Chapter 2
Bayesian Inference with Missing Data

Qing Zhou*

Contents

*UCLA Department of Statistics (email: zhou@stat.ucla.edu).

1. Bayesian Inference

Two major tasks of statistical inference is (i) to estimate unknown model parameters from data; (ii) to quantify the uncertainty in the estimates. Suppose we have collected data:

$$y_1, y_2, \cdots, y_n \overset{\text{iid}}{\sim} f(y \mid \theta),$$

where $f(y \mid \theta)$ is a pdf (or pmf) of a distribution parameterized by θ. Then we want to estimate θ and/or build a confidence interval for θ.

In general, denote the observed data by $\mathbf{y} = (y_1, y_2, \cdots, y_n)$. A common estimation method is the maximum likelihood estimate (MLE). Define the likelihood of θ give data $\mathbf{y}$ as

$$L(\theta \mid \mathbf{y}) := p(y_1, \cdots, y_n \mid \theta) = \prod_{i=1}^{n} f(y_i \mid \theta).$$

The MLE $\hat{\theta}_{\text{MLE}}$ is the maximizer of $L(\theta \mid \mathbf{y})$ over θ:

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}}\, L(\theta \mid \mathbf{y}).$$

Moreover, we often estimate the standard error of the MLE, denoted by $\widehat{\text{se}}$, and construct an approximate 95% confidence interval as

$$(\hat{\theta}_{\text{MLE}} - 2\widehat{\text{se}}, \hat{\theta}_{\text{MLE}} + 2\widehat{\text{se}})$$

as a way to quantify the uncertainty in our estimate. The interpretation of the interval is

$$\mathbb{P}[\theta \in (\hat{\theta}_{\text{MLE}} - 2\widehat{\text{se}}, \hat{\theta}_{\text{MLE}} + 2\widehat{\text{se}})] = 0.95.$$

Here, $\hat{\theta}_{\text{MLE}}$ is regarded as a random variable as a function of the random sample $\mathbf{y}$, while θ is an *unknown constant*.

1.1. Main steps

Bayesian inference relies on posterior distributions to provide solutions to the two inferential tasks (i) and (ii). The unknown parameter θ is regarded as a *random variable* and thus we need to specify a marginal distribution for θ, denoted by $p(\theta)$, which is called a prior distribution. Here, "prior" means before observing any data, as the prior distribution does not depend on the data $\mathbf{y}$. Therefore, a Bayesian model for the data $\mathbf{y}$ is set up by two distributions:

$$\text{Prior: } \theta \sim p(\theta), \tag{1}$$

$$\text{Data: } \mathbf{y} = (y_1, \ldots, y_n) \mid \theta \overset{\text{iid}}{\sim} f(y \mid \theta). \tag{2}$$

Together, they define a joint distribution for $(\theta, \mathbf{y})$:

$$p(\theta, \mathbf{y}) = p(\theta)p(\mathbf{y} \mid \theta) = p(\theta) \cdot \prod_{i=1}^{n} f(y_i \mid \theta). \tag{3}$$

Based on (3), we find the conditional distribution $[\theta \mid \mathbf{y}]$ to perform inference on θ. This conditional distribution of θ given the data $\mathbf{y}$ is called the posterior distribution, where "posterior" means the distribution of θ is now updated after observing the data and thus depends on $\mathbf{y}$. Applying Bayes formula,

$$p(\theta \mid \mathbf{y}) = \frac{p(\theta, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\theta)p(\mathbf{y} \mid \theta)}{p(\mathbf{y})} = \frac{p(\theta) \cdot \prod_{i=1}^{n} f(y_i \mid \theta)}{p(\mathbf{y})},$$

where the marginal density $p(\mathbf{y}) = \int p(\theta, \mathbf{y}) d\theta$ does not depend on θ and can be regarded as a normalizing constant. Consequently, it is more convenient to work with an unnormalized posterior density:

$$p(\theta \mid \mathbf{y}) \propto p(\theta)p(\mathbf{y} \mid \theta) = p(\theta) \cdot \prod_{i=1}^{n} f(y_i \mid \theta). \tag{4}$$

We may either recognize the posterior distribution via the unnormalized density on the right side or use Monte Carlo methods to draw samples given the unnormalized density.

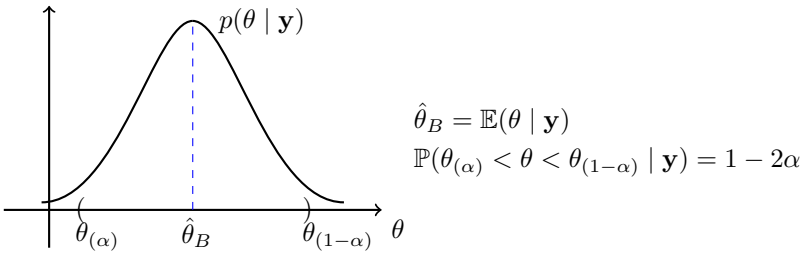
A Bayesian estimate of θ is usually constructed as the mean of the posterior distribution,

$$\hat{\theta}_B := \mathbb{E}(\theta|\mathbf{y}) = \int \theta \cdot p(\theta \mid \mathbf{y}) d\theta. \tag{5}$$

A $(1 - 2\alpha)$ Bayesian interval for θ can be constructed by the quantiles of the posterior distribution: $(\theta_{(\alpha)}, \theta_{(1-\alpha)})$, where $\theta_{(\alpha)}$ is the α-quantile for $\alpha \in (0, 1)$. The interpretation of a Bayesian interval is

$$\mathbb{P}(\theta \in (\theta_{(\alpha)}, \theta_{(1-\alpha)}) \mid \mathbf{y}) = 1 - 2\alpha, \tag{6}$$

where θ is a random variable following the posterior distribution $p(\theta \mid \mathbf{y})$.



$\hat{\theta}_B = \mathbb{E}(\theta \mid \mathbf{y})$
$\mathbb{P}(\theta_{(\alpha)} < \theta < \theta_{(1-\alpha)} \mid \mathbf{y}) = 1 - 2\alpha$

For complicated problems, Monte Carlo simulation, such as MCMC, is applied to draw samples of θ from the posterior distribution $p(\theta \mid \mathbf{y})$, regarding (4) as the target density. From the Monte Carlo samples, one can easily calculate the sample mean and sample quantiles to approximate $\hat{\theta}_B$ and $(\theta_{(\alpha)}, \theta_{(1-\alpha)})$.

In summary, the main steps of Bayesian inference are:

1. Choose a prior distribution $p(\theta)$.
2. Find the posterior distribution $p(\theta \mid \mathbf{y})$ by (4).
3. Apply a Monte Carlo algorithm to draw samples from $p(\theta \mid \mathbf{y})$.
4. Construct Bayesian estimates and intervals from the Monte Carlo samples.

1.2. Some basic models

We will demonstrate the main steps of Bayesian inference with a few simple examples.

Example 1 (Binomial distribution)**.** Consider independent coin tossing with $\theta \in (0, 1)$ being the probability of heads. Suppose we toss n times and observe heads x times. How to estimate θ?

Let X (random variable) be the number of times we observe heads. The distribution of X given θ is

$$X \mid \theta \sim \text{Bin}(n, \theta).$$

Thus, the likelihood

$$\mathbb{P}(X = x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

MLE: $\hat{\theta}_{\text{MLE}} = \dfrac{x}{n}$.

Bayesian inference:

1. Choose a prior distribution for θ: Without any prior knowledge on θ, we usually choose a flat prior,

$$\theta \sim \text{Unif}(0, 1), \quad \text{i.e. } p(\theta) = 1,\ \theta \in (0, 1).$$

2. Then find the posterior distribution:

$$\begin{aligned} p(\theta \mid X = x) &\propto p(\theta) \cdot \mathbb{P}(X = x \mid \theta) \\ &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \\ &\propto \theta^x (1-\theta)^{n-x}, \end{aligned} \tag{7}$$

where θ is the random variable.

3. From (7), we recognize that it is an unnormalized Beta density. Therefore, the posterior distribution is

$$\theta \mid x \sim \text{Beta}(x+1, n-x+1). \tag{8}$$

   As a reference, the pdf of $\text{Beta}(\alpha, \beta)$ is

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

   and its mean is $\mathbb{E}(\theta) = \frac{\alpha}{\alpha+\beta}$.

4. Given (8), we find Bayesian estimate

$$\hat{\theta}_B = \mathbb{E}(\theta|x) = \frac{x+1}{n+2}.$$

   To construct a 95% Bayesian interval, we use the 2.5% and 97.5% quantiles of $\text{Beta}(x+1, n-x+1)$. For example, if $n = 10, x = 3$, the posterior distribution is $\text{Beta}(4, 8)$, for which the two quantiles are

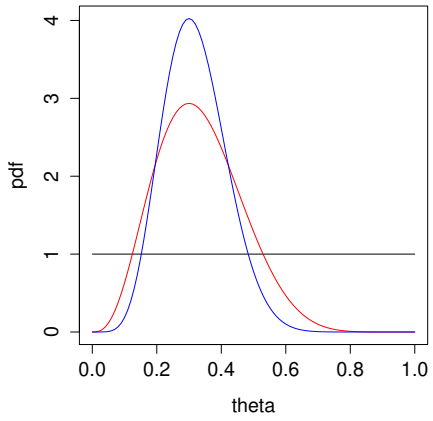
```
> qbeta(c(0.025,0.975),4,8)
[1] 0.1092634 0.6097426
```

   So the 95% Bayesian interval is $(0.109, 0.610)$. If $n = 20, x = 6$, the posterior distribution is $\text{Beta}(7, 15)$ with the quantiles given by

```
> qbeta(c(0.025,0.975),7,15)
[1] 0.1458769 0.5217511
```

   In this case, Bayesian interval is $(0.146, 0.522)$, which is shorter than the first case as the sample size n is larger.

The following figure shows the shape of the prior (black) and the posterior distributions: red for $n = 10, x = 3$ and blue for $n = 20, x = 6$.



A Bayesian interval can be used to do hypothesis test. Suppose we want to decide whether the coin is fair

$$H_0 : \theta = 0.5.$$

Based on the data $n = 20, x = 6$, the 95% Bayesian interval $(0.146, 0.522)$ covers 0.5, and therefore we will accept the null hypothesis H_0. If we collect more data and observe $n = 50, x = 15$, then $\theta \mid x \sim \text{Beta}(16, 36)$ and a 95% Bayesian interval will be $(0.191, 0.438)$. Because 0.5 falls outside this interval, we conclude with 95% probability that the coin is not fair (reject H_0).

The uniform distribution $\text{Unif}(0, 1)$ is equivalent to $\text{Beta}(1, 1)$. We may choose other Beta distribution as the prior for θ:

$$\theta \sim \text{Beta}(\alpha, \beta), \qquad p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Then the posterior distribution

$$\begin{aligned} p(\theta \mid X = x) &\propto p(\theta) \cdot \mathbb{P}(X = x \mid \theta) \\ &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \binom{n}{x} \theta^x (1-\theta)^{n-x} \\ &\propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}, \end{aligned}$$

and thus,

$$\theta \mid x \sim \text{Beta}(x + \alpha, n - x + \beta).$$

We see that the posterior is in the same family of the prior, both Beta distributions, in which case we say the prior is a *conjugate prior*. That is, Beta prior is conjugate to the Binomial distribution. The Bayesian estimate, i.e. the posterior mean, under this prior is

$$\hat{\theta}_B = \frac{x + \alpha}{n + \alpha + \beta}. \tag{9}$$

Compared to the MLE $\hat{\theta}_{\text{MLE}} = x/n$, the prior parameters (α, β) may be regarded as pseudo counts added to the two possible outcomes (heads or tails). If there is no prior knowledge about θ, we choose small pseudo counts, $\alpha, \beta \in (0, 1]$. If there is strong prior for θ, say from historical data, one may choose larger values of α, β to reflect such prior knowledge.

Example 2 (Multinomial distribution)**.** We generalize Example 1 to multinomial data. Let $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$ be the probabilities of k possible outcomes in a random experiment, $\theta_j > 0, \sum_{j=1}^k \theta_j = 1$. Suppose we have done this experiment n times independently and observed the jth outcome x_j times. So the observations follow a multinomial distribution:

$$\mathbf{x} = (x_1, x_2, \cdots, x_k) \mid \theta \sim \text{M}(n, \theta), \quad \sum x_j = n.$$

The likelihood is

$$p(\mathbf{x} \mid \theta) \propto \theta_1^{x_1}\theta_2^{x_2}\cdots\theta_k^{x_k}, \tag{10}$$

and the MLE

$$(\hat{\theta}_j)_{\text{MLE}} = \frac{x_j}{n}, \qquad j = 1, \ldots, k.$$

To do Bayesian inference, let us first find a conjugate prior.

Definition 1 (Dirichlet distribution)**.** Let $\theta = (\theta_1, \cdots, \theta_k) \in \mathbb{R}^k$ be a random vector such that $\theta_j \geq 0$ for all $j = 1, \ldots, k$ and $\sum_{j=1}^k \theta_j = 1$. Then θ follows the Dirichlet distribution $\text{Dir}(\alpha_1, \ldots, \alpha_k)$, $\alpha_j > 0$ for all j, if the pdf of θ is

$$p(\theta) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)}\theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1}\cdots\theta_k^{\alpha_k - 1}.$$

The mean of θ is

$$\mathbb{E}(\theta_j) = \frac{\alpha_j}{\alpha_1 + \alpha_2 + \cdots + \alpha_k}, \quad j = 1, \ldots, k. \tag{11}$$

How to sample θ from $\text{Dir}(\alpha_1, \ldots, \alpha_k)$?

1. Draw $v_j \sim \text{Gamma}(\alpha_j, 1)$ independently for $j = 1, \ldots, k$.
2. Put $S = \sum_{j=1}^k v_j$ and define

$$\theta_j = \frac{v_j}{S} = \frac{v_j}{v_1 + \cdots + v_k}, \qquad j = 1, \ldots, k.$$

Then $\theta = (\theta_1, \theta_2, \cdots, \theta_k) \sim \text{Dir}(\alpha_1, \alpha_2, \cdots, \alpha_k)$.

It turns out the Dirichlet is a conjugate prior for multinomial distribution. To see that, let us assume the prior is $\theta \sim \text{Dir}(\alpha_1, \cdots, \alpha_k)$, i.e.

$$p(\theta) \propto \theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1}\cdots\theta_k^{\alpha_k - 1}. \tag{12}$$

Then the posterior distribution, by multiplying (12) and (10),

$$\begin{aligned} p(\theta \mid \mathbf{x}) &\propto p(\theta)p(\mathbf{x} \mid \theta) \\ &\propto \theta_1^{x_1 + \alpha_1 - 1}\theta_2^{x_2 + \alpha_2 - 1}\cdots\theta_k^{x_k + \alpha_k - 1}, \end{aligned}$$

which is an unnormalized density of $\text{Dir}(x_1 + \alpha_1, \ldots, x_k + \alpha_k)$. Therefore,

$$\theta \mid \mathbf{x} \sim \text{ Dir}(x_1 + \alpha_1, \ldots, x_k + \alpha_k). \tag{13}$$

Put $\alpha_0 = \sum_{j=1}^k \alpha_j$. By (11), we find the Bayesian estimate of θ by the posterior mean:

$$(\hat{\theta}_j)_B = \frac{x_j + \alpha_j}{n + \alpha_0}, \quad j = 1, \ldots, k.$$

Similar to (9), here $\alpha_1, \cdots, \alpha_k$ are also interpreted as pseudo counts for the k possible outcomes. Without any prior knowledge, we choose $\alpha_j \in (0, 1]$. In particular, if $\alpha_j = 1$ for all j, the prior is a uniform distribution ($p(\theta) \propto 1$).

If we wish to build a Bayesian interval for θ_j, we can do so using the quantiles of the posterior distribution $[\theta_j \mid \mathbf{x}]$, which is simply a marginal distribution of the Dirichlet distribution (13). By properties of Dirichlet distributions, the marginal distribution is a Beta distribution:

$$\theta_j \mid \mathbf{x} \sim \text{Beta}(x_j + \alpha_j, n - x_j + \alpha_0 - \alpha_j).$$

Then we can use the same procedure in Example 1 to construct a Bayesian interval for each θ_j.

Example 3 (Normal data with known variance)**.** Suppose we have observed

$$y_1, \cdots y_n \mid \theta \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2),$$

where σ^2 is known. Our goal is to make inference on θ. The likelihood of θ is

$$\begin{aligned} p(y_1, \cdots, y_n \mid \theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \theta)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\right\}. \end{aligned}$$

The MLE $\hat{\theta}_{\text{MLE}} = \bar{y} = \frac{1}{n}\sum_i y_i$. The standard error (standard deviation) of $\bar{y}$ is $\text{se} = \sigma/\sqrt{n}$. Thus, we can construct a 95% confidence interval $(\bar{y} \pm 2\sigma/\sqrt{n})$.

Now consider Bayesian inference. A conjugate prior for θ is $\theta \sim \mathcal{N}(\mu_0, \tau_0^2)$. Let us consider a flat prior by choosing $\tau_0 \to \infty$:

$$p(\theta) \propto \exp\left\{-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right\} \to 1, \text{ as } \tau_0 \to \infty.$$

Then, the posterior distribution $[\theta \mid \mathbf{y} = (y_1, \ldots, y_n)]$ is

$$p(\theta \mid \mathbf{y}) \propto p(\theta)p(\mathbf{y}|\theta) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\theta - y_i)^2\right\}.$$

Recall that θ is the random variable and $\mathbf{y}$ is constant. Using the equality

$$\begin{aligned} \sum_{i=1}^{n}(\theta - y_i)^2 &= \sum_i (\theta - \bar{y} + \bar{y} - y_i)^2 \\ &= n(\theta - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2, \end{aligned}$$

we get

$$p(\theta \mid \mathbf{y}) \propto \exp\left\{-\frac{1}{2\sigma^2} n(\theta - \bar{y})^2\right\} = \exp\left\{-\frac{(\theta - \bar{y})^2}{2\sigma^2/n}\right\}.$$

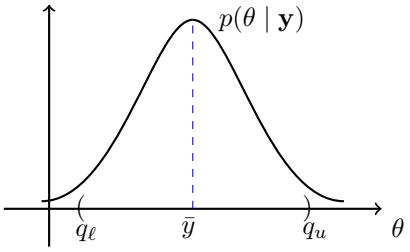
This shows that the posterior distribution

$$\theta \mid \mathbf{y} \sim \mathcal{N}(\bar{y}, \sigma^2/n).$$

Then, the Bayesian estimate is $\hat{\theta}_B = \mathbb{E}(\theta|\mathbf{y}) = \bar{y}$ and a 95% Bayesian interval, constructed by the quantiles (q_ℓ, q_u) of $\mathcal{N}(\bar{y}, \sigma^2/n)$, is

$$(\bar{y} - 2\sigma/\sqrt{n}, \bar{y} + 2\sigma/\sqrt{n}).$$

See below for illustration:



$\mathbb{E}(\theta \mid \mathbf{y}) = \bar{y}$
$\mathbb{P}(q_\ell < \theta < q_u \mid \mathbf{y}) = 0.95$

Again, the interval length $(4\sigma/\sqrt{n})$ shrinks when n increases. For this example, the Bayesian point and interval estimates both coincide with the MLE and the confidence interval.

2. Missing Data Problems

Suppose we have data

$$\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n \overset{\text{iid}}{\sim} f(\mathbf{y} \mid \theta),$$

where each data point $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{ip}) \in \mathbb{R}^p$. Put them into a data matrix $Y = (y_{ij})_{n \times p}$. However, some data points contain missing elements, shown as '?' in the following table, such as y_{2p} and y_{n1}.

	1	2	...	p
$\mathbf{y}_1$				
$\mathbf{y}_2$		?		?
...				
$\mathbf{y}_n$	?	?		

?: missing value (e.g. $y_{22}, y_{2p}, \cdots, y_{n2}$)
Y_{obs}: observed elements of Y (observed data).
Y_{mis}: missing elements of Y (missing data).
$Y = (Y_{obs}, Y_{mis})$: complete data.

Denote by Y_{obs} the observed elements of Y and Y_{mis} the missing elements of Y. We call Y_{obs} the observed data, Y_{mis} the missing data, and $Y = (Y_{obs}, Y_{mis})$ the complete data.

Assume the missing data mechanism is ignorable (Chapter 1, §1.1). Our goal is to estimate the model parameter θ based on the observed data Y_{obs}.

2.1. Data augmentation

Bayesian inference for missing data problems (1) estimates θ and (2) predicts missing data Y_{mis} based on the joint posterior distribution of (θ, Y_{mis}):

$$p(\theta, Y_{mis} | Y_{obs}) \propto p(\theta) p(Y_{obs}, Y_{mis} \mid \theta),$$

where $p(\theta)$ is the prior for θ and

$$p(Y_{obs}, Y_{mis} \mid \theta) = p(Y \mid \theta) = \prod_i f(\mathbf{y}_i \mid \theta)$$

is the complete-date likelihood.

Usually there are no closed-form formulas for posterior mean or quantiles of the posterior distribution of θ:

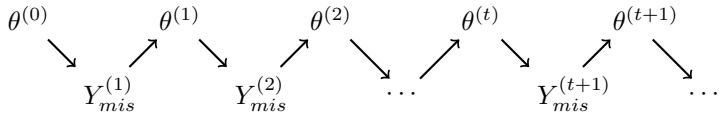
$$\begin{aligned} p(\theta \mid Y_{obs}) &\propto p(\theta) p(Y_{obs} \mid \theta) \\ &= p(\theta) \int p(Y_{obs}, Y_{mis} \mid \theta) dY_{mis}, \end{aligned}$$

which involves marginalization over the missing data Y_{mis}. We need to draw samples of (θ, Y_{mis}) from the joint posterior distribution $[\theta, Y_{mis} \mid Y_{obs}]$ to perform Bayesian inference. To do that, we develop a two-block Gibbs sampler, one iteration of which contains two conditional sampling steps:

1. Given $\theta^{(t)}$, draw $Y_{mis}^{(t+1)} \sim p(Y_{mis} \mid Y_{obs}, \theta^{(t)})$;

2. Given $Y_{mis}^{(t+1)}$, draw $\theta^{(t+1)} \sim p(\theta \mid Y_{obs}, Y_{mis}^{(t+1)}) = p(\theta \mid Y^{(t+1)})$, where $Y^{(t+1)} = (Y_{obs}, Y_{mis}^{(t+1)})$ is a complete data matrix with missing values imputed as $Y_{mis}^{(t+1)}$.

This two-block Gibbs sampler is illustrated by the following diagram:



Remark 1. This two-block Gibbs sampler can be viewed as a stochastic version of the EM algorithm and was first developed under the name of data augmentation by Tanner and Wong (1987).

For many commonly used models, both conditional sampling steps are easy to implement, as shown by the following examples.

2.2. Discrete data example

Example 4. Suppose $x_1, x_2, \cdots, x_n \overset{\text{iid}}{\sim} \text{Discrete}(\theta_1, \theta_2, \theta_3)$:

$$\mathbb{P}(x_i = k) = \theta_k, \ k = 1, 2, 3.$$

As shown in the following table, the data is coarsened, in which x_1, x_2, x_3 are only partially classified: $x_1 \in \{2, 3\}$, $x_2 \in \{1, 3\}$ and $x_3 \in \{1, 2\}$, while the other data points are fully classified: $x_4 = 1, \ldots, x_n = 2$.

	1	2	3
x_1	×	?	?
x_2	?	×	?
x_3	?	?	×
x_4	✓		
$\vdots$			
x_n		✓	

?: possible categories for an observation;

✓: observed category for an observation.

Prior: $\theta \sim Dir(\alpha_1, \alpha_2, \alpha_3), \quad (\theta_1 + \theta_2 + \theta_3 = 1)$

$$p(\theta_1, \theta_2, \theta_3) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \theta_3^{\alpha_3 - 1}.$$

Missing data in x_1, x_2, x_3, and $Y_{obs} = (x_1 \neq 1, x_2 \neq 2, x_3 \neq 3, x_4, \ldots, x_n)$. Let $C_j^{obs} = \sum_{i=4}^{n} I(x_i = j)$: observed counts for the jth category from x_4 to x_n.

1. Given $\theta = (\theta_1, \theta_2, \theta_3)$, $\mathbb{P}(x_1 = j|\theta) = \theta_j$ for $j = 1, 2, 3$,

$$\Rightarrow \mathbb{P}(x_1 = j | x_1 \neq 1, \theta) = \frac{\theta_j}{\theta_2 + \theta_3},\ j = 2, 3.$$

Similarly,

$$\mathbb{P}(x_2 = j | x_2 \neq 2, \theta) = \frac{\theta_j}{\theta_1 + \theta_3},\ j = 1, 3.$$
$$\mathbb{P}(x_3 = j | x_3 \neq 3, \theta) = \frac{\theta_j}{\theta_1 + \theta_2},\ j = 1, 2.$$

Draw x_1, x_2, x_3 independently according to the above conditional probabilities.

2. Given (x_1, x_2, x_3), $C_j^{(mis)} = \sum_{i=1}^{3} I(x_i = j)$,

then $p(\theta|x_1, \cdots, x_n) \propto \prod_{j=1}^{3} \theta_j^{C_j^{(Obs)} + C_j^{(mis)} + \alpha_j - 1}$. Draw θ from

$$\theta|\mathbf{x} \sim Dir(C_1^{(obs)} + C_1^{(mis)} + \alpha_1,\ C_2^{(obs)} + C_2^{(mis)} + \alpha_2,\ C_3^{(obs)} + C_3^{(mis)} + \alpha_3),$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ is complete data.

Iterate between steps 1 and 2 to generate $(\theta^{(t)}, x_{1,2,3}^{(t)})$ for $t = 1, \ldots, m$.

Bayesian estimates: $\hat{\theta}_B \approx \frac{1}{m} \sum_t \theta^{(t)}$ and histogram of $\theta_j^{(t)}$.

2.3. Gaussian data example

Example 5. $y_1, y_2, \cdots, y_n \overset{\text{iid}}{\sim} \mathcal{N}_2(\mu, \Sigma)$, $y_i = (y_{i1},\ y_{i2})$.

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \underbrace{\Sigma}_{\text{known}} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

	Y_1	Y_2
y_1	?	✓
y_2	✓	?
y_3	✓	✓
y_4	✓	✓
$\vdots$	$\vdots$	$\vdots$
y_n	✓	✓

? : missing value,
✓: observed value.

Improper flat prior: $p(\mu) \propto 1$.

Missing data $Y_{mis} = (y_{11},\ y_{22})$ and observed data $Y_{obs} = (y_{12}, y_{21}, y_3, \ldots, y_n)$.

Data augmentation for this problem:

1. Given μ, sample y_{11} and y_{22}, $[y_{11}|y_{12},\ \mu,\ \Sigma] \sim$? Recall $y_1 = (y_{11},\ y_{12})$.

$$
\begin{aligned}
&p(y_{11}|y_{12}, \mu, \Sigma) \propto p(y_{11}, y_{12}|\mu, \Sigma) \propto \exp\left[-\frac{1}{2}(y_1-\mu)^T\Sigma^{-1}(y_1-\mu)\right] \\
&= \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(y_{11}-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(y_{11}-\mu_1)(y_{12}-\mu_2)}{\sigma_1\sigma_2} + \frac{(y_{12}-\mu_2)^2}{\sigma_2^2}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2(1-\rho^2)\sigma_1^2}\left[(y_{11}-\mu_1)^2 - \frac{2\rho\sigma_1}{\sigma_2}(y_{12}-\mu_2)(y_{11}-\mu_1)\right]\right\} \\
&= \exp\left\{-\frac{1}{2(1-\rho^2)\sigma_1^2}\left[y_{11}-\mu_1-\frac{\rho\sigma_1}{\sigma_2}(y_{12}-\mu_2)\right]^2 + C\right\}.
\end{aligned}
$$

$\therefore y_{11}|y_{12}, \mu, \Sigma \sim \mathcal{N}\left(\mu_1 + \frac{\rho\sigma_1}{\sigma_2}(y_{12}-\mu_2), (1-\rho^2)\sigma_1^2\right)$.

Similarly, $y_{22}|y_{21}, \mu, \Sigma \sim \mathcal{N}\left(\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(y_{21}-\mu_1), (1-\rho^2)\sigma_2^2\right)$.

Given μ, draw y_{11} and y_{22} independently from the two normal distributions.

2. Given y_{11} and y_{22}, sample μ?

$$
\begin{aligned}
p(\mu|y_1, y_2, \cdots, y_n, \Sigma) &\propto p(y_1, \cdots, y_n|\mu, \Sigma) \\
&= |2\pi\Sigma|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)\right] \\
&\propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)\right].
\end{aligned}
$$

Let $\bar{y} = \sum_i y_i/n$.

$$
\begin{aligned}
&\sum_i (\mu-y_i)^T\Sigma^{-1}(\mu-y_i) \\
&= \sum_i (\mu-\bar{y}+\bar{y}-y_i)^T\Sigma^{-1}(\mu-\bar{y}+\bar{y}-y_i) \\
&= \sum_i \left[(\mu-\bar{y})^T\Sigma^{-1}(\mu-\bar{y}) + 2(\mu-\bar{y})^T\Sigma^{-1}(\bar{y}-y_i) + (\bar{y}-y_i)^T\Sigma^{-1}(\bar{y}-y_i)\right] \\
&= n(\mu-\bar{y})^T\Sigma^{-1}(\mu-\bar{y}) + C.
\end{aligned}
$$

Therefore, $\mu|y_1, \cdots, y_n \sim \mathcal{N}_2(\bar{y}, \frac{1}{n}\Sigma)$.

Iterate between steps 1 and 2.

References

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82** 528–540.