Two Applications of Multiple Imputation

Kole Butterer

May 4, 2025

Abstract

The multiple imputation (MI) algorithm serves two purposes: (1) to simplify the sampling of an intractable or unwieldy posterior, or (2) better approximate the posterior when data are partially missing. In both scenarios, we augment the observed data Y with a latent variable Z. We take advantage of a tractable augmented posterior $p(\theta|Y, Z)$ to recover $p(\theta|Y)$. We show an example for each purpose of MI. The first example simplifies the posterior of the genetic linkage model. The second example demonstrates how right-censored regression improves with MI in a small sample setting. The examples are adapted from *Tools for Statistical Inference* (Tanner, 1996).

1 Multiple imputation

In using the multiple imputation method, we augment the observed data Y with a latent variable Z. We take advantage of a tractable augmented posterior $p(\theta|Y,Z)$ to recover $p(\theta|Y)$. Specifically, assuming we can analytically evaluate or sample $p(\theta|Y,Z)$, we can integrate Z out yielding $p(\theta|Y)$ i.e.

$$p(\theta|Y) = \int_Z p(\theta|Y,Z)p(Z|Y)dZ. \tag{1}$$

We call (1) the posterior identity. But in order to evaluate the posterior identity, we need the form of $p(Z|Y)$. We can write

$$p(Z|Y) = \int_\theta p(Z|Y,\theta)p(\theta|Y)d\theta. \tag{2}$$

We call (2) the predictive identity. Note that $p(Z|Y)$ depends on $p(\theta|Y)$, and $p(\theta|Y)$ in turn depends on $p(Z|Y)$. Starting at an initial approximation to $p(\theta|Y)$, we can alternate between sampling latent variables Z using the predictive identity, and sampling θ using the posterior identity. This way, we iteratively improve our approximation to $p(\theta|Y)$. More precisely, the algorithm works as follows:

1. Sample $\theta_1^*...,\theta_m^*$ from the current approximation to the nonaugmented posterior $p(\theta|Y)$.
2. Sample z_j from the conditional predictive distribution $p(Z|\theta_j^*,Y)$ for each $j = 1,...,m$.
3. Update $p(\theta|Y)$ as the mixture of augmented posteriors so that $p(\theta|Y) = \frac{1}{m}\sum_{j=1}^{m} p(\theta|z_j,Y)$.
4. Repeat steps (1)-(3) until the approximation converges to $p(\theta|Y)$.

To be clear, each z_j for $j = 1,...,m$ determines an augmented posterior, from which we generate a corresponding θ_j^*. The approximation to the nonaugmented posterior we wish to recover is the mixture of these augmented posteriors. Meaning, our approximation is represented by θ_1^*, ..., θ_m^*, where each θ_j^* potentially comes from a different augmented posterior distribution. We demonstrate this process concretely below using the genetic linkage model.

1.1 Simplifying genetic linkage posterior

Consider the genetic linkage model. Suppose 197 animals are distributed into four categories as follows:

$$Y = (y_1,y_2,y_3,y_4) = (125,18,20,34)$$

according to cell probabilities

$$\left(\frac{1}{2}+\frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right).$$

We want to find the posterior for θ. Under a flat prior, the posterior distribution is

$$p(\theta|Y) \propto (2+\theta)^{y_1}(1-\theta)^{y_2+y_3}\theta^{y_4}.$$

Analytically evaluating expectations with respect to θ here is too complicated, and we have no way to directly sample θ either. But we can simplify the form of the posterior using data augmentation.

We augment the observed data Y with latent variable Z. We construct Z by spliting the first cell of Y into two cells as follows:

$$y_1 = z_1 + z_2,\ y_2 = z_3,\ y_3 = z_4,\ y_4 = z_5,$$

and giving Z cell probabilities

$$\left(\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right).$$

Under a flat prior, the augmented posterior is given by

$$p(\theta|Y,Z) \propto \theta^{z_2+z_5}(1-\theta)^{z_3+z_4}.$$

The augmented posterior admits a much simpler form. In fact, the augmented posterior is a beta distribution, which yields tractable expectations and direct sampling! Once we have the form for the conditional predictive distribution, we can employ the data augmentation algorithm.

The conditional predictive distribution $p(z_2|Y,\theta)$ for z_2 is given by

$$z_2 \sim \text{Binom}\left(n = 125, p = \frac{\theta}{\theta+2}\right).$$

Intuitively, we are flipping a coin to determine where to place the 125 counts originally belonging to y_1—each count goes to either z_1 or z_2. The probability of each count being placed in z_1 is given by the probability of being placed in z_1 divided by the probability of being in z_1 or z_2 which simplifies to $\frac{\theta}{\theta+2}$.

At each iteration of the algorithm, we generate $\theta_1^*, ..., \theta_m^*$ from the current approximation to the posterior. Then we generate $z_2^{(1)}, ..., z_2^{(m)}$ from the conditional predictive distribution using each $\theta_1^*, ..., \theta_m^*$. And we update the posterior distribution accordingly by sampling the beta distribution dependent on each $z_2^{(1)}, ..., z_2^{(m)}$. We repeat these two steps—impute, then update—until the algorithm converges. At that point, we have a sample from the desired posterior distribution! In this case, once the algorithm has been initialized, it works as follows:

1. Sample $\theta_1^*..., \theta_m^*$ from the current approximation to the posterior $p(\theta|Y)$.
2. Draw $z_2^{(i)} \sim \text{Binom}\left(125, \frac{\theta_i^*}{\theta_i^*+2}\right)$ for each $i = 1, ..., m$.
3. Update the approximation to the posterior as the mixture of beta distributions i.e. $p(\theta|Y) = \frac{1}{m}\sum_{i=1}^{m} \text{Beta}(z_2^{(i)} + z_5 + 1, z_3 + z_4 + 1)$.
4. Repeat steps (1)-(3) until the approximation to the posterior converges.

To initialize the algorithm, you might draw each θ^* from a uniform distribution over the interval (0, 1). To sample from the mixture of beta distributions, randomly select z_2^* from the imputed values $z_2^{(1)}, ..., z_2^{(m)}$ and then draw from $\text{Beta}(z_2^* + z_5 + 1, z_3 + z_4 + 1)$. Letting

$$x_1^{(i)} = z_2^* + z_5 + 1 \text{ and } x_2^{(i)} = z_3 + z_4 + 1,$$

we would draw the ith sample of θ^* from

$$\text{Beta}(x_1^{(i)}, x_2^{(i)})(\theta) = \frac{\Gamma(x_1^{(i)} + x_2^{(i)})}{\Gamma(x_1^{(i)})\Gamma(x_2^{(i)})}\theta^{x_1^{(i)}-1}(1-\theta)^{x_2^{(i)}-1}.$$

We can see in **Figure 1a** that the sample from the final augmented posterior agrees well with the true analytical, non-augmented posterior. Figure 1b emphasizes the dependence between θ_i^* and $z_2^{(i)}$. If we were to show the plots in **Figure 1b** across iterations (not just imputations), we would expect the black bands to occupy different regions of the parameter space over each iteration, and similarly for the red bands—especially in early iterations.

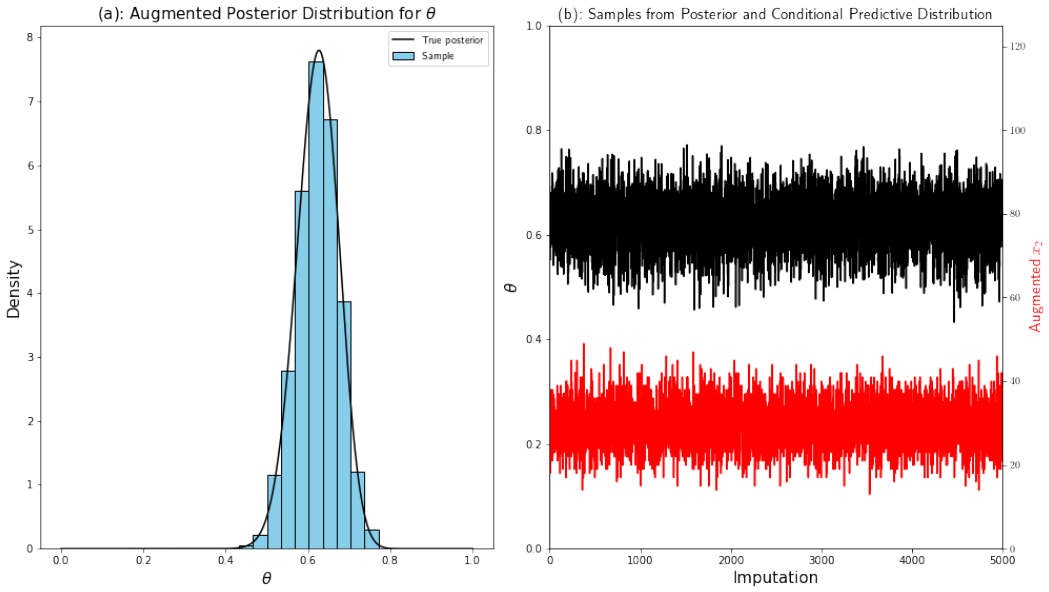

Figure 1: (a) Approximating true posterior with sample from mixture of augmented posteriors, (b) samples from mixture of augmented posteriors (black) and conditional predictive distributions for z_2 (red).

1.2 Partially missing data: right-censored regression

Moving to the second application of data augmentation, we demonstrate how to better approximate a posterior in cases where we have "incomplete" data. In the following data set, each patient had a heart transplant. The survival times and age at the time of the transplant were recorded. We want to regress log survival time onto age. However, half of the patients survived past the length of the study; we do not know the true value of their survival time.

At each iteration of the algorithm, we generate our best guess for the values of the missing data using the current approximation to the posterior distribution. Then we regress log survival time onto age using the complete data, and update the approximation to the posterior distribution as the mixture of the posteriors conditional on varying samples of augmented data. We repeat these steps—impute, then update—until the posterior approximations converge. In the end, we have a sample from a better approximation to the true posterior. Under a flat prior for the regression parameters, the algorithm works as follows once it has been initialized:

1. Sample the tuple $(\beta_0^*, \beta_1^*, \sigma_*^2)$ from the current approximation to the nonaugmented posterior $p(\theta|Y)$.
    a. Sample $\sigma_*^2 \sim$ Inverse-Gamma(α, γ) for $\alpha = \frac{1}{2}(n-k-1)$ and $\gamma = \frac{1}{2}\hat{\sigma}^2(n-k)$ where $\hat{\sigma}^2$ is the current sample error variance corresponding to the current regression line.
    b. Sample $(\beta_0^*, \beta_1^*) \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, \sigma_*^2 \mathbf{X^T X})$ where $\hat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$ i.e. the least squares solution to the regression problem.
2. Sample z from the conditional predictive distribution $p(z|\beta_0^*, \beta_1^*, \sigma_*^2, Y, X)$ i.e. sample $z \sim \mathcal{N}(\mathbf{X}^{(\text{unknown})}\beta^*, \sigma_*^2\mathbf{I})$ such that each imputed survival time is larger than the time of recording.
3. Repeat steps (1) and (2) m times yielding $z_1, ..., z_m$.
4. Update the approximation to the posterior as the mixture of augmented posteriors i.e. $p(\beta_0^*, \beta_1^*, \sigma_*^2|Y, X) = \frac{1}{m}\sum_{i=1}^{m} p(\beta_0^*, \beta_1^*, \sigma_*^2|z_j, Y, X)$.

We animate how multiple imputation works over 10 imputations during one iteration of the algorithm

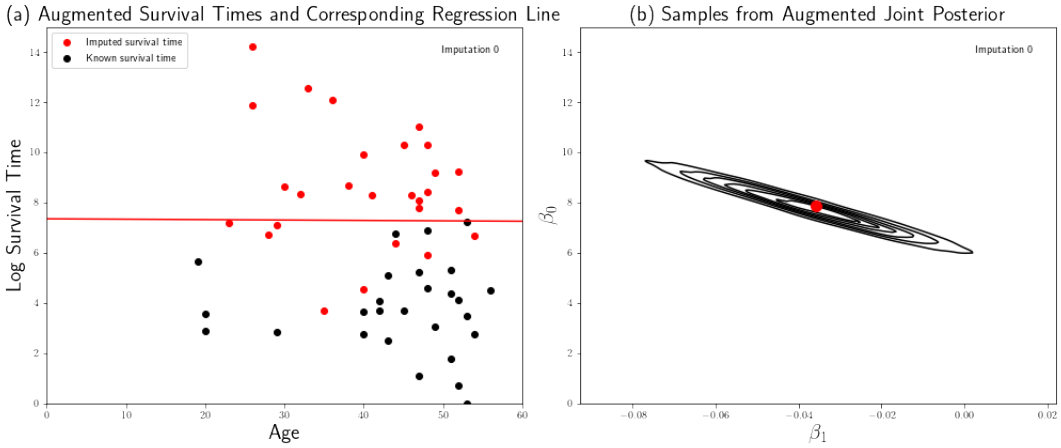

Figure 2: (a) Augmented and known survival times with augmented regression line, (b) sampled regression line from mixture of joint augmented posteriors.

in **Figure 2**. **Figure 2a** shows the augmented survival times in red, and the known survival times in black, as well as the regression line sampled from the mixture of joint augmented posteriors (for one fixed iteration) on the right. The regression line is also represented by the red dot on the mixture of joint augmented posteriors in Figure 2b.

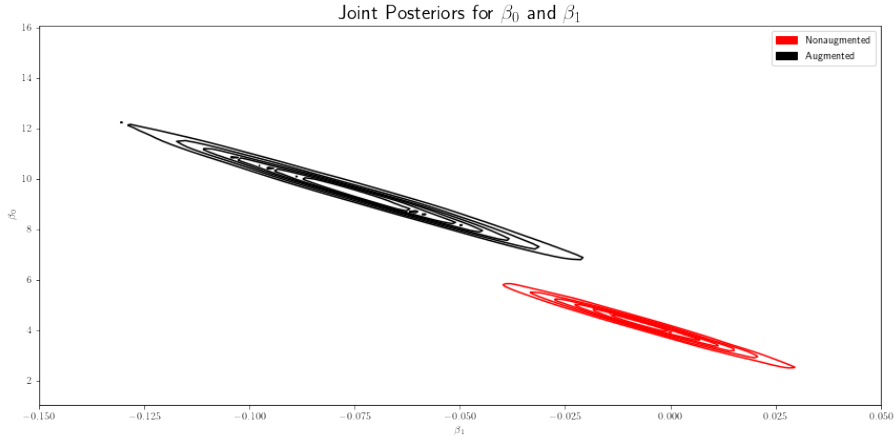

Figure 3: Comparing final mixture of augmented posteriors to initial, non-augmented posterior.

In **Figure 3**, we compare the final mixture of augmented posteriors to the initial, non-augmented posterior distribution for the regression parameters using only the known survival times.

In **Figure 4**, we compare the marginal augmented and non-augmented posteriors for each of the regression parameters.

1.3 Conclusion

We can use the data augmentation to make intractable posteriors tractable, and better estimate posterior distributions when data is partially missing. The augmented and non-augmented posteriors can be quite different in the case of censored regression data.

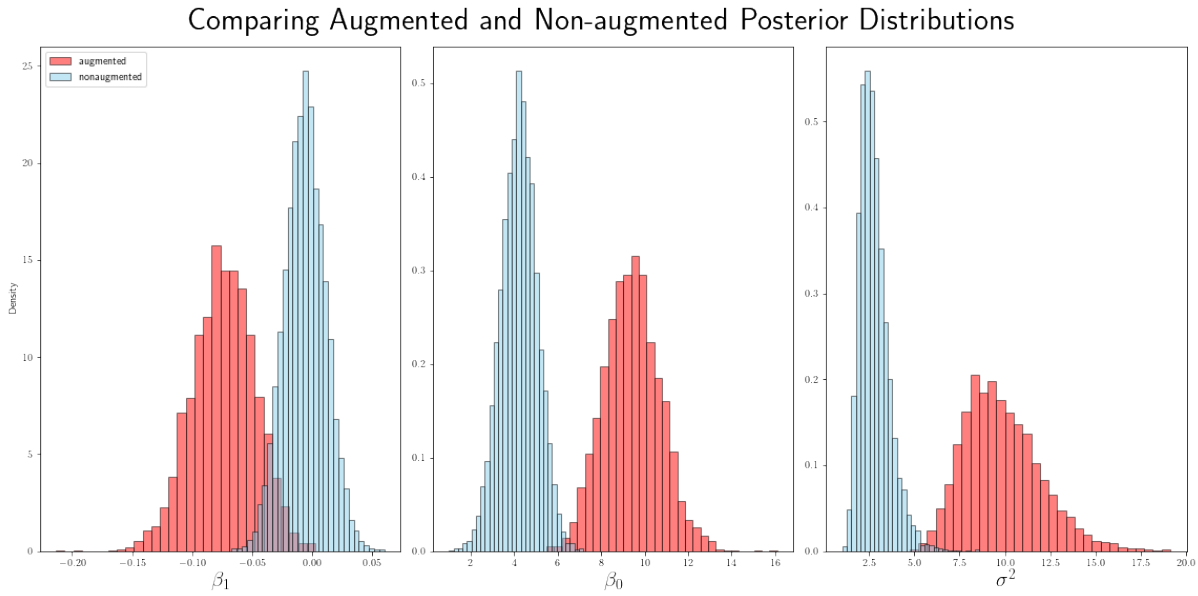

Figure 4: Comparing augmented and non-augmented marginals.