MSA101/MVE187 2021 Lecture 3
Discretization, low dim Bayesian inference
Mixtures
Some multivariate conjugacies

Petter Mostad

Chalmers University

September 6, 2021

Review

- Defined the Bayesian *paradigm*: Y_{data}, Y_{pred}, θ, etc.
- Defined some basic concepts and properties: Prior, posterior, predictive, sequential use of data, etc.
- Defined *conjugacy*; seen some examples.
- The exponential family of distributions.

Overview for today

- More tools for basic Bayesian inference; next time: Inference based on *simulation*.
- Discrete Bayes and discretization. Numerical integration.
- Mixtures.
- Some multivariate conjugacies.

Bayesian inference using discretization

- When θ has a finite (and manageable) number of possible values: Seen examples (in Albert) of Bayesian computations.
- *Discretization*: Approximating a continuous prior for θ with a discrete prior.
- **Presentation break for computations by hand**
- Summary:
  - The prior distribution $\pi(\theta)$ is represented by a vector.
  - The posterior distribution $\pi(\theta \mid y)$ is obtained by termwise multiplication of the vectors $\pi(y \mid \theta)$ and $\pi(\theta)$ and normalizing so the result sums to 1.
  - The prediction $\pi(y_{new} \mid y) = \int_\theta \pi(y_{new} \mid \theta)\pi(\theta \mid y)\, d\theta$ simplifies to taking the sum of the termwise product of the vectors $\pi(y_{new} \mid \theta)$ and $\pi(\theta \mid y)$.
- Very often a very good and accurate computational method, when *theta* has 1 (or 2 or 3) dimensions.
- Why does it not work when θ has many dimensions?

Bayesian inference using numerical integration

- The prediction we want to make can be expressed as a quotient of integrals:

$$
\begin{aligned}
\pi(y_{pred} \mid y_{data}) &= \int_\theta \pi(y_{pred} \mid \theta)\pi(\theta \mid y_{data})\, d\theta \\
&= \int_\theta \pi(y_{pred} \mid \theta) \frac{\pi(y_{data} \mid \theta)\pi(\theta)}{\int_\theta \pi(y_{data} \mid \theta)\pi(\theta)\, d\theta}\, d\theta \\
&= \frac{\int_\theta \pi(y_{pred} \mid \theta)\pi(y_{data} \mid \theta)\pi(\theta)\, d\theta}{\int_\theta \pi(y_{data} \mid \theta)\pi(\theta)\, d\theta}
\end{aligned}
$$

- One idea: Compute these integrals using numerical integration.
- **Presentation break for computations by hand**
- Can work well as long as the dimension of θ is low (max 2 or 3?) and the functions are well-behaved.

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

$$\pi(y \mid \theta) = e^{-\theta}\theta^y/y! \qquad \text{and} \qquad \pi(\theta) \propto_\theta \theta^{\alpha-1}\exp(-\beta\theta)$$

  so that $\pi(\theta \mid y) \propto_\theta \theta^{\alpha+y-1}\exp(-(\beta+1)\theta)$.
- Then a linear combination prior (C_1 and C_2 integration constants)

$$\pi(\theta) = w_1 C_1 \theta^{\alpha_1-1}\exp(-\beta_1\theta) + w_2 C_2 \theta^{\alpha_2-1}\exp(-\beta_2\theta)$$

  will result in a linear combination posterior

$$\pi(\theta \mid y) \propto_\theta w_1 C_1 \theta^{\alpha_1+y-1}\exp(-(\beta_1+1)\theta) + w_2 C_2 \theta^{\alpha_2+y-1}\exp(-(\beta_2+1)\theta).$$

- 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 priors: Formulas

- Assume $\pi(\theta \mid \lambda)$ is a family of conjugate priors to $\pi(y \mid \theta)$. Given $\lambda_1, \ldots, \lambda_n$, let $g_i(\theta \mid y)$ and $f_i(y)$ denote the posterior and the prior predictive, respectively, when using the prior $\pi(\theta \mid \lambda_i)$. Then
$$\pi(y \mid \theta)\pi(\theta \mid \lambda_i) = g_i(\theta \mid y) f_i(y).$$
- Assume we use a linear combination prior
$$\pi(\theta) = \sum_{i=1}^n w_i \pi(\theta \mid \lambda_i) \text{ where } \sum_{i=1}^n w_i = 1.$$
- For the prior predictive we get
$$\pi(y) = \int \pi(y \mid \theta) \sum_{i=1}^n w_i \pi(\theta \mid \lambda_i)\, d\theta = \sum_{i=1}^n w_i f_i(y).$$
- for the posterior we get
$$\pi(\theta \mid y) = \frac{\pi(y \mid \theta)\pi(\theta)}{\pi(y)} = \frac{\pi(y \mid \theta) \sum_{j=1}^n w_j \pi(\theta \mid \lambda_j)}{\sum_{i=1}^n w_i f_i(y)}$$
$$= \frac{\sum_{j=1}^n w_j f_j(y) g_j(\theta \mid y)}{\sum_{i=1}^n w_i f_i(x)} = \sum_{j=1}^n w_j' g_j(\theta \mid y) \text{ where } w_j' = \frac{w_j f_j(y)}{\sum_{i=1}^n w_i f_i(y)}.$$

Mixtures of priors

- NOTE: The formula on previous overhead is valid for any mixture of any set of priors. However: It is useful mostly when the posterior and predictive distributions are easily computable.
- NOTE: The $f_j(y)$ in the updated weights

$$w_j' = \frac{w_j f_j(y)}{\sum_{i=1}^{n} w_i f_i(y)}$$

can be interpreted as *the probability of observing the data y if we assume the prior $\pi(\theta \mid \lambda_i)$*.

Example of mixtures

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

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

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

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

- Thus the first updated weight becomes

$$w_1' = \frac{0.5 \cdot \binom{4}{3} \frac{\text{B}(2.5+3, 2.5+1)}{\text{B}(2.5, 2.5)}}{0.5 \cdot \binom{4}{3} \frac{\text{B}(2.5+3, 2.5+1)}{\text{B}(2.5, 2.5)} + 0.5 \cdot \binom{4}{3} \frac{\text{B}(11+3, 31+1)}{\text{B}(11, 31)}} = 0.7975$$

  and for the second updated weight $w_2' = 1 - w_1' = 0.2025$.
- The posterior becomes

$$\pi(\theta \mid y = 3) = 0.7975 \cdot \text{Beta}(\theta; 2.5+3, 2.5+1) + 0.2025 \cdot \text{Beta}(\theta; 11+3, 31+1).$$

Multivariate conjugacy example: The normal likelihood, no parameters known

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

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

- 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 $\mu \mid \tau \sim \text{Normal}(\mu_0, 1/(\lambda\tau))$.

Computing the posterior

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

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

- Computing the posterior density using our proportionality method, the result is a Normal-Gamma density which can be expressed as

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

- Computations like these can get hairy; if you are lazy like me, consult, e.g., Wikipedia.
- Using improper prior $\pi(\mu, \tau) \propto_{\mu,\tau} 1/\tau$ gives posterior $\tau \mid x \sim \text{Gamma}(\frac{n-1}{2}, \frac{1}{2}\sum_{i=1}^{n}(x_i - \overline{x})^2)$ and $\mu \mid \tau, x \sim \text{Normal}(\overline{x}, \frac{1}{n\tau})$.
- NOTE: The expectation of the posterior for τ then becomes 1 divided by the classical variance estimator, and the expectation for μ becomes $\overline{x}$.

Predictive distributions

- Given parameters $\nu > 0$, μ, and σ^2, a real variable x has a **generalized t-distribution**, $x \sim \mathrm{t}(\nu, \mu, \sigma^2)$, when the density is

$$t(x; \nu, \mu, \sigma^2) = \frac{1}{\sqrt{\nu\sigma^2} B(\nu/2, 1/2)} \left[1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{2}}$$

- When $x \mid \tau \sim \mathrm{Normal}(\mu, \frac{1}{\lambda\tau})$ and $\tau \sim \mathrm{Gamma}(\alpha, \beta)$, the marginal (i.e. prior predictive) becomes

$$\pi(x) = \mathrm{t}\left(x; 2\alpha, \mu, \frac{\beta}{\alpha\lambda}\right)$$

- When $x \mid \mu, \tau \sim \mathrm{Normal}(\mu, 1/\tau)$, $\mu \mid \tau \sim \mathrm{Normal}(\mu_0, \frac{1}{\lambda\tau})$, and $\tau \sim \mathrm{Gamma}(\alpha, \beta)$, then the marginal becomes

$$\pi(x) = \mathrm{t}\left(x; 2\alpha, \mu_0, \frac{\beta(\lambda+1)}{\alpha\lambda}\right).$$

- To derive this, marginalize first over the normal-normal conjugacy.

Example: Normal observations

A Normal$(\mu, 1/\tau)$ distribution is investigated.
We use a prior $\pi(\mu, \tau) \propto_{\mu,\tau} 1/\tau$.

- First question: If observations are 3.1, 4.2, 2.9, 3.7, 3.9, find the posterior and the posterior predictive.
- Second question: Given the additional information that we must have $\mu \in [3, 3.5]$, find the posterior and the posterior predictive.
- Third question: Then given the additional observations 2.5, 2.1, and 4.0, find the posterior and the posterior predictive.
- **Presentation break for computations by hand**

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

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

- $\theta = (\theta_1, \ldots, \theta_n)$ with $\theta_i > 0$ and $\sum_{i=1}^n \theta_i = 1$ 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!
- With a Dirichlet$(\alpha_1, \ldots, \alpha_n)$ prior, one can show that the probability of observing a type i result in the next trial becomes

$$\frac{\alpha_i + x_i}{\sum_{j=1}^n (\alpha_j + x_j)}.$$

Applied example: Forensic DNA matches

- DNA matching between a trace and a person may be used as proof in criminal cases: For this, one needs to compute the strength of evidence when there is a match at some investigated *loci*.
- At an *STR locus* in a chromosome, a person has a particular *allele* (variant): Variants there differ by the number of repetitions of a short sequence (such as CAAT).
- The probability that a random person has a particular allele at this chromosome needs to be computed.
- To do so, population databases of alleles are collected. A small database might look like

10	11	12	13	14	15	16	17	18
1	0	5	89	143	9	3	0	2

- What is the probability that a random person has 17 repetitions as his allele?
- It is common to use the Multinomial-Dirichlet model together with *pseudocounts*, i.e., values for α_i, for example $\alpha_i = 0.5$ or $\alpha_i = 1$.
- Probabilities get a reasonable value, instead of zero.

The multivariate normal distribution

- We say X has a multivariate (n-variate) normal distribution, if it is a real 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 θ_1 and θ_2 is multivariate normal. Then

- If we integrate out one of them, e.g. θ_2, the marginal for θ_1 is multivariate normal. The parameters can be read off the expectation and the covariance matrix of the joint distribution.
- If we fix θ_2, then the *conditional distribution* $\theta_1 \mid \theta_2$ is also multivariate normal. In fact, if

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim \text{Normal}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}^{-1} \right)$$

we have

$$\theta_1 \mid \theta_2 \sim \text{Normal}(\mu_1 - P_{11}^{-1} P_{12}(Y - \mu_2), P_{11}^{-1})$$

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 θ_1 and θ_2, and rewrite it as two factors, one depending only on θ_2.