

## Handout 7: Mixture priors

Lecturer &amp; author: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

**Aim:** Explain the mixture distribution. Explain, theorize, and construct conjugate mixture prior distribution.

### References:

- Berger, J. O. (2013; Sections 4.2.2). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.
- Robert, C. (2007; Sections 3, pp. 105-123, & pp. 127-141). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.

**Web applets:** [https://georgios-stats-1.shinyapps.io/demo\\_mixturepriors/](https://georgios-stats-1.shinyapps.io/demo_mixturepriors/)

## 1 Finite mixture distributions

**Definition 1.** Let  $\mathcal{P}_m = \{\Pi_l(\theta|\chi_l); l = 1, \dots, m\}$  be a collection of probability distributions where  $\{\chi_l\}_{l=1}^m$  are parameters of the  $l$ -th component  $\Pi_l(\theta|\chi_l)$ . Let  $\{\varpi_l\}_{l=1}^m$  be a set of weights where  $\varpi_l > 0$  and  $\sum_{l=1}^m \varpi_l = 1$ . The mixture distribution derived from the aforementioned collections is

$$\Pi(\theta|\varpi, \chi) = \sum_{l=1}^m \varpi_l \Pi_l(\theta|\chi_l), \quad \theta \in \Theta \quad (1)$$

where  $\chi := (\chi_l, l = 1 : m)$  and  $\varpi := (\varpi_l, l = 1, \dots, m)$ .  $\Pi_l(\theta|\chi_l)$  is called  $l$ -th mixture component with mixture weight  $\varpi_l$ .

**Example 2.** Let  $h(\cdot)$  be a function defined on  $\Theta$ . The expectation of  $h(\cdot)$  with respect to (1) is

$$E_{\Pi}(h(\theta)|\varpi, \chi) = \int \sum_{l=1}^m \varpi_l h(\theta) d\Pi_l(\theta|\chi_l) = \sum_{l=1}^m \varpi_l \int h(\theta) d\Pi_l(\theta|\chi_l) = \sum_{l=1}^m \varpi_l E_{\Pi_l}(h(\theta)|\chi_l)$$

where  $E_{\Pi_l}(h(\theta)|\chi_l) = \int h(\theta) d\Pi_l(\theta|\chi_l)$ .

**Example 3.** A mixture of probability distributions (1) is a probability distribution; i.e.

$$\int_{\Theta} \pi(\theta|\varpi, \chi) d\theta = \sum_{l=1}^m \varpi_l \underbrace{\int_{\Theta} d\Pi_l(\theta|\chi_l) d\theta}_{=1} = \sum_{l=1}^m \varpi_l = 1$$

**Definition 4.** The mixture is called finite mixture if  $m < \infty$ , and countably infinite mixture if  $m \rightarrow \infty$ . Here, we focus on finite mixtures.

**Definition 5.** A mixture model is called parametric mixture model if its components are members of the same parametric family of distributions eg.  $\mathcal{P}_m = \{\Pi(\theta|\chi_l); l = 1, \dots, m\}$ , and hence

$$\Pi(\theta|\varpi, \chi) = \sum_{l=1}^m \varpi_l \Pi(\theta|\chi_l)$$

**Example 6.** An example of a parametric mixture model is the Normal mixture model

$$\pi(\theta|\varpi, \mu, \sigma^2) = \sum_{l=1}^m \varpi_l \mathcal{N}(\theta|\mu_l, \sigma_l^2),$$

where all components belong to the Normal distribution family.

*Note 7.* Mixture distributions are useful because (among others):

- they can approximate complicate other distributions by using a combination of simpler distributions  $\{\Pi_l(\theta|\chi_l)\}$  which lead to more convenient computations.
- they can naturally model heterogeneity. E.g. consider a population which is heterogeneous in the sense that there are multiple sub-groups labeled by  $\ell \in \{1, \dots, m\}$ , each group represented in the population with proportion  $\varpi_\ell$ , and distributed as  $\Pi_l(\theta|\chi_l)$ . Then  $y \sim \Pi(\theta|\chi)$  can be realized by drawing  $\ell$  with probability  $P(\ell = l) = \varpi_l$  and drawing  $\theta$  from  $\Pi_\ell(\theta|\chi_\ell)$  given  $\ell$ , aka

$$\begin{cases} \theta|\ell & \sim \Pi_\ell(\theta|\chi_\ell) \\ \ell & \sim P(\ell) \end{cases} \quad \text{which implies } \theta \sim \Pi(\theta|\chi) \quad \text{since} \quad \Pi(\theta|\chi) = \sum_{\ell=1}^m \Pi(\theta|\chi_\ell)P(\ell) = \sum_{l=1}^m \varpi_l \Pi(\theta|\chi_l)$$

See Section 1 in Handout 2.

**Example 8.** A bimodal, or a right skewed (non-symmetric) distribution can be approximated by a Mixture of (unimodal and symmetric) Normal distributions with different parameter values. Also it describes a population with two groups Normally distributed with different parameters.



(a) Bimodal PDF:  
 $\pi(\theta) = 0.7\mathcal{N}(\theta|1, 0.9^2) + 0.3\mathcal{N}(\theta|6, 1.9^2)$



(b) Right skewed PDF:  
 $\pi(\theta) = 0.7\mathcal{N}(\theta|1, 0.9^2) + 0.3\mathcal{N}(\theta|4, 1.9^2)$

Figure 1: Normal mixture models.  $\mathcal{N}(x|\mu, \sigma^2)$  denotes the Normal distribution density at value  $x$  with mean  $\mu$  and variance  $\sigma^2$ .

## 2 Mixture prior distributions

*Note 9.* Mixture models can be used to specify priors distributions, either as a mean to approximate Your actual prior distribution with simpler & tractable distributions, or as a mean to represent heterogeneous prior believes.

**Theorem 10.** Let  $y := (y_1, \dots, y_n)$  be observables generated from the sampling distribution  $F(y|\theta)$ . Prior mixture distribution  $\Pi(\theta|\varpi)$  is called the prior with pdf/pmf

$$\pi(\theta|\varpi) = \sum_{l=1}^m \varpi_l \pi_l(\theta), \quad (2)$$

where,  $\mathcal{P}_m = \{\pi_l(\theta), \theta \in \Theta\}_{l=1}^m$  is a collection of distributions, and  $\{\varpi_l\}$  are weights such that  $\sum_{l=1}^m \varpi_l = 1$  and  $\varpi_l > 0$ . Then:

1. the posterior distribution  $\Pi(\theta|y, \varpi)$  has pdf/pmf

$$\pi(\theta|y, \varpi) = \sum_{l=1}^m \varpi_l^* \pi_l(\theta|y), \quad (3)$$

2. the predictive distribution of a future outcome  $z$  has pdf/pmf

$$g(z|y, \varpi) = \sum_{l=1}^m \varpi_l^* g_l(z|y)$$

where

$$\begin{aligned} \varpi_l^* &= \frac{\varpi_l f_l(y)}{\sum_{l=1}^m \varpi_l f_l(y)} \propto \varpi_l f_l(y) \\ f_l(y) &= \int_{\Theta} f(y|\theta) d\Pi_l(\theta) \\ \pi_l(\theta|y) &= \frac{f(y|\theta) \pi_l(\theta)}{\int_{\Theta} f(y|\vartheta) d\Pi_l(\vartheta)} \propto f(y|\theta) \pi_l(\theta) \\ g_l(z|y) &= \int_{\Theta} f(z|\theta) \pi_l(\theta|y) d\theta \end{aligned}$$

*Proof.* From Bayes theorem, we have:

$$\begin{aligned} d\Pi(\theta|y) &= \frac{f(y|\theta) d\Pi(\theta)}{\int_{\Theta} f(y|\vartheta) d\Pi(\vartheta)} = \frac{f(y|\theta) \sum_{l=1}^m \varpi_l d\Pi_l(\theta)}{\int_{\Theta} f(y|\vartheta) \sum_{l'=1}^m \varpi_{l'} d\Pi_{l'}(\vartheta)} = \frac{\sum_{l=1}^m \varpi_l f(y|\theta) d\Pi_l(\theta)}{\sum_{l'=1}^m \int_{\Theta} \varpi_{l'} f(y|\vartheta) d\Pi_{l'}(\vartheta)} \\ &= \frac{\sum_{l=1}^m \varpi_l f_l(y) \overbrace{\frac{f(y|\theta) d\Pi_l(\theta)}{f_l(y)}}^{\substack{=d\Pi_l(\theta|y)}}}{\sum_{l'=1}^m \int_{\Theta} \varpi_{l'} f_{l'}(y) \underbrace{\frac{f(y|\vartheta) d\Pi_{l'}(\vartheta)}{f_{l'}(y)}}_{=d\Pi_{l'}(\vartheta|y)}} = \frac{\sum_{l=1}^m \varpi_l f_l(y) d\Pi_l(\theta|y)}{\sum_{l'=1}^m \varpi_{l'} f_{l'}(y) \underbrace{\int_{\Theta} d\Pi_{l'}(\vartheta|y)}_{=1}} \\ &= \sum_{l=1}^m \underbrace{\frac{\varpi_l f_l(y)}{\sum_{l'=1}^m \varpi_{l'} f_{l'}(y)}}_{=\varpi_l^*} d\Pi_l(\theta|y) = \sum_{l=1}^m \varpi_l^* d\Pi_l(\theta|y). \end{aligned}$$

Also

$$g(z|y, \varpi) = \int_{\Theta} f(z|\theta) d\Pi_l(\theta|y) = \int_{\Theta} f(y|\theta) \sum_{l=1}^m \varpi_l^* d\Pi_l(\theta|y) = \sum_{l=1}^m \varpi_l^* g_l(z|y)$$

□

**Remark 11.** Theorem 10 shows that the posterior distribution  $\Pi(\theta|y)$  (derived by a mixture prior) is a mixture of 'individual posterior distributions'  $\Pi_l(\theta|y)$  weighted by  $\varpi_l^*$ . It is determined not only by the observables but also by the weights of the individual distributions. Note that, for  $l = 1, \dots, m$ , the prior weights  $\varpi_l$  and posterior weights  $\varpi_l^*$  may differ a lot, however they can be close each other.

**Remark 12.** Mixture priors  $\Pi(\theta)$  whose components  $\{\Pi_l(\theta)\}$  are conjugate to the likelihood  $f(y|\theta)$  can facilitate tractable Bayesian inference. In Theorem 10, if each  $\Pi_l(\theta)$  is conjugate to the likelihood  $f(y|\theta)$ , then obviously each  $\Pi_l(\theta|y)$  will belong to the same distribution family as the corresponding  $\Pi_l(\theta)$  because  $\pi_l(\theta|y) \propto f(y|\theta)\pi_l(\theta)$ . Then, provided that components  $\pi_l(\theta)$  are tractable, the components  $\pi_l(\theta|y)$  will be tractable too (as they belong to the same distr. family). Likewise, the posterior weights can be calculated in closed form i.e.,  $\varpi_l^* \propto \varpi_l f_l(y)$  since the integral  $f_l(y) = \int_{\Theta} f(y|\theta) d\Pi_l(\theta)$  will be tractable. Hence, the produced posterior mixture pdf/pmf will be tractable.

**Remark 13.** Mixtures of conjugate priors are as easy to manipulate as regular conjugate distributions, while leading to a greater freedom in the modeling of the prior information.

**Example 14.** Let  $y = (y_1, \dots, y_n)$  observable quantities, generated iid from a Bernoulli sampling distribution with unknown parameter  $\theta$ ; aka  $y_i|\theta \stackrel{\text{iid}}{\sim} \text{Br}(\theta)$ ,  $i = 1, \dots, n$ .

1. Find the likelihood function
2. Find the PDF of the conjugate prior mixture prior  $\pi(\theta) = \sum_{l=1}^m \varpi_l \pi_l(\theta)$ , with  $m$  components  $\{\pi_l(\theta)\}_{l=1}^m$ .
3. Compute the pdf of the posterior distribution, and recognize it.
4. Compute the predictive pdf for the next outcome  $z = (y_{n+1}, \dots, y_{n+m})$  given we have observed  $y$ ? What do you observe?

**Hint:** Beta distribution:  $x \sim \text{Be}(a, b)$  has pdf

$$f(x) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1} \mathbf{1}(x \in [0, 1]); \quad \text{where} \quad B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \quad a > 0, b > 0$$

**Solution.**

1. The likelihood function is

$$f(y|\theta) = \prod_{i=1}^n f(y_i|\theta) = \prod_{i=1}^n \text{Br}(x_i|\theta) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}$$

2. In previous examples, we found that

- the sampling distribution  $F(\cdot|\theta)$  is a Bernoulli distribution which is a member of the Exponential family.
- the conjugate prior is Beta distribution, aka  $\theta \sim \text{Be}(a, b)$   $a > 0$  and  $b > 0$ .

Therefore, the components  $\{\pi_l(\theta)\}_{l=1}^m$  in the prior mixture distribution will be from the family of Beta distributions  $\mathcal{P}_m = \{\text{Be}(\theta|a_l, b_l); l = 1, \dots, m\}$ .

Therefore, the conjugate mixture prior has pdf

$$\pi(\theta) = \sum_{l=1}^m \varpi_l \text{Be}(\theta|a_l, b_l)$$

where  $\text{Be}(\theta|a_l, b_l)$  is the pdf of  $\text{Be}(a_l, b_l)$ , and  $\{(a_l, b_l)\}$  are fixed prior hyper-parameters.

3. The posterior mixture posterior has PDF

$$\pi(\theta|y) = \sum_{l=1}^m \varpi_l^* \pi_l(\theta|y)$$

According to Theorem 10, the  $l$ -th components of the mixture posterior is

$$\begin{aligned} \pi_l(\theta|y) &\propto f(y|\theta)\pi_l(\theta) = \prod_{i=1}^n \text{Br}(y_i|\theta) \times \text{Be}(\theta|a_l, b_l) \propto \prod_{i=1}^n [\theta^{y_i}(1-\theta)^{1-y_i}] \times \theta^{a_l-1}(1-\theta)^{b_l-1} \\ &\quad \theta^{r_n+\sum_{i=1}^n y_i} \theta^{r_n+a_l-1}(1-\theta)^{n-r_n+b_l-1} \propto \text{Be}(\theta|a_l^*, b_l^*) \end{aligned}$$

with  $a_l^* = r_n + a_l$ ,  $b_l^* = n - r_n + b_l$ , and  $r_n = \sum_{i=1}^n y_i$ .

According to Theorem 10, the posterior weights can be calculated as

$$\begin{aligned} \varpi_l^* &\propto \varpi_l f_l(y) \propto \varpi_l \int_{(0,\infty)} \prod_{i=1}^n f(x_i|\theta)\pi_l(\theta)d\theta = \varpi_l \int_{(0,\infty)} \prod_{i=1}^n \text{Br}(y_i|\theta)\text{Be}(\theta|a_l, b_l)d\theta \\ &= \varpi_l \int_{(0,\infty)} \theta^{r_n}(1-\theta)^{n-r_n} \frac{\Gamma(a_l+b_l)}{\Gamma(a_l)\Gamma(b_l)} \theta^{a_l-1}(1-\theta)^{b_l-1} d\theta \\ &= \varpi_l \frac{\Gamma(a_l+b_l)}{\Gamma(a_l)\Gamma(b_l)} \int_{(0,1)} \theta^{r_n+a_l-1}(1-\theta)^{n-r_n+b_l-1} d\theta = \varpi_l \frac{\Gamma(a_l+b_l)}{\Gamma(a_l)\Gamma(b_l)} \frac{\Gamma(a_l^*)\Gamma(b_l^*)}{\Gamma(a_l^*+b_l^*)} \end{aligned} \quad (4)$$

namely,

$$\varpi_l^* = \frac{\varpi_l \frac{\Gamma(a_l+b_l)}{\Gamma(a_l)\Gamma(b_l)} \frac{\Gamma(a_l^*)\Gamma(b_l^*)}{\Gamma(a_l^*+b_l^*)}}{\sum_{l=1}^m \varpi_l \frac{\Gamma(a_l+b_l)}{\Gamma(a_l)\Gamma(b_l)} \frac{\Gamma(a_l^*)\Gamma(b_l^*)}{\Gamma(a_l^*+b_l^*)}}$$

4. It is ...

$$g(z|y) = \int_{\Theta} f(z|\theta)d\Pi(\theta|y) = \sum_{i=1}^m \varpi_i^*(y) \underbrace{\int_{\Theta} f(z|\theta)d\Pi_i(\theta|y)}_{=g_i(z|y)}$$

where the following is just copy-paste from Handout 3...

$$\begin{aligned} g_l(z|y) &= \int_{\Theta} f(z|\theta)\pi_l(\theta|y)d\theta = \int_{\Theta} \prod_{i=1}^m f(z_i|\theta)\pi_l(\theta|y)d\theta = \int_{(0,\infty)} \prod_{i=1}^m \text{Br}(z_i|\theta)\text{Be}(\theta|a_l^*, b_l^*)d\theta \\ &= \int_0^1 \left[ \theta^{\sum_{i=1}^m z_i} (1-\theta)^{m-\sum_{i=1}^m z_i} \right] \left[ \frac{\theta^{a_l^*-1}(1-\theta)^{b_l^*-1}}{B(a_l^*, b_l^*)} \right] d\theta \mathbf{1}(z \in \{0, 1\}^m) \\ &= \frac{1}{B(a_l^*, b_l^*)} \int_0^1 \theta^{\sum_{i=1}^m z_i + a_l^* - 1} (1-\theta)^{m - \sum_{i=1}^m z_i + b_l^* - 1} d\theta \mathbf{1}(z \in \{0, 1\}^m) \\ &= \frac{B(\sum_{i=1}^m z_i + a_l^*, m - \sum_{i=1}^m z_i + b_l^*)}{B(a_l^*, b_l^*)} \mathbf{1}(z \in \{0, 1\}^m) \end{aligned}$$

### 3 Practice

**Question 15.** Try the Exercise 51 from the Exercise sheet.