MSA101/MVE187 2020 Lecture 3.1 Mixtures

Petter Mostad

Chalmers University

September 15, 2020

Review of Lecture 2

- Basic concepts: Prior, posterior, ...
- Conjugate priors. Computations.
- The prior predictive and posterior predictive. Computations.
- The exponential family of distributions.

Mixtures of conjugate priors

- A family of conjugate priors, with limited flexibility, can be greatly extended by also considering linear combinations of these prior densities.
- Example: The Poisson-Gamma conjugacy: Assume $\pi(x \mid \theta) = e^{-\theta}\theta^x/x!$ and $\pi(\theta) \propto_\theta \theta^{\alpha-1}\exp(-\beta\theta)$ so that $\pi(\theta \mid x) \propto_\theta \theta^{\alpha+x-1}\exp(-(\beta+1)\theta)$. Then a linear combination prior

$$\pi(\theta) = C_1\theta^{\alpha_1-1}\exp(-\beta_1\theta) + C_2\theta^{\alpha_2-1}\exp(-\beta_2\theta)$$

  will result in a linear combination posterior

$$\pi(\theta \mid x) \propto_\theta C_1\theta^{\alpha_1+x-1}\exp(-(\beta_1+1)\theta) + C_2\theta^{\alpha_2+x-1}\exp(-(\beta_2+1)\theta).$$

- This works for any conjugate family, and any linear combination of priors from it.
- Note however that the weigts *of the densities* in the linear combination are updated!

Mixtures of conjugate priors: Formulas

- Assume $\pi(\theta \mid \lambda)$ is a family of conjugate priors to $\pi(x \mid \theta)$. Given $\lambda_1, \ldots, \lambda_n$, let $g_i(\theta \mid x)$ and $f_i(x)$ denote the posterior and the prior predictive, respectively, when using the prior $\pi(\theta \mid \lambda_i)$. Then

$$\pi(x \mid \theta)\pi(\theta \mid \lambda_i) = g_i(\theta \mid x) f_i(x).$$

- Assume we use a linear combination prior

$$\pi(\theta) = \sum_{i=1}^{n} w_i \pi(\theta \mid \lambda_i) \quad \text{where} \sum_{i=1}^{n} w_i = 1.$$

- For the prior predictive we get

$$\pi(x) = \int \pi(x \mid \theta) \sum_{i=1}^{n} w_i \pi(\theta \mid \lambda_i)\, d\theta = \sum_{i=1}^{n} w_i f_i(x).$$

- for the posterior we get

$$\pi(\theta \mid x) = \frac{\pi(x \mid \theta)\pi(\theta)}{\pi(x)} = \frac{\pi(x \mid \theta) \sum_{j=1}^{n} w_j \pi(\theta \mid \lambda_j)}{\sum_{i=1}^{n} w_i f_i(x)}$$

$$= \frac{\sum_{j=1}^{n} w_j f_j(x) g_j(\theta \mid x)}{\sum_{i=1}^{n} w_i f_i(x)} = \sum_{j=1}^{n} w_j' g_j(\theta \mid x) \quad \text{where } w_j = \frac{w_j f_j(x)}{\sum_{i=1}^{n} w_i f_i(x)}.$$

Example of mixtures

- We use a Binomial likelihood Binomial$(4, p)$, with 3 successes observed in 4 trials.
- We use a mixture prior

$$\pi(p) = 0.5 \cdot \text{Beta}(p; 2.5, 2.5) + 0.5 \cdot \text{Beta}(p; 11, 31)$$

- Recall that if $x \mid p \sim \text{Binomial}(n, p)$ and $p \sim \text{Beta}(\alpha, \beta)$ then the prior predictive becomes

$$\pi(x) = \binom{n}{x} \frac{\text{B}(\alpha + x, \beta + n - x)}{\text{B}(\alpha, \beta)}$$

- Thus the first updated weight becomes

$$w_1' = \frac{0.5 \cdot \binom{4}{3} \frac{\text{B}(2.5+3,2.5+1)}{\text{B}(2.5,2.5)}}{0.5 \cdot \binom{4}{3} \frac{\text{B}(2.5+3,2.5+1)}{\text{B}(2.5,2.5)} + 0.5 \cdot \binom{4}{3} \frac{\text{B}(11+3,31+1)}{\text{B}(11,31)}}$$

  and similar for the second updated weight w_2'.
- The posterior becomes

$$\pi(p \mid x) = w_1' \, \text{Beta}(x; 2.5 + 3, 2.5 + 1) + w_2' \, \text{Beta}(x; 11 + 3, 31 + 1).$$

MSA101/MVE187 2020 Lecture 3.2
Some multivariate conjugacies

Petter Mostad

Chalmers University

September 3, 2020

Multivariate conjugacy example: The normal likelihood, no parameters known

- Assume $X \sim \text{Normal}(\mu, 1/\tau)$, with both μ and τ uncertain. The likelihood becomes

$$\pi(x \mid \mu, \tau) \propto_{\mu,\tau} \tau^{1/2} \exp\left(-\frac{\tau}{2}(x-\mu)^2\right)$$

- Then the Normal-Gamma family is conjugate: The pair (μ, τ) has a Normal-Gamma distribution with parameters $\mu_0, \lambda > 0, \alpha > 0, \beta > 0$ if the density has the form

$$\pi(\mu, \tau \mid \mu_0, \lambda, \alpha, \beta) = \frac{\beta^\alpha \sqrt{\lambda}}{\Gamma(\alpha)\sqrt{2\pi}} \tau^{\alpha - 1/2} \exp\left(-\beta\tau - \frac{\lambda\tau}{2}(\mu - \mu_0)^2\right)$$

- Note: If (μ, τ) has the Normal-Gamma distribution above, we have $\tau \sim \text{Gamma}(\alpha, \beta)$ and $\mu \mid \tau \sim \text{Normal}(\mu_0, 1/(\lambda\tau))$.

Example

- Assume $x = (x_1, x_2, \ldots, x_n)$ observed from Normal$(\mu, 1/\tau)$.
- Assume prior

$$\tau \sim \text{Gamma}(\alpha, \beta) \quad \text{and} \quad \mu \mid \tau \sim \text{Normal}(\mu_0, 1/(\lambda\tau))$$

- The posterior becomes (compute or use Wikipedia...)

$$\begin{aligned} \tau \mid x &\sim \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{=1}^{n}(x_i - \overline{x})^2 + \frac{n\lambda}{\lambda + n}\frac{(\overline{x} - \mu_0)^2}{2}\right) \\ \mu \mid \tau, x &\sim \text{Normal}\left(\frac{\lambda\mu_0 + n\overline{x}}{\lambda + n}, \frac{1}{(\lambda + n)\tau}\right) \end{aligned}$$

- Numerical example: $x = (3.1, 4.2, 2.9, 3.7, 3.9)$. Plot:
  - Likelihood
  - Prior

$$\tau \sim \text{Gamma}(1, 1/3) \quad \text{and} \quad \mu \mid \tau \sim \text{Normal}(3, 3/\tau)$$

  - Posterior

$$\tau \sim \text{Gamma}(3.5, 0.9783) \quad \text{and} \quad \mu \mid \tau \sim \text{Normal}(3.525, 1/(5.33 \cdot \tau))$$

MSA101/MVE187 2020 Lecture 3.3
Some multivariate conjugacies

Petter Mostad

Chalmers University

September 3, 2020

The prior

- Where does the prior come from?
- From "other information" than the current data; possibly older data. The context.
- What if the context indicates "no prior information"?
- For many likelihood models there exist "non-informative priors".
- For the Normal-Gamma prior to the Normal distribution, try the prior $\alpha = -1/2, \beta = \lambda = \mu_0 = 0$.
- NOTE: The posterior now becomes $\tau \sim \text{Gamma}(\frac{n-1}{2}, \frac{1}{2}\sum_{i=1}^{n}(x_i - \overline{x})^2)$ and $\mu \mid \tau \sim \text{Normal}(\overline{x}, \frac{1}{n\tau})$, corresponding to standard frequentist analysis.
- NOTE: Plugging in the suggested parameters into the prior density gives $\pi(\mu, \tau) \propto_{\mu,\tau} \tau^{-1}$. The prior is an *improper prior*: It does not integrate to 1!

Stepwise Bayesian updating

- Assume $x = (x_1, \ldots, x_m)$ is a random sample, so that

$$\pi(x \mid \theta) = \prod_{i=1}^{m} \pi(x_i \mid \theta)$$

- Using a prior $\pi(\theta)$ the posterior becomes

$$\pi(\theta \mid x) \propto_\theta \prod_{i=1}^{m} \pi(x_i \mid \theta)\pi(\theta)$$

- If we first update only with the observations $x_1, \ldots, x_k$, we get the posterior

$$\pi(\theta \mid x_1, \ldots, x_k) \propto_\theta \prod_{i=1}^{k} \pi(x_i \mid \theta)\pi(\theta)$$

- We see that if we use this as the prior and update with the remaining data $(x_{k+1}, \ldots, x_m)$, we get the same result as before.
- In Bayesian statistics, we may subdivide the data into data subsets and update the model stepwise with the data, as long as all the data sets are mutually independent given the model parameter θ.

Multinomial-Dirichlet conjugacy

- Assume $x = (x_1, \ldots, x_n) \sim \text{Multinomial}(m, \theta_1, \theta_2, \ldots, \theta_n)$, with $\theta_1 + \cdots + \theta_n = 1$, so that x_i counts the number of results of type i in m independent trials, if results of type i have probability θ_i. The probability mass function is

$$\pi(x \mid \theta_1, \ldots, \theta_n) = \frac{m!}{x_1! \ldots x_k!} \theta_1^{x_1} \ldots \theta_n^{x_n}$$

- $(\theta_1, \ldots, \theta_n)$ has a Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_n$ if the density can be written as

$$\pi(\theta_1, \ldots, \theta_n \mid \alpha_1, \ldots, \alpha_n) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_n)} \theta_1^{\alpha_1 - 1} \ldots \theta_n^{\alpha_n - 1}$$

- Prove that the Dirichlet family is a conjugate family to the Multinomial likelhiood!

Applied example: Forensic DNA matches

- DNA matching between a trace and a person may be used as proof in criminal cases.
- Given that there is a match, how strong is the evidence? Could the match have been by chance?
- DNA traces collect information from (mostly independent) *loci* (i.e., locations) in the genome. At these loci, each person has two copies of an *allele*, various alleles may exist in the population.
- Popular forensic loci are STRs, where alleles differ by the number of repetitions of a short sequence (such as, e.g., CGTT).
- To answer question above, one needs to establish the *probaility of observing* in a random person each possible allele. For this *databases* are collected.
- Lets say a (toy) database of size 16 contans the counts: 14: 3; 15: 9, 16: 3, 17: 0, 18: 1. What is the population probabilities of each allele? How certain can you be?
- It is then common, and advantageous, to use the Multinomial-Dirichlet model, with *pseudocounts*.

MSA101/MVE187 2020 Lecture 3.4
Total expectation, total variance
The multivariate normal distribution

Petter Mostad

Chalmers University

September 4, 2020

Laws of total expectation, total variance

Assume you have a joint density $\pi(x, \theta)$. Note the formulas

$$\mathsf{E}(x) = \mathsf{E}(\mathsf{E}(x \mid \theta))$$

and

$$\mathsf{Var}(x) = \mathsf{E}(\mathsf{Var}(x \mid \theta)) + \mathsf{Var}(\mathsf{E}(x \mid \theta)).$$

You should be able to use and prove these formulas.

The multivariate normal distribution

- We say X has a multivariate (n-variate) normal distribution, if it is a vector of length n with density

$$\pi(X) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X-\mu)\Sigma^{-1}(X-\mu)^t\right)$$

  where the vector μ is the expectation and the $n \times n$ symmetric matrix Σ is the covariance matrix. $|2\pi\Sigma|$ is the determinant of $2\pi\Sigma$.
- We write $X \sim \text{Normal}(\mu, \Sigma)$.
- Just as in the 1-dimensional case: If $Y \mid X \sim \text{Normal}(AX + B, \Sigma_1)$ and $X \sim \text{Normal}(\mu, \Sigma_0)$, and if we look at $Y \mid X$ as a likelihood and $\pi(X)$ as a prior, then this is a conjugate prior.
- We usually express this by using that
  - In the case above, the *joint* density for X and Y is multivariate normal.
  - For a multivariate normal vector, the *conditional* vector when fixing one or more components in the vector is also multivariate normal.

The joint multivariate normal distribution

- Assume $Y \mid X \sim \text{Normal}(AX + B, \Sigma_1)$ and $X \sim \text{Normal}(\mu, \Sigma_0)$. Then

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \text{Normal}\left(\begin{bmatrix} \mu \\ A\mu + B \end{bmatrix}, \begin{bmatrix} \Sigma_0 & \Sigma_0 A^t \\ A\Sigma_0 & A\Sigma_0 A^t + \Sigma_1 \end{bmatrix} \right)$$

- One can prove this directly from the definitions, or use
  - Prove first that the joint distribution must be multivariate normal.
  - Then, compute the expectation and the covariance matrix of the joint vector, using, e.g., the formulas for total expectation and variation, or matrix algebra.

The conditional and the marginal in a multivariate normal distribution

Assume the joint distribution for two vectors X and Y is multivariate normal. Then

- If we integrate out one of them, e.g. Y, the marginal for X, the remaining one, is multivariate normal. The parameters can be read off the expectation and the covariance matrix of the joint distribution.
- If we fix Y, then the *conditional distribution* $X \mid Y$ is also multivariate normal. In fact, if

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \text{Normal}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}^{-1} \right)$$

  we have

$$X \mid Y \sim \text{Normal}(\mu_1 - P_{11}^{-1} P_{12}(Y - \mu_2), P_{11}^{-1})$$

Elements of a proof

- Prove the algebraic matrix identity

$$\begin{aligned} & \left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \right)^t \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \right) \\ = \quad & \left(\theta_1 - \mu_1 + P_{11}^{-1} P_{12}(\theta_2 - \mu_2)\right)^t P_{11} \left(\theta_1 - \mu_1 + P_{11}^{-1} P_{12}(\theta_2 - \mu_2)\right) \\ & + (\theta_2 - \mu_2)^t (P_{22} - P_{21} P_{11}^{-1} P_{12})(\theta_2 - \mu_2). \end{aligned}$$

- Use the definition of the joint density for X and Y, and rewrite it as two factors, one depending only on Y.