Data augmentation

Motivation: why missing data is awkward

- Real datasets are *rarely complete*: surveys, medical records, transaction logs, policing data, etc.
- "Frequentist quick fixes" often used in practice:
  1. **Discard** incomplete cases (complete-case analysis).
  2. **Impute** a value (mean/median, regression, last-observation-carried-forward, etc.) and then proceed "as if observed".
- These are often unsatisfying:
  - Discarding throws away information and can introduce bias.
  - Naive imputation can understate uncertainty (it pretends we truly observed what we did not).

Missingness mechanisms (high level)

- There is a large literature on how data go missing:
  - **MCAR**: missing completely at random (missingness independent of everything).
  - **MAR**: missing at random (missingness may depend on observed data, not on the missing values themselves).
  - **MNAR**: missing not at random (missingness depends on the missing values).
- In this section we focus on a practical computational message:

Treat missing / censored quantities as random variables and sample them.

Bayesian idea: promote missing data to an unknown variable

Suppose we have

parameter of interest θ, observed data y_{obs}, missing data y_{mis}.

Bayesian modelling says:

$$\pi(\theta, y_{\text{mis}} \mid y_{\text{obs}}) \propto \underbrace{\pi(y_{\text{obs}}, y_{\text{mis}} \mid \theta)}_{\text{complete-data likelihood}} \underbrace{\pi(\theta)}_{\text{prior}}.$$

- The missing value(s) y_{mis} are *unknown*, so we infer them jointly with θ.
- This naturally carries uncertainty about the missingness into posterior summaries.

Why Chapter 6 needs Chapter 5 (MCMC)

The above posterior is typically not available in closed form.

We therefore aim to construct an MCMC algorithm that samples from

$$\pi(\theta, y_{\text{mis}} \mid y_{\text{obs}}).$$

- Often we use a **Gibbs sampler** or **Metropolis-within-Gibbs**:

  $$\theta \mid y_{\text{mis}}, y_{\text{obs}} \qquad \text{and} \qquad y_{\text{mis}} \mid \theta, y_{\text{obs}}$$

  are updated in turn.
- Key benefit: we avoid an intractable likelihood by working with the *complete-data* model.

Observed-data vs complete-data likelihoods

Definition 6.11 (Observed-data likelihood).
Given a parametric model indexed by θ and observed data y_{obs}, the *observed-data likelihood* is

$$L_{\text{obs}}(\theta; y_{\text{obs}}) \equiv \pi(y_{\text{obs}} \mid \theta).$$

Definition 6.12 (Complete-data likelihood).
If we also introduce missing/unobserved data y_{mis}, then the *complete-data likelihood* is

$$L_{\text{comp}}(\theta; y_{\text{obs}}, y_{\text{mis}}) \equiv \pi(y_{\text{obs}}, y_{\text{mis}} \mid \theta).$$

- L_{comp} is sometimes called the **augmented likelihood**.

A toy example: X observed, Y unobserved

Suppose (X, Y) has a joint model $\pi(x, y \mid \theta)$.

We observe $X = x$ but do not observe Y.

Complete-data likelihood

$$L_{\text{comp}}(\theta; x, y) = \pi(X = x, Y = y \mid \theta).$$

Observed-data likelihood

$$L_{\text{obs}}(\theta; x) = \pi(X = x \mid \theta).$$

But these are linked by *marginalisation*:

$$\pi(X = x \mid \theta) = \int \pi(X = x, Y = y \mid \theta)\, dy,$$

(or a sum if Y is discrete).

Why the observed-data likelihood is hard

If

$$\pi(X = x \mid \theta) = \int \pi(X = x, Y = y \mid \theta)\, dy,$$

then **maximum likelihood** needs

$$\frac{d}{d\theta}\pi(X = x \mid \theta) = \frac{d}{d\theta}\int \pi(X = x, Y = y \mid \theta)\, dy.$$

- The dependence on θ is *inside the integral/sum*.
- This often yields messy expressions and expensive computation.
- Similar pain occurs in MCMC if each Metropolis–Hastings step requires evaluating the integral/sum.

Takeaway: the observed-data likelihood is the one we can write down, but it is often the one we cannot work with efficiently.

Data augmentation: the core computational trick

- Introduce y_{mis} explicitly and target the joint posterior:

$$\pi(\theta, y_{\text{mis}} \mid y_{\text{obs}}) \propto \pi(y_{\text{obs}}, y_{\text{mis}} \mid \theta)\, \pi(\theta).$$

- Then design MCMC updates for:

$$\theta \mid y_{\text{obs}}, y_{\text{mis}} \qquad \text{and} \qquad y_{\text{mis}} \mid \theta, y_{\text{obs}}.$$

- This replaces an intractable marginal likelihood by tractable complete-data pieces.

What you gain

You gain access to the same likelihood form as if you had fully observed the data *at the cost of* sampling additional variables.

Example 6.2: censored bank transaction counts

A bank checks transactions in batches of $n = 1000$.

- Let p be the probability that a transaction is suspicious.
- Let Y_i be the number of suspicious transactions in batch i.
- Model:

$$Y_i \mid p \sim \text{Binomial}(1000, p), \qquad i = 1, \ldots, 5.$$

- We observe $Y_1, \ldots, Y_4$ fully.
- Batch 5 is corrupted, but we know it is **censored**:

$$Y_5 < 6 \quad (\text{i.e. } Y_5 \in \{0, 1, 2, 3, 4, 5\}).$$

Observed-data likelihood for the censored model

Write $y_{1:4} = (y_1, y_2, y_3, y_4)$.

The observed information is:

$$(Y_1, \ldots, Y_4) = (y_1, \ldots, y_4), \qquad Y_5 < 6.$$

Observed-data likelihood

$$\begin{aligned}
\pi(y_{1:4}, Y_5 < 6 \mid p) &= \left[\prod_{i=1}^{4} \binom{1000}{y_i} p^{y_i} (1-p)^{1000-y_i} \right] \mathbb{P}(Y_5 < 6 \mid p) \\
&= \left[\prod_{i=1}^{4} \binom{1000}{y_i} p^{y_i} (1-p)^{1000-y_i} \right] \sum_{j=0}^{5} \binom{1000}{j} p^{j} (1-p)^{1000-j}.
\end{aligned}$$

- This is valid but involves an extra (ugly) term.

Posterior with a Beta prior (not conjugate here)

A natural prior is Beta because $p \in (0,1)$:

$$p \sim \text{Beta}(a,b), \qquad \pi(p) \propto p^{a-1}(1-p)^{b-1}.$$

Then the posterior is

$$\pi(p \mid y_{1:4}, Y_5 < 6) \propto \pi(y_{1:4}, Y_5 < 6 \mid p)\, \pi(p).$$

- If Y_5 were fully observed, the Beta prior would be conjugate.
- With censoring ($Y_5 < 6$), conjugacy breaks because of $\mathbb{P}(Y_5 < 6 \mid p)$.

Complete-data likelihood (the "wish list")

If we *had* observed $Y_5 = y_5$ as well, the likelihood would be

$$\pi(y_{1:5} \mid p) = \prod_{i=1}^{5} \binom{1000}{y_i} p^{y_i} (1-p)^{1000-y_i}.$$

This is simple and matches the standard Binomial model.

Data augmentation move

Treat Y_5 as an *unknown* variable constrained to $\{0, 1, 2, 3, 4, 5\}$, and sample it inside an MCMC algorithm.

Augmented posterior: sample (p, Y_5) jointly

We now target

$$\pi(p, y_5 \mid y_{1:4},\, y_5 < 6) \propto \pi(y_{1:5} \mid p)\, \pi(p), \quad \text{with } y_5 \in \{0, 1, 2, 3, 4, 5\}.$$

- Notice: we do *not* add a separate prior for Y_5.
- Y_5 is a data variable; its distribution is already specified by the likelihood model.
- The only additional information is the constraint $Y_5 < 6$ (support restriction).

Full conditional for p (general Beta prior)

Given $y_{1:5}$, the complete-data likelihood is Binomial product, so:

$$\pi(p \mid y_{1:5}) \propto \left[\prod_{i=1}^{5} p^{y_i} (1-p)^{1000-y_i} \right] p^{a-1} (1-p)^{b-1}.$$

Collect powers:

$$\pi(p \mid y_{1:5}) \propto p^{\left(\sum_{i=1}^{5} y_i\right)+a-1} \, (1-p)^{\left(5000-\sum_{i=1}^{5} y_i\right)+b-1}.$$

Result

$$p \mid y_{1:5} \sim \text{Beta}\left(a + \sum_{i=1}^{5} y_i, \;\; b + 5000 - \sum_{i=1}^{5} y_i \right).$$

Special case: Uniform prior $p \sim \text{Unif}(0, 1)$

Uniform$(0, 1)$ corresponds to Beta$(1, 1)$, i.e. $a = b = 1$.

Then

$$p \mid y_{1:5} \sim \text{Beta}\left(1 + \sum_{i=1}^{5} y_i,\ \ 1 + 5000 - \sum_{i=1}^{5} y_i\right).$$

- This is the clean "conjugate-looking" update you would use *if* Y_5 *were known*.
- In augmentation, we *make* Y_5 known by sampling it.

Full conditional for Y_5 (a truncated Binomial)

We want

$$\pi(y_5 \mid p, y_{1:4}, y_5 < 6).$$

Given p, the batches are independent, so conditioning on $y_{1:4}$ is irrelevant for Y_5:

$$\pi(y_5 \mid p, y_{1:4}, y_5 < 6) = \pi(y_5 \mid p, y_5 < 6).$$

By conditional probability (and truncation),

$$\pi(y_5 \mid p, y_5 < 6) = \frac{\pi(y_5 \mid p)}{\mathbb{P}(Y_5 < 6 \mid p)} \quad \text{for } y_5 \in \{0, 1, 2, 3, 4, 5\}.$$

Here

$$\pi(y_5 \mid p) = \binom{1000}{y_5} p^{y_5}(1-p)^{1000-y_5}, \quad \mathbb{P}(Y_5 < 6 \mid p) = \sum_{j=0}^{5} \binom{1000}{j} p^j (1-p)^{1000-j}.$$

How to sample Y_5 in practice

Y_5 takes only 6 possible values: $\{0, 1, 2, 3, 4, 5\}$.

1. Compute unnormalised weights for $k = 0, 1, 2, 3, 4, 5$:

$$w_k \propto \binom{1000}{k} p^k (1-p)^{1000-k}.$$

2. Normalise:

$$\tilde{w}_k = \frac{w_k}{\sum_{j=0}^{5} w_j}.$$

3. Draw Y_5 from the discrete distribution

$$\mathbb{P}(Y_5 = k \mid p, Y_5 < 6) = \tilde{w}_k, \qquad k = 0, \ldots, 5.$$

Interpretation

This is a Binomial$(1000, p)$ distribution **truncated** to $\{0, \ldots, 5\}$.

Gibbs sampler for the augmented model

We now have two full conditionals:

$$p \mid y_{1:5} \sim \text{Beta}\left(a + \sum_{i=1}^{5} y_i,\ \ b + 5000 - \sum_{i=1}^{5} y_i\right),$$

$$Y_5 \mid p,\ (Y_5 < 6) \sim \text{Truncated-Binomial}(1000, p; \{0, \ldots, 5\}).$$

Algorithm (Gibbs)

1. Initialise $p^{(0)} \in (0,1)$ and $y_5^{(0)} \in \{0, \ldots, 5\}$.
2. For $t = 1, 2, \ldots$:
   1. Sample
   $$p^{(t)} \sim \pi(p \mid y_1, \ldots, y_4, y_5^{(t-1)}).$$
   2. Sample
   $$y_5^{(t)} \sim \pi(y_5 \mid p^{(t)},\ y_5 < 6).$$

What does this buy us?

- We **avoid** repeatedly evaluating the observed-data likelihood term $\mathbb{P}(Y_5 < 6 \mid p)$ inside a Metropolis ratio.
- Each step becomes straightforward:
  - p update is a standard Beta draw (fast).
  - Y_5 update is a small discrete draw on $\{0, \ldots, 5\}$ (very fast).
- Posterior uncertainty in Y_5 is **propagated** into uncertainty about p.

Summary

Data augmentation turns a hard marginal likelihood problem into a larger but easy-to-sample joint problem.

Posterior summaries from the Gibbs output

From samples $\{(p^{(t)}, y_5^{(t)})\}_{t=1}^{T}$ (after burn-in):

- Posterior mean of p:

$$\mathbb{E}[p \mid y_{1:4}, Y_5 < 6] \approx \frac{1}{T} \sum_{t=1}^{T} p^{(t)}.$$

- Credible interval for p: use empirical quantiles of $\{p^{(t)}\}$.
- Posterior distribution of the censored count:

$$\mathbb{P}(Y_5 = k \mid y_{1:4}, Y_5 < 6) \approx \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{y_5^{(t)} = k\}.$$

Interpretation: you are not "filling in a single imputed value": you are learning a *distribution* over the missing/censored quantity.

Big picture: observed vs complete likelihoods

- Observed-data likelihood:

$$\pi(y_{\text{obs}} \mid \theta) = \int \pi(y_{\text{obs}}, y_{\text{mis}} \mid \theta)\, dy_{\text{mis}} \quad \text{(or sum)}.$$

- Complete-data likelihood:

$$\pi(y_{\text{obs}}, y_{\text{mis}} \mid \theta) \quad \text{(usually simple / standard family)}.$$

- Data augmentation + MCMC:

$$\text{sample } (\theta, y_{\text{mis}}) \text{ from } \pi(\theta, y_{\text{mis}} \mid y_{\text{obs}}).$$

One-sentence takeaway

When missingness makes the observed likelihood messy, augment the data and sample the missing pieces.