MSA101/MVE187 2021 Lecture 3
Low-dimensional Bayesian inference
Mixtures
Some multivariate conjugacies

Petter Mostad

Chalmers University

September 5, 2022

Review: Bayesian framework

- Prediction variable Y_{pred}, data Y_{data}, parameter θ.
- Specify a complete model by specifying prior $\pi(\theta)$, likelihood $\pi(Y_{data} \mid \theta)$, and prediction distribution $\pi(Y_{pred} \mid \theta)$.
- Derive the posterior $\pi(\theta \mid Y_{data})$.
- Make predictions using

$$\pi(Y_{pred} \mid Y_{data}) = \int \pi(Y_{pred} \mid \theta)\pi(\theta \mid Y_{data})\, d\theta$$

Ideas for practical computations

- Last time: Both likelihood and prior are from a list of elementary distributions, and are conjugate.
- Extension: Use discretization and computers: Works in low dimensions.
- Small extension: Use *mixtures* of priors.
- Small extension: Use multivariate conjugacies.
- Next time: Huge extension: Use simulation.

Bayesian inference with a discrete parameter θ

Assume θ has possible values $\theta_1, \ldots, \theta_n$.

- The prior $\pi(\theta)$ is represented as a vector $v = (v_1, \ldots, v_n)$:

$$v_i = \pi(\theta_i).$$

- The likelihood $\pi(y \mid \theta)$ is represented as a vector $w = (w_1, \ldots, w_n)$:

$$w_i = \pi(y \mid \theta_i).$$

- The posterior is represented as a vector $z = (z_1, \ldots, z_n)$:

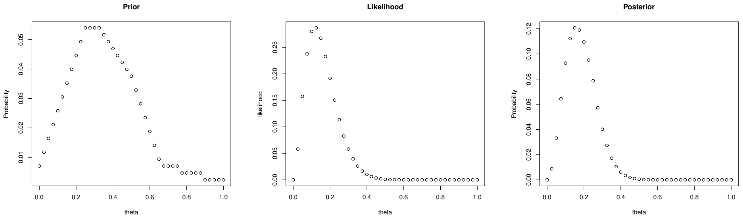
$$z_i = \frac{v_i \cdot w_i}{\sum_{j=1}^{n} v_j \cdot w_j}.$$

- The posterior predictive distribution can be computed for all values of Y_{pred} as a sum:

$$\pi(Y_{\text{pred}} \mid Y_{\text{data}}) = \sum_{i=1}^{n} \pi(Y_{pred} \mid \theta_i) z_i.$$

Example: An experimental production process

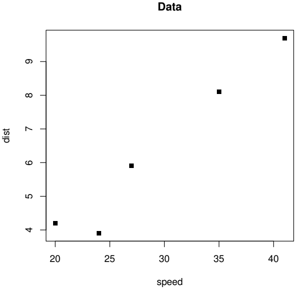
An experimental production process for an electronic component produces faulty components at a rate θ; 17 tests have produced 2 faulty components; you want to predict probability of at most 1 faulty component in the next batch of 10.

- Prior (constructed based on earlier experience)
- Likelihood: Binomial$(2; 17, \theta)$
- 

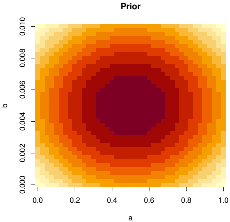
- Prediction
  $\sum_\theta (\text{Binomial}(0; 10, \theta) + \text{Binomial}(1; 10, \theta))\pi(\theta \mid \text{data}) = 0.4642503$

Example: Braking distance for a bike, depending on speed

Braking distance for a bike has been measured at 5 different speeds: Data is $(x_1, y_1), \dots, (x_5, y_5)$. At speed 30, what is the probability that breaking distance will be more than 5?

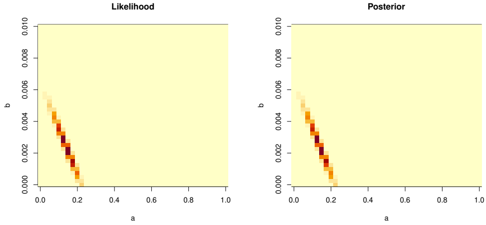


- Model: We assume $y_i \mid a, b \sim \text{Normal}(ax_i + bx_i^2, 0.8^2)$, and use a discrete prior on a grid for parameters (a, b).
- Prior:



Example: Braking distance for a bike

- Likelihood: $\pi(\text{data} \mid (a,b)) = \prod_{i=1}^{5} \text{Normal}(y_i; ax_i + bx_i^2, 0.8^2)$
- Likelihood and posterior computed:



- Prediction:

$$\begin{aligned}
& \Pr(y > 5 \mid x = 30, \text{data}) \\
= \quad & \sum_{a,b} \left(\int_5^\infty \text{Normal}(y; a30 + b30^2, 0.8^2)\, dy \right) \pi(a, b \mid \text{data}) \\
= \quad & 0.9396133
\end{aligned}$$

Choosing the prior

- Two main strategies: Aiming for *non-informative* or *informative* priors.
- Non-informative examples:
  - With conjugacies, using improper distributions like Gamma$(0, 0)$ or Beta$(0, 0)$
  - "Flat" densities...(but depends on scale!)
- Informative:
  - Use posteriors based on previous data, or
  - Check out the prior predictive: Does it "look reasonable" compared to what you expect for such data?

The curse of dimensionality

- **What happens with the discretization if θ is a high-dimensional variable?**
- In practice, we have to find other methods than discretization.

Numerical integration

The integrals of Bayesian inference

$$\pi(\theta \mid Y_{\text{data}}) = \frac{\pi(Y_{\text{data}} \mid \theta)\pi(\theta)}{\int_\theta \pi(Y_{\text{data}} \mid \theta)\pi(\theta)\, d\theta}$$

and

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

can be computed with numerical integration.

- Can work slightly better than discretization (after all discretization is a primitive form of numerical integration).
- Suffers from the same curse of dimensionality as discretization.

Mixtures

A density written as a linear combination of other densities is called a mixture (where $\sum_{i=1}^{n} \nu_i = 1$):

$$\pi(\theta) = \sum_{i=1}^{n} \nu_i \pi_i(\theta).$$

- Using a mixture prior gives a mixture prior predictive distribution:

$$\pi(y) = \int \pi(y \mid \theta) \sum_{i=1}^{n} \nu_i \pi_i(\theta)\, d\theta = \sum_{i=1}^{n} \nu_i \int \pi(y \mid \theta) \pi_i(\theta)\, d\theta = \sum_{i=1}^{n} \nu_i \pi_i(y).$$

- Defining $\pi_i(\theta \mid y) = \frac{\pi(y|\theta)\pi_i(\theta)}{\pi_i(y)}$, we also get a mixture posterior:

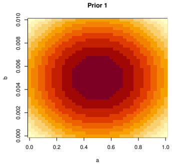
$$\begin{aligned}
\pi(\theta \mid y) &= \frac{\pi(y \mid \theta)\pi(\theta)}{\pi(y)} = \frac{\sum_{i=1}^{n} \nu_i \pi(y \mid \theta) \pi_i(\theta)}{\sum_{j=1}^{n} \nu_j \pi_j(y)} = \frac{\sum_{i=1}^{n} \nu_i \pi_i(y) \pi_i(\theta \mid y)}{\sum_{j=1}^{n} \nu_j \pi_j(y)} \\
&= \sum_{i=1}^{n} \left(\frac{\nu_i \pi_i(y)}{\sum_{j=1}^{n} \mu_j \pi_j(y)} \right) \pi_i(\theta \mid y).
\end{aligned}$$

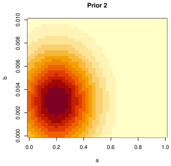
- Finally, a mixture posterior predictive distribution:

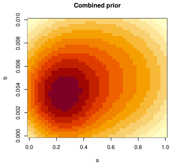
$$\pi(y_{\text{pred}} \mid y) = \int \pi(y_{\text{pred}} \mid \theta) \pi(\theta \mid y)\, d\theta = \sum_{i=1}^{n} \left(\frac{\nu_i \pi_i(y)}{\sum_{j=1}^{n} \mu_j \pi_j(y)} \right) \pi_i(y_{\text{pred}} \mid y).$$

Example: More on braking bikes

We now use the following pixture of priors:

$0.95\cdot$ 

$+0.05\cdot$ 

$=$ 

- We get the updated weights

$$\frac{0.95 \cdot \pi_1(\text{data})}{0.95 \cdot \pi_1(\text{data}) + 0.05 \cdot \pi_2(\text{data})} = 0.8530929$$

  for Prior 1 and $1 - 0.8530929 = 0.1469071$ for Prior 2.

- Using combined prior, the prediction hardly changes: 0.9399131 (before it was 0.9396133).

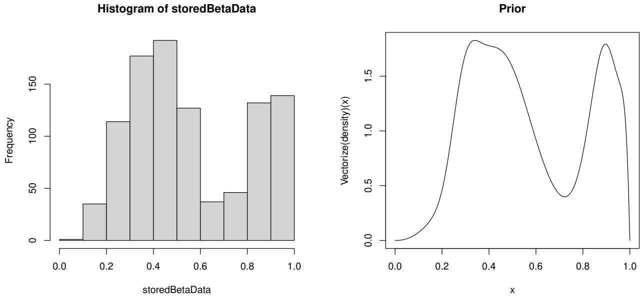
Mixtures when priors are conjugate

- When all the priors $\pi_1(\theta), \ldots, \pi_n(\theta)$ are conjugate to the likelihood $\pi(\text{data} \mid \theta)$, the mixture is also conjugate!
- Is a very powerful way to make families of conjugate priors more flexible!
- Note that we have formulas for all the weights and the posteriors occurring, no integration necessary.

Example: Using mixtures

If $y \sim \text{Geometric}(p)$ with $0 < p < 1$ then $\pi(y \mid p) = p(1-p)^y$, and $p \sim \text{Beta}(\alpha, \beta)$ is a conjugate family.

- If in some applied context our prior information is represented by, e.g., a histogram, we can model it as a Beta mixture:



-
- In this simple case, you could alternatively use discretization.

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.

Multinomial-Dirichlet conjugacy

- Assume $x = (x_1, \dots, x_n) \sim \text{Multinomial}(m, \theta_1, \theta_2, \dots, \theta_n)$, 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

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

- $\theta = (\theta_1, \dots, \theta_n)$ with $\theta_i > 0$ and $\sum_{i=1}^n \theta_i = 1$ has a Dirichlet distribution with parameters $\alpha_1, \dots, \alpha_n$ if the density can be written as

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

- Prove that the Dirichlet family is a conjugate family to the Multinomial likelhiood!
- With a Dirichlet$(\alpha_1, \dots, \alpha_n)$ 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.