More on Continuous Random Variables

- Exponential family of distributions
- Bayesian Inference and Conjugate Priors

Textbook: Hisashi Kobayashi, Brian L. Mark and William Turin, ***Probability, Random Processes and Statistical Analysis*** (Cambridge University Press, 2012)

4.4 Exponential Family of Distributions

A family of PDFs (or PMFs) of the form

$$\boxed{f_{\mathbf{X}}(\boldsymbol{x};\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^\top(\boldsymbol{\theta})\boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\theta})\},} \tag{4.126}$$

is called an **exponential family**. The function ***T***(*x*) is called the **sufficient statistic**.

$$\boxed{f_{\mathbf{X}}(\boldsymbol{x};\boldsymbol{\eta}) = h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^\top\boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\eta})\},} \tag{4.127}$$

is called the **canonical** (or **natural**) **exponential family**.

- The exponential family of distributions includes the exponential, gamma, normal, Poisson, binomial distributions, etc.

Example 4.3: Normal distribution. Consider a normal RV $X \sim N(\mu, \sigma^2)$. With $\boldsymbol{\theta} = (\mu, \sigma)$ we write the PDF of each sample $x_i \ \ (i = 1, 2, \dots)$ as

$$\begin{aligned} f(x_i; \boldsymbol{\theta}) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2\sigma^2} + \frac{x_i \mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log \sigma\right), \ \ i = 1, 2, \dots, n. \end{aligned}$$

As in the previous example, we can present the normal distribution in the canonical exponential family form by identifying

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} \\ \frac{\mu}{\sigma^2} \end{bmatrix}, \ \ \boldsymbol{T}(x) = \begin{bmatrix} -\frac{x^2}{2} \\ x \end{bmatrix},$$

$$h(\boldsymbol{X}) = \frac{1}{\sqrt{2\pi}}, \ \ A(\boldsymbol{\eta}) = \frac{\mu^2}{2\sigma^2} + \log \sigma.$$

We can write the original parameter as $\boldsymbol{\theta} = (\mu, \sigma^2)$, where $\mu = \frac{\eta_2}{\eta_1}$ and $\sigma^2 = \frac{1}{\eta_1}$. Hence,

$$A(\boldsymbol{\eta}) = \frac{\eta_2^2}{2\eta_1} - \frac{\log \eta_1}{2}.$$

4.5 Bayesian Inference and Conjugate Priors

- Suppose that an observed sample *X* is drawn from a certain family of distributions specified by parameter θ.
- The Bayesian treats this parameter as a RV Θ, which is assigned a **prior** PDF **$\pi(\theta)=f_\Theta(\theta)$**.
- If RV *X* is a discrete RV, we have from Bayes' theorem (2.63)

$$\boxed{\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)},} \tag{4.133}$$

where $p(x) = \sum_\theta p(x|\theta)\pi(\theta)$. $p(x) = \int_\theta p(x|\theta)\pi(\theta)d\theta.$

- If the RV *X* is a continuous RV,

$$\boxed{\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)},} \tag{4.134}$$

where $f(x) = \int_\theta f(x|\theta)\pi(\theta)\, d\theta$. $f(x) = \sum_\theta f(x|\theta)\pi(\theta)$

- The conditional PDF $f(x|\theta)$ is called the **likelihood function**, when it is viewed as a function of θ with given *x*, and is denoted as

$$L_x(\theta) = f(x|\theta) \text{ or } L_x(\theta) = p(x|\theta), \tag{4.135}$$

- Then the posterior distribution can be written as

$$\pi(\theta|x) \propto L_x(\theta)\pi(\theta). \tag{4.137}$$

- For certain choices of the prior distribution, the posterior distribution has the same mathematical form as the prior distribution. Such prior distribution is called a **conjugate prior** (**distribution**) of the given likelihood function.

Example 4.4: The Bernoulli distribution and its conjugate prior, the beta distribution

- Write the probability of success as θ (instead of *p*).
- Define the binary variable X_i which takes on 1 or 0, depending on the *i*th trial is a success (s) or failure (f).
- Then, we can write $p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}.$
- For n independent trials we observe the data $\boldsymbol{x} \triangleq (x_1, x_2, \ldots, x_n)^\top$

  The likelihood function of θ given ***x*** is

$$L_{\boldsymbol{x}}(\theta) = p(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$
$$= \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}. \tag{4.139}$$

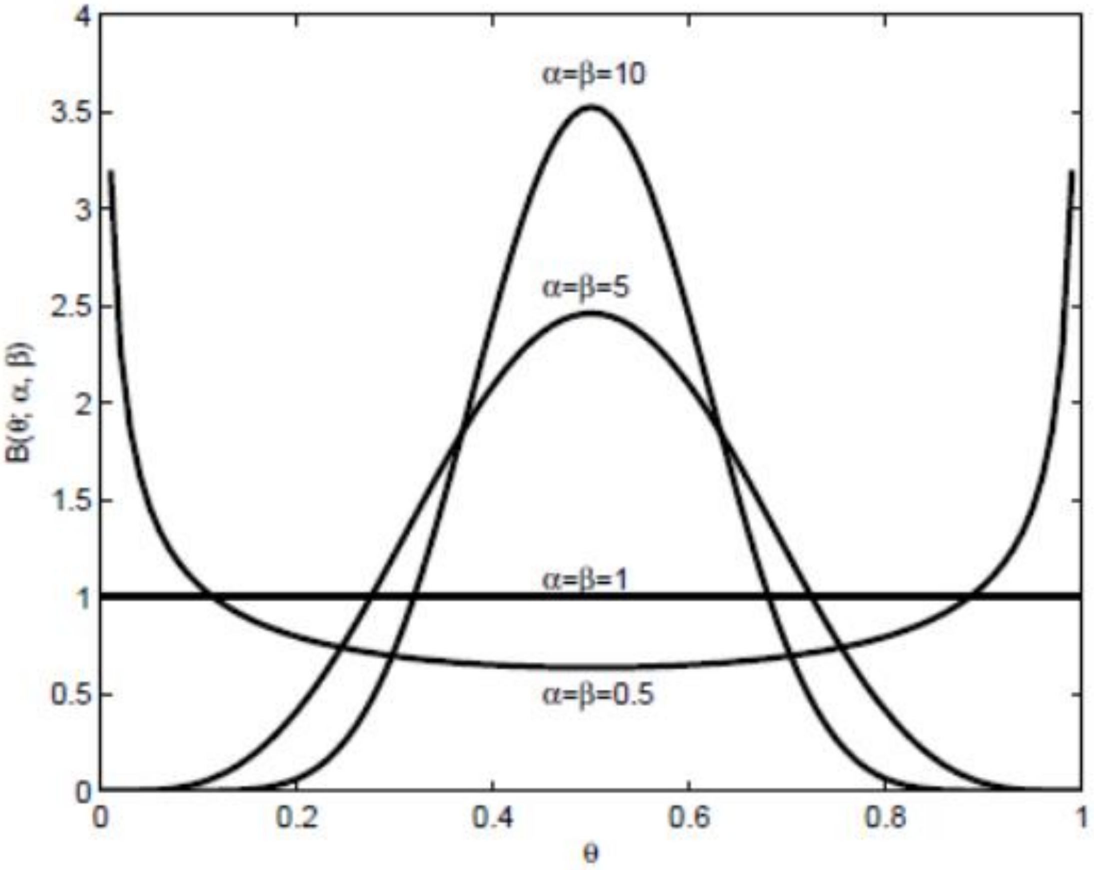
- As a prior distribution, consider the **beta distribution**:

$$\pi(\theta) = \mathrm{Beta}(\theta;\alpha,\beta) \triangleq \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}, \ \ 0 \le \theta \le 1, \ \alpha > 0, \ \beta > 0, \tag{4.140}$$

  where

$$B(\alpha,\beta) = \int_0^{\infty} \theta^{\alpha-1}(1-\theta)^{\beta-1}\, d\theta \tag{4.141}$$

  α and β are called **prior hyperparameters** (cf, the model parameter θ).



(a)

Figure 4.8 The PDF of beta distribution $\mathrm{Beta}(\theta; \alpha, \beta)$ of (4.140) for (a) $\alpha = \beta = 0.5, 1.0, 5$ and 10;

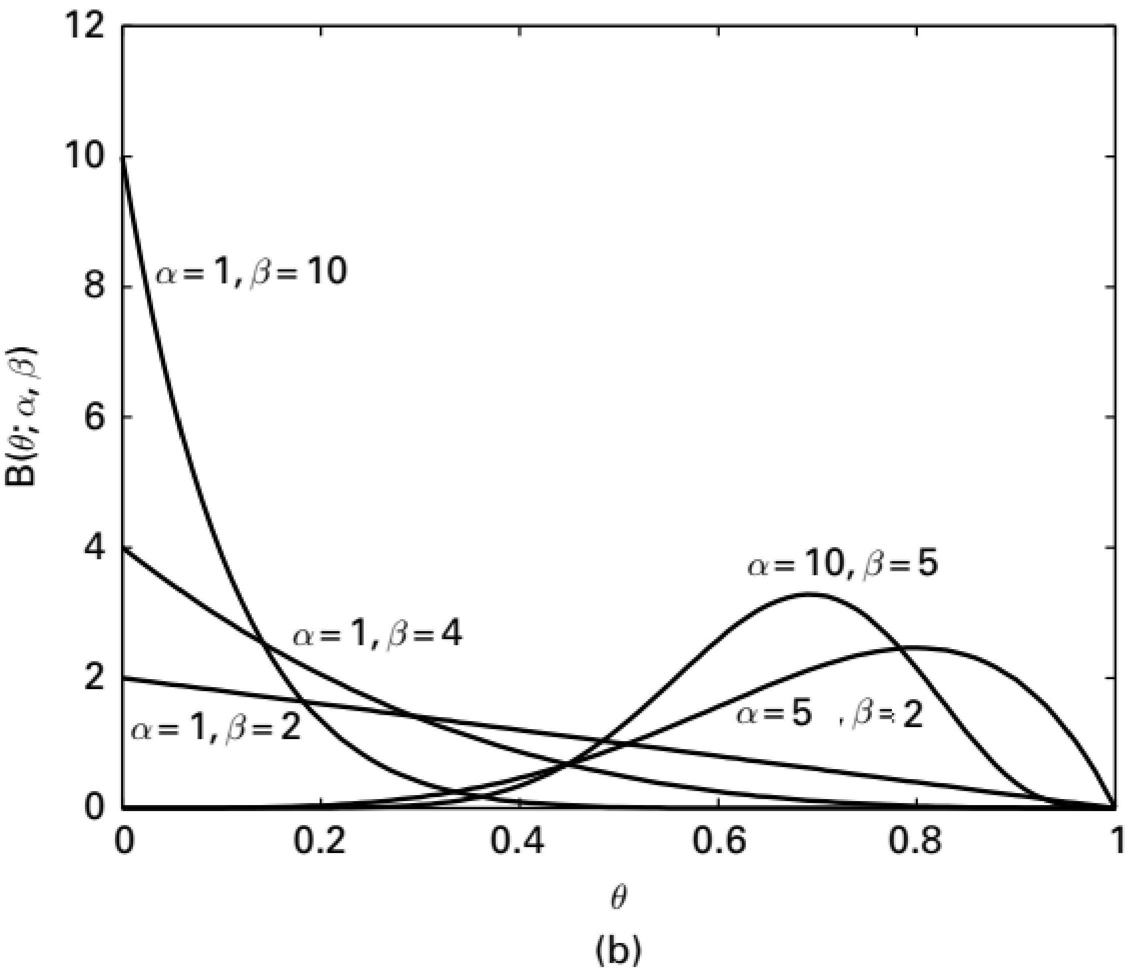


Figure 4.8 The PDF of beta distribution $\mathrm{Beta}(\theta; \alpha, \beta)$ of (4.140) for (a) $\alpha = \beta = 0.5, 1.0, 5$, and 10; (b) $(\alpha, \beta) = (1, 2), (1, 4), (1, 10), (10, 5)$, and $(5, 2)$.

- The **beta function** is related to the **gamma function** (see (4.31) of p. 78)

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{4.142}$$

- The mean and variance of this prior distribution are

$$E[\Theta]=\frac{\alpha}{\alpha+\beta},\ \text{and}\ \mathrm{Var}[\Theta]=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{4.143}$$

- The posterior probability can be evaluated as

$$\begin{aligned}\pi(\theta|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\theta)\pi(\theta) \propto \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\theta^{\alpha-1}(1-\theta)^{\beta-1}\\ &\propto \theta^{(\alpha+\sum_{i=1}^n x_i)-1}(1-\theta)^{(\beta+n-\sum_{i=1}^n x_i)-1},\end{aligned} \tag{4.144}$$

- Thus, the posterior probability is also a beta distribution Beta(θ; α_1, β_1),

$$\alpha_1=\alpha+\sum_{i=1}^n x_i,\ \text{and}\ \beta_1=\beta+n-\sum_{i=1}^n x_i, \tag{4.145}$$

$$\begin{aligned}E[\Theta|\boldsymbol{x}] &= \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\frac{\alpha}{\alpha+\beta}+\left(\frac{n}{\alpha+\beta+n}\right)\overline{x}_n,\\ &= \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)E[\Theta]+\left(\frac{n}{\alpha+\beta+n}\right)\hat{\theta}_{\mathrm{MLE}}(\boldsymbol{x}),\end{aligned} \tag{4.146}$$

where we call α_1 and β_1 the **posterior hyperparameters**, and

$$\hat{\theta}_{\mathrm{MLE}}(x) = \overline{x}_n \triangleq \frac{x_1 + x_2 + \ldots + x_n}{n}$$

very important

is the **maximum likelihood estimate (MLE)** of θ, which is the value that maximizes the likelihood function $L_x(\theta)$ of (4.139).

- As the sample size n increases, the weight on the prior means diminishes, whereas the weight on the MLE approaches one. This behavior illustrates how **Bayesian inference** generally works.
- For a likelihood function that belongs to the **exponential family**, i.e.,

$$L_{\boldsymbol{x}}(\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^\top(\boldsymbol{\theta})\boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\theta})\}, \tag{4.147}$$

conjugate priors can be constructed as follows:

$$f(\boldsymbol{\theta}; \boldsymbol{\alpha}, \beta) \propto \exp\{\boldsymbol{\eta}^\top(\boldsymbol{\theta})\boldsymbol{\alpha} - \beta A(\boldsymbol{\theta})\}, \tag{4.148}$$

then the posterior distribution takes the form

$$f(\boldsymbol{\theta}|x; \boldsymbol{\alpha}, \beta) \propto \exp\{\boldsymbol{\eta}^\top(\boldsymbol{\theta})[\boldsymbol{\alpha} + \boldsymbol{T}(x)] - (1+\beta)A(\boldsymbol{\theta})\}, \tag{4.149}$$

i.e., $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha} + \boldsymbol{T}(\mathbf{x})$, and $\beta_1 = 1 + \beta$.

Example 4.5: **Conjugate prior for the exponential distribution.** The likelihood function for the exponential distribution has the form (cf. (4.25))

$$L_x(\lambda) = \lambda \exp(-\lambda x), \quad x \geq 0, \tag{4.150}$$

where λ is the model parameter. We choose a conjugate prior having the form of a gamma distribution (cf. (4.30)):

$$f(\lambda; \alpha, \beta) = \frac{\alpha(\alpha\lambda)^{\beta-1}e^{-\alpha\lambda}}{\Gamma(\beta)}, \quad \lambda \geq 0, \tag{4.151}$$

where α and β are the prior hyperparameters. Using (4.137), the posterior distribution is computed as

$$f(\lambda|x; \alpha, \beta) = \frac{\alpha(\alpha\lambda)^{\beta}e^{-(\alpha+x)\lambda}}{\Gamma(\beta+1)}, \quad \lambda \geq 0, \tag{4.152}$$

which is a gamma distribution such that the posterior hyperparameters are $\alpha_1 = \alpha + x$ and $\beta_1 = \beta + 1$. If M independent samples $x_1, \ldots, x_M$, are drawn from an exponential distribution, the likelihood function for the vector $\boldsymbol{x} = (x_1, \ldots, x_M)^\top$ has the form

$$L_{\boldsymbol{x}}(\lambda) = \lambda^M \exp\left(-\lambda \sum_{i=1}^{M} x_i\right), \quad \boldsymbol{x} \geq \boldsymbol{0}. \tag{4.153}$$

Using the conjugate prior given by (4.151), we find that the posterior distribution is a gamma distribution with posterior hyperparameters $\alpha_1 = \alpha + M$ and $\beta_1 = \beta + \sum_{i=1}^{M} x_i$. $\square$

Example 4.6: Conjugate prior for a normal distribution with fixed variance σ^2. The likelihood function for a normal family of distributions with fixed variance σ^2 has the form (cf. (4.25))

$$L_x(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{4.154}$$

where μ is the model parameter. Choosing a normal distribution as the conjugate prior, we have

$$f(\mu; \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right], \tag{4.155}$$

with prior hyperparameters μ_0 and σ_0^2. Applying (4.137), we find that the posterior distribution has the form

$$f(\lambda|x;\mu_0,\sigma_0^2) \propto \exp\left\{-\frac{1}{2}\left[\frac{(\mu-x)^2}{\sigma^2}+\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right]\right\}. \tag{4.156}$$

After some algebraic manipulations, we obtain

$$f(\lambda|x;\mu_0,\sigma_0^2) \propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\mu-\frac{\frac{x}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2}+\frac{1}{\sigma^2}}\right)^2\right]. \tag{4.157}$$

Hence, the posterior hyperparameters are

$$\mu_1 = \frac{\frac{x}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}} \text{ and } \sigma_1^2 = \left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}.$$

Generalizing to the case of n independent samples, i.e., $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^\top$, we can show that the posterior hyperparameters are given by

$$\mu_1 = \frac{\frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} \text{ and } \sigma_1^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}. \tag{4.158}$$

The second posterior hyperparameter σ_1^2 in the last expression is the harmonic mean of the prior σ_0^2 and the variance of data. For notational conciseness, the inverse of the variance, $h \triangleq \sigma^{-2}$, called the **precision**, is often used in the Bayesian statistics literature. From the last expression, for instance, the posterior precision is simply given by $h_1 = nh + h_0$, where $h_0 = \sigma_0^{-2}$ is the precision of the prior distribution. Use of precision instead of variance eliminates most of the inversions in the equations presented above. $\square$