

Notes
on
Bayesian Inference

Albert Y. Lo
Department of Information and Systems Management
The University of Science & Technology
Clear Water Bay, Hong Kong

ISOM 359A, Spring 2009

Topics

- 1 Model distributions
- 2 Large-sample behaviors of a posterior distribution
- 3 IID Monte Carlo approximations
- 4 Markov chain Monte Carlo approximations
- 5 Model based clustering
- 6 Bayesian nonparametrics: Gamma typed process priors
- 7 Mixture models

1 Prior and posterior analysis

1.1 Model distributions.

X - observation X ; θ - an unknown parameter.

X and θ are related by a conditional distribution of X given θ , called a model distribution. Specifically, for each given θ ,

$$F(x|\theta) \equiv P\{X \leq x | \theta\}, \text{ } x \text{ varies on the real line.}$$

For each θ , $F(x|\theta)$ increases from 0 to 1 as x increases from $-\infty$ to ∞ .

$F(x|\theta)$ is called the model distribution of X given θ .

For each given θ , the derivative $f(x|\theta) \equiv (d/dx)F(x|\theta)$, if exists, is nonnegative, and for each x ,

$$F(x|\theta) = \int_{\{z \leq x\}} f(z|\theta) dz$$

Model checking: Does the model $F(x|\theta)$ explain the relationship between X and θ well? Topics discussed by other courses.

How to uncover information about the parameter θ based on an observed value of X given θ ?

Example 1.1

(i) The model distribution of $X|\theta$ is given by

$$\Pr\{X=\theta|\theta\}=1 \text{ for each } \theta \in (-\infty, \infty).$$

The observed X is 0.1. According to the model distribution, the only case that an observed X is 0.1 is that θ is also 0.1 - any other value of θ would not yield an observed X value of 0.1. Hence, one concludes that θ must be 0.1.

(ii) We know that $\theta=0$ or 5 in the following two Normal models.

(a) $X|\theta$ is $N(\theta, 0.04)$. If the observed X is close to 0, say within 0.8

of zero, θ is very likely to be 0; if the observed X is within 0.8 of 5, θ is very likely to 5. We are less-certain about the value of θ if the observed X is between 2 and 4.

(b) $X|\theta$ is $N(\theta, 9)$. Not too sure about whether θ is 0 or 5 if the observed X is between -1 and 6.

Example 1.2. Some standard model distributions.

- (i) Given $\theta \in (0,1)$, $X|\theta$ has a Bernoulli (θ) distribution.
- (ii) Given $\theta \in (0,1)$, $X|\theta$ has a Binomial ($25,\theta$) distribution.
- (iii) Given $\theta=(\alpha,\beta) \in (0, \infty)\times(0, \infty)$, $X|\theta$ has a Beta (α,β) distribution.
- (iv) Given $\theta \in (0, \infty)$, $X|\theta$ has a Poisson (θ) distribution.
- (v) Given $\theta \in (0, \infty)$, $X|\theta$ has an exponential (θ) distribution.
- (vi) Given $\theta=(\alpha,\beta) \in (0, \infty)\times(0, \infty)$, $X|\theta$ has a Gamma ($\alpha;1/\beta$) distribution (with mean α/β).
- (vii) Given $\theta \in (0, \infty)$, $X|\theta$ has a Uniform($0,\theta$) distribution.
- (viii) Given $\theta=(\mu,\sigma) \in (-\infty,\infty)\times(0, \infty)$, $X|\theta$ has a $N(\mu,\sigma^2)$ distribution.
- (ix) Given $\theta=(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in (\mathbb{R}^k, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}$ is a column vector with k entries and $\boldsymbol{\Sigma}$ is a k by k covariance matrix (and denoted by $\boldsymbol{\tau} \equiv \boldsymbol{\Sigma}^{-1}$, which exists,)

$\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}$ has a $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution if

\mathbf{X} is a random column vector with k entries, and its density is

$$f(x|\boldsymbol{\mu}, \boldsymbol{\tau}) \propto |\boldsymbol{\tau}|^{1/2} \times \exp\{-(1/2) \sum_{1 \leq i \leq n} (\mathbf{x}_i - \boldsymbol{\mu})^t \boldsymbol{\tau} (\mathbf{x}_i - \boldsymbol{\mu})\}.$$

$\boldsymbol{\tau}=(\sigma^2)^{-1}$ is called the precision of a $N(\mu,\sigma^2)$ distribution. Likewise,

$\boldsymbol{\tau}=\boldsymbol{\Sigma}^{-1}$ is called the precision matrix of a $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution.

The inverse of a symmetric and positive definite k by k matrix exists; see Section 5.4, De Groot.

Notes: Compare this with a Gamma density model in (6).

- (x) Let \mathbb{S} be the set of symmetric and positive definite k by k

matrices - the set of precision matrices. Given $\theta=(\alpha,\psi)$, $\alpha \leq k$ and $\psi \in \mathbb{R}^+$, $\tau|\theta$ has a Wishart distribution with degrees of freedom α ($\alpha \leq k$) and precision parameter matrix $\psi \in \mathbb{R}^+$, i.e. $\tau|\alpha,\psi$ is $W(\alpha,\psi)$, if its density is proportional to (Section 5.5, De Groot.)

$$\tau|\alpha,\psi \propto |\tau|^{(\alpha-k)/2} \times \exp\{-(1/2)\text{trace}(\psi\tau)\}.$$

The trace of a k by k matrix is the sum of its diagonal entries.

Note: Compare this with a Gamma density model in (6).

Model distribution for observations/data. The unknown parameter is denoted by θ . Data/observations are denoted by x_1, \dots, x_n , which is the observed initial segment of a sequence of random variables $X_1, \dots, X_n, X_{n+1}, \dots$

Sampling distribution/Model distribution for observations $X_1, \dots, X_n|\theta$ is the conditional distribution of X_1, \dots, X_n given θ , defined by, for all A_i 's of the half-interval form $(-\infty, x_i]$,

$$(1.1) \quad \begin{aligned} F(x_1, \dots, x_n|\theta) &\equiv \Pr\{X_1 \in A_1, \dots, X_n \in A_n|\theta\} \\ &= \Pr\{X_1 < x_1, \dots, X_n < x_n|\theta\}. \end{aligned}$$

For statistical inference, the model distribution $F(x_1, \dots, x_n|\theta)$ relates the data to the unknown, and is a "must-have." Note that (1.1) is just $F(x|\theta)$ above in which x is a vector $\mathbf{x}=(x_1, \dots, x_n)$, and they share similar properties: For each θ ,

$$\begin{aligned} F(x_1, \dots, x_n|\theta) &\text{ increases in each coordinate,} \\ F(-\infty, \dots, -\infty|\theta) &= 0, \text{ and } F(\infty, \dots, \infty|\theta) = 1. \end{aligned}$$

If in addition the joint densities (given θ) exist, successive derivatives of $F(x_1, \dots, x_n|\theta)$ with respect to the x_i 's, yields a product rule for densities:

$$(1.2) \quad f(\mathbf{x}|\theta) \equiv f(x_1, x_2, \dots, x_n|\theta)$$

$$= f(x_1|\theta) \times f(x_2|\theta, x_1) \times f(x_3|\theta, x_1, x_2) \times \dots \times f(x_n|\theta, x_1, x_2, \dots, x_{n-1})$$

Example 1.3. Some sampling plans with model joint densities.

(i) Sampling with replacements model. $X_1, \dots, X_n|\theta$ are iid with a distribution $F(x|\theta)$ that has a density $f(x|\theta)$, the joint density of $X_1, \dots, X_n|\theta$ is given by

$$f(\mathbf{x}|\theta) = \prod_{1 \leq i \leq n} f(x_i|\theta).$$

(ii) Linear regression: $\theta = (\beta_0, \beta_1)$ and $Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$, $i=1, 2, \dots, n$ are observable, and $\varepsilon_i|s|\theta$ are iid with a unimodal (error) density $f(z|\theta)$. Here given θ , the observable Y_i s are not iid, their linear transforms $\varepsilon_i = Y_i - \beta_0 - \beta_1 x_{i1}$ are

$$f(\mathbf{y}|\theta) = \prod_{1 \leq i \leq n} f(y_i - \beta_0 - \beta_1 x_{i1}|\theta).$$

Can you simplify $f(\mathbf{y}|\theta)$ if $\varepsilon_i|s|\theta$ are iid $N(0, \sigma^2)$?

What is $f(\mathbf{y}|\theta)$ if $(\varepsilon_1, \dots, \varepsilon_n)$ is $N(\mathbf{0}, \Sigma)$ where Σ is a diagonal matrix with the i -th entry being σ_i^2 ?

(iii) An AR(1) time series: Given θ , $Y_i = \theta Y_{i-1} + \varepsilon_i$, $i=1, 2, \dots$, $|\theta| \leq 1$, $Y_0 = 1$, and X_i are iid $N(0, \sigma^2)$. Since $\varepsilon_i = Y_i - \theta Y_{i-1}$, $i=1, 2, \dots$ (given θ) are iid,

$$f(\mathbf{y}|\theta) = \prod_{1 \leq i \leq n} f(y_i - \theta y_{i-1}|\theta) = ?$$

How about if the X_i s are iid standard Cauchy?

(iv) $X_1, \dots, X_n|\theta$ is a Markov chain sequence with stationary distribution $Q(x|\theta)$ and transition probability $K_\theta(y|x)$: for all $n=1, 2, \dots$, and all reals a and b

$$P\{X_n \leq a \text{ and } X_{n+1} \leq b\} = \iint_{\{x \leq a \text{ and } z \leq b\}} K_\theta(dz|x) Q(dx|\theta),$$

where $Q(x|\theta)$ is called the initial distribution of X_1 given θ ,
 and $K_\theta(z|x)$ is the conditional distribution of X_2 given $(\theta$ and
 $X_1=x)$

If in addition, the corresponding densities $q(.|\theta)$ for $Q(.|\theta)$ and $k_\theta(.|x)$ for
 $K_\theta(.|x)$ exist,

$$f(\mathbf{x}|\theta) = q(x_1|\theta) \times k_\theta(x_2|x_1) \times k_\theta(x_3|x_2) \times \dots \times k_\theta(x_n|x_{n-1}). \quad (\text{Why?})$$

(v) Sample survey: Given a finite population of N objects each numbered
 either "yes" or "no". The percentage of "yes" is $\theta \in (0,1)$;

$N\theta \in \{0,1,\dots,N\}$. Sample without replacement n ($n < N$) times from this
 population and record X_1, \dots, X_n , where $X_i=1$ if "yes" appears at the i th
 draw; otherwise $X_i=0$, $i=1, \dots, n$. X_i s are discrete random variables. Let S
 denote the sum $\sum_{1 \leq i \leq n} X_i$. The joint density of $X_1, \dots, X_n | \theta$ is

$$f(\mathbf{x}|\theta) = C(N\theta, S) \times C(N(1-\theta), n-S) / C(N, n),$$

$$0 \leq S \leq N\theta \text{ and } 0 \leq n-S \leq N(1-\theta).$$

(vi) Right censored data. The experiment is an life testing on or a group
 of n patients (or of n IC chips) and X_i is the life time of the i -th patient.
 The model assumption is $T_1, \dots, T_n | \theta$ are iid $F(t|\theta)$ that has a density $f(t|\theta)$.
 However, the only k of the data are completely observed; the rest of the
 observation one only knows that the life time T_j exceeds a fixed threshold
 c_j , i.e., the i th patient withdrew from the experiment at time c_j . To derive
 the model density of the "observable," we rearrange the data so that the
 first k ($k < n$) life times are completely observed, while the rest of the life
 times are righted censored. Let

$$Y_j \equiv \min\{T_j, c_j\}, \quad \delta_j \equiv I\{T_j \leq c_j\}, \quad j=k+1, \dots, n.$$

The δ_j s capture information provided by the censored data: $\delta_j = 0$ means

that the j -th observation is right-censored at c_j . The observed data are n pairs (Y_j, δ_j) , $j=1, \dots, n$. First, given θ , the joint density of (Y_j, δ_j) , $j=1, \dots, k | \theta$ is given by $\prod_{1 \leq i \leq k} f(y_i | \theta)$. Next, given $(\theta, Y_1, \dots, Y_k)$, the joint density of (Y_j, δ_j) , $j=k+1, \dots, n$ is

$$\prod_{k+1 \leq j \leq n} \{ [1-F(c_j | \theta)]^{1-\delta_j} \times F(c_j | \theta)^{\delta_j} \}$$

Hence, the joint "density" of (Y_j, δ_j) , $j=1, \dots, n$ is

$$\prod_{1 \leq i \leq k} f(y_i | \theta) \times \prod_{k+1 \leq j \leq n} \{ [1-F(c_j | \theta)]^{1-\delta_j} \times F(c_j | \theta)^{\delta_j} \}.$$

For right censored data $X_j > c_j$, i.e. $\delta_j = 0$, $j=k+1, \dots, n$; hence the joint density $|\theta$ is

$$\prod_{1 \leq i \leq k} f(y_i | \theta) \times \prod_{k+1 \leq j \leq n} [1-F(c_j | \theta)].$$

Exercise. What is the model density of the observable if k observations are complete data, n_1 observations are right censored, and the remaining n_2 observations are left censored (i.e., one only knows $Y_j \leq c_j$.)

Exercise. What is the model density of the observable if k observations are complete data, n_1 observations are right censored, and the remaining n_2 observations are interval-censored (one only knows that $a_i < Y_i \leq b_i$.)

(vii) Random partition model.

(viii) A graphical structure model.

1.2. Prior and posterior distributions.

Bayesian statistics assumes a prior distribution $\pi(\theta)d\theta$ on the parameter space $\Theta = \{\theta\}$. One feature of Bayesian statistics is that a Bayesian can do statistical inference using his/her own prior distribution. Since the conditional density of $\mathbf{x} | \theta$ is known, the product rule for densities gives

$$(1.3) \quad \pi(\theta) \times f(\mathbf{x} | \theta) = \text{the joint density of } (\mathbf{x}, \theta) = q(\mathbf{x}) \times \pi(\theta | \mathbf{x}),$$

where $q(\mathbf{x})$ is the (marginal) density of \mathbf{x} , and $\pi(\mathbf{x} | \theta)$, is a conditional

density of θ given $\mathbf{x}|\theta$. As a consequence, this product rule leads to the so-called Bayes theorem: A conditional distribution of $\theta|\mathbf{x}$ given by

$$(1.4) \quad \pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta) \times \pi(\theta),$$

$$\text{i.e.,} \quad \pi\{\theta \in A|\mathbf{x}\} = \int_A f(\mathbf{x}|\theta) \times \pi(\theta) d\theta / [\int f(\mathbf{x}|\theta) \times \pi(\theta) d\theta]$$

Note. If prior probability of an event is zero, $\pi(\theta \in A) = 0$, the posterior probability of that event remains to be zero: $\pi\{\theta \in A|\mathbf{x}\} = 0$.

Remark 1.1. Suppose another statistician chooses a different prior, say $\pi_a(\theta)d\theta$, and updates his/her prior to a posterior denoted by $\pi_a(\theta|\mathbf{x})d\theta$. The two different posterior densities are related by

$$\pi_a(\theta|\mathbf{x}) \propto \pi(\theta|\mathbf{x}) \times \pi_a(\theta)/\pi(\theta), \quad (\text{Check this.})$$

$$\text{i.e.,} \quad \pi\{\theta \in A|\mathbf{x}\} \propto \int_A [\pi_a(\theta)/\pi(\theta)] \times \pi(\theta|\mathbf{x}) d\theta \text{ for each event } A \subseteq \Theta.$$

Suppose an additional variable Z is observed. Conditional on $Z=z$, Bayes theorem (with obvious notation) yields

$$\begin{aligned} \pi(\theta|z) \times f(\mathbf{x}|z, \theta) &= \text{the joint density of } (\mathbf{x}, \theta)|z \\ &= q(\mathbf{x}|z) \times \pi(\mathbf{x}|z, \theta), \end{aligned}$$

and (a Bayes theorem conditional on $Z=z$)

$$\pi(\theta|z, \mathbf{x}) \propto f(\mathbf{x}|z, \theta) \times \pi(\theta|z).$$

Note. This is but a product rule conditional on on $Z=z$.

Conjugate prior family of priors for sampling a model density $f(\mathbf{x}|\theta)$: A family of distributions on Θ that is closed under sampling from $f(\mathbf{x}|\theta)$ is called a conjugate family of priors for the model $f(\mathbf{x}|\theta)$, $\theta \in \Theta$.

Example 1.4. X_1, \dots, X_n are iid from the density $f(\mathbf{x}|\theta)$, $\theta \in \Theta$.

For (i) to (vi), let $S = \sum_{1 \leq i \leq n} x_i$ and $x(n) = \max \{x_i, 1 \leq i \leq n\}$.

	θ	$f(\mathbf{x} \theta) \propto$	$\pi(\theta) \propto$	$\pi(\theta \mathbf{x}) \propto$
(i)	Binomial	$\theta^S(1-\theta)^{n-S}$	$\theta^{\alpha-1}(1-\theta)^{\beta-1}$	$\theta^{\alpha+S-1}(1-\theta)^{\beta+n-S-1}$
(ii)	Poisson.	$\theta^S e^{-n\theta}$	$\theta^{\alpha-1} e^{-\beta\theta}$	$\theta^{\alpha+S-1} e^{-(\beta+n)\theta}$
(iii)	Exponential	$\theta^n e^{-S\theta}$	$\theta^{\alpha-1} e^{-\beta\theta}$	$\theta^{\alpha+n-1} e^{-(\beta+S)\theta}$
(iv)	Uniform..	$\theta^{-n} \mathbf{I}_{\{0 < x(n) < \theta\}}$	$\theta^{-\alpha} \mathbf{I}_{\{0 < m < \theta\}}$	$\theta^{-\alpha+n} \mathbf{I}_{\{0 < \max(m,$

$x(n) < \theta\}$

(v) $N(0,1/\tau): f(\mathbf{x}|\tau) \propto \tau^{n/2} \times \exp\{-\tau \sum_{1 \leq i \leq n} x_i^2/2\}$
 $\pi(\tau) \propto \tau^{\alpha-1} e^{-\beta\tau}$, and $\pi(\tau|\mathbf{x}) \propto \tau^{\alpha+n/2-1} \times \exp\{-\tau[\beta + \sum_{1 \leq i \leq n} x_i^2/2]\}$.

(vi) $N(\mu,1/\tau): f(\mathbf{x}|\mu,\tau) \propto \tau^{n/2} \times \exp\{-\tau \sum_{1 \leq i \leq n} (x_i - \mu)^2/2\}$

The prior is $\pi(\mu,\tau) = \pi(\tau) \times \pi(\mu|\tau)$, (Product rule.)

where $\pi(\tau) \propto \tau^{\alpha-1} \exp\{-\beta\tau\}$, $\pi(\mu|\tau) \propto \tau^{1/2} \times \exp\{-v\tau(\mu-m)^2/2\}$, $v > 0$.

That is, τ is Gamma ($\alpha; 1/\beta$) and $\mu|\tau$ is $N(m,1/(\tau v))$.

[We call this Gamma-Normal ($\alpha, 1/\beta, m, 1/v$).]

The posterior is $\pi(\mu,\tau|\mathbf{x}) = \pi(\tau|\mathbf{x}) \times \pi(\mu|\mathbf{x},\tau)$. [Product rule (1.5)]

Using identities (\hat{a} denotes the sample average.)

$$\sum_{1 \leq i \leq n} (x_i - \mu)^2 = \sum_{1 \leq i \leq n} (x_i - \hat{a})^2 + n(\hat{a} - \mu)^2, \quad (\text{Id.1})$$

and $n(\mu - \hat{a})^2 + v(\mu - m)^2 = (\mu - m^*)^2 v^* + (m - \hat{a})^2 n v / v^*$. (Id.2)

$\tau|\mathbf{x}$ is Gamma ($\alpha^*; 1/\beta^*$) and $\mu|\mathbf{x},\tau$ is $N(m^*, (\tau v^*)^{-1})$,

$$v^* = v+n, \text{ and } m^* = (vm+n\hat{a})/v^*, \alpha^* = \alpha+n/2,$$

and $\beta^* = \beta + (1/2)[\sum_{1 \leq i \leq n} (x_i - \hat{a})^2 + (m - \hat{a})^2(nv/v^*)]$.

(vii) The k -dimensional column vector $\mathbf{Y} = (Y_1, \dots, Y_k)^T | (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

with a density $f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\tau}) \propto |\boldsymbol{\tau}|^{1/2} \times \exp\{-(1/2) (\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\tau} (\mathbf{y} - \boldsymbol{\mu})\}$, $\boldsymbol{\tau} \equiv \boldsymbol{\Sigma}^{-1}$.

The n iid copies of it, $\mathbf{X}_1, \dots, \mathbf{X}_n$, has a joint density

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\tau}) \propto |\boldsymbol{\tau}|^{n/2} \times \exp\{-(1/2) \sum_{1 \leq i \leq n} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\tau} (\mathbf{x}_i - \boldsymbol{\mu})\}.$$

The prior is: $\boldsymbol{\tau}$ is $W(\alpha, \boldsymbol{\psi}): \pi(\boldsymbol{\tau}) \propto |\boldsymbol{\tau}|^{(\alpha-k-1)/2} \times \exp\{-(1/2) \text{trace}(\boldsymbol{\psi} \boldsymbol{\tau})\}$, and

$\mu|\tau$ is $N(\mathbf{m}, (v\tau)^{-1})$, $t>0$: $f(\mu|\mathbf{m},\tau) \propto |\tau|^{1/2} \times \exp\{-v(\mu-\mathbf{m})^T\tau(\mu-\mathbf{m})/2\}$.

Note that $t, \mathbf{m}, \alpha, \psi$ are prior parameters. The posterior density is

$$\pi(\mu, \tau|\mathbf{x}) = \pi(\tau|\mathbf{x}) \times \pi(\mu|\mathbf{x}, \tau) \text{ where}$$

$$\tau|\mathbf{x} \text{ is } W(\alpha^*, \psi^*) \text{ and } \mu|\mathbf{x}, \tau \text{ is } N(\mathbf{m}^*, (t^*\tau)^{-1});$$

$$v^* = v+n, \mathbf{m}^* = (v\mathbf{m}+\mathbf{a})/v^*, \alpha^* = \alpha+n,$$

$$\psi^* = \psi + \sum_{1 \leq i \leq n} (\mathbf{x}_i - \mathbf{a})(\mathbf{x}_i - \mathbf{a})^T + (\mathbf{m} - \mathbf{a})(\mathbf{m} - \mathbf{a})^T (nv/v^*),$$

and \mathbf{a} is the column vector of average $(\mathbf{X}_1 + \dots + \mathbf{X}_n)/n$.

[The proof is identical to that of (vi), using matrix identities analogous to (Id.1) and (Id.2); see Section 9.10 in De Groot.]

(viii) A categorical data (Multinomial) model.

Sample with replacement n times from a finite population with k different groups, and the proportion of the j -th group in the population is p_j , $j=1, \dots, k$. ($\sum_{1 \leq j \leq k} p_j = 1$.) Let x_j = the number of selections from the j -th group. Let \mathbf{x} denote (x_1, \dots, x_k) . Let θ denotes $(p_1, \dots, p_k) \in \Theta$, where

$$\Theta = \{(p_1, \dots, p_k): \sum_{1 \leq j \leq k} p_j = 1\} \text{ is the parameter space.}$$

The joint density of $\mathbf{x}|\theta$ is

$$f(\mathbf{x}|\theta) = \prod_{1 \leq j \leq k} p_j^{x_j}.$$

A conjugate prior density on $\theta = (p_1, \dots, p_k)$ would be proportional to a function of the form $\prod_{1 \leq j \leq k} p_j^{\alpha_j}$, and a Dirichlet $(\alpha_1, \dots, \alpha_k)$ density on Θ fits the requirement: ($\alpha_j > 0$.)

$$\pi(p_1, \dots, p_k) \propto \prod_{1 \leq j \leq k} p_j^{\alpha_j - 1}, \sum_{1 \leq j \leq k} p_j = 1.$$

It follows then

$$\pi(p_1, \dots, p_k|\mathbf{x}) \propto \prod_{1 \leq j \leq k} p_j^{\alpha_j^* - 1}, \sum_{1 \leq j \leq k} p_j = 1,$$

which is Dirichlet $(\alpha_1^*, \dots, \alpha_k^*)$ where $\alpha_j^* = \alpha_j + x_j$, $j=1, \dots, k$.

(ix) Right censored data model with a discrete lifetime. Suppose the

life

time can take only values $s=1, \dots, k$. Let θ denotes the k categorical probabilities: $\theta \equiv (p_1, \dots, p_k) \in \Theta = \{(p_1, \dots, p_k): \sum_{1 \leq s \leq k} p_s = 1\}$. The model density is [See Example 1.3 (vi).]

$$\{\prod_i f(x_i|\theta)\} \times \prod_j [1-F(c_j|\theta)];$$

the index i runs through the indices of the uncensored data, and j runs through the indices of the censored data. Both products can be written as products with the index running through the categories as

$$\prod_i f(x_i|\theta) = \prod_{1 \leq s \leq k} p_s^{d_s} \text{ and } \prod_j [1-F(c_j|\theta)]$$

$$= \prod_{1 \leq j \leq k} (p_{s+1} + \dots + p_k)^{t_s},$$

$d_s =$ number of deaths (complete data) observed at s ,

$t_s =$ number of right censored data observed at s ,

and

$$r_s = d_s + t_s + \dots + d_k + t_k \text{ for } s=1, \dots, k.$$

The model density $\{\prod_i f(x_i|\theta)\} \times \prod_j [1-F(c_j|\theta)]$ becomes

$$\prod_{1 \leq j \leq k} p_s^{d_s} \times \prod_{1 \leq j \leq k} (p_{s+1} + \dots + p_k)^{t_s}$$

$$= \prod_{1 \leq j \leq k} \lambda_s^{d_s} \times \prod_{1 \leq j \leq k} [(1-\lambda_1) \dots (1-\lambda_s)]^{t_s + d_s} \times \prod_{1 \leq j \leq k} (1-\lambda_s)^{r_s - d_s}$$

d_s

$$= \prod_{1 \leq j \leq k} [\lambda_s^{d_s} \times (1-\lambda_s)^{r_s - d_s}];$$

where the $k-1$ p_s s are re-parametrized to the $k-1$ hazard rates

$$\lambda_s = \Pr\{\text{a death at } s | \text{death at } s, s+1, \dots, \text{ or } k\}$$

$$= p_s / (p_s + \dots + p_k), \quad s=1, \dots, k-1.$$

A conjugate prior on λ_j s is: λ_j s are independent Beta (α_j, β_j) s. Therefore, $\lambda_j | (r_j, d_j)$, $j=1, \dots, k$, are independent Beta (α_j^*, β_j^*) s where $\alpha_j^* = \alpha_j + d_j$ and $\beta_j^* = \beta_j + r_j - d_j$.

(x) A cluster model. Let x_1, \dots, x_n be iid data from an arbitrary unknown density. Let $\mathbf{p} = C_1, \dots, C_{n(p)}$ be a partition of $\{1, \dots, n\}$. We construct a

cluster model via a prescribed symmetric joint density $k(x_1, \dots, x_n)$:
 $[k(x_1, \dots, x_n)]$ also yields all sub-sample joint densities denoted
descriptively by $k(x_j, j \in C)$.]

$$f(\mathbf{x}|\mathbf{p}) \equiv f(x_1, \dots, x_n | \mathbf{p}) \propto \prod_{1 \leq i \leq n(\mathbf{p})} k(x_j, j \in C_i).$$

Let $\Theta = \{\mathbf{p}: \mathbf{p} \text{ is a partition of } \{1, \dots, n\}\}$; Θ is a finite set. Let g be a
positive-valued function defined on subsets of $\{1, \dots, n\}$. The prior density

$$\pi(\mathbf{p}|g) \propto \prod_{1 \leq i \leq n(\mathbf{p})} g(C_i) \text{ is a density on } \Theta.$$

(a) What is the proportional constant in $\pi(\mathbf{p}|g) \propto \prod_{1 \leq i \leq n(\mathbf{p})} g(C_i)$?

(b) Is $\{\pi(\mathbf{p}|g): g \text{ varies}\}$ a family of conjugate prior densities for
the cluster model $f(\mathbf{x}|\mathbf{p})$?

(c) What is the posterior density $\pi(\mathbf{p}|\mathbf{x})$ of \mathbf{p} if the prior is $\pi(\mathbf{p}|g)$?

(d) Let $e_0 > 0$. If $g(C) \equiv e_0 \times (\# \text{ of elements in } C - 1)!$,

and $g(\text{empty set}) \equiv 1$,

$\pi(\mathbf{p}|g)$ is (the distribution of) a Chinese restaurant process (e_0).

Bayesian interval estimation. Take iid sampling from the $N(\mu, 1/\tau)$
density as in (vi) of the previous example. The shortest interval A such
that $\pi\{\mu \in A | \mathbf{x}\} = 95\%$ could be a 95% posterior interval for the Normal
mean μ . For a Gamma-Normal $(\alpha, 1/\beta, m, 1/t)$ prior on (τ, μ) , the marginal
density of μ is

$$\begin{aligned} \pi(\mu) &\propto \int \tau^{\alpha-1} \exp\{-\beta\tau\} \times \tau^{1/2} \exp\{-(\tau v/2)(\mu-m)^2\} d\tau \\ &\propto \int \tau^{\alpha+1/2-1} \exp\{-\tau[\beta+(v/2)(\mu-m)^2]\} d\tau \\ &\propto 1/\{\beta+(t/2)(\mu-m)^2\}^{\alpha+1/2} \\ &\propto 1/\{1+(t/2\beta)(\mu-m)^2\}^{\alpha+1/2}, \end{aligned}$$

which is a t-density with degrees of freedom 2α , location m , and
precision $t(\alpha/\beta)$. Hence, a posterior density of μ is a t-density with
degrees of freedom $2\alpha^*$, location m^* , and precision $v^*(\alpha^*/\beta^*)$. If $t_{2.5\%}$ is

the 97.5 percentile point of the t-density with degrees of freedom $2\alpha^*$, the posterior (shortest) 95% interval is

$$(1.5) \quad m^* \pm t_{2.5\%} \times 1/\sqrt{[v^*(\alpha^*/\beta^*)]}.$$

(Compare with the classic t-confidence interval for the mean.)

In general, if the posterior density is unimodal, the choice of a "shortest" posterior interval is possible; otherwise, there could be several choices of "shortest" intervals. Consult a text in statistical theory for further discussions.

1.3. A Bayesian hypothesis test. The general strategy for Bayesian inference is to assign a prior distribution to the range of the unknown, and use the Bayes theorem to turn the prior and the model density to derive posterior. The situation in the testing case can be summarized as follows. A sampling plan yields data x that has a model density given by $f(x|\theta)$, $\theta \in \Theta$ could be multi-dimensional. (See Examples 1.2.), θ is an unknown parameter. For two disjoint subsets of Θ , say Θ_1 and Θ_2 , we have to make a choice whether

$$H_1: \theta \in \Theta_1 \text{ vs } H_2: \theta \in \Theta_2.$$

We start by assuming a prior distribution on $\Theta_1 \cup \Theta_2$, This prior must guarantee that the two hypotheses have positive probabilities, i.e., that $\pi\{\theta \in \Theta_j\} > 0$ for $j=1$ and 2 . [Otherwise the corresponding posterior probabilities remains zero.] This requires a "mixture prior": $\pi_1 > 0$ and $\pi_2 = 1 - \pi_1 > 0$,

$$(1.6) \quad \pi(\theta)d\theta = \pi_1 \times \pi(\theta|1)d\theta + \pi_2 \times \pi(\theta|2)d\theta;$$

$$\pi(\theta|j)d\theta \text{ is a distribution on } \Theta_j, j=1,2.$$

For the mixture prior (1.6), the prior probabilities for the two choices are

$$\pi\{\theta \in \Theta_1\} = \pi_1 \text{ and } \pi\{\theta \in \Theta_2\} = \pi_2 = 1 - \pi_1$$

Since $\theta \sim \pi(\theta)d\theta$, and $x|\theta \sim f(x|\theta)$, the joint distribution of (θ, x) is

$$\pi(\theta)f(x|\theta)dx d\theta,$$

and the posterior distribution of $\theta|x$ is

$$(1.7) \quad \pi(\theta|x)d\theta \propto \pi_1 \times f(x|\theta) \pi(\theta|1)d\theta + \pi_2 \times f(x|\theta) \pi(\theta|2)d\theta;$$

implying $\pi\{\theta \in \Theta_j | x\} \propto \pi_j \times \int_{\Theta_j} f(x|\theta) \pi(\theta|j)d\theta$ for $j=1, 2$.

[Note that $\int_{\Theta_j} f(x|\theta) \pi(\theta|j)d\theta$ is the conditional density of x given $\theta \in \Theta_j$.]

A decision criterion: The Θ_j with a higher posterior probability is preferred.

Example 1.5.

(i) $X_i | (\mu, \tau)$ are iid $N(\mu, 1/\tau)$. μ_0 is a fixed number < 1 . It is desired to test

$$H_1: \mu = \mu_0 \text{ vs } H_2: \mu > 1.$$

Here $\theta = (\tau, \mu)$, $\Theta_1 = \{(\tau, \mu): \tau > 0, \mu = \mu_0\}$ with a distribution $\pi(\tau|1)d\tau$, and $\Theta_2 = \{(\tau, \mu): \tau > 0, \mu > 1\}$ with a distribution $\pi(\tau, \mu|2)d\tau d\mu$. Check that Θ_1 and Θ_2 are disjoint. Hence, by (1.5)

$$\pi\{\theta \in \Theta_1 | \mathbf{x}\} \propto \pi_1 \times \int_{\Theta_1} f(\mathbf{x}|\tau, \mu_0) \pi(\tau|1)d\tau \text{ and}$$

$$\pi\{\theta \in \Theta_2 | \mathbf{x}\} \propto \pi_2 \times f(\mathbf{x}|2) = \iint_{\Theta_2} f(\mathbf{x}|\tau, \mu) \pi(\tau, \mu|2) d\tau d\mu.$$

The idea of conjugate priors facilitates the evaluation of the integrals.

Assume that $\pi(\tau|1)$ is Gamma (α, β) restricted to Θ_1 , and $\pi(\tau, \mu|2) \propto \pi(\tau, \mu)$ which is the Gamma-Normal $(\alpha, 1/\beta, m, 1/v)$ as specified in Example 1.4 (vi), and restricted to Θ_2 . It follows then

$$\begin{aligned} \pi\{\theta \in \Theta_1 | \mathbf{x}\} & \propto \pi_1 \times \int \tau^{n/2} \exp\{-\tau \sum (x_i - \mu_0)^2 / 2\} \times [\beta^\alpha / \Gamma(\alpha)] \tau^{\alpha-1} \exp\{-\tau\beta\} d\tau \\ & \propto \pi_1 \times [\beta^\alpha / \Gamma(\alpha)] \times \int \tau^{\alpha-1} \exp\{-\tau\beta'(\mu_0)\} d\tau \end{aligned}$$

$$\propto \pi_1 \times [\beta^\alpha / \Gamma(\alpha)] \times [\Gamma(\alpha') / [\beta'(\mu_0)^{\alpha'}]],$$

where $\alpha' = \alpha + n/2$ and $\beta'(\mu_0) = \beta + (1/2)\sum(x_i - \mu_0)^2$.

Next, $\pi\{\theta \in \Theta_2 | \mathbf{x}\} \propto \pi_2 \times (A^*/A)$, where

$$\begin{aligned} A &\equiv \iint_{\Theta_2} \tau^{\alpha-1} \exp\{-\tau\beta\} \tau^{1/2} \exp\{-\tau v(\mu-m)^2/2\} d\tau d\mu \\ &= \Gamma(\alpha+1/2) \times \int_{\{\mu>1\}} [\beta + v(\mu-m)^2/2]^{-(\alpha+1/2)} d\mu \\ &= \Gamma(\alpha-1/2) (v)^{-1} \times [\beta + v(1-m)^2/2]^{-(\alpha-1/2)} \end{aligned}$$

and

$$\begin{aligned} A^* &\equiv \iint_{\Theta_2} \tau^{n/2} \exp\{-\tau[n(\mu-S/n)^2 + \sum(x_i-S/n)^2]/2\} \\ &\times \tau^{\alpha-1} \exp\{-\tau\beta\} \tau^{1/2} \exp\{-\tau v(\mu-m)^2/2\} d\tau d\mu \\ &= \iint_{\Theta_2} \tau^{\alpha^*-1} \exp\{-\tau\beta^*\} \exp\{-\tau v^*(\mu-m^*)^2/2\} d\tau d\mu \end{aligned}$$

where α^*, β^*, m^* , and v^* are defined in Example 1.4 (vi). A^* has the same expression as A with α, β, m , and v replaced by the corresponding "*"s.

(ii) (A two sample problem.) The model is: $X_i, i=1, \dots, p | (\mu_1, \tau)$ are iid $N(\mu_1, 1/\tau)$ and $Y_j, j=1, \dots, q | (\mu_2, \tau)$ are iid $N(\mu_2, 1/\tau)$; $p+q=n$. Given $\theta = (\tau, \mu_1, \mu_2)$, X_i s and Y_j s are independent

By independence, the model density (as a function of θ) is

$$\begin{aligned} f(\mathbf{x}, \mathbf{y} | \tau, \mu_1, \mu_2) &\propto \tau^{p/2} \exp\{-\tau[p(\mu_1 - S_x/p)^2 + S_{xx}]/2\} \\ &\times \tau^{q/2} \exp\{-\tau[q(\mu_2 - S_y/q)^2 + S_{yy}]/2\}. \end{aligned}$$

Notation: $a_x = \sum_i x_i/p$, $a_y = \sum_j y_j/q$, $S_{xx} = \sum_i (x_i - a_x)^2$, $S_{yy} = \sum_j (y_j - a_y)^2$, $S = pa_x + qa_y$.

We are interested in the difference between two means $\Delta = \mu_2 - \mu_1$. The test is

$$H_1: \Delta = 0 \text{ vs } H_2: \Delta > 0.$$

Since Θ_1 is $\{(\tau, \mu_1, \mu_2): \tau > 0, \mu_1 = \mu_2\}$, and $\Theta_2 = \{(\tau, \mu_1, \mu_2): \tau > 0, \mu_2 - \mu_1 > 0\}$, Θ_1 and Θ_2 are disjoint. It remains to compute $\pi\{\theta \in \Theta_j | \mathbf{x}\}$ s in (1.7). Let $\mu = \mu_1 = \mu_2$, and write Θ_1 as $\{(\tau, \mu): \tau > 0, -\infty < \mu < \infty\}$.

For $\pi\{\theta \in \Theta_1 | \mathbf{x}\}$, the relevant integral is $\iint_{\Theta_1} f(\mathbf{x}, \mathbf{y} | \tau, \mu)$
 $\pi(\tau, \mu | 1) d\tau d\mu$,

where

$$\begin{aligned}
f(\mathbf{x}, \mathbf{y} | \tau, \mu) &\propto \tau^{n/2} \exp\{-(\tau/2)(S_{xx} + S_{yy})\} \\
&\quad \times \exp\{-(\tau/2)[p(\mu - a_x)^2 + q(\mu - a_y)^2]\} \\
&\propto \tau^{n/2} \exp\{-(\tau/2)[S_{xx} + S_{yy} + (a_x - a_y)^2 pq/n]\} \\
&\quad \times \exp\{-(\tau/2)n(\mu - S/n)^2\}.
\end{aligned}$$

By inspection, the idea of conjugate priors suggest that letting $\pi(\tau, \mu | 1) d\tau d\mu$ be Gamma-Normal $(\alpha, 1/\beta, m, 1/v)$ simplifies this integral.

Next, for $\pi\{\theta \in \Theta_2 | \mathbf{x}\}$, the integral is

$$\iint_{\Theta_2} f(\mathbf{x}, \mathbf{y} | \tau, \mu_1, \mu_2) \pi(\tau, \mu_1, \mu_2 | 2) d\tau d\mu,$$

where

$$\begin{aligned}
f(\mathbf{x}, \mathbf{y} | \tau, \mu_1, \mu_2) &\propto \tau^{n/2} \exp\{-\tau(S_{xx} + S_{yy})/2\} \\
&\quad \times \exp\{-\tau p(\mu_1 - a_x)^2/2\} \times \exp\{-\tau q(\mu_2 - a_y)^2/2\}.
\end{aligned}$$

By inspection, the idea of conjugate priors suggest that τ is Gamma $(\alpha, 1/\beta)$,

and $\mu_1, \mu_2 | \tau$ are independent $N(m_1, 1/(\tau v_1))$ and $N(m_2, 1/(\tau v_2))$, respectively.

This gives the distribution $\pi(\tau, \mu_1, \mu_2) d\tau d\mu_1 d\mu_2$. Finally conditional on Θ_2 ,

$$\pi(\tau, \mu_1, \mu_2 | 2) \propto \pi(\tau, \mu_1, \mu_2) \times I\{\tau, \mu_1, \mu_2 \in \Theta_2\}.$$

Exercise. Evaluate $\pi\{\theta \in \Theta_j | \mathbf{x}\}$, $j=1, 2$ in (ii).

Remark 1.3. (A multiple decision problem.) For a model density $f(\mathbf{x} | \theta)$, we would like to see if the unknown θ is in $\Theta_j, j=1, \dots, k$. The Θ_j s are disjoint. With a mixture prior

$$(1.6') \quad \pi(\theta) d\theta = \pi_1 \times \pi(\theta | 1) d\theta + \dots + \pi_k \times \pi(\theta | k) d\theta,$$

where $\pi(\theta | j) d\theta$, is a distribution on $\Theta_j, j=1 \Delta k$. The posterior probabilities are

$$\pi\{\theta \in \Theta_j | \mathbf{x}\} \propto \pi_j \times \int_{\Theta_j} f(\mathbf{x} | \theta) \pi(\theta | j) d\theta \text{ for } j=1, \dots, k.$$

As in the testing case, the Θ_j with the highest posterior probability is preferred.

1.4. Bayesian prediction. The data is assumed to have a joint density $k(x_1, \dots, x_n)$ which is known. This is the marginal density of x_1, \dots, x_n . In case the sampling distribution is $f(x_1, \dots, x_n | \theta)$ such that θ is unknown, we integrate out the θ with the help of a prior $\pi(\theta)d\theta$ to get

$$(1.8) \quad k(x_1, \dots, x_n) = \int f(x_1, \dots, x_n | \theta) \pi(\theta) d\theta.$$

Averaging out the unknown θ results in depending x_i s that have a joint density $k(x_1, \dots, x_n)$, which is the basis of Bayesian prediction. For example, given $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n$, how would one predict the observation x_k ? Since (1.8) gives the joint density of x_1, \dots, x_n , we can derive all conditional densities. In particular, the conditional density of x_k given $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n$ is proportional to the joint density of all variables while the variables being conditional on, i.e. $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n$, are treated as fixed constants:

$$k(x_k | x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n) \propto k(x_1, \dots, x_{k-1}, x_k, x_{k+1}, \dots, x_n);$$

the proportional constant depends on the given $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n$.

Therefore, the predictive probability of $x_k \in A$ given $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n$ is

$$(1.9) \quad \begin{aligned} & \Pr\{x_k \in A | x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n\} \\ &= \int_A k(x_k | x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n) dx_k. \end{aligned}$$

For example, a shortest interval A with 95% predictive probability is a 95% predictive interval for x_k given the values of $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n$.

If the $x_i | \theta$ are iid with distribution iid $F(x | \theta)$, the conjugate prior idea facilitates the computation as follows.

$$(1.10) \quad \begin{aligned} & \Pr\{X_{n+1} \leq t | x_1, \dots, x_n\} \\ &= \int \Pr\{X_{n+1} \leq t | \theta, x_1, \dots, x_n\} \pi(\theta | x_1, \dots, x_n) d\theta \\ &= \int \Pr\{X_{n+1} \leq t | \theta\} \pi(\theta | x_1, \dots, x_n) d\theta \quad (x_i | \theta \text{ are} \end{aligned}$$

independent.)

$$= \int F(t|\theta) \pi(\theta|x_1, \dots, x_n) d\theta \quad [x_{n+1}|\theta \sim F(.|\theta).]$$

Assuming that one can switch (Math Q: Conditions ?) the derivative (with respect to t) and the integral, the predictive density of $x_{n+1}|x_1, \dots, x_n$ is

$$k(t|x_1, \dots, x_n) = \int f(t|\theta) \pi(\theta|x_1, \dots, x_n) d\theta.$$

The conjugate prior idea suggests that we compute the predictive density based on no data $\int f(t|\theta) \pi(\theta) d\theta$, and then simply update the prior parameters to get $\int f(t|\theta) \pi(\theta|x_1, \dots, x_n) d\theta$.

Example 1.6. (i) The sun rised for the last n days, what is the (predictive) probability that the sun will rise tomorrow? We take a model-based statistical approach. Suppose $x_i=1$ if the sun rises on the i-th day, and θ is the unknown probability of this event. The model assumption is $x_1, \dots, x_n|\theta$ are iid Bernoulli (θ), $0 < \theta < 1$. Here the model density for x_i is for $x_i=t$, $f(t|\theta) = \theta^t(1-\theta)^{1-t}$, t is 0 or 1. Suppose a Bayesian determines that the unknown θ is equally likely to be any number between 0 and 1, and assumes a Uniform (0,1) prior for θ . Since a Uniform (0,1) density is Beta (1,1), we shall derive the answer using the Beta (α, β) conjugate prior, and then letting the prior parameters be 1. In this case,

$$\Pr\{x_1=1\} = \int_{\{0 < \theta < 1\}} \theta \pi(\theta) d\theta = \alpha/(\alpha+\beta),$$

And the predictive probability is: ($\alpha^* = \alpha + S$, $\beta^* = \beta + n - S$, and $S = \sum_{1 \leq i \leq n} x_i$.)

$$\Pr\{x_{n+1}=1|x_1, \dots, x_n\} = \alpha^*/(\alpha^* + \beta^*) = (\alpha + S)/(\alpha + \beta + n).$$

For a uniform prior $\alpha = \beta = 1$, and S is the number that the sun has risen for the last n days, hence the answer is $\Pr\{x_{n+1}=1|x_1, \dots, x_n\} = (1+S)/(2+n)$.

(ii) The data $x_i|\theta$ are iid $N(\mu, 1/\tau)$, $\theta = (\tau, \mu)$. Assume a Gamma-Normal ($\alpha, 1/\beta, m, 1/v$) conjugate prior $\pi(\tau, \mu) d\tau d\mu$. [See Example 1.4 (vi) for notation.] The predictive density based on no data is

$$\begin{aligned}
& \iint f(t|\theta) \pi(\theta) d\theta \\
& \propto \iint \tau^{1/2} \exp\{-(\tau/2)(t-\mu)\} \times \tau^{\alpha-1} \exp\{-\tau\beta\} \tau^{1/2} \exp\{-(\tau/2)v(\mu-m)^2\} \\
& d\tau d\mu \\
& \propto \iint \tau^{\alpha+1/2-1} \exp\{-\tau\beta\} \tau^{1/2} \exp\{-(\tau/2)[(t-\mu)^2 + v(\mu-m)^2]\} d\tau d\mu \\
& \propto \iint \tau^{\alpha+1/2-1} \exp\{-\tau[\beta+(1/2)(t-m)^2v/(v+1)]\} \\
& \quad \times \tau^{1/2} \exp\{-(\tau v^*/2)[(\mu-m^*)^2] d\mu d\tau \\
& \propto \int \tau^{\alpha+1/2-1} \exp\{-\tau[\beta+(1/2)(t-m)^2v/(v+1)]\} d\tau \\
& \propto \{\beta+(1/2)(t-m)^2v/(v+1)\}^{-(\alpha+1/2)} \\
& \propto \{1+(1/2\beta)(t-m)^2v/(v+1)\}^{-(\alpha+1/2)},
\end{aligned}$$

which is a t-density with degrees of freedom 2α , location m , and precision $(\alpha/\beta)v/(v+1)$. Hence, the predictive density of $x_{n+1}|x_1, \dots, x_n$ is a t-density with degrees of freedom $2\alpha^*$, location m^* , and precision $(\alpha^*/\beta^*)\times v^*/(v^*+1)$.

A 95% predictive interval for the next observation x_{n+1} is

$$(1.11) \quad m^* \pm t_{2.5\%} \times \sqrt{(v^*+1)/[(\alpha^*/\beta^*)\times v^*]}.$$

[Compare with the classic predictive interval, and with the predictive interval for the unknown mean μ given by (1.5).]

Remark 1.4. While a Bayesian solves a prediction problem by averaging out the unknown parameter using a prior distribution, a non-Bayesian could eliminate the unknown conditioning on a sufficient statistics. The sufficient statistics of choice is preferred to be the "minimal" sufficient statistics (if it exists.)

2. Large-sample behaviors of a posterior distribution

2.1. Doob's (1949) law of large numbers for a posterior distribution.

Some preliminaries in probability. We have learned from the product rule that for two random variables X and θ . $Y \sim \pi(\theta)d\theta$ [also denoted by $\pi(d\theta)$], and $x|\theta \sim f(x|\theta)dx \equiv F(dx|\theta)$, the product rule for densities states that if

$$q(x) \equiv \int f(x|\theta) \pi(\theta)d\theta$$

is the "normalization" constant making $\pi(\theta|x)d\theta$ a distribution of $\theta|x$,

$$\pi(\theta) \times f(x|\theta) = \pi(\theta|x) \times q(x).$$

The product rule actually is a change of order of integration in Calculus:

$$\iint h(\theta,x) f(x|\theta)dx \pi(\theta)d\theta = \iint h(\theta,x) \pi(\theta|x)d\theta q(x)dx,$$

for any non-negative (measurable) function $h(\theta,x)$.

In the distribution form, [$Q(dx) \equiv q(x)dx$ is the marginal dist. of x .]

$$(2.1) \quad \iint h(\theta,x) F(dx|\theta) \pi(d\theta) = \iint h(\theta,x) \pi(d\theta|x) Q(dx), \text{ any } h(\cdot) \geq 0.$$

It also gives the familiar double expectation formula:

$$(2.1') \quad E\{E[h((\theta,X)|\theta)]\} = E[h((\theta,X))] = E\{E[h((\theta,X)|X)]\}, \text{ any}$$

$h(\cdot) \geq 0$.

[The equality also defines the "Palm distribution" in random measure theory. It is also the basis of a theory called disintegration of the joint distribution of (x,θ) .] We simply call this switching of order of integration (2.1) or (2.1') a "Fubini theorem."

Example 2.1. Let X and Y be two random variables with a joint distribution $F(x,y)$, and C is an event on the x - y plane. Discuss the meanings of $P\{C\}=1$, and $P\{C\}=0$ using (2.1) for

(i) $C=\{(x,y): x=y\}$, and (ii) $C=\{(x,y): y=g(x)\}$ where $g(x)$ is a function of

x.

Exercise 2.1. Let $h(y)$ be a function on the line, and $0 \leq h(y) \leq 1$. Let Y a random variable on the line. Suppose $E[h(Y)] = 1$. Discuss the value of $h(y)$, $-\infty < y < \infty$ if

(i) Y is the # obtained in a roll of a die, (ii) Y is Poisson (1), (iii) Y is Gamma (1,1), and (iv) Y is $N(0,1)$.

In traditional statistics, the model distribution of $x_1, x_2, \dots, x_n | \theta$ is specified as $f(x_1, x_2, \dots, x_n | \theta) d(x_1, x_2, \dots, x_n)$ which is written as $F(dx | \theta)$, where \mathbf{x} denotes (x_1, x_2, \dots, x_n) . This holds for arbitrary n , and there is a joint distribution (conditional on θ) on the space of \mathbf{x} s, where $\mathbf{x} \equiv (x_1, x_2, \dots)$. An estimator $T_n \equiv T(x_1, \dots, x_n)$ is said to be a consistent estimator of $h(\theta)$ if T_n converges to $h(\theta)$ in $P(\cdot | \theta)$ -probability:

$$(2.2) \quad P\{|T_n - h(\theta)| \geq \epsilon | \theta\} \rightarrow 0 \text{ for each } \theta.$$

This states that the collection of θ s such that $P\{|T_n - \theta| \geq \epsilon | \theta\} \rightarrow 0$ is equal to Θ . The fact that there exists a consistent classical estimator of an unknown parameter in the sense of (2.2) has tremendous implications in Bayesian statistics. A Bayesian assumes a prior $\pi(d\theta)$ for θ to make $(\theta, x_1, \dots, x_n)$ has a joint distribution for each n , and a joint distribution on the space of $(\theta, x_1, x_2, \dots)$ s. From (2.2), $E[P\{|T_n(x_1, x_2, \dots, x_n) - \theta| \geq \epsilon | \theta\}] \rightarrow 0$. (Bounded convergence theorem.) However,

$$(2.3) \quad \begin{aligned} E\{P(|T_n(x_1, x_2, \dots, x_n) - \theta| \geq \epsilon | \theta)\} \\ = \int P(|T_n(x_1, x_2, \dots, x_n) - \theta| \geq \epsilon | \theta) \pi(\theta) d\theta \\ = P(|T_n(x_1, x_2, \dots, x_n) - \theta| \geq \epsilon) \end{aligned} \quad (\text{by 2.1'})$$

This means that the random variable T_n converges to a target random variable θ in probability [i.e., in joint-probability of $(\theta, x_1, x_2, \dots)$ and for

whatever prior distribution of θ .] Hence, there is a subsequence $\{T_{n'}\} \subseteq \{T_n\}$, and $T_{n'}$ converges to θ with (joint-) probability one. Since $T_{n'}$ is a function of $(x_1, \dots, x_{n'})$, it is also a function of the whole sequence (x_1, \dots, x_n, \dots) . Hence the limit of $T_{n'}$ should also be a function of the whole sequence with probability one. [Another way of saying this is that the limit of $T_{n'}$ is in the completion of the sigma field of $(\theta, x_1, x_2, \dots)$ s.] This also means that one can compute the parameter θ based on a whole sequence of T_n s and hence of observations x_1, x_2, \dots . Let us state this as a basic assumption:

Main assumption. There exists a measurable (w.r.t. completion sigma field) function $g(\cdot)$ of the sequence such that $\theta = g(x_1, \dots, x_n, \dots)$.

Theorem 2.1 (Doob, 1949): Let x_1, \dots, x_n, \dots be a sequence of observations, and θ is a parameter. The model distribution is $x_1, \dots, x_n | \theta$ given by, for each n , $f(x_1, x_2, \dots, x_n | \theta) d(x_1, x_2, \dots, x_n)$, and $\theta \sim \pi(d\theta) d\theta$. Assume that θ is a (measurable) function of the observations x_1, \dots, x_n, \dots , $\int h(\theta) \pi(\theta) d\theta$ is finite implies

$$P\{E[h(\theta)|x_1, \dots, x_n] \rightarrow h(\theta) | \theta\} = 1, \text{ for almost all } \theta.$$

Poof. First, the forward martingale convergence theorem [See Remark 2.1 (a) below] states that with probability one,

$$e_n(x_1, \dots, x_n) \equiv E[h(\theta)|x_1, \dots, x_n] \rightarrow E[h(\theta)|x_1, x_2, \dots].$$

Here probability is the joint distribution of $\omega \equiv (\theta, x_1, x_2, \dots)$.

Next, according to the assumption, there is a (measurable) function g of the sequence of observations such that $g(x_1, x_2, \dots) = \theta$. The existence of such a function $g(\cdot)$ implies that $h(\theta) = h(g(x_1, x_2, \dots))$, and $h(\theta)$ is a (composite) function of $x_1, x_2, \dots, x_n, \dots$, and hence (almost surely)

$$\begin{aligned} E[h(\theta)|x_1, x_2, \dots] &= E[\text{hog}(x_1, x_2, \dots)|x_1, x_2, \dots] = \text{hog}(x_1, x_2, \dots) \\ &= h(\theta). \end{aligned}$$

Thus, $e_n(x_1, \dots, x_n) \equiv E[h(\theta)|x_1, \dots, x_n] \rightarrow h(\theta)$ almost surely.

Finally, let $C = \{\omega: e_n \rightarrow h(\theta)\}$,

$$1 = P\{C\} = E[I_{\{\omega \in C\}}] = E[E\{I_{\{\omega \in C\}}|\theta\}] = E[P\{C|\theta\}].$$

[The third equality follows from the Fubini theorem (2.1).]

The proof is concluded by noting that $1 = E[P\{C|\theta\}] = \int P\{C|\theta\}\pi(\theta)d\theta$ is equivalent to $P\{C|\theta\}=1$ for almost all θ , while the "almost all" is respect to the prior distribution $\pi(\theta)d\theta$. (See Esxercise 2.1.) This completes the proof. \parallel

There are three key steps in the proof. The first is based on the martingale convergence, and the last is based on Fubini theorem. The second step involves the main assumption that the unknown θ can be computed based on the sequence of observations $x_1, x_2, \dots, x_n, \dots$, and it deserves additional comments.

The main assumption and iid sampling. We have already seen that if there exists an consistent estimator of θ in the traditional sense, then the main assumption is satisfied. Doob (1949) made another contribution: He showed that in the case of iid sampling, i.e. $x_i|s| \theta$ are iid $f(x|\theta)dx$ and the model $f(x|\theta)dx$ is "identifiable" in the sense that (Identifiability condition)

$$\theta_1 \neq \theta_2 \text{ implies there is an event } A \text{ such that } \int_A f(x|\theta_1)dx \neq \int_A f(x|\theta_2)dx.$$

The main idea is that the identifiability condition means that there is a one to one Borel function on $\{\theta s\}$ to $\{F(.|\theta)s\}$. Kuratowski's fundamental theorem on one to one Borel functions (that states that for an one to one

Borel map from a complete and separable metric space to another complete and separable metric space, the inverse map is also Borel) implies that the inverse map from $\{F(.|\theta)s\}$ to $\{\theta s\}$ is also Borel. (Both spaces equipped with appropriate metrics to make them complete and separable metric spaces.) Therefore, θ is a Borel function of the distribution function $F(.|\theta)$. Next, note that the empirical distribution function is a consistent estimator of $F(.|\theta)$ in the sense of (2.2). Therefore, θ is a measurable function of (x_1, \dots, x_n, \dots) , i.e. a Borel function of a measurable function is measurable, and the main assumption is satisfied under the identifiability condition.

On the null set that Bayesian consistency fails*.

(i) The case of a countable parameter space. If the parameter space is countable, $\Theta = \{\theta_1, \theta_2, \dots\}$, and θ_t is the "true" parameter.

We only need to assume a prior on Θ that is positive on each θ_j , $j=1, 2, \dots$, then we are assured of Bayesian consistency:

$$P(E[I_{\{\theta=\theta_t\}} | x_1, \dots, x_n] \rightarrow 1 \equiv I_{\{\theta_t=\theta_t\}} | \theta_t) = 1, \text{ for each } t=1, \dots$$

(Note. An example of Bahadur shows that the maximum likelihood estimator of θ_t can be inconsistent for a model with a countable parameter space.)

- The case of invariant statistical models. In an invariant statistical model, using an invariant (Haar) prior measure on the parameter space often yields a posterior distribution of θ that is also "invariant". Doob's 1949 result implies that posterior consistency occurs at at least one point. By posterior invariance, it holds for ALL points. (Lo 1984, Lo & Sazonov 2003.)

Remark 2.1. Levy convergence theorem. Let $Y, X_1, \dots, X_n, \dots$ be random variables, and $E[|Y|]$ is finite,

- (a) $E[Y|X_1, \dots, X_n] \rightarrow E[Y|X_1, \dots]$ with probability one, and
- (b) $E[Y|X_n, \dots, X_{n+k}, \dots] \rightarrow Z$ with probability one, and $EY=EZ$.

See for example page 242 in "An Introduction to Probability Theory and its Applications" by W. Feller (1971).

Blackwell and Dubins (1965) discuss how one can replace Y by Y_n in Levy's theorem. [See Chung (1974), page 340.]

If $|Y_n| \leq Z$ where $E[Z]$ is finite, $Y_n \rightarrow Y$ with probability one

- (a) $E[Y_n|X_1, \dots, X_n] \rightarrow E[Y|X_1, \dots]$ with probability one, and
- (b) $E[Y_n|X_n, \dots, X_{n+k}, \dots] \rightarrow Z$ with probability one, and $EY=EZ$.

2.2. Normal approximation to a posterior distribution.

We assume iid sampling from $f(x|\theta)$ in this section. Let us look at the behavior of a posterior distribution in two cases.

Example 2.2.

(i) Convergence to an exponential distribution. The model density

$f(x|\theta) = \theta^{-1} I_{\{0 < x < \theta\}}$, and $\pi(\theta) \propto \theta^{-\alpha} I_{\{0 < m < \theta\}}$. Since $f(\mathbf{x}|\theta) \propto \theta^{-n} I_{\{0 < x(n) < \theta\}}$,

$\pi(\theta|\mathbf{x}) \propto \theta^{-(\alpha+n)} I_{\{0 < t_n < \theta\}}$, where $t_n = \max\{m, x(n)\}$. Hence, for each $z > 0$,

$$\pi\{\theta > z|\mathbf{x}\} = (t_n/z)^{(\alpha+n)}, \text{ all } z > 0,$$

and $\pi\{n(\theta/t_n - 1) > y|\mathbf{x}\} = (1+y/n)^{-(\alpha+n)} \rightarrow \exp\{-y\}$ for each $y > 0$.

Note that $L(\theta|\mathbf{x}) = \log[f(\mathbf{x}|\theta)]$ has a jump point at $\theta = x(n)$.

(ii) Convergence to $N(0,1)$. Sampling from the $N(\mu, 1/\tau)$ model. For a Gamma-Normal $(\alpha, 1/\beta; m, 1/t)$ prior, a posterior distribution of $\mu|\mathbf{x}$ is a t-density with degrees of freedom $2\alpha+n$, location $m^* = (tm + \sum_i x_i)/(t+n)$, and precision $(t+n)(\alpha^*/\beta^*)$. See Example 1.4 (vi). Note that the posterior

density of μ is a smooth density. Given \mathbf{x} , the standardized μ is $[(t+n)(\alpha^*/\beta^*)]^{1/2} (\mu-m^*)$, which has a t-density with degrees of freedom $2\alpha+n$. This t-density tends to a $N(0,1)$ density if n increases.

We shall show that smoothness of the model density $f(\mathbf{x}|\theta)$ as a function in θ and $\pi(\theta)$ ensure that the posterior distribution of θ is approximately Normal. Example 2.2 (ii) is a case in point. One formulation of such Normal approximation to posterior densities states that at the neighborhood of the root of the likelihood equation (called the maximum likelihood estimator), the posterior density looks like a Normal density. Again, a smoothness condition that the derivatives (in θ) of $f(\mathbf{x}|\theta)$ are assumed to be differentiable in a neighborhood I of θ_n , is essential. Let

$$(2.3) \quad L(\theta|\mathbf{x}) \equiv \sum_{1 \leq i \leq n} \log f(x_i|\theta),$$

$$L^{(k)}(\theta|\mathbf{x}) \equiv \sum_{1 \leq i \leq n} (d/d\theta)^k \log f(x_i|\theta), \quad k=1,2 \text{ and } 3;$$

Here we assume that the derivatives (in θ) of $f(\mathbf{x}|\theta)$ are 3-times differentiable in a neighborhood I of θ_n , where $L^{(1)}(\theta_n|\mathbf{x})=0$. Furthermore, θ_n and its limit are in I .

(i) **Domination assumption** on the model density $f(\mathbf{x}|\theta)$:

$$\text{Sup}_{\theta \in I} |(d/d\theta)^k \log f(x_i|\theta)| \leq M_k(x) \text{ where } E[M_k(x)|\theta_0] \text{ is finite,}$$

$$k=2,3.$$

(ii) **Prior assumption** on the prior density $\pi(\theta)$:

The prior density is continuous and positive on I .

Theorem 2.2. Assume (i) and (ii). The posterior distribution of θ is approximately $N(\theta_n, 1/[-L^{(2)}(\theta_n|\mathbf{x})])$.

Proof. Since $\pi(\theta|\mathbf{x}) \propto \pi(\theta) \times \exp\{-L(\theta|\mathbf{x})\}$, a Taylor expansion for

$L(\theta|\mathbf{x})$ at the point $\theta = \theta_n$ yields

$$\pi(\theta|\mathbf{x}) \propto \pi(\theta) \exp\left\{\frac{1}{2} L^{(2)}(\theta_n|\mathbf{x}) (\theta - \theta_n)^2 + \frac{1}{3!} (\theta - \theta_n)^3 L^{(3)}(\theta^*|\mathbf{x})\right\}$$

where θ^* is between θ and θ_n . The posterior density of

$$T = (\theta - \theta_n) [-L^{(2)}(\theta_n|\mathbf{x})]^{1/2}$$

is
$$\pi_T(t|\mathbf{x}) \propto \pi(\theta_n + t [-L^{(2)}(\theta_n|\mathbf{x})]^{-1/2}) \times \exp\{-t^2/2\} \times \exp\{t^3 r_n\},$$

where
$$r_n \equiv \frac{1}{3!} L^{(3)}(\theta^*|\mathbf{x}) \times [L^{(2)}(\theta_n|\mathbf{x})]^{-3/2}.$$

Use $e^u = 1 + u + o(u)$ to write

$$(2.3) \quad \pi_T(t|\mathbf{x}) \propto \pi(\theta_n + t [-L^{(2)}(\theta_n|\mathbf{x})]^{-1/2}) \times \exp\{-t^2/2\} \times \{1 + t^3 r_n + o(r_n)\}.$$

It remains to estimate r_n and $\pi(\theta_n + t [-L^{(2)}(\theta_n|\mathbf{x})]^{-1/2})$ for each fixed t .

For brevity, let $l(\theta|z) \equiv \log f(z|\theta)$ and $l^{(k)}(\theta|z) \equiv (d/d\theta)^k \log f(z|\theta)$, $k=2,3$. Under the boundedness condition (i) on $l^{(k)}(\theta|z)$, $k=2,3$, (Notation: $E[\dots]$ means $E[\dots|\theta_0]$.)

$$(2.4) \quad \begin{aligned} (1/n) \times L^{(2)}(\theta_n|\mathbf{x}) &= E[l^{(2)}(\theta_n|X_1)|O_n] \\ &\rightarrow E[l^{(2)}(\theta_0|X_1)] \text{ as } \theta_n \rightarrow \theta_0, \end{aligned}$$

and
$$\begin{aligned} (1/n) \times \sup_{\theta \in I} |L^{(3)}(\theta|\mathbf{x})| &= (1/n) \times \sup_{\theta \in I} |\sum_i l^{(3)}(\theta|x_i)| \\ &\leq (1/n) \times \sum_i \sup_{\theta \in I} |l^{(3)}(\theta|x_i)| \\ &\leq (1/n) \times \sum_i M_3(x_i) \\ &= E[M_3(X_1)|O_n] \rightarrow E[M_3(X_1)] < \infty. \end{aligned}$$

Hence, eventually $|r_n| \leq C(\theta_0) \times n^{-1/2}$.

If $\pi(\theta)$ is continuous, $\pi(\theta_n + t [-L^{(2)}(\theta_n|\mathbf{x})]^{-1/2}) = \pi(\theta_n) + o(1)$. (2.3) becomes

$$\pi_T(t|\mathbf{x}) \propto \exp\{-t^2/2\} \times \{1 + o(1)\}. \quad \parallel$$

Theorem 2.2 states that for a smooth model, the posterior

distribution of $\theta|\mathbf{x}$ is approximately $N(\theta_n, 1/[-L^{(2)}(\theta_n|\mathbf{x})])$. An approximate $1-\alpha$ interval for θ is

$$\theta_n \pm z_{\alpha/2} \times [-L^{(2)}(\theta_n|\mathbf{x})]^{-1/2}.$$

Remark 2.2. A quick proof of (2.4) can be resulted from a backward martingale representation for averages. Suppose θ_n is a symmetric statistic. The average $(1/n)\sum_{1 \leq i \leq n} h(\theta_n, X_i)$ equals $E[h(\theta_n, X_1)|O_n]$ where O_n denotes the ordered statistics $O_n \equiv (X(1) \leq \dots \leq X(n))$ that is independent of X_{n+1}, X_{n+2}, \dots . Furthermore,

$$\begin{aligned} E[h(\theta_n, X_1)|O_n] &= E[h(\theta_n, X_1)|O_n, X_{n+1}, X_{n+2}, \dots] \\ &= E[h(\theta_n, X_1)|O_n, O_{n+1}, O_{n+2}, \dots] \end{aligned}$$

If $h(\theta_n, X_1) \rightarrow h(\theta_0, X_1)$, by the Blackwell and Dubins martingale convergence theorem, if $|h(\theta_n, X_1)|$ is dominated by an integrable random variable, i.e.,

$$\sup_{\theta} |h(\theta, X_1)| \leq M(X_1) \text{ and } E[M(X_1)] \text{ is finite,}$$

and $h(\theta_n, X_1) \rightarrow h(\theta_0, X_1)$, $E[h(\theta_n, X_1)|O_n] \rightarrow E[h(\theta_0, X_1)|O_\infty] = E[h(\theta_0, X_1)]$; the last equality follows from the Hewitt-Savage zero-one law (which states that for iid x_i s, any random variable that is a function of O_∞ must be a constant.) The convergence (2.4) follows from setting $h(\theta, X_1) \equiv l^{(k)}(\theta|X_1)$ for $k=2,3$. This method extends to multiple averages involving symmetric statistics (i.e., relatives of U-statistics).

Remark 2.3. By taking more terms for the Taylor expansions of $L(\theta|\mathbf{x})$ and the prior density $\pi(\theta)$, the previous arguments extend to obtain higher-term expansions for the posterior density in which the standard normal density as the leading term. It is known that the 1st term of the expansion depends on the derivative of the prior density due to centering

at the mle. Bertail and Lo (1996, unpublished manuscript) showed that if one centers at the posterior mean, the one-term expansion is prior-free.

