '1's in $\{X_1, X_2, \ldots, X_n\}$; what is $\mathbb{P}[H|\mathcal{M}, \theta]$?

the Bernoulli likelihood function

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution

likelihood: $\mathcal{L}(\Theta) \triangleq p(D|\mathcal{M}, \theta)$: function of Θ summarizing the data

log-likelihood, sufficient statistics, MLE

cromwell's rule

how should we choose the prior?

the zeroth rule of Bayesian statistics

never set $p(\theta|\mathcal{M}) = 0$ or $p(\theta|\mathcal{M}) = 1$ for any θ

also see:

– Jacob Bronowski on Cromwell's Rule and the scientific method

– Richard Feynman on the scientific method (at Cornell!)

from where do we get a prior?

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution

option 1: from the 'problem statement'

Mackay example 2.6

- eleven urns labeled by $u \in \{0, 1, 2, \ldots, 10\}$, each containing ten balls
- urn u contains u red balls and $10 - u$ blue balls
- select urn u uniformly at random and draw n balls with replacement, obtaining n_R red and $n - n_R$ blue balls

from where do we get a prior

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0, 1\}^n$
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution

option 2: the maximum entropy principle

choose $p(\theta|\mathcal{M})$ to be distribution with maximum entropy given $\mathcal{M}$

we know $\theta \in [0, 1]$

from where do we get the prior, take 2

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution

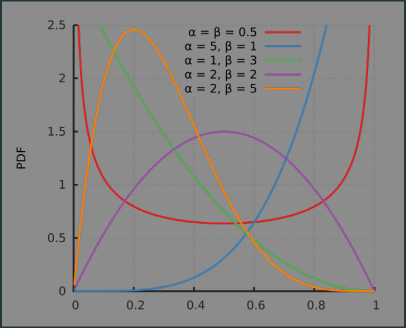
option 3: easy updates via conjugate priors

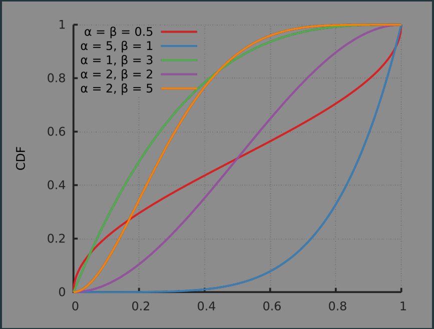
- prior $p(\theta)$ is said to be conjugate to likelihood $p(D|\theta)$ if corresponding posterior $p(\theta|D)$ has same functional form as $p(\theta)$
- natural conjugate prior: $p(\theta)$ has same functional form as $p(D|\theta)$
- conjugate prior family: closed under Bayesian updating

the Beta distribution

Beta distribution

- $x \in [0,1]$, parameters: $\Theta = (\alpha, \beta) \in \mathbb{R}^+$ ('# ones'+1,'# zeros'+1)
- pdf: $p(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$
- normalizing constant: $\frac{1}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$





Beta-Bernoulli prior and updates

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution

Beta-Bernoulli model

- prior parameters: $\Theta_0 = (\alpha, \beta) \in \mathbb{R}^+$ (hyperparameters)
- Beta-Bernoulli prior: $Beta(\alpha, \beta) \sim p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$
- likelihood: $p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$

  then via Bayesian update we get

- posterior:

$$p(\theta|D) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{N_1}(1-\theta)^{N_0} \sim Beta(\alpha + N_1, \beta + N_0)$$

the Beta distribution: getting familiar

$Beta(\alpha, \beta)$ distribution

$$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

properties of $\Gamma(\alpha)$

the Beta distribution: mean and mode

$Beta(\alpha, \beta)$ distribution

$$p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha - 1}(1 - x)^{\beta - 1}$$

Beta-Bernoulli model: what should we report?

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution
- prior: $p(\theta) \sim Beta(\alpha, \beta)$ posterior: $p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_0)$

decision theory

decision theory in a nutshell

Bayesian decision theory in learning

given prior F on θ, choose 'action' $\hat{\theta}$ to minimize loss function $\mathbb{E}_F[L(\theta, \hat{\theta})]$

decision theory in a nutshell

Bayesian decision theory in learning

given prior F on θ, choose 'action' $\hat{\theta}$ to minimize loss function $\mathbb{E}_F[L(\theta, \hat{\theta})]$

examples

- L_0 loss: $L(\theta, \hat{\theta}) = \mathbb{1}_{\{\theta \neq \hat{\theta}\}} \Rightarrow \hat{\theta_{L_0}} =$ mode of F
- L_1 loss: $L(\theta, \hat{\theta}) = ||\theta - \hat{\theta}||_1 \Rightarrow \hat{\theta_{L_1}} =$ median of θ under F
- L_2 loss: $L(\theta, \hat{\theta}) = ||\theta - \hat{\theta}||_2 \Rightarrow \hat{\theta_{L_2}} = \mathbb{E}_F[\theta]$

decision theory in a nutshell

Bayesian decision theory in learning

given prior F on θ, choose 'action' $\hat{\theta}$ to minimize loss function $\mathbb{E}_F[L(\theta, \hat{\theta})]$

examples

- L_0 loss: $L(\theta, \hat{\theta}) = \mathbb{1}_{\{\theta \neq \hat{\theta}\}} \Rightarrow \hat{\theta_{L_0}} =$ mode of F
- L_1 loss: $L(\theta, \hat{\theta}) = ||\theta - \hat{\theta}||_1 \Rightarrow \hat{\theta_{L_1}} =$ median of θ under F
- L_2 loss: $L(\theta, \hat{\theta}) = ||\theta - \hat{\theta}||_2 \Rightarrow \hat{\theta_{L_2}} = \mathbb{E}_F[\theta]$

in general 'decision-making'

given prior F on X, choose 'action' $a \in \mathcal{A}$ to minimize loss, i.e.

$$a^* = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{X \sim F}[L(a, X)]$$

Beta-Bernoulli model: posterior mean

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution
- prior: $p(\theta) \sim Beta(\alpha, \beta)$ posterior: $p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_0)$

posterior mean: $\mathbb{E}[\theta|\alpha, \beta, N_0, N_1] =$

Beta-Bernoulli model: posterior mode (MAP estimation)

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution
- prior: $p(\theta) \sim Beta(\alpha, \beta)$ posterior: $p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_0)$

posterior mode: $\max_{\theta \in [0,1]} p(\theta|\alpha, \beta, N_0, N_1) =$

Beta-Bernoulli model: posterior prediction (marginalization)

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution
- prior: $p(\theta) \sim Beta(\alpha, \beta)$ posterior: $p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_0)$

posterior prediction: $\mathbb{P}[X = 1|D] =$

the black swan

summarizing the posterior

model $\mathcal{M}+$ prior $p(\Theta)+$ data $D \Rightarrow$ posterior $p(\Theta|D)$

summarizing $p(\Theta|D)$

- posterior mean $\widehat{\theta}_{mean} = \mathbb{E}[\Theta|D]$
- posterior mode (or MAP estimate) $\widehat{\theta}_{MAP} = \arg\max_{\Theta} p(\Theta|D)$
- posterior median $\widehat{\theta}_{median} = \min\{\Theta : p(\Theta|D) \geq 0.5\}$
- Bayesian credible intervals: given $\delta > 0$, want (ℓ_Θ, u_Θ) s.t.

$$\mathbb{P}[\ell_\Theta \leq \Theta \leq u_\Theta | D] > 1 - \delta$$

summarizing the posterior

Bayesian credible intervals

given posterior $p(\Theta|D)$ and any $\delta > 0$, want (ℓ_Θ, u_Θ) s.t.

$$\mathbb{P}[\ell_\Theta \leq \Theta \leq u_\Theta | D] > 1 - \delta$$

marginal likelihood (evidence)

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution

marginal likelihood

$$p(D) = \frac{p(\theta)p(D|\theta)}{p(\theta|D)} = \frac{\text{prior} \times \text{likelihood}}{\text{posterior}}$$

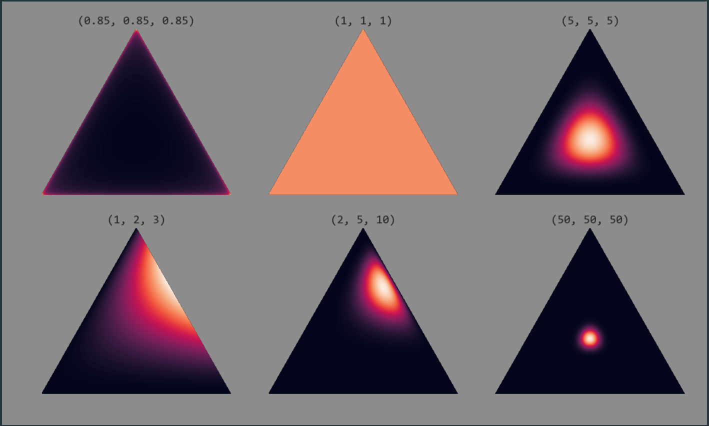
multiclass data

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{1, 2, \ldots, K\}^n$
- for $i \in [K]$, data D contains N_i copies of type i
- model $\mathcal{M}$: X_i generated i.i.d. from $\mathit{Multinomial}(\theta_1, \theta_2, \ldots, \theta_K)$ distn

the Dirichlet distribution

Dirichlet distribution

- $x \in \{x_i \in [0,1], \sum_{i=1}^{K} x_i = 1\}$, parameters: $\Theta = (\alpha_1, \alpha_2, \ldots, \alpha_K)$
- pdf: $p(x) \propto \prod_{i=1}^{K} x_i^{\alpha_i - 1}$



the Dirichlet distribution

Dirichlet distribution

- $x \in \{x_i \in [0,1], \sum_{i=1}^{K} x_i = 1\}$, parameters: $\Theta = (\alpha_1, \alpha_2, \ldots, \alpha_K)$
- denote $\alpha = \{\alpha_i\}_{i=1}^{K})$; Dirichlet pdf

$$p(x) = \frac{1}{Z(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

- normalizing constant: $\frac{1}{Z(\alpha)} = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)}$

multiclass data and Dirichlet priors

- for $i \in [K]$, data D contains N_i copies of type i
- model $\mathcal{M}$: X_i generated i.i.d. from $Multinomial(\theta_1, \theta_2, \ldots, \theta_K)$ distn

Dirichlet-Multinomial model

- prior parameters: $\Theta_0 = (\alpha_1, \alpha_2, \ldots, \alpha_K) \in \mathbb{R}_+^K$ (hyperparameters)
- Dirichlet prior: $Dir(\alpha_1, \alpha_2, \ldots, \alpha_K) \sim p(\theta) \propto \prod_{i=1}^K \theta_i^{\alpha_i - 1}$
- likelihood: $p(D|\theta) = \prod_{i=1}^K \theta_i^{N_i}$
- posterior: $p(\theta|D) \sim Dir(\alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_K + N_K)$
- marginal likelihood: let $m = \sum_{i=1}^K \alpha_i$

$$p(D) = \frac{\Gamma(m)}{\Gamma(n+m)} \prod_{i=1}^{K} \frac{\Gamma(N_i + \alpha_i)}{\Gamma(\alpha_i)}$$

generative models for continuous data

continuous data and Gaussian priors

- data $D = \{X_1, X_2, \ldots, X_n\} \in \mathbb{R}^n$
- model $\mathcal{M}$: X_i generated i.i.d. from $\mathcal{N}(\mu, \sigma^2)$ distribution

Gaussian prior

- $x \in \mathbb{R}$, parameters: $\Theta = (\mu, \sigma)$
- pdf: $\mathbb{P}[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] \propto \frac{1}{(\sigma^2)^{n/2}} \exp\left(\frac{-\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right)$
- normalizing constant: $(2\pi)^{-n/2}$

3 options:

1. μ unknown, σ^2 known
2. σ^2 unknown, μ known
3. μ unknown, σ^2 unknown

notation: define precision $\tau = \frac{1}{\sigma^2}$

case 1: unknown μ

- data $D = \{X_1, X_2, \dots, X_n\} \in \mathbb{R}^n$
- model $\mathcal{M}$: X_i i.i.d. from $\mathcal{N}(\mu, 1/\tau)$, with unknown μ, known $\tau = 1/\sigma^2$

normal-normal model

- likelihood:

$$p(D|\mu) \propto \tau^{n/2} \exp\left(-\tau \sum_{i=1}^{n} (x_i - \mu)^2/2\right)$$

- prior parameter: $\Theta_0 = (m_\mu, 1/\tau_\mu)$ (mean, precision for μ)
- Gaussian prior for μ: $\mu \sim \mathcal{N}(m_\mu, \tau_\mu)$, where $\tau_\mu = 1/Var(\mu)$

$$p(\mu|m_\mu, \tau_\mu) \propto \tau_\mu^{1/2} \exp\left(-\tau_\mu(\mu - m_\mu)^2/2\right)$$

normal-normal model: posterior

normal-normal model: posterior

normal-normal model: posterior predictive distribution

normal-normal model: posterior predictive distribution

- data $D = \{X_1, X_2, \dots, X_n\} \in \mathbb{R}^n$
- model $\mathcal{M}$: X_i i.i.d. from $\mathcal{N}(\mu, \tau)$, with unknown μ, known $\tau = 1/\sigma^2$
- thus we have

$$X_i = \mu + \sigma Z_1$$
$$\mu = m_\mu + \sigma_\mu Z_2$$

normal-normal model for unknown μ

- data $D = \{X_1, X_2, \ldots, X_n\} \in \mathbb{R}^n$
- model $\mathcal{M}$: X_i i.i.d. from $\mathcal{N}(\mu, \tau)$, with unknown μ, known $\tau = 1/\sigma^2$

normal-normal model

- likelihood: $p(D|\mu) \propto \exp\left(-\tau \sum_{i=1}^{n}(x_i - \mu)^2/2\right)$
- prior: $\mu \sim \mathcal{N}(M_\mu, 1/\tau_\mu) \propto \exp\left(-\tau_\mu(\mu - m_\mu)^2/2\right)$
- posterior: let $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $m_D = \frac{n\tau \cdot \overline{x} + \tau_\mu \cdot m_\mu}{n\tau + \tau_\mu}$ and $\tau_D = n\tau + \tau_\mu$

$$p(\mu|D) \sim \mathcal{N}\left(m_D, 1/\tau_D\right)$$

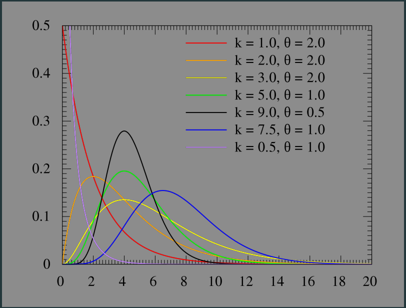
- posterior predictive distribution:

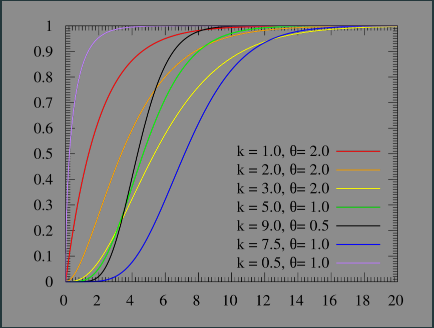
$$p(x|D) \sim \mathcal{N}\left(m_D, 1/\tau + 1/\tau_D\right)$$

the gamma distribution

gamma distribution

- $x \in (0, \infty)$, parameters: $\Theta = (\alpha, \beta) \in \mathbb{R}^+$ ('shape,rate')
- pdf of $Gamma(\alpha, \beta)$: $p(x) \propto x^{\alpha-1} e^{-\beta x}$
- normalizing constant: $\frac{1}{Z(\alpha,\beta)} = \frac{\beta^\alpha}{\Gamma(\alpha)}$





case 2: unknown σ

- data $D = \{X_1, X_2, \dots, X_n\} \in \mathbb{R}^n$
- model $\mathcal{M}$: X_i i.i.d. from $\mathcal{N}(\mu, 1/\tau)$, with unknown $\tau = 1/\sigma$, known μ

normal-gamma model

- likelihood:

$$p(D|\theta) \propto \tau^{n/2} \exp\left(-\tau \sum_{i=1}^{n} (x_i - \mu)^2/2\right)$$

- prior parameters: $\Theta_0 = (\alpha, \beta)$
- gamma prior for τ: $\tau \sim Gamma(\alpha, \beta)$

$$p(\tau|\alpha, \beta) \propto \tau^{\alpha-1} e^{-\beta\tau}$$

normal-gamma model: posterior

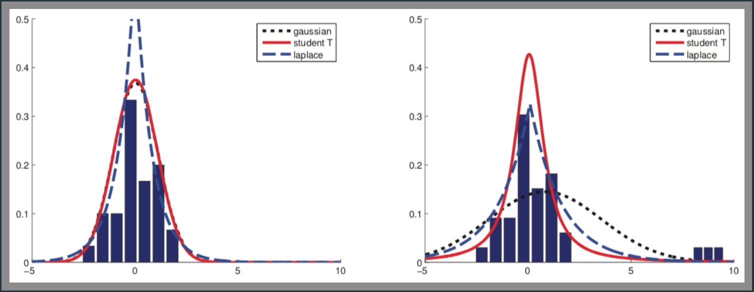
normal-gamma model: posterior predictive distribution

normal-gamma model: posterior predictive distribution

the Student-t distribution

Student-t distribution

- $x \in \mathbb{R}$, parameter: $\mu \in \mathbb{R}, \nu > 0$ (mean,'degrees of freedom')
- pdf of student-t(μ, ν): $p(x) \propto \left(1 + \frac{(x-\mu)^2}{\nu}\right)^{\frac{\nu+1}{2}}$
- normalizing constant: $\frac{1}{Z(\mu,\nu)} = \frac{\Gamma(\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}$



robustness of student-t to outliers

normal-gamma model for unknown τ

- data $D = \{X_1, X_2, \dots, X_n\} \in \mathbb{R}^n$
- model $\mathcal{M}$: X_i i.i.d. from $\mathcal{N}(\mu, \sigma^2)$, with unknown $\tau = 1/\sigma^2$, known μ

normal-gamma model

- likelihood: $p(D|\theta) \propto \exp\left(-\tau \sum_{i=1}^{n}(x_i - \mu)^2/2\right)$
- prior for τ: $\tau \sim gamma(\alpha, \beta)$
- posterior: let $\alpha_D = \alpha + \frac{n}{2}$ and $\beta_D = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$

$$p(\tau|D) \sim gamma\left(\alpha_D, \beta_D\right)$$

- posterior predictive distribution:

$$p(x|D) \sim \text{student-t}$$

case 3: unknown μ and σ^2

case 3: unknown μ and σ^2

case 3: unknown μ and σ^2

case 3: unknown μ and σ^2