

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. $\tau|\theta$ is $W(\alpha,\psi)$, if its density is proportional to (Section 5.5, De Groot.)

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

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 distributions. 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:

$$(1.2) \quad \begin{aligned} f(\mathbf{x}|\theta) &\equiv f(x_1, x_2, \dots, x_n|\theta) \\ &= f(x_1|\theta) \times f(x_2|\theta, x_1) \times f(x_3|\theta, x_1, x_2) \times \dots \times f(x_n|\theta, x_1, x_2, \dots, x_{n-1}) \end{aligned}$$

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 $\varepsilon_i | \theta$ 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

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

Can you simplify $\text{Lik}(\mathbf{y}|\theta)$ if $\varepsilon_i | \theta$ 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

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

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) , $j=1, \dots, k$ | θ 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 | θ 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 $\pi(\theta | \mathbf{x})$, 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 $f(x|\theta)$, $\theta \in \Theta$.

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

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

- (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.. $\theta^{-n} I_{\{0 < x(n) < \theta\}}$ $\theta^{-\alpha} I_{\{0 < m < \theta\}}$ $\theta^{-\alpha+n} I_{\{0 < \max(m, x(n)) < \theta\}}$
(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 (α ; $1/\beta$) 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.)

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

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

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

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

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

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

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

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

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

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

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

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

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

and $\bar{\mathbf{x}}$ 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

$$\Theta = \{(p_1, \dots, p_k) : \sum_{1 \leq j \leq k} p_j = 1\}$$

is the parameter space, which is a $k-1$ dimensional simplex. 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 $\prod_{1 \leq j \leq k} p_j^{b_j}$, 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 $j=1, \dots, k$. Let θ denotes the k categorical

probabilities: $\theta = (p_1, \dots, p_k) \in \Theta = \{(p_1, \dots, p_k) : \sum_{1 \leq s \leq k} p_s = 1\}$. The model density is [See Example 1.3 (vi).]

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

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

$$\prod_i f(y_i | \theta) = \prod_{1 \leq s \leq k} p_s^{d_s}$$

and
$$\prod_j [1 - F(c_j | \theta)] = \prod_{1 \leq s \leq k} (p_s + \dots + p_k)^{e_s},$$

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

$e_s =$ number of right censored data (incomplete data)

observed at s ,

and
$$r_s = d_s + e_s + \dots + d_k + e_k \text{ for } s=1, \dots, k.$$

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

$$\begin{aligned} & \prod_{1 \leq s \leq k} p_s^{d_s} \times \prod_{1 \leq s \leq k} (p_s + \dots + p_k)^{e_s} \\ &= \prod_{1 \leq s \leq k} \lambda_s^{d_s} \times \prod_{1 \leq s \leq k} [(1 - \lambda_1) \dots (1 - \lambda_s)]^{d_s + e_s} \times \prod_{1 \leq s \leq k} (1 - \lambda_s)^{-d_s} \\ &= \prod_{1 \leq s \leq k} [\lambda_s^{d_s} \times (1 - \lambda_s)^{r_s - d_s}], \quad (0^0 \equiv 1) \end{aligned}$$

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

$$\begin{aligned} \lambda_s &= \Pr\{\text{a death at } s | \text{death at } s, s+1, \dots, \text{ or } k\} \\ &= p_s / (p_s + \dots + p_k), \quad s=1, \dots, k-1. \end{aligned}$$

A conjugate prior on λ_j s is: λ_j s are independent Beta (α_j, β_j) s. Therefore,

$\lambda_j | (r_k, d_k), j=1, \dots, k$, are independent Beta (α_k^*, β_k^*) s where $\alpha_k^* = \alpha_k + d_k$

and $\beta_k^* = \beta_k + r_k - d_k$.

(x) A cluster model. Let x_1, \dots, x_n be iid data from an arbitrary unknown density. Let $\mathbf{p} = C_1, \dots, C_{n(\mathbf{p})}$ 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 $k(x_j, j \in C)$.]

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

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 with parameter e_0 .

A Bayesian interval estimation problem. 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 $(\alpha, 1/\beta, m, 1/v)$ prior on (τ, μ) , the marginal density of μ is

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

which is a t-density with degrees of freedom 2α , location m , and precision $v(\alpha/\beta)$. 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, a “highest posterior density region” (HDR) is preferred. If the posterior density is unimodal, the HDR corresponds to a "shortest" posterior interval; otherwise, there could be several choices of "shortest" intervals. Consult a text in statistical theory for further discussions.

A Bayesian regression problem. Simple linear regression: $\theta=(\beta_0,\beta_1)$ and $Y_i=\beta_0+\beta_1x_i+\varepsilon_i$, $i=1,2,\dots,n$ are observable, and $\varepsilon_i|\theta$ are iid with a unimodal (error) density $f(z)$. We consider the case that $f(z)$ is $N(0,1/\tau)$. Here given θ , the observable Y_i s are not iid since the conditional means (given x_i s)

$$\begin{aligned} \text{Lik}(\mathbf{y}|\theta) &= \prod_{1 \leq i \leq n} f(y_i - \beta_0 - \beta_1 x_i | \theta) \\ &\propto \tau^{n/2} \exp\{-(\tau/2) \sum [y_i - \bar{y} + \bar{y} - \beta_0 - \beta_1(x_i - \bar{x})]^2\} \\ &\propto \tau^{n/2} \exp\{-(\tau/2) \sum (y_i - \bar{y})^2 + [\bar{y} - \beta_0 - \beta_1(x_i - \bar{x})]^2 - 2[\sum (y_i - \bar{y})\beta_1(x_i - \bar{x})]\} \\ &\propto \tau^{n/2} \exp\{-(\tau/2) \{s_{yy} [1 - (s_{xy}/s_{xx})^2] + (\beta_0 - \bar{y})^2 + s_{xx}(\beta_1 - s_{xy}/s_{xx})^2\}. \end{aligned}$$

Exercise. Show that τ is Gamma (α,β) , and $\beta_0,\beta_1|\tau$ are independent Normal($m_1,1/\tau t_1$) and Normal($m_2,1/\tau t_2$) respectively form a family of conjugate priors for the Bayesian regression problem. Find the posterior distribution of (τ,β_0,β_1) given (x_i,y_i) , $i=1,\dots,n$.

A Bayesian variable selection problem*.

Observations $\mathbf{y} = (y_1, \dots, y_n)'$, y_i is the observed i th response, and $\mathbf{x}^{n \times p} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$. $\mathbf{x}_i' = (x_{i1}, \dots, x_{ip})$: the observed value of the i th p -dimensional predictor, $i=1, \dots, n$; $\beta = (\beta_1, \dots, \beta_p)' \in \mathbb{R}^p$, the regression parameter, and the regression equation is

$$\mathbf{y} = \mathbf{x} \boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$ and $\varepsilon_1, \dots, \varepsilon_n | \mathbf{x}$ are iid errors.

Let S denote a subset of the indexes $\{1, \dots, p\}$, and

$$\beta_j = 0 \text{ if } j \notin S, \text{ and } \beta_j \neq 0 \text{ if } j \in S.$$

The joint distribution of (parameter, data) is defined by

(i) $S \sim \pi(S)$, a density on the space of all subsets of S ,

(ii) given S , $\boldsymbol{\beta}$ is defined by

$$\beta_j \sim N(0, \sigma_j^2), j \in S \text{ and } \beta_j = 0, j \notin S,$$

(iii) given $(S, \boldsymbol{\beta})$, the pairs (\mathbf{x}_i, y_i) s are independent

(iv) given $(S, \boldsymbol{\beta}, \mathbf{x}_i)$, $\varepsilon_i = y_i - \mathbf{x}_i \boldsymbol{\beta} = y_i - \sum_{t \in S} x_{it} \beta_t$, and

$$\varepsilon_i \text{ has a Normal } (0, 1/\tau) \text{ density, } i=1, \dots, n.$$

Computer the posterior distribution of $S, \boldsymbol{\beta} | \mathbf{y}, \mathbf{x}$ and the subset that carries the highest posterior probability is the subset of choice.

[George and McCulloch (93, 97).]

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 \mathbf{x} that has a model density given by $f(\mathbf{x} | \theta)$, $\theta \in \Theta$ 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

probability remains zero.] This requires a "mixture prior": $0 < \pi_1 < 1$ and $\pi_2 = 1 - \pi_1 > 0$,

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

$\pi(\theta|j)d\theta$ 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

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

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 “*” versions.

(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

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

Notation: $\bar{x} = \sum_i x_i / p$, $\bar{y} = \sum_j y_j / q$, $s_{xx} = \sum_i (x_i - \bar{x})^2$, $s_{yy} = \sum_j (y_j - \bar{y})^2$, $S = p\bar{x} + q\bar{y}$.

We are interested in the difference between two means $\Delta \equiv \mu_2 - \mu_1$. 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 $\mu \equiv \mu_1 = \mu_2$, and write Θ_1 as $\{(\tau, \mu): \tau>0, -\infty < \mu < \infty\}$.

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

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

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

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

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

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 Θ_j , $j=1, \dots, k$. 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 $F(x|\theta)$, the conjugate prior idea facilitates the computation as follows.

$$\begin{aligned}
 (1.10) \quad & \Pr\{X_{n+1} \leq t | x_1, \dots, x_n\} \\
 &= \int \Pr\{X_{n+1} \leq t | \theta, x_1, \dots, x_n\} \pi(\theta | x_1, \dots, x_n) d\theta \\
 &= \int \Pr\{X_{n+1} \leq t | \theta\} \pi(\theta | x_1, \dots, x_n) d\theta \\
 & \hspace{15em} (x_i | \theta \text{ are independent.}) \\
 &= \int F(t|\theta) \pi(\theta | x_1, \dots, x_n) d\theta \quad [x_{n+1} | \theta \sim F(\cdot | \theta).]
 \end{aligned}$$

Switching 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 coverage/confidence 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 $\{T_{n_i}\} \subseteq \{T_n\}$, and T_{n_i} converges to θ with (joint-) probability one. Since T_{n_i} is a function of (x_1, \dots, x_{n_i}) , it is also a function of the whole sequence (x_1, \dots, x_n, \dots) . Hence the limit of T_{n_i} should also be a function of the whole sequence with probability one. [Another way of saying this is that the limit of T_{n_i} 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)

$$E[h(\theta)|x_1, x_2, \dots] = E[hog(x_1, x_2, \dots)|x_1, x_2, \dots] = hog(x_1, x_2, \dots) \\ = h(\theta).$$

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 $\{F(\cdot|\theta)s\}$ 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 $F(\cdot|\theta)$. Next, note that the empirical distribution function is a consistent estimator of $F(\cdot|\theta)$ 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 $\log[f(\mathbf{x}|\theta)]$ 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 $f(x|\theta)$ 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 $f(x|\theta)$ are assumed to be differentiable in a neighborhood I of θ_n , is essential. Let

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

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

Here we assume the derivatives of θ , $l^{(k)}(\theta|\mathbf{x})$, is 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 $f(x|\theta)$: For $k=2,3$,

$$\text{Sup}_{\theta \in I} |(d/d\theta)^k \log f(x_1|\theta)| \leq M_k(x_1) \text{ where } E[M_k(X_1)|\theta_0] \text{ is finite.}$$

(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

$$\pi(\theta|\mathbf{x}) \propto \pi(\theta) \exp\{(1/2) l^{(2)}(\theta_n|\mathbf{x}) (\theta - \theta_n)^2$$

$$+ (1/3!)(\theta-\theta_n)^3 l^{(3)}(\theta^*|\mathbf{x})\},$$

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

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 $l(\theta|x_i) \equiv \log f(x_i|\theta)$ and $l^{(k)}(\theta|x_i) \equiv (d/d\theta)^k \log f(x_i|\theta)$, $k=2,3$. Under the boundedness condition (i) on $l^{(k)}(\theta|x_i)$, $k=2,3$,

(Notation: $E[\dots]$ means $E[\dots|\theta_0]$.)

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

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

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

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

becomes $\pi_T(t|\mathbf{x}) \propto \exp\{-t^2/2\} \times \{1 + O(n^{-1/2})\}$. \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 $1-\alpha$ posterior 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 $O_n = (X(1) \leq \dots \leq X(n))$ 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 $h(\theta, X_1) = l^{(k)}(\theta|X_1)$ 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 $L(\theta|x)$ 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.

3. IID Monte Carlo approximations

The posterior distribution summarizes information for Bayesians, and it would be important to calculate posterior quantities such as mean, SD, higher moments, percentiles and etc. The posterior quantities often are written as (ratios of) integrals. For example, the posterior k-th moment of θ is

$$\int \theta^k f(\mathbf{x}|\theta)\pi(\theta)d\theta / \int f(\mathbf{x}|\theta) \pi(\theta)d\theta,$$

and the posterior probability that $\theta \in (a,b]$ is

$$\int_a^b f(\mathbf{x}|\theta)\pi(\theta)d\theta / \int f(\mathbf{x}|\theta) \pi(\theta)d\theta.$$

These are ratio of two integrals, and this chapter discusses the use of Monte Carlo method to evaluate such integrals.

Suppose one wants to evaluate a finite integral

$$\zeta = \int h(\theta)d\theta.$$

The Monte Carlo method proposes that one takes an iid sample $\theta_1, \dots, \theta_M$ from a trial density $g(\theta)$ which satisfies a

Target criterion: The trial density $g(\theta)$ has a larger support than the integrand $h(\theta)$:

$$\{\theta: h(\theta) \neq 0\} \subseteq \{\theta: g(\theta) > 0\}.$$

According to the law of large numbers,

$$(3.1) \quad \zeta_M = (1/M) \sum_{1 \leq m \leq M} h(\theta_m)/g(\theta_m)$$

approximates its expectation

$$\zeta = \int h(\theta)d\theta .$$

Remark 3.1. In the case that the support of $g(\theta)$ does not cover the support of $h(\theta)$, ζ_M converges (as M grows) to

$$E[\zeta_M] = \int I\{\theta: g(\theta) > 0\} [h(\theta)/g(\theta)]g(\theta)d\theta = \int I\{\theta: g(\theta) > 0\} h(\theta)d\theta$$

that is not the target $\zeta = \int h(\theta)d\theta$. In the extreme case, the support of $g(\theta)$ and that of $h(\theta)$ are disjoint, $\zeta_M = 0$ for all M and ζ_M has no chance of approximating $\zeta = \int h(\theta)d\theta$.

The Target criterion gives a condition that ensures a correct target. If the target is correct, then a less variable MC estimator would be preferable: The variance of ζ_M is

$$(3.2) \quad \text{Var}(\zeta_M) = (1/M) \times \int [h(\theta)/g(\theta) - \zeta]^2 g(\theta)d\theta .$$

$\text{Var}(\zeta_M)$ is small if $h(\theta)/g(\theta)$ is close to a constant function – approximately. This is to be interpreted as the "main part" of $h(\theta)$ is approximately $g(\theta)$.

Variance Reduction criterion: A trial density \propto the integrand reduces $\text{Var}(\zeta_M)$. Note however that $h(\theta) \neq \zeta \times g(\theta)$ where ζ is a constant; otherwise the Monte carlo average is the constant ζ .

In practice it may be difficult to select a trial density proportional approximately to the integrand. The criterion states that one picks a trial density proportional to the integrand approximately.

Sampling a posterior distribution.

For posterior inference, the quantities to be approximated are integrals of the form

$$\zeta = \int h(\theta)f(\mathbf{x}|\mathbf{q}) \pi(\theta)d\theta / \int f(\mathbf{x}|\theta) \pi(\theta)d\theta .$$

The posterior density $\pi(\theta|x) = f(\mathbf{x}|\theta)\pi(\theta) / \int f(\mathbf{x}|\theta)\pi(\theta)d\theta$ can be used as a trial density. Sample from the posterior density to get the MC sample $\theta_1, \dots, \theta_M$, the variability criterion becomes

$$1 \propto h(\theta).$$

While $1 \propto \eta(\theta)$ requires that $\eta(\theta)$ is a constant, the variance reduction

condition is stable in the sense that it does not depend on n . By (3.1), the MC approximation to ζ is,

$$\zeta_M = (1/M) \sum_{1 \leq m \leq M} h(\theta_m)$$

The contribution of the prior distribution and the likelihood function are both accounted for during the course of sampling a posterior distribution. The resulting Monte Carlo estimate ζ_M depends on the data only thru the MC sample $\theta_1, \dots, \theta_M$. Set $h(\theta) = \theta^k$, the corresponding ζ_M is an approximation to the posterior moments $E[\theta^k | \mathbf{x}]$, $k=1, 2, \dots$. Set $h(\theta) = I_{\{a < \theta < b\}}$, ζ_M is an approximation to the posterior probability that θ is in the interval (a, b) ; in particular, $a = E[\theta | \mathbf{x}] - 2\sigma[\theta | \mathbf{x}]$, and $b = E[\theta | \mathbf{x}] + 2\sigma[\theta | \mathbf{x}]$ correspond to the posterior probability that $\theta \in E[\theta | \mathbf{x}] \pm 2\sigma[\theta | \mathbf{x}]$.

Remark 3.1. In the case that the prior density $\pi_0(\theta)$ is not a conjugate prior, the posterior expectation of $h(\theta)$ can be written as

$$\int h(\theta) \pi(\theta) \pi(\theta)^{-1} \pi(\theta | \mathbf{x}) d\theta / \int \pi(\theta)^{-1} \pi(\theta | \mathbf{x}) d\theta,$$

where $\pi(\theta | \mathbf{x})$ is the posterior density with respect to the conjugate prior $\pi(\theta)$. An iid sample $\theta_1, \dots, \theta_M$ from a trial density $\pi(\theta | \mathbf{x})$ results in an iid Monte Carlo weighted average approximation

$$\zeta_M = \left\{ \sum_{1 \leq m \leq M} h(\theta_m) \pi_0(\theta_m) \times \pi(\theta_m)^{-1} \right\} / \sum_{1 \leq m \leq M} \pi(\theta_m)^{-1}.$$

Sampling the likelihood density

The form of the posterior distribution may be complicated, and it is difficult to sample from it. However, sampling from a density that is "almost proportional" to the posterior density would do. That sampling accounts for the likelihood function $f(\mathbf{x} | \theta)$ is the key criterion for an efficient Monte Carlo approximation to posterior quantities since $f(\mathbf{x} | \theta)$ could be very peaky. If the factor $f(\mathbf{x} | \theta)$ is not accounted for during the course of sampling, the variance reduction criterion will be violated

eventually. With this in mind, one can sample from a "likelihood density"

$$f(\mathbf{x}|\theta) / \int f(\mathbf{x}|\theta)d\theta,$$

if the denominator is non zero. The variance reduction criteria for approximations to the denominator and numerator are

$$1 \propto h(\theta)\pi(\theta) \text{ and } 1 \propto \pi(\theta),$$

respectively. The variance reduction criteria are independent of the sample size n . Sample $\theta_1, \dots, \theta_M$ from a "likelihood density" $\propto f(\mathbf{x}|\theta)$, MC approximations to the numerator and denominator are $(1/M) \sum_{1 \leq m \leq M} h(\theta_m) \pi(\theta_m)$ and $\sum_{1 \leq m \leq M} \pi(\theta_m)$, respectively. A MC weighted average approximation to ζ is

$$\zeta_M \equiv \sum_{1 \leq m \leq M} h(\theta_m) \pi(\theta_m) / \sum_{1 \leq m \leq M} \pi(\theta_m).$$

Sampling the prior density.

Sampling the prior density usually does not fare well. The integrand for the numerator is $h(\theta)f(\mathbf{x}|\theta)\pi(\theta)$. Consider a MC approximation to ζ based on a sample $\theta_1, \dots, \theta_M$ from $\pi(\theta)$ is

$$\sum_{1 \leq m \leq M} h(\theta_m)f(\mathbf{x}|\theta_m) / \sum_{1 \leq m \leq M} f(\mathbf{x}|\theta_m).$$

The variance reduction criterion is $1 \propto h(\theta)f(\mathbf{x}|\theta)$ for the numerator alone, and it is increasingly violated as n grows since $f(\mathbf{x}|\theta)$, as a function of θ , often peaks at the maximum likelihood estimator.

4. Markov chain Monte Carlo approximations

The numerical results in the last chapter suggest that sampling from a posterior distribution is preferable. The Bayes theorem expresses a posterior distribution as a ratio, where the denominator is normalization constant. In practice, the normalization constant may be unknown, and sampling from a posterior distribution may not be possible. The Markov chain Monte Carlo method can be used to produce an approximate observation from this posterior distribution without knowledge of the normalization constant. This chapter discusses Markov chain Monte Carlo approximations to integrals.

Markov chain preliminaries. A sequence Y_0, Y_1, \dots of random variables (or vectors) is called a Markov chain if the following Markov property is satisfied:

Given Y_k : Y_0, \dots, Y_{k-1} and Y_{k+1}, \dots are independent.

By the product rule, the joint distribution of Y_0, Y_1, \dots can be determined by

(i) An initial distribution for Y_0 , $\pi(x)dx$,

And (ii) a conditional distribution of $Y_{n+1}|Y_n=x$, say, $k(y|x)dy$ for $n=1, \dots$

Note that we only consider the case that the conditional distribution in (ii) is independent of n and in this case $k(y|x)dy$ is called a stationary transition function (Markov kernel). Consider the first two terms of the Markov chain, Y_0 and Y_1 . The marginal distribution of Y_1 is $\pi_1(y)dy$ where

$$\pi_1(y) = \int k(y|x) \pi(x)dx.$$

Stationary distribution being a fixed point of the transition kernel.

The transition function takes a distribution $\pi(x)dx$ to another distribution $\pi_1(y)dy$. The marginal distributions of Y_0 and Y_1 , $\pi(x)dx$ and $\pi_1(x)dx$, are different. However, if $\pi(x)dx$ and $\pi_1(x)dx$ are identical,

$$(4.1) \quad \pi(y) = \int k(y|x) \pi(x)dx,$$

the distribution $\pi(x)dx$ is called a stationary distribution of the Markov chain. The word "stationary" is justified: By induction the marginal distributions of Y_n , $n \geq 0$ are all identically $\pi(x)dx$.

Subject to some "ergodic" conditions on $k(y|x)$, there is a unique stationary distribution $\pi(x)dx$ of $k(y|x)$. In addition the following law of large numbers prevails: Suppose $\zeta = \int h(x) \pi(x)dx$ is finite,

$$(4.2) \quad (1/M) \sum_{1 \leq m \leq M} h(Y_m) \text{ converges to } \zeta = \int h(x) \pi(x)dx \text{ almost surely.}$$

Similar to the iid case, the convergence in (4.2) is the backbone of the Markov chain Monte Carlo method.

The convergence does not depend on the marginal distribution of Y_0 . In practice, one can conveniently set Y_0 to a prescribed value y_0 (i.e., the marginal distribution of Y_0 is a point mass at y_0), which is then discarded in computing the average (4.2). Discarding an initial segment of the Markov chain before averaging (called warm-up or burn-in) often results in a better approximation to ζ .

The Markov chain sample $\{Y_m, m=1, \dots, M\}$ is not independent. The variance of the average (4.2) is

$$(1/M) \sum_{1 \leq i \leq M} [h(Y_i) - \zeta]^2 + (2/M^2) \sum \sum_{1 \leq i < j \leq M} [h(Y_i) - \zeta][h(Y_j) - \zeta].$$

The second sum is the correlation term that often affects the convergence in (4.2) substantially.

Equation (4.1) states that the distribution $\pi(x)dx$, being taken by the transition kernel $k(y|x)$ to itself, is a "fixed point" of this transition

kernel. A sufficient condition that a distribution $\pi(y)dy$ is the fixed point of a transition kernel $k(.|.)$ is a "reversibility" condition:

$$(4.3) \quad k(y|x) \pi(x) = g_y(x) \pi(y),$$

where for each y , $g_y(x)$ is a density in x . [The conditional density, $g_y(x)$, is given by the Bayesian theorem, and (4.7) is just a product rule for a joint density!] Integrating (4.3) with respect to dx yields (4.1).

There could be many transition kernels, $k(.|.)$, that admit the same $\pi(x)dx$ as a fixed point. Except in a few isolated cases, not much is known in determining which transition kernel provides the "best" Markov chain approximation (4.2).

Remark 4.1. (i) In case that density does not exist, (4.1) is to be interpreted as

$$(4.1') \quad \int I_{\{y \in B\}} \pi(y)dy = \int I_{\{y \in B\}} \int k(y|x) \pi(x)dx dy \text{ for all events } B,$$

or equivalently, $\pi(y)dy = \int k(y|x) \pi(x)dx$, almost all y (Lebesgue).

(ii) If $Y_n \in S$ that is a finite set, a sufficient condition for an ergodic kernel is that for any two fixed states x and y in the state space, starting from any x , with positive probability the Markov chain will reach $y \in S$, i.e., $P\{Y_n=y|Y_0=x\} > 0$ for some $n \geq 1$.

Summary of the Markov chain preliminaries. Suppose an integration with respect to a distribution $\pi(y)dy$ is required. The Markov chain Monte Carlo method concerns the construction of an "ergodic transition kernel" $k(y|x)dy$ so that $\pi(y)dy$ is the unique fixed point of it. Pick an initial value y_0 . A Markov chain sequence y_0, Y_1, \dots, Y_M can then be simulated as follows: Given the present value x , the next value is an observation from the conditional distribution $k(y|x)dy$, and repeat to get

$$y_0, Y_1, \dots, Y_L, Y_{L+1}, \dots, Y_{L+M}.$$

Discard the initial segment (of length L ; L is called a warm-up time) of the chain, the average of $h(y_{L+1}), \dots, h(y_{L+M})$

$$(1/M) \sum_{L+1 \leq m \leq L+M} h(Y_m) \text{ approximates } \zeta = \int h(y) \pi(y) dy.$$

4.1. The Gibbs sampler. The Gibbs sampler [see for example Geman and Geman (1984), IEEE] is constructed based on rather natural predictions between the multiple variables which defines the domain of the multiple integral. We first illustrate the Gibbs sampler concept in the two variable case. Suppose integration with respect to a bivariate density $\pi(x,y)$ proportional to a known $h(x,y)$ is desired. That is,

$$\pi(x,y)=h(x,y)/Z$$

where Z is a perhaps unknown normalization constant. The Markov chain Monte Carlo method concerns the construction of a conditional density of the next pair (x,y) given the present pair (x',y') , called a transition function $k(x,y|x',y')$, such that $\pi(x,y)$ is the fixed point/stationary density of $k(x,y|x',y')$. Typically there are different Markov chains that have this $\pi(x,y)$ as the stationary distribution (i.e., density). The Gibbs sampler is one such chain. The main advantage of the Gibbs sampler over the iid Monte Carlo is that the knowledge of the normalization constant Z is not required in identifying the two predictive densities.

The transition function of a Gibbs chain is defined through predictions between x and y , and is particularly appealing. To describe it, given the present value of the pair (x',y') . Select x and y consecutively to form the next pair (x,y) according to the following two predictive densities:

Step 1. $x|(x',y')$ has density $\pi_{x|y}(x|y') = C(y') \times h(x,y')$,

(viewed as a density in x .)

Step 2. $y|(x',y',x)$ has density $\pi_{y|x}(y|x) = D(x) \times h(x,y)$,

(viewed as a density in y .)

By the product rule, the conditional density of $(x,y)|(x',y')$ is then

$$(4.4) \quad k(x,y|x',y') = \pi_{x|y}(x|y') \times \pi_{y|x}(y|x).$$

[That is, x', y', x, y is a Markov chain in the order written.]

Note that

$$\begin{aligned} & k(x, y | x', y') \times \pi(x', y') \\ &= C(y') \times h(x, y') \times D(x) \times h(x, y) \times \pi(x', y') \\ &= C(y') \times h(x', y') \times D(x) \times h(x, y') \times h(x, y) / Z \\ &= \pi_{x|y}(x' | y') \times \pi_{y|x}(y' | x) \times \pi(x, y) \end{aligned}$$

Hence, $\iint k(x, y | x', y') \pi(x', y') dx' dy' = \pi(x, y)$.

That is, $\pi(x, y) dx dy$ is a fixed point of the transition kernel $k(x, y | x', y')$.

Pick an initial value (x_0, y_0) to simulate the next pair (X_1, Y_1) by going through Steps 1 and 2 (called a Gibbs cycle.) Use (X_1, Y_1) as the present value and go through the Gibbs cycle to get (X_2, Y_2) . Repeat to get a Markov chain sequence $(x_0, y_0), (X_1, Y_1), \dots, (X_L, Y_L), (X_{L+1}, Y_{L+1}), \dots, (X_{L+M}, Y_{L+M})$.

Example 4.1. Let us consider the gamma-Normal $(\alpha, 1/\beta; m, 1/v)$ conjugate prior when sampling from $N(\mu, 1/\tau)$ population. The posterior distribution of (τ, μ) is Gamma-Normal $(\alpha^*, 1/\beta^*; m^*, 1/v^*)$. Suppose it is desirable to approximate the posterior coefficient of variation $\iint \mu \tau^{1/2} \pi(\tau, \mu | x) d\tau d\mu$. Here $h(\tau, \mu) = \mu \times \tau^{1/2}$ in (4.2). We only need to find the predictive densities. But these are already given in Example 1.3 (vi):

$$\pi(\mu | x, \tau) \text{ is } N(m^*, 1/(v^* \tau)),$$

and $\pi(\tau | x, \mu) \text{ is Gamma } (\alpha^* + 1/2, 1/[\beta^* + v^*(\mu - m^*)^2/2])$.

These two predictive densities define a Gibbs cycle as follows: Suppose the present value is (τ', μ') . Conditional on the present value (τ', μ') , to select the next (τ, μ) , we go through the two prediction steps,

Step 1. $\tau | \mu'$ from Gamma $(\alpha^* + 1/2, 1/[\beta^* + v^*(\mu' - m^*)^2/2])$;

Step 2. $\mu | \tau$ from $N(m^*, 1/(v^* \tau))$.

Remark 4.2. The Gibbs sampler can be applied to sample a proper posterior distributions of more than one parameter variables. In previous chapters, we sometimes assume a flat prior "density" for the parameter and, combining with the data likelihood, to get a true posterior density. For example, in the previous example, starting with a Gamma-Normal $(\alpha, 1/\beta; m, 1/v)$ flat prior in which $\alpha=\beta=m=v=0$, the posterior is Gamma-Normal $(n/2, 1/\beta^*; m^*, 1/v^*)$ and a Gibbs cycle can be constructed. The posterior distribution is proper while the prior is "flat" and is not a proper density. This is a situation that one can use a Gibbs sampler to sample the posterior distribution, and not the "flat" prior that is not a distribution.

The Gibbs sampler for two variables extends to several variables as well. The case of three variables, where $\pi(x,y,z)=h(x,y,z)/Z$, suffices to illustrate the necessary modifications. Given the present value of the triple (x',y',z') , one goes through a 'prediction circle' of length three to get the next triple (x,y,z) , as follows:

$$\begin{aligned} x|y',z' \text{ has density } \pi(x|y',z') &\propto h(x,y',z'); & (y',z' \text{ are constants}) \\ y|x,z' \text{ has density } \pi(y|z',x) &\propto h(x,y,z'); & (x,z' \text{ are constants}) \\ z|x,y \text{ has density } \pi(z|x,y) &\propto h(x,y,z). & (x,y \text{ are constants}) \end{aligned}$$

In words, the 'prediction circle' is summarized by

'Each variable is simulated consecutively from the predictive density of it given the present values of all other variables.'

By the product rule, the conditional density of $(x,y,z)|(x',y',z')$ is

$$(4.5) \quad k(x,y,z|x',y',z') = \pi_{x|y',z'}(x|y',z') \times \pi_{y|z',x}(y|z',x) \times \pi_{z|x,y}(z|x,y).$$

Similar to the Gibbs sampler for two variables, integrating

$$k(x,y,z|x',y',z') \times \pi(x',y',z')$$

with respect to $dx'dy'dz'$ yields $\pi(x,y,z)$. Hence $\pi(.,.,.)$ is a fixed point of the transition kernel $k(x,y,z|x',y',z')$.

Example 4.2. Consider the two sample Normal model with two unknown means μ_1 and μ_2 , and common precision $\tau_1=\tau_2\equiv\tau$ that is unknown; see Example 1.4 (ii). Describe a Gibbs sampler for three variables to evaluate posterior integrals if the prior for τ is Gamma $(\alpha, 1/\beta)$, and $\pi(\mu_1, \mu_2|\tau) \equiv 1$ is 'flat'. The posterior density of (τ,μ_1,μ_2) is proportional to

$$\begin{aligned} & \tau^{\alpha-1} \exp\{-\beta\tau\} \times \tau^{(n_1+n_2)/2} \\ & \times \exp\{-\tau n_1(\mu_1-\hat{a}_x)^2/2\} \times \exp\{-\tau n_2(\mu_2-\hat{a}_y)^2/2\} \\ & \times \exp\{-\tau[\sum_i(x_i-\hat{a}_x)^2 + \sum_j(y_j-\hat{a}_y)^2]/2\}, \end{aligned}$$

where \hat{a}_x is the average of the x_i s and \hat{a}_y is the average of the y_j s. Let

$$\beta(\mu_1,\mu_2) = \beta + (1/2)[n_1(\mu_1-\hat{a}_x)^2 + n_2(\mu_2-\hat{a}_y)^2 + \sum_i(x_i-\hat{a}_x)^2 + \sum_j(y_j-\hat{a}_y)^2].$$

Suppose the present value is (τ',μ_1',μ_2') . The next value (τ,μ_1,μ_2) will result at the completion of a prediction cycle of length three. Inspecting the above expression for the properly defined posterior density, the prediction cycle is, conditional on (τ',μ_1',μ_2') ,

Step 1. $\tau|(\mu_1', \mu_2')$ from Gamma $(\alpha + (n_1+n_2)/2; 1/\beta(\mu_1',\mu_2'))$;

Step 2. $\mu_1|(\mu_2', \tau)$ from $N(\hat{a}_x, 1/(n_1\tau))$;

Step 3. $\mu_2|(\tau, \mu_1)$ from $N(\hat{a}_y, 1/(n_2\tau))$.

4.2. The Metropolis-Hastings algorithm. The Metropolis-Hastings algorithm is a general Markov chain Monte Carlo method for approximating single or multiple integrals. The method has many applications, and is originally due to Metropolis and etc. (1953, Jour, Chem. Physics) and Hastings (1970, Biometrika). It has the flexibility to allow for situations in which cross-predictions, which is the basis of a

Gibbs cycle, is a source of difficulties. An examination of the fixed point property of the Metropolis-Hastings algorithm are delayed to Section 4.4, and we shall discuss its applications in this section. Suppose an integral with respect to a distribution function $\pi(x)dx$ is to be approximated:

$$(4.6) \quad \zeta = \int h(x) \pi(x)dx.$$

The goal is to find a conditional distribution of $Y|X=x$ denoted by $k(y|x)dy$ so that both Y and X have identical (marginal) distribution $\pi(x)dx$. Equivalently, we look for a transition kernel $k(y|x)$ so that $\pi(x)dx$ is the fixed point of the kernel $k(y|x)$; see (4.1). The following Lemma 4.1 is the key to the construction of $k(y|x)dy$.

Lemma 4.1. Suppose X has distribution $\pi(x)dx$, and that $Y|X=x$ has a mixture distribution: Given $X=x$, $Y=x$ with probability $1-c(x)$; otherwise, Y has a (conditional) distribution $g_x(y)dy$. The fixed point condition (see Remark 4.1)

$$(4.7) \quad \int c(x) \pi(x) g_x(y) dx = c(y) \pi(y), \text{ almost all } y \text{ (Lebesgue),}$$

implies that Y has the same distribution as X .

Proof. We shall show $P\{Y \in B\} = P\{X \in B\}$ for all events B .

First,

$$P\{Y \in B | X = x\} = E[I_{\{Y \in B\}} | X=x] = [1-c(x)] I_{\{x \in B\}} + c(x) \int I_{\{t \in B\}} g_x(t) dt.$$

Next, integrating with respect to $\pi(x)dx$ to get

$$\begin{aligned} P\{Y \in B\} &= \int P\{Y \in B | X = x\} \pi(x) dx \\ &= \int [1-c(x)] I_{\{x \in B\}} \pi(x) dx + \int c(x) \int I_{\{t \in B\}} g_x(t) dt \pi(x) dx \\ &= \int [1-c(x)] I_{\{x \in B\}} \pi(x) dx + \int I_{\{t \in B\}} \int c(x) \pi(x) g_x(t) dx dt \\ &= \int [1-c(x)] I_{\{x \in B\}} \pi(x) dx + \int I_{\{t \in B\}} c(t) \pi(t) dt \\ &= \int I_{\{x \in B\}} \pi(x) dx = P\{X \in B\}. \quad \parallel \end{aligned}$$

A sufficient condition for the (4.7) is a reversibility condition

$$c(x) \pi(x) g_x(y) = c(y) \pi(y) g_y(x).$$

Note that we already pointed out in (4.3) that the $g_y(x)$ in right hand side of (4.8) can be replaced by any density in x , and the fixed point property remains.

To describe the Metropolis-Hastings algorithm from the point of Lemma 4.1, we consider three random variables X, Y^*, Y with a joint distribution:

- (i) X has a distribution $\pi(x)dx$,
- (ii) Given $X=x$, Y^* has a distribution $q(y^*|x)dy^*$, and
- (iii) Given $X=x$ and $Y^*=y^*$, $Y = y^*$ with probability $p(y^*,x)$; otherwise $Y=x$.

The conditional distribution of $Y|X=x$ can be determined as follows. Conditional on $X=x$ and $Y^*=y^*$,

$$\begin{aligned} P\{Y \in B | X = x, Y^* = y^*\} &= E[I_{\{Y \in B\}} | X = x, Y^* = y^*] \\ &= [1 - p(y^*, x)] \times I_{\{x \in B\}} + p(y^*, x) \times I_{\{y^* \in B\}}. \end{aligned}$$

Hence,

$$\begin{aligned} P\{Y \in B | X = x\} &= \int E[I_{\{Y \in B\}} | X = x, Y^* = y^*] q(y^*|x) dy^* \\ &= [1 - c(x)] I_{\{x \in B\}} + c(x) \int I_{\{t \in B\}} g(t|x) dt \end{aligned}$$

where $g(t|x) \equiv p(t,x)q(t|x)/c(x)$ and is a density in t .

According to Lemma 4.1, X and Y are symmetric if the kernel $g(t|x)$ has a fixed point $c(x) \pi(x)$:

$$\int c(x) \pi(x) g(y|x) dx = c(y) \pi(y), \text{ almost all } y \text{ (Lebesgue.)}$$

That is

$$(4.9) \quad \int \pi(x) p(y,x)q(y|x) dx = \int p(x,y)q(x|y) \pi(y) dx.$$

The classic reversibility condition is "the two integrands in (4.9) are

identical":

$$(4.10) \quad \pi(x) p(y,x)q(y|x) = p(x,y)q(x|y) \pi(y).$$

Possible choices of $p(y,x)$ that satisfy the reversibility condition:

(4.11)

a. $p(y,x) = \text{Min}\{[q(x|y) \pi(y)]/[q(y|x) \pi(x)], 1\}$. (Metropolis-Hastings.)

b. $p(y,x) = q(x|y) \pi(y)/[q(y|x) \pi(x)+q(x|y) \pi(y)]$.

Example 4.3. The M-H approximation to posterior quantities. This example discusses a transition function that admits a posterior distribution as a fixed point. Suppose an integral, say $\zeta = \int h(\theta) \pi(\theta|\mathbf{x})d\theta$, with respect to a (posterior) distribution

$$\pi(\theta|\mathbf{x})d\theta \propto f(\theta)q(\theta)d\theta$$

is to be approximated where $f(\theta)$ denotes $[f(\mathbf{x}|\theta)\pi(\theta)/q(\theta)]$, and $q(\theta)d\theta$ is a density function that can be sampled easily. A transition function $k(\theta|\theta')$ for the M-H algorithm, which admits $\pi(\theta|\mathbf{x})d\theta$ as a fixed point, can be defined as follows: Given θ' , select a θ^* from $q(\theta)$, and compute $p(\theta^*,\theta') \equiv \min \{f(\theta^*)/f(\theta'), 1\}$. Let $\theta = \theta^*$ with probability $p(\theta^*,\theta')$; otherwise $\theta = \theta'$.

How to pick the trial density $q(\theta)$? Here the target criterion and the variance reduction criterion give a clue: The trial density $q(\theta)$ has to have a support larger than $\pi(\theta|\mathbf{x})$, and that it is proportional to $\pi(\theta|\mathbf{x})$.

Example 4.4. Describe the Metropolis-Hastings algorithm for a testing problem for $f(\mathbf{x}|\mu)$ vs $f(\mathbf{x}|\nu)$. The data \mathbf{x} is given and fixed. Let us assume that $f(\mathbf{x}|\mu) < f(\mathbf{x}|\nu)$. The trial density $q(\theta)$ is such that $q(\mu)+q(\nu)=1$. Select a θ^* from $\{\mu, \nu\}$ where $\theta^*=\mu$ with probability $q(\mu)$.

- a. $\theta_0=\mu$: $\theta =\theta^*$ with probability 1 (Ascending from the lowest valley with certainty.)
- b. $\theta_0=\nu$: If $\theta^*=\nu$, $\theta =\theta^*$ with probability 1. If $\theta^*=\mu$, $\theta =\theta^*$ with probability $f(\mathbf{x}|\mu)/f(\mathbf{x}|\nu)$ [Descending the peak $\theta_0=\nu$ with probability $f(\mathbf{x}|\mu)/f(\mathbf{x}|\nu)$.]

It is interesting to view a Gibbs sampler from a Metropolis-Hastings MCMC viewpoint. It suffices to study a Gibbs sampler of three variables x,y and z . Consider the second step of a Gibbs cycle consisting of moving (x,y',z') to (x,y,z') ; the y value is simulated based on the predictive distribution

$$\pi(y|z',x) = \pi(x,y,z') / \pi(x,z') \quad (x, z' \text{ are constants.})$$

Denote the present value (x,y',z') by \mathbf{y}_0 , and the test value (x,y,z') by \mathbf{y}^* .

With an abuse of notation,

$$q(\mathbf{y}^*|\mathbf{y}_0) \text{ is } \pi(x,y,z') / \pi(x,z'), \pi(\mathbf{y}_0) \text{ is } \pi(x,y',z'),$$

$$q(\mathbf{y}_0|\mathbf{y}^*) \text{ is } \pi(x,y',z') / \pi(x,z'), \text{ and } \pi(\mathbf{y}^*) \text{ is } \pi(x,y,z').$$

Hence, $[q(\mathbf{y}^*|\mathbf{y}_0) \times \pi(x,y',z')] / [q(\mathbf{y}_0|\mathbf{y}^*) \times \pi(x,y,z')] = 1$,

implying $p(\mathbf{y}^*, \mathbf{y}_0) = \text{Min} \{1,1\}=1$.

4.3. A combination of Markov chains. The Gibbs sampler for two variables x and y suggests to form a Markov chain moving from the present pair to the next pair based on cross predictions between x and y . Often one or both of the two predictive densities are difficult to sample. A method based on combining two (or more) conditional Markov chains can be useful. The two Markov chains are defined by conditioning at y' and x in a line up of present and next values (x',y',x,y) as follows: Denote the conditional distribution of $x|y$ and $y|x$ by $\pi_y(x)$ and $\pi_x(y)$.

- (i) given y' , $\pi_{y'}(x)dx$ is a fixed point for the transition function $k_{y'}(x|x')$;
- (ii) given x , $\pi_x(y)dy$ is a fixed point for the transition function $h_x(y|y')$.

Consider a transition function for moving the pair (x',y') to (x,y) defined by

$$k(x,y|x',y') \equiv k_{y'}(x|x') \times h_x(y|y').$$

With obvious notation, the joint density of x and y is (product rule)

$$\pi(x,y) = \pi_x(y) \times \pi(x) = \pi_{y'}(x) \times \pi(y).$$

Proposition 4.1. $\pi(x,y)dxdy$ is a fixed point for the transition function $k(x,y|x',y')$.

Proof. The right hand side of the fixed point equation (4.1) is

$$\begin{aligned} & \iint k_{y'}(x|x') \times h_x(y|y') \times \pi_{y'}(x') \pi(y') dx' dy' \\ &= \int h_x(y|y') \times \left[\int k_{y'}(x|x') \pi_{y'}(x') dx' \right] \times \pi(y') dy' \\ &= \int h_x(y|y') \times [\pi_{y'}(x)] \times \pi(y') dy' && \text{[by (i)]} \\ &= \pi(x) \int h_x(y|y') \pi(y'|x) dy' \\ &= \pi(x) \pi_x(y) && \text{[by (ii)]} \\ &= \pi(x,y). \quad \parallel \end{aligned}$$

A Gibbs sampler is based on exact sampling from univariate predictive distributions given all other variables. In the case that some of these predictive distributions are difficult to sample, one can use an M-H chain instead. The case of two variables (x,z) suffices to illustrate the idea. A Gibbs sampler requires sampling from $\pi(x|z)dx$ and $\pi(z|x)dz$, and suppose that it is difficult to sample the latter. Given the present value (x',z') of the pair, To obtain the next pair, one first sample x from $\pi(x|z')$. Next is to sample a density in z

$$\pi(z|x) \propto \pi(x|z) \pi(z).$$

Conditional on x , we construct another transition function $k_x(z|z')$ which has $\pi(z'|x)dz' \equiv \pi_x(z')dz'$ as a fixed point. A Metropolis-Hastings Markov chain with $f(z) = \pi(x|z)$ as in Example 4.3 will do the job: Sample from the marginal density of z , $\pi(z)dz$, to get z^* , and compute $p(z^*) = \min \{ \pi(x|z^*)/\pi(x|z'), 1 \}$. (See Example 4.3.) Let $z=z^*$ with probability $p(z^*)$; otherwise $z=z'$. In completion of this cycle, one arrives at the new value (x,z) .

If the marginal density of z , say $\pi(z)$, is difficult to sample, it is necessary to choose a different trial density $q(z)$. In this case

$$p(z^*) = \min \{ [\pi(x|z^*)\pi(z^*)/q(z^*)]/[\pi(x|z') \pi(z')/q(z')], 1 \}.$$

Example 4.5. Consider a model with two variables (\mathbf{x},z) , where $\mathbf{x}=(x_1,x_2)$. z is Normal $(m,1/v)$, $x_1,x_2|z$ are independent and $x_i|z$ has density $\pi_i(x_i|z)$ which is a t-density with $df=\alpha_i$, location= z , and precision τ_i , $i=1,2$. Constants m, v, α_i 's, and τ_i 's are known. There are obvious difficulties in sampling from $\pi(z|\mathbf{x})$. (Why?) The Metropolis-Hastings algorithm can be used to bypass this difficulty. Suppose the present value is (\mathbf{x}',z') . Next is to simulate $\mathbf{x}|z_0$. This does not present difficulties as $x_1,x_2|z$ are independent t- random variables. To complete the prediction cycle, it remains to draw from $\pi(z|\mathbf{x})$. Here we need a Metropolis-Hastings step: Select a z^* from Normal $(m,1/v)$, and compute

$$p(z^*) = \min \{ \pi(\mathbf{x}|z^*)/\pi(\mathbf{x}|z'), 1 \} .$$

Let $z=z^*$ with probability $p(z^*)$; otherwise $z=z'$. At the completion of the Metropolis-Hastings step, we reach (\mathbf{x},z) .

Example 4.6. An AR(1) time series model. The data, y_1, \dots, y_n obey the

recursive relationship $y_t = \phi y_{t-1} + z_t$; z_t s are iid $N(0, 1/\tau)$; $t=1, \dots, n$; $y_0=0$; $|\phi| \leq 1$ (a stationary condition). The joint density for $\mathbf{y}=(y_1, \dots, y_n)$ is

$$f(\mathbf{y}|\phi, \tau) = \tau^{n/2} \exp\{-\tau[y_1^2 + \sum_{2 \leq t \leq n} (y_t - \phi y_{t-1})^2]/2\}.$$

As a function of ϕ , $f(\mathbf{y}|\phi, \tau)$ looks like a truncated Normal density restricted (truncated) to $|\phi| < 1$, suggesting that Gamma-truncated Normal (τ, ϕ) s are conjugate priors. For simplicity, we assume that the prior density for (ϕ, τ) is $\pi(\phi)$ (the τ component is 'flat'). The posterior density is

$$\pi(\phi, \tau|\mathbf{y}) \propto \pi(\phi) \tau^{n/2} \exp\{-\tau[y_1^2 + \sum_{2 \leq t \leq n} y_{t-1}^2 (y_t/y_{t-1} - \phi)^2]/2\}$$

Assume this conjugate prior where $\pi(\phi)$ is truncated Normal, $\phi|(\mathbf{y}, \tau)$ has a Normal density restricted (truncated) to $|\phi| < 1$, whereas $\tau|(\mathbf{y}, \phi)$ is Gamma. A two variable Gibbs sampler can be applied to evaluate posterior quantities.

If a non-conjugate prior is used,

$$\pi(\phi|\mathbf{y}, \tau) \propto \exp\{-\tau s(\phi - c)^2/2\} \times \pi(\phi),$$

where $c = \sum_{2 \leq t \leq n} y_t y_{t-1} / \sum_{2 \leq t \leq n} y_{t-1}^2$ and $s = \sum_{2 \leq t \leq n} y_{t-1}^2$. $\pi(\phi|\mathbf{y}, \tau)$ has a product form and a Metropolis-Hastings step can be used to sample ϕ from $\pi(\phi|\mathbf{y}, \tau)$. The data are summarized in $N(c, 1/(\tau s))$ which is then the preferred trial density (i.e., a variability criterion applied to the Markov chain setting. The trial density in the previous example, Normal $(m, 1/v)$, is not data dependent and the resulting Markov sample is less efficient.) The probability of accepting the next ϕ is $\min\{\pi(\phi)/\pi(\phi'), 1\}$.

4.4. The fixed point of a transition kernel. This section discusses sufficient conditions for a transition kernel $k(y|x)$ to have a fixed point $\pi(x)dx$. It would be convenient to discuss the fixed point in terms of two random (or vectors) say, X (present value) and Y (next value).

The main result below, Theorem 4.1, gives a sufficient condition to ensure that X and Y have an identical marginal distribution.

Definition 4.1. Let $g(y|x)dy$ be a conditional distribution (transition kernel) of $Y|X=x$. A function $a(x) \geq 0$ is a fixed point of the kernel $g(y|x)$ if

$$\int a(x) g(y|x) dx = a(y), \text{ for almost all (Lebesgue) } y;$$

that is, the above equality is to be interpreted as

$$\int I_{\{y \in B\}} \int a(x) f(y|x) dx dy = \int I_{\{y \in B\}} a(y) dy$$

for all event B ; see also Remark 4.1.

Theorem 4.1. Suppose X has distribution $\pi(x)dx$, and that $Y|X=s$ has a mixture distribution: Given $X=s$, with probability $1-c(s)$, Y has a (conditional) distribution $h_s(y)dy$; otherwise, Y has a distribution $g_s(y)dy$. The fixed point conditions

$$(4.12) \quad [1-c(s)] \times \pi(s) \text{ is a fixed point of the transition kernel } h_s(y)$$

$$\text{and} \quad c(s) \pi(s) \text{ is a fixed point of the transition kernel } g_s(y)$$

imply that X and Y have identical marginal distribution.

Proof Let $J=0$ denotes that $Y|X=s$ has distribution $h_s(y)dy$. Hence, $\Pr\{J=0|X=s\}$ be $1-c(s)$ and $\Pr\{J=1|X=s\} = c(s)$. By the double expectation formula,

$$\begin{aligned} \Pr\{Y \in B\} &= \int [\Pr(Y \in B | J=0, X=s) \Pr(J=0 | X=s) \\ &\quad + \Pr(Y \in B | J=1, X=s) \Pr(J=1 | X=s)] \pi(s) ds \\ &= \int \{ [1-c(s)] \times \int I_{\{y \in B\}} h_s(y) dy \pi(s) ds + \int c(s) \times \int I_{\{y \in B\}} g_s(y) dy \pi(s) ds \\ &= \int I_{\{y \in B\}} [1-c(y)] \pi(y) dy + \int I_{\{y \in B\}} \int c(y) \pi(y) dy \quad [\text{by (4.10)}] \\ &= \int I_{\{y \in B\}} \pi(y) dy = \Pr\{X \in B\}. \quad \parallel \end{aligned}$$

Lemma 4.1 follows from Theorem 4.1 by setting $h_x(y)dx$ to be $\delta_x(y)dy$,

which is a point mass at x and satisfies, for all $a(t) \geq 0$:

$$\int a(y) \delta_x(y) dy = a(x) \text{ for all } x.$$

That is, the point mass transition kernel $\delta_x(y) dy$ makes all functions $a(x)$ its fixed point in the sense of Definition 4.1.

Remark 4.3. The first fixed point property in (4.12) states that

$$\int [1-c(s)] \pi(s) h_s(y) ds = [1-c(y)] \pi(y).$$

This follows if a "reversibility condition" on densities is satisfied: [see (4.9)]

$$(4.13) \quad [1-c(s)] \pi(s) \times h_s(y) = [1-c(y)] \pi(y) \times h_y(s).$$

Similarly, a reversibility condition on densities

$$(4.14) \quad c(s) \pi(s) \times g_s(y) = c(y) \pi(y) \times g_y(s)$$

also yields the second fixed point property in (4.12).

Remark 4.4. If X and Y have identical marginal distribution, X and Y can be made to have a symmetric joint distribution in the sense that

$$\Pr\{X \in A \text{ and } Y \in B\} = \Pr\{X \in B \text{ and } Y \in A\}$$

for any two events A and B . To do this, just define a conditional distribution of $X|Y=x$ by $P\{X \in B|Y=x\} \equiv P\{Y \in B|X=x\}$ for all event B and x to get symmetric X and Y .

5. Model-based clustering

A model-based method for statistical clustering of n objects assumes that the numerical measurements, $\mathbf{x} = \{x_1, \dots, x_n\}$, of the n objects have a joint model density: Given a partition $\mathbf{p} = \{C_1, \dots, C_{n(\mathbf{p})}\}$ of the indices $\{1, \dots, n\}$ of the n objects, the measurements of the objects are modeled by a "classification likelihood" given \mathbf{p} has a product form

$$(5.1) \quad f(\mathbf{x}|\mathbf{p}) = \prod_{1 \leq j \leq n(\mathbf{p})} k(x_j, i \in C_j),$$

where $k(x_i, i \in \{1, \dots, n\})$ is a joint density of the data \mathbf{x} . The joint density $k(\cdot)$ yields a marginal density of $\{x_q, q \in C\}$, i.e., $k(x_q, q \in C)$, for each subset C of $\{1, \dots, n\}$. For $j \notin C$, with an abuse of notation, $k(x_j | x_q, q \in C)$ - denotes the predictive density of x_j given $\{x_q, q \in C\}$ evaluated at x_j .

Assume a prior density $\pi(\mathbf{p})$ on \mathbf{p} s, and compute a posterior distribution on partitions given data as

$$\pi(\mathbf{p}|\text{data}) \propto f(\mathbf{x}|\mathbf{p}) \times \pi(\mathbf{p}), \quad (\sum_{\mathbf{p}} \pi(\mathbf{p}|\text{data}) = 1.)$$

Here the posterior mode \mathbf{p}^* , i.e. $\pi(\mathbf{p}^*|\text{data}) = \max_{\mathbf{p}} \pi(\mathbf{p}|\text{data})$, could be used as a Bayesian estimator of the "true" partition. (There are difficulties in defining the posterior mean and posterior medium.) A conjugate prior is used to facilitate computations. A conjugate prior density of \mathbf{p} is of the product form

$$(5.2) \quad \pi(\mathbf{p}|g) \propto \prod_{1 \leq j \leq n(\mathbf{p})} g(C_j),$$

where $g(\cdot)$ is a nonnegative function defined on the collection of subsets of $\{1, \dots, n\}$. In this case, the posterior distribution $\pi(\mathbf{p}|\text{data})$ is also of the product form (5.2) in which g is updated to

$$(5.3) \quad \pi(\mathbf{p}|\text{data}) = \pi(\mathbf{p}|g^*) \propto \prod_{1 \leq j \leq n(\mathbf{p})} g^*(C_j),$$

where $g^*(C_j) = g(C_j) \times k(x_i, i \in C_j)$.

An inspection of (5.3) suggests that the search of the posterior mode \mathbf{p}^* ,

$$\pi(\mathbf{p}^*|\text{data}) = \max_{\mathbf{p}} \{\pi(\mathbf{p}|\text{data})\}$$

is an integer programming problem as the maximization is over the finite, but large, combinatoric set of partitions. [The maximum likelihood estimator is but a posterior mode with respect to a uniform prior!]

We shall prescribe Monte Carlo methods to approximate a posterior mode \mathbf{p}^* . One convenient way to locate \mathbf{p}^* is through sampling $\pi(\mathbf{p}|\text{data})$: If $\mathbf{p}_1, \dots, \mathbf{p}_M$ are iid from $\pi(\mathbf{p}|\text{data})$, a \mathbf{p}_k with a larger $\pi(\mathbf{p}_k|\text{data})$ will have a higher probability to be sampled. Compute $\pi(\mathbf{p}_k|\text{data})$; $k = 1, \dots, M$; the \mathbf{p}_k that gives the largest $\pi(\mathbf{p}_k|\text{data})$ is an approximation to \mathbf{p}^* . However, the exact (perfect) sampling [Propp and Wilson (1996)] from $\pi(\mathbf{p}|\text{data})$ does not seem to be available. A sequential seating algorithm that samples a partition from a distribution close to $\pi(\mathbf{p}|\text{data})$ is a possibility; its relative, the Gibbs sampler, is also a reasonable alternative.

Remark 5.1. The traditional model-based clustering considers the case that $n(\mathbf{p})$ (i.e., the number of clusters) is a known integer and $k(\cdot)$ is a multivariate-Normal density. The product over the $n(\mathbf{p})$ Normal component densities gives a "classification likelihood," which suggests the maximum likelihood estimation for the mean and covariance structure of the Normal densities. See Gordon, A. D. (1999). *Classification*, 2nd edition. Chapman & Hall.

5.1. The Chinese restaurant process.

A rather natural way to construct a distribution on partitions is through the "Chinese restaurant process." The Chinese restaurant process (CR) takes its name from a sequential seating procedure allegedly observed in a Bay Area Chinese restaurant by L. Dubins and J. Pitman.

Imagine customers 1,...,n arrive at a restaurant one after the other. The first customer one is seated at an unoccupied table. When the second customer arrives, with some probability he (i.e., or she) is seated at the table with customer one; otherwise a new empty table is set up for him. When customer three arrives, he may be seated with customer number one, or with customer number two (or with both, if customer two was seated with customer one), or a third table may be opened for him. The process continues until n customers are seated. Note that all customers, except for the first one to arrive, are seated randomly according to a seating probability.

The seating probability of a Chinese restaurant process with parameter $e_0 > 0$ is defined as follows. Suppose $i-1$ customers are seated. Customer i is seated at an empty table with probability proportional to e_0 ; otherwise the customer is seated at an occupied table with probability proportional to the number of individuals at that table.

After seating n customers, the resulting partition $\mathbf{p} = \{C_1, \dots, C_{n(\mathbf{p})}\}$ is called a Chinese restaurant process with parameter $e_0 > 0$. It has a probability mass function

$$(5.4) \quad \pi(\mathbf{p}) \propto \prod_{1 \leq j \leq n(\mathbf{p})} [e_0 \times (e_j - 1)!], \quad (\text{Product rule of probability})$$

where e_j is the number of customers in C_j , $j=1, \dots, n(\mathbf{p})$. This is of the product form (5.2) with $g(C_j) \equiv [e_0 \times (e_j - 1)!]$, suggesting that simulating a random partition from the prior (5.2) [and posterior (5.3)] is not that different from the sequential seating Chinese restaurant process. We shall see why the CR density has the product form in the next Section.

Note that the normalization constant has a compact form

$$\sum_{\mathbf{p}} \prod_{1 \leq j \leq n(\mathbf{p})} [e_0 \times (e_j - 1)!] = \Gamma(e_0 + n) / \Gamma(e_0).$$

In particular, it does not depend on the seating order of the n customers.

5.2. The sequential seating weighted Chinese restaurant process.

What is the key randomness component of the assortment of weighted Chinese restaurant processes? Following the sequential seating CR, we consider a sequential seating WCR approximation to $\pi(\mathbf{p}|g)$ first. (This exemplifies the method of *sequential importance sampler*; see also Remark 5.2 in the next Section.) To initiate the sequential seating, customer 1 is seated at an empty C_0 , resulting in partition $\mathbf{p}[1]$. Suppose the first $i-1$ customers are seated, resulting in partition $\mathbf{p}[i-1] = \{C_1, \dots, C_{n(\mathbf{p}[i-1])}\}$. The next customer, Customer i , will be seated at table C_j with probability proportional to the "predictive ratios"

$$g(\{i\}|C_j) \equiv g(\{i\} \cup C_j) \div g(C_j), j=0, \dots, n(\mathbf{p}[i-1]);$$

the normalization constant of the conditional probability at seating i is

$$(5.5) \quad \lambda(i-1) = \sum_{0 \leq j \leq n(\mathbf{p}[i-1])} g(\{i\}|C_j).$$

The process terminates after seating n customers, resulting in $\mathbf{p}[n]$.

The sequential build-up of a partition structure is that $\mathbf{p}[i-1]$ is a function of $\mathbf{p}[i]$, and the product rule gives the probability distribution of $\mathbf{p}[n]$, denoted concisely by $\mathbf{p} = \{C_1, \dots, C_{n(\mathbf{p})}\}$, as

$$(5.6) \quad q(\mathbf{p}|g) = [\prod_{1 \leq j \leq n(\mathbf{p})} g(C_j)] / [\prod_{1 \leq i \leq n} \lambda(i-1)], \lambda(0) \equiv 1.$$

The density $q(\mathbf{p}|g)$ is not exactly $\pi(\mathbf{p}|g)$. But the numerator of $q(\mathbf{p}|g)$ agrees with that of $\pi(\mathbf{p}|g)$, and in this sense $q(\mathbf{p}|g)$ is close (or proportional) to $\pi(\mathbf{p}|g)$ and is an "importance sampler". (See Remark 5.2 below.) The denominator $\prod_{1 \leq i \leq n} \lambda(i-1)$ depends on the seating order; this dependence seems to retard the closeness.

Letting g^* play the role of g in the seating algorithm, the resulting sequential WCR distribution is denoted by $q(\mathbf{p}|g^*)$, which is close to the posterior distribution $\pi(\mathbf{p}|g^*)$. Hence, if $\mathbf{p} = \{C_j, j = 1, \dots, n(\mathbf{p})\}$ results

after sequentially seating customers $1, \dots, i-1$, the seating probability of Customer i is proportional to

$$(5.7) \quad k(x_i|x_q, q \in C_j) \times g(\{i\}|C_j), j = 0, 1, \dots, n(\mathbf{p});$$

replacing g by g^* in $\lambda^{(i-1)}$ results in the normalization constant $\lambda^{*(i-1)}$.

[For an empty table C_0 , $k(x_i|x_q, q \in C_0)$ is defined to be $k(x_i)$.]

Sequentially seating the n customers repeatedly and independently to obtain $\mathbf{p}_1, \dots, \mathbf{p}_M$, the partition that maximizes $\{\pi(\mathbf{p}_k|g^*), k=1, \dots, M\}$ is an approximation to the posterior mode \mathbf{p}^* , that provides the required clustering of the n objects.

Example 5.1. A Chinese restaurant process with parameter e_0 corresponds to $g(C_j) = e_0 \times (e_j-1)!$, $j=1, \dots, n(\mathbf{p})$. After sequentially seating $i-1$ customers, the sequential seating prior probability for customer i is proportional to [Recall $g(C_0)=1$.]

$$g(\{i\}|C_j) \equiv g(\{i\} \cup C_j) \div g(C_j) = e_j, j=0, 1, \dots, n(\mathbf{p}).$$

Hence, the posterior sequential seating probability for customer i is proportional to

$$k(x_i|x_q, q \in C_j) \times e_j, j = 0, 1, \dots, n(\mathbf{p}).$$

This is the Chinese restaurant seating probability e_j multiplied by a data-driven predictive "weight" $k(x_i|x_q, q \in C_j)$. This is the origin of the name *weighted* Chinese restaurant process (WCR).

Remark 5.2. The (sequential) importance sampler means, literally, that the "important part" of the integrand is sampled in the course of the (sequential) sampling. This is but a restatement of the variance reduction criterion discussed in the iid Monte Carlo method Chapter. The sequential WCR sampling takes care of the "peaky" numerator of the

posterior (5.3). It produces normalization weights that are dealt with by the Monte Carlo weighted average. Numerical examples suggest that while the sequential importance sampler WCR performs satisfactorily. However, its relative, the Gibbs WCR discussed in the next section has a slight edge.

5.3. The Gibbs weighted Chinese restaurant process.

A Gibbs sampler version of the WCR, called Gibbs WCR, has $\pi(\mathbf{p}|\mathbf{g})$ as the stationary distribution. The general theory of Gibbs sampler states that the Markov transition consists of a Gibbs cycle which is based on the idea of "predicting a new variable given the rest of the variables and then rotating (cycling) the new variable among the existing ones for predictions." In the present situation, the prediction part could be read off from the last seating step in the sequential seating WCR, and the cycling part follows from rotating the role of n through $1, \dots, n$. To be precise, let i be an integer in $\{1, \dots, n\}$. Given a partition \mathbf{q} of $\{1, \dots, n\}$, remove the Customer i from \mathbf{q} to obtain a partition $\mathbf{p} = \{C_1, \dots, C_{n(\mathbf{p})}\}$ that partitions $\{1, \dots, i-1, i+1, \dots, n\}$. Note that the two partitions differ only at the table in \mathbf{q} that contains Customer r . Since \mathbf{p} is a function of \mathbf{q} , a distribution of \mathbf{q} induces a distribution of \mathbf{p} .

Lemma 5.1. Assume that \mathbf{q} partitions $\{1, \dots, n\}$ and $\pi(\mathbf{q}|\mathbf{g}) \propto [\prod_{1 \leq j \leq n(\mathbf{q})} g(C_j)]$. Remove i from \mathbf{q} to get $\mathbf{p} = \{C_1, \dots, C_{n(\mathbf{p})}\}$ which is a partition of $\{1, \dots, n\} - \{i\}$. \mathbf{p} is a function of \mathbf{q} and

$$(5.8) \quad \mathbf{q}|\mathbf{p} \text{ has probability proportional to } g(\{i\}|C_j), j=0, 1, \dots, n(\mathbf{p}),$$

$$\text{and } \pi(\mathbf{p}|\mathbf{g}) \propto \prod_{1 \leq j \leq n(\mathbf{p})} g(C_j).$$

Proof. (i) Customer i has to be in one of $C_0, C_1, \dots, C_{n(\mathbf{p})}$. Call that table C_r .

$$\begin{aligned}\pi(\mathbf{q}|\mathbf{g}) &\propto g(\{i\} \cup C_r) \times \prod_{1 \leq j \leq n(\mathbf{p}), j \neq r} g(C_j) \\ &\propto [g(\{i\} \cup C_r) / g(C_r)] \times \prod_{1 \leq j \leq n(\mathbf{p})} g(C_j),\end{aligned}$$

Given \mathbf{p} , the part $\prod_{1 \leq j \leq n(\mathbf{p})} g(C_j)$ is a constant, hence $\mathbf{q}|\mathbf{p}$ has probability distribution proportional to $g(\{i\} \cup C_r) \div g(C_r) \equiv g(\{i\}|C_r)$, $r=0,1,\dots,n(\mathbf{p})$.

||

The Gibbs WCR cycle of moving the current partition to the next partition is completed by cycling through the following two steps for $i=1,2,\dots,n$.

(5.9) Step 1. Delete integer i from (the table containing i in) the current partition \mathbf{q} to get a partition $\mathbf{p} = \{C_0, C_1, \dots, C_{n(\mathbf{p})}\}$.

Step 2. Reseat integer i to the table C_j , with a reseating probability proportional to $g(\{i\}|C_j)$, $j = 0,1,\dots,n(\mathbf{p})$.

The stationary of this Gibbs WCR chain is $\pi(\mathbf{p}|\mathbf{g})$.

Letting g^* play the role of g in the cycling algorithm, the Gibbs WCR has a reseating probability proportional to (5.7):

$$g^*(\{i\}|C_j) \equiv k(x_i|x_q, q \in C_j) \times g(\{i\}|C_j), j = 0,1,\dots,n(\mathbf{p}).$$

The stationary distribution of this Gibbs WCR chain is $\pi(\mathbf{p}|g^*)$.

Reseating the n customers cyclically. Repeat to obtain $\mathbf{p}_1, \dots, \mathbf{p}_M$, the partition that maximizes $\pi(\mathbf{p}_k|g^*)$ is an approximation to the posterior mode \mathbf{p}^* , and provides the required clustering of the n objects.

Remark 5.3. The following is a converse to Lemma 5.1. Let $\mathbf{p} = \{C_1, \dots, C_{n(\mathbf{p})}\}$ partitions $\{1, \dots, n\} - \{i\}$ and $\pi(\mathbf{p}|\mathbf{g}) \propto [\prod_{1 \leq j \leq n(\mathbf{p})} g(C_j)]$. Given \mathbf{p} , seat Customer i to $\{C_0, C_1, \dots, C_{n(\mathbf{p})}\}$ to get a \mathbf{q} which is a partition of $\{1, \dots, n\}$ according to the seating probability (5.8). Then $\pi(\mathbf{q}|\mathbf{g}) \propto [\prod_{1 \leq j \leq n(\mathbf{q})} g(C_j)]$.

5.4. A key predictive property of WCR clustering.

The use of a seating probability (5.7) to generate a random partition is of interest and deserves a careful examination. For the rest of the Section, we shall discuss the case of the Gibbs WCR, as the sequential seating WCR case is rather similar. The way a new customer, say i , is assigned to occupied table C_j s, reveals a clustering process for a set of data $\{x_i, i=1, \dots, n\}$ by means of predictive properties between measurements rather than the traditional one based on a distance function defined between (groups of) data. The "next" customer i is assigned to table C_j with seating probability proportional to (5.7). Identify the observation x_i with integer i , $i = 1, \dots, n$, and regard C_j as a cluster of data. The predictive weight $k(x_i|x_q, q \in C_j)$ is the value of a predictive density, conditional on $\{x_q, q \in C_j\}$, and evaluated at a new observation x_i . Note that the predictive density $k(x_i|x_q, q \in C_j)$ is large if i is close to C_j (that is, if x_i is close to $\{x_q, q \in C_j\}$); otherwise $k(x_i|x_q, q \in C_j)$ is small. Hence if i is close to C_j , the seating probability that it will be grouped into C_j is also large. The second ratio $g(\{i\} \cup C_j)/g(C_j)$ is a similar "predictive" quantity based on the prior. The product of the data driven $k(x_i|x_q, q \in C_j)$ and the prior predictive ratio provides a rather natural balancing effect to the random seating, and it also defines the closeness of the clustering process. The following example illustrates the balancing effect of the seating probabilities based on various prior $\pi(\mathbf{p}|g)$ s.

Example 5.1 (continued.) A Chinese restaurant process prior. Example 5.1 states that the seating probability for the posterior (5.3) is proportional to

$$k(x_i|x_q, q \in C_j) \times e_j, j = 0, 1, \dots, n(\mathbf{p}).$$

Here a short distance between x_i and $\{x_q, q \in C_j\}$, that is a large value of $k(x_i | x_q, q \in C_j)$, could be balanced out by a small table size e_j , yielding a seating probability of moderate size.

Example 5.2. The case of a uniform prior on \mathbf{p} gives another perspective of the seating probability. Let $g(\cdot)$ be the constant 1, $\pi(\mathbf{p} | g)$ becomes a uniform prior on the space of partitions. The prior predictive ratio at (5.4) vanishes, and the Gibbs cycle is entirely "data driven": The seating probability for the posterior (5.3) is proportional to the predictive densities evaluated at x_i .

$$k(x_i | x_q, q \in C_j), j = 0, 1, \dots, n(\mathbf{p}).$$

The resulting Gibbs sampler yields an approximation to a maximum likelihood estimator of the clustering partition.

Example 5.3. Clustering data (error measurements) from a symmetric and unimodal density (with a mode at the origin) that is "smooth."

Clustering of such data x_i s can be achieved by:

$$k(x_i, i \in C_j) \equiv \int \left\{ \prod_{i \in C_j} \tau^{1/2} \varphi(\tau^{1/2} x_i) \right\} \gamma(\tau) d\tau.$$

[Work in mixture models (next Chapters) shows that clustering by this $k(\cdot)$ is appropriate if the error x_i s are iid with a scale-mixture of Normal density.] The idea of conjugate priors suggests that $\gamma(\tau) d\tau$ is a Gamma (α, β) distribution and $\varphi(t)$ is a standard Normal density. Simple algebra gives

$$k(x_i, i \in C_j) = (2\pi)^{-e_j/2} \times \beta^\alpha / \Gamma(\alpha) \times (\beta_j^*)^{\alpha_j^*} / \Gamma(\alpha_j^*),$$

where $\alpha_j^* = \alpha + e_j/2$, $\beta_j^* = \beta + (1/2) \sum_{i \in C_j} x_i^2$, and $t_j^* = t + e_j$; e_j is the number of integers in C_j . The data-driven predictive density $k(x_i | x_q, q \in C_j)$ is a t -density with $df = 2\alpha_j^*$, location = 0, and precision $(\alpha_j^* / \beta_j^*) t_j^* / (t_j^* + 1)$,

evaluated at x_i . If e_j is large, $k(x_i|x_q, q \in C_j)$ is approximately a $N(0, \sum_{i \in C_j} x_i^2/e_j)$ density.

Example 5.4. Clustering data from an arbitrary density. In this case the data x_i s are modeled as iid with a density that is a location and scale mixture of Normals. Work in mixture models show that the following $k(\cdot)$ is appropriate:

$$k(x_i, i \in C_j) \equiv \int \left\{ \prod_{i \in C_j} \tau^{1/2} \varphi(\tau^{1/2}[x_i - \mu]) \right\} \gamma(\tau, \mu) d\tau d\mu,$$

where $\gamma(\tau, \mu) d\tau d\mu$ is the Gamma-Normal $(\alpha, 1/\beta; m, 1/t)$ distribution.

Simple algebra [Example 1.4 (vi)] shows that

$$k(x_i, i \in C_j) = (2\pi)^{-e_j/2} \times \beta^\alpha / \Gamma(\alpha) \times \Gamma(\alpha_j^*) / (\beta_j^*)^{\alpha_j^*},$$

where $\alpha_j^* = \alpha + e_j/2$, $\beta_j^* = \beta + (1/2) \{ \sum_{i \in C_j} (y_i - m_j^*)^2 + e_j(m - a_j)^2 / t_j^* \}$, $t_j^* = t + e_j$,

$m_j^* = (tm + a_j) / t_j^*$, and $a_j = \sum_{i \in C_j} x_i / e_j$ is the sample average at the j th table.

The predictive density $k(x_i|x_q, q \in C_j)$ is a t -density with $df = 2\alpha_j^*$, location $= m_j^*$, and precision $(\alpha_j^* / \beta_j^*) t_j^* / (t_j^* + 1)$, evaluated at x_i . If e_j is large, $k(x_i|x_q, q \in C_j)$ is approximately a $N(a_j, \sum_{i \in C_j} x_i^2/e_j)$ density.

The clustering of multivariate data amounts to a change of notation; see Example 1.4 (vii).

Model (5.1) can be fine-tuned to accommodate additional regression-typed parameter θ . Assume that the "error" has a density $k(\cdot)$.

The observed data are (x_i, y_i) , $i = 1, \dots, n$, and x_i is the "independent variable" that is regarded as given and fixed. The model assumption is, for each subset C_j of $\{1, \dots, n\}$,

$$\varepsilon_i \equiv y_i - h(\theta, x_i), \quad i \in C_j | (\mathbf{p}, \theta) \text{ have a joint density } k(\varepsilon_i, i \in C_j),$$

and $\pi(\mathbf{p}, \theta | \mathbf{g}) \propto \psi(\theta) \prod_{1 \leq j \leq n(\mathbf{p})} g(C_j, \theta)$.

$g(C_j, \theta)$ may or may not depend on θ , and $h(\theta, v)$ is a known function of (θ, v) . It follows then given \mathbf{y} , the joint posterior density of $(\mathbf{p}, \theta) | \mathbf{y}$ is

$$(5.8) \quad \pi(\mathbf{p}, \theta | \mathbf{y}) \propto \psi(\theta) \prod_{1 \leq j \leq n(\mathbf{p})} g^*(C_j, \theta),$$

where $g^*(C_j, \theta)$ denotes $k(y_i - h(\theta, x_i), i \in C_j) \times g(C_j, \theta)$. That is, a Markov chain that has a stationary distribution being the posterior distribution (5.8) is obtained by alternately sampling \mathbf{p} and θ using the combination of Markov chains method:

(i) Given θ , the density of \mathbf{p} is proportional to $\prod_{1 \leq j \leq n(\mathbf{p})} g^*(C_j, \theta)$; move the present \mathbf{p}' to the next \mathbf{p} by cycle through Steps 1 and 2 for $j=1, \dots, n$.

(ii) Given \mathbf{p} , sample a next θ from the density of $\theta | \mathbf{p}$, which is proportional to $\psi(\theta) \times \prod_{1 \leq j \leq n(\mathbf{p})} k(y_i - h(\theta, x_i), i \in C_j)$.

If sampling θ in step (ii) is difficult, we apply the combination of MCMCs method (Proposition 4.1) and run the Metropolis-Hastings algorithm to move the present θ' to the next θ .

Example 5.3 (continued.) Regression model with a smooth symmetric and unimodal error density. The y_i s are independent and $y_i = \theta^T x_i + \varepsilon_i$, where $\theta \equiv (\theta_0, \theta_1, \dots, \theta_p)^T$ is the vector of regression parameters, $x_i \equiv (1, x_{i1}, \dots, x_{ip})^T$ is the vector of predictors for the i th observation, and $\theta^T x_i$ denotes $\sum_{0 \leq k \leq p} \theta_k x_{ik}$. Suppose we use the following $k(\cdot)$ for clustering:

$$(5.10) \quad k(\varepsilon_i, i \in C_j | \theta) \equiv \int \left\{ \prod_{i \in C_j} \tau^{1/2} \varphi(\tau^{1/2} [y_i - \theta^T x_i]) \right\} \gamma(\tau) d\tau.$$

[Work in mixture models (next Chapters) shows that clustering by this $k(\cdot)$ is appropriate if the error ε_i s are iid with a scale-mixture of Normal density.] The idea of conjugate priors suggests that $\gamma(\tau) d\tau$ is a Gamma (α, β) distribution. Simple algebra gives

$$k(\varepsilon_i, i \in C_j | \theta) = (2\pi)^{-e_j/2} \times \beta^\alpha / \Gamma(\alpha) \times (\beta_j^*)^{\alpha_j^*} / \Gamma(\alpha_j^*),$$

where $\beta_j^* = \beta + (1/2) \sum_{i \in C_j} (y_i - \theta^T x_i)^2$, and $\alpha_j^* = \alpha + e_j/2$; e_j is the number of integers in C_j . Apply the combination of MCMCs method and run the

Metropolis-Hastings algorithm to move θ' to the next θ .

Example 5.4. Clustering AR time series data. Time series data are almost regression data and therefore should be treated similarly. Consider the AR(1) case. Given θ , $Y_i = \theta Y_{i-1} + \varepsilon_i$, $i=1,2,\dots$, $|\theta| \leq 1$, $Y_0 = 1$, and X_i are $\varepsilon_i = Y_i - \theta Y_{i-1}$, $i=1,2,\dots$ (given θ) are iid with a unimodal and symmetric error density. (The error density could be heavy-tailed, such as Cauchy.)

Following Example 5.3,

$$k(\varepsilon_i, i \in C_j | \theta) \equiv \int \left\{ \prod_{i \in C_j} \tau^{1/2} \varphi(\tau^{1/2} [y_i - \theta y_{i-1}]) \right\} \gamma(\tau) d\tau.$$

A vector θ covers several other time series models; but the treatment amounts to a change of notation.

6. Bayesian nonparametrics: Gamma typed process priors

6.1. The Gamma and Dirichlet random vectors and sequences. Roll a

die with b faces, and the probability that face j shows up is $p_j, j=1, \dots, b;$

$\sum_{1 \leq j \leq b} p_j = 1$. Repeatedly rolling the die n times and

$n_j =$ the number of times that face j shows up, $j=1, \dots, b$.

Let $\mathbf{n} \equiv (n_1, \dots, n_b)$, and $\theta \equiv (p_1, \dots, p_b) \in \Theta; \Theta = \{(p_1, \dots, p_b): \sum_{1 \leq j \leq b} p_j = 1\}$. The

joint density of $\mathbf{n}|\theta$ is

$$f(\mathbf{n}|\theta) = \prod_{1 \leq j \leq b} p_j^{n_j}.$$

This can be restated "statistically" as: $x_1, \dots, x_n | \theta$ are iid from a distribution with mass p_j on the points $a_j, j=1, \dots, b$. That is, x_1, \dots, x_n are independent and

$$P\{x_i = a_j | p_1, \dots, p_b\} = p_j, \quad j=1, \dots, b.$$

The density of x_i s given θ is $f(x_1, \dots, x_n | \theta) = \prod_{1 \leq j \leq b} p_j^{n_j}$,

where $n_j = \sum_{1 \leq i \leq n} I\{x_i = a_j\}, j=1, \dots, b$.

[This density is proportional to that obtained from multinomial sampling:

$n_1, \dots, n_b | (\theta, n)$ has a Multinomial ($n; p_1, \dots, p_b$) distribution.]

A Dirichlet $(\alpha_1, \dots, \alpha_b)$ prior ($\alpha_j > 0$) for θ

$$\pi(p_1, \dots, p_b) \propto \prod_{1 \leq j \leq b} p_j^{\alpha_j - 1} \quad (\text{on } p_j\text{s such that } \sum_{1 \leq j \leq b} p_j = 1)$$

results in a Dirichlet $(\alpha_1^*, \dots, \alpha_b^*)$ posterior; $\alpha_j^* = \alpha_j + n_j, j=1, \dots, b$. To

summary,

Proposition A. $x_1, \dots, x_n | \theta$ are iid and $P\{x_i = a_j | p_1, \dots, p_b\} = p_j, j=1, \dots, b$.

$$\pi(p_1, \dots, p_b) \propto \prod_{1 \leq j \leq b} p_j^{\alpha_j - 1}$$

implies

$$\pi(p_1, \dots, p_b | x_1, \dots, x_n) \propto \prod_{1 \leq j \leq b} p_j^{\alpha_j^* - 1}, \quad \alpha_j^* = \alpha_j + \sum_{1 \leq i \leq n} I\{x_i = a_j\}.$$

Can the prior to posterior analysis be generalized to the case of

infinitely many k a_s? We shall see that the answer is yes, and generally for points a_j in quite arbitrary spaces. The Dirichlet vector does not have independent components (since they sum up to one,) and it is easier to discuss them in terms of independent Gamma random variables.

Lemma A. Suppose y_1, \dots, y_b are independent Gamma $(\alpha_j; 1)$, $j=1, \dots, b$ rvs. Let $y_+ = \sum_j y_j$ and $\alpha_+ = \sum_j \alpha_j$. Then, (i) y_+ is Gamma $(\alpha_+; 1)$, (ii) $(y_1/y_+, \dots, y_b/y_+)$ is Dirichlet $(\alpha_1, \dots, \alpha_b)$, and (iii) y_+ is independent of $(y_1/y_+, \dots, y_b/y_+)$.

Proof. We shall show this result in the case that $b=2$; the case for general b can be handled similarly (using the multi-dimension Jacobian method).

Suppose u and v are Gamma $(\alpha; 1)$ and Gamma $(\beta; 1)$ random variables, respectively, and u and v are independent. Let $S=u+v$, $W=u/(u+v)$. S is Gamma $(\alpha+\beta; 1)$, W is Beta (α, β) , and S and W are independent. Note that $u=SW$, $v=S(1-W)$, and let the joint density of u, v be $f_{U,V}(u, v)$. Then,

$$f(s, w) = f_{U,V}(sw, s(1-w)) \times |J(u, v)|,$$

where $J(u, v)$ has rows $(\partial u / \partial s, \partial v / \partial s) = (w, 1-w)$ and $(\partial u / \partial w, \partial v / \partial w) = (s, -s)$.

Hence, $|J(u, v)| = | -sw - s(1-w) | = | -s | = s$, and

$$\begin{aligned} f(s, w) &\propto (sw)^{\alpha-1} e^{-sw} [s(1-w)]^{\beta-1} e^{-s(1-w)} s \\ &= s^{\alpha+\beta-1} e^{-s} w^{\alpha-1} (1-w)^{\beta-1}, \end{aligned}$$

which is a product of the stated Gamma and Beta density.

The proof of (ii) is similar (using a multi-dimension Jacobian,) \parallel

Example 6.1. (Updating by a change of measures.) In the setting of Lemma A, for any $h(\cdot) \geq 0$,

$$(i) \quad E[y_k h(y_1, \dots, y_b)] = \alpha_k \times E^\# [h(y_1, \dots, y_b)],$$

where under the expectation $E^\#$, y_1, \dots, y_b are independent Gamma $(\alpha_j^*; 1)$ rvs where $\alpha_j^* \equiv \alpha_j + I\{j=k\}$, $j=1, \dots, b$.

(ii) Let $p_j \equiv y_j/y^+$. $E[p_k h(p_1, \dots, p_b)] = (\alpha_k/\alpha_+) \times E^\#[h(p_1, \dots, p_b)]$,

where under the expectation $E^\#$, p_1, \dots, p_b is a Dirichlet $(\alpha_1^*, \dots, \alpha_b^*)$ vector, and $\alpha_j^* \equiv \alpha_j + I\{j=k\}$, $j=1, \dots, b$.

Exercise 6.1. (Updating by a change of random variables.) In the setting of Lemma A, for any $h(\cdot) \geq 0$,

(i) $E[y_k h(y_1, \dots, y_b)] = \alpha_k \times E[h(y_1^*, \dots, y_b^*)]$,

where $y_j^* \equiv y_j + u_j \times I\{j=k\}$, $j=1, \dots, b$, and u_j s are iid Gamma $(1; 1)$ rvs independent of the y_j s.

(ii) Let p_j denotes y_j/y^+ . $E[p_k h(p_1, \dots, p_b)] = (\alpha_k/\alpha_+) \times E[h(p_1^*, \dots, p_b^*)]$,

where $p_j^* \equiv y_j^*/(y_1^* + \dots + y_b^*)$, $y_j^* \equiv y_j + u_j \times I\{j=k\}$, $j=1, \dots, b$, and u_j s are iid Gamma $(1; 1)$ rvs independent of the y_j s.

Exercise 6.2. Partition the set $\{1, \dots, b\}$ to C_1, \dots, C_m .

(i) (Consistency property of a Gamma vector.) y_1, \dots, y_b are independent Gamma $(\alpha_j; 1)$, $j=1, \dots, b$. Then $\sum_{j \in C_1} y_j, \dots, \sum_{j \in C_m} y_j$ are independent Gamma $(\sum_{j \in C_1} \alpha_j, \dots, \sum_{j \in C_m} \alpha_j; 1)$ random variables.

(ii) (Consistency property of a Dirichlet vector.) (p_1, \dots, p_b) is Dirichlet $(\alpha_1, \dots, \alpha_b)$. Then $(\sum_{j \in C_1} p_j, \dots, \sum_{j \in C_m} p_j)$ is Dirichlet $(\sum_{j \in C_1} \alpha_j, \dots, \sum_{j \in C_m} \alpha_j)$

[Hint: (ii) follows from (i) and Lemma A.]

Exercise 6.3. Let $\theta \equiv (\mu_1, \dots, \mu_b)$ and $\mu_j > 0$, and $p_j \equiv \mu_j / \mu_+$. Suppose $n_1, \dots, n_b | \theta$ are independent Poisson (μ_j) s. Then μ_1, \dots, μ_b are independent Gamma $(\alpha_j; 1)$ s implies $\mu_1, \dots, \mu_b | (n_1, \dots, n_b)$ are independent Gamma $(\alpha_j^*; 1)$ s; $\alpha_j^* \equiv \alpha_j + n_j$.

It follows from the last statement and Lemma A that $p_1, \dots, p_b | (n_1, \dots, n_b)$ is Dirichlet $(\alpha_1^*, \dots, \alpha_b^*)$. This supplies another way of proving Proposition A. [The original proof uses the fact that $n_1, \dots, n_b | (\theta, n_+)$ is Multinomial $(n_+; p_1, \dots, p_b)$, and p_1, \dots, p_b is Dirichlet $(\alpha_1, \dots, \alpha_b)$.]

In a standard Bayesian prior to posterior updating, $x_1, \dots, x_n | \theta$ iid $F(x|\theta)$, θ is a parameter. A prior distribution $\pi(\theta)d\theta$ is assumed for θ , and the posterior distribution $\pi(\theta|x) d\theta$ completes the Bayesian updating. In nonparametric and semiparametric problems, the parameter of interest θ often has an infinite dimension. For example, the parameter could be the distribution function itself, $F = \{F(t), t \in (-\infty, +\infty)\}$, and this is the "completely nonparametric" case. The problem would be the construction of a prior distribution on the space of F s.

If F is discrete and supported by b known points $\{a_1, \dots, a_b\}$, the parameter F is determined by the size of jumps at a_j , i.e., $p_j = P\{X_1 = a_j | F\}$, $j=1, \dots, b$. That is sampling from F is like rolling a b -faced die with numbers a_1, \dots, a_b on it (a_j s being known). The Dirichlet vector distribution discussed is a conjugate prior; see Proposition A.

In the nonparametric case, F is entirely unknown. How to construct a prior on F ? An intuitive approach is to assume a discrete F , and let the number of a_j s grows to infinity. Sampling from F is like rolling a die which has infinitely many faces with numbers a_1, a_2, \dots on it. Hopefully the main results could be restated as

Proposition B. Suppose $x_1, \dots, x_n | \theta$ are iid with $\Pr\{x_i = a_j | \theta\} = p_j$, $j = 1, \dots$
 $\theta \equiv (p_1, p_2, \dots)$ is Dirichlet $(\alpha_1, \alpha_2, \dots)$ implies $(p_1, p_2, \dots) | \mathbf{x}$ is Dirichlet $(\alpha_1^*, \alpha_2^*, \dots)$, where $\alpha_j^* = \alpha_j + n_j$, $n_j = \sum_{1 \leq i \leq n} I\{x_i = a_j\}$, $j=1, 2, \dots$. (Note: there can be no more than n non-zero n_j s.)

Lemma B. Suppose y_1, y_2, \dots are independent Gamma $(\alpha_j; 1)$, $j=1, 2, \dots$ rvs and $\alpha_+ = \sum_j \alpha_j$ is finite. Let $y_+ = \sum_j y_j$. Then, (i) y_+ is Gamma $(\alpha_+; 1)$, (ii) $(y_1/y_+, y_2/y_+, \dots)$ is Dirichlet $(\alpha_1, \alpha_2, \dots)$, and (iii) y_+ is independent of $(y_1/y_+, y_2/y_+, \dots)$.

To validate these results, one has to discuss the meaning of a "Dirichlet $(\alpha_1, \alpha_2, \dots)$ sequence."

The problem is about the existence of a distribution function [here a Dirichlet $(\alpha_1, \alpha_2, \dots)$ distribution] on the infinite dimension parameter space

$$\{(p_1, p_2, \dots): p_j \geq 0, \sum_j p_j = 1\}.$$

The existence problem can be resolved by assuming

- (i) the existence of a sequence of independent Gamma $(\alpha_j; 1)$ y_j s,
- (ii) $\alpha_+ = \sum_j \alpha_j$ is finite.

Under these two assumptions, it is not difficult to convince oneself that $y_+ = \sum_j y_j$ is Gamma $(\alpha_+; 1)$. If so, to complete the construction, we just let

$$p_j = y_j / \sum_j y_j, j=1, 2, \dots (\sum_j p_j = 1.)$$

The parameters in Proposition B are the size of jumps at a_j , i.e., $p_j = P\{X=a_j | p_1, p_2, \dots\}$ $j=1, 2, \dots$. We could let the parameter space to be the sequence of partial sums $p_1, p_1+p_2, \dots, 1$ ($= \sum_j p_j$) rather than the sequence p_1, p_2, \dots . (The two sequences are one to one corresponding to each other.) One could view the sequence of partial sums as a distribution (i.e., rises from zero to one), and the p_1, p_2, \dots sequence a "density" function (sums up to one.) It turns out that for an extension to a completely nonparametric case (i.e., a_j s being also unknown), the distribution sequence is the appropriate one for generalization. The problem is then to

find a nonnegative stochastic process $F(t)$, $-\infty < t < \infty$, such that $F(t)$ rises from zero to one as t runs from $-\infty$ to ∞ . We shall use the construction dictated by Lemma A. That is, we shall construct a "Dirichlet process" from a "Gamma process"; the latter has an independence property and is easier to understand.

6.2. A Gamma process $\mu(t)$ with shape measure $\alpha(t)$, $-\infty < t < \infty$.

Suppose y_1, y_2, \dots are independent random variables. The y_j s are one to one corresponding to the sequence of partial sums y_1, y_1+y_2, \dots . A special property of the partial sum sequence is the following: Take any term, say, $y_1+y_2+\dots+y_r$, it is a sum of independent random variables. That is, the "increments" of the partial sum sequence are independent. With this in mind, the statement that

" y_j s, ... are independent Gamma $(\alpha_j; 1)$ s"

is equivalent to

"the partial sum sequence $\sum_{1 \leq j \leq k} y_j$, $k=1, \dots$ has independent increments and $\sum_{1 \leq j \leq k} y_j$ is Gamma $(\sum_{1 \leq j \leq k} \alpha_j; 1)$, $k=1, \dots$ "

A Gamma process is a continuous time extension to the last statement. First is a statement on "independent increment" process. A stochastic process $y(t)$, $-\infty < t < \infty$, is called an independent increment process if for any grips $t_1 < \dots < t_b$, the increments

$$y(t_2)-y(t_1), y(t_3)-y(t_2), \dots, y(t_b)-y(t_{b-1})$$

are independent. Next, a nondecreasing function $\alpha(t)$ plays the role of a nondecreasing sequence of partial sums $\sum_{1 \leq j \leq k} \alpha_j$, $k=1, 2, \alpha(-\infty)=0$.

Furthermore, " $\alpha(\infty)$ is finite" is the same as " α_+ is finite."

Unless stated otherwise, a nondecreasing function on $(-\infty, \infty)$ is

assumed to be right continuous. For a right continuous $h(t)$, the size of the jump of $h(t)$ at $t=a$ is $h(\{a\}) \equiv h(a)-h(a^-)$.

Remark 6.1. A function $h(t)$ is called right continuous if $\lim_{t \downarrow s} h(t) = h(s)$.

For such $h(\cdot)$, $h(s,t]$ represents the increment of $h(\cdot)$ on the grip $s < t$:

$$h(s,t] \equiv h(t) - h(s).$$

Definition 6.1. A Gamma process $\mu(t)$ with a nondecreasing shape function $\alpha(t)$, $-\infty < t < \infty$, is an independent increment process such that for each t , $\mu(t)$ is Gamma $(\alpha(t); 1)$. This is denoted concisely by $\mu \sim G(d\mu|\alpha)$, or μ is Gamma $(\alpha; 1)$.

Properties of a Gamma process $\mu(t)$ with a shape $\alpha(t)$

- (6.2) (i) nonnegative: $0 \leq \mu(t)$
(ii) nondecreasing: $s \leq t$ implies $\mu(s) \leq \mu(t)$
(iii) $\mu(-\infty) = 0$
(iv) $\mu(t) - \mu(s)$ is Gamma $(\alpha(t) - \alpha(s); 1)$, $s \leq t$.
(v) $\mu(t)$ is Gamma $(\alpha(s); 1)$ if $\alpha(t)$ is constant on $[s, t]$
(vi) $\alpha(\{a\}) > 0$, $\mu(\{a\})$ is Gamma $(\alpha(\{a\}); 1)$

Moment computation $\mu(t)$ is a Gamma process with shape $\alpha(t)$

$$E[\mu(t)] = \alpha(t), \text{Var}[\mu(t)] = \alpha(t),$$

$$E[\mu(s)\mu(t)] = E[\mu(s)^2] + E[\mu(s)] \times E[\mu(t) - \mu(s)], s \leq t$$

$$E[\mu(s)h(\mu(t) - \mu(s))] = \alpha(s) \times E[h(\mu(t) - \mu(s))], s \leq t$$

Updating (change of measures). Suppose $h(\cdot) \geq 0$.

$$E[\mu(s) \times h(\mu(s), \mu(\infty) - \mu(s))] = \alpha(s) \times E^\# [h(\mu(s), \mu(\infty) - \mu(s))],$$

where under the expectation $E^\#$; $\mu(s)$ is Gamma $(\alpha(s)+1, \alpha(\infty) - \alpha(s))$.

6.3. A Dirichlet process $F(t)$ with shape $\alpha(t)$, $-\infty < t < \infty$.

A Gamma process $\mu(t)$ with shape function $\alpha(t)$ is nondecreasing and if $\alpha(\infty) [\equiv \sup_t \alpha(t)]$ is finite, we define $\mu(\infty)$ to be Gamma $(\alpha(\infty); 1)$. In this case, analogous to the sequence case, $\mu(t)$, $-\infty < t < \infty$, defines another stochastic process $F(t)$ that is a (random) distribution function rising from 0 to 1:

$$F(t) \equiv \mu(t) / \mu(\infty), \quad -\infty < t < \infty.$$

$F = \{F(t), -\infty < t < \infty\}$ is called a Dirichlet process with shape $\alpha = \{\alpha(t), -\infty < t < \infty\}$, denoted concisely by $F \sim D(dF|\alpha)$, or F is Dirichlet (α) .

Remark 6.2. (i) $F(t)$, $-\infty < t < \infty$, is the continuous time analogy to the discrete time partial sum sequence $p_1 + \dots + p_k$, $k=1, 2, \dots$

(ii) The definition of $\mu(t)$ and the corresponding $F(t)$ extends to include $t = \infty$; see Section 6.5 below.

Example 6.2. F is Dirichlet (α) .

(i) $F(t)$ is a Beta $(\alpha(t), \alpha(\infty) - \alpha(t))$ random variable.

(ii) The "increments" of F is a Dirichlet vector: Let $-\infty = t_0 < t_1 < \dots < t_b = \infty$ be grips in $(-\infty, \infty]$, the increments $(F(t_0, t_1], F(t_1, t_2], \dots, F(t_{b-1}, t_b])$ is a Dirichlet $(\alpha(t_0, t_1], \alpha(t_1, t_2], \dots, \alpha(t_{b-1}, t_b])$ vector.

(iii) The "tails" of F , $F(t_0, t_1], F(t_1, t_2]/F(t_1, t_b], \dots, F(t_{b-2}, t_{b-1}]/F(t_{b-2}, t_b]$, are independent Beta $(\alpha(t_{j-1}, t_j], \alpha(t_j, t_b))$, $j=1, \dots, b-1$ rvs.

Properties of a Dirichlet process.

From the properties of the corresponding Gamma process, a Dirichlet process F with shape α inherits similar properties

(6.3) (i) nonnegative: $0 \leq F(t)$

(ii) nondecreasing: $s \leq t$ implies $F(s) \leq F(t)$

- (iii) $F(-\infty)=0$ and $F(\infty)=1$
- (iv) $F(t)-F(s)$ is Beta $(\alpha(t)-\alpha(s), \alpha(\infty)-[\alpha(t)-\alpha(s)])$
- (v) $F(t)$ is Beta $(\alpha(s), \alpha(\infty)-\alpha(s))$ if α is constant on $[s,t]$
- (vi) $\alpha(t)$ jumps at $t=a$ with size $\alpha(\{a\})>0$,

$$F(\{a\}) \text{ is Beta } (\alpha(\{a\}), \alpha(\infty)-\alpha(\{a\}))$$

Moment computation. F is Dirichlet (α) .

$$E[F(t)] = \alpha(t)/\alpha(\infty), \text{Var}[F(t)] = [\alpha(t)[\alpha(\infty)-\alpha(t)]/[\alpha(\infty)^2(\alpha(\infty)+1)];$$

$$E[F(s)F(t)] = E[F(s)^2] + E[F(s)\times(F(t)-F(s))], s \leq t.$$

Updating (change of measures). Suppose $h(\cdot)$ is nonnegative,

$$E[F(s)\times h(F(s), 1-F(s))] = [\alpha(s)/\alpha(\infty)] \times E^\# [h(F(s), 1-F(s))],$$

where under the expectation $E^\#$, $F(s)$ is Beta $(\alpha(s)+1, \alpha(\infty)-\alpha(s))$.

Example 6.3. (Updating/Change of measures.) Suppose $h(\cdot) \geq 0$.

(i) $\mu(t)$, $-\infty < t < \infty$ is a Gamma process with shape function $\alpha(t)$. For $s < t < u$

$$\begin{aligned} E[\{\mu(t) - \mu(s)\} \times h(\mu(s), \mu(t)-\mu(s), \mu(u)-\mu(t))] \\ = [\alpha(t)-\alpha(s)] \times E^\# [h(\mu(s), \mu(t)-\mu(s), \mu(u)-\mu(t))], \end{aligned}$$

where under the expectation $E^\#$, $\mu(s)$, $\mu(t)-\mu(s)$, $\mu(u)-\mu(t)$ are independent Gamma random variables with parameters $\alpha(s)$, $\alpha(t)-\alpha(s)+1$, $\alpha(u)-\alpha(t)$.

(ii) $F(t)=\mu(t)/\mu(\infty)$, $-\infty < t < \infty$ is a Dirichlet process with shape $\alpha(t)$. For $s < t < u$,

$$\begin{aligned} E[\{F(t)-F(s)\} \times h(F(s), F(t)-F(s), 1-F(t))] \\ = \{[\alpha(t)-\alpha(s)]/\alpha(\infty)\} \times E^\# [h(F(s), F(t)-F(s), 1-F(t))], \end{aligned}$$

where under the expectation $E^\#$, $F(s)$, $F(t)-F(s)$, $1-F(t)$ is a Dirichlet $(\alpha(s), \alpha(t)-\alpha(s)+1, \alpha(\infty)-\alpha(t))$ vector.

6.4. Bayesian nonparametric: Gamma and Dirichlet process priors.

Bayesian statistics is on the use of a posterior distribution to do statistical inference. From a technical viewpoint, the posterior distribution is a solution to a Fubini-type theorem. To describe it, suppose $\theta \sim \pi(d\theta)$, $x|\theta \sim F(dx|\theta)$. (x and θ could be multi-dimensional.) The pair $\pi(d\theta)$ and $F(dx|\theta)$ define the marginal distribution of x , $Q(t)$, $-\infty < t < \infty$ by

$$Q(t) = \int I\{x \leq t\} Q(dx) = \Pr\{x \leq t\} = \iint I\{x \leq t\} F(dx|\theta) \pi(d\theta).$$

A posterior distribution of θ given x , $\pi(d\theta|x)$, is defined by a change of the order of integration formula.

Definition 6.2. A posterior distribution $\pi(d\theta|x)$ is the conditional distribution of $\theta|x$ satisfying the following Fubini-typed theorem: For any nonnegative function $h(\cdot)$,

$$(6.4) \quad \iint h(x, \theta) F(dx|\theta) \pi(d\theta) = \iint h(x, \theta) \pi(d\theta|x) Q(dx)$$

that could be stated concisely as

$$F(dx|\theta) \pi(d\theta) = \pi(d\theta|x) Q(dx).$$

This equality is the usual double expectation

$$E\{E[h(x, \theta)|\theta]\} = E\{E[h(x, \theta)|x]\}.$$

Remark 6.3. Standard arguments in Lebesgue integral states that

(i) To ensure that $\pi(d\theta|x)$ satisfies (6.4), it suffices to check (6.4) for all $h(x, \theta)$ s of the form $h(x, \theta) = I\{x \in A, \theta \in B\} \equiv I\{(x, \theta) \in A \times B\}$: Since then (6.4) is satisfied for $h(x, \theta)$ s of the form $\sum_j w_j I\{(x, \theta) \in I_j\}$ I_j s are of the form $A \times B$, and $w_j \geq 0$. A limiting argument concludes (6.4) for all $h(x, \theta) \geq 0$.

(ii) If $\pi(d\theta|x)$ satisfies (6.4), it satisfies (6.4) for all $h(x, \theta)$ s that are "quasi-integrable": For any function $h(x, \theta)$, write it as a sum of the positive part and negative part $h = h^+ - h^-$, where $0 \leq h^+ = h(x, \theta) I\{h \geq 0\}$ and $0 \leq h^- = -$

$h(x,\theta)I\{h<0\}$ are both nonnegative functions. Note that $h=h^+-h^-$. Since both h^+ and h^- are nonnegative,

$$\iint h^+ F(dx|\theta)\pi(d\theta) = \iint h^+ \pi(d\theta|x)Q(dx),$$

and
$$\iint h^- F(dx|\theta)\pi(d\theta) = \iint h^- \pi(d\theta|x)Q(dx).$$

If one of the two equalities is finite (called such $h(\cdot)$ s quasi-integrable,) one can subtract (i.e., there is no " $\infty-\infty$ " possibility) to conclude

$$\iint (h^+-h^-) F(dx|\theta)\pi(d\theta) = \iint (h^+-h^-) \pi(d\theta|x)Q(dx).$$

Suppose $F(x|\theta)$ is $F(x)$ itself, i.e., the distribution function $F(\cdot)$ itself is the parameter θ . The model becomes: $x|F \sim F$, i.e.,

$$P\{x \leq t|F\} = F(t), \text{ all } t.$$

A prior on the space of F s, $\pi(dF)$ [another notation for it is $d\pi(F)$,] is assumed for F , and the posterior distribution of $F|x$, $\pi(dF|x)$, completes the Bayesian updating. If the model density $f(x)=dF(x)/dx$ exists, the posterior distribution

$$\pi(dF|x) \propto f(x) \pi(dF) \quad [= f(x) \pi(dF) / \int f(x) \pi(dF).]$$

is a ratio. Otherwise, Definition 6.2 is used to identify the $\pi(dF|x)$ that satisfies, for any nonnegative function $h(x,F)$ of x and F ,

$$\iint h(x,F)F(dx) \pi(dF) = \iint h(x,F) \pi(dF|x)Q(dx),$$

where
$$Q(t) = \Pr\{x \leq t\} = E[P\{x \leq t|F\}] = E[F(t)], \text{ all } t$$

is the marginal distribution of x . The question is what prior distribution $\pi(dF)$ leads to a tractable posterior distribution $\pi(dF|x)$?

Example 6.4. Let us look at the case that θ is the distribution function $\theta=F(\cdot)$ in the categorical model Proposition A. In this case F only increases in jumps at known locations a_1, \dots, a_b with $P\{x=a_j|p_1, \dots, p_b\} = p_j$, $j=1, \dots, b$ being the sizes of the jumps of $F(t)$. As the a_j s are known,

$F = \{(p_1, p_1 + p_2, \dots, 1)\}$, and (p_1, \dots, p_b) is a Dirichlet $(\alpha_1, \alpha_2, \dots, \alpha_b)$ vector. That is, $F \sim D(dF|\alpha)$ where the shape function $\alpha(t)$ is a step function and jumps up at the a_j s:

$$\alpha(t) = \sum_{1 \leq j \leq b} \alpha_j \delta_{a_j}(t),$$

where $\delta_y(t) \equiv I\{y \leq t\}$. (A distribution in t with a point mass at y .) Using this notation, the translation of Proposition A in the case of one observation ($n=1$) is particularly simple: For all nonnegative $h(\cdot)$ s,

$$\iint h(x, F) F(dx) D(dF|\alpha) = \iint h(x, F) D(dF|\alpha + \delta_x) Q(dx),$$

where the marginal distribution $Q(t) = \alpha(t)/\alpha(\infty)$. This result

$$"F \sim D(dF|\alpha) \text{ and } x|F \sim F(dx) \text{ imply } F|x \sim D(dF|\alpha + \delta_x)"$$

is freed from the a_j s. Therefore, it is reasonable to guess that it could be true in general. Since we have guessed the posterior distribution of $F|x$ as a Dirichlet process with shape $\alpha + \delta_x$, we could work on a formal verification of this result. Namely, we need to check that $\pi(dF|x) = D(dF|\alpha + \delta_x)$ satisfies equation (6.4).

Lemma 6.1. For any nonnegative function $h(x, F)$,

$$(6.5) \quad \iint h(x, F) F(dx) D(dF|\alpha) = \iint h(x, F) D(dF|\alpha + \delta_x) \alpha(dx) / \alpha(\infty).$$

Proof. It suffices to check the equality for $h(\cdot)$ s of the form

$I\{x \leq t\} \times I\{F \in B\}$, where B is a (measurable) set of F s. Since finite

dimensional distributions determines the (infinite dimensional)

distribution of a stochastic process, the distribution of F is determined by the distribution of its finite dimensional sets of the form

$$(6.6) \quad B = \{F: F(I_1) \leq z_1, \dots, F(I_b) \leq z_b\}, \quad [\text{Note: } \sum_{1 \leq j \leq b} F(I_j) = 1.]$$

where $I_j = (t_{j-1}, t_j]$, $j = 1, \dots, b$ are adjacent intervals partitioning $(-\infty, \infty]$ and

$F(I_j)$ denotes $F(t_j) - F(t_{j-1})$. It suffices to check the equality for $h(\cdot)$ of the

form

$$h(x,F)=I_{\{x \leq t\}} \times I_B.$$

Since I_j s form a partition of $(-\infty, \infty]$, the number t must be in one of the I_j s.

Suppose

$$t \text{ is in } I_k = (t_{k-1}, t] + (t, t_k] = I_{k'} + I_{k''}. \quad [\text{Hence } F(I_k) = F(I_{k'}) + F(I_{k''}).]$$

The left hand side of Lemma 6.1 is

$$\begin{aligned} & \iint I_{\{x \leq t\}} \times I_B F(dx) D(dF|\alpha) \\ &= \iint I_{\{x \leq t\}} F(dx) I_B D(dF|\alpha) \\ &= \int F(t) I_B D(dF|\alpha) \\ &= \int \{ \sum_{1 \leq j \leq b} F((-\infty, t] \cap I_j) \} I_B D(dF|\alpha) \quad (I_j \text{ s form a partition.}) \\ &= \int \{ [\sum_{1 \leq j \leq k-1} F(I_j)] + F(I_{k'}) \} I_B D(dF|\alpha) \\ &= \{ \sum_{1 \leq j \leq k-1} [\alpha(I_j)/\alpha(\infty)] E_j^\# [I_B] \} + [\alpha(I_{k'})/\alpha(\infty)] E_{k'}^\# [I_B], \end{aligned}$$

where for $j=1, \dots, k-1$, under $E_j^\#$, $\{F(I_j), F(I_m), m \neq j\}$ is a Dirichlet $(\alpha(I_j)+1, \alpha(I_m), m \neq j)$ vector, and under $E_{k'}^\#$, $\{F(I_{k'}), F(I_{k''}), F(I_m), m \neq k\}$ is a Dirichlet $(\alpha(I_{k'})+1, \alpha(I_{k''}), \alpha(I_m), m \neq k)$ vector. See Example 6.3 (ii).

The right hand side of (6.5) is

$$\begin{aligned} & \iint I_{\{x \leq t\}} I_B D(dF|\alpha + \delta_x) \alpha(dx)/\alpha(\infty) \\ &= \sum_{1 \leq j \leq b} \int I_{\{x \in (-\infty, t] \cap I_j\}} \int I_B D(dF|\alpha + \delta_x) \alpha(dx)/\alpha(\infty) \\ &= \{ \sum_{1 \leq j \leq k-1} \int I_{\{x \in I_j\}} \int I_B D(dF|\alpha + \delta_x) \alpha(dx)/\alpha(\infty) \\ & \quad + \int I_{\{x \in I_{k'}\}} \int I_B D(dF|\alpha + \delta_x) \alpha(dx)/\alpha(\infty) \} \\ &= \{ \sum_{1 \leq j \leq k-1} \int I_{\{x \in I_j\}} E_j^\# [I_B] \alpha(dx)/\alpha(\infty) \} \\ & \quad + \int I_{\{x \in I_{k'}\}} E_{k'}^\# [I_B] \alpha(dx)/\alpha(\infty); \end{aligned}$$

the last equality follows from the fact that if $x \in I$, $(\alpha + \delta_x)(I) = \alpha(I) + 1$.

Hence

$$I_{\{x \in I_j\}} \times \int I_B D(dF|\alpha + \delta_x) = I_{\{x \in I_j\}} \times E_j^\# [I_B]$$

and $I_{\{x \in I_{k'}\}} \times \int I_B D(dF|\alpha + \delta_x) = I_{\{x \in I_{k'}\}} \times E_{k'}^\# [I_B]. \quad \parallel$

Example 6.6. Suppose F is Dirichlet with shape α .

$$E[\int g(x)F(dx)] = \int g(x) \alpha(dx)/\alpha(\infty)$$

$$E[\int g(x)F(dx)] = \iint g(x)F(dx)D(dF|\alpha)$$

$$= \iint g(x)D(dF|\alpha+\delta_x) \alpha(dx)/\alpha(\infty) \quad (\text{Lemma 6.1.})$$

$$= \int g(x) \int D(dF|\alpha+\delta_x) \alpha(dx)/\alpha(\infty) = \int g(x) \alpha(dx)/\alpha(\infty).$$

$$\text{Var}\{\int g(x)F(dx)\}$$

$$= [\alpha(\infty)+1]^{-1} \times \{ \int g(x)^2 \alpha(dx)/\alpha(\infty) - [\int g(x) \alpha(dx)/\alpha(\infty)]^2 \}$$

$$\text{Var}\{\int g(x)F(dx)\} = E[\int g(x)F(dx)]^2 - \{E[\int g(x) F(dx)]\}^2.$$

$$E[\int g(x)F(dx)]^2 = \int \int [g(x) \int g(y)F(dy)] F(dx) D(dF|\alpha)$$

$$= \int g(x) \int \int g(y) F(dy) D(dF|\alpha+\delta_x) \alpha(dx)/\alpha(\infty)$$

$$= [\alpha(\infty)+1]^{-1} \int g(x) \int g(y) \int D(dF|\alpha+\delta_x+\delta_y) (\alpha+\delta_x)(dy) \alpha(dx)/\alpha(\infty)$$

$$= [\alpha(\infty)+1]^{-1} \int g(x) \int g(y) (\alpha+\delta_x)(dy) \alpha(dx)/\alpha(\infty)$$

$$= [\alpha(\infty)+1]^{-1} [\int g(x) \alpha(dx) \int g(y) \alpha(dy) + \int g(x)^2 \alpha(dx)]/\alpha(\infty);$$

the 2nd equality is from applying Lemma 6.1 to $F(dx) D(dF|\alpha)$, and the 3rd equality is from applying Lemma 6.1 (for a fixed x) to $F(dy) D(dF|\alpha+\delta_x)$.

Example 6.7. A Dirichlet process F with shape α is discrete with probability one. That is, it increases only in jumps with probability one. The situation is complicated as both the sizes and locations of the jumps could be random; the book on Poisson Processes by Kingman (1993) discusses Blackwell's proof of this result, as well as other references. Lemma 6.1 gives a simpler proof. Each distribution F has a set of "atoms,"

$$A_F = \{x: F(\{x\}) > 0\}. \quad (\text{Note: } x \in A_F \text{ iff } F(\{x\}) > 0.)$$

A distribution function F is discrete if and only if F only jumps up on the set A_F , i.e., the distribution F gives mass 1 to A_F :

F is discrete if and only if $F(A_F) = \int I\{x \in A_F\} F(dx) = 1$.

Hence the probability that F is discrete is

$$\Pr\{F: F \text{ is discrete}\} = \Pr\{F: F(A_F) = 1\},$$

and this probability is one if $E[F(A_F)] = 1$!

$$\begin{aligned} E[F(A_F)] &= \int F(A_F) D(dF|\alpha) = \iint I\{x \in A_F\} F(dx) D(dF|\alpha) \\ &= \iint I\{x \in A_F\} D(dF|\alpha + \delta_x) \alpha(dx)/\alpha(\infty) \quad (\text{Lemma 6.1.}) \\ &= \iint I\{F(\{x\}) > 0\} D(dF|\alpha + \delta_x) \alpha(dx)/\alpha(\infty) \quad (x \in A_F \text{ iff} \end{aligned}$$

$F(\{x\}) > 0$.)

However, for each fixed x , $F \sim D(dF|\alpha + \delta_x)$ implies that $F(\{x\})$ is Beta $(\alpha(\{x\}) + 1, \alpha(\infty) - \alpha(\{x\}))$. Such a Beta random variable is positive with probability one [even if $\alpha(\{x\}) = 0$.] That is, the inner integral is always 1, i.e.,

$$\int I\{F(\{x\}) > 0\} D(dF|\alpha + \delta_x) = 1 \text{ for each } x.$$

[The application of the Fubini-type theorem requires that the integrand $h(x, F) = I\{x \in A_F\} = I\{F(\{x\}) > 0\}$ is a "measurable" function of (x, F) .

Dubins and Freedman, in "Measurable sets of measures," Pacific J. Math., 1964, showed that this is indeed the case since R is a complete and separable metric space.]

Theorem 6.1. Suppose $x_1, \dots, x_n | F$ are iid F . Then $F \sim D(dF|\alpha)$ implies that $F|x_1, \dots, x_n \sim D(dF|\alpha + nF_n)$ where $F_n(t) = (1/n) \sum_{1 \leq i \leq n} I\{x_i \leq t\}$.

6.5. Gamma and Dirichlet process on a general space. In the case of multivariate observations, $F(t)$, $t = (t_1, \dots, t_q) \in R$ is a multivariate distribution function on a q -dimension Euclidean space R . The Gamma and Dirichlet process theory extends readily if we interpret $s \leq t$ as coordinatewise inequality; i.e., each coordinate in the s -vector is less than or equal to the corresponding coordinate in the t -vector.

Construction. The theory also extends to the case of a general complete and separable metric space R . In this case the definitions of Gamma and Dirichlet processes using random measures and random probabilities, for which the role of intervals $(s,t]$ is played by (measurable) subsets of R . A sketch of this development goes as follows: A stochastic process indexed by (measurable) subsets $\mu(\cdot)$ is called a Gamma process (or Gamma random measure) with shape measure $\alpha(\cdot)$ if

- (i) $\mu(\cdot)$ is "independent increment" in the sense that if B_j 's are disjoint, then $\mu(B_j)$'s are independent, and
- (ii) $\mu(A)$ is Gamma $(\alpha(A);1)$ for each subset A of R .

The system of stochastic processes indexed by (measurable) sets $\{\mu(A): A \text{ is a measurable subset of } R\}$ defined by (i) and (ii) is "consistent" in the sense of Kolmogorov (1933), and hence the existence of its joint distribution is guaranteed by Kolmogorov's existence theorem.

If the space R is also a nice space (a complete and separable metric space equipped with Borel sigma field) one can show that (i) and (ii) also imply

- (iii) $\Pr\{\mu: \mu \text{ is a countably additive measure}\}=1$.

Remark 6.4. Conclusion (iii) is subject to measurability conditions; see Harris, T. (1968) "Counting measures monotone random set functions." *Z. Wahr. verw Gebiete* **10**, 102-119.

The key Lemma 6.1. With the existing questions (i), (ii), and (iii) settled, all results in the previous two sections are valid (with a slight change of notation) and we say that $\mu(\cdot)$ is a Gamma random measure

with shape $\alpha(\cdot)$. The definition for the corresponding Dirichlet process $F(\cdot)$ follows:

$$F(\cdot) \equiv \mu(\cdot)/\mu(\mathbb{R}), \text{ denoted by } F \sim D(dF|\alpha).$$

[Analogous to the sequence case, this definition of $F(\cdot)$ requires that $\alpha(\mathbb{R})$ is finite.]

Lemma 6.1 translates to $F \sim D(dF|\alpha)$ and $x|F \sim F$. Then $F|x \sim D(dF|\alpha + \delta_x)$. The proof of this result for this general case follows the same way, and is in fact simpler in terms of notation. It suffices to check the equality there for $h(\cdot)$ s of the form $I_{\{x \in A\}} \times I_{\{F \in B\}}$, where A is a (measurable) set of x s, and B is a (measurable) set of F s. Since the finite-dimensional sets determines the distribution of a stochastic process, it suffices to check B of the form

$$B = \{F: F(B_1) \leq z_1, \dots, F(B_b) \leq z_b\},$$

where $B_j, j=1, \dots, b$ form a partition of the space \mathbb{R} . Note that

$$B_j = A \cap B_j + A^c \cap B_j, \text{ and hence } F(B_j) = F(A \cap B_j) +$$

$F(A^c \cap B_j)$. The left hand side of the equality in Lemma 6.1 is

$$\begin{aligned} & \iint h(x, F) F(dx) D(dF|\alpha) \\ &= \int I_{\{x \in A\}} I_{\{F(B_1) \leq z_1, \dots, F(B_b) \leq z_b\}} F(dx) D(dF|\alpha) \\ &= \int \left\{ \sum_{1 \leq j \leq b} F(A \cap B_j) \right\} I_{\{F(B_1) \leq z_1, \dots, F(B_b) \leq z_b\}} D(dF|\alpha) \\ &= \sum_{1 \leq j \leq b} \int F(A \cap B_j) I_{\{F(B_1) \leq z_1, \dots, F(B_b) \leq z_b\}} D(dF|\alpha) \\ &= \sum_{1 \leq j \leq b} \alpha(A \cap B_j) E_j^\# [I_{\{F(B_1) \leq z_1, \dots, F(B_b) \leq z_b\}}], \end{aligned}$$

where for each $j=1, \dots, b$ under $E_j^\#$, (Example 6.3)

$$F(A \cap B_j), F(A^c \cap B_j), F(A \cap B_m), F(A^c \cap B_m), m=1, \dots, j-1, j+1, \dots, b$$

is a Dirichlet vector with parameters

$$\alpha(A \cap B_j) + 1, \alpha(A^c \cap B_j), \alpha(A \cap B_m), \alpha(A^c \cap B_m), m=1, \dots, j-1, j+1, \dots, b.$$

A simple exercise, identical to the proof of Lemma 6.1, shows that the right hand side of the equality reduces to the same expression:

$$\begin{aligned}
& \iint I\{x \in A\} \int_{\mathbf{I}_B} D(dF|\alpha + \delta_x) \alpha(dx) / \alpha(\infty) \\
&= \sum_{1 \leq j \leq b} \int I\{x \in A \cap B_j\} \int_{\mathbf{I}_B} D(dF|\alpha + \delta_x) \alpha(dx) / \alpha(\infty) \\
&= \sum_{1 \leq j \leq b} \int I\{x \in A \cap B_j\} \int_{\mathbf{I}_B} D(dF|\alpha + \delta_x) \alpha(dx) / \alpha(\infty).
\end{aligned}$$

To complete the proof, note that for each $j \in \{1, \dots, b\}$,

$$I\{x \in A \cap B_j\} \int_{\mathbf{I}_B} D(dF|\alpha + \delta_x) = I\{x \in A \cap B_j\} \times E_j^\#[\mathbf{I}_B]. \quad \parallel$$

Exercise 6.4. Let $\mu \sim G(d\mu|\alpha)$ be a Gamma process with shape function α .

For any nonnegative function $h(x, \mu)$,

$$\iint h(x, \mu) \mu(dx) G(d\mu|\alpha) = \iint h(x, \mu) G(d\mu|\alpha + \delta_x) \alpha(dx).$$

Hint: Use the same proof as Lemma 6.1.

Exercise 6.5. Show that the above Fubini theorem for a gamma process [for any nonnegative functions $h(x, \mu)$] is equivalent to Lemma 6.1.