Introduction to Bayesian modeling

lecture notes

Fall 2022

Jüri Lember

1 Bayesian model

In what follows, $\mathcal{X}$ stands for *sample space*, which might be (a subset of) $\mathbb{R}$, $\mathbb{R}^d$, $\mathbb{R}^n$, $(\mathbb{R}^d)^n$. Thus $x \in \mathcal{X}$ might be a real number, a d-dimensional vector, a sample (vector) $(x_1, \ldots, x_n)$, where $x_i \in \mathbb{R}^d$.

Parametric model. In statistics, *a model* $\mathcal{P}$ is any set of probability distributions on $\mathcal{X}$ (more precisely, on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, where $\mathcal{B}(\mathcal{X})$ stands for Borel σ-algebra). Observe that a set of probability distribution can be formally represented as $\{P_\theta : \theta \in \Theta\}$, where Θ is a set. The model $\{P_\theta : \theta \in \Theta\}$ is *parametric*, when Θ is *finite dimensional*, w.l.o.g. $\Theta \subset \mathbb{R}^p$. In what follows, we shall consider parametric model, and we assume that all distributions are absolutely continuous with respect to some common reference measure (typically either Lebesgue or counting measure), let $f(\cdot|\theta)$ be the density of P_θ. The integration with respect to the reference measure will be denoted by $P_\theta(A) = \int_A f(x|\theta)dx$, as it is typically done when integrating with respect to the Lebesgue'i measure. However, when the reference measure is the counting measure, then the integral above is sum and $f(\cdot|\theta)$ is the probability mass function. So the densities of continuous and discrete random variables will be denoted identically.

1.1 Examples of parametric models

Here $\mathcal{X} \subset \mathbb{R}^d$ and $\{f(\cdot|\theta)\}$ is a class of distributions. In the following we list the models that will be used later.

Discrete models.

- Binomial distribution:
  $\mathcal{X} = \{0, 1, \ldots\}$, $\theta = (n, p) \in \mathbb{N} \times [0, 1] = \Theta$, (notation $B(n, p)$)

  $$f(k|\theta) = f(k|n, p) = \binom{n}{k}\theta^k(1-\theta)^{n-k}, \quad k = 0, \ldots, n; \quad f(k|\theta) = 0, \quad k > n;$$

- Poisson distribution:
  $\mathcal{X} = \{0, 1, \ldots\}$, $\theta = \lambda \in (0, \infty) = \Theta$, (notation $P_o(\lambda)$)

  $$f(k|\theta) = f(k|\lambda) = e^{-\lambda}\frac{\lambda^k}{k!};$$

- Negative binomial distribution:
  $\mathcal{X} = \{0, 1, \ldots\}$, $\theta = (r, p) \in (0, \infty) \times (0, 1) = \Theta$, (notation $\mathrm{NB}(r, p)$, alternative parametrizations)

  $$f(k|\theta) = f(k|r, p) = \frac{\Gamma(k + r)}{k!\Gamma(r)}(1 - p)^k p^r;$$

when r is integer, then

$$\frac{\Gamma(k+r)}{k!\Gamma(r)} = \binom{k+r-1}{k}.$$

- Discrete distributions with (at most) k atoms (categorical distributions):
  $\mathcal{X} = \{1, \dots, k\}$,

  $\Theta = \left\{(p_1, \dots, p_k) : \sum_{i=1}^{k} p_i = 1, p_i \geq 0\right\} =: S_{k-1}, \quad (k-1)$-dim simplex,

  $f(l|\theta) = f(l|p_1, \dots, p_k) = p_l, \quad l \leq k, \quad f(l|\theta) = 0, \quad l > k;$

- Multinomial distribution:

  $\mathcal{X} = \{(n_1, \dots, n_k) : n_1 + \dots + n_k = n\};$
  $\theta = (n, (p_1, \dots, p_k)) \in \mathbb{N} \times S_{k-1} = \Theta \quad$ (notation Multin$(n; p_1, \dots, p_k)$)

  $$f(n_1, \dots, n_k|\theta) = f(n_1, \dots, n_k|n; p_1, \dots, p_k) = \binom{n}{n_1 \cdots n_k} \prod_{i=1}^{k} p_i^{n_i};$$

Continuous models.

- Beta distribution:
  $\mathcal{X} = [0,1]$, $\theta = (\alpha, \beta) \in (0, \infty) \times (0, \infty) = \Theta$, (notation Beta$(\alpha, \beta)$)

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

  $$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

  Thus Beta$(1,1)$ is $U[0,1]$ – uniform distribution.

- Dirichlet distribution:
  $\mathcal{X} = S_{k-1}$, $\theta = (\alpha_1, \dots, \alpha_k) \in (0, \infty)^k = \Theta$ (notation Dir$(\alpha_1, \dots, \alpha_k)$)

  $$f(x|\theta) = f(x|\alpha_1, \dots, \alpha_k) = \frac{1}{B(\alpha_1, \dots, \alpha_k)} \prod_{i=1}^{k} x_i^{\alpha_i - 1},$$

  $$B(\alpha_1, \dots, \alpha_k) = \int_{S_k} \prod_{i=1}^{k} (x_i)^{\alpha_i - 1} dx =$$

  $$\int_0^1 \int_0^{1-x_1} \cdots \int_0^{1-x_1-\cdots-x_{k-2}} \prod_{i=1}^{k-1} (x_i)^{\alpha_i-1} (1 - x_1 - \cdots - x_{k-1})^{\alpha_k - 1} dx_1 \cdots dx_{k-1} =$$

  $$\frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)}{\Gamma(\alpha_1 + \cdots + \alpha_k)}.$$

- Uniform distribution:
  $\mathcal{X} = [0, \infty)$, $\theta \in (0, \infty) = \Theta$ (notation $U(0, \theta)$):

$$f(x) = \frac{1}{\theta} I_{[0,\theta]}(x).$$

- Gamma distribution:
  $\mathcal{X} = [0, \infty)$, $\theta = (\alpha, \beta) \in (0, \infty)^2 = \Theta$ (notation $\text{Gamma}(\alpha, \beta)$, alternative parametrizations)

$$f(x|\theta) = f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp[-\beta x].$$

- Exponential distribution:
  $\mathcal{X} = [0, \infty)$, $\theta = \lambda \in (0, \infty) = \Theta$ (notation $\text{Exp}(\lambda)$)

$$f(x|\theta) = f(x|\lambda) = \lambda \exp[-\lambda x].$$

  $\text{Gamma}(1, \beta) = \text{Exp}(\beta)$.

- Shifted exponential distributions:
  $\mathcal{X} = [0, \infty)$, $\theta = (x_o, \lambda) \in (0, \infty)^2 = \Theta$ (notation $\text{Exp}(x_o, \lambda)$)

$$f(x|\theta) = f(x|x_o, \lambda) = \lambda \exp[-\lambda(x - x_o)] I_{[x_o,\infty)}(x).$$

  $\text{Exp}(0, \lambda) = \text{Exp}(\lambda)$.

- Chi-squared distribution:
  $\mathcal{X} = (0, \infty)$, $\theta = k \in \{1, 2, \ldots\}$ (notation χ^2_k)

$$f(x|\theta) = f(x|k) = \frac{1}{2^{k/2}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} \exp[-\frac{x}{2}], \quad x > 0.$$

  $\chi^2_k = \text{Gamma}(\frac{k}{2}, \frac{1}{2})$.

- Inverse Gamma distribution:
  $\mathcal{X} = [0, \infty)$, $\theta = (\alpha, \beta) \in (0, \infty)^2 = \Theta$ (notation $\text{InvGamma}(\alpha, \beta)$),

$$f(x|\theta) = f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp[-\beta/x].$$

  When $X \sim \text{Gamma}(\alpha, \beta)$, then $X^{-1} \sim \text{InvGamma}(\alpha, \beta)$.
  When parametrized $\alpha = \frac{\nu}{2}$ and $\beta = \frac{\nu\tau^2}{2}$, then it is called scaled inverse chi-squared distribution, thus

$$\text{InvGamma}(\frac{\nu}{2}, \frac{\nu\tau^2}{2}) = \text{ScaleInv}{-}\chi^2(\nu, \tau^2).$$

- Normal distribution:
  $\mathcal{X} = \mathbb{R}$, $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$ (notation $\mathcal{N}(\mu, \sigma^2)$)

$$f(x|\theta) = f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{(x-\mu)^2}{2\sigma^2}].$$

- Student t-distribution:
  $\mathcal{X} = \mathbb{R}$, $\theta = \nu \in \mathbb{R} = \Theta$ (notation t_ν),

$$f(x|\theta) = f(x|\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})}\Big(1 + \frac{x^2}{\nu}\Big)^{\frac{-(\nu+1)}{2}}.$$

  Since $\Gamma(1/2) = \sqrt{\pi}$,

$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} = \frac{1}{\sqrt{\nu}B(1/2, \nu/2)}.$$

- Location-scale t-distribution:
  $\mathcal{X} = \mathbb{R}$, $\theta = (\mu, \tau^2, \nu) \in \mathbb{R} \times (0, \infty) \times \mathbb{R} = \Theta$ (notation $\mathrm{lst}(\mu, \tau^2, \nu)$),

$$f(x|\theta) = f(x|\mu, \tau^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu\tau^2}\Gamma(\frac{\nu}{2})}\Big(1 + \frac{1}{\nu}\frac{(x-\mu)^2}{\tau^2}\Big)^{\frac{-(\nu+1)}{2}}.$$

  When $X \sim t_\nu$, then $\mu + \tau X \sim \mathrm{lst}(\mu, \tau^2, \nu)$.

- Pareto distribution:
  $\mathcal{X} = [0, \infty)$, $\theta = (\alpha, x_o) \in (0, \infty) \times (0, \infty) = \Theta$ (notation $\mathrm{Pa}(\alpha, x_o)$)

$$f(x|\theta) = f(x|\alpha, x_o) = \alpha(x_o)^\alpha \frac{1}{x^{\alpha+1}} I_{[x_o, \infty)}(x).$$

- Lomax distribution:
  $\mathcal{X} = [0\infty)$, $\theta = (\alpha, \lambda) \in (0, \infty) \times (0, \infty) = \Theta$ (notation: $\mathrm{Lomax}(\alpha, \lambda)$),

$$f(x|\theta) = f(x|\alpha, \lambda) = \frac{\alpha\lambda^\alpha}{(x+\lambda)^{\alpha+1}}, \quad x \geq 0.$$

However, a parametric model can also be defined via a class of distributions $\{f(\cdot|\theta)\}$ as above, but $x = (x_1, \ldots, x_n)$ can be an iid sample from $f(\cdot|\theta)$. In this case $\mathcal{X} \subset \mathbb{R}^n$, $x = (x_1, \ldots, x_n)$ and $f(x|\theta) = \prod_{i=1}^n f(x_i|\theta)$. It can also be that $x = (x_1, \ldots, x_n)$ is an output of any other model rather than iid model, for example Markov chain.

Bayesian model. Given a parametric model $\{f(\cdot|\theta)\}$ and a parameter θ, the data are modeled as a random vector (or random variable) X having density $f(\cdot|\theta)$. Any realization x of X are called *observations* and $f(x|\theta)$ is the *likelihood* of x (we call it likelihood even when the reference measure is the counting measure and $f(x|\theta)$ is actually a probability).

In parametric Bayesian model, the parameter is modeled random as well, thus (X, θ) is a random vector on $(\mathcal{X}, \Theta)$. The model presupposes that Θ is equipped with Borel σ-algebra $\mathcal{B}(\Theta)$ and (X, θ) is $\mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\Theta)$-measurable. The distribution of θ shall be denoted by π, thus for any $A \in \mathcal{B}(\mathcal{X})$ and $B \in \mathcal{B}(\Theta)$, it holds (law of total probability)

$$P(X \in A, \theta \in B) = \int_B \int_A f(x|\theta) dx \pi(d\theta).$$

Clearly the integral above presupposes that the map $(x, \theta) \mapsto f(x|\theta)$ is $\mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\Theta)$-measurable, and this is so for all models in these notes. Observe that we use the same θ for a random parameter and its realization. The distribution π is called the *prior distribution (eelmõõt)*. Now the (marginal) distribution of data X has density $f(x) = \int f(x|\theta)\pi(d\theta)$, because Fubini theorem allows to change the order of integration and so

$$P(X \in A) = \int_\Theta \int_A f(x|\theta) dx \pi(d\theta) = \int_A \int_\Theta f(x|\theta)\pi(d\theta) dx = \int_A f(x) dx.$$

Observe that the density $f(x)$ do not (in general) belong to the class $\{f(\cdot|\theta)\}$, moreover, the marginal distribution of X might be very different form any distribution from $\{f(\cdot|\theta)\}$.

The central object in Bayesian statistics is the conditional distribution of parameter given the observations: $P(\theta \in \cdot | X = x)$. This distribution is called *posterior distribution (järelmõõt)*. By *Bayes formula* (Exercise 1)

$$P(\theta \in B | X = x) = \frac{\int_B f(x|\theta)\pi(d\theta)}{f(x)} = \frac{\int_B f(x|\theta)\pi(d\theta)}{\int f(x|\theta)\pi(d\theta)}, \quad \forall B \in \mathcal{B}(\Theta). \quad (1.1)$$

When π has density $\pi(\theta)$ with respect to some reference measure $d\theta$ on $(\Theta, \mathcal{B}(\Theta))$, then also posterior distribution has density (with respect to the same measure), because

$$P(\theta \in B | X = x) = \frac{\int_B f(x|\theta)\pi(\theta)d\theta}{f(x)} = \int_B \frac{f(x|\theta)\pi(\theta)}{f(x)} d\theta = \int_B \pi(\theta|x) d\theta,$$

where

$$\pi(\theta|x) := \frac{f(x|\theta)\pi(\theta)}{f(x)} = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}.$$

In this note, a prior measure always has a density, thus $\pi(\theta|x)$ is always defined and, in a sense, it is the main object of interest. Since the denominator $f(x)$ does not depend on θ, the posterior density is

$$\pi(\theta|x) \propto \pi(\theta)p(x|\theta).$$

To recapitulate: A Bayesian model is made of a parametric statistical model $f(x|\theta)$, and a prior density $\pi(\theta)$ of parameters. When $\pi(\theta)$ belongs to a parametric model, then these parameters are called *hyperperameters*.

Posterior predicative density. The standard parametric model is the following: $x = (x_1, \ldots, x_n)$ is a realization of iid random variables $X_1, \ldots, X_n$ from the distribution $f(\cdot|\theta)$, where θ has density $\pi(\theta)$. In Bayesian literature this model is often written as

$$\begin{aligned} \theta &\sim \pi \\ X_1, \ldots, X_n|\theta &\overset{i.i.d.}{\sim} f(\cdot|\theta) \end{aligned} \tag{1.2}$$

Thus θ is a random variable with density $\pi(\theta)$, and given θ, $x_1, \ldots, x_n$ are realizations of iid random variables $X_1, \ldots, X_n$, where X_i has a density $f(\cdot|\theta)$. Now, let X_{n+1} another independent observation with the same distribution. We are interested in the conditional distribution of X_{n+1} given the previous observations x (here $X = (X_1, \ldots, X_n)$):

$$\begin{aligned} P(X_{n+1} \in A|X = x) &= \int P(X_{n+1} \in A|X = x, \theta)\pi(\theta|x)d\theta \\ &= \int \int_A f(x_{n+1}|\theta)dx_{n+1}\pi(\theta|x)d\theta \\ &= \int_A \int f(x_{n+1}|\theta)\pi(\theta|x)d\theta \ dx_{n+1} \\ &= \int_A f(x_{n-1}|x)dx_{n+1}, \quad f(x_{n-1}|x) := \int f(x_{n+1}|\theta)\pi(\theta|x)d\theta. \end{aligned}$$

The second equality holds because given θ, the random variable X_{n+1} is independent of $X_1, \ldots, X_n$ and has density $f(\cdot|\theta)$. The order of integration can be reversed due to Fubini theorem. The conditional distribution $P(X_{n+1} \in \cdot|X = x)$ is called *posterior predicative distribution* and its density $f(\cdot|x)$ is called *posterior predicative density*. When the posterior density $\pi(\theta|x)$ in the definition of $f(x_{n-1}|x)$ is replaced by prior density π, then one obtains *prior predicative density*, which is just density of X_1.

Exchangeable random variables. The random variables $X_1, \ldots, X_n$ are said to be *exchangeable (vahelduvad)* if the joint distribution of $(X_1, \ldots, X_n)$ is the same as that of $\{X_{\sigma_1}, \ldots, X_{\sigma_n}\}$ for any permutation $(\sigma_1, \sigma_2, \ldots, \sigma_n)$ of the indices $\{1, 2, \cdots, n\}$. For example, when $n = 3$, then six random

vectors $(X_1, X_2, X_3), (X_2, X_1, X_3), (X_1, X_3, X_2), (X_2, X_3, X_2), (X_3, X_1, X_2)$, (X_3, X_2, X_1) have the same distribution.
A sequence of random variables $X_1, X_2, \ldots$ is called *(infinitely) exchangeable*, if for any n, $X_1, \ldots, X_n$ are exchangeable.

When the vector $(X_1, \ldots, X_n)$ has the (joint) density $f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n)$ then the random variables are exchangeable when for every $(x_1, \ldots, x_n)$ and for every permutation $(\sigma_1, \sigma_2, \ldots, \sigma_n)$ it holds:

$$f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = f_{(X_1,\ldots,X_n)}(x_{\sigma_1}, \ldots, x_{\sigma_n}).$$

Clearly iid random variables are exchangeable. Also the random variables $X_1, \ldots, X_n$ defined as in (1.2) are exchangeable, because the density

$$f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = \int \prod_{i=1}^{n} f(x_i|\theta)\pi(\theta)d\theta$$

is invariant with respect to the permutation of the arguments.

When $X_1, \ldots, X_n$ are exchangeable, then they are identically distributed, but not generally independent. For example, when $X_1, \ldots, X_n$ are defined as in (1.2) and

$$E\big(E[X_1|\theta]\big)^2 = \int \Big(\int x f(x|\theta)dx \Big)^2 \pi(\theta)d\theta < \infty,$$

then for every $i \neq j$ (Exercise 1), $\mathrm{Cov}(X_i, X_j) = \mathrm{Var}(E[X_1|\theta]) > 0$. Also observe that when the random vector $(X_1, \ldots, X_n)$ is as in (1.2), then the distribution of X_i is the same as with $n = 1$, i.e. X_i has prior predicative density (Exercise 1)

$$f(x) = \int f(x|\theta)\pi(\theta)d\theta. \tag{1.3}$$

Hewitt-Savage-deFinetti theorem. Let us recall once again our standard parametric model

$$\begin{aligned} \theta &\sim \pi \\ X_1, X_2, \ldots|\theta &\overset{i.i.d.}{\sim} f(\cdot|\theta) \end{aligned} \tag{1.4}$$

We know that the sequence $X_1, X_2, \ldots$ is exchangeable. For every n and θ, $X_1, X_2, \ldots, X_n$ are conditionally independent, i.e.

$$P(X_1 \in A_1, \ldots, X_n \in A_n|\theta) = \prod_{i=1}^{n} P(X_i = A_i|\theta), \quad \forall A_i \in \mathcal{B}(\mathcal{X}), \quad i = 1, \ldots, n.$$

Since $P(X_i = A_i|\theta) = P(X_1 = A_i|\theta)$, we denote $P(X_1 \in A|\theta) =: Q(A, \theta)$, so that the equation above reads

$$P(X_1 \in A_1, \ldots, X_n \in A_n|\theta) = \prod_{i=1}^{n} Q(A_i, \theta), \quad \forall A_i \in \mathcal{B}(\mathcal{X}), \quad i = 1, \ldots, n.$$

Unconditionally thus

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \int \prod_{i=1}^{n} Q(A_i, \theta)\pi(d\theta), \quad \forall A_i \in \mathcal{B}(\mathcal{X}), \quad i = 1, \ldots, n.$$

Let $\mathcal{P}$ be the set of all probability measures on $\mathcal{X}$. It can be equipped with a suitable metric (Prokhorov one) so that the mapping

$$T : \Theta \to \mathcal{P}, \quad T(\theta) = Q(\cdot, \theta)$$

is $\mathcal{B}(\Theta)/\mathcal{B}(\mathcal{P})$-measurable. Then by the change of variables formula, it holds

$$\int_{\Theta} \prod_{i=1}^{n} Q(A_i, \theta)\pi(d\theta) = \int_{\mathcal{P}} \prod_{i=1}^{n} Q(A_i)\pi T^{-1}(dQ),$$

where πT^{-1} is the pushforward measure, let us denote it $\pi^* = \pi T^{-1}$. Thus, there exists a probability measure (prior) π^* on $\mathcal{P}$ so that it holds

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{\mathcal{P}} \prod_{i=1}^{n} Q(A_i)\pi^*(dQ), \quad \forall n, \quad \forall A_i \in \mathcal{B}(\mathcal{X}). \tag{1.5}$$

Since π^* can be considered as the distribution of a random probability measure (defined on some probability space), the random variables $X_1, X_2, \ldots$ can be considered as outputs of the following model: drawing a probability measure Q from the prior π^* (a realization of random measure) and the sampling from Q. Formally:

$$\begin{aligned} Q &\sim \pi^* \\ X_1, X_2 \ldots |Q &\overset{i.i.d.}{\sim} Q \end{aligned} \tag{1.6}$$

To summarize: (1.4) is a special case of (1.6). Since under (1.6) also (1.5) holds, so it follows that $X_1, X_2, \ldots$ are infinitely exchangeable.

Let P_n be the empirical measure of $X_1, \ldots, X_n$, i.e.

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(X_i).$$

When $X_1, X_2, \ldots$ are iid random variables from the distribution Q, then $P_n \Rightarrow Q$, a.s., where $\Rightarrow$ stands for the weak convergence of probability measures (convergence in Prokhorov metric). Hence in the model (1.6) the random measure Q is the limit of empirical measures.

The celebrated deFinetti-Hewitt-Savage theorem states that whenever $X_1, X_2, \ldots$ is a infinitely exchangeable sequence of random variables, there exists a probability measure π^* such that (1.6) holds. In other words: all (infinitely) exchangeable sequences are mixtures of iid sequences – the prior measure exists!

Theorem 1.1 (deFinetti-Hewitt-Savage) *Let $X_1, X_2, \ldots$ be infinitely exchangeable random variables. Then there exists a probability measure π^* on $\mathcal{P}$ so that (1.5) holds. Moreover, π^* is the distribution of a random measure Q, where $Q = \lim_n P_n$, a.s. and given Q, the random variables $X_1, X_2, \ldots$ are iid with distribution Q:*

$$P(X_1 \in A_1, \ldots, X_n \in A_n | Q) = \prod_{i=1}^{n} Q(A_i) \quad \forall n, \quad \forall A_i \in \mathcal{B}(\mathcal{X}).$$

When $|\mathcal{X}| = k < \infty$, i.e. the random variables have categorical distribution, then $\mathcal{P} = S_{k-1}$ (simplex) and the random measure Q is then a just a k-dim random vector and we have a parametric case. In case $k = 2$, the simplex is $[0, 1]$ and random measure Q can be identified with a random variable θ taking values in $[0, 1]$. Then the theorem above can be stated as follows.

Theorem 1.2 (deFinetti) *Let $X_1, X_2, \ldots$ be infinitely exchangeable Bernoulli random variables. Then there exists a random variable θ taking values in $[0, 1]$ such that for every n, every $x_1, \ldots, x_n \in \{0, 1\}$, it holds*

$$P(X_1 = x_1, \ldots, X_n = x_n | \theta) = \theta^s (1 - \theta)^{n-s}, \quad s = x_1 + \cdots + x_n. \tag{1.7}$$

Moreover, θ is the a.s. limit of sample mean: $\frac{1}{n} \sum_{i=1}^{n} X_i \to \theta$, a.s..

Taking expectation in (1.7), we get the special case of (1.5):

$$P(X_1 = x_1, \ldots, X_n = x_n) = E[P(X_1 = x_1, \ldots, X_n = x_n | \theta)] = \int \theta^s (1-\theta)^{n-s} \pi(d\theta),$$

where $\theta \sim \pi$.

For the proofs of these theorems, see [4], sec 1.5.

1.2 Exercises

1. The conditional probability $P(\theta \in B | X)$ is defined as any function $g(X)$ satisfying

$$E\big(I_A(X) g(X)\big) = P(X \in A, \theta \in B), \quad \forall A \in \mathcal{B}(\mathcal{X}). \tag{1.8}$$

Show that

$$g(X) = \frac{\int_B f(X|\theta)\pi(\theta)d\theta}{\int f(X|\theta)\pi(\theta)d\theta}$$

satisfies (1.8) and so Bayes formula (1.1) holds.

2. Let $X_1, \ldots, X_n$ as in (1.2). Prove that when $E\big(E[X_1|\theta]\big)^2 < \infty$, then for every $i \neq j$, $\mathrm{Cov}(X_i, X_j) = \mathrm{Var}[E(X_1|\theta)] \geq 0$. Show that X_i has density (1.3).

2 Examples of standard parametric models

In this section we give overview of some standard parametric models in form (1.4) used in practice.

2.1 Beta-Bernoulli and Beta-Binomial model

From deFinetti theorem, it follows that for any infinitely exchangeable sequence of Bernoulli random variables $X_1, X_2, \ldots$, there exists a probability measure π on $[0,1]$ (prior) so that the model (1.4) hold:

$$\begin{aligned} \theta &\sim \pi \\ X_1, X_2, \ldots|\theta &\overset{i.i.d.}{\sim} B(1,\theta). \end{aligned} \tag{2.1}$$

Hence when modeling exchangeable Bernoulli random variables, one has to choose π. A common choice for π is Beta distribution.

Beta distribution. Beta distribution is supported on $[0,1]$, hence suitable for modeling probabilities. When $\theta \sim \mathrm{Beta}(\alpha, \beta)$, then

$$E(\theta) = \frac{\alpha}{\alpha+\beta}, \quad \mathrm{Var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\frac{1}{\alpha+\beta+1}.$$

Hence the bigger $\alpha+\beta$, the smaller variance and the more is prior concentrated around its mean. When $\alpha > 1$ and $\beta > 1$ then the density $\pi(\theta)$ has unique maximum at

$$\frac{\alpha-1}{\alpha+\beta-2},$$

this is then the mode of the distribution.

Observe that any density in form $f(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ must be the density of $\mathrm{Beta}(\alpha, \beta)$ distribution.

Beta-Bernoulli and Beta Binomial models. With Beta distribution as π, we obtain *Beta-Bernoulli model*:

$$\theta \sim \text{Beta}(\alpha, \beta)$$
$$X_1, \ldots, X_n | \theta \overset{iid}{\sim} B(1, \theta).$$

Thus $x = (x_1, \ldots, x_n) \in \{0,1\}^n$, the hyperparameters are α and β. Since the sum of n iid Bernoulli random variables has binomial distribution, then for given θ, $X = \sum_{i=1}^n X_i \sim B(n, \theta)$, hence with the sum X instead of $X_1, \ldots, X_n$, we obtain *Beta-Binomial model*:

$$\theta \sim \text{Beta}(\alpha, \beta)$$
$$X|\theta \sim B(n, \theta).$$

When $X_1, \ldots, X_n$ are obtained by Beta-Bernoulli model – we shall call them Beta-Bernolli random variables for short – then $\sum_{i=1}^n X_i$ follows Beta-Binomial model.

(Marginal) distribution of X. Let us find the joint distribution of Beta-Bernoulli random variables $X = (X_1, \ldots, X_n)$. As noted above, in Bayesian terminology it is a marginal distribution, because the full model is (θ, X). The sample space is $\mathcal{X} = \{0,1\}^n$, and for any $x \in \{0,1\}^n$, let $n_1 = \sum_{i=1}^n x_i$. The distribution of X is thus

$$\begin{aligned} P(X = x) &= \int_0^1 P(X = x|\theta)\pi(\theta)d\theta = \int_0^1 \theta^{n_1}(1-\theta)^{n-n_1} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} d\theta \\ &= \frac{1}{B(\alpha,\beta)} \int_0^1 \theta^{\alpha+n_1-1}(1-\theta)^{\beta+n-n_1-1} d\theta \\ &= \frac{B(\alpha+n_1, \beta+n-n_1)}{B(\alpha,\beta)}. \end{aligned}$$

Observe that the probability $P(X = x)$ depend on the number of ones n_1, so that clearly any permutation of vector x has the same probability – the exchangeability. Since Beta-Bernoulli model is a special case of (1.2), we already know that $X_1, \ldots, X_n$ are identically distributed with $\text{B}(1, \alpha/(\alpha+\beta))$ distribution random variables that are positively correlated (Exercise 1).

Let now X be as in Beta-Binomial model, i.e. $X \in \{0, \ldots, n\}$. Since $X = \sum_{i=1}^n X_i$, where $(X_1, \ldots, X_n)$ follow Beta-Bernoulli model, then (now x instead n_1)

$$P(X = x) = \binom{n}{x} \frac{B(\alpha+x, \beta+n-x)}{B(\alpha,\beta)}, \quad x \in \{0, 1, \ldots, n\}. \tag{2.2}$$

The obtained distribution is known as *Beta-binomial* distribution with parameters α, β, n, denoted $X \sim \text{BetaBin}(n, \alpha, \beta)$. When $\alpha = \beta$, then this

distribution is symmetric in the sense that $f(0) = f(n)$, $f(1) = f(n-1)$, $f(2) = f(n-2)$, etc. Here f the density (probability mass function) of Beta-binomial law. When $X \sim \text{BetaBin}(n, \alpha, \beta)$, then (Exercise 2)

$$EX = \frac{n\alpha}{\alpha+\beta}, \quad \text{Var}(X) = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta+1)(\alpha+\beta)^2}.$$

When $\alpha = \beta = 1$ (the prior distribution is uniform), then Beta-binomial law is uniform over $\{0, \dots, n\}$ (Exercise 2).

Polya urn. Let $\alpha \geq 1$ and $\beta \geq 1$ be integers. Imagine an urn containing α red balls and β black balls, where random draws are made. If a red ball is observed, then two red balls are returned to the urn. Likewise, if a black ball is drawn, then two black balls are returned to the urn. Let $X_1 = 1$, when the ball taken out is red and $X_1 = 0$, otherwise. Then repeat it n times. Let $X_i \in \{0, 1\}$ be the color of i-th ball taken out of urn. When $\alpha > 0$ and $\beta > 0$ are not integer, the *Polya urn model/scheme* mimics the above-described urn: $X_1 \sim B(1, \frac{\alpha}{\alpha+\beta})$,

$$P(X_2 = 1 | X_1 = x_1) = \frac{\alpha + x_1}{\alpha+\beta+1}, \cdots$$
$$P(X_{i+1} = 1 | X_1 = x_1, \dots, X_i = x_i) = \frac{\alpha + (x_1 + \cdots + x_i)}{\alpha+\beta+i}, \quad i = 3, \dots, n.$$

So we shall speak about Polya urn with initial number of balls α and β even when they are not integers. It is easy to see that the random variables $X_1, X_2, \dots$ are infinitely exchangeable (the order the balls are taken out of urn do not matter), hence by deFinetti theorem there exists a prior distribution π so that the model (2.1) holds. It turns out that the prior distribution is Beta(α, β) distribution. To see that observe that for every n, $(X_1, \dots, X_n)$ has the same distribution as in Beta-Bernoulli model, and hence the total number of red balls taken out of urn (during n draws) has BetaBin(n, α, β)-distribution (Exercise 3). Hence, from deFinetti theorem again, the proportion of red balls out of all balls taken out of urn (as well as the proportion of red balls in urn) converges a.s. to a random variable θ, where $\theta \sim \text{Beta}(\alpha, \beta)$.

The observation that Beta-Bernoulli model is equivalent to Polya urn is important in generating observations $X_1, X_2, \dots$ from Beta-Bernoulli model – *sampling*. According to the definition, one could generate a random parameter $\theta \sim \text{Beta}(\alpha, \beta)$, and then generate an iid sample from $B(1, \theta)$. The alternative is just to use Polya urn scheme. Polya urn model is an example of a model, where a prior measure naturally exists (whether we believe in Bayesian approach or not).

Posterior predicative distribution. Polya urn allows to calculate the posterior predicative distribution in Beta-Bernoulli model without actually calculating the posterior distribution! Indeed: let $X_1, \dots, X_n$ be Beta-Bernoulli random variables, and let us find the posterior predicative distribution. Since $X_{n+1} \in \{0, 1\}$, it suffices to find the conditional probability

$$P(X_{n+1} = 1 | X_1 = x_1, \dots, X_n = x_n) = \frac{\alpha + n_1}{\alpha + \beta + n}, \quad n_1 = x_1 + \dots + x_n.$$

In particular, when the prior has uniform distribution, i.e. $\alpha = \beta = 1$, then

$$P(X_{n+1} = 1 | X_1 = x_1, \dots, X_n = x_n) = \frac{n_1 + 1}{n + 2}.$$

The estimate above is known as *Laplace rule of succession* and it does not exclude the possibility that $X_{n+1} = 1$ (or $X_{n+1} = 0$) even when $n_1 = 0$ i.e. only zeros so far (or $n_1 = n$ i.e. only ones so far):

$$P(X_{n+1} = 1 | X_1 = 0, \dots, X_n = 0) = \frac{1}{n + 2}$$
$$P(X_{n+1} = 0 | X_1 = 1, \dots, X_n = 1) = \frac{n + 1}{n + 2}.$$

When data consists of only ones (or zeros), then maximum likelihood estimate n_1/n would be 1 (or 0) and with such estimates the possibility of observing something else would be zero.

Posterior density. Consider the Beta-Bernoulli model with hyperparameters α and β. Let $x = (x_1, \dots, x_n) \in \{0, 1\}^n$, where $n_1 = \sum_{i=1}^n x_i$ be the observations. The posterior density:

$$\begin{aligned} \pi(\theta | x) \propto \pi(\theta) f(x | \theta) &= \frac{1}{B(\alpha, \beta)} \theta^{\alpha - 1} (1 - \theta)^{\beta - 1} \theta^{n_1} (1 - \theta)^{n - n_1} \\ &= \frac{1}{B(\alpha, \beta)} \theta^{n_1 + \alpha - 1} (1 - \theta)^{n - n_1 + \beta - 1}. \end{aligned}$$

Hence

$$\pi(\theta | x) \propto \theta^{n_1 + \alpha - 1} (1 - \theta)^{n - n_1 + \beta - 1},$$

so posterior distribution must be Beta distribution:

$$\theta | x \sim \text{Beta}(n_1 + \alpha, n - n_1 + \beta). \tag{2.3}$$

Suppose $\alpha > 1$ and $\beta > 1$ are integers. Then $\text{Beta}(\alpha, \beta)$ distribution can be considered as posterior distribution of a sample that has $\alpha - 1$ ones and $\beta - 1$ zeros and uniform $\text{Beta}(1, 1)$ prior. This observation allows to interpret the hyperparameters – the prior distribution can be considered as the posterior of some "prior sample".

We can now calculate the posterior predicative distribution directly (by definition, without Polya urn scheme):

$$P(X_{n+1}=1|X=x)=\int f(1|\theta)\pi(\theta|x)d\theta=\int\theta\pi(\theta|x)d\theta=E[\theta|x]=\frac{\alpha+n_1}{\alpha+\beta+n}.$$

Since the posterior distribution depends on data through n_1 only, the posterior is the same as in Beta-binomial model. Now the observation x is the number of ones, i.e. $x\in\{0,1,\dots,n\}$ and

$$\begin{aligned}\pi(\theta|x)\propto\pi(\theta)p(x|\theta)&=\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\binom{n}{x}\theta^x(1-\theta)^{n-x}\\&=\frac{1}{B(\alpha,\beta)}\binom{n}{x}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}.\end{aligned}$$

Again, we see that

$$\pi(\theta|x)\propto\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1},$$

hence the posterior distribution must be $\text{Beta}(x+\alpha,n-x+\beta)$:

$$\pi(\theta|x)=\frac{1}{B(x+\alpha,n-x+\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}. \tag{2.4}$$

Posterior mean, mode and variance, interpretation. Since posterior distribution is the same for Beta-Bernoulli and Beta-binomial model, we shall consider the latter, thus $x\in\{0,\dots,n\}$ is the number of ones in observations and $\theta|x\sim\text{Beta}(\alpha+x,\beta+n-x)$. Recall the posterior mean:

$$E[\theta|x]=\int_0^1\theta\pi(\theta|x)d\theta=\frac{x+\alpha}{n+\alpha+\beta}=\frac{x}{n}\cdot\frac{n}{\alpha+\beta+n}+\frac{\alpha}{\alpha+\beta}\cdot\frac{\alpha+\beta}{\alpha+\beta+n}.$$

We see that the posterior mean is the weighted average of sample mean $\frac{x}{n}$ and prior mean $\frac{\alpha}{\alpha+\beta}$, the bigger is n, the smaller is prior influence. Thus when n is relatively big in comparison with $\alpha+\beta$, then $E[\theta|x]\approx\frac{x}{n}$. The posterior variance is

$$\text{Var}[\theta|x]=\frac{(x+\alpha)(n-x+\beta)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}=\frac{E[\theta|x](1-E[\theta|x])}{\alpha+\beta+n+1},$$

and when n is relatively big then $\text{Var}[\theta|x]\approx\frac{1}{n}\frac{x}{n}(1-\frac{x}{n})$. We see that when the data have distribution $X\sim B(n,\theta^*)$, where θ^* is so called *true parameter*, then by SLLN, $\frac{x}{n}\approx\theta^*$, so that the bigger n the more posterior distribution concentrates around θ^*.

When the goal is to obtain a point-estimate $\hat{\theta}$ of parameter (in Bayesian

setup), then posterior mean and mode (analogue of MLE) are standard estimates. Observe that for Beta-Bernoulli model they do not coincide, because the posterior mode is

$$\frac{\alpha + x - 1}{\alpha + \beta + n - 2}.$$

Observe that when the prior is uniform ($\alpha = \beta = 1$), then the posterior mode is MLE: x/n, the posterior mean is $(x+1)/(n+2)$.

The beginning of Bayesian analysis: T. Bayes and P.S. Laplace. The article: *An Essay towards solving a Problem in the Doctrine of Chances* by presbyterian minister Thomas Bayes (published in 1763, two years after Bayes death). The model in the article: A ball W is randomly thrown (according to a uniform distribution) on the table of unit area. The horizontal position of the ball on the table is θ. Then a second ball O is randomly thrown n times; X denotes the number of times O stopped on the left of W. T. Bayes asks: what inference can we make on θ given X?

In modern terms it is a Beta-Bernoulli model with uniform prior. T. Bayes succeeded to calculate posterior and found that the random variable X has uniform distribution – BetaBinomial$(n, 1, 1)$.

Independently of Thomas Bayes, the Bayes formula was used by Pierre S. Laplace. One of his applications (published in 1786) was to test whether the probability θ of a male birth is above $1/2$. From 1745–1770, a total 251527 boys and 241945 girls were born in Paris. Assuming uniform prior and Beta-Binomial model with $n = 251527 + 241945 = 493472$, $x = 251527$ (thus $x/n = 0.5097\cdots$), Laplace showed that

$$P(\theta \leq 0.5|x) = \int_0^{0.5} \pi(\theta|x)d\theta \approx 1.15 \times 10^{-42}.$$

He then deduces that θ is more than likely to be above $1/2$. We know that such a small probability is due to the very small posterior variance.

Empirical prior: rat tumor example. Consider the rat tumor example in [5], section 5.1. The problem is estimating the probability of tumor θ. The data consists of small (current) estimate – 4/14 (number of rats with tumor)/(total number of rats) and 70 estimates (x_j/n_j), $j = 1, \ldots, 70$ constructed from similar previous (historical) experiments (see Table 5.1 in [5]). There are several ways to proceed.
One option is to consider all experiments equal – all 71 estimates (the current estimate is 71-th) are considered as independent outputs from the same

Beta-Binomial model:

$$\theta \sim \text{Beta}(\alpha, \beta)$$
$$X_j|\theta \overset{ind}{\sim} B(n_j, \theta), \quad j = 1, \ldots, 71.$$

This model can be considered as a single Beta-Bernoulli model with $n = n_1 + \cdots + n_{71}$ trials and $x = x_1 + \cdots + x_{71}$ successes and the posterior distribution is then $\text{Beta}(\alpha + x, \beta + n - x)$.
Another option is to consider all historical experiment being realizations of independent binomial random variables with different parameters. The parameters are iid from $\text{Beta}(\alpha, \beta)$ distribution:

$$\theta_j \overset{iid}{\sim} \text{Beta}(\alpha, \beta) \quad j = 1, \ldots, 70.$$
$$X_j|\theta_j \overset{ind}{\sim} B(n_j, \theta_j).$$

Then the historical data are used to determine the hyperparameters α, β. A way for estimating α and β is *the method of moments* – given a sample $\theta_1, \ldots, \theta_n$ from $\text{Beta}(\alpha, \beta)$ one calculates the sample mean $\bar{\theta} = \frac{1}{n}\sum_{j=1}^{n} \theta_j$ and variance $\frac{1}{n}\sum_{j=1}^{n}(\theta - \bar{\theta})^2$ and treats them as they were $E\theta$ and $\text{Var}(\theta)$. For Beta distribution, the mean and variance uniquely determine the parameters (Exercise 4) and so α, β can be estimated. In our case, we do not have a sample from prior, hence instead of θ_j, we use x_j/n_j, $j = 1, \ldots, 70$ (ratio x_j/n_j is MLE for θ_j). As reported in [5], in rat tumor case, the sample mean and variances (for 70 historical data) are 0.136 and 0.103^2, resp. Using these estimates instad of mean and variance, one gets the estimates for hyperparameters $\hat{\alpha} = 1.383$, $\hat{\beta} = 8.787$. With these hyperparameters the posterior distribution for the current estimate is $\theta_{71} \sim \text{Beta}(\hat{\alpha} + x_{71}, \hat{\beta} + (n_{71} - x_{71}))$, where $x_{71} = 4$ and $n_{71} = 14$, thus $\theta_{71} \sim \text{Beta}(5.383, 18.787)$. The posterior mean, mode and variance are thus:

$$\frac{5.383}{5.383 + 18.787} \approx 0.223, \quad \frac{5.383 - 1}{5.383 + 18.787 - 2} \approx 0.198,$$
$$\frac{5.383}{24.17} \cdot \frac{8.787}{24.17} \cdot \frac{1}{25.17} \approx 0.0069.$$

We see that the posterior mean 0.223 is lower than $4/14 = 0.2857\cdots$ – the historical data suggest that in the current estimate of the rate us untypically high. On the other hand the posterior mean is higher that 0.136 – the mean of the historical data, which also the prior mean. This is because the posterior mean averages the observed ratio 4/14 (high) and the prior mean 0.136 (low).

2.2 Dirichlet-categorical and Dirichlet-multinomial model

Dirichlet distribution. Dirichlet distribution is the generalization of Beta-distribution. *Aggregation*: when $(\theta_1, \ldots, \theta_k) \sim \text{Dir}(\alpha_1, \ldots, \alpha_k)$, then

$$(\theta_1 + \theta_2, \theta_3, \ldots, \theta_k) \sim \text{Dir}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_k).$$

From aggregation, it follows that the marginals (components) have Beta-distribution:

$$\theta_i \sim \text{Beta}(\alpha_i, |\alpha| - \alpha_i), \quad |\alpha| := \alpha_1 + \cdots + \alpha_k.$$

Hence the moments (recall the moments of Beta distribution)

$$E\theta_i = \frac{\alpha_i}{|\alpha|}, \quad \text{Var}(\theta_i) = \frac{\alpha_i}{|\alpha|}\frac{|\alpha| - \alpha_i}{|\alpha|}\frac{1}{|\alpha| + 1}, \quad i = 1, \ldots, k.$$

When $\alpha_i > 1 \ \forall i$, then the mode of Dirichlet distribution is

$$\left(\frac{\alpha_1}{|\alpha| - k}, \ldots, \frac{\alpha_k}{|\alpha| - k}\right).$$

The moments (Exercise 5):

$$E\left(\theta_1^{r_1} \cdots \theta_k^{r_k}\right) = \frac{B(\alpha_1 + r_1, \ldots, \alpha_k + r_k)}{B(\alpha_1, \ldots, \alpha_k)}.$$

From the formula above, it follows that (Exercise 5)

$$\text{Cov}(\theta_i, \theta_j) = -\frac{\alpha_i \alpha_j}{|\alpha^2|(|\alpha| + 1)}.$$

Given a vector $(p_1, \ldots, p_k) \in S_{k-1}$ (a vector of probabilities), the *categorical distribution* $\text{Cat}(p_1, \ldots, p_k)$ is the discrete distribution with atoms in set $\{1, \ldots, k\}$ and density (probability mass function) $f(i) = p_i$, $i = 1, \ldots, k$. (The enumeration of values is not important, to get a direct generalization of Bernoulli distribution, the atoms should actually be $\{0, \ldots, k-1\}$, but $\{1, \ldots, k\}$ is more conventual. Another option is to label the categories as the vectors $e_1, \ldots, e_k$, where e_i is a k-dimensional vector with i-th coordinate being 1 and rest are zeros. Then then sum of n iid categorical random vectors has multinomial distribution).

Dirichlet-categorical model. The *Dirichlet-categorical* model is now a straightforward generalization of Beta-Bernoulli model:

$$\theta = (\theta_1, \ldots, \theta_k) \sim \text{Dir}(\alpha_1, \ldots, \alpha_k)$$
$$X_1, \ldots, X_n | \theta \overset{iid}{\sim} \text{Cat}(\theta).$$

Thus $x = (x_1, \ldots, x_n) \in \{1, \ldots, k\}^n$, the hyperparameters are $\alpha_1, \ldots, \alpha_k$. Let $(n_1, \ldots, n_k)$ be the counts in $x_1, \ldots, x_n$, i.e. n_i is the number of i-s in $x_1, \ldots, x_n$. When $X_1, \ldots, X_n$ are iid from $\text{Cat}(\theta)$, then the random counts $N = (N_1, \ldots, N_k) \sim \text{Multin}(n, \theta_1, \ldots, \theta_k)$ ($N_i = \sum_{j=1}^n I_{\{i\}}(X_j)$.)

Hence with the counts $N = (N_1, \ldots, N_k)$ instead of $X_1, \ldots, X_n$, we obtain *Dirichlet-multinomial model*:

$$\theta \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$$
$$N|\theta \sim \mathrm{Multin}(n, \theta).$$

When $X_1, \ldots, X_n$ are obtained by Dirichlet-categorical model – we shall call them Dirichlet-categorical random variables for short – then the counts N follow Dirichlet-Multinomial model.

(Marginal) distribution. The joint distribution of Dirichlet-categorical random variables $X = (X_1, \ldots, X_n)$ (marginal distribution in Bayes jargon) is now following. The sample space is $\mathcal{X} = \{1, \ldots, k\}^n$ and for any $x \in \mathcal{X}$,

$$\begin{aligned} P(X = x) &= \int_{S_{k-1}} P(X = x|\theta)\pi(\theta)d\theta = \int_{S_{k-1}} \prod_{j=1}^{k} \theta_j^{n_j} \frac{\prod_{j=1}^{k} \theta_j^{\alpha_j - 1}}{B(\alpha_1, \ldots, \alpha_k)} d\theta \\ &= \frac{1}{B(\alpha_1, \ldots, \alpha_k)} \int_{S_{k-1}} \prod_{j=1}^{k} \theta_j^{\alpha_j + n_j - 1} d\theta \\ &= \frac{B(\alpha_1 + n_1, \ldots, \alpha_k + n_k)}{B(\alpha_1, \ldots, \alpha_k)}. \end{aligned}$$

We already know that $X_1, \ldots, X_n$ are exchangeable, hence they are identically distributed ($X_1 \sim \mathrm{Cat}(\alpha_1/|\alpha|, \ldots, \alpha_k/|\alpha|)$ but positively correlated (see Exercise 6).

Since the number of vectors $x \in \{1, \ldots, k\}^n$ with counts $n_1, \ldots, n_k$ is $\frac{n!}{n_1! \cdots n_k!}$, we get the marginal distribution Dirichlet-multinomial model:

$$P(N = (n_1, \ldots, n_k)) = \frac{n!}{n_1! \cdots n_k!} \frac{B(\alpha_1 + n_1, \ldots, \alpha_k + n_k)}{B(\alpha_1, \ldots, \alpha_k)} = \frac{\Gamma(|\alpha|)\Gamma(n+1)}{\Gamma(n + |\alpha|)} \prod_{i=1}^{k} \frac{\Gamma(n_i + \alpha_i)}{\Gamma(\alpha_i)\Gamma(n_i + 1)}.$$

This distribution is known as *Dirichlet-multiomial distribution*, denoted as $\mathrm{DirMult}(n; \alpha_1, \ldots, \alpha_k)$. Observe that with $\alpha_i = 1$, $i = 1, \ldots, k$, the distribution is uniform over its all possible values. When

$$(N_1, \ldots, N_k) \sim \mathrm{DirMult}(n; \alpha_1, \ldots, \alpha_k),$$

then $N_i \sim \mathrm{BetaBin}(n, \alpha_i, |\alpha| - \alpha_i)$ (Exercise 7) so that

$$EN_i = n\frac{\alpha_i}{|\alpha|}, \quad \mathrm{Var}(N_i) = \frac{n\alpha_i(|\alpha| - \alpha_i)}{|\alpha|^2} \frac{(|\alpha| + 1)}{(|\alpha| + n)}.$$

The correlation (Exercise 7):

$$\mathrm{Cov}(N_i, N_j) = -\frac{\alpha_i \alpha_j}{|\alpha|^2} \frac{(|\alpha| + n)}{(|\alpha| + 1)}.$$

Just like in the case of Dirichlet distribution, Dirichlet-multinomial distribution has aggregation property (Exercise 8): when

$$(N_1, \ldots, N_k) \sim \text{DirMult}(n; \alpha_1, \ldots, \alpha_k),$$

then

$$(N_1 + N_2, N_3, \ldots, N_k) \sim \text{DirMult}(n; \alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_k).$$

Polya urn and posterior predicative distribution. In the beginning there are $|\alpha|$ balls with k different colors enumerated as $1, \ldots, k$: α_i balls of color i (not necessarily integers), after a ball with color i has been taken put of the urn it will be returned to the urn together with another ball of the same color. Let X_t be the color of t-th ball taken out of the urn. Thus

$$P(X_1 = i) = \frac{\alpha_i}{|\alpha|}, \quad P(X_{t+1} = i | X_1 = x_1, \ldots, X_t = x_t) = \frac{\alpha_i + n_i(t)}{|\alpha| + t} \quad (2.5)$$

where $n_i(t) := \sum_{j=1}^{t} I_{\{i\}}(x_j)$.

Just in the case of two colors, it is easy to see that the random variables are infinitely exchangeable so that according to deFinetti-Hewitt-Savage theorem there exists a prior measure π on S_{k-1} so that so that the model (1.6) holds:

$$\theta = (\theta_1, \ldots, \theta_k) \sim \pi$$
$$X_1, \ldots, X_n | \theta \overset{iid}{\sim} \text{Cat}(\theta)$$

Since for any n, the random vector $(X_1, \ldots, X_n)$ has the same distribution as in Dirichlet-categorical model (Exercise 3), the measure π is $\text{Dir}(\alpha_1, \ldots, \alpha_k)$, so Polya urn corresponds to Dirichlet-categorical model. Hence, when N_i is the number of i-color balls taken out of the urn (during n draws), then $(N_1, \ldots, N_k) \sim \text{DirMult}(n; \alpha_1, \ldots, \alpha_k)$. From deFinetti-Hewitt-Savage theorem it also follows that

$$(N_1/n, \ldots, N_k/n) \to (\theta_1, \ldots, \theta_k), \quad \text{a.s. where } (\theta_1, \ldots, \theta_k) \sim \text{Dir}(\alpha_1, \ldots, \alpha_k).$$

The probabilities (2.5) (for $i = 1, \ldots, k$) form the posterior predicative distribution. The generalization of Laplace rule of succession (the case $\alpha_i = 1$) is thus

$$P(X_{n+1} = i | X_1 = x_1, \ldots, X_n = x_n) = \frac{1 + n_i}{k + n}, \quad i = 1, \ldots, k.$$

Posterior density. Like for Beta-Bernoulli and Beta-binomial model, also for Dirichlet-categorical and Dirichlet-multinomial case the posterior distribution is Dirichlet distribution. Let $x = (x_1, \ldots, x_n) \in \{1, \ldots, k\}^n$ (and

$n_i = \sum_{j=1}^n I_{\{i\}}(x_j)$ as usually):

$$\pi(\theta|x) \propto \pi(\theta) f(x|\theta) = \frac{1}{B(\alpha_1, \ldots, \alpha_k)} \prod_{i=1}^k \theta_i^{\alpha_i - 1} \prod_{i=1}^k \theta_i^{n_i} = \frac{1}{B(\alpha_1, \ldots, \alpha_k)} \prod_{i=1}^k \theta_i^{\alpha_i + n_i - 1},$$

hence the posterior distribution is $\mathrm{Dir}(\alpha_1 + n_1, \ldots, \alpha_k + n_k)$. The same hods for Dirichlet-multinomial model, where observations are the counts: $x = (n_1, \ldots, n_k)$.

Posterior mean, mode and variance, interpretation. Let $x = (n_1, \ldots, n_k)$. The posteriror distribution is Dirichlet distribution, so it is easy to find posertior mean, variances and covariances of components. Posterior mean is the vector $E[\theta|x] = (E[\theta_1|x], \ldots, E[\theta_k|x])'$, where

$$E[\theta_i|x] = \frac{\alpha_i + n_i}{n + |\alpha|} = \frac{n_i}{n} \cdot \frac{n}{n + |\alpha|} + \frac{\alpha_i}{|\alpha|} \cdot \frac{|\alpha|}{n + |\alpha|}.$$

Again, $E[\theta_i|x]$ is the weighted average of sample mean n_i/n and prior mean $\alpha_i/|\alpha|$, when n increases, the prior has less influence. The posterior variance of θ_i is

$$\mathrm{Var}[\theta_i|x] = \frac{E[\theta_i|x](1 - E[\theta_i|x])}{n + |\alpha| + 1} = \frac{n_1 + \alpha_i}{n + |\alpha|} \cdot \frac{n - n_1 + |\alpha| - \alpha_i}{n + |\alpha|} \cdot \frac{1}{n + |\alpha| + 1}.$$

The posterior covariance

$$\mathrm{Cov}[(\theta_i, \theta_j)|x] = -\frac{E[\theta_i|x] E[\theta_j|x]}{|\alpha| + n + 1}.$$

We see that the covariance decreases as n grows.

Pre-election polling example. See [5], sec 3.4. A survey: 1447 adults from U.S.A. were asking their preferences in upcoming presidental election (1988). Out of 1447 person, $n_1 = 727$ supported G. Bush, $n_2 = 583$ supported M. Dukakis and $n_3 = 137$ supported other candidates or expressed no opinion. Assuming all responses are independent and every respondent is a categorical random variable (three categories) with the same probabilities/distribution $\theta = (\theta_1, \theta_2, \theta_3)$, we end up with (n_1, n_2, n_3) being a realization of a multinomial random vector. Assuming that every possible vector of probabilities is a priori alike, we end up with Dirichlet-multinomial model with hyperparameters $\alpha_1 = \alpha_2 = \alpha_3 = 1$. Then the posterior distribution of θ is $\mathrm{Dir}(728, 584, 138)$. The goal of the survey is to estimate the difference $\theta_1 - \theta_2$. In [5], 1000 draws from $\mathrm{Dir}(728, 584, 138)$-distribution (each draw is a 3-dim vector) has been made, and for every draw $\theta_1 - \theta_2$ is computed. In this way, the posterior distribution of $\theta_1 - \theta_2$ is estimated (since posterior distribution has analytic form, that distribution could ce calculated exactly as well). In fact, far all 1000 draws it hold that $\theta_1 > \theta_2$ so the estimated posterior probability that G. Bush has more support than M. Dukakis is practically 1.

2.3 Gamma-Poisson model

When $\theta \sim \text{Gamma}(\alpha, \beta)$, then

$$E\theta = \frac{\alpha}{\beta}, \quad \text{Var}(\theta) = \frac{\alpha}{\beta^2}.$$

Mode of Gamma-distribution is 0 for $\alpha < 1$ and $\frac{(\alpha-1)}{\beta}$ for $\alpha \geq 1$.

Gamma-Poisson model:

$$\theta \sim \text{Gamma}(\alpha, \beta)$$
$$X_1, \ldots, X_n | \theta \overset{iid}{\sim} \text{Po}(\theta)$$

Hence, for given θ

$$f(x|\theta) = e^{-n\theta} \frac{\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}, \quad x = (x_1, \ldots, x_n) \in \mathbb{N}^n.$$

Marginal distribution is negative multinomial. Marginal distribution of $X_1, \ldots, X_n$ – the joint density:

$$\begin{aligned} f(x) &= \int_0^\infty f(x|\theta)\pi(\theta)d\theta = \int_0^\infty \exp[-n\theta] \frac{\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp[-\beta\theta] d\theta \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{\prod_{i=1}^n x_i!} \int_0^\infty \exp[-\theta(n+\beta)] \theta^{\sum_{i=1}^n x_i + \alpha - 1} d\theta \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{\prod_{i=1}^n x_i!} \frac{\Gamma(\sum_i x_i + \alpha)}{(\beta+n)^{\sum_i x_i + \alpha}} \\ &= \frac{\Gamma(\sum_i x_i + \alpha)}{\Gamma(\alpha) \prod_{i=1}^n x_i!} \Big(\frac{\beta}{\beta+n}\Big)^\alpha \Big(\frac{1}{\beta+n}\Big)^{\sum_i x_i}. \end{aligned}$$

When $n = 1$, then with $x = x_1$, we get the density of NB(α, $\frac{\beta}{\beta+1}$):

$$f(x) = \frac{\Gamma(x+\alpha)}{\Gamma(\alpha)x!} \Big(\frac{\beta}{\beta+1}\Big)^\alpha \Big(\frac{1}{\beta+1}\Big)^x.$$

For that reason negative binomial distribution is sometimes referred to as *Gamma-Poisson* distribution.
When α is integer, then

$$\frac{\Gamma(\sum_{i=1}^n x_i + \alpha)}{\Gamma(\alpha) \prod_{i=1}^n x_i!} = \binom{\sum_i x_i + \alpha - 1}{x_1 \cdots x_n \ \alpha - 1},$$

so that for $n > 1$ the marginal distribution of $(X_1, \ldots, X_n)$ is a kind of multinomial generalization of negative binomial, called *negative multinomial*

distribution. So the (marginal) distribution $(X_1, \ldots, X_n)$ is negative multinomial distribution: the random variables $X_1, \ldots, X_n$ are identically distributed, $X_i \sim \mathrm{NB}(\alpha, \frac{\beta}{\beta+1})$, thus (exercise 10)

$$EX_i = \frac{\alpha}{\beta}, \quad \mathrm{Var}(X_i) = \frac{\alpha(1+\beta)}{\beta^2} \tag{2.6}$$

and they are positively correlated: $\mathrm{Cov}(X_i, X_j) = \frac{\alpha}{\beta^2}$.

Posterior and posterior predicative distribution. Posterior density: $x = (x_1, \ldots, x_n)$,

$$\begin{aligned} f(x|\theta)\pi(\theta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{\prod_{i=1}^n x_i!} \exp[-\theta(n+\beta)]\theta^{\sum_{i=1}^n x_i + \alpha - 1} \\ &\propto \exp[-\theta(n+\beta)]\theta^{\sum_{i=1}^n x_i + \alpha - 1}. \end{aligned}$$

Therefore $\pi(\theta|x)$ is the density of $\mathrm{Gamma}(\sum_{i=1}^n x_i + \alpha, n+\beta)$, so that

$$\theta|x \sim \mathrm{Gamma}(\sum_{i=1}^n x_i + \alpha, n+\beta). \tag{2.7}$$

Posterior predicative distribution (exercise 10) is $\mathrm{NB}(\sum_i x_i + \alpha, \frac{\beta+n}{\beta+n+1})$:

$$P(X_{n+1} = i | X_1 = x_1, \ldots, X_n = x_n) = \frac{\Gamma(i + \sum_{i=1}^n x_i + \alpha)}{\Gamma(\sum_{i=1}^n x_i + \alpha) i!} \Big(\frac{\beta+n}{\beta+n+1}\Big)^{\sum_i x_i + \alpha} \Big(\frac{1}{\beta+n+1}\Big)^i.$$

The interpretation of hyperparameters: β is the size of "prior sampleänd α is the sum of observations in that sample.

Posterior mean, mode, variance. Posterior mean, variance and mode – the mean and variance of $\mathrm{Gamma}(\sum_{i=1}^n x_i + \alpha, n+\beta)$ distribution – is

$$E[\theta|x] = \frac{\sum_{i=1}^n x_i + \alpha}{n+\beta} = \frac{\sum_{i=1}^n x_i}{n} \frac{n}{n+\beta} + \frac{\alpha}{\beta} \frac{\beta}{n+\beta}, \quad \mathrm{Var}[\theta|x] = \frac{\sum_{i=1}^n x_i + \alpha}{(n+\beta)^2}.$$

The posterior mode is

$$\frac{\sum_{i=1}^n x_i + \alpha - 1}{n+\beta}.$$

We see that posterior mean is again the convex combination of sample mean $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ and prior mean $\frac{\alpha}{\beta}$. The bigger is n, the smaller is the prior influence.

Kidney cancer example. The parameter of Poisson distribution is also its mean. Therefore, in practice often the i-th observation is generated from $\mathrm{Po}(m_i\theta)$-distribution, where m_i is known *exposure* and θ is the parameter of interest (sometimes called *rate*). The Gamma-Poisson model then

$$\begin{aligned} \theta &\sim \mathrm{Gamma}(\alpha, \beta) \\ X_1, \dots, X_n|\theta &\overset{ind}{\sim} \mathrm{Po}(m\theta). \end{aligned}$$

It is easy to see that posterior distribution of θ is again Gamma-distribution (Exercise 10):

$$\theta|x \sim \mathrm{Gamma}\big(\sum_{i=1}^{n} x_i + \alpha, nm + \beta\big). \tag{2.8}$$

In kidney cancer example ([5], sec. 2.7), the model is the following:

$$\begin{aligned} \theta_1, \dots, \theta_n &\overset{iid}{\sim} \mathrm{Gamma}(\alpha, \beta) \\ X_j|\theta_j &\overset{ind}{\sim} \mathrm{Po}(10m_j\theta_j), \quad j = 1, \dots, n. \end{aligned}$$

Here n is then umber of counties, m_j is the population of the county, θ_j is the underlying rate in units of death per person per year in the j-th county and X_j is the number of kidney cancer deaths in county j during 10 years. Given the observation x_j, the frequentist (MLE) estimate of θ_j is just $x_j/10m_j$.
For determining the hyperparameters α and β, the empirical prior can be used. In [5] it has been done as follows: first observe that (Exercise 10) that for every j,

$$X_j \sim \mathrm{NB}\Big(\alpha, \frac{\beta}{10m_j + \beta}\Big). \tag{2.9}$$

Thus

$$EX_j = 10m_j\frac{\alpha}{\beta}, \quad \mathrm{Var}(X_j) = \frac{\alpha}{\beta^2}10m_j(10m_j + \beta),$$

so that

$$E\big(\frac{X_j}{10m_j}\big) = \frac{\alpha}{\beta}, \quad \mathrm{Var}\Big(\frac{X_j}{10m_j}\Big) = \frac{1}{10m_j}\frac{\alpha}{\beta} + \frac{\alpha}{\beta^2}.$$

The idea is to match the theoretical and observed moments (method of moments). For means it is easy:

$$\bar{x} := \frac{1}{n}\sum_{j=1}^{n}\frac{x_j}{10m_j} = \frac{\alpha}{\beta}.$$

Variances of X_j depend on m_j, so after calculating sample variance

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{x_j}{10m_j} - \bar{x}\Big)^2,$$

it will be matched with average variance:

$$s^2 = \frac{\alpha}{\beta^2} + \frac{\alpha}{\beta}\frac{1}{n}\sum_{j=1}^{n}\frac{1}{10m_j}.$$

Now find α and β. In [5], the values $\alpha = 20$ and $\beta = 430000$ are obtained. The prior mean is $\alpha/\beta = 4.65 \times 10^{-5}$. For each $j = 1, \ldots, n$, from (2.8),

$$\theta_j|x_j \sim \text{Gamma}(20 + x_j, 430000 + 10m_j),$$
$$E[\theta_j|x_j] = \frac{20 + x_j}{430000 + 10m_j}, \quad \text{Var}[\theta_j|x_j] = \frac{20 + x_j}{(430000 + 10m_j)^2}.$$

The posterior mean is the weighted average of the raw rate $x_j/10m_j$ and prior mean α/β:

$$E[\theta_j|x_j] = \frac{20 + x_j}{430000 + 10m_j} = \frac{20}{430000}\cdot\frac{430000}{430000 + 10m_j} + \frac{x_j}{10m_j}\cdot\frac{10m_j}{430000 + 10m_j}.$$

The smaller is population m_j, the closer is the posterior mean to prior mean 4.65×10^{-5}, this reduces the effect of the small population, see Sec 2.7 of [5] for closer discussion.

Remark. Since m_j is typically rather big, it holds that $s^2 \approx \frac{\alpha}{\beta^2}$ and so the empirical prior method is actually almost the same as in previously considered rat tumor example: since $\theta_1, \ldots, \theta_n$ are iid from Gamma(α, β), one would like to match the empirical mean and variance of parameters with $\frac{\alpha}{\beta}$ (mean of Gamma) and $\frac{\alpha}{\beta^2}$ (variance of Gamma) respectively. However, the parameters θ_j are not directly observable, so we use $x_j/10m_j$ instead. This yields to the system of equations $\bar{x} = \alpha/\beta$, $s^2 = \alpha/\beta^2$.

2.4 Gamma-Exponential model

The model:

$$\theta \sim \text{Gamma}(\alpha, \beta)$$
$$X_1, \ldots, X_n|\theta \overset{iid}{\sim} \text{Exp}(\theta)$$

Hence, for given θ

$$f(x|\theta) = \theta^n \exp[-\theta\sum_{i=1}^{n} x_i], \quad x = (x_1, \ldots, x_n) \in [0, \infty)^n.$$

Marginal distribution. The joint density

$$f(x,\theta) = \theta^n \exp[-\theta \sum_{i=1}^n x_i] \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp[-\beta\theta] = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{n+\alpha-1} \exp[-\theta(\sum_{i=1}^n x_i + \beta)].$$

The marginal density ($x = (x_1, \ldots, x_n)$),

$$\begin{aligned} f(x) &= \int_0^\infty f(x,\theta) d\theta = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \theta^{n+\alpha-1} \exp[-\theta(\sum_{i=1}^n x_i + \beta)] d\theta \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(n+\alpha)}{(\sum_{i=1}^n x_i + \beta)^{n+\alpha}} = (\alpha+n-1)(\alpha+n-2)\cdots\alpha \frac{\beta^\alpha}{(\sum_{i=1}^n x_i + \beta)^{\alpha+n}}. \end{aligned}$$

For $n = 1$, thus

$$f(x) = \frac{\alpha\beta^\alpha}{(x+\beta)^{\alpha+1}},$$

hence $X_i \sim \text{Lomax}(\alpha, \beta)$. Thus $X_1, X_2, \ldots, X_n$ are positively correlated Lomax-distributed random variables with mean, variances and covariation as follows (exercises 11):

$$EX_i = \frac{\beta}{\alpha - 1}, \quad \text{Var}X_i = \begin{cases} \frac{\beta^2\alpha}{(\alpha-1)^2(\alpha-2)}, & \alpha > 2; \\ \infty, & \alpha \le 2. \end{cases} \tag{2.10}$$

$$\text{Cov}(X_i, X_j) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \quad \alpha > 2. \tag{2.11}$$

Posterior distribution, mean, variance.

$$\pi(\theta|x) \propto \theta^{n+\alpha-1} \exp[-\theta(\sum_{i=1}^n x_i + \beta)],$$

so that the posterior distribution is Gamma:

$$\theta|x \sim \text{Gamma}\big(n+\alpha, \sum_{i=1}^n x_i + \beta\big). \tag{2.12}$$

The posterior mean is

$$E[\theta|x] = \frac{n+\alpha}{\sum_{i=1}^n x_i + \beta} = \frac{n}{\sum_{i=1}^n x_i} \cdot \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i + \beta} + \frac{\alpha}{\beta} \cdot \frac{\beta}{\sum_{i=1}^n x_i + \beta} = \bar{x}^{-1} \frac{\bar{x}}{\bar{x} + \frac{\beta}{n}} + \frac{\alpha}{\beta} \frac{\frac{\beta}{n}}{\bar{x} + \frac{\beta}{n}},$$

where $\bar{x} = n^{-1} \sum_{i=1}^n$ (sample average). When $X \sim \text{Exp}(\theta)$, then $EX = \frac{1}{\theta}$, hence $\bar{x}^{-1}$ is the sample based estimate of θ (also MLE). We see that posterior mean is again a convex combination of sample based estimate and prior mean $\frac{\alpha}{\beta}$. The bigger n, the smaller the prior influence.

Posterior variance:

$$\mathrm{Var}[\theta|x] = \frac{n+\alpha}{(\sum_{i=1}^n x_i + \beta)^2} = \frac{1+\frac{\alpha}{n}}{n(\bar{x}+\frac{\beta}{n})^2}.$$

Interpretation of hyperparameters: α is now the size of "prior sample" and β is the sum of observation in that prior sample. Recall that in the Gamma-Poisson model the meaning of hyperparameters was opposite (β the sample size and α the observations).

Posterior predicative distribution. Prior predicative distribution is Lomax:

$$X_{n+1}|x \sim \mathrm{Lomax}(n+\alpha, \sum_{i=1}^n x_i + \beta), \tag{2.13}$$

Thus the posterior predicative density is

$$f(x_{n+1}|x) = \frac{(\alpha+n)\big(\sum_{i=1}^n x_i + \beta\big)^{\alpha+n}}{\big(\sum_{i=1}^{n+1} x_i + \beta\big)^{\alpha+n+1}}.$$

Remark. Observe that in the joint (hence marginal and posterior) distribution, the data $x = (x_1, \ldots, x_n)$ enter to the formulas through the sum $\sum_{i=1}^n x_i$, only. In other words, when instead of x only the sum $\sum_{i=1}^n x_i$ were observed, nothing would change (the sum is sufficient statistic here). It is also well known that when $X_1, \ldots, X_n$ are iid random variables with $\mathrm{Exp}(\theta)$ distribution, then the sum follows Gamma distribution (Exercise (11)) $S = X_1 + \cdots + X_n \sim \mathrm{Gamma}(n, \theta)$. Hence the Gamma-exponential model can be also represented as *Gamma-Gamma model*:

$$\begin{aligned} \theta &\sim \mathrm{Gamma}(\alpha, \beta) \\ X|\theta &\sim \mathrm{Gamma}(n, \theta). \end{aligned}$$

An example: earthquakes. Exponential distribution is a common model for waiting times. Assume that waiting times between earthquakes follow exponential distribution with parameter θ. The parameter θ is the average rate an earthquakes occur per hour (because $1/\theta$ is the average waiting time). Assume $\mathrm{Gamma}(4, 3)$ prior. Hence according to our prior knowledge, the average rate is $E\theta = 4/3$. Since $\theta \sim \mathrm{Gamma}(4, 3)$, $1/\theta \sim \mathrm{InvGamma}(4, 3)$ and $E(1/\theta) = 1$ (because the mean of $\mathrm{InvGamma}(\alpha, \beta)$ distribution is $\beta/(\alpha-1)$). Hence, based on our prior knowledge the expected waiting time between earthquakes is 1. In fact, we know that the waiting time between any two earthquakes has $\mathrm{Lomax}(4, 3)$-distribution, implying that the mean is 1 (because the mean of $\mathrm{Lomax}(\alpha, \beta)$ is $\beta/(\alpha-1)$).

Suppose that two earthquakes were observed: the first has waiting time 3.2 hours, the second 1.6 hours. Thus $x_1 = 3.2$ and $x_2 = 1.6$. The sample average $\bar{x} = 4.8/2 = 2.4$. Hence, the empirical estimate of expected waiting time is 2.4 with the corresponding rate estimate $1/2.4 = 0.4166$. We see that data based estimate differ from prior estimate. The Bayesian approach combines prior knowledges with data. The posterior distribution is now $\theta|x \sim \text{Gamma}(6, 7.8)$ with mean $E[\theta|x] = 0.769$. We know that the posterior mean is a convex combination of prior mean $E\theta$ and $1/\bar{x}$, hence it must be between these numbers: $1/\bar{x} = 0.4166 < 0.769 < 4/3 = E\theta$. The conditional distribution of the waiting time to the third earthquake given the data is Lomax: $X_3|x \sim \text{Lomax}(6, 7.2)$ with mean $7.2/5 = 1.44$. Again, we see that Bayesian estimate is between prior and data based estimates: $\bar{x} = 2.4 > 1.44 > 1 = EX_3$.

Now consider the case 100 earthquakes were observed with total waiting time 63.09 ($\sum_{i=1}^{n} x_i = 63.09$). Hence the empirical estimates of expected waiting time and rate are $\bar{x} = 0.6309$ and $1/\bar{x} = 1.585$. The posterior distribution is now $\theta|x \sim \text{Gamma}(104, 66.09)$ with mean $E[\theta|x] = 104/66.09 = 1.5736$. Thus $1/\bar{x} = 1.585 > 1.5736 > 1.333 = E\theta$. The conditional waiting time to the 101-th earthquake has Lomax distribution: $X_{101} \sim \text{Lomax}(104, 66.09)$ with mean $E[X_{101}|x] = 66.09/103 = 0.641$. Thus $\bar{x} = 0.6309 < 0.641 < 1 = EX_{101}$. We see that Bayesian estimates are pretty close to the empirical estimates. This is due to the fact that $n = 100$ is rather big in comparison with $\beta = 3$.

2.5 Normal models

Recall: any density of θ in the form $f(\theta) = \exp[-A\theta^2 + B\theta + C]$, $A > 0$ is a normal density with mean $\mu = \frac{B}{2A}$ and variance $\sigma^2 = \frac{1}{2A}$, because

$$f(\theta) = \exp[-A\theta^2 + B\theta + C] = \exp[-\frac{(\theta^2 - 2\mu\theta + \mu^2)}{2\sigma^2} + D] \propto \exp[-\frac{(\theta - \mu)^2}{2\sigma^2}].$$

2.5.1 Known variance

A standard model (variance known, prior on mean)

$$\theta \sim \mathcal{N}(\mu, \tau^2)$$
$$X_1, \ldots, X_n|\theta \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2).$$

Marginal distribution. The joint density of $(X_1, \dots, X_n, \theta)$ has form ($f(x|\theta)$ is normal density)

$$f(x_1, \dots, x_n, \theta) = \prod_{i=1}^{n} f(x_i|\theta)\pi(\theta)$$
$$\propto \exp\Big[-\frac{1}{2}\Big(\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2} - 2\sum_{i=1}^{n} \frac{(x_i-\mu)(\theta-\mu)}{\sigma^2} + (\theta-\mu)^2(\frac{n}{\sigma^2} + \frac{1}{\tau^2})\Big)\Big].$$

We see that $f(x_1, \dots, x_n, \theta)$ is jointly normal density with precision matrix (inverse of covariance matrix):

$$\begin{pmatrix} \frac{1}{\sigma^2} & 0 & \cdots & 0 & -\frac{1}{\sigma^2} \\ 0 & \frac{1}{\sigma^2} & \cdots & 0 & -\frac{1}{\sigma^2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{1}{\sigma^2} & -\frac{1}{\sigma^2} \\ -\frac{1}{\sigma^2} & -\frac{1}{\sigma^2} & \cdots & -\frac{1}{\sigma^2} & \frac{n}{\sigma^2} + \frac{1}{\tau^2} \end{pmatrix}.$$

Marginals of jointly normal random variables are jointly normal as well, hence $(X_1, \dots, X_n)$ is multivariate normal random vector. The mean vector and covariance matrix of $(X_1, \dots, X_n)$ are

$$\mu = \begin{pmatrix} \mu \\ \mu \\ \cdots \\ \mu \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma^2+\tau^2 & \tau^2 & \cdots & \tau^2 \\ \tau^2 & \sigma^2+\tau^2 & \cdots & \tau^2 \\ \cdots & \cdots & \cdots & \cdots \\ \tau^2 & \tau^2 & \cdots & \sigma^2+\tau^2 \end{pmatrix}. \qquad (2.14)$$

To see that the covariance matrix is just as above, recall that $X_1, \dots, X_n$ are identically distributed random variables with

$$\mathrm{Cov}(X_i, X_j) = \mathrm{Var}\big(E[X_1|\theta]\big) = \mathrm{Var}(\theta) = \tau^2,$$
$$\mathrm{Var}(X_1) = \mathrm{Var}\big(E[X_1|\theta]\big) + E[\mathrm{Var}(X_1|\theta)] = \tau^2 + \sigma^2.$$

Posterior distribution, mean and variance. From the joint density $f(x_1, \dots, x_n, \theta)$ we see that ($x = (x_1, \dots, x_n)$)

$$\pi(\theta|x) \propto \exp\Big[-\frac{1}{2}(\frac{n}{\sigma^2} + \frac{1}{\tau^2})\Big((\theta-\mu)^2 - 2\sum_{i=1}^{n} \frac{(x_i-\mu)(\theta-\mu)}{\sigma^2}(\frac{n}{\sigma^2} + \frac{1}{\tau^2})^{-1}\Big)\Big].$$

Denote

$$\tau_n^2 = (\frac{n}{\sigma^2} + \frac{1}{\tau^2})^{-1} = \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}$$

Thus

$$\pi(\theta|x) \propto \exp\Big[-\frac{1}{2\tau_n^2}\Big(\theta^2 - 2\theta\mu - \frac{2}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)\theta\tau_n^2\Big)\Big]$$
$$= \exp\Big[-\frac{1}{2\tau_n^2}\Big(\theta^2 - 2\theta\big(\mu + \sum_{i=1}^{n}(x_i-\mu)\frac{\tau_n^2}{\sigma^2}\big)\Big)\Big].$$

Therefore $\pi(\theta|x)$ is normal density with variance τ_n^2 and mean ($\bar{x} = \sum_{i=1}^n x_i/n$)

$$\mu_n := \mu + (n\bar{x} - n\mu)\frac{\tau^2}{n\tau^2 + \sigma^2} = \frac{\mu\sigma^2 + n\tau^2\bar{x}}{n\tau^2 + \sigma^2} = \frac{\frac{\mu}{\tau^2} + \frac{n}{\sigma^2}\bar{x}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} = (\frac{\mu}{\tau^2} + \frac{n\bar{x}}{\sigma^2})\tau_n^2.$$

Thus

$$\theta|x \sim \mathcal{N}(\mu_n, \tau_n^2).$$

We see that again posterior mean is the average of sample mean and prior mean:

$$\mu_n = \bar{x}\frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} + \mu\frac{\frac{1}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}},$$

the bigger n, the smaller is prior influence. We also see that $\tau_n^2 \to 0$ when $n \to \infty$. When $\sigma^2 = 1$, then posterior mean is

$$\mu_n = \frac{\mu/\tau^2 + \sum_{i=1}^n x_i}{n + 1/\tau^2}$$

Thus $1/\tau^2$ can be considered as the size of "prior sample "with prior sample average μ.

Posterior predicative distribution. Recall the posterior predicative density

$$f(x_{n+1}|x) = \int f(x_{n+1}|\theta)\pi(\theta|x)d\theta.$$

Hence the posteriror predicative distribution is the marginal distribution with $\pi(\theta|x)$ as prior density, so that we immediately get (exercise 12) that $f(x_{n+1}|x)$ is a normal density with mean μ_n and variance $\tau_n^2 + \sigma^2$, i.e.

$$X_{n+1}|x \sim \mathcal{N}(\mu_n, \tau_n^2 + \sigma^2). \tag{2.15}$$

2.5.2 Known mean

When

$$\theta \sim \text{ScaleInv}{-}\chi^2(\nu, \tau^2),$$

then

$$E\theta = \frac{\nu\tau^2}{\nu - 2}, \quad \nu > 2, \quad \text{Var}\theta = \frac{2\nu^2\tau^4}{(\nu-2)^2(\nu-4)} = \frac{2(E\theta)^2}{(\nu-4)}, \quad \nu > 4$$

and mode is

$$\frac{\nu\tau^2}{\nu + 2}.$$

Recall the density

$$\pi(\theta|\nu, \tau^2) = \frac{1}{\Gamma(\nu/2)}\Big(\frac{\nu\tau^2}{2}\Big)^{\frac{\nu}{2}}\theta^{-(\nu/2+1)}\exp[-\frac{\nu\tau^2}{2\theta}].$$

A standard model (mean know, prior on variance)

$$\theta \sim \text{ScaleInv}{-}\chi^2(\nu, \tau^2)$$
$$X_1, \ldots, X_n | \theta \overset{iid}{\sim} \mathcal{N}(\mu, \theta).$$

Marginal distribution. Joint density ($x = (x_1, \ldots, x_n)$)

$$\begin{aligned} f(x,\theta) &= (2\pi\theta)^{-n/2} \exp[-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\theta}] \frac{1}{\Gamma(\nu/2)} \Big(\frac{\nu\tau^2}{2}\Big)^{\frac{\nu}{2}} \theta^{-(\nu/2+1)} \exp[-\frac{\nu\tau^2}{2\theta}] \\ &= c \cdot \exp\big[-\frac{\sum_{i=1}^n (x_i-\mu)^2 + \nu\tau^2}{2\theta}\big] \theta^{-\big((n+\nu)/2+1\big)}, \end{aligned}$$

where

$$c = (2\pi)^{\frac{-n}{2}} \frac{1}{\Gamma(\nu/2)} \Big(\frac{\nu\tau^2}{2}\Big)^{\frac{\nu}{2}}.$$

Since

$$\int_0^\infty \theta^{-(\alpha+1)} \exp[-\beta/\theta] d\theta = \frac{\Gamma(\alpha)}{\beta^\alpha},$$

it follows that

$$\int_0^\infty \exp\big[-\frac{\sum_{i=1}^n (x_i-\mu)^2 + \nu\tau^2}{2\theta}\big] \theta^{-(\frac{(n+\nu)}{2}+1)} d\theta = \Gamma\big(\frac{\nu+n}{2}\big) \big(\frac{\sum_{i=1}^n (x_i-\mu)^2 + \nu\tau^2}{2}\big)^{-\frac{\nu+n}{2}}.$$

Therefore, the joint density of $X_1, \ldots, X_n$ is

$$f(x) = \frac{\Gamma\big(\frac{\nu+n}{2}\big)}{\Gamma\big(\frac{\nu}{2}\big)} \big(\pi\nu\tau^2\big)^{-n/2} \Big(\frac{\sum_{i=1}^n (x_i-\mu)^2}{\nu\tau^2} + 1\Big)^{-\frac{\nu+n}{2}}. \tag{2.16}$$

Hence, $X_i \sim \text{lst}(\mu, \tau^2, \nu)$, $i = 1, \ldots, n$. Observe that

$$\text{Cov}(X_i, X_j) = \text{Var}(E[X_1|\theta]) = \text{Var}(\mu) = 0,$$

i.e. the random variables are uncorrelated but not independent.

Posterior distribution, mean and variance. From the joint density $f(x, \theta)$, it follows:

$$\pi(\theta|x) \propto \theta^{-\big((n+\nu)/2+1\big)} \exp\big[-\frac{\sum_{i=1}^n (x_i-\mu)^2 + \nu\tau^2}{2\theta}\big] = \theta^{-(\nu_n/2+1)} \exp[-\frac{\nu_n\tau_n^2}{2\theta}],$$

where

$$\nu_n := \nu + n, \quad \tau_n^2 := \frac{\sum_{i=1}^n (x_i-\mu)^2 + \nu\tau^2}{\nu_n}.$$

Hence

$$\theta|x \sim \text{ScaleInv}{-}\chi^2(\nu_n, \tau_n^2). \tag{2.17}$$

Posterior mean:

$$E[\theta|x] = \frac{\nu_n \tau_n^2}{\nu_n - 2} = \frac{(\nu+n)\sum_{i=1}^n (x_i-\mu)^2 + \nu\tau^2}{\nu+n-2} = \frac{\sum_{i=1}^n (x_i-\mu)^2}{n}\frac{n}{\nu+n-2} + \frac{\nu-2}{n+\nu-2}\frac{\nu\tau^2}{\nu-2}.$$

Since

$$E\theta = \frac{\nu\tau^2}{\nu-2},$$

we see again that posterior mean is the average of prior mean $E\theta$ and sample variance (MLE estimate) $\frac{\sum_{i=1}^n (x_i-\mu)^2}{n}$. The bigger n, the smaller is prior influence.

The meaning of hyperparameters ν and τ^2: ν is the size of "prior sampleänd τ^2 is the "sample variance of prior sample". Thus $\nu_n = \nu + n$ is the "total"sample size and τ_n is the "total sample variance".

Posterior variance:

$$\mathrm{Var}[\theta|x] = \frac{2(E[\theta|x])^2}{(\nu_n - 4)}.$$

Posterior predicative distribution. Posterior predicative distribution is location scale t-distribution, with parameters μ, τ_n^2, ν_n (Exercise 12) i.e.

$$X_{n+1}|x \sim \mathrm{lst}(\mu, \tau_n^2, \nu_n). \tag{2.18}$$

2.5.3 Unknown mean and variance

Now $\theta = (\mu, \sigma^2)$. A standard model:

$$\begin{aligned} \sigma^2 &\sim \mathrm{ScaleInv}{-}\chi^2(\nu, \tau^2) \\ \mu|\sigma^2 &\sim \mathcal{N}(\mu_o, \frac{\sigma^2}{\kappa}) \\ X_1, \ldots, X_n|(\mu, \sigma^2) &\overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2). \end{aligned}$$

Hence, μ and σ^2 are not independent random variables, their joint density is

$$\begin{aligned} \pi(\mu, \sigma^2) &= c(\nu, \tau^2, \kappa)(\sigma^2)^{-(\nu/2+1)} \exp[-\frac{\nu\tau^2}{2\sigma^2}]\frac{1}{\sigma}\exp[-\frac{\kappa(\mu-\mu_o)^2}{2\sigma^2}] \\ &= c(\nu, \tau^2, \kappa)(\sigma^2)^{-(\frac{\nu+1}{2}+1)} \exp[-\frac{\kappa(\mu-\mu_o)^2 + \nu\tau^2}{2\sigma^2}], \\ c(\nu, \tau^2, \kappa) &:= \frac{1}{\Gamma(\frac{\nu}{2})}\Big(\frac{\nu\tau^2}{2}\Big)^{\frac{\nu}{2}}\sqrt{\frac{\kappa}{2\pi}}. \end{aligned}$$

Sometimes called *normal Inverse* χ^2 or *NIX* density with (hyper)parameters ν, μ_o, τ^2, κ.

The prior mean: $E(\mu) = E\big(E[\mu|\sigma^2]\big) = \mu_o$, $E(\sigma^2) = \frac{\nu\tau^2}{\nu-2}$. The parameter κ shows how strongly we believe in μ_o, parameter ν shows how strongly we believe that the variance is τ^2.

In this model (provided $\nu > 2$, exercise 12)

$$EX_i = \mu_o,\ \mathrm{Var}(X_i) = \frac{\nu(1+\kappa)\tau^2}{\kappa(\nu-2)},\ \mathrm{Cov}(X_i, X_j) = \mathrm{Var}(\mu) = \frac{1}{\kappa}\frac{\nu\tau^2}{\nu-2}. \tag{2.19}$$

Posterior distribution. Joint density ($x = (x_1, \ldots, x_n)$)

$$\begin{aligned} f(x,\mu,\sigma^2) &= (2\pi)^{-\frac{n}{2}} c(\nu,\tau^2,\kappa)(\sigma^2)^{-n/2} \exp[-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}](\sigma^2)^{-(\frac{\nu+1}{2}+1)} \exp[-\frac{\kappa(\mu-\mu_o)^2+\nu\tau^2}{2\sigma^2}] \\ &= (2\pi)^{-\frac{n}{2}} c(\nu,\tau^2,\kappa)(\sigma^2)^{-(\frac{\nu+n+1}{2}+1)} \exp[-\frac{\sum_{i=1}^n (x_i-\mu)^2+\kappa(\mu-\mu_o)^2+\nu\tau^2}{2\sigma^2}] \\ &= (2\pi)^{-\frac{n}{2}} c(\nu,\tau^2,\kappa)(\sigma^2)^{-(\frac{\nu+n+1}{2}+1)} \exp[-\frac{\sum_{i=1}^n (x_i-\bar{x})^2+n(\bar{x}-\mu)^2+\kappa(\mu-\mu_o)^2+\nu\tau^2}{2\sigma^2}] \end{aligned}$$

Define

$$\mu_n = \frac{\kappa}{\kappa+n}\mu_o + \frac{n}{\kappa+n}\bar{x}$$

and observe that (exercise 12):

$$\kappa(\mu-\mu_o)^2 + n(\mu-\bar{x})^2 = (\kappa+n)\big(\mu-\mu_n\big)^2 + \frac{\kappa n}{\kappa+n}(\mu_o-\bar{x})^2. \tag{2.20}$$

It follows that the posterior density is also NIX-density with parameters:

$$\begin{aligned} \nu_n &= \nu + n \\ \mu_n &= \frac{\kappa}{\kappa+n}\mu_o + \frac{n}{\kappa+n}\bar{x} \\ \kappa_n &= \kappa + n \\ \nu_n\tau_n^2 &= \nu\tau^2 + \sum_{i=1}^n (x_i-\bar{x})^2 + \frac{\kappa n}{\kappa+n}(\bar{x}-\mu_o)^2 \end{aligned}$$

so that

$$\sum_{i=1}^n (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2 + \kappa(\mu-\mu_o)^2 + \nu\tau^2 = \kappa_n(\mu-\mu_n)^2 + \nu_n\tau_n^2. \tag{2.21}$$

With $\mu = \mu_n$, the equality (2.21) gives another formula for $\nu_n\tau_n^2$:

$$\nu_n\tau_n^2 = \nu\tau^2 + \sum_{i=1}^n (x_i-\bar{x})^2 + n(\bar{x}-\mu_n)^2 + \kappa(\mu_n-\mu_o)^2 = \sum_{i=1}^n (x_i-\mu_n)^2 + \kappa(\mu_n-\mu_o)^2 + \nu\tau^2. \tag{2.22}$$

Therefore from (2.21),

$$\pi(\mu, \sigma^2|x) = c(\nu_n, \tau_n^2, \kappa_n)(\sigma^2)^{-(\frac{\nu_n+1}{2}+1)} \exp[-\frac{\kappa_n(\mu-\mu_n)^2 + \nu_n\tau_n^2}{2\sigma^2}]. \quad (2.23)$$

Thus, given σ^2, the posterior distribution of μ is normal with mean μ_n and variance σ^2/κ_n and the posterior distribution of σ^2 is scaled inverse χ^2 with parameters ν_n and τ_n^2:

$$\sigma^2|x \sim \text{ScaleInv}-\chi^2(\nu_n, \tau_n^2)$$
$$\mu|\sigma^2, x \sim \mathcal{N}(\mu_n, \frac{\sigma^2}{\kappa_n}).$$

It is easy to see that when $X \sim \text{ScaleInv}-\chi^2(\nu, \tau)$, then for any $a > 0$, $aX \sim \text{ScaleInv}-\chi^2(\nu, a\tau)$. Therefore,

$$\frac{\sigma^2}{\kappa_n}|x \sim \text{ScaleInv}-\chi^2(\nu_n, \frac{\tau_n^2}{\kappa_n})$$

and from (2.18) it follows that when integrating σ^2 out, we obtain (Exercise 12)

$$\mu|x \sim \text{lst}(\mu_n, \frac{\tau_n^2}{\kappa_n}, \nu_n). \quad (2.24)$$

When $X \sim \text{lst}(\mu, \tau^2, \nu)$, then

$$EX = \mu, \quad \text{Var}(X) = \tau^2\frac{\nu}{\nu-2}, \quad \text{provided } \nu > 2. \quad (2.25)$$

Therefore the posterior means are (recall (2.22)):

$$E[\mu|x] = \mu_n = \frac{\kappa}{\kappa+n}\mu_o + \frac{n}{\kappa+n}\bar{x},$$
$$E[\sigma^2|x] = \frac{\tau_n^2\nu_n}{\nu_n-2} = \frac{\nu\tau^2 + \sum_{i=1}^n(x_i-\mu_n)^2 + \kappa(\mu_n-\mu_o)^2}{\nu+n-2}.$$

We can now interpret the parameters: there is a "prior sample (for expectation)" with size κ and all observations being equal to μ_o. Adding to that prior sample our observations, we obtain "posterior sample (for expectation)" with size $\kappa + n = \kappa_n$. The sample mean of the "posterior sample" is μ_n. The bigger is the size of prior sample, the more μ_o influences μ_n. The sum of squares of "posterior sample" is $\sum_{i=1}^n(x_i-\mu_n)^2 + \kappa(\mu_n-\mu_o)^2$. For variance, there is another "prior sample (for variance)" with sample mean μ_n, size ν and sum of squares $\nu\tau^2$. So the sum of squares of both prior samples and the actual observations is $\nu\tau^2 + \sum_{i=1}^n(x_i-\mu_n)^2 + \kappa(\mu_n-\mu_o)^2 = \nu_n\tau_n^2$. For posterior mean and mode of σ^2, the sample size of the first "prior sample for expectation", namely κ, has not been taking into account (sum of squares is divided by $\nu+n-2$). We see that the bigger ν (the sample size for "prior sample for variance"), the closer $E[\sigma^2|x]$ is to τ^2.

Marginal and posteriror predicative distribution. Since $f(\theta|x) = f(x,\theta)/f(x)$, it holds

$$f(x) = \frac{f(x,\theta)}{f(\theta|x)}.$$

Hence the marginal density

$$f(x) = \frac{(2\pi)^{-\frac{n}{2}} c(\nu,\tau^2,\kappa)}{c(\nu_n,\tau_n^2,\kappa_n)} = \frac{\Gamma(\frac{\nu_n}{2})}{\Gamma(\frac{\nu}{2})} \frac{\Big(\frac{\nu\tau^2}{2}\Big)^{\frac{\nu}{2}}\sqrt{\kappa}}{(2\pi)^{\frac{n}{2}}\Big(\frac{\nu_n\tau_n^2}{2}\Big)^{\frac{\nu_n}{2}}\sqrt{\kappa_n}} = \frac{\Gamma(\frac{\nu_n}{2})}{\Gamma(\frac{\nu}{2})} \frac{\Big(\nu\tau^2\Big)^{\frac{\nu}{2}}}{\pi^{\frac{n}{2}}\Big(\nu_n\tau_n^2\Big)^{\frac{\nu_n}{2}}} \sqrt{\frac{\kappa}{\kappa_n}}. \tag{2.26}$$

Since (take $\nu = \nu_n, \mu_o = \mu_n, \kappa = \kappa_n, \tau^2 = \tau_n^2$ and $n = 1$)

$$\begin{aligned}
\nu_{n+1} &= \nu_n + 1 \\
\mu_{n+1} &= \frac{\kappa_n}{\kappa_n+1}\mu_n + \frac{1}{\kappa_n+1}x_{n+1} \\
\kappa_{n+1} &= \kappa_n + 1 \\
\nu_{n+1}\tau_{n+1}^2 &= \nu_n\tau_n^2 + \frac{\kappa_n}{\kappa_n+1}(x_{n+1}-\mu_n)^2.
\end{aligned}$$

We see that

$$\begin{aligned}
\frac{\nu_{n+1}\tau_{n+1}^2}{\nu_n\tau_n^2} &= 1 + \frac{\kappa_n}{(\kappa_n+1)\nu_n\tau_n^2}(x_{n+1}-\mu_n)^2 \\
\Big(\frac{\nu_{n+1}\tau_{n+1}^2}{\nu_n\tau_n^2}\Big)^{-\frac{\nu_{n+1}}{2}} &= \Big(1 + \frac{\kappa_n}{(\kappa_n+1)\nu_n\tau_n^2}(x_{n+1}-\mu_n)^2\Big)^{-\frac{\nu_{n+1}}{2}}.
\end{aligned}$$

The posterior predicative density is thus

$$\begin{aligned}
f(x_{n+1}|x) &= \frac{\Gamma(\frac{\nu_{n+1}}{2})}{\Gamma(\frac{\nu_n}{2})} \frac{\Big(\nu_n\tau_n^2\Big)^{\frac{\nu_n}{2}}}{\pi^{\frac{1}{2}}\Big(\nu_{n+1}\tau_{n+1}^2\Big)^{\frac{\nu_{n+1}}{2}}} \sqrt{\frac{\kappa_n}{\kappa_{n+1}}} \\
&= \frac{\Gamma(\frac{\nu_n+1}{2})}{\Gamma(\frac{\nu_n}{2})} \sqrt{\frac{\kappa_n}{\pi(\kappa_n+1)\nu_n\tau_n^2}} \Big(1 + \frac{\kappa_n}{(\kappa_n+1)\nu_n\tau_n^2}(x_{n+1}-\mu_n)^2\Big)^{-\frac{\nu_n+1}{2}}.
\end{aligned}$$

We recognize the local scaled t-distribution density with parameters $\mu_n, \tau_n^2\frac{1+\kappa_n}{\kappa_n}, \nu_n$, i.e.

$$X_{n+1}|x \sim \mathrm{lst}\big(\mu_n, \frac{\tau_n^2(1+\kappa_n)}{\kappa_n}, \nu_n\big).$$

Taking $n = 0$, i.e. replacing $\nu_n, \tau_n^2, \kappa_n, \mu_n$ by their prior values $\nu, \kappa, \tau^2, \mu_0$, we see that the marginal distribution of X_i is also location-scale t-distribution:

$$X_i \sim \mathrm{lst}\big(\mu_0, \frac{\tau^2(1+\kappa)}{\kappa}, \nu\big), \quad i = 1, 2, \ldots.$$

The mean and variance of the location-scale t-distribution above are μ_o and $\frac{\nu(1+\kappa)\tau^2}{\kappa(\nu-2)}$, as we already obtained by direct calculation in (2.19).

2.6 Exercises

1. Let $X_1, \ldots, X_n$ be Beta-Bernoulli random variables. Show that $X_i \sim B(1, \frac{\alpha}{\alpha+\beta})$ and $\mathrm{Cov}(X_i, X_j) = \mathrm{Var}[E(X_1|\theta)] = \mathrm{Var}(\theta)$. Find the correlation coefficient $\rho(X_i, X_j)$ for Beta-Bernoulli model. What is the covariance and correlation for uniform prior?

2. Let $X \sim \mathrm{BetaBin}(n, \alpha, \beta)$. Find EX and $\mathrm{Var}(X)$. Show that $\mathrm{BetaBin}(n, 1, 1)$ is uniform over $\{0, \ldots, n\}$

3. Polya urn.

   - Let $X_1, \ldots, X_n$ be the outcomes of Polya urn with α and β initial balls (not necessarily integers). Show that $(X_1, \ldots, X_n)$ has the same joint distribution as the Beta-Bernoulli random variables. You might use the formula $\Gamma(z+1) = z\Gamma(z)$, $\forall z > 0$.
   - Generalize the proof for $k > 2$: $(X_1, \ldots, X_n)$ has the same joint distribution as the Dirichlet-categorical random variables.

4. Let $\theta \sim \mathrm{Beta}(\alpha, \beta)$. Show that α and β can be found from $E\theta$ and $\mathrm{Var}(\theta)$ as follows:

$$\alpha + \beta = \frac{E\theta(1 - E\theta)}{\mathrm{Var}(\theta)} - 1, \quad \alpha = (\alpha + \beta)E\theta.$$

5. Let $(X_1, \ldots, X_k) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$. Show that

$$E\big(X_1^{r_1} \cdots X_k^{r_k}\big) = \frac{B(\alpha_1 + r_1, \ldots, \alpha_k + r_k)}{B(\alpha_1, \ldots, \alpha_k)}.$$

   Show

$$\mathrm{Cov}(X_i, X_j) = -\frac{\alpha_i \alpha_j}{|\alpha|^2(|\alpha| + 1)}$$

   and find correlation $\rho(X_i, X_j)$.

6. Let $X_1, \ldots, X_n$ be the Dirichlet-categorical random variables. Show that

$$\mathrm{Cov}(X_1, X_2) = \frac{\sum_{i=1}^k i^2 \alpha_i(|\alpha| - \alpha_i) - 2\sum_{i=1}^k \sum_{j=i+1}^k ij\alpha_i\alpha_j}{|\alpha^2|(|\alpha| + 1)}.$$

7. Show that the marginal distribution of Dirichlet-multinomial distribution are Beta-binomial: when $(N_1, \ldots, N_k) \sim \mathrm{DirMult}(n; \alpha_1, \ldots, \alpha_k)$, then N_i has Beta-binomial distribution. Show that

$$EN_iN_j = n(n-1)\frac{\alpha_i\alpha_i}{|\alpha|(|\alpha| + 1)}, \quad \mathrm{Cov}(N_i, N_j) = -\frac{\alpha_i\alpha_j}{|\alpha|^2}\frac{(|\alpha| + n)}{(|\alpha| + 1)}.$$

8. Prove the aggregation: when $(N_1, \ldots, N_k) \sim \text{DirMult}(n; \alpha_1, \ldots, \alpha_k)$, then

$$(N_1 + N_2, N_3, \ldots, N_k) \sim \text{DirMult}(n; \alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_k).$$

9. Find the posterior predicative density for Dirichlet-categorical model without Polya urn, i.e. show that

$$P(X_{n+1} = i | X = x) = \int f(i|\theta)\pi(\theta|x)d\theta = \frac{\alpha_i + n_i}{|\alpha| + n}.$$

10. Gamma-Poisson model

    - Show (2.6) and $\text{Cov}(X_i, X_j) = \frac{\alpha}{\beta^2}$.
    - Show that the posterior predicative distribution is negative binomial distribution.
    - Show (2.8).
    - Show (4.29).

11. Gamma-Exponential model

    - Show (2.10) and (2.11).
    - Show (2.13).
    - Consider Gamma-exponential model with $n = 1$, i.e. $X|\theta \sim \text{Exp}(\theta)$ and the prior distribution is $\text{Gamma}(\alpha, \beta)$. Suppose we observe that $X \geq c$, but we do not observe the exact value of X. Find the posterior distribution $\theta | X \geq c$ ($\text{Gamma}(\alpha, \beta + c)$).
    - Let $X_1, \ldots, X_n$ be iid random variables, $X_i \sim \text{Exp}(\theta)$. Let $S = \sum_{i=1}^n X_i$. Show that the moment generating function of S is that of $\text{Gamma}(n, \theta)$ and deduce that $S \sim \text{Gamma}(n, \theta)$.

12. Normal models

    - Show (2.19).
    - Show (2.15).
    - Show (2.18).
    - Show (2.20) and deduce (2.23).
    - Show (2.24).

3 Conjugate prior

In all the examples the prior and posterior distribution belonged to the same parametric model/class: in Beta-Bernoulli model prior and posterior both have Beta-distribution, in Dirichlet-multinomial model both have Dirichlet distribution, in gamma-Poisson and gamma-exponential model both (prior and posterior) have gamma distribution, and also in normal models the prior and posterior distribution belong to the same class. In fact, in all these examples the prior distribution was deliberately chosen so that such a property would hold.

Let $\{f(\cdot|\theta)\}$, $\theta \in \Theta$ be a parametric model and let $\mathcal{P}$ be a class of distributions on Θ. The class $\mathcal{P}$ is *conjugate for* $\{f(\cdot|\theta)\}$ if for any $x \in \mathcal{X}$ and for any prior $\pi \in \mathcal{P}$ also the posterior belongs to $\mathcal{P}$: $P(\theta \in \cdot|X = x) \in \mathcal{P}$. In terms of densities: when $\pi(\cdot) \in \mathcal{P}$, then $\pi(\cdot|x) \in \mathcal{P}$, $\forall x \in \mathcal{X}$.

Of course, when $\mathcal{P}$ consists of all probability distributions on Θ, then it is trivially conjugate for any model. Therefore the possibly small and parametric conjugate classes are of interest. We already know that the parameters of priors are called *hyperparameters* and with conjugate prior switching from prior to posterior distribution is reduced to an updating of the corresponding hyperparameters. This property alone can explain why conjugate priors are so popular, since the posterior distributions are always computable. Another advantage of conjugate priors is being interpretable as additional data (prior sample), as we have seen in all examples above.

Exponential family. The parametric model $\{f(\cdot|\theta)\}$ belongs to an *exponential family*, when all its members have form

$$f(x|\theta) = h(x)g(\theta)\exp[\phi(\theta)\cdot u(x)], \tag{3.1}$$

where $\phi : \Theta \to \mathbb{R}^k$ and $u : \mathcal{X} \to \mathbb{R}^k$. For $a, b \in \mathbb{R}^k$, $a \cdot b$ stands for inner product. The vector $\phi(\theta)$ is called the *natural parameter* for the family. When $\phi(\theta) = \theta$, then the exponential family is said to be in *canonical form*. By the change of parameters $\phi = \phi(\theta)$, it is always possible to convert an exponential family to canonical form. Note that the canonical form is non-unique. When $\phi(\theta) = \theta$ and $u(x) = x$ (hence $\phi(\theta)\cdot u(x) = \theta \cdot x$) the family is called *natural exponential family*. Defining $A(\theta) = -\ln g(\theta)$, and $B(x) = \ln h(x)$, we can rewrite (3.1) as

$$f(x|\theta) = h(x)\exp[\phi(\theta)\cdot u(x) - A(\theta)] = \exp[\phi(\theta)\cdot u(x) - A(\theta) + B(x)]. \tag{3.2}$$

Under the natural parametrization, it holds (Exercise 1) for every η such that $\theta + \eta \in \Theta$

$$M_u(\eta) = \int \exp[\eta \cdot u(x)] f(x|\theta)dx = e^{A(\eta+\theta)-A(\theta)} \tag{3.3}$$

The function M_u is the moment generating function of $u(X)$, where X has density $f(\cdot|\theta)$, its logarithm $\Lambda_u(\eta) = \ln M_u(\eta)$ is the cumulant generating function. In univariate case, i.e. when $\Theta \subset \mathbb{R}$, it follows that

$$A'(\theta) = \Lambda'_u(0) = E[u(X)|\theta], \quad A''(\theta) = \Lambda''_u(0) = \mathrm{Var}[u(X)|\theta]. \tag{3.4}$$

When $x_1, \ldots, x_n$ is an iid sample from exponential family (3.1), then with $x = (x_1, \ldots, x_n)$ and $f(x|\theta) = \prod_{i=1}^n f(x_i|\theta)$, we get that $f(x|\theta)$ belongs to an exponential family as well:

$$f(x|\theta) = \big(\prod_{i=1}^n h(x_i)\big) g^n(\theta) \exp[\phi(\theta) \cdot t(x)], \quad t(x) = \sum_{i=1}^n u(x_i). \tag{3.5}$$

The statistic $t(x) = \sum_{i=1}^n u(x_i)$ is called *sufficient statistic*.

Examples of exponential families.

1. Poisson distribution:

$$f(x|\theta) = e^{-\theta}\frac{\theta^x}{x!} = \frac{1}{x!}\exp[x\ln\theta - \theta], \quad x \in \mathbb{N}.$$

   Here $A(\theta) = \theta$ (or $g(\theta) = e^{-\theta}$), $h(x) = \frac{1}{x!}$, $u(x) = x$, $\phi(\theta) = \ln\theta$.

2. Negative binomial distribution: r is fixed, $\theta = p$ is the parameter

$$f(x|\theta) = f(k|r,\theta) = \frac{\Gamma(x+r)}{x!\Gamma(r)}(1-\theta)^x\theta^r.$$

   Here $u(x) = x$, $\phi(\theta) = \ln(1-\theta)$, $g(\theta) = \theta^r$ $(A(\theta) = -r\ln\theta)$.

3. Gamma(α, θ). Here α is fixed, θ is a parameter.

$$f(x|\theta) = \frac{\theta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\theta x}.$$

   Here $g(\theta) = \theta^\alpha$, $(A(\theta) = -\alpha\ln\theta)$, $u(x) = -x$, $\phi(\theta) = \theta$ – canonical form.

4. Gamma(α, β). The parameters $\theta = (\alpha, \beta)$.

$$f(x|\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} = \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{1}{x}\exp[\alpha\ln x - \beta x].$$

   Thus

$$\phi(\theta) = \theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad u(x) = \begin{pmatrix} \ln x \\ -x \end{pmatrix}.$$

5. Dirichlet distribution: $\theta = (\theta_1, \ldots, \theta_k)$

$$f(x|\theta) = \frac{1}{B(\theta_1, \ldots, \theta_k)} \prod_{i=1}^{k} x_i^{\theta_i - 1}, \quad x = (x_1, \ldots, x_k) \in \mathrm{S}_k$$
$$= \frac{1}{B(\theta_1, \ldots, \theta_k)} \Big(\prod_{i=1}^{k} \frac{1}{x_i} \Big) \exp[\sum_{i=1}^{k} \theta_i \ln x_i].$$

Hence Dirichlet distribution constitute an exponential family canonical form: $\phi : \Theta \to \mathbb{R}^k$ is identity function $\phi(\theta) = \theta$,

$$u = (u_1, \ldots, u_k) : \mathcal{X} \to \mathbb{R}^k, \quad u_i(x) = \ln x_i.$$

6. Beta distribution: $\theta = (\theta_1, \theta_2)$

$$f(x|\theta) = \frac{1}{B(\theta_1, \theta_2)} \Big(\frac{1}{x} \cdot \frac{1}{1-x} \Big) \exp[\theta_1 \ln(x) + \theta_2 \ln(1-x)], \quad x \in (0,1).$$

Hence $\phi(\theta) = \theta$ and

$$u(x) = \begin{pmatrix} \ln(x) \\ \ln(1-x) \end{pmatrix}.$$

7. Binomial $B(n, \theta)$, $\theta \in (0,1)$.

$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = \binom{n}{x} (1-\theta)^n \exp[x \ln \frac{\theta}{1-\theta}].$$

Thus $\phi(\theta) = \ln \frac{\theta}{1-\theta}$, $u(x) = x$.

8. Multinom$(n; \theta_1, \ldots, \theta_k)$, $x = (x_1, \ldots, x_k)$

$$f(x|\theta_1, \ldots, \theta_k) = \binom{n}{x_1 \cdots x_k} \prod_{i=1}^{k} \theta_i^{x_i} = \binom{n}{x_1 \cdots x_k} \exp[\sum_{i=1}^{n} x_i \ln \theta_i].$$

Hence $u(x) = x$, $\phi_i(\theta) = \ln \theta_i$.

9. Normal distribution $\mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$.

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{x^2 - 2x\mu + \mu^2}{2\sigma^2}].$$

Hence

$$u(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}, \quad \phi(\theta) = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix}.$$

10. Normal distribution $\mathcal{N}(\theta, \sigma^2)$

$$g(\theta) = \exp[-\frac{\theta^2}{2\sigma^2}], \quad \phi(\theta) = \theta, \quad u(x) = \frac{x}{\sigma^2}. \tag{3.6}$$

11. Normal distribution $\mathcal{N}(\mu, \theta)$

$$g(\theta) = \frac{1}{\sqrt{\theta}} \exp[-\frac{\mu^2}{2\theta}], \quad u(x) = \begin{pmatrix} x^2 \\ x \end{pmatrix}, \quad \phi(\theta) = \begin{pmatrix} -\frac{1}{2\theta} \\ \frac{\mu}{\theta} \end{pmatrix}.$$

If the support of the distribution depends on parameter, it cannot belong to an exponential family. Hence uniform $U(0, \theta)$ distribution does not belong to an exponential family. Also Student t-distribution cannot be represented as (3.1).

Conjugate prior for exponential families. Let us consider the model (3.5) that obviously includes also (3.1) when $n = 1$. A conjugate family of priors is given by

$$\pi(\theta|\mu, \lambda) \propto g(\theta)^\lambda \exp[\phi(\theta) \cdot \mu], \tag{3.7}$$

where $\lambda \in \mathbb{R}$ and $\mu \in \mathbb{R}^k$ are hyperparameters. Then the posterior distribution of the model (1.2) belongs to the same family, because with $x = (x_1, \ldots, x_n)$

$$\begin{aligned} \pi(\theta|x) &\propto \big(\prod_{i=1}^n h(x_i)\big) g^n(\theta) \exp[\phi(\theta) \cdot t(x)] g(\theta)^\lambda \exp[\phi(\theta) \cdot \mu] \\ &\propto g(\theta)^{\lambda+n} \exp[\phi(\theta) \cdot (\mu + t(x))]. \end{aligned}$$

Thus $\pi(\theta|x) = \pi(\theta|\mu + t(x), \lambda + n)$.

Examples of conjugate priors (3.7) for exponential families.

1. Bernoulli distribution $\mathrm{B}(1, \theta)$:

$$\phi(\theta) = \ln \frac{\theta}{1-\theta}, \quad g(\theta) = 1 - \theta, \quad t(x) = \sum_{i=1}^n x_i, \quad x = (x_1, \ldots, x_n)$$

$$\pi(\theta|\mu, \lambda) \propto g(\theta)^\lambda \exp[\mu\phi(\theta)] = (1-\theta)^\lambda \frac{\theta^\mu}{(1-\theta)^\mu} = (1-\theta)^{\lambda-\mu}\theta^\mu.$$

   We recognize the $\mathrm{Beta}(\mu + 1, \lambda - \mu + 1)$ density (provided $\mu + 1 > 0$ and $\lambda - \mu + 1 > 0$). Hence posterior

$$\theta|x \sim \mathrm{Beta}(\mu + t(x) + 1, \lambda - \mu - t(x) + n + 1). \tag{3.8}$$

   Denoting $\mu + 1 = \alpha$ and $\lambda - \mu + 1 = \beta$, we that (3.8) is the same as (2.3).

2. Multinomial distribution $\mathrm{Multinom}(n; \theta_1, \ldots, \theta_k)$:
   Since the sample size n is included to multinomial distribution, we take

in (3.5) sample size equal to 1.

$$\phi(\theta) = \begin{pmatrix} \ln\theta_1 \\ \ln\theta_2 \\ \cdots \\ \ln\theta_k \end{pmatrix}, \quad u(x) = x \in \mathbb{R}^k, \quad g(\theta) = 1, \quad \mu \in \mathbb{R}^k$$

$$\pi(\theta|\mu,\lambda) \propto \exp[\mu\phi(\theta)] = \exp[\sum_{i=1}^{k} \mu_i \ln\theta_i] = \prod_{i=1}^{k} \theta_i^{\mu_i}.$$

We recognize Dirichlet distribution $\mathrm{Dir}(\mu_1+1,\ldots,\mu_k+1)$. Hence, with $\alpha_i = \mu_i + 1$, we get the posterior

$$\theta|x \sim \mathrm{Dir}(\alpha_1 + x_1, \cdots, \alpha_k + x_k). \tag{3.9}$$

3. Poisson distribution $\mathrm{Po}(\theta)$:

$$\phi(\theta) = \ln\theta, \quad g(\theta) = e^{-\theta}, \quad t(x) = \sum_{i=1}^{n} x_i, \quad x = (x_1,\ldots,x_n)$$
$$\pi(\theta|\mu,\lambda) \propto \exp[\mu\phi(\theta) - \lambda\theta] = \theta^{\mu} e^{-\lambda\theta}.$$

We recognize $\mathrm{Gamma}(\mu+1,\lambda)$-density. Denoting $\mu+1 = \alpha$ and $\lambda = \beta$, we get the posterior

$$\theta|x \sim \mathrm{Gamma}(\alpha + t(x), \beta + n). \tag{3.10}$$

we see that (3.10) is the same as (2.7).

4. $\mathrm{Gamma}(\nu,\theta)$:

$$\phi(\theta) = \theta, \quad g(\theta) = \theta^{\nu}, \quad t(x) = -\sum_{i=1}^{n} x_i,$$
$$\pi(\theta|\mu,\lambda) \propto \theta^{\nu\lambda} \exp[\mu\theta].$$

We recognize $\mathrm{Gamma}(\nu\lambda+1,-\mu)$ density (provided $\mu < 0$). Denoting $\alpha = \nu\lambda + 1$ and $\beta = -\mu$, we get the posterior

$$\theta|x \sim \mathrm{Gamma}\big(\alpha + \nu n, \beta + \sum_{i=1}^{n} x_i\big). \tag{3.11}$$

Since $\mathrm{Gamma}(1,\theta) = \mathrm{Exp}(\theta)$, with $\nu = 1$, we see that (3.11) is the same as (2.12).

5. Normal $\mathcal{N}(\theta, \sigma^2)$

$$g(\theta) = \exp[-\frac{\theta^2}{2\sigma^2}], \quad \phi(\theta) = \theta, \quad t(x) = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i$$

$$\pi(\theta|a, \lambda) \propto \exp[-\frac{\lambda\theta^2}{2\sigma^2} + a\theta].$$

We recognize the normal distribution with mean μ_o and variance τ^2, where

$$\mu_o = \frac{a\sigma^2}{\lambda}, \quad \tau^2 = \frac{\sigma^2}{\lambda}.$$

because

$$\frac{(\theta - \mu_o)^2}{2\tau^2} = \frac{\theta^2}{2\tau^2} - \frac{\mu_o\theta}{\tau^2} + \frac{\mu_o^2}{2\tau^2} = \frac{\lambda\theta^2}{2\sigma^2} - a\theta + \frac{\mu_o}{2\tau^2}.$$

The posterior density

$$\pi(x|a + \theta(x), \lambda + n) \propto \exp[-\frac{(\lambda + n)\theta^2}{2\sigma^2} + (a + t(x))\theta].$$

The posterior mean and variance are thus

$$\mu_n = \frac{(a + \frac{n\bar{x}}{\sigma^2})\sigma^2}{\lambda + n} = \frac{\sigma^2\frac{\mu_o}{\tau^2} + n\bar{x}}{\frac{\sigma^2}{\tau^2} + n} = \frac{\frac{\mu_o}{\tau^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}$$

$$\tau_n^2 = \frac{\sigma^2}{\lambda + n}, \quad \frac{1}{\tau_n^2} = \frac{1}{\tau^2} + \frac{n}{\sigma^2}.$$

We get the same posterior distribution as in the subsection 2.5.1.

6. Normal $\mathcal{N}(\mu, \theta)$

$$g(\theta) = \frac{1}{\sqrt{\theta}}\exp[-\frac{\mu^2}{2\theta}], \quad t(x) = \begin{pmatrix} \sum_i x_i^2 \\ \sum_i x_i \end{pmatrix}, \quad \phi(\theta) = \begin{pmatrix} -\frac{1}{2\theta} \\ \frac{\mu}{\theta} \end{pmatrix}.$$

$$\pi(\theta|(a_1, a_2), \lambda) \propto \frac{1}{\theta^{\frac{\lambda}{2}}}\exp[-\frac{\lambda\mu^2 + a_1 - 2a_2\mu}{2\theta}].$$

We recognize Scale $-$ Inv$\chi^2(\nu, \tau^2)$-distribution, with

$$\nu = \lambda, \quad \nu\tau^2 = \lambda\mu^2 + a_1 - 2a_2\mu.$$

Posterior density is

$$\pi\big(\theta|(a_1 + \sum_i x_i^2, a_2 + \sum_i x_i), \lambda + n\big) \propto \frac{1}{\theta^{\frac{\lambda+n}{2}}}\exp[-\frac{(\lambda + n)\mu^2 + a_1 + \sum_i x_i^2 - 2(a_2 + \sum_i x_i)\mu}{2\theta}]$$

Since

$$(\lambda+n)\mu^2+a_1+\sum_i x_i^2-2(a_2+\sum_i x_i)\mu=\sum_{i=1}^n(x_i-\mu)^2+\nu\tau^2,$$

we obtain the posterior distribution

$$\theta|x\sim \text{ScaleInv}\chi^2(\nu_n,\tau_n^2),\quad \nu_n=\nu+n,\quad \nu_n\tau_n^2=\sum_{i=1}^n(x_i-\mu)^2+\nu\tau^2,$$

just like in (2.17).

Conjugate priors and sufficient statistics. Let $\{f(\cdot|\theta):\theta\in\Theta\}$ be a parametric family, let $x=(x_1,\dots,x_n)$ be observations. Recall that a statistic $t(x)$ is sufficient, if the density $f(x|\theta)$ factorizes $f(x|\theta)=h(x)g(\theta,t(x))$. Then the conditional density of data given the value of sufficient statistic $t(x)$ is independent of parameter, because for any x such that $t(x)=t$

$$f(x|t(x)=t,\theta)=\frac{h(x)g(\theta,t)}{\int_{\{x':t(x')=t\}}h(x')g(\theta,t)dx'}=\frac{h(x)}{\int_{\{x':t(x')=t\}}h(x')dx'}.$$

In Bayesian setting it means that the posterior distribution depends on x through $t(x)$:

$$\pi(\theta|x)=\frac{h(x)g(\theta,t(x))\pi(\theta)}{f(x)}=\frac{h(x)g(\theta,t(x))\pi(\theta)}{\int h(x)g(\theta,t(x))\pi(\theta)d\theta}=\frac{g(\theta,t(x))\pi(\theta)}{\int g(\theta,t(x))\pi(\theta)d\theta}.$$

Thus, when $t(x)$ is a sufficient statistic, then $\pi(\theta|x)\propto g(\theta,t(x))\pi(\theta)$ and

$$\pi(\theta|x)=\pi(\theta|t(x)). \tag{3.12}$$

On the other hand, when (3.12) holds, then $f(\theta|x)f(x)=\pi(\theta|t(x))f(x)$ and $f(x|\theta)=f(x)\pi(\theta|t(x))/\pi(\theta)$. We see that with $h(x)=f(x)$ and $g(\theta,t(x))=\pi(\theta|t(x))/\pi(\theta)$ the factorization $f(x|\theta)=h(x)g(\theta,t(x))$ holds and by definition $t(x)$ is a sufficient statistic.

Suppose that $\{\pi(\theta|\alpha):\alpha\in\Lambda\}$, where $\Lambda\subset\mathbb{R}^d$ is a conjugate family for $\{f(\cdot|\theta):\theta\in\Theta\}$. Then for every x, $\pi(\theta|x)=\pi(\theta|\alpha(x))$, where $\alpha(x)\in\Lambda$. Hence (3.12) holds and $\alpha(x)$ must be a sufficient statistic. Hence a parametric conjugate priors exist if and only if there exists a sufficient statistic. The following lemma states than when $\{f(\cdot|\theta):\theta\in\Theta\}$ is such that the support of $f(\cdot|\theta)$ does not depend on θ, then a sufficient statistic exists only if $\{f(\cdot|\theta):\theta\in\Theta\}$ is an exponential family. In the lemma by a sample the iid sample is meant.

Lemma 3.1 *(***Pitman-Koopman lemma***) Let $\{f(\cdot|\theta) : \theta \in \Theta\}$ be a family with the property that the support of $f(\cdot|\theta)$ does not depend on θ. If the following holds: for a sample size large enough, there exists a sufficient statistic of constant dimension, then $\{f(\cdot|\theta) : \theta \in \Theta\}$ is an exponential family.*

Thus, for the families where support does not depend on parameters, the existence of parametric (i.e. finite dimensional) conjugate priors is equivalent to being exponential. When the support depends on parameter, then the family cannot be exponential. However, the conjugate priors might still exist: the class of Pareto distributions constitute conjugate priors for $\mathrm{U}(0,\theta)$ (Exercise 2); the class of Pareto distributions $\mathrm{Pa}(\alpha,\theta)$ has conjugate priors (Exercise 4); shifted exponential distributions $\mathrm{Exp}(\theta,\lambda)$ have conjugate priors (Exercise 5).

Mixtures of conjugate priors are conjugate. Let $\mathcal{P}$ be a class of conjugate priors for a parametric family $\{f(\cdot|\theta)\}$. Consider the class of K-mixtures from $\mathcal{P}$:

$$\mathcal{P}_K = \Big\{ \sum_{k=1}^{K} q_i \pi_k, \quad q_k > 0, \quad \sum_{k=1}^{K} q_k = 1, \quad \pi_k \in \mathcal{P} \Big\}. \tag{3.13}$$

Here K is the number of mixture components. It is easy to see that $\mathcal{P}_K$ is conjugate as well, because (Exercise 8), when $\pi(\theta) = \sum_{k=1}^{K} q_k \pi_k(\theta)$, then the posterior belongs to $\mathcal{P}_K$ as well:

$$\pi(\theta|x) = \sum_{k=1}^{K} q_k(x)\pi_k(\theta|x), \quad q_k(x) = \frac{q_k f_k(x)}{\sum_{j=1}^{K} q_j f_j(x)}, \quad f_k(x) := \int \pi_k(\theta) f(x|\theta) d\theta. \tag{3.14}$$

When $\{f(\cdot|\theta)\}$ is an exponential family (3.1) and $\mathcal{P}$ is the corresponding class of conjugate priors as in (3.7), then (Exercise 8)

$$\int f(x|\theta)\pi(\theta|\mu,\lambda)d\theta = h(x)\frac{K(\mu,\lambda)}{K(\mu+u(x),\lambda+1)}, \quad K^{-1}(\mu,\lambda) = \int g(\theta)^{\lambda} e^{\phi(\theta)\cdot\mu} d\theta. \tag{3.15}$$

Thus, when $\{f(\cdot|\theta)\}$ is an exponential family (3.1), $\mathcal{P}$ is the corresponding class of conjugate priors (3.7) , $\pi(\theta) = \sum_{k=1}^{K} q_k \pi(\theta|\mu_k,\lambda_k)$ and $x = (x_1,\ldots,x_n)$ are observations from (1.2), then the posterior density is (exercise 8):

$$\begin{aligned} \pi(\theta|x) &= \sum_{k=1}^{K} q_k(x)\pi_k(\theta|\mu_k + t(x), \lambda_k + n), \\ q_k(x) &= \frac{q_k K(\mu_k,\lambda_k)/K(\mu_k + t(x), \lambda_k + n)}{\sum_{j=1}^{K} q_j K(\mu_j,\lambda_j)/K(\mu_j + t(x), \lambda_j + n)}. \end{aligned} \tag{3.16}$$

Observe that $\theta \sim \sum_{k=1}^{K} q_i \pi_k$ can be written in Bayesian language as follows:

$$\begin{aligned} k &\sim \mathrm{Cat}(q_1, \dots, q_K) \\ \theta|k &\sim \pi_k \end{aligned}$$

Hence the model (1.2) with mixture prior is a simple example of *hierarchial model*

$$\begin{aligned} k &\sim \mathrm{Cat}(q_1, \dots, q_K) \\ \theta|k &\sim \pi_k \\ X_1, \dots, X_n|\theta &\overset{iid}{\sim} f(\cdot|\theta). \end{aligned}$$

Example of Beta-Binomial mixtures. As an example, consider the mixture of Beta-binomial models, where $\pi_k = \mathrm{Beta}(\alpha_k, \beta_k)$ and given parameter θ, the observation has $\mathrm{B}(n, \theta)$ distribution:

$$\begin{aligned} k &\sim \mathrm{Cat}(q_1, \dots, q_K) \\ \theta|k &\sim \mathrm{Beta}(\alpha_k, \beta_k) \\ X|\theta &\sim \mathrm{B}(n, \theta). \end{aligned}$$

We know that the posterior distribution is the mixture, and by (3.14) the posterior density is

$$\pi(\theta|x) = \sum_{k=1}^{K} q_k(x) \pi_k(\theta|x),$$

where

- $\pi_k(\theta|x)$ is the posterior density corresponding to prior $\pi_k = \mathrm{Beta}(\alpha_k, \beta_k)$, i.e. (recall (2.4))

$$\pi_k(\theta|x) = \frac{1}{B(\alpha_k + x, \beta_k + n - x)} \theta^{\alpha_k + x} (1-\theta)^{\beta_k + (n-x)}$$

- The weights $q_k(x)$ are

$$q_k(x) = \frac{q_k f_k(x)}{\sum_{j=1}^{K} q_j f_j(x)},$$

where $f_k(x)$ is the density of Beta-binomial distribution (2.2), i.e.

$$f_k(x) = \binom{n}{x} \frac{B(\alpha_k + x, \beta_k + n - x)}{B(\alpha_k, \beta_k)}.$$

Thus

$$q_k(x) \propto q_k \frac{B(\alpha_k + x, \beta_k + n - x)}{B(\alpha_k, \beta_k)}, \quad \text{i.e.} \quad q_k(x) = \frac{q_k \frac{B(\alpha_k + x, \beta_k + n - x)}{B(\alpha_k, \beta_k)}}{\sum_{j=1}^{K} q_j \frac{B(\alpha_j + x, \beta_j + n - x)}{B(\alpha_j, \beta_j)}}.$$

More about this example (motivation, pictures), see [6], Example 3.4.1.

3.1 Exercises

1. Let $f(x|\theta)$ be an exponential family in canonical form, i.e. $\phi(\theta) = \theta$. Assume $\theta \in \mathbb{R}$ and consider the model (1.2).

   - Prove (3.3) and (3.4).
   - $\pi(\theta|\mu, \lambda)$ ba conjugate prior as in (3.7). Knowing that

     $$\int \pi'(\theta|\mu,\lambda)d\theta = \Big(\int \pi(\theta|\mu,\lambda)d\theta\Big)' = 0,$$

     show that

     $$E[u(X_1)] = EA'(\theta) = \int A'(\theta)\pi(\theta)d\theta = \mu/\lambda.$$

   - Show that with $x = (x_1, \ldots, x_n)$ being iid sample,

     $$E[u(X_{n+1})|x] = \frac{\mu + t(x)}{\lambda + n}.$$

2. Let $\{f(x|\theta) = 1/\theta, \quad \theta \geq \theta_o\}$ be a family of uniform distributions. Prove that Pareto distributions constitute a conjugate family. Find the posterior density $\pi(\theta|x)$, where $x = (x_1, \ldots, x_n)$ are iid observations.

3. Let

   $$\{f(x|\theta) = \theta x^{-(\theta+1)} I_{[1,\infty)}(x), \quad \theta > 0\}$$

   be a family of Pareto distributions. Find conjugate priors. Find the posterior density $\pi(\theta|x)$, where $x = (x_1, \ldots, x_n)$ are iid observations.

4. Let

   $$\{f(x|\theta) = \alpha\theta^{\alpha} x^{-(\alpha+1)} I_{[\theta,\infty)}(x), \quad \theta > 0\}$$

   be a family of Pareto distributions. Prove that the following family of priors constitute a conjugate family:

   $$\Big\{\pi(\theta) = \frac{a}{b^a}\theta^{a-1} I_{[0,b]}(\theta), \quad a > 0, b > 0\Big\}.$$

   Find the posterior density $\pi(\theta|x)$, where $x = (x_1, \ldots, x_n)$ are iid observations.

5. Prove that the following family of priors constitute a conjugate family for the shifted exponential distributions $\mathrm{Exp}(\theta, \lambda)$:

   $$\Big\{\pi(\theta) = \alpha e^{\alpha(\theta-\beta)} I_{(-\infty,\beta]}(\theta), \quad \alpha > 0, \beta > 0\Big\}.$$

   Find the posterior density $\pi(\theta|x)$, where $x = (x_1, \ldots, x_n)$ are iid observations.

6. Let $\pi(\theta|\lambda,\mu)$ be the prior density (3.7), thus

$$\pi(\theta|\mu,\lambda) = K(\mu,\lambda)g(\theta)^{\lambda}\exp[\phi(\theta)\cdot\mu], \quad K^{-1}(\mu,\lambda) = \int g(\theta)^{\lambda}\exp[\phi(\theta)\cdot\mu]d\theta.$$

   Show that the posterior predicative density is

$$f(x_{n+1}|x) = \frac{h(x_{n+1})K(\mu+t(x),\lambda+n)}{K(\mu+t(x)+u(x_{n+1}),\lambda+n+1)}.$$

7. Find conjugate prior for negative binomial distribution, the parameter is p. Find the posterior density $\pi(\theta|x)$, where $x=(x_1,\ldots,x_n)$ are iid observations.

8. Prove (3.14), (3.15), (3.16).

4 Noninformative and unproper priors

4.1 Flat prior

When no prior information is available, and the parameter space Θ is bounded, then it is tempting to use uniform distribution with constant density $\pi(\theta)\equiv$ const. In this case all parameters are "equiprobable" and the researcher has no preferences. That was the reasoning of P. Laplace for using uniform distribution in his studies about birth rates. Observe that the uniform prior also matches with the maximum likelihood principle, because MLE estimator is the posterior mode under uniform prior:

$$\arg\max_{\theta}\pi(\theta|x) = \arg\max_{\theta}\frac{f(x|\theta)\pi(\theta)}{f(x)} = \arg\max_{\theta} f(x|\theta)\pi(\theta) = \arg\max_{\theta} f(x|\theta).$$

The uniform (and uniform-like priors with possibly flat densities) are called *noninformative*. Noninformative priors are often a popular choice in practice.

When Θ is not bounded, there is no uniform prior on Θ. For example, there is no uniform distribution on $\mathbb{R}$ or $[0,\infty)$. There is, however, Lebesgue measure and although it is not a probability measure, it is tempting to use it as it were. This means plugging $\pi(\theta)\equiv c$ into Bayes formula and calculating posterior

$$\pi(\theta|x) = \frac{f(x|\theta)c}{\int f(x|\theta)cd\theta} = \frac{f(x|\theta)}{\int f(x|\theta)d\theta}. \tag{4.1}$$

When x is such that $\int f(x|\theta)d\theta < \infty$, then $\pi(\theta|x)$ is still a density of a probability measure.

Improper prior. In general, a (prior) density function $\pi(\cdot)$ (i.e. non-negative and measurable function) on Θ is called a density of *improper prior measure*, when

$$\int \pi(\theta)d\theta = \infty.$$

Thus the corresponding prior measure is not a finite measure and certainly not a probability measure. When the observations x satisfy:

$$\int f(x|\theta)\pi(\theta)d\theta < \infty, \tag{4.2}$$

then the posterior density

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$$

is well defined probability density and, therefore, posterior probability measure (for that particular x) exists. However, even the posterior exists for all x, the Bayesian models with improper prior cannot be considered as standard probability model, where (X, θ) is a random vector and posterior probability is the conditional probability of parameter θ given $X = x$.

4.2 Example: normal models with flat unproper priors

Since normal models with flat unproper priors are very often used in practice, we study them closer. We see that although the priors are not conjugate, the posterior distributions can be analytically found. It is partially, because they can be considers as a limits of conjugate priors.

4.2.1 One parameter normal models

Known variance. Consider the normal model with known variance $\mathcal{N}(\theta, \sigma^2)$. We know that conjugate prior is normal $\mathcal{N}(\mu, \tau^2)$. The smaller is τ, the flatter is the prior but there will be always more prior mass around μ. Let us calculate $\pi(\theta|x)$ according to (4.1) (with $x = (x_1, \ldots, x_n)$):

$$\sum_{i=1}^{n}(x_i - \theta)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \theta)^2$$

so that

$$f(x) = \int f(x|\theta)d\theta = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp[-\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{2\sigma^2}] \int \exp[-\frac{(\bar{x} - \theta)^2}{2\sigma^2/n}]d\theta.$$

Therefore

$$f(x) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp[-\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{2\sigma^2}]\sqrt{\frac{2\pi\sigma^2}{n}} = \frac{1}{\sqrt{n}(2\pi\sigma^2)^{\frac{n-1}{2}}} \exp[-\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{2\sigma^2}]. \tag{4.3}$$

$$\pi(\theta|x) = \frac{f(x|\theta)}{\int f(x|\theta)d\theta} = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp[-\frac{(\bar{x}-\theta)^2}{2\sigma^2/n}] \tag{4.4}$$

and we recognize the density of $\mathcal{N}(\bar{x}, \sigma^2/n)$ distribution. Hence the posterior is normal:

$$\theta|x \sim \mathcal{N}(\bar{x}, \sigma^2/n). \tag{4.5}$$

Observe that with (posterior variance and mean under $\mathcal{N}(\mu, \tau^2)$ prior)

$$\tau_n^2 = (n/\sigma^2 + 1/\tau^2)^{-1}, \quad \mu_n = \frac{\mu/\tau^2 + \sum x_i}{n + 1/\tau^2},$$

it holds

$$\bar{x} = \lim_{\tau^2 \to \infty} \mu_n \quad \sigma^2/n = \lim_{\tau^2 \to \infty} \tau_n^2.$$

Hence the flat prior can be considered as a limit of normal prior as $\tau^2 \to \infty$ (in a sense), and so is the posterior as the limit $\tau^2 \to \infty$.

Known mean. Another classical example of improper and non-informative prior is the normal model with unknown variance $\mathcal{N}(\mu, \theta)$. We know that the conjugate prior is scaled-inverse χ^2-distribution with density

$$\pi(\theta|\nu, \tau^2) = \frac{1}{\Gamma(\nu/2)} \Big(\frac{\nu\tau^2}{2}\Big)^{\frac{\nu}{2}} \theta^{-(\nu/2+1)} \exp[-\frac{\nu\tau^2}{2\theta}].$$

The density is becomes flatter as $\nu \to 0$, the limit:

$$\lim_{\nu \to 0} \theta^{-(\nu/2+1)} \exp[-\frac{\nu\tau^2}{2\theta}] = \theta^{-1}.$$

Since $\int_0^\infty \theta^{-1} d\theta = \infty$, we see that $\pi(\theta) = 1/\theta$ is a density of an improper prior. It still puts more prior mass to smaller variance but in a sense it is less informative than any $\pi(\theta|\nu, \tau^2)$. Since

$$\int_0^\infty \theta^{-\frac{n}{2}} \exp\big[-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\theta}\big] \frac{1}{\theta} d\theta = \int_0^\infty \theta^{-(\frac{n}{2}+1)} \exp\big[-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\theta}\big] d\theta < \infty,$$

we see that (4.2) holds for $n \geq 1$ (at least one observation) and so the posterior distribution is well defined. Since

$$\pi(\theta|x) \propto \theta^{-(\frac{n}{2}+1)} \exp\big[-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\theta}\big], \tag{4.6}$$

we see that $\pi(\theta|x)$ is the density of $\text{ScaleInv} - \chi^2(\nu_n, \tau_n^2)$, where

$$\nu_n = n, \quad \tau_n^2 = \frac{\sum_{i=1}^n (x_i-\mu)^2}{n}.$$

This justifies the interpretation of the improper prior $\frac{1}{\theta}$ as ScaleInv$-\chi^2(0,\tau^2)$, because when $\nu > 0$, then ScaleInv $-\ \chi^2(\nu,\tau^2)$ prior gives the posterior ScaleInv $-\ \chi^2(\nu_n,\tau_n^2)$, where

$$\nu_n = \nu + n, \quad \tau_n^2 = \frac{\sum_{i=1}^n (x_i-\mu)^2 + \nu\tau^2}{n+\nu}.$$

Observe that in this example $\int f(x|\theta)d\theta < \infty$, when $n > 2$, but for $n \leq 2$ $\int f(x|\theta)d\theta = \infty$ so the constant prior density (Lebesgue measure) would not always give a well-defined posterior distribution for sample size $n = 1, 2$.

4.2.2 Unknown mean and variance

A popular flat choice for the prior for mean and variance is the product of previously obtained priors, i.e. $\pi(\mu,\sigma^2) = (\sigma^2)^{-1}$. This prior measure corresponds to the product measure and is sometimes interpreted as the independence of mean and variance. However, the concept of independence holds for probability measures only, so for improper priors such an interpretation is incorrect.
Since $(x = (x_1,\ldots,x_n))$

$$\begin{aligned}
&\int\int f(x|\mu,\sigma^2)\pi(\mu,\sigma^2)d\mu d\sigma^2 = \\
&\int (2\pi\sigma^2)^{-\frac{n}{2}} \int \exp[-\frac{\sum_i (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}](\sigma^2)^{-1}d\mu d\sigma^2 = \\
&(2\pi)^{-\frac{n}{2}} \int (\sigma^2)^{-\frac{(n+2)}{2}} \exp[-\frac{\sum_i (x_i-\bar{x})^2}{2\sigma^2}] \int \exp[-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}]d\mu d\sigma^2 = \\
&(2\pi)^{-\frac{n}{2}} \int (\sigma^2)^{-\frac{(n+2)}{2}} (2\pi\frac{\sigma^2}{n})^{\frac{1}{2}} \exp[-\frac{\sum_i (x_i-\bar{x})^2}{2\sigma^2}]d\sigma^2 = \\
&(2\pi)^{-\frac{n-1}{2}} \int (\sigma^2)^{-(\frac{(n-1)}{2}+1)} \exp[-\frac{\sum_i (x_i-\bar{x})^2}{2\sigma^2}]d\sigma^2 < \infty,
\end{aligned}$$

because with

$$\nu = n-1, \quad \tau^2 = \sum_i (x_i-\bar{x})^2/(n-1)$$

it holds

$$(\sigma^2)^{-(\frac{(n-1)}{2}+1)} \exp[-\frac{\sum_i (x_i-\bar{x})^2}{2\sigma^2}] = (\sigma^2)^{-(\nu/2+1)} \exp[-\frac{\nu\tau^2}{2\sigma^2}]$$

and so the right hand side is proportional to ScaleInv $-\ \chi^2(\nu,\tau^2)$ density. Thus (4.2) holds and $\pi(\mu,\sigma^2|x)$ is a probability density. Let us find this distribution.

Posterior distribution. To find posterior, we factorize it:

$$\pi(\mu, \sigma^2|x) = \pi(\sigma^2|x)\pi(\mu|\sigma^2, x).$$

Observe that

$$\pi(\mu|\sigma^2, x) \propto \exp[-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}],$$

so that given σ^2, the mean is normally distributed:

$$\mu|\sigma^2, x \sim \mathcal{N}(\bar{x}, \sigma^2/n).$$

Since

$$\pi(\sigma^2|x) \propto \int f(x|\mu,\sigma^2)\pi(\mu,\sigma^2)d\mu \propto (\sigma^2)^{-(\frac{(n-1)}{2}+1)} \exp[-\frac{\sum_i (x_i-\bar{x})^2}{2\sigma^2}],$$

we see that

$$\sigma^2|x \sim \text{ScaleInv} - \chi^2\Big(n-1, \frac{\sum_i (x_i-\bar{x})^2}{n-1}\Big).$$

The obtained distributions are used in sampling: first generate σ^2 from $\pi(\sigma^2|x)$ and given σ^2, generate μ from $\pi(\mu|\sigma^2, x)$.

So the posterior is NIX with parameters

$$\nu_n = n-1, \quad \tau_n^2 = \frac{\sum_i (x_i-\bar{x})^2}{n-1}, \quad \mu_n = \bar{x}, \quad \kappa_n = n.$$

Sometimes it is of interest find the posterior of mean $\pi(\mu|x)$ (σ^2 is a kind of nuisance parameter). To find that distribution, integrate

$$\begin{aligned}
\int f(x|\mu,\sigma^2)\pi(\mu,\sigma^2)d\sigma^2 &= (2\pi)^{-\frac{n}{2}} \int (\sigma^2)^{-(\frac{n}{2}+1)} \exp[-\frac{\overbrace{\sum (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}^{\nu\tau^2}}{2\sigma^2}]d\sigma^2 \\
&= (2\pi)^{-\frac{n}{2}} \Gamma(n/2) \Big(\frac{\sum_i (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2}\Big)^{-n/2} \\
&\propto \Big(1 + \frac{n(\bar{x}-\mu)^2}{\sum_i (x_i-\bar{x})^2}\Big)^{-\frac{n-1+1}{2}} \\
&\propto \Big(1 + \frac{1}{n-1}\frac{(\bar{x}-\mu)^2}{\frac{\sum_i (x_i-\bar{x})^2}{n(n-1)}}\Big)^{-\frac{n-1+1}{2}}. \qquad (4.7)
\end{aligned}$$

Here we used:

$$\int_0^\infty (\sigma^2)^{-(\nu/2+1)} \exp[-\frac{\nu\tau^2}{2\sigma^2}]d\sigma^2 = \big(\frac{\nu\tau^2}{2}\big)^{-\frac{\nu}{2}} \Gamma(\frac{\nu}{2}). \qquad (4.8)$$

Hence

$$\mu|x \sim \text{lst}\Big(\bar{x}, \frac{\sum_i (x_i-\bar{x})^2}{n(n-1)}, n-1\Big).$$

Posterior predicative distribution.

$$f(x_{n+1}|x) = \int\int f(x_{n+1}|\mu,\sigma^2)\pi(\mu|\sigma^2,x)d\mu\pi(\sigma^2|x)d\sigma^2.$$

Using (2.15), and the fact that $\pi(\mu|\sigma^2,x)$ is Gaussian with mean $\bar{x}$ and variance σ^2/n, we obtain

$$X_{n+1}|x,\sigma^2 \sim \mathcal{N}(\bar{x},\sigma^2(1+1/n)),$$

i.e.

$$\int f(x_{n+1}|\mu,\sigma^2)\pi(\mu|\sigma^2,x)d\mu = (2\pi\sigma^2(1/n+1))^{-\frac{1}{2}}\exp[-\frac{(x_{n+1}-\bar{x})^2}{2\sigma^2(1/n+1)}].$$

Now integrate σ^2 out

$$\int (2\pi\sigma^2(1/n+1))^{-\frac{1}{2}}\exp[-\frac{(x_{n+1}-\bar{x})^2}{2\sigma^2(1/n+1)}]\cdot$$
$$\cdot\frac{1}{\Gamma(n-1)}\Big(\frac{\sum_{i=1}^n(x_i-\bar{x})^2}{2}\Big)^{\frac{n-1}{2}}(\sigma^2)^{-(\frac{n-1}{2}+1)}\exp[-\frac{\sum_{i=1}^n(x_i-\bar{x})^2}{2\sigma^2}]d\sigma^2.$$

Since (recall (4.8))

$$\int(\sigma^2)^{-(\frac{n}{2}+1)}\exp\Big[-\frac{n}{2\sigma^2}\overbrace{\Big(\frac{(x_{n+1}-\bar{x})^2}{n+1}+\frac{\sum_{i=1}^n(x_i-\bar{x})^2}{n}\Big)}^{\tau^2}\Big]d\sigma^2 = \Gamma(\frac{n}{2})(\frac{n\tau^2}{2})^{-\frac{n}{2}},$$

we see that

$$f(x_{n+1}|x) \propto \Big(\frac{(x_{n+1}-\bar{x})^2}{n+1}+\frac{\sum_{i=1}^n(x_i-\bar{x})^2}{n}\Big)^{-\frac{n}{2}} \propto \Big(1+\frac{1}{n-1}\frac{(x_{n+1}-\bar{x})^2}{\frac{\sum_{i=1}^n(x_i-\bar{x})^2}{n-1}}\frac{n}{(n+1)}\Big)^{-\frac{n-1+1}{2}}.$$

Hence

$$X_{n+1}|x \sim \mathrm{lst}\Big(\bar{x},\frac{\sum_{i=1}^n(x_i-\bar{x})^2}{n-1}(1+\frac{1}{n}),n-1\Big). \tag{4.9}$$

See also [5], sec 3.2.

4.2.3 Application: ordinary least squared regression

The setting. Let $x_1,\ldots,x_n \in \mathbb{R}^k$ be fixed covariates/explanatory variables and consider the standard regression model:

$$\begin{aligned}(\beta,\sigma^2) &\sim \pi\\ Y_i|\beta,\sigma^2 &\overset{ind.}{\sim} \mathcal{N}(\beta' x_i,\sigma^2),\quad i=1,\ldots,n\end{aligned} \tag{4.10}$$

Here β is k-dim random vector. Observe that when $k=1$ and $x_i=1$ for every i, then we obtain the model considered on previous paragraph (iid Gaussian).

Let X be $n \times k$ matrix (i-th row is x_i') and let Y be n-dimensional random vector of responses $Y = (Y_1, \ldots, Y_n)'$. Then with I_n being $n \times n$ identity matrix, (4.10) is

$$\begin{aligned} &(\beta, \sigma^2) \sim \pi \\ &Y|\beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 I_n), \quad i = 1, \ldots, n \end{aligned} \tag{4.11}$$

Hence, the density of Y given the parameters is

$$f(y_1, \ldots, y_n|\beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\big[-\frac{(y - X\beta)'(y - X\beta)}{2\sigma^2}\big], \quad y = (y_1, \ldots, y_n)'.$$

Let $\hat{\beta} = (X'X)^{-1}X'y$ be the OLS regression estimate. Then (Exercise 1)

$$(y - X\beta)'(y - X\beta) = (y - X\hat{\beta})'(y - X\hat{\beta}) + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \tag{4.12}$$

and so

$$f(y, \beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\big[-\frac{(y - X\hat{\beta})'(y - X\hat{\beta}) + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})}{2\sigma^2}\big]\pi(\beta, \sigma^2).$$

With the flat prior as in the previous example:

$$\pi(\beta, \sigma^2) = (\sigma^2)^{-1},$$

we obtain

$$f(y, \beta, \sigma^2) = (2\pi\sigma^2)^{-(\frac{n}{2}+1)} \exp\big[-\frac{(y - X\hat{\beta})'(y - X\hat{\beta}) + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})}{2\sigma^2}\big].$$

Posterior. We now find the posterior distribution of parameters (the posterior is proper as soon as $n > k$):

$$\pi(\beta, \sigma^2|y) = \pi(\beta|\sigma^2, y)\pi(\sigma^2|y)$$

To find $\pi(\beta|\sigma^2, y)$, observe (Exercise 1)

$$\pi(\beta|y, \sigma^2) \propto \exp[-\frac{(\beta - \hat{\beta})'X'X(\beta - \hat{\beta})}{2\sigma^2}], \tag{4.13}$$

so that

$$\beta|y, \sigma^2 \sim \mathcal{N}\big(\hat{\beta}, (X'X)^{-1}\sigma^2\big).$$

In order to find $\pi(\sigma^2|y)$, observe that (Exercise 1) with c being independent of σ^2,

$$\int \exp[-\frac{(\beta - \hat{\beta})'X'X(\beta - \hat{\beta})}{2\sigma^2}]d\beta = c \times (\sigma^2)^{\frac{k}{2}} \tag{4.14}$$

and so (Exercise 1)

$$\int f(y,\beta,\sigma^2)d\beta \propto (\sigma^2)^{-\left(\frac{n-k}{2}+1\right)} \exp\Big[-\frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{2\sigma^2}\Big] \tag{4.15}$$

$$= (\sigma^2)^{-\left(\frac{\nu}{2}+1\right)} \exp[-\frac{\nu\tau^2}{2\sigma^2}], \quad \nu = n-k, \quad \tau^2 = \frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n-k}.$$

Therefore

$$\sigma^2|y \sim \text{ScaleInv} - \chi^2(\nu,\tau^2).$$

Easy to sample: first σ^2 from ScaleInv$-\chi^2(\nu,\tau^2)$ and then $\beta|\sigma^2$ from $\mathcal{N}\big(\hat{\beta},(X'X)^{-1}\sigma^2\big)$.

Observe that when $k=1$ and $x_i=1$, $i=1,\ldots,n$, then $\hat{\beta}=\bar{y}$ and then the obtained distribution is exactly the same as in the previous example.

To get $\pi(\beta|y)$, integrate σ^2 out (Exercise 1):

$$\int_0^\infty f(y,\beta,\sigma^2)d\sigma^2 \propto \Big(1+\frac{1}{(n-k)}\frac{(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})}{(y-X\hat{\beta})'(y-X\hat{\beta})}(n-k)\Big)^{-\frac{(n-k)+k}{2}}. \tag{4.16}$$

Hence $\beta|y$ follows *multivariate lst* distribution with parameters $\hat{\beta}$ (mean),

$$\frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{(n-k)}(X'X)^{-1}$$

(scale matrix) and $n-k$ (degrees of freedom). Observe that when $k=1$ and $x_i=1$, $i=1,\ldots,n$, then (4.16) reduces to (4.7). The marginal of multivariate lst-distribution is univariate lst-distribution:

$$\beta_1|y \sim \text{lst}\big(\hat{\beta}_1, \frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{(n-k)}(X'X)^{-1}_{11}, n-k\big),$$

where $(X'X)^{-1}_{11}$ is the first row and column element of $(X'X)^{-1}$.

Predicative distribution. Let x_{n+1} be one more covariate and let us find the predicative distribution of the response Y_{n+1}.

$$f(y_{n+1},\beta|\sigma^2,y) = f(y_{n+1}|\beta,\sigma^2,y)\pi(\beta|\sigma^2,y)$$

$$= (2\pi\sigma^2)^{-\frac{1}{2}} \exp[-\frac{(y_{n+1}-\beta' x_{n+1})^2}{2\sigma^2}](2\pi)^{-\frac{k}{2}}|(X'X)^{-1}\sigma^2|^{-\frac{1}{2}} \exp[-\frac{(\beta-\hat{\beta})'(X'X)(\beta-\hat{\beta})}{2\sigma^2}]$$

$$\propto \exp\Big[-\frac{(y_{n+1}-\beta' x_{n+1})^2+(\beta-\hat{\beta})'(X'X)(\beta-\hat{\beta})}{2\sigma^2}\Big].$$

We recognize that (Y_{n+1}, β) must be jointly normally distributed, hence $Y_{n+1}|\sigma^2, y \sim \mathcal{N}$. To find expectation and variance, observe

$$\begin{aligned} E[Y_{n+1}|\sigma^2, y] &= E[E[Y_{n+1}|\beta, \sigma^2, y]|\sigma^2, y] = E[\beta' x_{n+1}|\sigma^2, y] = \hat{\beta}' x_{n+1} \\ \mathrm{Var}[Y_{n+1}|\sigma^2, y] &= E[\mathrm{Var}[Y_{n+1}|\beta, \sigma^2, y]|\sigma^2, y] + \mathrm{Var}\big[E[Y_{n+1}|\beta, \sigma^2, y]|\sigma^2, y\big] \\ &= \sigma^2 + \mathrm{Var}[\beta' x_{n+1}|\sigma^2, y] = \sigma^2(1 + x'_{n+1}(X'X)^{-1}x_{n+1}). \end{aligned}$$

$$Y_{n+1}|\sigma^2, y \sim \mathcal{N}\big(\hat{\beta}' x_{n+1}, \sigma^2(1 + x'_{n+1}(X'X)^{-1}x_{n+1})\big).$$

So the expected value of the new response variable (given σ^2) is $\hat{\beta}x_{n+1}$, just like in ordinary regression, but the due to the random β, the variance is bigger than σ^2.

Denote $\mu = \hat{\beta}' x_{n+1}$, and $a = (1 + x'_{n+1}(X'X)^{-1}x'_{n+1})$. To obtain the distribution $Y_{n+1}|y$, integrate σ^2 out:

$$\int_0^\infty f(y_{n+1}|\sigma^2, y)\pi(\sigma^2|y)d\sigma^2 = \int_0^\infty (2\pi\sigma^2 a)^{-\frac{1}{2}} \exp[-\frac{(y_{n+1}-\mu)^2}{2\sigma^2 a}](\sigma^2)^{-\left(\frac{\nu}{2}+1\right)} \exp[-\frac{\nu\tau^2}{2\sigma^2}]d\sigma^2.$$

Since (recall (4.8))

$$\int_0^\infty (\sigma^2)^{-(\frac{\nu+1}{2}+1)} \exp[-\frac{a^{-1}(y_{n+1}-\mu)^2 + \nu\tau^2}{2\sigma^2}]d\sigma^2 = \Big(\frac{a^{-1}(y_{n+1}-\mu)^2 + \nu\tau^2}{2}\Big)^{-\frac{\nu+1}{2}} \Gamma(\frac{\nu+1}{2}),$$

we obtain (Exercise 1), that

$$Y_{n+1}|y \sim \mathrm{lst}\Big(\hat{\beta}x_{n+1}, \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n-k}(1 + x'_{n+1}(X'X)^{-1}x_{n+1}), n-k\Big). \tag{4.17}$$

Observe that when $k = 1$ and $x_i = 1$, $i = 1, \ldots, n$, then (4.17) reduces to (4.9).

The expected value of the new response variable is $\hat{\beta}x_{n+1}$, the variance is (when $n > k + 2$) is

$$\mathrm{Var}[Y_{n+1}|y] = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n-k-2}(1 + x'_{n+1}(X'X)^{-1}x_{n+1}).$$

As an application of such regression in political science, see [5], sec. 14.2.

4.3 Jeffrey's prior

4.3.1 Change of variables

Univariate. Let Θ, Θ_η be open subsets of $\mathbb{R}$ and let $g : \Theta \to \Theta_\eta$ be a one-to-one (monotone) function. When $\pi(\theta)$ is a density of θ with respect to the Lebesgue measure, then the random variable $\eta = g(\theta)$ has the density

$$\pi^*(\eta) = \pi(h(\eta))|h'(\eta)|, \quad h(\eta) := g^{-1}(\eta).$$

Since $h'(\eta) = 1/g'(h(\eta))$ (because $g(h(\eta)) = \eta$ and $g'(h(\eta))h'(\eta) = 1$), the formula above reads

$$\pi^*(\eta) = \frac{\pi(h(\eta))}{|g'(h(\eta))|}$$

Since $h(\eta) = \theta$, we obtain the change of variable formula also in other direction

$$\pi^*(\eta) = \frac{\pi(\theta)}{|g'(\theta)|} \quad \Leftrightarrow \quad \pi(\theta) = \pi^*(g(\theta))|g'(\theta)|. \tag{4.18}$$

Example: Let $\Theta = (0,1)$ and $\pi(\theta) \equiv 1$ be the density of uniform distribution. Let $\theta \sim U(0,1)$ and define $\eta = \frac{\theta}{1-\theta}$. Thus $\Theta_\eta = (0,\infty)$. Now

$$g(\theta) = \frac{\theta}{1-\theta}, \quad h(\eta) = \frac{\eta}{1+\eta}, \quad h'(\eta) = \frac{1}{(1+\eta)^2}.$$

The density of η is thus

$$\pi^*(\eta) = \frac{1}{(1+\eta)^2}, \quad \eta \in (0,\infty).$$

We see that the distribution of η is not uniform – more weights for small η (corresponding to small θ).

Similarly taking

$$\eta = g(\theta) = \ln \frac{\theta}{1-\theta}, \quad h(\eta) = \frac{e^\eta}{1+e^\eta}, \quad h'(\eta) = \frac{e^\eta}{(1+e^\eta)^2}.$$

Thus

$$\pi^*(\eta) = \frac{e^\eta}{(1+e^\eta)^2}.$$

Again, the measure (although symmetric) is not uniform – more weights around zero (corresponding $\theta = 0.5$).

Multivariate. Let $\Theta, \Theta_\eta \subseteq \mathbb{R}^d$ (open subsets) and let $g : \Theta \to \Theta_\eta$ be a one-to-one differentiable mapping. Let $\pi(\theta)$ be the density (w.r.t. Lebesgue measure) of θ. Denote by $h = g^{-1}$ the inverse of g and let $J(\eta)$ be Jacobian matrix of h evaluated at η, i.e. with $h(\eta) = (h_1(\eta), \ldots, h_d(\eta))'$ (transposed)

$$J(\eta) = J(\eta_1, \ldots, \eta_d) = \Big(\frac{\partial h_i(\eta)}{\partial \eta_j}\Big)_{ij}.$$

Then with $|J(\eta)|$ being absolute value of the the determinant of Jacobian, the random vector $\eta = g(\theta)$ has density

$$\pi^*(\eta) = \pi(h(\eta))|J(\eta)| = \pi(\theta)|J(\eta)|.$$

In other direction:

$$\pi(\theta) = \pi^*(g(\theta))|J(\theta)| = \pi^*(\eta)|J(\theta)|.$$

Hence $|J(\theta)||J(\eta)| = 1$.

For example, let θ_1, θ_2 be iid standard normal variables. Let us find the distribution of the random vector (η_1, η_2), where $\eta_1 = \theta_1 + \theta_2$ and $\eta_2 = \theta_1 - \theta_2$. The inverse mapping is

$$h(\eta) = h\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \frac{\eta_1+\eta_2}{2} \\ \frac{\eta_1-\eta_2}{2} \end{pmatrix}, \quad J(\eta) = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix}, \quad |J(\theta)| = 1/2$$

Let $\phi(\theta)$ theta be the density of standard normal. Thus $\pi(\theta) = \phi(\theta_1)\phi(\theta_2)$ and

$$\pi^*(\eta) = \frac{1}{2}\phi(\frac{\eta_1+\eta_2}{2})\phi(\frac{\eta_1-\eta_2}{2}) = \frac{1}{4\pi}\exp[-\frac{1}{2}(\frac{\eta_1+\eta_2}{2})^2+(\frac{\eta_1-\eta_2}{2})^2] = \frac{1}{4\pi}\exp[-\frac{\eta_1^2+\eta_2^2}{4}].$$

Thus

$$\pi^*(\eta_1, \eta_2) = \frac{1}{\sqrt{2\pi 2}}\exp[-\frac{\eta_1^2}{4}]\frac{1}{\sqrt{2\pi 2}}\exp[-\frac{\eta_2^2}{4}]$$

and hence η_1,η_2 are iid $\mathcal{N}(0, \sqrt{2}^2)$ distributed random variables.

Reparametrization (switching from θ to η) changes the prior distribution and this is the main criticism against the flat (uninformative) prior – it is flat under a certain parametrization, only. Having uniform prior in Beta-Bernoulli model might be interpreted as having no prior information about the success probability θ. In terms of odds ratio $\eta = \frac{\theta}{1-\theta}$, the same prior has density $(1+\eta)^{-2}$ so that more weights are put on smaller odds. In terms of log odds, however, more weights are put around zero.

4.3.2 Fisher information

Let $\Theta \subset \mathbb{R}$ (open) and let us assume that $\theta \mapsto f(x|\theta)$ are differentiable functions for every x. The *Fisher information* is the following quantity:

$$I(\theta) := E\big[\big(\frac{\partial}{\partial\theta}\ln f(X|\theta)\big)^2\big|\theta\big].$$

Observe that

$$\frac{\partial}{\partial\theta}\ln f(x|\theta) = \frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}$$

so that

$$E\big[\frac{\partial}{\partial\theta}\ln f(X|\theta)\big|\theta\big] = \int \frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}f(x|\theta)dx = \int \frac{\partial}{\partial\theta}f(x|\theta)dx.$$

Under so-called *regularity conditions* the order of taking derivatives and integral can be switched (i.e. on can move with derivation under the integral), then

$$\int \frac{\partial}{\partial\theta} f(x|\theta)dx = \frac{\partial}{\partial\theta}\int f(x|\theta)dx = 0.$$

Then $I(\theta)$ is the conditional (on θ) variance of a random variable $\frac{\partial}{\partial\theta}\ln f(X|\theta)$, because the expectation of this random variable is 0. Usually Fisher information is considered for regular families, otherwise it is meaningful to define it using variance instead of second moments.

An alternative definition. If $\theta \mapsto \ln f(x|\theta)$ (and hence $\theta \mapsto \ln f(x|\theta)$) are twice differentiable functions for every x then under regularity (that allows to change the order of differentiation and integration)

$$\int \frac{\partial^2}{\partial\theta^2} f(x|\theta)dx = \frac{\partial^2}{\partial\theta^2}\int f(x|\theta)dx = 0$$

implying that (Exercise 3)

$$I(\theta) = -E\big[\frac{\partial^2}{\partial\theta^2}\ln f(X|\theta)\big|\theta\big]. \tag{4.19}$$

Fisher information of independent observations. When $X_1, \dots, X_n$ are independent random variables with density $X_i \sim f_i(\cdot|\theta)$ (the same parameter θ for every i), then under regularity the Fisher information of the random vector $(X_1, \dots, X_n)$ (i.e. the model $f(x|\theta) = \prod_{i=1}^n f_i(x_i|\theta)$, $x = (x_1, \dots, x_n)$) – let it be $I_n(\theta)$ – is the sum of components information (Exercise 3):

$$I_n(\theta) = \sum_{i=1}^n I_i(\theta), \quad I_i(\theta) = E\big[\big(\frac{\partial}{\partial\theta}\ln f_i(X_i|\theta)\big)^2\big|\theta\big]. \tag{4.20}$$

Hence, when $X_1, \dots, X_n$ are iid from $f(\cdot|\theta)$, then $I_n(\theta) = nI(\theta)$.

Change of variables and Fisher information. Let Θ, Θ_η be open subsets of $\mathbb{R}$ and let $g : \Theta \to \Theta_\eta$ be a differentiable one-to-one function. Let us denote the density in η-parametrization as $p(x|\eta)$. Thus $p(x|\eta) := f(x|g^{-1}(\eta))$ or, equivalently, $f(x|\theta) = p(x|g(\theta))$. Then

$$\frac{\partial}{\partial\theta}\ln f(x|\theta) = \frac{\partial}{\partial\theta}\ln p(x|g(\theta)) = \frac{\partial}{\partial\eta}\ln p(x|\eta)g'(\theta), \tag{4.21}$$

so that with $g(\theta) = \eta$.

$$\begin{aligned} I(\theta) &= \int \big(\frac{\partial}{\partial\theta}\ln f(x|\theta)\big)^2 f(x|\theta)dx = (g'(\theta))^2\int \big(\frac{\partial}{\partial\eta}\ln p(x|\eta)\big)^2 f(x|\theta)dx \\ &= (g'(\theta))^2\int \big(\frac{\partial}{\partial\eta}\ln p(x|\eta)\big)^2 p(x|\eta)dx = (g'(\theta))^2 I_\eta(\eta), \end{aligned}$$

where

$$I_\eta(\eta) := \int \big(\frac{\partial}{\partial \eta} \ln p(x|\eta)\big)^2 p(x|\eta) dx.$$

Hence, we have

$$I(\theta) = (g'(\theta))^2 I_\eta(g(\theta)) = (g'(\theta))^2 I_\eta(\eta). \tag{4.22}$$

Multivariate case. Let $\Theta \subset \mathbb{R}^d$, thus $\theta = (\theta_1, \ldots, \theta_d)'$. Then $\frac{\partial}{\partial \theta} \ln f(x|\theta)$ is the following vector:

$$\frac{\partial}{\partial \theta} \ln f(x|\theta) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \ln f(x|\theta) \\ \frac{\partial}{\partial \theta_2} \ln f(x|\theta) \\ \cdots \\ \frac{\partial}{\partial \theta_d} \ln f(x|\theta) \end{pmatrix}$$

The Fisher information is matrix

$$I(\theta) = E\Big[\big(\frac{\partial}{\partial \theta} \ln f(X|\theta)\big)\big(\frac{\partial}{\partial \theta} \ln f(X|\theta)\big)'|\theta\Big].$$

Thus $I(\theta)$ is $d \times d$-dimensional matrix with i,j-th element being

$$I(\theta)_{ij} = E\Big[\big(\frac{\partial}{\partial \theta_i} \ln f(X|\theta)\big)\big(\frac{\partial}{\partial \theta_j} \ln f(X|\theta)\big)\Big|\theta\Big].$$

The multivariate version of (4.19) is (holds under regularity):

$$I(\theta)_{ij} = -E\big[\frac{\partial^2}{\partial \theta_i \partial_j} \ln f(X|\theta)\big)|\theta\big] \quad \forall i,j. \tag{4.23}$$

Multivariate version of change of variable: Let Θ, Θ_η be open subsets of $\mathbb{R}^d$, let $g : \Theta \to \Theta_\eta$ be a differentiable one-to-one function. Thus $\theta = (\theta_1, \ldots, \theta_d)$ and $g(\theta) = (g_1(\theta), \ldots, g_d(\theta))$. By the chain rule, the multivariate version of (5.9) is

$$\frac{\partial}{\partial \theta} \ln f(x|\theta) = \frac{\partial}{\partial \theta} \ln p(x|g(\theta)) = \begin{pmatrix} \frac{\partial \ln p(x|\eta)}{\partial \eta_1} \frac{\partial g_1(\theta)}{\partial \theta_1} + \cdots + \frac{\partial \ln p(x|\eta)}{\partial \eta_d} \frac{\partial g_d(\theta)}{\partial \theta_1} \\ \frac{\partial \ln p(x|\eta)}{\partial \eta_1} \frac{\partial g_1(\theta)}{\partial \theta_2} + \cdots + \frac{\partial \ln p(x|\eta)}{\partial \eta_d} \frac{\partial g_d(\theta)}{\partial \theta_2} \\ \cdots \\ \frac{\partial \ln p(x|\eta)}{\partial \eta_1} \frac{\partial g_1(\theta)}{\partial \theta_d} + \cdots + \frac{\partial \ln p(x|\eta)}{\partial \eta_d} \frac{\partial g_d(\theta)}{\partial \theta_d} \end{pmatrix}.$$

In matrix notation

$$\frac{\partial}{\partial \theta} \ln f(x|\theta) = \begin{pmatrix} \frac{\partial g_1(\theta)}{\partial \theta_1} & \cdots & \frac{\partial g_d(\theta)}{\partial \theta_1} \\ \cdots & \cdots & \cdots \\ \frac{\partial g_1(\theta)}{\partial \theta_d} & \cdots & \frac{\partial g_d(\theta)}{\partial \theta_d} \end{pmatrix} \begin{pmatrix} \frac{\partial \ln p(x|\eta)}{\partial \eta_1} \\ \cdots \\ \frac{\partial \ln p(x|\eta)}{\partial \eta_d} \end{pmatrix} = J'(\theta)\big(\frac{\partial \ln p(x|\eta)}{\partial \eta}\big),$$

where $J(\theta)$ is the Jacobian matrix, i.e. $J(\theta)_{ij} = \frac{\partial g_i(\theta)}{\partial \theta_j}$. Therefore

$$\Big(\frac{\partial}{\partial\theta}\ln f(x|\theta)\Big)\Big(\frac{\partial}{\partial\theta}\ln f(x|\theta)\Big)' = J'(\theta)\Big(\frac{\partial \ln p(x|\eta)}{\partial\eta}\Big)\Big(\frac{\partial \ln p(x|\eta)}{\partial\eta}\Big)' J(\theta)$$

and so (expectation is linear) and with $\eta = g(\theta)$ we obtain:

$$I(\theta) = J'(\theta) I_\eta(\eta) J(\theta). \tag{4.24}$$

4.3.3 Examples of Fisher information

Binomial distribution. Here $x \in \{0, 1, \dots, n\}$ and

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad \frac{\partial}{\partial\theta}\ln f(x|\theta) = \frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)} = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$
$$\Big(\frac{\partial}{\partial\theta}\ln f(x|\theta)\Big)^2 = \frac{x^2}{\theta^2} + \frac{(n-x)^2}{(1-\theta)^2} - 2\frac{x(n-x)}{\theta(1-\theta)} = \frac{x^2 - 2xn\theta + \theta^2 n^2}{\theta^2(1-\theta)^2}.$$

Since

$$E[X^2|\theta] = n\theta(1-\theta) + n^2\theta^2, \quad E[X|\theta] = n\theta$$

we obtain

$$I(\theta) = \frac{E[X^2|\theta] - 2nE[X|\theta] + \theta^2 n^2}{\theta^2(1-\theta)^2} = \frac{n}{\theta(1-\theta)}.$$

Since

$$\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta) = -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2},$$

it holds

$$E\big[\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)|\theta\big] = -\frac{n}{\theta} - \frac{n}{1-\theta} = -\frac{n}{\theta(1-\theta)} = -I(\theta),$$

so that the formula (4.19) holds.

Negative binomial model. $X \sim \mathrm{NB}(r, \theta)$,

$$f(x|\theta) = \frac{\Gamma(x+r)}{x!\Gamma(r)}(1-\theta)^x\theta^r, \quad \ln f(x|\theta) = \mathrm{const} + x\ln(1-\theta) + r\ln\theta.$$

$$\frac{\partial}{\partial\theta}\ln f(x|\theta) = -\frac{x}{1-\theta} + \frac{r}{\theta}, \quad \frac{\partial^2}{\partial\theta^2}\ln f(x|\theta) = -\frac{x}{(1-\theta)^2} - \frac{r}{\theta^2}.$$

When $X \sim \mathrm{NB}(r, \theta)$, then

$$EX = \frac{r(1-\theta)}{\theta}, \quad EX^2 = r\frac{1-\theta}{\theta^2} + r^2\frac{(1-\theta)^2}{\theta^2},$$

so that according to definition the Fisher information is,

$$I(\theta) = \frac{EX^2}{(1-\theta)^2} - \frac{2rEX}{\theta(1-\theta)} + \frac{r^2}{\theta^2} = \frac{r}{\theta^2(1-\theta)} - \frac{2r^2}{\theta^2} + \frac{2r^2}{\theta^2} = \frac{r}{\theta^2(1-\theta)}.$$

The formula (4.19) holds:

$$-E\Big(\frac{\partial^2}{\partial\theta^2}\ln f(X|\theta)\Big) = \frac{EX}{(1-\theta)^2} + \frac{r}{\theta^2} = \frac{r}{\theta^2(1-\theta)} = I(\theta).$$

Poisson model. $X_1, \ldots, X_n$ are iid with $Po(\lambda)$ distribution. Let us find Fisher information $I(\theta)$ for single observation $X \sim \mathrm{Po}(\lambda)$, since the information for the sample is $nI(\theta)$. When $X \sim \mathrm{Po}(\lambda)$, then $EX = \lambda$, $EX^2 = \lambda + \lambda^2$. Fisher information is $I(\lambda) = \lambda^{-1}$, since

$$\begin{aligned}
&\ln f(x|\lambda) = -\lambda + x\ln\lambda - \ln x! \\
&\frac{\partial}{\partial\lambda}\ln f(x|\lambda) = \frac{x}{\lambda} - 1, \quad E\big(\frac{X}{\lambda} - 1\big)^2 = \frac{1}{\lambda^2}EX^2] - 1 = \frac{1}{\lambda} \\
&\frac{\partial^2}{\partial\lambda^2}\ln f(x|\lambda) = -\frac{x}{\lambda^2}, \quad -\frac{EX}{\lambda^2} = -\frac{1}{\lambda},
\end{aligned}$$

so the formula (4.19) holds.

Exponential family. Recall exponential family in canonical form

$$f(x|\theta) = \exp[\theta u(x) - A(\theta) + B(x)], \quad \theta \in \mathbb{R}.$$

When regular, then according to (4.19) $I(\theta) = A''(\theta)$.

Normal with known variance: The density of $\mathcal{N}(\theta, \sigma^2)$ is in canonical form with $A(\theta) = \frac{\theta^2}{2\sigma^2}$ (recall (3.6)), so that $A''(\theta) = (\sigma^2)^{-1}$ that is independent of θ.

When the exponential family is not canonical form, i.e.

$$f(x|\theta) = \exp[\phi(\theta)u(x) - A(\theta) + B(x)], \quad \phi(\theta) \in \mathbb{R},$$

then one can use canonical parametrization $\eta = \phi(\theta)$ (here ϕ must be monotone), so that

$$p(x|\eta) = \exp[\eta u(x) - A(\phi^{-1}(\eta)) + B(x)],$$

find the information matrix

$$I_\eta(\eta) = \frac{d^2}{d\eta^2}A(\phi^{-1}(\eta))$$

and then use the change of variable formula for Fisher information (4.22):

$$I(\theta) = (\phi'(\theta))^2 I_\eta(\phi(\theta)).$$

Binomial distribution belongs to exponential family

$$\ln f(x|\theta) = x \ln \frac{\theta}{1-\theta} + n \ln(1-\theta) + \ln \binom{n}{x} = x\eta - n \ln(1+e^\eta) + \ln \binom{n}{x}, \quad \eta = \ln \frac{\theta}{1-\theta} =: \phi(\theta).$$

Since

$$\big(\ln(1+e^\eta)\big)'' = \frac{e^\eta}{1+e^\eta} - \big(\frac{e^\eta}{1+e^\eta}\big)^2,$$

we obtain

$$I_\eta(\eta) = n\Big(\frac{e^\eta}{1+e^\eta} - \big(\frac{e^\eta}{1+e^\eta}\big)^2\Big).$$

To get $I(\theta)$ use the change of variable formula for information (4.22):

$$I(\theta) = (\phi'(\theta))^2 I_\eta(\phi(\theta)) = \frac{1}{\theta^2(1-\theta)^2} n(\theta - \theta^2) = \frac{n}{\theta(1-\theta)}.$$

Poisson distribution belongs to exponential family

$$f(x|\theta) = \exp[x \ln \theta - \theta - \ln(x!)] = \exp[x\eta + e^\eta - \ln x!], \quad \eta = \ln \theta =: \phi(\theta).$$

Therefore $I_\eta(\eta) = (e^\eta)'' = e^\eta$ and by (4.22)

$$I(\theta) = \frac{1}{\theta^2} I_\eta(\ln \theta) = \frac{1}{\theta}.$$

Exponential distribution belongs to exponential family. Information (Exercise 5)

$$I(\theta) = \frac{1}{\theta^2}. \tag{4.25}$$

Normal model with μ and σ^2 unknown. Thus $X_1, \dots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$-distributed random variables, $\theta = (\mu, \sigma^2)$. The model is regular, so to find the information matrix, the formula (4.23) can be used. Since $(x = (x_1, \dots, x_n))$.

$$\ln f(x|\mu, \sigma^2) = \text{const} - \frac{n}{2} \ln \sigma^2 - \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{2\sigma^2} - \frac{n(\bar{x} - \mu)^2}{2\sigma^2},$$

the derivatives are

$$\begin{aligned}
\frac{\partial^2}{\partial \mu^2} \ln f(x|\mu, \sigma^2) &= -\frac{n}{\sigma^2} \\
\frac{\partial^2}{\partial \mu \partial \sigma^2} \ln f(x|\mu, \sigma^2) &= -\frac{n(\bar{x} - \mu)}{\sigma^4} \\
\frac{\partial^2}{\partial \sigma^2 \partial \mu} \ln f(x|\mu, \sigma^2) &= -\frac{n(\bar{x} - \mu)}{\sigma^4} \\
\frac{\partial^2}{\partial (\sigma^2)^2} \ln f(x|\mu, \sigma^2) &= \frac{n}{2(\sigma^2)^2} - \frac{\sum_{i=1}^n (x_i - \mu)^2}{(\sigma^2)^3}.
\end{aligned}$$

Taking expectations (and multiplying with -1), we obtain the information matrix:

$$I_n(\theta) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} = n \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} = nI(\theta). \tag{4.26}$$

4.3.4 Jeffreys prior

Univariate case. Given the model $f(x|\theta)$, where $\theta \in \Theta \subset \mathbb{R}$ (open subset), *Jeffreys prior* denoted by $\pi^J(\theta)$ is proportional to square root of Fisher information

$$\pi^J(\theta) \propto \sqrt{I(\theta)}.$$

When $\int \sqrt{I(\theta)}d\theta < \infty$, then

$$\pi^J(\theta) = \frac{\sqrt{I(\theta)}}{\int \sqrt{I(\theta)}d\theta} < \infty,$$

otherwise Jeffreys prior is improper.

Jeffreys prior is invariant with respect to reparametrization – when $\theta \sim \pi^J(\theta)$, then $\eta := g(\theta) \sim \pi^J(\eta)$. Indeed, when $\theta \sim \pi^J(\theta)$, then $g(\theta) \sim \pi^*(\eta)$, where by the change of variable (4.18)

$$\pi^*(\eta) = \frac{\pi^J(\theta)}{|g'(\theta)|} \propto \frac{\sqrt{I(\theta)}}{|g'(\theta)|} = \sqrt{I_\eta(\eta)},$$

where the last equality follows from (4.22).

Multivariate case. Given the model $f(x|\theta)$, where $\theta \in \Theta \subset \mathbb{R}^d$ (open subset), *Jeffreys prior* denoted by $\pi^J(\theta)$ is proportional to square root of the determinant of Fisher information:

$$\pi^J(\theta) \propto \sqrt{|I(\theta)|}.$$

When $\theta \sim \pi^J(\theta)$, then $g(\theta) \sim \pi^*(\eta)$, where by (4.24)

$$\pi^*(\eta) = \pi^J(h(\eta))|J(\eta)| = \pi^J(\theta)|J(\eta)| \propto \sqrt{|I(\theta)|}|J(\eta)| = \sqrt{|J'(\theta)I_\eta(\eta)J(\theta)|}|J(\eta)|.$$

Since (recall $|J(\eta)||J(\theta)| = 1$)

$$|J'(\theta)I_\eta(\eta)J(\theta)| = |J'(\theta)||I_\eta(\eta)||J(\theta)| = |J(\theta)|^2|I_\eta(\eta)| = |J(\eta)|^{-2}|I_\eta(\eta)|,$$

we obtain $\sqrt{|J'(\theta)I_\eta(\eta)J(\theta)|}|J(\eta)| = \sqrt{|I_\eta(\eta)|}$.

The invariance with respect to the one-to-one transformations holds also for posterior: Let $\pi(\theta|x) \propto \pi^J(\theta)f(x|\theta)$ be posterior with respect to the Jeffreys prior, and let g be one-to-one (univariate) transformation. Then after change of variable, the posterior has density (recall (4.18))

$$\pi^*(\eta|x) = \frac{\pi(\theta|x)}{|g'(\theta)|} \propto \frac{f(x|\theta)\pi^J(\theta)}{|g'(\theta)|} = f(x|\theta)\pi^J(\eta) = p(x|\eta)\pi^J(\eta).$$

4.3.5 Examples of Jeffreys priors

Location parameter. Let $f(x|\theta) = f(x-\theta)$, where $f(x) > 0$ for every x and f is differentiable probability density. The parameter θ is called *location parameter*, because when $X \sim f$, then $f(\cdot|\theta)$ is the density of $X+\theta$. Therefore a noninformative prior should be location invariant as well, i.e. for every $c \in \mathbb{R}$ the random variables $c+\theta$ and θ should have the same distribution, i.e. $\pi(\theta - c) = \pi(\theta)$ implying that $\pi(\theta) = \text{const}$. It turns out that this is also Jeffreys prior, because $I(\theta) = \text{const}$. (Exercise 4). Thus $\pi^J \propto 1$ (uniform).
For example, when θ is the mean of Gaussian model, then Jeffreys prior is improper. In other words, when $X_1, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$ and σ^2 is known (fixed), then Jeffreys prior for μ is constant (uniform). However, as showed in Section 4.2, the posterior $\pi(\mu|x)$ (here $x = (x_1, \ldots, x_n)$) is (recall (4.5))

$$\theta|x \sim \mathcal{N}(\bar{x}, \frac{\sigma^2}{n}),$$

so posterior is proper and, as pointed out in Section 4.2, it is the limit of posterior of conjugate prior (normal) when the variance tends to infinity: $\tau^2 \to \infty$.

Scale parameter. Let $f(x|\theta) = \frac{1}{\theta} f(\frac{x}{\theta})$, where f is differentiable probability density. The parameter $\theta > 0$ is called *scale parameter*. When $X \sim f$, then $f(\cdot|\theta)$ is the density of θX. Therefore a noninformative prior should be invariant with respect to the multiplication with positive scalar, i.e. $c\theta$ and θ should have the same distribution for every $c > 0$: $\pi(\theta) = \frac{1}{c}\pi(\frac{\theta}{c})$. This holds only if $\pi(\theta) \propto \frac{1}{\theta}$. It turns out that this is also Jeffreys prior, because $I(\theta) = \frac{a}{\theta^2}$, where $a > 0$ is a constant that is independent of θ (Exercise 4). Thus $\pi^J(\theta) \propto \frac{1}{\theta}$.
For example, when $\theta = \sigma$ is the standard deviation of Gaussian model, then Jeffreys prior is proportional to $1/\sigma$ and this is again improper. In other words, when $X_1, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, the Jeffreys prior for σ is proportional to $\frac{1}{\sigma}$ (improper). After change of variable, we see that Jeffreys prior for $\eta = \sigma^k$, $k = 2, 3, \ldots$ must be proportional $\frac{1}{\eta}$ (Exercise 4). Thus Jeffreys prior for variance $\theta = \sigma^2$ is proportional to $\frac{1}{\theta}$. In Section 4.2, it was shown that this improper prior can be interpreted as ScaleInv$-\chi^2(0, \tau^2)$ since leads to the proper posterior ScaleInv$-\chi^2(\nu_n, \tau_n^2)$, where $\nu_n = n$ and $\tau_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2$.

Normal model with μ and σ^2 unknown. Thus $X_1, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$-distributed random variables, $\theta = (\mu, \sigma^2)$. We already have found the information matrix (4.26), so the determinant and Jeffreys prior are as follows:

$$I(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}, \quad |I(\theta)| = \frac{1}{2\sigma^6}, \quad \pi^J(\theta) \propto \frac{1}{\sigma^6} = \frac{1}{(\sigma^2)^{\frac{3}{2}}}.$$

This is improper prior with proper posterior, because

$$\pi^J(\theta)f(x|\theta) \propto (\sigma^2)^{-\frac{n}{2}-\frac{3}{2}} \exp[-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}]$$
$$= (\sigma^2)^{-\left(\frac{(n+1)}{2}+1\right)} \exp[-\frac{\sum_{i=1}^n (x_i-\bar{x})^2 + n(\bar{x}-\mu)}{2\sigma^2}]$$

so that posterior distribution is NIX distribution with parameters

$$\nu_n = n, \quad \mu_n = \bar{x}, \quad \tau_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2, \quad \kappa_n = n$$

Thus the posterior:

$$\sigma^2|x \sim \text{ScaleInv}-\chi^2(\nu_n, \tau_n^2)$$
$$\mu|\sigma^2, x \sim \mathcal{N}(\mu_n, \frac{\sigma^2}{\kappa_n}).$$

Recall the posterior updating formulas for NIX prior:

$$\nu_n = \nu + n$$
$$\mu_n = \frac{\kappa}{\kappa+n}\mu_o + \frac{n}{\kappa+n}\bar{x}$$
$$\kappa_n = \kappa + n$$
$$\nu_n\tau_n^2 = \nu\tau^2 + \sum_{i=1}^n (x_i-\bar{x})^2 + \frac{\kappa n}{\kappa+n}(\bar{x}-\mu_o)^2.$$

This justifies calling Jefferys prior as NIX prior with $\nu = \kappa = 0$.
Recall that with $\pi(\theta) \propto (\sigma^2)^{-1}$ (Section 4.2), the posterior os NIX with

$$\nu_n = n-1, \quad \tau_n^2 = \frac{\sum_i (x_i-\bar{x})^2}{n-1}, \quad \mu_n = \bar{x}, \quad \kappa_n = n.$$

Poisson model. $X_1, \ldots, X_n$ are iid with $Po(\lambda)$ distribution. We already know that the Fisher information is $I(\theta) = \frac{1}{\lambda}$, so that Jeffreys prior $\pi^J(\lambda) \propto \frac{1}{\sqrt{\lambda}}$ (improper).
We also know already that under the parametrization $\eta = g(\lambda) = \ln(\lambda)$ (natural parametrization for exponential model) $I_\eta(\eta) = e^\eta$. Therefore corresponding Jeffreys prior is $\pi^{*J}(\eta) \propto e^{\frac{\eta}{2}}$ and after the change of variable formula (4.18)

$$\pi^J(\lambda) = \pi^{*J}(g(\lambda))|g'(\lambda)| = \frac{\pi^{*J}(\ln\lambda)}{\lambda} \propto \frac{\sqrt{\lambda}}{\lambda} = \frac{1}{\sqrt{\lambda}}.$$

Sometimes Jeffreys prior for Poisson model is interpreted as Gamma$(1/2, 0)$, because $\theta^{1/2-1}e^{-0\theta} = \theta^{-1/2}$. Posterior is proper (Exercise 5)):

$$\theta|x_1, \ldots, x_n \sim \text{Gamma}\Big(\sum_{i=1}^n x_i + 1/2, n\Big) \tag{4.27}$$

an analogue with (2.7) that justifies Gamma$(1/2, 0)$-notation.

Binomial model. $X \sim \mathrm{B}(n, \theta)$. We already know that

$$I(\theta) = \frac{n}{\theta(1-\theta)},$$

so that

$$\pi^J(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}},$$

implying that Jeffreys prior is Beta$(\frac{1}{2}, \frac{1}{2})$ distribution.

Exponential model. $X_1, \ldots, X_n$ are iid with Exp(θ) distribution.

$$I(\theta) = \frac{1}{\theta^2}$$

and so Jeffreys prior $\pi^J(\theta) \propto \frac{1}{\theta}$ (improper), sometimes denoted as Gamma$(0, 0)$. Posterior (Exercise 5):

$$\theta | x_1, \ldots, x_n \sim \mathrm{Gamma}(n, \sum_{i=1}^{n} x_i). \tag{4.28}$$

Again analogue with (2.12) justifying the Gamma$(0, 0)$-notation.

Negative binomial model. $X \sim \mathrm{NB}(r, \theta)$, We already know that

$$I(\theta) = \frac{r}{\theta^2(1-\theta)}$$

and so the Jeffreys prior $\pi^J(\theta) \propto \frac{1}{\theta\sqrt{(1-\theta)}}$ (improper).

Posterior (Exercise 5)

$$\theta | x_1, \ldots, x_n \sim \mathrm{Beta}(nr, \sum_{i=1}^{n} x_i + \frac{1}{2}). \tag{4.29}$$

Multinomial model. $X \sim \mathrm{Multinom}(n, \theta_1, \ldots, \theta_k)$. Jeffreys prior is Dir$(1/2, \ldots, 1/2)$ (Exercise 6).

4.4 Exercises

1. Work out the formulas in subsection 4.2.3: prove (4.12), (4.13), (4.14), (4.15), (4.16), (4.17).
2. Let $\theta \sim \pi(\theta)$, where π is univariate density with respect to the Lebesgue maesure. Find the density of linear combination $\eta = a\theta + b$.
3. Prove (4.19), (4.20).

4.
- Let $f(x|\theta) = f(x-\theta)$, where $f(x)$ is a differentiable probability density. Prove that $I(\theta)$ =const.
- Let $f(x|\theta) = \frac{1}{\theta} f(\frac{x}{\theta})$, where $f(x)$ is a differentiable probability density. Prove that $I(\theta) \propto \theta^{-2}$.
- Let $\pi^J(\theta) \propto 1/\theta$. Show that for any $k = 2, 3, \ldots$ $\pi^J(\eta) \propto 1/\eta$, where $\eta = \theta^k$.

5. Prove (4.25), (4.27), (4.28), (4.29).

6. Jeffreys prior for multinomial. Take $n = 1$ and find

$$\mathrm{Cov}\Big(\frac{\partial}{\partial\theta} \ln f(x|\theta)\Big) = \mathrm{diag}(\frac{1}{\theta_1}, \ldots, \frac{1}{\theta_k}) - 1_k =: J,$$

where 1_k is a matrix where every entry is 1. Then show that for $n > 1$,

$$I_n(\theta) = nJ + n^2 1_k = \mathrm{diag}(\frac{n}{\theta_1}, \ldots, \frac{n}{\theta_k}) + (n^2 - n)1_k.$$

Then use matrix determinant lemma to show that

$$|I_n(\theta)| = c \cdot \prod_{i=1}^{k} \frac{1}{p_i}.$$

5 Testing hypotheses

5.1 Bayes factor

Suppose we have two alternative Bayesian models to explain observations x: the 0-model $\{f_0(\cdot|\theta)\}$ with prior π_0 and 1-model $\{f_1(\cdot|\theta)\}$ with prior π_1. These priors might have densities with respect to the different reference measures, so the integration with respect to π_i will be denoted by $\pi_i(d\theta)$. The Bayes factor is ratio of marginal densities (marginal likelihoods):

$$B_{01}(x) := \frac{f_0(x)}{f_1(x)}, \quad f_i(x) := \int f_i(x|\theta)\pi_i(d\theta), \quad i = 0, 1. \tag{5.1}$$

Hence Bayes factor measures how much the 0-model describes the observations better (or worse) than the alternative 1-model. Observe that when $f_0(\cdot|\theta) = f_1(\cdot|\theta) = f(\cdot|\theta)$, but $\pi_0 = \delta_{\theta_0}$ and $\pi_1 = \delta_{\theta_1}$ then Bayes factor is just the *likelihood ratio:*

$$B_{01}(x) = \frac{f(x|\theta_0)}{f(x|\theta_1)}.$$

Example (binomial): Suppose that x is the number of successes out of n trials. Assume that under 0-model, the observation come from binomial distribution with parameter 0.5, under 1-model the parameter has uniform distribution: $\pi_1(\theta) = 1$. In this example, $f_0(x|\theta) = f_1(x|\theta)$ (both are binomial), but $\pi_0 = \delta_{0.5}$ and π_1 is uniform. Hence f_0 is the density (probability) of binomial $\mathrm{B}(n, 0.5)$ distribution and f_1 is the density of Beta-binomial distribution $\mathrm{BetaBin}(n, 1, 1)$ and we know that this distribution is uniform. Hence

$$B_{01}(x) = \frac{f_0(x)}{f_1(x)} = \frac{\binom{n}{x}(0.5)^n}{(n+1)^{-1}} = (n+1)\binom{n}{x}(0.5)^n.$$

Some numerical examples:

1. When $n = 200$ and $x = 115$, $(x/n = 0.575)$ then $f_0(115) \approx 0.006$ and $f_1(x) = 1/201 \approx 0.005$, so that $B_{01} \approx 1.2$ – slightly supporting 0-model.

2. When $n = 98451$ and $x = 49581$, $(x/n \approx 0.5036)$,

$$B_{01}(x) \approx \frac{1.95 \times 10^{-4}}{1.02 \times 10^{-5}} \approx 19.$$

   Thus there are very strong evidences favoring 0-model.

5.2 Hypotheses in Bayesian setting

Assume $\Theta = \Theta_0 \cup \Theta_1$, where $\Theta_0 \cap \Theta_1 = \emptyset$. Suppose we have model $\{f(\cdot|\theta)\}$ with prior density $\pi(\theta)$. We would like to test hypotheses

$$H_0 : \ \theta \in \Theta_0$$
$$H_1 : \ \theta \in \Theta_1$$

In Bayesian approach it is natural to calculate the posterior probability

$$P(\Theta_i|x) = \int_{\Theta_i} \pi(\theta|x)d\theta = \int_{\Theta_i} \frac{f(x|\theta)\pi(\theta)d\theta}{f(x)}, \quad i = 0, 1$$

and compare these probabilities. The following decision theoretic argument justifies that comparison.

Loss and risk. Let $\phi(x) \in \{0, 1\}$ be a test so that when $\phi(x) = 0$, then H_0 is accepted and when $\phi(x) = 1$, H_1 is accepted. It means the existence of set A such that $\phi = I_A$, meaning that H_0 is rejected whenever $x \in A$. Let L be the following *loss function* (here $\phi \in \{0, 1\}$)

$$L(\theta, \phi) = \begin{cases} 0, & \text{if } \phi = I_{\Theta_1}(\theta); \\ a_0, & \text{if } \theta \in \Theta_0,\ \phi = 1; \\ a_1, & \text{if } \theta \in \Theta_1,\ \phi = 0. \end{cases} \tag{5.2}$$

Hence the loss is a_0, when $\theta \in \Theta_0$, but $\phi(x) = 1$ (first type error); the loss is a_1, when $\theta \in \Theta_1$, but $\phi(x) = 0$ (second type error); otherwise the loss is zero. Given x, the test $\phi(x)$ is fixed (either 0 or 1), let us find the test that minimizes the expected loss when expectation is taken with respect to posterior measure $\pi(\cdot|x)$:

$$R(\phi|x) := \int L(\theta, \phi)\pi(\theta|x)d\theta.$$

It is easy to see that the optimal (best) test that minimizes $R(\phi|x)$ over $0, 1$ is the following (Exercise 1):

$$\phi(x) = \arg\min_{0,1} R(\phi|x) = \begin{cases} 0, & \text{if } P(\Theta_0|x) \geq \frac{a_1}{a_1+a_0}; \\ 1, & \text{else.} \end{cases} \tag{5.3}$$

The ratio $\frac{a_0}{a_0+a_1}$ is sometimes known as *acceptance level*. How $\phi(x)$ is defined when $P(\Theta_0|x) = \frac{a_1}{a_1+a_0}$ does not matter, because the risk is the same, but as it is common in statistics, the ties are broken in favor of H_0. Also observe that only the ratio $\frac{a_1}{a_1+a_0}$, not the constants a_i, matter. Hence for *symmetric loss* $a_0 = a_1$, w.l.o.g. we take $a_0 = a_1 = 1$. Observe that the rule (5.3) is equivalent to (Exercise 1)

$$\phi(x) = 0 \quad \Leftrightarrow \quad a_0 P(\Theta_0|x) \geq a_1 P(\Theta_1|x) \tag{5.4}$$

Hence for symmetric loss ($a_1 = a_0$) the acceptance level is 0.5 and the decision is H_0, i.e. $\phi(x) = 0$, when

$$P(\Theta_0|x) \geq 0.5 \quad \Leftrightarrow \quad P(\Theta_0|x) \geq P(\Theta_1|x).$$

It is easy to see that for symmetric loss $a_0 = a_1 = 1$, the loss function (5.2) can be written as L_1-loss

$$L(\theta, \phi) = |I_{\Theta_1}(\theta) - \phi| \tag{5.5}$$

and the solution is the same if ϕ is minimized over the interval $[0, 1]$ instead of $\{0, 1\}$ (Exercise 2):

$$\arg\min_{0,1} R(\phi|x) = \arg\min_{[0,1]} R(\phi|x) \tag{5.6}$$

Hypotheses testing as the comparison of the models. The hypotheses testing can be considered as the comparison of two models: $f_0(\cdot|\theta) = f_1(\cdot|\theta) = f(\cdot|\theta)$, $\forall\theta$, but the prior measures π_i have densities proportional to $I_{\Theta_i}(\theta)\pi(\theta)$, i.e.

$$\pi_i(\theta) = I_{\Theta_i}(\theta)\frac{\pi(\theta)}{q_i}, \quad q_i := \int_{\Theta_i} \pi(\theta)d\theta = \pi(\Theta_i), \quad i = 0, 1.$$

Thus the prior is mixture of two (model) priors and q_i are prior model probabilities:

$$\pi(\theta) = q_0\pi_0(\theta) + q_1\pi_1(\theta). \tag{5.7}$$

Hence

$$f_i(x) = \int f(x|\theta)\pi_i(d\theta) = \frac{\int_{\Theta_i} f(x|\theta)\pi(d\theta)}{q_i} = P(\Theta_i|x)\frac{f(x)}{q_i}, \quad i = 0, 1.$$

Therefore the Bayes factor is

$$B_{01}(x) = \frac{f_0(x)}{f_1(x)} = \frac{P(\Theta_0|x)}{P(\Theta_1|x)}\frac{q_1}{q_0} = \frac{P(\Theta_0|x)}{P(\Theta_1|x)}\frac{\pi(\Theta_1)}{\pi(\Theta_0)}. \tag{5.8}$$

We now see that in terms of Bayes factor (5.4) is

$$\phi(x) = 0 \quad \Leftrightarrow \quad B_{01}(x) \geq \frac{a_1 q_1}{a_0 q_0}. \tag{5.9}$$

Hence the prior probabilities and losses give a meaningful threshold.

Example (normal mean, one sided). Consider the normal model with known variance and conjugate prior:

$$\theta \sim \mathcal{N}(\mu, \tau^2)$$
$$X_1, \ldots, X_n|\theta \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2).$$

We know the posterior is normal (subsection 2.5.1):

$$\theta|x \sim \mathcal{N}(\mu_n, \tau_n^2),$$

with

$$\mu_n = \frac{\mu\sigma^2 + n\tau^2\bar{x}}{n\tau^2 + \sigma^2} = \frac{\frac{\mu}{\tau^2} + \frac{n}{\sigma^2}\bar{x}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \qquad \tau_n^2 = (\frac{n}{\sigma^2} + \frac{1}{\tau^2})^{-1} = \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}. \tag{5.10}$$

Observe that τ_n^2 is independent of observations x, but μ_n depends on x through the sample mean, hence we can write $\mu_n(\bar{x})$.
We aim to test

$$H_0 : \ \theta < 0$$
$$H_1 : \ \theta \geq 0$$

Thus $\Theta_0 = (-\infty, 0)$ and $\Theta_1 = [0, \infty)$. Calculate ($x = (x_1, \ldots, x_n)$)

$$P\big(\theta < 0|x\big) = P\big(\frac{\theta - \mu_n}{\tau_n} < -\frac{\mu_n}{\tau_n}|x\big) = P(Z < -\frac{\mu_n}{\tau_n}), \quad Z \sim \mathcal{N}(0, 1).$$

With Φ being the distribution function of standard normal, thus

$$P\big(\theta < 0|x\big) = \Phi\big(-\frac{\mu_n}{\tau_n}\big). \tag{5.11}$$

Therefore, by (5.3), H_0 is accepted when

$$\Phi\big(-\frac{\mu_n}{\tau_n}\big) \geq \frac{a_1}{a_1+a_0}$$

With k_{a_0,a_1} being $a_1/(a_1+a_0)$-quantile, i.e. $\Phi(z_{a_0,a_1}) = a_1/(a_1+a_0)$, the rule (5.3) is: H_0 is accepted when

$$-\mu_n(\bar{x}) > k_{a_0,a_1}\tau_n.$$

More general setting. We generalize the argument above allowing the conditional densities $f_i(\cdot|\theta)$ being different and also π_i can be different probability measures (prior distributions) on Θ_i, respectively $i = 0, 1$. With q_i being prior model probabilities, the overall prior measure on $\Theta_0 \cup \Theta_1$ is now mixture $\pi = q_0\pi_0 + q_1\pi_1$. Thus we have the following hierarchial model

$$\begin{aligned} Z &\sim \mathrm{B}(1, q_1) \\ \theta|Z &\sim \pi_Z \\ X|\theta, Z &\sim f_Z(\cdot|\theta) \end{aligned} \tag{5.12}$$

When both priors π have a density with respect to same measure, denoted as $\pi_i(\theta)$, then π has also density that is a mixture of model densities as in (5.7). However, it might be that the two priors have densities with respect to different measures. Typically it is the case when testing point-null hypotheses $H_0 : \theta = \theta_0$, i.e. $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \Theta \setminus \{\theta_0\}$, where $\Theta \subset \mathbb{R}$. Then $\pi_0 = \delta_{\theta_0}$ has density with respect to counting measure and π_i might have density with respect to Lebesgue measure. This was exactly the case in the binomial example above. Observe that the posterior is still well defined:

$$P(A|x) = P_0(A \cap \Theta_0|x)q_0(x) + P_1(A \cap \Theta_1|x)q_1(x), \tag{5.13}$$

where $P_i(\cdot|x)$ is posterior under the model i, i.e. $P_i(\cdot|x)$ has density (on Θ_i) proportional to $\pi_i(\theta)f_i(x|\theta)$ and

$$q_i(x) := \frac{q_i f_i(x)}{q_0 f_0(x) + q_1 f_1(x)}, \quad f_i(x) := \int_{\Theta_i} f_i(x|\theta)\pi_i(\theta)d\theta, \quad i = 0, 1.$$

Taking $A = \Theta_i$ in (5.13), we see that

$$q_i(x) = P(\Theta_i|x), \quad i = 0, 1$$

and in terms of (5.12), we see that $q_i(x) = P(Z = i|X = x)$ (Exercise 3). With loss $L(\theta, \phi)$ the risk function is

$$R(\phi|x) = q_0(x)\int_{\Theta_0} L(\theta, \phi)P_0(d\theta|x) + q_1(x)\int_{\Theta_1} L(\theta, \phi)P_1(d\theta|x)$$

and when $L(\theta, \phi)$ as in (5.2), it is easy to see that (5.4) and (5.3) still holds (Exercise 3), i.e.

$$\phi(x) = 0 \quad \Leftrightarrow \quad q_0(x) \geq \frac{a_1}{a_1 + a_0}. \tag{5.14}$$

Hence the optimal decision is the model 0 when the posterior probability of 0-model is sufficiently high (determined by losses). In terms of Bayes factor $B_{01}(x)$, the posterior probability $q_0(x)$ is (Exercise 3)

$$q_0(x) = \Big(1 + \frac{q_1}{q_0 B_{01}(x)}\Big)^{-1}. \tag{5.15}$$

In terms of Bayes factor, the rule (5.14) is (5.9) (Exercise 3).

Observe that with $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, $\pi_i = \delta_{\theta_i}$, we have $f_i(x) = f_i(x|\theta_i)$ (feature densities) and so

$$q_i(x) = \frac{q_i f_i(x|\theta_i)}{q_0 f_0(x|\theta_0) + q_1 f_1(x|\theta_1)}.$$

Then (5.14) is known as *Bayes rule* in classification.

Example (binomial). Let us continue the binomial example above.

$$H_0 : \ \theta = 0.5$$
$$H_1 : \ \theta \neq 0.5$$

The posterior probability of 0-model is

$$q_0(x) = \frac{q_0 f_0(x)}{q_0 f_0(x) + q_1 f_1(x)} = \frac{q_0 \binom{n}{x} (0.5)^n}{q_0 \binom{n}{x} (0.5)^n + q_1 (n+1)^{-1}} = \Big(1 + \frac{q_1}{q_0} 2^n \frac{x!(n-x)!}{(n+1)!}\Big)^{-1}.$$

1. When $n = 200$, $x = 115$, $q_0 = 0.5$, from (5.15) we obtain (recall $B_{01}(115) = 1.2$)

   $$q_0(115) = \big(1 + \frac{1}{1.2}\big)^{-1} \approx 0.54.$$

   With $a_1 = a_2$, by the rule (5.14) we decide for $H_0 : \theta = 0.5$.

   The probability that $\mathrm{B}(200, 0.5)$-distributed random variable X takes values at least 115 is 0.02002, hence $P(X \geq 115) + P(X \leq 85) = 0.04$ so that frequentist test would reject the hypotheses $H_0 : \ \theta = 0.5$ at level 0.05.

2. When $n = 98451$, $x = 49581$, $q_0 = 0.5$, with $B_{01}(49581) = 19$, we obtain

   $$q_0(49581) = \big(1 + \frac{1}{19}\big)^{-1} = 0.95$$

So the posterior probability for H_0 is (approximatively) 0.95.

The probability that $\mathrm{B}(98451, 0.5)$-distributed random variable takes values at least 49851 is approximatively 0.01183, hence the frequentist test would reject the hypotheses $H_0 : \theta = 0.5$ at level 0.05 ($2 \times 0.01183 < 0.05$).

We see that Bayesian and frequentist approach give different solutions and as it is evident from the last example, the difference can be remarkable big.

Example (normal mean, two-sided). Let $f(\cdot|\theta)$ be the normal density with known variance σ^2. The parameter θ is the mean and we test the hypotheses:

$$\begin{aligned} H_0 &: \ \theta = 0 \\ H_1 &: \ \theta \neq 0 \end{aligned}$$

We take π_1 as $\mathcal{N}(0, \tau^2)$ (conjugate) and $\pi_0 = \delta_0$. Hence the model is

$$\begin{aligned} Z &\sim \mathrm{B}(1, q_1) \\ \theta|Z = 0 &\sim \delta_0 \quad (\theta = 0) \\ \theta|Z = 1 &\sim \mathcal{N}(0, \tau^2) \\ X_1, \ldots, X_n|\theta, Z &\overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2) \end{aligned}$$

We know that $f_1(x_1, \ldots, x_n)$ is the density of jointly normally distributed random variables with mean vector zero and covariance matrix as in (2.14):

$$\Sigma = \begin{pmatrix} \sigma^2+\tau^2 & \tau^2 & \cdots & \tau^2 \\ \tau^2 & \sigma^2+\tau^2 & \cdots & \tau^2 \\ \cdots & \cdots & \cdots & \cdots \\ \tau^2 & \tau^2 & \cdots & \sigma^2+\tau^2 \end{pmatrix},$$

$$\Sigma^{-1} = \frac{1}{\sigma^2} \begin{pmatrix} 1-\frac{1}{\sigma^2/\tau^2+n} & -\frac{1}{\sigma^2/\tau^2+n} & \cdots & -\frac{1}{\sigma^2/\tau^2+n} \\ -\frac{1}{\sigma^2/\tau^2+n} & 1-\frac{1}{\sigma^2/\tau^2+n} & \cdots & -\frac{1}{\sigma^2/\tau^2+n} \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{1}{\sigma^2/\tau^2+n} & -\frac{1}{\sigma^2/\tau^2+n} & \cdots & 1-\frac{1}{\sigma^2/\tau^2+n} \end{pmatrix}, \quad |\Sigma| = \big(1+\frac{n}{\sigma^2/\tau^2}\big)(\sigma^2)^n.$$

Thus

$$\begin{aligned} f_1(x_1, \ldots, x_n) &= (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp[-\frac{1}{2} x' \Sigma^{-1} x] \\ &= (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \big(1+\frac{n}{\sigma^2/\tau^2}\big)^{-\frac{1}{2}} \exp\Big[-\frac{1}{2\sigma^2}\big(\sum_{i=1}^n x_i^2 - \frac{1}{\sigma^2/\tau^2+n}(\sum_i x_i)^2\big)\Big]. \end{aligned}$$

Clearly

$$f_0(x_1, \ldots, x_n) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp[-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}].$$

Hence with $x = (x_1, \ldots, x_n)$

$$\begin{aligned} B_{01}(x) = \frac{f_0(x)}{f_1(x)} &= \big(1 + \frac{n}{\sigma^2/\tau^2}\big)^{\frac{1}{2}} \exp\Big[- \frac{1}{2\sigma^2(\sigma^2/\tau^2 + n)} (\sum_i x_i)^2 \Big] \\ &= \sqrt{\frac{n + \sigma^2/\tau^2}{\sigma^2/\tau^2}} \exp\Big[- \frac{1}{2} \frac{n}{(\sigma^2/\tau^2 + n)} \big(\frac{\bar{x}}{\sigma/\sqrt{n}}\big)^2 \Big] \\ &= \sqrt{\frac{\tau^2 n + \sigma^2}{\sigma^2}} \exp\Big[- \frac{1}{2} \frac{\tau^2 n}{(\sigma^2 + \tau^2 n)} \big(\frac{\bar{x}}{\sigma/\sqrt{n}}\big)^2 \Big]. \end{aligned} \tag{5.16}$$

Plugging $B_{01}(x)$ into (5.15), we obtain $q_0(x)$. Observe that $B_{01}(x)$ is a function of *z-score* $z := \frac{\bar{x}}{\sigma/\sqrt{n}}$ that in frequentist statistics is used in calculating p-value for testing the hypotheses $H_0 : \theta = 0$.

Also observe:

- When n increases and z remains constant, then $B_{01}(x) \to \infty$, hence $q_0(x) \to 1$ whatever the z-score is. Bayesian choice would be H_0 (for very big n).
- When τ^2 increases (everything else is remains constant), then $B_{01}(x) \to \infty$, hence $q_0(x) \to 1$ whatever the z-score is. Bayesian choice would be H_0 – flat prior do not work!
- When $\tau^2 \to 0$ (everything else is remains constant), then $B_{01}(x) \to 1$ and $q_0(x) \to q_0$ (prior probability).

In Table 5.2.2 in [6], the values for $q_0(x)$ for different z-scores are found $n = 1, \sigma^2 = \tau^2$, $q_0 = 0.5$. For $z = 1.96$ (p-value 0.05) $q_0(x) = 0.351$, for $z = 1.28$ (p-value 0.2), $q_0(x) = 0.484$. In Table 5.2.3 the same is done for $\tau^2 = 10\sigma^2$. The probabilities are now bigger: for $z = 1.96$ (p-value 0.05) $q_0(x) = 0.366$, for $z = 1.28$ (p-value 0.2), $q_0(x) = 0.612$.

Improper priors. Consider the case of testing point-null hypotheses $H_0 : \theta = \theta_0$ with $\pi_0 = \delta_{\theta_0}$, but π_1 is improper with $f_1(x) = \int_{\Theta_1} f_1(x|\theta)\pi_1(d\theta) < \infty$. Then $q_i(x)$ can be calculated for $i = 0, 1$, but, unfortunately, it is not invariant with respect to scaling and might be biased. Indeed, multiplying an improper prior π by a constant $c > 0$ would not normally change the posterior, because

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(d\theta)}{\int f(x|\theta) d\pi(\theta)} = \frac{c f(x|\theta)\pi(d\theta)}{c \int f(x|\theta)\pi(d\theta)} = \frac{f(x|\theta) c\pi(d\theta)}{\int f(x|\theta) c\pi(d\theta)},$$

but it might change $q_i(x)$ in hypotheses testing, because with $f_0(x) = f_0(x|\theta_0)$,

$$q_0(x) = \frac{q_0 f_0(x)}{q_0 f_0(x) + q_1 f_1(x)} = \frac{q_0 f_0(x)}{q_0 f_0(x) + q_1 \int f_1(x|\theta)\pi_1(d\theta)} \neq \frac{q_0 f_0(x)}{q_0 f_0(x) + c q_1 \int f_1(x|\theta)\pi_1(d\theta)}.$$

When c increases, then $q_0(x)$ tends to 0.

Example (normal mean, two-sided). Let us modify the previous example by taking π_1 Lebesgue measure (Jeffrey's prior) and $\sigma^2 = 1$. Thus $\theta_0 = 0$, $f_0(x)$ is standard normal density, $f_1(x|\theta)$ is normal density with mean θ and unit variance. Let $\pi_1(\theta) \equiv 1$. Then with $n = 1$ and $x_1 = x$

$$f_1(x) = \int f_1(x|\theta)d\theta = \frac{1}{\sqrt{2\pi}} \int \exp[-\frac{1}{2}(x-\theta)^2]d\theta = 1. \tag{5.17}$$

Therefore (the generalization is Exercise 6)

$$q_0(x) = \frac{q_0(2\pi)^{-\frac{1}{2}} \exp[\frac{-x^2}{2}]}{q_0(2\pi)^{-\frac{1}{2}} \exp[\frac{-x^2}{2}] + q_1} = \frac{q_0}{q_0 + q_1\sqrt{2\pi}\exp[\frac{x^2}{2}]}. \tag{5.18}$$

Hence the maximal value of $q_0(x)$ is $q_0(0) = \left(1 + (q_1/q_0)\sqrt{2\pi}\right)^{-1}$, which for $q_0 = 0.5$ is 0.285.

In Tables 5.2.5 and 5.2.6 in [6], the probabilities $q_0(x)$ are calculated for $\pi_1(\theta) \equiv 1$ and $\pi_1(\theta) \equiv 10$, resp ($q_0 = 0.5$). For $\pi_1(\theta) \equiv 1$, $q_0(0) = 0.28$ and $q_0(1.96) = 0.055$; for $\pi_1(\theta) \equiv 10$, $q_0(0) = 0.0384$ (upper bound for $q_0(x)$) and $q_0(1.96) = 0.00581$.

So far we have observed that improper prior can be considered as limit of proper priors and the posterior behaves accordingly. Recall, for example, the normal model with known variance:

$$\theta \sim \mathcal{N}(\mu, \tau^2)$$
$$X_1, \ldots, X_n|\theta \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2).$$

The posterior $\theta|x \sim \mathcal{N}(\mu_n, \tau_n^2)$ and when $\tau^2 \to \infty$, then the posterior converges to $\mathcal{N}(\bar{x}, \frac{\sigma^2}{n})$ that is the posterior under improper (Lebesgue) prior $\pi(\theta) \equiv c$. Since Lebesgue measure can (in a sense) be considered as a limit of normals as $\tau^2 \to \infty$, we obtain a certain continuity. There are many examples like that where improper prior can be considered (in a sense) as a continuation of proper prior and the posterior of improper prior is a limit of corresponding posteriors. In hypotheses testing the situation is different. Recall the normal example – with π_1 being $\mathcal{N}(\mu, \tau^2)$, for any x, it holds $q_0(x) \to 1$. Hence in the limit the posterior probability of 0-model is 1. By the continuity argument, one could expect the same for improper Lebesgue prior $\pi(\theta) \equiv c$. However, we saw totally different phenomenon: $q_0(x)$ is upper bounded by as small constant, and the bound decreases with c. This is sometimes known as *Jeffreys-Lindley paradox*. It typically holds for two-sided tests, because then only one of the model priors, typically π_1, is improper.

Example (normal mean, one-sided). Consider normal model $\mathcal{N}(\theta, \sigma^2)$. Hypotheses

$$H_0 : \ \theta < 0$$
$$H_1 : \ \theta \geq 0.$$

Let $\pi(\theta) \equiv 1$, $x = (x_1, \ldots, x_n)$. Since (recall (4.5)) $\theta|x \sim \mathcal{N}(\bar{x}, \frac{\sigma^2}{n})$, it holds

$$q_0(x) = P(\theta < 0|x) = \Phi(-z), \quad z = \frac{\bar{x}}{\sigma/\sqrt{n}},$$

where Φ is the distribution function of $\mathcal{N}(0,1)$. Therefore, $q_0(x)$ is the p-value of frequentist test. Recall from (5.11) than for conjugate normal prior $\pi = \mathcal{N}(\mu, \tau^2)$,

$$P(\theta < 0|x) = \Phi\Big(-\frac{\mu_n}{\tau_n}\Big), \quad \lim_{\tau\to\infty} \tau_n^2 = \sigma^2/n, \quad \lim_{\tau\to\infty} \mu_n = \bar{x},$$

so that when $\tau \to \infty$, then the posterior probability $P(\theta < 0|x)$ converges to that of flat prior.

Pseudo-Bayes factors. Let π be a improper prior. Following [6], we say that $(x_1, \ldots, x_n)$ is a *a training sample*, if the corresponding posterior $\pi(\cdot|x_1, \ldots, x_n)$ is proper and is a minimal training sample if no subsample is a training sample.

Example. Normal model $\mathcal{N}(\mu, \sigma^2)$. The parameters $\theta = (\mu, \sigma^2)$.

1. With Jeffreys prior $\pi(\theta) = (\sigma^2)^{-3/2}$, the posterior is NIX with parameters (Subsection 4.3.5)

$$\nu_n = n, \quad \tau_n^2 = \frac{\sum_i (x_i - \bar{x})^2}{n}, \quad \mu_n = \bar{x}, \quad \kappa_n = n.$$

   The minimal training sample size is 1.

2. With prior $\pi(\theta) = (\sigma^2)^{-1}$, the posterior is NIX with parameters (Subsection 4.2.2)

$$\nu_n = n - 1, \quad \tau_n^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}, \quad \mu_n = \bar{x}, \quad \kappa_n = n.$$

   Hence the the size of minimal training sample is 2.

3. With prior $\pi(\theta) = (\sigma)^{-1} = (\sigma^2)^{-1/2}$, the minimal training sample size is 3 (Exercise 4).

Consider the general setting (5.12), where π_0 is a proper prior on Θ_0 and π_1 improper prior on Θ_1. Given a sample $x = (x_1, \ldots, x_n)$, let $x_{(\ell)}$ be the minimal training sample for π_1; let $x_{(-\ell)}$ be the rest of the sample. Hence both posteriors $\pi_i(\cdot|x_{(\ell)})$ $(i = 0, 1)$ are proper and independent of normalizing constants.

The *pseudo-Bayes factor* is

$$B_{01}^{(\ell)} := \frac{\int_{\Theta_0} f_0(x_{(-\ell)}|\theta)\pi_0(d\theta|x_{(\ell)})}{\int_{\Theta_1} f_1(x_{(-\ell)}|\theta)\pi_1(d\theta|x_{(\ell)})}. \tag{5.19}$$

It depends on the choice of $x_{(\ell)}$.

The following holds ($x = (x_1, \ldots, x_n)$, Exercise 5):

$$\begin{aligned} B_{01}^{(\ell)} &= B_{01}(x) \cdot B_{10}(x_{(l)}) \quad \text{where} \\ B_{01}(x) := \frac{\int_{\Theta_0} f_0(x|\theta)\pi_0(d\theta)}{\int_{\Theta_1} f_1(x|\theta)\pi_1(d\theta)}, &\quad B_{10}(x_{(l)}) := \frac{\int_{\Theta_1} f_1(x_{(l)}|\theta)\pi_1(d\theta)}{\int_{\Theta_0} f_0(x_{(l)}|\theta)\pi_0(d\theta)}. \end{aligned} \tag{5.20}$$

Example (normal mean, two-sided). The model $\mathcal{N}(\theta, 1)$: Hypotheses

$$\begin{aligned} H_0 &: \theta = 0, \quad \pi_0 = \delta_0 \\ H_1 &: \theta \neq 0, \quad \pi_1(\theta) \equiv 1. \end{aligned}$$

The minimal sample size is 1. Hence (recall (5.17))

$$B_{10}(x_{(1)}) = \sqrt{2\pi}\exp[\frac{x_1^2}{2}]$$

Since

$$f_0(x_1, \ldots, x_n) = (2\pi)^{-\frac{n}{2}}\exp[-\frac{\sum_{i=1}^n x_i^2}{2}]$$

and (by (4.3))

$$f_1(x_1, \ldots, x_n) = \int (2\pi)^{-\frac{n}{2}}\exp[-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2}]d\theta = (2\pi)^{\frac{(1-n)}{2}} n^{-\frac{1}{2}}\exp[-\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{2}]$$

so that

$$B_{01}(x) = (\frac{2\pi}{n})^{-\frac{1}{2}}\exp[-\frac{(\bar{x})^2}{2/n}]. \tag{5.21}$$

By Formula (5.20), thus

$$B_{01}^{(1)} = B_{01}(x)\cdot B_{10}(x_{(l)}) = (\frac{2\pi}{n})^{-\frac{1}{2}}\exp[-\frac{(\bar{x})^2}{2/n}]\sqrt{2\pi}\exp[\frac{x_1^2}{2}] = \sqrt{n}\exp[\frac{x_1^2 - n(\bar{x})^2}{2}].$$

Let us also calculate $B_{01}(x)$ directly, i.e. without formula (5.20). Recall (4.4):

$$\pi_1(\theta|x_1) = \frac{1}{\sqrt{2\pi}} \exp[-\frac{(x_1-\theta)^2}{2}].$$

Then (recall (4.3))

$$\begin{aligned} f_1(x_2,\dots,x_n) &= (2\pi)^{-\frac{(n-1)}{2}} \int \exp[-\frac{1}{2}\sum_{i=2}^{n}(x_i-\theta)^2] \frac{1}{\sqrt{2\pi}} \exp[-\frac{(x_1-\theta)^2}{2}] d\theta \\ &= (2\pi)^{-\frac{n}{2}} \int \exp[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2] d\theta \\ &= (2\pi)^{-\frac{(n-1)}{2}} n^{-1/2} \exp[-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{2}]. \end{aligned}$$

Since

$$f_0(x_2,\dots,x_n) = (2\pi)^{-\frac{(n-1)}{2}} \exp[-\frac{\sum_{i=2}^{n} x_i^2}{2}],$$

we get the same result

$$B_{01}^{(1)} = \frac{f_0(x_2,\dots,x_n)}{f_1(x_2,\dots,x_n)} = \sqrt{n} \exp[\frac{x_1^2 - n(\bar{x})^2}{2}].$$

We see that $B_{01}^{(1)}$ depend on x_1 – the choice of training sample.

Example (normal mean, unknown variance, two-sided). Hypotheses

$$\begin{aligned} H_0 &: \mu = 0, \quad \pi_0(\sigma^2) \propto \frac{1}{\sigma^2}. \\ H_1 &: \mu \neq 0, \quad \pi_1(\mu,\sigma^2) \propto \frac{1}{\sigma^2}. \end{aligned}$$

The minimal sample size under H_1 is $n = 2$ and then the posterior is NIX with parameters (section 4.2.2)

$$\nu = 1, \quad \tau^2 = \sum_{i=1}^{2}(x_i - \bar{x}_1)^2 = \frac{(x_1-x_2)^2}{2}, \quad \mu_o = \bar{x}_1 := \frac{x_1+x_2}{2}, \quad \kappa = 2.$$

Applying formula (2.26) (marginal density under NIX prior):

$$f(x) = \frac{\Gamma(\frac{\nu_n}{2})}{\Gamma(\frac{\nu}{2})} \frac{\left(\nu\tau^2\right)^{\frac{\nu}{2}}}{\pi^{\frac{n}{2}} \left(\nu_n\tau_n^2\right)^{\frac{\nu_n}{2}}} \sqrt{\frac{\kappa}{\kappa_n}}$$

with (sample size is $n-2$)

$$\nu_n = n-2+\nu = n-1, \quad \kappa_n = n-2+\kappa = n,$$
$$\nu_n\tau_n^2 = \tau^2 + \sum_{i=3}^{n}(x_i-\bar{x}_3)^2 + \frac{2(n-2)}{n}(\bar{x}_3-\bar{x}_1)^2, \quad \bar{x}_3 := \frac{1}{n-2}\sum_{i=3}^{n} x_i,$$

the marginal density is

$$f_1(x_3,\ldots,x_n) = \int_{-\infty}^{\infty} f(x_3,\ldots,x_n|\sigma^2,\mu)\pi_1(\mu,\sigma^2|x_1,x_2)d\mu d\sigma^2 = \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{1}{2})}\pi^{\frac{(2-n)}{2}}\frac{(\tau^2)^{\frac{1}{2}}}{(\nu_n\tau_n^2)^{\frac{n-1}{2}}}\sqrt{\frac{2}{n}},$$

Under H_0 the posterior is (recall (4.6)) ScaleInv-$\chi^2(\nu,\tau^2)$ where

$$\nu = 2, \quad \nu\tau^2 = \sum_{i=1}^{2} x_i^2.$$

In particular,

$$\pi_0(\sigma^2|x_1,x_2) \propto (\sigma^2)^{-(1+1)}\exp[-\frac{x_1^2+x_2^2}{2\sigma^2}].$$

Applying formula (2.16) (the marginal density under scaled inverse χ^2 with sample size $n-2$, we obtain :

$$\begin{aligned} f_0(x_3,\ldots,x_n) &= \frac{\Gamma\big(\frac{n}{2}\big)}{\Gamma\big(\frac{2}{2}\big)}\Big(\sum_{i=1}^{2} x_i^2\pi\Big)^{-(n-2)/2}\Big(\frac{\sum_{i=3}^{n} x_i^2}{\sum_{i=1}^{2} x_i^2}+1\Big)^{-\frac{n}{2}} \\ &= \Gamma\big(\frac{n}{2}\big)(\pi)^{\frac{2-n}{2}}\Big(\frac{\sum_{i=1}^{2} x_i^2}{\sum_{i=1}^{n} x_i^2}\Big)^{\frac{n}{2}}\Big(\sum_{i=1}^{2} x_i^2\Big)^{\frac{2-n}{2}}. \end{aligned}$$

Now the pseudo-Bayes factor

$$B_{(01)}^{(2)}(x) = \frac{f_0(x_3,\ldots,x_n)}{f_1(x_3,\ldots,x_n)} = \frac{\Gamma\big(\frac{n}{2}\big)}{\Gamma(\frac{n-1}{2})}\sqrt{n\pi}\frac{(\nu_n\tau_n^2)^{\frac{n-1}{2}}}{(\sum_{i=1}^{n} x_i^2)^{\frac{n}{2}}}\frac{(x_1^2+x_2^2)}{|x_1-x_2|},$$

which depends on x_1 and x_2.

A way to remove this dependence is to average all possible pseudo-Bayes factors over all possible training samples. One option is so called *arithmetic intrinsic-Bayes factor*

$$B_{01}^{A} = \frac{1}{L}\sum_{x^{(l)}} B_{01}^{(l)} = B_{01}(x)\frac{1}{L}\sum_{x^{(l)}} B_{10}(x^{(l)}).$$

Example (normal). For the normal unknown mean example we obtain

$$B_{01}^A := B_{01}(x)\sqrt{2\pi}\Big(\frac{1}{n}\sum_{l=1}^{n}\exp[x_l^2/2]\Big) = \exp[-\frac{n\bar{x}^2}{2}]\Big(\frac{1}{\sqrt{n}}\sum_{l=1}^{n}\exp[x_l^2/2]\Big).$$

When π_1 is $\mathcal{N}(0,2)$, then the regular Bayes factor is (recall (5.16), $\tau^2 = 2$, $\sigma^2 = 1$)

$$\sqrt{\frac{\tau^2 n + \sigma^2}{\sigma^2}}\exp\Big[-\frac{1}{2}\frac{\tau^2 n}{(\sigma^2+\tau^2 n)}\big(\frac{\bar{x}}{\sigma/\sqrt{n}}\big)^2\Big] = \sqrt{2n+1}\exp\Big[-\frac{n\bar{x}^2}{2}\cdot\frac{2n}{(1+2n)}\Big].$$

5.3 Credible intervals

Credible sets are Bayesian analogue of confidence sets. Definition (from [6]): For a prior π, a set C_x is said to be an *α-credible region* if

$$P(\theta \in C_x|x) = 1-\alpha.$$

This region is called an *HPD a-credible region* (for highest posterior density), denoted by C_x^α if it can be written under the form

$$\{\theta : \pi(\theta|x) > k_\alpha\} \subset C_x^\alpha \subset \{\theta : \pi(\theta|x) \geq k_\alpha\},$$

where k_α is the largest bound such that

$$P(\theta \in C_x^\alpha|x) = 1-\alpha.$$

To consider only HPD regions is motivated by the fact that they minimize the volume (when density is with respect to Lebesgue measure) among a-credible regions and, therefore, it is in a sense an optimal solutions. Credible intervals can be used with improper priors (as long as posterior is proper), as the following examples show, they might coincide with frequentist confidence intervals.

Example (normal mean). Consider the normal model with known variance and conjugate prior

$$\theta \sim \mathcal{N}(\mu,\tau^2)$$
$$X_1,\ldots,X_n|\theta \overset{iid}{\sim} \mathcal{N}(\theta,\sigma^2).$$

The posterior is normal with mean $\mu_n(\bar{x})$ and variance τ_n^2 as in (5.10)

$$\mu_n(\bar{x}) = \mu + (n\bar{x}-n\mu)\frac{\tau^2}{n\tau^2+\sigma^2} = \frac{\mu\sigma^2+n\tau^2\bar{x}}{n\tau^2+\sigma^2} = \frac{\frac{\mu}{\tau^2}+\frac{n}{\sigma^2}\bar{x}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}$$
$$\tau_n^2 = (\frac{n}{\sigma^2}+\frac{1}{\tau^2})^{-1} = \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}.$$

Then, with e k_α being $\alpha/2$ quantile of standard normal distribution, the HPD region is

$$C_x^\alpha = [\mu_n(x) - k_\alpha \tau_n, \mu_n(x) + k_\alpha \tau_n].$$

With improper flat (Jefferys) prior $\pi(\theta) \equiv 1$, the posterior (recall (4.5)) is normal with mean $\bar{x}$ and variance σ^2/n (the limits of μ_n and τ_n^2 as $\tau^2 \to \infty$) so that HPD region is the same as the frequentist confidence interval

$$C_x^\alpha = [\bar{x} - k_\alpha \sigma/\sqrt{n}, \bar{x} + k_\alpha \sigma/\sqrt{n}].$$

Example (normal, unknown mean and variance). Consider the normal model with unknown variance and flat improper prior $\pi(\mu, \sigma^2) \propto (\sigma^2)^{-1}$. We are interested in HDP region for μ. The posterior (subsection 4.2.2)

$$\mu | x \sim \mathrm{lst}\Big(\bar{x}, \frac{\sum_i (x_i - \bar{x})^2}{n(n-1)}, n-1\Big).$$

Therefore

$$\frac{\sqrt{n}(\mu - \bar{x})}{\sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}} = \frac{\mu - \bar{x}}{\sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n(n-1)}}} \sim t_{n-1},$$

so that with k_α being $\alpha/2$ quantile of student t_{n-1} distribution, the HDP region is the same as the standard confidence interval in frequentist statistics:

$$C_x^\alpha = \Big[\bar{x} - k_\alpha \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n(n-1)}}, \bar{x} + k_\alpha \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n(n-1)}}\Big].$$

To find the HDP-region for conjugate NIX prior is Exercise 7.

Exercises:

1. Prove (5.3), (5.4).
2. Prove that for $a_0 = a_1 = 1$ the equations (5.5) and (5.6) hold.
3. Let Θ_0 and Θ_1 form a partition of Θ; let π_i be a prior density on Θ_i and let $f_i(\cdot|\theta)$ be conditional densities for observation. With model (5.12) show that $q_i(x) = P(Z = i | X = x)$ and prove that (5.4), (5.15) and (5.9) hold.
4. Prove that for normal model with $\pi(\theta) = (\sigma)^{-1}$, the minimal training sample size is 3.
5. Prove (5.20).
6. Generalize (5.18) for $n > 1$ and $\sigma^2 \neq 1$.
7. Consider the normal model with unknown μ and σ^2. Find the HDP region for mean μ under NIX-prior with hyperparameters $\nu, \mu_o, \tau^2, \kappa$.

6 Posterior consistency

Let $\{f(\cdot|\theta); \theta \in \Theta\}$ be a parametric model, and let $X_1, X_2, \ldots$ be iid random variables with density $f(\cdot|\theta_0)$, where $\theta_0 \in \Theta$ is so called *true parameter*. Since the true data generating parameter θ_0 is unknown, a prior probability measure π is assumed and so one ends up with our standard model (1.4):

$$\theta \sim \pi$$
$$X_1, X_2, \ldots | \theta \overset{i.i.d.}{\sim} f(\cdot|\theta).$$

The existence of true parameter θ_0 is called *frequentist setting*. Then, formally, θ is not a random variable any more. However, the posterior measure

$$P(A|X_1, \ldots, X_n) := \int_A \pi(\theta|X_1, \ldots, X_n)d\theta = \frac{\int_A \prod_{i=1}^n f(X_i|\theta)\pi(\theta)d\theta}{\int \prod_{i=1}^n f(X_i|\theta)\pi(\theta)d\theta}$$

still exists and we ask: how does the posterior distribution behave when the sample size n grows? It is natural to expect that when the prior is not very badly chosen, then the true parameter θ_0 should be recovered as n grows. In classical frequentist statistics, the property of recovering the true parameter as sample size increases is called *consistency*: an estimator $\hat{\theta}_n$ (that is a function of $X_1, \ldots, X_n$, hence random variable) is *consistent* if it converges to the true parameter θ_0 in some sense (as n grows). When $\hat{\theta}_n \to \theta_0$ in probability, then $\hat{\theta}_n$ is called *consistent*, and when $\hat{\theta}_n \to \theta_0$, a.s., then $\hat{\theta}_n$ is called *strongly consistent*. In Bayesian case, the main idea behind the consistency is the same but since instead of random variables (estimators) there are random measures $P(\cdot|X_1, \ldots, X_n)$, the definitions are should be slightly modified. There are (slightly) different versions of consistency in the literature, in parametric case the most common version is as follows: The posterior distribution $P(\cdot|X_1, \ldots, X_n)$ is called *strongly consistent at θ_0* if for every $\epsilon > 0$,

$$P\big(\|\theta - \theta_0\| > \epsilon | X_1, \ldots, X_n\big) \to 0, \quad f(\cdot|\theta_0) - \text{a.s.}. \tag{6.1}$$

When the convergence in (6.1) holds in probability, then the posterior distribution $P(\cdot|X_1, \ldots, X_n)$ is called *consistent (mõjus)*.

Remarks:

1. In the frequentist setting θ is not a random variable, so the conditional probability in (6.1) is defined as follows

   $$P\big(\|\theta - \theta_0\| > \epsilon | X_1, \ldots, X_n\big) := P\big(B^c(\theta_0, \epsilon)\big| X_1, \ldots, X_n\big),$$

   where $B(\theta_0, \epsilon) \subset \Theta$ is the closed ball centered at θ_0 and having radius $\epsilon > 0$. The convergence (6.1) means that the posterior mass outside ϵ-neighborhood tends to 0 and it holds for any neighborhood. However, to

facilitate the reading, in what follows, we use the full Bayesian notation as in (6.1), i.e $P(\theta \in \cdot|X_1,\dots,X_n) := P(\cdot|X_1,\dots,X_n)$. So we use this notation even when the prior is improper (so that the conditional distribution $P(\theta \in \cdot|X_1,\dots,X_n)$ does not exist), but posterior measure in proper probability measure and $P(\cdot|X_1,\dots,X_n)$ exists.

2. $f(\cdot|\theta_0)$ – a.s. means that $X_1, X_2, \dots$ are iid with density $f(\cdot|\theta_0)$. It does not necessarily presuppose the frequentist setting (true parameter exists). When θ is random variable (as in fully Bayesian model (1.4)), then θ_0 is just a realization of θ.

3. By definition of a.s. convergence, for any $\epsilon > 0$ there exists the exceptional null set where the convergence fails. However, since we can restrict ourselves with rational ϵ-s, a single null set always works.

4. When $P_1, P_2, \dots$ is a sequence of probability measures on Θ such that for every $\epsilon > 0$, $P_n\big(B^c(\theta_0,\epsilon)\big) \to 0$, then (see e.g. Prop 6.2 in [2]) $P_n \Rightarrow \delta_{\theta_0}$, where $\Rightarrow$ stands for the weak convergence of probability measures and δ_{θ_0} is Dirac measure at θ_0. Thus (6.1) can be equivalently stated as follows:

$$P(\theta \in \cdot|X_1,\dots,X_n) \Rightarrow \delta_{\theta_0}, \quad f(\cdot|\theta_0) - \text{a.s.}. \tag{6.2}$$

Moreover, with d being a metric in Θ that metrizes the weak convergence (e.g. Prokhorov metric), then (6.2) is

$$d\big(P(\theta \in \cdot|X_1,\dots,X_n), \delta_{\theta_0}\big) \to 0, \quad f(\cdot|\theta_0) - \text{a.s.}. \tag{6.3}$$

6.1 A simple criterion for consistency

In the following proposition, we assume $X_1, X_2, \dots$ to be iid random variables with density $f(\cdot|\theta_0)$ and the a.s. convergence will be with respect to θ_0 (i.e. $f(\cdot|\theta_0)$-a.s. will be dropped). We shall assume $\Theta \subset \mathbb{R}^1$ and we denote denote posterior mean and variance by μ_n and σ_n^2, resp. Thus

$$\mu_n = E[\theta|X_1,\dots,X_n] = \int \theta\pi(\theta|X_1,\dots,X_n)d\theta,$$
$$\sigma_n^2 = \mathrm{Var}[\theta|X_1,\dots,X_n] = \int (\theta-\mu_n)^2\pi(\theta|X_1,\dots,X_n)d\theta.$$

Proposition 6.1 *Let $\sigma_n^2 \to 0$ and $\mu_n \to \theta_0$, a.s.. Then the posterior distribution is strongly consistent.*

Proof. By triangular inequality

$$\begin{aligned} P(|\theta-\theta_0| > \epsilon|X_1,\dots,X_n) \le & P(|\theta-\mu_n| > \epsilon/2|X_1,\dots,X_n) + \\ & + P(|\mu_n-\theta_0| > \epsilon/2|X_1,\dots,X_n). \end{aligned}$$

Since $\mu_n \to \theta_0$, a.s., $P(|\mu_n - \theta_0| > \epsilon/2|X_1, \ldots, X_n) = 0$ eventually, a.s. (meaning that for almost every ω, $P(|\mu_n - \theta_0| > \epsilon/2|X_1, \ldots, X_n) = 0$, eventually). By Chebyshev inequality

$$P(|\theta - \mu_n| > \epsilon/2|X_1, \ldots, X_n) \leq \frac{\sigma_n^2}{(\epsilon/2)^2} \to 0, \quad \text{a.s.}$$

Hence (6.1) holds. ■

The proof relies on the posterior measure only (Chebyshev inequality does not require the random variables), hence Proposition 6.1 also holds when the prior is improper but posterior is proper.

In multivariate case, i.e. $\theta = (\theta_1, \ldots \theta_d)$ the consistency holds, when the assumptions of Proposition 6.1 holds for every component (Exercise 1): for every $i = 1, \ldots, d$:

$$E[\theta_i|X_1, \ldots, X_n] \to \theta_{0,i}, \quad \text{Var}[\theta_i|X_1, \ldots, X_n] \to 0, \quad \text{a.s.}. \tag{6.4}$$

Examples.

1. Beta-Bernoulli model: $X_1, X_2, \ldots$ iid $X_i \sim \text{B}(1, \theta_0)$. Prior: Beta$(\alpha, \beta)$. Posterior is
$$\text{Beta}\big(\sum_{i=1}^n X_i + \alpha, n - \sum_{i=1}^n X_i + \beta\big)$$
with mean
$$\mu_n = \frac{\sum_{i=1}^n X_i}{n} \cdot \frac{n}{\alpha + \beta + n} + \frac{\alpha}{\alpha + \beta + n}.$$
By SLLN, $\frac{\sum_{i=1}^n X_i}{n} \to \theta_0$, a.s, and so $\mu_n \to \theta_0$, a.s.
Posterior variance
$$\sigma_n^2 = \frac{\mu_n(1 - \mu_n)}{\alpha + \beta + n + 1} \to 0, \quad \text{a.s.}.$$

2. Gamma-Poisson model: $X_1, X_2, \ldots$ iid $X_i \sim \text{Po}(\theta_0)$. Prior: Gamma$(\alpha, \beta)$. Posterior is
$$\text{Gamma}\big(\sum_{i=1}^n X_i + \alpha, n + \beta\big),$$
with mean
$$\mu_n = \frac{\sum_{i=1}^n X_i}{n} \cdot \frac{n}{\beta + n} + \frac{\alpha}{\beta + n}.$$

By SLLN, $\frac{\sum_{i=1}^n X_i}{n} \to \theta_0$, a.s., so $\mu_n \to \theta_0$, a.s..
Posterior variance:

$$\sigma_n^2 = \frac{\sum_{i=1}^n X_i + \alpha}{(n+\beta)^2} = \frac{\sum_{i=1}^n X_i}{n} \cdot \frac{n}{(n+\beta)^2} + \frac{\alpha}{(n+\beta)^2} \to 0, \quad \text{a.s..}$$

With Jeffreys prior: Gamma$(1/2, 0)$ (improper) posterior is

$$\text{Gamma}\big(\sum_{i=1}^n X_i + 1/2, n\big)$$

with mean and variance

$$\mu_n = \frac{\sum_{i=1}^n X_i + 1/2}{n} \to \theta_0, \quad \sigma_n^2 = \frac{\sum_{i=1}^n X_i + 1/2}{n^2} \to 0, \quad \text{a.s..}$$

3. Gamma-Exponential model: $X_1, X_2, \ldots$ iid $X_i \sim \text{Exp}(\theta_0)$. Prior: Gamma$(\alpha, \beta)$. Posterior is

$$\text{Gamma}\big(n+\alpha, \sum_{i=1}^n X_i + \beta\big),$$

with mean:

$$\mu_n = \frac{n+\alpha}{\sum_{i=1}^n X_i + \beta} \to \theta_0, \quad \text{a.s.,}$$

since

$$\frac{\sum_{i=1}^n X_i}{n} \to \frac{1}{\theta_0} \quad \text{a.s..}$$

Posterior variance:

$$\sigma_n^2 = \frac{n+\alpha}{(\sum_{i=1}^n X_i + \beta)^2} = \frac{n+\alpha}{n^2(\sum_{i=1}^n X_i/n + \beta/n)^2} \to 0, \quad \text{a.s..}$$

With Jeffreys prior: Gamma$(0, 0)$ (improper).
Posterior is Gamma$\big(n, \sum_{i=1}^n X_i\big)$, with mean and variance

$$\mu_n = \frac{n}{\sum_{i=1}^n X_i} \to \theta_0, \quad \sigma_n^2 = \frac{n}{(\sum_{i=1}^n X_i)^2} \to 0 \quad \text{a.s..}$$

4. Normal model, unknown mean: $X_1, X_2, \ldots$ iid $X_i \sim \mathcal{N}(\theta_0, \sigma^2)$. Prior: normal $\mathcal{N}(\mu, \tau^2)$. Posterior is normal with mean:

$$\mu_n = \frac{\sum_{i=1}^n X_i}{n} \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} + \mu \frac{\frac{1}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \to \theta_0, \quad \text{a.s.}$$

and variance

$$\sigma_n^2 = \frac{\sigma^2 \tau^2}{n\tau^2 + \sigma^2} \to 0.$$

With Jeffreys prior (constant) posterior is normal with mean and variance

$$\mu_n = \frac{\sum_{i=1}^n X_i}{n} \to \theta_0, \quad \sigma_n^2 = \frac{\sigma^2}{n} \to 0.$$

5. Normal model, unknown variance: $X_1, X_2, \ldots$ iid $X_i \sim \mathcal{N}(\mu, \theta_0)$. Prior: scale-inversed $\chi^2(\nu, \tau^2)$. Posterior is ScaleInv$-\chi^2(\nu_n, \tau_n^2)$, where

$$\nu_n := \nu + n, \quad \nu_n \tau_n^2 := \sum_{i=1}^n (X_i - \mu)^2 + \nu\tau^2.$$

Posterior mean:

$$\mu_n = \frac{\sum_{i=1}^n (X_i - \mu)^2 + \nu\tau^2}{\nu + n - 2} = \frac{\sum_{i=1}^n (X_i - \mu)^2}{n} \frac{n}{\nu + n - 2} + \frac{\nu\tau^2}{n + \nu - 2} \to \theta_0, \quad \text{a.s.},$$

because

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \to \theta_0, \quad \text{a.s.}.$$

Posterior variance

$$\sigma_n^2 = \frac{2\mu_n^2}{\nu_n - 4} \to 0, \quad \text{a.s.}.$$

With Jeffreys prior (proportional to $1/\theta$) ScaleInv$-\chi^2(0, \tau^2)$ the posterior is ScaleInv$-\chi^2(\nu_n, \tau_n^2)$, where

$$\nu_n := n, \quad n\tau_n^2 := \sum_{i=1}^n (X_i - \mu)^2.$$

Posterior mean and variance

$$\mu_n = \frac{\sum_{i=1}^n (X_i - \mu)^2}{n - 2} \to \theta_0, \quad \sigma_n^2 = \frac{2\mu_n^2}{n - 4} \to 0, \quad \text{a.s.}.$$

6. Normal model, unknown mean and variance: $X_1, X_2, \ldots$ iid $X_i \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Prior: NIX with ν, m, τ^2, κ, for Jeffreys prior $\nu = \kappa = 0$.
Posterior: NIX with

$$\begin{aligned}
\nu_n &= \nu + n \\
\mu_n &= \frac{\kappa}{\kappa + n} m + \frac{n}{\kappa + n} \frac{\sum_{i=1}^n X_i}{n} \\
\kappa_n &= \kappa + n \\
\nu_n \tau_n^2 &= \nu\tau^2 + \sum_{i=1}^n \Big(X_i - \frac{\sum_{i=1}^n X_i}{n} \Big)^2 + \frac{\kappa n}{\kappa + n} \Big(\frac{\sum_{i=1}^n X_i}{n} - m \Big)^2.
\end{aligned}$$

Posterior means (subsection 2.5.3)

$$\begin{aligned}
E[\mu | X_1, \ldots, X_n] &= \mu_n = \frac{\kappa}{\kappa + n} m + \frac{n}{\kappa + n} \frac{\sum_{i=1}^n X_i}{n} \to \mu_0, \quad \text{a.s.}, \\
E[\sigma^2 | X_1, \ldots, X_n] &= \frac{\nu\tau^2 + \sum_{i=1}^n (X_i - \mu_n)^2 + \kappa(\mu_n - m)^2}{\nu + n - 2} \to \sigma_0^2, \quad \text{a.s.}.
\end{aligned}$$

Posterior variances: Since $\sigma^2|X_1,\ldots,X_n \sim \text{ScaleInv}{-}\chi^2(\nu_n,\tau_n^2)$

$$\text{Var}[\sigma^2|X_1,\ldots,X_n] = \frac{2\mu_n^2}{\nu_n - 4} \to 0, \quad \text{a.s..}$$

Since

$$\mu|X_1,\ldots,X_n \sim \text{lst}\big(\mu_n, \frac{\tau_n^2}{\kappa_n}, \nu_n\big)$$

we have

$$\text{Var}[\mu|X_1,\ldots,X_n] = \frac{\nu_n\tau_n^2}{\kappa_n(\nu_n-2)} \to 0, \quad \text{a.s..}$$

6.2 Doob's consistency theorem

The celebrated Doob's consistency theorem assumes (only) that our model $\{f(\cdot|\theta)\}$ is identifiable meaning that to different parameters corresponds different probability distributions: when $\theta \neq \theta'$, then there exists $A \in \mathcal{B}(\mathcal{X})$ such that

$$P_\theta(A) = \int_A f(x|\theta)dx \neq \int_A f(x|\theta')dx = P_{\theta'}(A).$$

Recall that $g \in L_1(\pi)$ means that $g : \Theta \to \mathbb{R}$ is a measurable function such that $\int |g(\theta)|\pi(d\theta) < \infty$.

Theorem 6.1 (Doob's consistency theorem) *Let π be an arbitrary proper prior on Θ and let $g \in L_1(\pi)$. Then there exists a set $\Theta_0 \subset \Theta$ with $\pi(\Theta_0) = 1$ such that the posterior is strongly consistent at every $\theta_0 \in \Theta_0$. Moreover,*

$$E[g(\theta)|X_1,\ldots,X_n] \to g(\theta_0), \quad f(\cdot|\theta_0) - \text{a.s..}$$

In full Bayesian setting (1.4), θ is a random variable with distribution π and Doob's theorem simply states that for almost every realization of θ, the consistency as well as the convergence of conditional expectation of g holds without further assumptions. In frequentist setting, there is an unknown true parameter θ_0. Doob's theorem then states that whatever prior one chooses, there is a set Θ_0 with prior probability one so that strong consistency holds provided $\theta_0 \in \Theta_0$. Unfortunately in practice one does not know in advance whether the unknown θ_0 belongs to Θ_0 or not, and that constitutes the main criticism of Doob's theorem. An exception is countable Θ. Then any prior assigning positive mass to all parameters on Θ guarantees the posterior consistency, because for any such prior, $\Theta_0 = \Theta$.

Formally, Doob's theorem assumes proper prior. With improper prior, π one can use (minimal) training sample $x_{(l)}$, define a proper posterior

$$\pi_{(\ell)}(\cdot) := P(\cdot|x_{(\ell)}) \tag{6.5}$$

and use this as the prior for the rest of the observations to get posterior $P_{(\ell)}(\cdot|X_{\ell+1},\dots,X_n)$. Since (Exercise 2) $P_{(\ell)}(\cdot|X_{\ell+1},\dots,X_n) = P(\cdot|X_1,\dots,X_n)$, the consistency prevails. The set Θ_0, however, might depend on $X_{(l)}$. When π has an atom θ_0, and $f(X_{(\ell)}|\theta_0) > 0$, then θ_0 is an atom of $\pi_{(\ell)}$ as well (Exercise 2). Then $\theta_0 \in \Theta_0$.

For the proof of Doob's theorem as well as a nice discussion with further references, see [1].

6.3 Schwartz's theorem

Kullbak-Leibler support. The *Kullback Leibler (KL)divergence* between two densities $f(x) := f(x|\theta)$ and $f_0(x) := f(x|\theta_0)$ is defined as follows:

$$D(f_0\|f) := \int \ln\Big(\frac{f_0(x)}{f(x)}\Big) f_0(x)dx.$$

KL-divergence is always nonnegative, $D(f_0\|f) = 0$ if and only if $f = f_0$, a.s.. KL divergence measures the difference between f and f_0, but it is not symmetric and transitive, hence not a metric. Since we only consider densities from our parametric model $\{f(\cdot|\theta)\}$, we define $D(\theta_0\|\theta) := D(f_0\|f)$, where $f(x) = f(x|\theta)$ and $f_0(x) = f(x|\theta_0)$. Since we assume identifiability, $D(\theta_0\|\theta) = 0$ iff $\theta = \theta_0$ otherwise $D(\theta_0\|\theta) > 0$.

When our model $\{f(\cdot|\theta)\}$, $\Theta \subset \mathbb{R}^d$ is such that for a.e. $x \in \mathcal{X}$, $\theta \mapsto f(x|\theta)$ is continuous at θ_0 and $\exists \epsilon > 0$ such that

$$\int \sup_{\theta \in B(\theta_0,\epsilon)} \Big|\ln\Big(\frac{f(x|\theta_0)}{f(x|\theta)}\Big)\Big| f(x|\theta_0)dx < \infty \tag{6.6}$$

then by dominated convergence $\theta \mapsto D(\theta_0\|\theta)$ is continuous at θ_0. To see that let $\theta_n \to \theta_0$ and observe that for almost any x,

$$\ln\Big(\frac{f(x|\theta_n)}{f(x|\theta_0)}\Big) \to 0.$$

Condition (6.6) allows to go with limit under the integral (this is what dominated convergence is) and so

$$\begin{aligned}\lim_{\theta\to\theta_0} D(\theta_0\|\theta) &= \lim_{\theta\to\theta_0} \int \ln\Big(\frac{f(x|\theta_0)}{f(x|\theta)}\Big) f(x|\theta_0)dx \\ &= \int \lim_{\theta\to\theta_0} \ln\Big(\frac{f(x|\theta_0)}{f(x|\theta)}\Big) f(x|\theta_0)dx = 0.\end{aligned}$$

The continuity of $D(\theta_0\|\cdot)$, identifiability (i.e. $D(\theta_0\|\theta) = 0$ only if $\theta = \theta_0$) and relative compactness (boundedness) of Θ implies

$$D(\theta_0\|\theta_n) \to 0 \Rightarrow \theta_n \to \theta_0. \tag{6.7}$$

The *support (kandja)* of a probability distribution π on Θ is the minimal closed set $\mathrm{supp}(\pi) \subseteq \Theta$ having probability one. Equivalently the support is the set of such elements whose every neighborhood has positive probability:

$$\theta \in \mathrm{supp}(\pi) \quad \Leftrightarrow \quad \forall \epsilon > 0,\ \pi(B(\theta, \epsilon)) > 0. \tag{6.8}$$

Let for every $\epsilon > 0$

$$B_{\mathrm{KL}}(\theta_0, \epsilon) := \{\theta : D(\theta_0 \| \theta) \leq \epsilon\}$$

be the Kullback-Leibler ball. We say that a parameter θ belongs to *Kullback-Leibler support of* π, denoted by $\theta \in \mathrm{supp}_{KL}(\pi)$, if $\forall \epsilon > 0$, $\pi(B_{\mathrm{KL}}(\theta, \epsilon)) > 0$.

Proposition 6.2 *If $\theta \mapsto D(\theta_0 \| \theta)$ is continuous at θ_0 and (6.7) holds, then $\theta_0 \in \mathrm{supp}_{KL}(\pi)$ if and only if $\theta_0 \in \mathrm{supp}(\pi)$.*

Proof. If $\theta \mapsto D(\theta_0 \| \theta)$ is continuous at θ_0, then by definition of continuity for every $\epsilon > 0$ there exists $\delta > 0$ such that $B(\theta_0, \delta) \subset B_{KL}(\theta_0, \epsilon)$. This means:

$$\theta_0 \in \mathrm{supp}(\pi) \Rightarrow \theta_0 \in \mathrm{supp}_{KL}(\pi). \tag{6.9}$$

On the other hand, (6.7) implies that for every $\epsilon > 0$, there exists $\delta > 0$ so that $B_{KL}(\theta_0, \delta) \subset B(\theta_0, \epsilon)$. Indeed, if not, there would exists a $\epsilon > 0$ so that for every δ there exists a θ_δ such that $D(\theta_0 \| \theta_\delta) \leq \delta$, but $\|\theta_\delta - \theta\| > \epsilon$. This contradicts (6.7). Consequently:

$$\theta_0 \in \mathrm{supp}_{KL}(\pi) \Rightarrow \theta_0 \in \mathrm{supp}(\pi). \tag{6.10}$$

■

Examples of Kullback-Leibler support.

- (Beta-)Bernoulli model. For Bernoulli densities with parameters θ, θ', KL-divergence is (Exercise 3)

$$D(\theta \| \theta') = \theta \ln\Big(\frac{\theta}{\theta'}\Big) + (1-\theta) \ln\Big(\frac{1-\theta}{1-\theta'}\Big). \tag{6.11}$$

  Here $0 \cdot \infty = \infty \cdot 0 = 0$. Thus $D(\theta \| \cdot)$ is a continuous function and (6.7) holds. For any θ and therefore for any prior $\mathrm{supp}_{KL}(\pi) = \mathrm{supp}(\pi)$. The support of Beta-distribution is $[0, 1] = \Theta$.

- (Gamma-)Poisson model. For Poisson densities with parameters θ, θ', KL-divergence is (Exercise 3)

$$D(\theta \| \theta') = \theta' - \theta + \ln\Big(\frac{\theta}{\theta'}\Big)\theta. \tag{6.12}$$

Formally $\Theta = (0, \infty)$ (for $\theta = 0$, there is no Poisson distribution). Thus $D(\theta\|\cdot)$ is a continuous function for any θ and clearly (6.7) holds. Therefore for any prior $\text{supp}_{KL}(\pi) = \text{supp}(\pi)$. The support of Gamma-distribution is then $\Theta = (0, \infty)$ (even when the support of Gamma distribution on $\mathbb{R}$ is $[0, \infty)$). One can define $P_o(0) = \delta_0$ (has density with respect to the counting measure), then $\Theta = [0, \infty)$, continuity and (6.7) holds (because $D(0, \|\theta') = \theta'$) and $\text{supp}_{KL}(\pi) = \text{supp}(\pi)$. In this case, for Gamma distribution $\text{supp}(\pi) = [0, \infty)$.

- (Gamma-)Exponential model. Here $\Theta = (0, \infty)$. For exponential densities with parameters θ, θ', KL-divergence is (Exercise 3)

$$D(\theta\|\theta') = \frac{\theta'}{\theta} - 1 + \ln\big(\frac{\theta}{\theta'}\big). \tag{6.13}$$

  Thus $D(\theta\|\cdot)$ is a continuous function for any $\theta > 0$ and (6.7) holds. Therefore for any prior, $\text{supp}_{KL}(\pi) = \text{supp}(\pi)$.

- Normal model. For normal densities with parameters (μ, σ^2) and (μ', σ'^2), KL-divergence is (Exercise 3)

$$D(\mu, \sigma^2\|\mu', \sigma'^2) = \frac{1}{2}\Big(\ln\big(\frac{\sigma'^2}{\sigma^2}\big) + \frac{(\mu - \mu')^2}{\sigma'^2} + \frac{\sigma^2}{\sigma'^2} - 1\Big). \tag{6.14}$$

  Here $\Theta = \mathbb{R} \times (0, \infty)$, $D((\mu, \sigma^2)\|\cdot)$ is continuous for any (μ, σ^2), (6.7) holds and so $\text{supp}_{KL}(\pi) = \text{supp}(\pi)$.

Tests. Let $X_1, \ldots, X_n$ be iid sample from $f(\cdot|\theta)$ with unknown parameter θ. For testing the hypotheses $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$, one typically determines a set $A_n \subset \mathcal{X}^n$ so that H_0 is rejected whenever the sample $(x_1, \ldots, x_n) \in A_n$. Hence $I_{A_n}(x_1, \ldots, x_n) = 1$ iff H_0 is rejected, otherwise $I_{A_n}(x_1, \ldots, x_n) = 0$. Let $X = (X_1, \ldots, X_n)$. Thus the probability

$$\int_{A_n} \prod_{i=1}^{n} f(x_i|\theta_0) dx_1 \cdots dx_n = P(X \in A_n|\theta = \theta_0) = E\big[I_{A_n}(X)|\theta = \theta_0\big]$$
$$=: E_{\theta_0}\big(I_{A_n}(X)\big)$$

is the probability of the error of the first type. The test is good when that probability is small. The function $\theta \mapsto E_\theta\big(I_{A_n}(X)\big)$ is called the *power function* of the test and it is desirable when that function is big when $\theta \neq \theta_0$.

In the theory of statistics, *test* is any measurable function $\phi_n : \mathcal{X}^n \to [0, 1]$. The fact that $\phi_n(x_1, \ldots, x_n)$ can take any values in $[0, 1]$, not just 0 or 1 indicates the fact that a statistical test can be *randomized* and then

$\phi_n(x_1,\dots,x_n)$ is just the probability (over the extra randomness) that 0-hypothesis will be rejected when the sample is $x_1,\dots,x_n$. Now

$$E_\theta\big(\phi_n(X)\big) = E_\theta\big(\phi_n(X_1,\dots,X_n)\big)$$

is again the probability of rejecting H_0 under θ (i.e. $X_1,\dots,X_n$ be iid sample from $f(\cdot|\theta)$). A test ϕ_n is good if $E_{\theta_0}\big(\phi_n(X)\big)$ is small (close to 0) and for any $\theta \neq \theta_0$, $E_\theta\big(\phi_n(X)\big)$ is close to one or, equivalently, $E_\theta\big(1-\phi_n(X)\big)$ is close to 0.

Theorem 6.2 (Schwartz) *Let $\theta_0 \in \mathrm{supp}_{KL}(\pi)$. Suppose that for every $\epsilon > 0$ and n, there exists tests ϕ_n such that as $n \to \infty$,*

$$E_{\theta_0}\big(\phi_n(X_1,\dots,X_n)\big) \to 0, \quad \sup_{\theta:\|\theta-\theta_0\|\geq\epsilon} E_\theta\big(1-\phi_n(X_1,\dots,X_n)\big) \to 0. \quad (6.15)$$

Then the posterior is strongly consistent at θ_0.

For the proof see [2], sec 6.

The existence of test required in Schwartz theorem is nor very restrictive. It holds true, when there exist estimators $T_n = T(X_1,\dots,X_n)$ that are *uniformly consistent*: $\forall\epsilon > 0$

$$\sup_\theta P_\theta(\|T_n-\theta\| > \epsilon) \to 0 \quad (6.16)$$

(Exercise 4). Here, obviously, $P_\theta(\|T_n-\theta\| > \epsilon) := P(\|T_n-\theta\| > \epsilon|\theta)$. Consistent tests typically exist, to prove the uniform consistency might be a challenge and the proof might depend on the concrete family. For example, when $\theta \in \mathbb{R}$ is mean and variance of $f(\cdot|\theta)$ is bounded by M for every θ (like for Bernoulli model), then by Chebyshev inequality, the sample mean $T_n = \frac{1}{n}(X_1+\cdots+X_n)$ is universally consistent:

$$P_\theta(|T_n-\theta| > \epsilon) \leq \frac{M}{n\epsilon^2} \to 0.$$

Another example is Lemma 10.4 in [3] states that when the distribution functions of $f(\cdot|\theta)$, let them be F_θ satisfy the following condition

$$\inf_{\|\theta-\theta'\|>\epsilon} \sup_t |F_\theta(t) - F_{\theta'}(t)| > 0, \quad (6.17)$$

then universally consistent test exists. The condition (6.17) is often met in practice. Finally observe that the existence of the tests depends on the model $\{f(\cdot|\theta)\}$, only (not on the prior). Hence, even when the Schwartz theorem assumes proper proof, for improper π, one can replace π by $\pi_{(\ell)}$ (recall (6.5)), just as in the case of Doob's theorem, provided $\theta_0 \in \mathrm{supp}_{KL}(\pi_{(\ell)})$.

6.4 Point-null hypotheses and consistency

Recall the point-null hypotheses:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0.$$

Let π_1 be a (possible improper) prior on $\Theta_1 = \Theta \setminus \{\theta_0\}$. We take $\pi_0 = \delta_{\theta_0}$ and fix the model priors q_0, q_1 (both positive). The overall prior on Θ is thus mixture $\pi = q_0\delta_{\theta_0} + q_1\pi_1$ that has an atom at θ_0. The posterior (recall 5.13) is also a mixture, since $\Theta_0 = \{\theta_0\}$, the posterior is now (x is replaced now by $X_1, \ldots, X_n$)

$$P(A|X_1, \ldots, X_n) = I_A(\theta_0)q_0(X_1, \ldots, X_n) + P_1(A\setminus\{\theta_0\}|X_1, \ldots, X_n)q_1(X_1, \ldots, X_n).$$

Thus (take $A = \{\theta_0\}$), $P(\{\theta_0\}|X_1, \ldots, X_n) = q_0(X_1, \ldots, X_n)$, and we show that under H_0 (i.e. when $X_1, X_2, \ldots$ are iid from $f(\cdot|\theta_0)$)

$$q_0(X_1, \ldots, X_n) \to 1, \quad \text{a.s..}$$

Assuming, for a moment, that π_1 is proper, hence π is proper as well, we apply Doob's consistency theorem with $g(\theta) = I_{\theta^0}(\theta)$ to obtain

$$P(\{\theta_0\}|X_1, \ldots, X_n) = E(g(\theta)|X_1, \ldots, X_n) \to g(\theta_0) = 1, \quad \text{a.s..} \qquad (6.18)$$

Doob's theorem applies, since θ_0 is an atom. When π_1 is improper, we additionally assume that all densities in the model $\{f(\cdot|\theta)\}$ have the same support and posterior under π_1 is finite. The additional assumption that all densities $\{f(\cdot|\theta)\}$ have the same support ensures that for almost every training sample $X_{(\ell)}$, the posterior $\pi_{(\ell)}$ (recall (6.5)) has the atom at θ_0 as well and then (6.18) holds.

Suppose now that H_1 holds, i.e. $X_1, X_2, \ldots$ are iid random variables from $f(\cdot|\theta_1)$, with $\theta_1 \neq \theta_0$. Assume that π_1 is proper and $\theta_1 \in \text{supp}_{KL}(\pi_1)$. Thus $\forall\epsilon > 0$, $\pi_1(B_{\text{KL}}(\theta_1, \epsilon)) > 0$ and since π is mixture, clearly $\pi(B_{\text{KL}}(\theta_1, \epsilon)) > 0$ as well. Hence $\theta_1 \in \text{supp}_{KL}(\pi)$. It can be shown (see [2], Thm 10.24) that the tests exists and so by Schwartz's theorem posterior consistency holds. Hence, for any ball $B := B(\theta_1, \epsilon) \subset \Theta_1$ (thus $\theta_0 \notin B(\theta_1, \epsilon)$),

$$P(B|X_1, \ldots, X_n) = P_1(B|X_1, \ldots, X_n)q_1(X_1, \ldots, X_n) \to 1, \quad \text{a.s..}$$

Hence $P_1(B|X_1, \ldots, X_n) \to 1$, a.s. – it follows from Schwartz's theorem when applied to π_1 – but also (and most importantly)

$$q_1(X_1, \ldots, X_n) \to 1, \quad \text{a.s..}$$

When π_1 is improper, one can replace π by $\pi_{(\ell)}$.

Now recall the Bayes factor and formula (5.15):

$$q_0(X_1,\dots,X_n) = \Big(1 + \frac{q_1}{q_0 B_{01}(X_1,\dots,X_n)}\Big)^{-1}.$$

Since $q_0(X_1,\dots,X_n)$ tends to 1 (to 0) if and only if $B_{01}(X_1,\dots,X_n) \to \infty$ (to 0), we have shown that under H_0, $B_{01}(X_1,\dots,X_n) \to \infty$, a.s. and (provided π_1 is not too wrongly chosen) under H_1, $B_{01}(X_1,\dots,X_n) \to 0$, a.s.. These convergences are sometimes known as *Bayes factor consistency*.

Example (normal mean, known variance). Let $f(\cdot|\theta)$ be normal density with known variance $\sigma^2 = 1$ and unknown mean θ. Hypotheses:

$$H_0 : \theta = 0$$
$$H_1 : \theta \neq 0.$$

With $\pi_1(\theta) \equiv 1$ it holds (recall (5.21)):

$$B_{01}(X_1,\dots,X_n) = \sqrt{\frac{n}{2\pi}}\exp[-\frac{n(\bar{X})^2}{2}]. \tag{6.19}$$

When $X_1, X_2, \dots$ are iid with $\theta = 0$ (i.e. H_0 holds), then by LIL, for every $\epsilon > 0$,

$$\sqrt{n}\bar{X} = \frac{\sum_{i=1}^n X_i}{\sqrt{n}} \leq (1+\epsilon)\sqrt{2\ln\ln n}, \quad \text{ev}, \quad \text{a.s.}$$
$$\exp[\frac{n(\bar{X})^2}{2}] = \exp\Big[\frac{n}{2}\Big(\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\Big)^2\Big] \leq (\ln n)^{(1+\epsilon)^2} \quad \text{ev}, \quad \text{a.s.}.$$

Therefore, under H_0, $B_{01}(X_1,\dots,X_n) \to \infty$, a.s. and $q_0(X_1,\dots,X_n) \to 1$, a.s. (recall (5.18) gives upper bound to $q_0(X_1)$, only).

When $X_1, X_2, \dots$ are iid with $\theta_1 \neq 0$ (i.e. H_1 holds), then by LIL, for every $\epsilon > 0$,

$$\frac{\sum_{i=1}^n (X_i - \theta_1)}{\sqrt{n}} \leq (1+\epsilon)\sqrt{2\ln\ln n}, \quad \text{ev}, \quad \text{a.s.}$$
$$\frac{\sum_{i=1}^n (X_i - \theta_1)}{\sqrt{n}} \geq -(1+\epsilon)\sqrt{2\ln\ln n}, \quad \text{ev}, \quad \text{a.s.}$$

This means

$$\Big|\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\Big| = \Big|\frac{\sum_{i=1}^n (X_i - \theta_1)}{\sqrt{n}} + \sqrt{n}\theta_1\Big| \leq \sqrt{n}\frac{|\theta_1|}{2} \quad \text{ev}, \quad \text{a.s.}.$$

Therefore

$$\frac{1}{\sqrt{n}}\exp\Big[\frac{1}{2}\Big(\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\Big)^2\Big]\to\infty, \qquad \text{a.s.}$$

implying that under H_1, $B_{01}(X_1,\dots,X_n)\to 0$, a.s. Hence Bayes factor consistency holds.

When $\pi_1=\mathcal{N}(0,\tau^2)$, then by (5.16),

$$B_{01}(X_1,\dots,X_n)=\sqrt{(\tau^2n+1)}\exp\Big[-\frac{1}{2}\frac{\tau^2n}{(1+\tau^2n)}\cdot n\big(\bar{X}\big)^2\Big].$$

We see again that Bayes factor consistency holds.

6.5 Bernstein-von Mieses theorem

Total variation distance. Recall the *total variation distance* between two probability measures P and Q on $\mathcal{B}$:

$$\|P-Q\|_{TV}=\sup_{A\in\mathcal{B}}|P(A)-Q(A)|.$$

When Q and P have densities p and q w.r.t. some reference measure μ, then

$$\|P-Q\|_{TV}=\frac{1}{2}\int|p(x)-q(x)|\mu(dx).$$

Sometimes the total variation distance is defined as $\int|p(x)-q(x)|\mu(dx)$. When measuring the difference between distributions of random variables X,Y observe that the total variation distance is invariant under linear transformation (location and scale changes):

$$\|P(X\in\cdot)-P(Y\in\cdot)\|_{TV}=\|P(aX+b\in\cdot)-P(aY+b\in\cdot)\|_{TV},\quad a\neq 0. \tag{6.20}$$

The total variation distance is strong: the convergence in total variation implies the convergence in distribution (weak convergence):

$$\|P_n-P\|_{TV}\to 0\Rightarrow P_n\Rightarrow P.$$

When the model $\{f(\cdot|\theta)\}$ is such that $\theta\to f(x|\theta)$ is continuous at (almost) every x, then from Sheffe's theorem it follows that the model is *continuous in total variation*, i.e.

$$\theta_n\to\theta\quad\Rightarrow\quad\int|f(x|\theta_n)-f(x|\theta)|dx\to 0. \tag{6.21}$$

Most of the models considered above satisfy (6.21). In particular, it means that for every $\epsilon>0$ there exists $\delta>0$ such that

$$\sup_A|P_\theta(A)-P_{\theta'}(A)|=\frac{1}{2}\int|f(x|\theta_n)-f(x|\theta)|dx\le\epsilon,$$

whenever $\|\theta-\theta'\|\le\delta$.

Asymptotic normality of MLE. Consider the frequentist setting: $X_1, \ldots, X_n$ is iid sample from $f(\cdot|\theta_0)$. Let $\hat{\theta}_n$ be maximum likelihood estimate (MLE), i.e.

$$\hat{\theta}_n = \arg\max_{\theta} f(X_1, \ldots, X_n|\theta).$$

As it is well known, under fairly general conditions, $\hat{\theta}_n \to \theta_0$, a.s., (consistency) and it is asymptotically normal:

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \Rightarrow \mathcal{N}\big(0, I^{-1}(\theta_0)\big), \tag{6.22}$$

where $I(\theta_0)$ is Fisher information matrix. Hence, for large n, $\hat{\theta}_n$ has approximatively $\mathcal{N}(\theta_0, n^{-1}I^{-1}(\theta_0))$-distribution. Observe that $n^{-1}I^{-1}(\theta_0) = I_n^{-1}(\theta_0)$. For a rigorous statement and proof of (6.22), see e.g. [3], Thm 5.39.

Bernstein-von Mieses theorem. Let $X_1, X_2, \ldots$ be iid random variables from $f(\cdot|\theta_0)$ and let $\hat{\theta}_n$ be MLE. The Bernstein-von Mieses theorem states that under some general assumptions and under the existence of tests satisfying (6.15) for any prior distribution that is absolutely continuous (with respect to Lebesgue'i measure) and bounded away from zero in a neighborhood of θ_0, the following holds:

$$E_{\theta_0}\big\|P(\sqrt{n}(\theta - \theta_0) \in \cdot|X_1, \ldots, X_n) - \mathcal{N}(\sqrt{n}(\hat{\theta}_n - \theta_0), I^{-1}(\theta_0))\big\|_{TV} \to 0. \tag{6.23}$$

Remarks:

1. Since we are in frequentist setting, the probability

   $$P(\sqrt{n}(\theta - \theta_0) \in \cdot|X_1, \ldots, X_n)$$

   should be interpreted as $P(\theta : \sqrt{n}(\theta - \theta_0) \in \cdot|X_1, \ldots, X_n)$, where $P(\cdot|X_1, \ldots, X_n)$ is posterior (just like in the consistency results).

2. Since $\hat{\theta}_n$ is a function of $X_1, \ldots, X_n$, so both measures in (6.23) depend on $X_1, \ldots, X_n$. Hence the total variation distance, let us denote it by Y_n, i.e.

   $$Y_n := \big\|P(\sqrt{n}(\theta - \theta_0) \in \cdot|X_1, \ldots, X_n) - \mathcal{N}(\sqrt{n}(\hat{\theta}_n - \theta_0), I^{-1}(\theta_0))\big\|_{TV}$$

   is a function of $X_1, \ldots, X_n$ – a random variable. The theorem claims that when $X_1, X_2, \ldots$ are iid random variables from $f(\cdot|\theta_0)$, then the expectation of these random variable converges to 0: $E_{\theta_0}Y_n \to 0$. Recall that E_{θ_0} indicates that $X_1, X_2, \ldots$ are iid random variables from $f(\cdot|\theta_0)$, i.e. the true parameter is θ_0.

3. "Bounded away from 0 in a neighborhood of θ_0" means that there exists a neighborhood of θ_0, say U, (i.e U is open and $\theta_0 \in U$) and an $\epsilon > 0$ such that $\pi(\theta) > \epsilon$, $\forall\theta \in U$. It holds when π is continuous at θ_0 and $\pi(\theta_0) > 0$.

4. The total variation distance is bounded by 1 and so $Y_n \leq 1$. For bounded random variables the convergence in probability is equivalent to the convergence to in expectation, i.e.

$$E_{\theta_0} Y_n \to 0 \quad \Leftrightarrow \quad P_{\theta_0}(Y_n > \epsilon) \to 0, \quad \forall \epsilon > 0. \tag{6.24}$$

This means that (6.23) is equivalent to the convergence to 0 in probability, and sometimes in the literature it is so stated.

5. Let $Z \sim \mathcal{N}(0, I^{-1}(\theta_0))$. Then

$$Y_n = \big\| P(\sqrt{n}(\theta - \theta_0) \in \cdot | X_1, \ldots, X_n) - P(Z + \sqrt{n}(\hat{\theta}_n - \theta_0) \in \cdot) \big\|_{TV}$$

From (6.20), (take $a = 1$ and $b = -\sqrt{n}(\hat{\theta}_n - \theta_0)$), it follows

$$Y_n = \big\| P(\sqrt{n}(\theta - \hat{\theta}_n) \in \cdot | X_1, \ldots, X_n) - P(Z \in \cdot) \big\|_{TV}. \tag{6.25}$$

Applying (6.20) again (now take $a = 1/\sqrt{n}$ and $b = \hat{\theta}_n$), (6.25) can be restated:

$$Y_n = \big\| P(\theta \in \cdot | X_1, \ldots, X_n) - \mathcal{N}(\hat{\theta}, I_n^{-1}(\theta_0)) \big\|_{TV}.$$

Bernstein-von Mieses theorem provides Bayesian counterpart of asymptotic normality of MLE. Indeed, (6.22) states:

$$\sqrt{n}\big(\hat{\theta}_n - \theta\big) | \theta_0 \Rightarrow \mathcal{N}\big(0, I^{-1}(\theta_0)\big)$$

and (6.25) states (with some reservation)

$$\sqrt{n}\big(\theta - \hat{\theta}_n\big) | X_1, \ldots, X_n \Rightarrow \mathcal{N}\big(0, I^{-1}(\theta_0)\big).$$

6. In [2, 3], the Y_n in the statement of Bernstein-von Mieses is defined differently:

$$Y_n = \| P(\sqrt{n}(\theta - \theta_0) \in \cdot | X_1, \ldots, X_n) - \mathcal{N}(\Delta_{n,\theta_0}, I^{-1}(\theta_0)) \|, \tag{6.26}$$

where

$$\Delta_{n,\theta_0} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I^{-1}(\theta_0) \frac{\partial \ln f(X_i|\theta_0)}{\partial \theta}.$$

Observe that (under regularity)

$$E_{\theta_0}\big[\frac{\partial \ln f(X_i|\theta_0)}{\partial \theta}\big] = 0, \quad \mathrm{Var}_{\theta_0}\big[\frac{\partial \ln f(X_i|\theta_0)}{\partial \theta}\big] = I(\theta_0).$$

Hence

$$\frac{1}{n} \sum_{i=1}^{n} I^{-1}(\theta_0) \frac{\partial \ln f(X_i|\theta_0)}{\partial \theta} \to 0, \quad \Delta_{n,\theta_0} \Rightarrow \mathcal{N}(0, I^{-1}(\theta_0)) \tag{6.27}$$

It can be shown that under θ_0

$$|\Delta_{n,\theta_0} - \sqrt{n}(\hat{\theta}_n - \theta_0)| \xrightarrow{P} 0,$$

and from this, it follows that under θ_0

$$\|\mathcal{N}(\Delta_{n,\theta_0}, I^{-1}(\theta_0)) - \mathcal{N}(\sqrt{n}(\hat{\theta}_n - \theta_0), I^{-1}(\theta_0))\|_{TV} \xrightarrow{P} 0.$$

Now, it is clear that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ in (6.23) can be replaced by Δ_{n,θ_0}.

For detailed assumptions, theorems and proof, see [2], Ch 12, [3], Ch 10, [4], 7.4.2.

6.6 Exercises

1. Show that when (6.4) holds for every component, then the posterior is strongly consistent i.e. (6.1) holds.

2. Prove that for any training sample $x_{(\ell)}$

   $$P_{(\ell)}(\cdot|x_{\ell+1}, \ldots, x_n) = P(\cdot|x_1, \ldots, x_n).$$

   Show that when π has an atom θ_0, and $f(X_{(\ell)}|\theta_0) > 0$, then θ_0 is an atom of $\pi_{(\ell)}$.

3. KL divergence. Prove (6.11),(6.12),(6.13), (6.14).

4. Prove that when (6.16) holds, there for every $\epsilon > 0$ and n there exists ϕ_n such that (6.15) holds.

7 Dirichlet process model

7.1 Non-parametric setting: preliminaries

The space $(\mathcal{P}, \mathcal{B}(\mathcal{P}), \pi)$. Let $\mathcal{X} = \mathbb{R}^d$, let $\mathcal{P}$ be the set of all probability measures on $\mathcal{X}$. In this section, we denote the elements of $\mathcal{P}$ as p. Recall the *weak convergence of probability measures* $p_n \Rightarrow p$ if and only if $\int f dp_n \to \int f dp$ for any $f \in C_b(\mathcal{X})$ (set of bounded continuous functions). By portmanteu theorem, this is one of the many equivalent definitions. The weak convergence is metrizable by many metrics, one of them is *Prokhorov distance*:

$$d(p, q) = \inf\{\epsilon > 0 : p(A) < q(A^\epsilon) + \epsilon, q(A) < p(A^\epsilon) + \epsilon\},$$

where $A^\epsilon := \{x : \|x - y\| < \epsilon, y \in A\}$. It can be shown that $(\mathcal{P}, d)$ is complete and separable (hence Polish) metric space ([2], Thm A3 or [7], Thms 1.9, 1.15). The distance d generates open sets and, hence, also the Borel σ-algebra $\mathcal{B}(\mathcal{P})$ on $\mathcal{P}$.

For every $A \in \mathcal{B}(\mathcal{X})$ and for every $f \in C_b(\mathcal{X})$, we define

$$T_A : \mathcal{P} \mapsto [0,1], \quad T_A(p) = p(A)$$
$$T_f : \mathcal{P} \mapsto \mathbb{R}, \quad T_f(p) = \int f dp.$$

It can be shown that $\mathcal{B}(\mathcal{P})$ is the smallest σ-algebra on $\mathcal{P}$ making all maps T_A measurable (for every measurable A). Also $\mathcal{B}(\mathcal{P})$ is the smallest σ-algebra on $\mathcal{P}$ making all maps T_f measurable (for every bounded continuous f). ([2], Prop. A5 or [7], Prop 1.16, 1.18). From that the following proposition can be deduced ([2], Prop. A5 or [7], Thm. 1.19 and 1.21):

Proposition 7.1

1. *When π and π' are two different probability measures on $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$, then there exists a finite measurable partition $A_1, \ldots, A_k$ of $\mathcal{X}$ and a k-dimensional Borel set $B \in \mathcal{B}(\mathbb{R}^k)$ such that*

$$\pi\{p : (T_{A_1}(p), \ldots, T_{A_k}(p)) \in B\} \neq \pi'\{p : (T_{A_1}(p), \ldots, T_{A_k}(p)) \in B\}.$$

2. *When π and π' are two different probability measures on $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$, then there exists a function $f \in C_b(\mathcal{X})$ and a Borel set $B \in \mathcal{B}(\mathbb{R})$ such that*

$$\pi\{p : \int f dp \in B\} \neq \pi'\{p : \int f dp \in B\}.$$

In other words, 1. states that any probability measure π on $\mathcal{B}(\mathcal{P})$ is completely determined by the set of distributions $p \to \big(p(A_1), \ldots, p(A_k)\big)$ for every measurable partition $A_1, \ldots, A_k$ of $\mathcal{X}$ (given π, every partition generates a k-dimensional random vector and the distributions of all these vectors uniquely determine π).
In other words, 2. states that any probability measure π on $\mathcal{B}(\mathcal{P})$ is completely determined by the set of distributions $p \to \int f dp$ for every bounded and continuous f (given π, every function generates a random variable and the distributions of all these random variables uniquely determine π).

Random measure and its law as the prior. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Recall $\mathcal{P}$ is the set of all probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and also recall that $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$ is a measurable space (set of all probability measures equipped with Borel σ-algebra induced by Prokhorov metric). *A random measure P on $\mathcal{X}$* is a measurable mapping $P : \Omega \to \mathcal{P}$. The measurability here means that for every $E \in \mathcal{B}(\mathcal{P})$, $P^{-1}(E) \in \mathcal{F}$. Since for every $A \in \mathcal{B}(\mathcal{X})$, T_A is $\mathcal{B}(\mathcal{P})$ measurable, we get that the composition

$$T_A \circ P : \quad \Omega \to [0,1], \quad T_A \circ P(\omega) = P(\omega)(A)$$

is $\mathcal{F}$-measurable, hence a random variable. On the other hand, since $\mathcal{B}(\mathcal{P})$ is the smallest σ-algebra on $\mathcal{P}$ making all maps T_A measurable, it follows that a mapping $P : \Omega \to \mathcal{P}$ is $\mathcal{F}$-measurable (hence random measure) in and only if the composition above is a random variable for every $A \in \mathcal{B}(\mathcal{X})$ (Exercise 1.) Hence the mapping $P(\omega, A) := P(\omega)(A)$ is a *Markov kernel*:

$$P(\cdot, A) \text{ is } \mathcal{F}\text{-measurable for every } A \in \mathcal{B}(\mathcal{X})$$
$$P(\omega, \cdot) \text{ is a measure on } \mathcal{B}(\mathcal{X}) \text{ for every } \omega.$$

Every random measure P induces a pushforward measure on $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$– the *law (distribution) of* P. Let that be π:

$$\pi(E) := \mathbf{P}(P^{-1}(E)), \quad \forall E \in \mathcal{B}(\mathcal{P}).$$

We shall write $P \sim \pi$. In non-parametric Bayesian setting, the measure π is the prior measure on the set of all probability measures $\mathcal{P}$. On the other hand, given a probability measure π on $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$, one can take the underlying probability space $(\Omega, \mathcal{F}, \mathbf{P})$ as $(\mathcal{P}, \mathcal{B}(\mathcal{P}), \pi)$ and define $P(p) = p$. Then the law of P is π. So any prior measure π can be considered as a law of (at least one) random measure. In Bayesian setting, the distribution of a random measure matters, and so the setting above, i.e. $(\Omega, \mathcal{F}, \mathbf{P}) = (\mathcal{P}, \mathcal{B}(\mathcal{P}), \pi)$ (hence $\omega = p$) is typically used.

By proposition 7.1, any probability measure π on $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$ is determined by the distributions of all vectors $(T_{A_1}, \dots, T_{A_k})$ corresponding to all measurable partitions $(A_1, \dots, A_k)$ of $\mathcal{X}$. Equivalently, the the distribution of any random measure P is determined by the laws of $(P(A_1), \dots, P(A_k))$.

For any random measure $P \sim \pi$, let

$$\mu(A) := E[P(A)] = \int p(A)\pi(dp).$$

Clearly (Exercise 2) that μ is a measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ – *mean measure*.

7.2 Diriclet process

Definition. Let α be a finite measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. A random measure P on $\mathcal{X}$ is a *Dirichlet process with base measure* α if for any finite measurable partition $A_1, \dots, A_k$ of $\mathcal{X}$,

$$(P(A_1), \dots, P(A_k)) \sim \mathrm{Dir}(\alpha(A_1), \dots, \alpha(A_k)). \tag{7.1}$$

We know that (7.1) uniquely defines the distribution of Dirichlet process: if $P \sim \pi$ and $P' \sim \pi'$ are two different Dirichlet distributions with the same base measure (both satisfying (7.1) for all finite measurable partitions), then

$\pi = \pi'$. In what follows, the distribution of a Dirichlet process with base measure α will be denoted by $\mathrm{DP}(\alpha)$. Hence, if P is a Dirichlet process with base measure α, we shall write $P \sim \mathrm{DP}(\alpha)$.

We write $|\alpha| := \alpha(\mathcal{X}) < \infty$ for the total mass of α and we define $\bar{\alpha} := \alpha/|\alpha|$. The measure $\bar{\alpha}$ – often refereed to as *center measure* – is a probability measure.

For the proof of the existence of Dirichlet process see ([2], Sec 4.2).

Moment properties of Dirichlet process. Let $P \sim \mathrm{DP}(\alpha)$. Then for any two measurable sets $A, B \in \mathcal{B}(\mathcal{X})$ and for any two measurable functions $f, g : \mathcal{X} \to \mathbb{R}$ having the needed integrals,

1. $P(A) \sim \mathrm{Beta}(\alpha(A), \alpha(A^c))$;
2. $\mu(A) = EP(A) = \bar{\alpha}(A)$;
3. $\mathrm{Var}(P(A)) = \frac{\bar{\alpha}(A)\bar{\alpha}(A^c)}{1+|\alpha|}$;
4. $\mathrm{Cov}(P(A), P(B)) = \frac{\bar{\alpha}(A\cap B)-\bar{\alpha}(A)\bar{\alpha}(A^c)}{1+|\alpha|}$;
5. $E[\int f dP] = \int f d\bar{\alpha}$;
6. $\mathrm{Var}[\int f dP] = \frac{\int (f-\int f d\bar{\alpha})^2 d\bar{\alpha}}{1+|\alpha|}$;
7. $\mathrm{Cov}[\int f dP, \int f dP] = \frac{\int f g d\bar{\alpha} - \int f d\bar{\alpha} \int g d\bar{\alpha}}{1+|\alpha|}$.

Assertions 1-3 are immediate from the definition and the properties of Dirichlet distribution (Exercise 4). The proof of 4 can be found in ([2], Prop. 4.2). The proofs of 5,6,7 follow from the standard machinery argument when noticing that with $f = I_A$ and $g = I_B$, the properties 5,6,4 are the same as the properties 2,3,4, (resp.) (see [2], Prop 4.3).

Properties 2 and 3 show that the bigger is $|\alpha|$, the smaller the variance and the more the Dirichlet process is concentrated around its mean $\bar{\alpha}$. This justifies calling $M := |\alpha|$ as the *precision parameter* and rewriting $\alpha = M\bar{\alpha}$. Thus when specifying the prior, one can start with center measure $\bar{\alpha}$ and then choosing the precision parameter M reflecting the degree of confidence. The properties 2 and 3 also show that when $\alpha \neq \alpha'$, then $\mathrm{DP}(\alpha) \neq \mathrm{DP}(\alpha')$: either means or variances differ.

Stick-breaking representation. *Stick-breaking* is a technique to construct a random probability measure on countable set i.e. a random sequence $(W_1, W_2, \ldots)$ of non-negative random variables summing a.s. up to 1. Let

$V_1, V_2, \ldots$ be a sequence of random variables $V_i \in [0,1]$ for every i. Define new random variables $W_1, W_2, \ldots$ as follows

$$W_1 = V_1, \quad W_2 = (1-V_1)V_2, \quad W_3 = (1-V_1)(1-V_2)V_3, \cdots W_k = \prod_{i=1}^{k-1}(1-V_i)V_k. \tag{7.2}$$

The interpretation: a stick with length 1 is broken at a point given by V_1. The first weight is $W_1 = V_1$. The length of remaining stick is $1 - V_1$ and it is broken into two pieces of relative lengths V_2 and $1 - V_2$. The second weight is the first part of it, thus $W_2 = (1 - V_1)V_2$. The length of the remaining stick (after two breakings) is $(1 - V_1)(1 - V_2)$ that is broken into two pieces of relative lengths V_3 and $1 - V_3$, so that $W_3 = (1 - V_1)(1 - V_2)V_3$ and so on. It can be shown that when $V_1, V_2, \ldots$ are iid random variables such that $P(V_1 > 0) > 0$, then $\sum_{i=1}^{\infty} W_i = 1$ a.s. ([2], Lemma 3.4).

It turns out that the stick-breaking construction can be used to construct a Dirichlet process. Let $\alpha = M\bar{\alpha}$ be a base measure.

Theorem 7.1 (Sethuraman) *Let $\theta_1, \theta_2, \ldots$ be iid random variables, $\theta_i \sim \bar{\alpha}$ and $V_1, V_2, \ldots$ are iid random variables, $V_i \sim \mathrm{Beta}(1, M)$, then*

$$\sum_{j=1}^{\infty} W_j \delta_{\theta_j} \sim DP(\alpha),$$

where $W_1, W_2, \ldots$ are as in (7.2).

For the proof see ([2], Thm 4.12). Since

$$EW_k = \frac{M^{k-1}}{(1+M)^k},$$

we see that the bigger M the more uniform (and smaller) are the masses $W_1, W_2, \ldots$; recall also that the variance of W_i decreases when M increases.

The stick-breaking construction gives an easy method to simulate a Dirichlet process, at least approximatively. It also shows that almost every realization of $\mathrm{DP}(\alpha)$ is a discrete measure:

$$\mathrm{DP}(\alpha)\{p \in \mathcal{P} : p \text{ is discrete}\} = 1.$$

So the set of all discrete measures is a subset of $\mathcal{P}$ having $\mathrm{DP}(\alpha)$ measure 1. Recall the support of any probability measure π on $\mathcal{P}$ is the smallest closed set (in Prokhorov metric) having π-measure 1. It turns out that the support of $\mathrm{DP}(\alpha)$ is much larger set than just discrete measures – it can be shown ([2], Thm 4.15) that the support of $\mathrm{DP}(\alpha)$ is the set

$$\{p \in \mathcal{P} : \mathrm{supp}(p) \subset \mathrm{supp}(\bar{\alpha})\}.$$

Hence if $\bar{\alpha}$ or equivalently α is fully supported on $\mathcal{X}$ ($\mathrm{supp}(\bar{\alpha}) = \mathcal{X}$), then the support of $\mathrm{DP}(\alpha)$ is $\mathcal{P}$.

7.3 Dirichlet process model

With $\mathrm{DP}(\alpha)$ as the prior measure, we get the non-parametric version of our basic model (1.2):

$$\begin{aligned} P|\alpha &\sim \mathrm{DP}(\alpha) \\ X_1, \dots, X_n | P = p &\overset{i.i.d.}{\sim} p. \end{aligned} \tag{7.3}$$

With some abuse of terminology, the random variables $X_1, \dots, X_n$ is called *a sample from Dirichlet process.*

Conjugacy. We now argue that DP is conjugate prior, i.e. the posterior distribution is DP as well. Let $x = (x_1, \dots, x_n)$ be a realization of $(X_1, \dots, X_n)$ defined as in (7.3). Let $A_1, \dots, A_k$ be a partition of $\mathcal{X}$ and let $(N_1, \dots, N_k)$ be the cell-counts, i.e. $N_j = \sum_{i=1}^n I_{A_j}(X_i)$, $j = 1, \dots, k$. Hence, given $A_1, \dots, A_k$, we end up with Dirichlet-multinomial model:

$$\begin{aligned} \big(P(A_1), \dots, P(A_k)\big)|\alpha &\sim \mathrm{Dir}(\alpha(A_1), \dots, \alpha(A_k)) \\ (N_1, \dots, N_k)|\big(P(A_1), \dots, P(A_k)\big) = p(A_1), \dots, p(A_k)\big) &\sim \mathrm{Multin}\big(n, (p(A_1), \dots, p(A_k)\big). \end{aligned}$$

We know that Dirichlet distribution is conjugate prior for multinomial distribution, hence given a realization $(n_1, \dots, n_k)$ of $(N_1, \dots, N_k)$, (i.e. $n_j = \sum_{i=1}^n I_{A_j}(x_i)$, $j = 1, \dots, k$), the posterior distribution of $\big(P(A_1), \dots, P(A_k)\big)$ is Dirchlet distribution as well:

$$\big(P(A_1), \dots, P(A_k)\big)|n_1, \dots, n_k \sim \mathrm{Dir}(\alpha(A_1) + n_1, \dots, \alpha(A_k) + n_k).$$

It can be shown ([2], Thm 4.6) that the distribution $\big(P(A_1), \dots, P(A_k)\big)|n_1, \dots, n_k$ is the same as the distribution of $\big(P(A_1), \dots, P(A_k)\big)|x_1, \dots, x_k$. Thus for every partition $A_1, \dots, A_k$,

$$\big(P(A_1), \dots, P(A_k)\big)|x \sim \mathrm{Dir}\big(\alpha(A_1) + \sum_{i=1}^n I_{A_1}(x_i), \dots, \alpha(A_k) + \sum_{i=1}^n I_{A_k}(x_i)\big).$$

Since for any $A \in \mathcal{B}(\mathcal{X})$

$$\alpha(A) + \sum_{i=1}^n I_A(x_i) = \alpha(A) + \sum_{i=1}^n \delta_{x_i}(A),$$

we conclude that

$$P|x \sim \mathrm{DP}(\alpha)\Big(\alpha + \sum_{i=1}^n \delta_{x_i}\Big).$$

With $(M, \bar{\alpha})$-parametrization, the posterior updating $\alpha \mapsto \alpha + \sum_{i=1}^n \delta_{x_i}$ is the following

$$M \mapsto M + n, \quad \bar{\alpha} \mapsto \frac{M}{M+n}\bar{\alpha} + \frac{n}{n+M}P_n,$$

where

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$

is the empirical measure. Hence the center measure of posterior distribution is the convex combination of prior center measure and empirical measure, empirical measure becomes more relevant as n grows. We see that M can be interpreted as "prior sample size". For any (Borel) A, we get thus from moment-properties of DP (here $X = (X_1, \ldots, X_n)$)

$$E(P(A)|X = x) = \frac{M}{M+n}\bar{\alpha}(A) + \frac{n}{n+M}P_n(A);$$
$$\mathrm{Var}(P(A)|X = x) = \frac{E(P(A)|X = x)E(P(A^c)|X = x)}{1+M+n} \leq \frac{1}{4(1+M+n)}.$$

Marginal and predicative distributions. The marginal distribution of X_i is $\bar{\alpha}$, because (moment property 2 of DP) with P_X denoting the marginal distribution of X

$$P_X(A) = \int p(A) d\mathrm{DP}(\alpha)(p) = EP(A) = \bar{\alpha}(A).$$

Hence $X_1 \sim \bar{\alpha}$. Let us now find the distribution of $X_2|X_1 = x_1$,denoted by $P_{X_2|x_1}$. Clearly,

$$X_2|P = p, X_1 = x_1 \sim p.$$

Since $P|X_1 = x_1 \sim \mathrm{DP}(\alpha + \delta_{x_1})$, we get

$$\begin{aligned} P_{X_2|x_1}(A) &= \int p(A) d\mathrm{DP}(\alpha + \delta_{x_1})(p) = E(P(A)|X_1 = x_1) \\ &= \frac{M}{M+1}\bar{\alpha}(A) + \frac{1}{1+M}P_1(A). \end{aligned}$$

Thus

$$X_2|X_1 = x_1 \sim \frac{M}{M+1}\bar{\alpha} + \frac{1}{M+1}\delta_{x_1}.$$

Repeating this argument, we obtain

$$\begin{aligned} P_{X_{n+1}|X_1=x_1,\ldots,X_n=x_n}(A) &= E(P(A)|X_1 = x_1, \ldots, X_n = x_n) \\ &= \frac{M}{M+n}\bar{\alpha}(A) + \frac{n}{n+M}P_n(A) \end{aligned}$$

or, equivalently,

$$X_{n+1}|X_1 = x_1, \ldots, X_n = x_n \sim \frac{\alpha + \sum_{i=1}^{n}\delta_{x_i}}{n+M} = \frac{M}{M+n}\bar{\alpha} + \frac{n}{n+M}P_n. \tag{7.4}$$

Hence, the predicative distribution is the center (or mean) measure of posterior.

Generalized Polya urn scheme. The formula (7.4) gives an easy algorithm to generate a sample from Dirichlet process:

- generate X_1 from $\bar{\alpha}$, i.e. $X_1 \sim \bar{\alpha}$;
- given $X_1 = x_1$ generate X_2 as follows:

  with probability $\frac{1}{M+1}$ generate take $X_2 = x_1$,
  with probability $\frac{M}{M+1}$ generate X_2 from $\bar{\alpha}$;
- given $X_1 = x_1, X_2 = x_2$ generate X_3 as follows:

  with probability $\frac{1}{M+2}$ take $X_3 = x_1$,
  with probability $\frac{1}{M+2}$ take $X_3 = x_2$
  with probability $\frac{M}{M+2}$ generate X_3 from $\bar{\alpha}$;
- $\cdots$
- given $x_1, \ldots, x_n$ generate X_{n+1} is as follows:

  with probability $\frac{1}{M+n}$ take $X_{n+1} = x_i$ $(i = 1, \ldots, n)$
  with probability $\frac{M}{M+n}$ generate X_{n+1} from $\bar{\alpha}$.

Let $x_1^*, \ldots, x_k^*$ be distinct values in $x_1, \ldots, x_n$ with $n_1, \ldots, n_k$ being the repetitions. Then the same rule for generating X_{n+1} is the following:

with probability $\frac{n_j}{M+n}$ take $X_{n+1} = x_j^*$ $(j = 1, \ldots, k)$
with probability $\frac{M}{M+n}$ generate X_{n+1} from $\bar{\alpha}$.

Hence there is a clear analogue with Polya urn scheme: initially, in the urn there are M balls with "colors"distributed as $\bar{\alpha}$. After the first choice there are $M+1$ balls in urn and the colors are distributed accordingly $\frac{M}{M+1}\bar{\alpha}(A) + \frac{1}{1+M}\delta_{x_1}$. Proceeding like that, after n-th choice there are n additional balls in the urn (so $M+n$ balls in total) and the additional balls are of k different colors: n_j balls are of color x_j^* $(j = 1, \ldots, k)$. Thus the distribution of colors after n-th draw is

$$\frac{M}{M+n}\bar{\alpha} + \frac{\sum_{i=1}^n \delta_{x_i}}{n+M} = \frac{M}{M+n}\bar{\alpha} + \sum_{j=1}^k \frac{n_j}{n+M}\delta_{x_j^*}.$$

The described algorithm is known as generalized Polya urn scheme. When $\bar{\alpha}$ is non-atomic, then every draw from $\bar{\alpha}$ gives a new distinct value (new color) to the sample.

Number of distinct values. Let $X_1, \ldots, X_n$ be a sample from Dirichlet process (or from generalized Polya urn scheme) with *atomless* base measure α. There are clearly ties in the sample, let K_n be the number of distinct values among $X_1, \ldots, X_n$. Following ([2], Prop 4.8) we define D_i as a Bernoulli random variable, with $D_i = 1$ if X_i is the new value, i.e. $X_i \notin \{X_1, \ldots, X_{i-1}\}$. Thus $D_1 \equiv 1$ and from Polya urn scheme, it follows (recall α is atomless, i.e. every value generated from $\bar{\alpha}$ is new) $P(D_i = 1) = M/(M+i-1) = E(D_i)$. Clearly the random variables D_i are independent and clearly $K_n = \sum_{i=1}^n D_i$. Thus the expected number of distinct values and the variance of that number are:

$$E(K_n) = \sum_{i=1}^{n} \frac{M}{M+i-1}, \quad \mathrm{Var}(K_n) = \sum_{i=1}^{n} \frac{M(i-1)}{(M+i-1)^2}. \tag{7.5}$$

Since $\sum_{i=1}^n \frac{1}{i} \sim \ln n$ ($a_n \sim b_n$ means $a_n/b_n \to 1$), it follows that

$$E(K_n) \sim M \ln n, \quad \mathrm{Var}(K_n) \sim \ln n.$$

Since $\sum_i \frac{1}{i(\ln i)^2} < \infty$, we get that

$$\sum_{i=1}^{\infty} \frac{\mathrm{Var}(D_i)}{(\ln i)^2} = \sum_{i=1}^{n} \frac{M(i-1)}{(M+i-1)^2 (\ln i)^2} < \infty,$$

by SLLN (Kolmogorov II thm)

$$\frac{\sum_{i=1}^n (D_i - ED_i)}{\ln n} \to 0, \quad \text{a.s.} \quad \Rightarrow \quad \frac{K_n}{\ln n} \to M, \quad \text{a.s.}$$

(the implication above follows from $E(K_n) \sim M \ln n$).

The random variables D_i are bounded (Bernoulli) with $\sum_{i=1}^{\infty} \mathrm{Var}(D_i) = \infty$. Then CLT holds, i.e.

$$\frac{K_n - E(K_n)}{\sqrt{\mathrm{Var}(K_n)}} = \frac{\sum_{i=1}^n (D_i - ED_i)}{\sqrt{\sum_{i=1}^n \mathrm{Var}(D_i)}} \Rightarrow \mathcal{N}(0,1).$$

Hence, asymptotically K_n behaves like $M \ln n$, for the exact distribution of K_n see ([2], prop. 4.9).

Chinese restaurant process. A random sample $X_1, \ldots, X_n$ from Dirichlet process induces a random partition of $\{1, \ldots, n\}$ – equivalence classes $i \sim j$ if and only if $X_i = X_j$. Hence K_n is the number sets (equivalence classes) in the partition. Let $(S_1, \ldots, S_k)$ be a partition of $\{1, \ldots, n\}$ with $(n_1, \ldots, n_k)$ being the sizes of the sets (i.e. $|S_i| = n_i$). Suppose, for a moment, that the partition is the following:

$$S_1 = \{1, \ldots, n_1\}, S_2 = \{n_1{+}1, \ldots, n_1{+}n_2\}, \ldots, S_k = \{n_1{+}\cdots{+}n_{k-1}{+}1, \ldots, n\}.$$

From Polya urn scheme, it is evident that the probability of that partition is (α is atomless)

$$\underbrace{\frac{M}{M}\frac{1}{M+1}\cdots\frac{n_1-1}{M+n_1-1}}_{S_1}\cdot\underbrace{\frac{M}{M+n_1}\frac{1}{M+n_1+1}\cdots\frac{n_2-1}{M+n_1+n_2-1}}_{S_2}\cdots$$
$$\cdots\underbrace{\frac{M}{M+n_1+\cdots+n_{k-1}}\frac{M+1}{M+n_1+\cdots+n_{k-1}+1}\cdots\frac{n_k-1}{M+n-1}}_{S_k}=$$
$$\frac{M^k(n_1-1)!(n_2-1)!\cdots(n_k-1)!}{M(M+1)\cdots(M+n-1)}=\frac{M^k\prod_{i=1}^k\Gamma(n_i)\Gamma(M)}{\Gamma(n+M)}. \tag{7.6}$$

Form Polya urn scheme, it also follows that for any other partition with the same set sizes $(n_1,\ldots,n_k)$, the probability is the same (see also [2], Prop 4.11).

The sequence of random partitions corresponding to the random samples $X_1, X_2\ldots$ from Dirichlet process with atomless α is called the *Chinese restaurant process*. To explain the name, we quote ([2], sec 14.11): "The name derives from the following metaphor. Suppose that customers arrive sequentially in a Chinese restaurant with an infinite number of tables, each with infinite seating capacity. The first customer chooses an arbitrary table. The second customer has two options, sit at the table opened by customer 1 or open a new table, between which he decides with probabilities $1/(M+1)$ and $M/(M+1)$. More generally, the $(n+1)$st customer finds n customers seated at k tables in groups of $n_1,\ldots,n_k$, where $\sum_{j=1}^k n_j = n$, and chooses to sit at the jth open table with probability $n_j/(M+n)$, or open a new table with probability $M/(M+n)$. The gravitational effect of this scheme – more massive tables, apparently with more known faces, attract a newcomer with a higher probability – is valuable for clustering variables together in groups."

With Chinese restaurant process model, a sample $X_1,\ldots,X_n$ from Dirichlet process with atomless α can be constructed via random partition as follows: Generate a random partition from distribution (7.6). Every set in this partition shares a common X_j^* distributed as $\bar{\alpha}$ independently of other common values. Then take $X_i = X_j^*$ where $i \in S_j$.

7.4 Dirichlet process mixtures

Recall our parametric model $\{f(\cdot|\theta)\}$, where $\Theta \subset \mathbb{R}^d$. For any probability measure (prior) $p \in \mathcal{P}$, we defined the marginal density

$$f_p(x) := \int f(x|\theta)p(d\theta). \tag{7.7}$$

In statistics, densities like that are called *mixture densities* (segujaotused) . The measure p is called the *mixing measure*, when it is discrete with finite number of atoms, i.e. $p = \sum_i^k p_i \delta_{\theta_i}$ then the corresponding mixture – called as the *finite mixture* – is

$$f_p(x) = \sum_{i=1}^{k} f(x|\theta_i) p_i.$$

Clearly, an i.i.d sample $X_1, \ldots, X_n$ form the mixture density $f_p(\cdot)$ as in (7.7) can be obtained as follows:

$$\begin{aligned} \theta_1, \ldots, \theta_n | p &\overset{i.i.d.}{\sim} p \\ X_i | \theta_i &\overset{ind}{\sim} f(\cdot|\theta_i), \quad i = 1, \ldots, n. \end{aligned}$$

Here the use of *latent variables* $\theta_1, \ldots, \theta_n$ often simplifies the analysis. Mixtures form a flexible and large class of densities and in Bayesian setting, the mixing measures are considered random, i.e. the prior π is on $\mathcal{P}$. Hence P is a random probability measure and then

$$f_P(x) := \int f(x|\theta) P(d\theta)$$

is a random density. Hence we have the following Bayesian model:

$$\begin{aligned} P &\sim \pi \\ X_1, \ldots, X_n | P = p &\overset{i.i.d.}{\sim} f_p(\cdot) \end{aligned}$$

With the latent variables, the same model can be represented as the following hierarchical model:

$$\begin{aligned} P &\sim \pi \\ \theta_1, \ldots, \theta_n | P = p &\overset{i.i.d.}{\sim} p \\ X_i | \theta_i &\overset{ind}{\sim} f(\cdot|\theta_i), \quad i = 1, \ldots, n. \end{aligned} \tag{7.8}$$

When $\pi = \mathrm{DP}(\alpha)$, the model above is known as *Dirichlet process mixture* model. From Sethuraman representation, it follows that the random mixture f_P is then a.s. discrete with infinitely many atoms:

$$f_P(x) = \sum_{j=1}^{\infty} f(x|\theta_j) W_j,$$

where $\theta_1, \theta_2, \ldots$ are iid random variables from $\theta_i \sim \bar{\alpha}$; $W_1, W_2, \ldots$ are the stick-breaking weights as in (7.2).

Dirichlet process mixtures in density estimation. Dirichlet process mixtures have may applications, one of them is density estimation. With Dirichlet process P, one defines random density $f_P(\cdot)$ and given the observation $x = (x_1, \ldots, x_n)$, a meaningful point-estimate of unknown density would be (here $X = (X_1, \ldots, X_n)$)

$$\hat{f}(t) = E[f_P(t)|X = x].$$

Observe that

$$\begin{aligned} E[f_P(t)|X = x] &= E\big[E[f_P(t)|\theta_1, \ldots, \theta_n, X = x]\big|X = x\big] \\ &= E\big[E[f_P(t)|\theta_1, \ldots, \theta_n]\big|X = x\big], \end{aligned}$$

because given the latent variables, the posterior distribution of P is independent of x. Now, since $\theta_1, \ldots, \theta_n$ are the observations of Dirichlet process, we know that

$$P|\theta_1, \ldots, \theta_n \sim \mathrm{DP}\Big(\alpha + \sum_{i=1}^{n} \delta_{\theta_i}\Big).$$

Then (the 5.-th moment property of DP),

$$\begin{aligned} E[f_P(t)|\theta_1, \ldots, \theta_n] &= E\Big[\int f(t|\theta)P(d\theta)\Big|\theta_1, \ldots, \theta_n\Big] \\ &= \int f(t|\theta)\Big(\frac{\alpha + \sum_{i=1}^{n} \delta_{\theta_i}}{M+n}\Big)(d\theta) \\ &= \frac{1}{M+n}\Big(\int f(t|\theta)\alpha(d\theta) + \sum_{i=1}^{n} f(t|\theta_i)\Big) \\ &= \frac{M}{M+n}\int f(t|\theta)\bar{\alpha}(d\theta) + \frac{1}{M+n}\sum_{i=1}^{n} f(t|\theta_i). \end{aligned}$$

Hence the estimate is

$$\hat{f}(t) = \frac{M}{M+n}\int f(t|\theta)\bar{\alpha}(d\theta) + \frac{1}{M+n}\sum_{i=1}^{n} E[f(t|\theta_i)|X = x].$$

The integral $\int f(t|\theta)\bar{\alpha}(d\theta) = f_{\bar{\alpha}}(t) = E[f_P(t)]$ is the prior expectation and M shows how much we believe that the unknown density is (close to) $f_{\bar{\alpha}}$. The analytic form of $\hat{f}(t)$ is known ([2], prop. 5.2), but it is not very practical.

Due to the Polya urn scheme, it is rather easy to apply Gibbs samplers to generate a sample $\theta_1, \ldots, \theta_n$ from the model (7.8) *given* $X = x$. There are many algorithms in literature, see ([2], thm. 5.3, [5], sec 23.3; [8], sec 3). Given m samples $(\theta_1^{(l)}, \ldots, \theta_n^{(l)})$, $l = 1, \ldots, m$, the conditional expectation $E[f(t|\theta_i)|X = x]$ can be estimated by

$$\hat{E}[f(t|\theta_i)|X = x] = \frac{1}{m}\sum_{l=1}^{m} f(t|\theta_i^{(l)}).$$

In other words, after sampling we have mn densities $f(\cdot|\theta_i^{(l)})$ (they are not all different, because every sample has many repeating elements), after averaging them, we get purely empirical estimate

$$\hat{f}_{mn}(\cdot) := \frac{1}{mn}\sum_{i=1}^{n}\sum_{l=1}^{m} f(\cdot|\theta_i^{(l)}),$$

the final estimate is the convex combination of prior mean $f_{\bar{\alpha}}$ and empirical estimate $\hat{f}_{mn}$:

$$\hat{f} = \frac{M}{M+n} f_{\bar{\alpha}} + \frac{n}{M+n}\hat{f}_{mn}.$$

Dirichlet process mixtures in clustering. Dirichlet process mixtures have may applications, one of them is modeling clusters. Consider (7.8). Since the latent variables $\theta_1, \theta_2, \ldots$ can be considered as the observations from Dirichlet process, they form (random) partitions – *clustering* – a clustering $\mathcal{S} = \{S_1, \ldots, S_k\}$ occurs with probability (7.6):

$$Q(\mathcal{S}) = M^k \prod_{j=1}^{k} (|S_j| - 1)! \cdot \frac{\Gamma(M)}{\Gamma(M+n)}. \tag{7.9}$$

Every cluster shares a common latent variable θ_j, $j = 1, \ldots, k$ and the variables are iid from atomless ((7.9) assumes atomless base-measure) $\bar{\alpha}$. Given the clustering $\mathcal{S} = \{S_1, \ldots, S_k\}$ and $\theta_1, \ldots, \theta_k$, the observations $X_1, \ldots, X_n$ are independent, $X_i \sim f(\cdot|\theta_j)$ when $i \in S_j$. Hence, inside a cluster S_j, the random variables $\{X_i : i \in S_j\}$ are iid, the variables belonging to different clusters are independent. Formally, (7.8) can equivalently stated as follows:

$$\begin{aligned} \mathcal{S} = (S_1, \ldots, S_{K_n}) &\sim Q \\ \theta_1, \ldots, \theta_k | K_n = k &\overset{i.i.d.}{\sim} \bar{\alpha} \\ X_i | \theta_1, \ldots, \theta_k, \mathcal{S} = \{S_1, \ldots, S_k\} &\overset{ind}{\sim} \sum_{j=1}^{k} f(\cdot|\theta_j) I_{S_j}(i), \quad i = 1, \ldots, n. \end{aligned} \tag{7.10}$$

Given $x = (x_1, \ldots, x_n)$, the main interest now is the posterior distribution of clusters $P(\mathcal{S}|X = x)$ (here $X = (X_1, \ldots, X_n)$) or some other clustering related questions like $P(K_n|X = x)$ (posterior distribution of the number of clusters) or finding the posterior probability that the observations x_{i_1}, x_{i_2} are from the same cluster. i.e. i_1, i_2 belong to the same cluster etc. Like in the density estimation, these posterior distributions are estimated via Gibbs sampling. One can generate the samples $(\theta_1^{(l)}, \ldots, \theta_n^{(l)})$, $l = 1, \ldots, m$, and read the clusters from the samples: there are typically $k < n$ distinct values in $(\theta_1^{(l)}, \ldots, \theta_n^{(l)})$ and i_1, i_2 belong to the same cluster whenever they share the common value: $\theta_{i_1}^{(l)} = \theta_{i_2}^{(l)}$. However, there also algorithms that generate only the distinct values in a sample and corresponding clusters ([2], sec 5.2, [5], sec. 23.3, [8], sec. 3).

Exercises.

1. Let $(\Omega, \mathcal{F})$ be a measurable space. Show that a mapping $P : \Omega \to \mathcal{P}$ is $\mathcal{F}$-measurable if and only if $\omega \to P(\omega)(A)$ is $\mathcal{F}$-measurable for every $A \in \mathcal{B}(\mathcal{X})$.

2. Prove that μ is a measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.

3. Let $g : \mathcal{X} \to \mathbb{R}$ be a Borel-measurable mapping. Let $P \sim \mathrm{DP}(\alpha)$ be a Dirichlet process on $\mathcal{X}$. Define a random measure Q on $\mathbb{R}$ by $Q = P \circ g^{-1}$, i.e. $Q(B) = P(g^{-1}(B))$, $B \in \mathcal{B}(\mathbb{R})$. Prove: $Q \sim \mathrm{DP}(\alpha \circ g^{-1})$.

4. Prove the properties 1–3.

Viited

[1] J. Miller,
A detailed treatment of Doob's theorem,
arXiv preprint arXiv:1801.03122, 2018

[2] S. Ghosal, A. van der Vaart,
Fundamentals of nonparametric Bayesian inference,
Cambridge University Press, 2017

[3] A. van der Vaart,
Asymptotic statistics,
Cambridge University Press, 2006

[4] M. Schervish,
Theory of Statistics,
Springer, 1997

[5] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari,
Bayesian data analysis,
CRC, 2013.

[6] C. Robert,
The Bayesian choice,
Springer, 2007

[7] M. Raihhelgauz
Dirichlet protsess,
TÜ, 2020 (bakalureusetöö).

[8] M. Raihhelgauz
Bayesi segumudelid tiheduse hindamseks,
TÜ, 2022 (magistritöö).