Missing data lecture 10: Flawed approaches to missing data and EM

Flawed approach to missing data

One generally flawed approach to inference in missing data problems is to treat the missing values as unknown parameters and to maximize the following function of parameters *and* missing values:

$$L_{\text{mispar}}(\theta, y_{(1)} \mid y_{(0)}) = f_Y(y_{(1)}, y_{(0)} \mid \theta)$$

The MLE using this distribution would require you to jointly maximize the likelihood with respect to θ *and* $y_{(1)}$. Let's take this approach with the censored exponential samples and see what we get. We had that the likelihood was

$$\prod_{i=1}^{r} \theta^{-1} e^{-y_i/\theta} \mathbb{1}\left(y_i < c\right) \prod_{i=r+1}^{n} \theta^{-1} e^{-y_i/\theta} \mathbb{1}\left(y_i \geq c\right) f(y \mid \theta)$$

Because $e^{-y_i/\theta}$ is monotonically decreasing in y_i for any θ, $\hat{y}_i$ is c for the missing observations.

Plugging this into the log-likelihood gives

$$-r \log \theta - (n-r) \log \theta - \frac{\sum_{i=1}^{r} y_i + (n-r)c}{\theta}$$

Taking derivatives and setting this equal to zero gives:

$$\hat{\theta}_{\text{mispar}} = \frac{\sum_{i=1}^{r} y_i + (n-r)c}{n} = \frac{r}{n}\hat{\theta}$$

This understates the true value θ, and one can show that this estimator isn't consistent for θ. The only way this estimator is consistent for θ is if $r/n \to 1$.

This example shows that the goal in missing data analysis isn't to predict missing values, it is to account for the uncertainty in missing values by integrating over the distribution of missing values.

Here is another example.

Let y_i be normally distributed with unknown mean μ and variance σ^2. Suppose there are r observed values and $n - r$ missing values. We assume that the data are MAR and the parameters μ and σ^2 are distinct from the parameters of the missingness distribution.

The MLE from the ignorable likelihood is just the MLE on the complete cases:

$$\hat{\mu} = \sum_{i=1}^{r} \frac{y_i}{r}, \hat{\sigma}^2 = \sum_{i=1}^{r} \frac{(y_i - \hat{\mu}^2)}{r}$$

If we write down the mistaken likelihood for the data we get:

$$\ell_{\text{mispar}}(\mu, \sigma^2, y_{r+1}, \dots, y_n \mid y_{(0)}) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{r} (y_i - \mu)^2 - \frac{1}{2\sigma^2} \sum_{i=r+1}^{n} (y_i - \mu)^2$$

We can maximize this by setting $y_{r+1}, \dots, y_n$ to μ, thereby eliminating the second sum. This leads to an MLE for μ that is equal to $\hat{\mu}$ above. The variance, however, is incorrectly estimated. Taking gradients and solving for σ^2 gives

$$\hat{\sigma}^2 = \sum_{i=1}^{r} \frac{(y_i - \hat{\mu})^2}{n}$$

Why is this wrong? Intuitively, we can see that the expected value of the second term isn't zero, and thus we'll understate the variance Thus, the MLE in the misparametrized model yields a variance estimate that is too low.

Again, we want to integrate over our uncertainty in the missing values, rather than predict missing values.

EM algorithm

The last section discussed how maximizing the likelihood as a function of the missing data points was an incorrect way to go about doing likelihood inference when there is missing data.

The "right" way to do so is to integrate over the uncertainty in your likelihood stemming from the missing observations, and then maximize this object.

$$L_{\text{ign}}(\theta \mid Y_{(0)} = \tilde{y}_{(0)}) \propto \int_{\mathcal{Y}_{(1)}} f_Y(Y_{(0)} = \tilde{y}_{(0)}, Y_{(0)} = y_{(1)} \mid \theta) dy_{(1)}$$

We'll call the log-likelihood above:

$$\ell(\theta \mid Y_{(0)} = \tilde{y}_{(0)})$$

Sometimes we can't do this maximization directly in one step because it is tricky. Instead, of doing the maximization directly we do so iteratively, and this is the motivation for the Expectation-Maximization algorithm: Let $\ell_Y(\theta \mid Y)$ be the complete data log-likelihood, and $f(Y_{(1)} \mid Y_{(0)}, \theta^{(t)})$ be the distribution of missing values given the observed values and the current best guess at the parameters $\theta^{(t)}$. The object, which we'll call $Q(\theta \mid \theta^{(t)})$, we want to maximize is the expected log-likelihood, where we take the expectation over the conditional distribution of the missing values

$$Q(\theta \mid \theta^{(t)}) = \int_{\mathcal{Y}_{(1)}} \ell_Y(\theta \mid Y_{(1)}, Y_{(0)} = \tilde{y}_{(0)}) f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta^{(t)}) dY_{(1)}$$

Normal example, again

Continuing the normal example from above, we have that our complete data log-likelihood is:

$$\ell_Y(\mu, \sigma^2 \mid Y_{(1)}, Y_{(0)}) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{r}(y_i - \mu)^2 - \frac{1}{2\sigma^2}\sum_{i=r+1}^{n}(y_i - \mu)^2$$

Because we've assumed that the data are MAR (which in univariate settings equals MCAR), we have that

$$y_i \sim \text{Normal}(\mu^{(t)}, (\sigma^2)^{(t)})$$

Then for each y_i,

$$\mathbb{E}\left[y_i^2 - 2\mu y_i + \mu^2\right] = (\mu^{(t)})^2 + (\sigma^t)^2 - 2\mu\mu^{(t)} + \mu^2$$

Then the expected log-likelihood is:

$$Q(\theta \mid \theta^{(t)}) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{r}(y_i - \mu)^2 - \frac{1}{2\sigma^2}(n-r)((\mu^{(t)})^2 + (\sigma^t)^2 - 2\mu\mu^{(t)} + \mu^2)$$

Taking partial derivatives with respect to μ gives:

$$\begin{aligned} 0 &= \frac{1}{\sigma^2}\sum_{i=1}^{r}(y_i - \mu) - \frac{1}{\sigma^2}(n-r)(-2\mu^{(t)} + 2\mu) \\ &= \frac{1}{\sigma^2}(\sum_{i=1}^{r} y_i + (n-r)\mu^{(t)}) - n\frac{\mu}{\sigma^2} \end{aligned}$$

Solving for μ gives the update rule for $\mu^{(t)}$:

$$\mu^{(t+1)} = \frac{\sum_{i=1}^{r} y_i + (n-r)\mu^{(t)}}{n}$$

Taking partials with respect to σ^2

$$\frac{\partial Q(\theta \mid \theta^{(t)})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\sum_{i=1}^{r}(y_i - \mu)^2 + (n-r)((\mu^{(t)})^2 + (\sigma^{(t)})^2 - 2\mu\mu^{(t)} + \mu^2))$$

Setting this equal to zero and simplifying gives:

$$0 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\sum_{i=1}^{r}(y_i - \mu^{(t+1)})^2 + (n-r)((\mu^{(t)})^2 + (\sigma^{(t)})^2 - 2\mu^{(t+1)}\mu^{(t)} + (\mu^{(t+1)})^2))$$

This yields:

$$(\sigma^2)^{(t+1)} = \frac{(\sum_{i=1}^{r}(y_i - \mu^{(t+1)})^2 + (n-r)((\mu^{(t)})^2 + (\sigma^{(t)})^2 - 2\mu^{(t+1)}\mu^{(t)} + (\mu^{(t+1)})^2))}{n}$$

which simplifies to

$$(\sigma^2)^{(t+1)} = \frac{(\sum_{i=1}^{r} y_i^2 + (n-r)((\mu^{(t)})^2 + (\sigma^{(t)})^2)}{n} - (\mu^{(t+1)})^2$$

This shows why the prior procedure didn't work; namely it ignores the extra variability in imputations for the missing y_i^2 terms.

To see why this simplifies, expand out

$$\sum_{i=1}^{r}(y_i - \mu^{(t+1)})^2 = \sum_{i=1}^{r} y_i^2 - 2\mu^{(t+1)}\sum_{i=1}^{r} y_i + r(\mu^{(t+1)})^2$$

and sub in the following expression for $\sum_i y_i$

$$\sum_{i=1}^{r} y_i = n\mu^{(t+1)} - (n-r)\mu^{(t)}$$

Plugging in our estimate for $\mu^{(t+1)}$ gives the solution above.

One can show that the EM solution converges to the complete data result.

Nontrivial example

Here's an example: Let's say we have two outcomes, so Y is an $n \times 2$ matrix. For simplicity's sake, we assume the missingness is ignorable, and assume we have a general missingness pattern, i.e. there are 3 patterns of missingness, assuming all participants had at least one measurement. The parameters of interest are μ_1, μ_2, Σ, and we'd like to use all the available data to do inference. If we had a complete dataset, we know from an earlier lecture the ML solutions for these quantities:

$$\hat{\mu}_1 = \bar{y}_1,\ \hat{\mu}_2 = \bar{y}_2,\ \hat{\Sigma} = 1/n\sum_i y_i y_i^T - \hat{\mu}\hat{\mu}^T$$

Let

$r_1 = \#(y_{1i}$ obs, , y_{2i} missing $), r_2 = \#(y_{2i}$ obs, , y_{1i} missing $), n - r_1 - r_2 = \#(y_{2i}$ obs, , y_{1i} obs $)$

, and suppose we have arranged our indices i so $i \in \{1, \dots, r_1\}$ have y_1 observed, $i \in \{r_1+1, \dots, r_1+r_2\}$ have y_2 observed and $i \in \{r_1+r_2+1, \dots, n\}$ have all data observed. Let $f_{Y_j}(y_{ji} \mid \mu_j, \Sigma_{j,j}), j = 1, 2$ be the univariate normal density, while $f_Y(y_i \mid \mu_1, \mu_2, \Sigma)$ is the bivariate normal density. The observed data likelihood is

$$L(\mu_1, \mu_2, \Sigma \mid Y_{(0)}) = \prod_{i=1}^{r_1} f_{Y_1}(y_{1i} \mid \mu_1, \Sigma_{1,1}) \prod_{i=r_1+1}^{r_1+r_2} f_{Y_2}(y_{2i} \mid \mu_2, \Sigma_{2,2}) \prod_{i=r_1+r_2+1}^{n} f_{Y_2}(y_i \mid \mu_1, \mu_2, \Sigma)$$

We can find the MLEs for this expression, but it isn't standard, and it'll involve some thinking, whereas if we had a complete dataset we could just do the maximization very easily. The two variable dataset seems pretty tractable, but imagine we had many variables with lots of missingness patterns, and then the maximization would be hard.

Why does EM work?

Why does this work? We want to show that maximizing Q is equivalent to maximizing L_{ign}.

We can write the complete data distribution as the product of two factors:

$$f_Y(Y_{(1)}, Y_{(0)} \mid \theta) = f(Y_{(0)} \mid \theta) f(Y_{(1)} \mid Y_{(0)}, \theta)$$

Taking logs gives us the complete data likelihood in terms of the observe data likelihood and the log likelihood of the conditional distribution of the missing data given the observed data and a parameter value.

$$\ell_Y(Y_{(1)}, Y_{(0)} \mid \theta) = \log(f(Y_{(0)} \mid \theta)) + \log(f(Y_{(1)} \mid Y_{(0)}, \theta))$$

The first term is the observed data likelihood, which is what we want to maximize. Rearranging gives:

$$\log(f(Y_{(0)} \mid \theta)) = \ell_Y(Y_{(1)}, Y_{(0)} \mid \theta) - \log(f(Y_{(1)} \mid Y_{(0)}, \theta))$$

Taking expectations for a current iterate θ^r over $f(Y_{(1)} \mid Y_{(0)}, \theta^t)$, gives

$$\mathbb{E}_{f(Y_{(1)} \mid Y_{(0)}, \theta^t)} \left[\log(f(Y_{(0)} \mid \theta)) \right] = Q(\theta \mid \theta^t) - H(\theta \mid \theta^t)$$

where Q is as above and $H(\theta \mid \theta^t)$ is:

$$H(\theta \mid \theta^{(t)}) = \int_{\mathcal{Y}_{(1)}} \log f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta) f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta^t) dY_{(1)}$$

We can show that $H(\theta \mid \theta^t) \leq H(\theta^t \mid \theta^t)$ for all θ.

$$
\begin{aligned}
H(\theta \mid \theta^t) - H(\theta^t \mid \theta^t) &= \int_{\mathcal{Y}_{(1)}} \log \frac{f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta)}{f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta^t)} f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta^t) dY_{(1)} \\
&= \mathbb{E}_{f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta^t)} \left[\log \frac{f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta)}{f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}} \right] \\
&\leq \log \mathbb{E}_{f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta^t)} \left[\frac{f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta)}{f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}} \right] \\
&= \log 1 = 0
\end{aligned}
$$

This is the same proof that shows that the KL divergence is always positive!

$$
\mathrm{KL}(f \mid g) = \int_{\mathcal{Y}} \log \frac{f(y)}{g(y)} f(y) dy
$$

Now we look at the value of the maximized oberved data likelihood:

$$
\mathbb{E}_{f(Y_{(1)} \mid Y_{(0)}, \theta^t)} \left[\log(f(Y_{(0)} \mid \theta^t)) \right]
$$

and how this changes between steps t and $t+1$ so that θ^{t+1} as $\theta^{t+1} = \mathrm{argmax}_\theta Q(\theta \mid \theta^t)$.

The difference between the observed likelihoods

$$
\begin{aligned}
\mathbb{E}_{f(Y_{(1)} \mid Y_{(0)}, \theta^{t+1})} \left[\log(f(Y_{(0)} \mid \theta^{t+1})) \right] - & \mathbb{E}_{f(Y_{(1)} \mid Y_{(0)}, \theta^t)} \left[\log(f(Y_{(0)} \mid \theta^t)) \right] \\
&= Q(\theta^{t+1} \mid \theta^t) - Q(\theta^t \mid \theta^t) - (H(\theta^{t+1} \mid \theta^t) - H(\theta^t \mid \theta^t))
\end{aligned}
$$

Thus, by maximizing the expected complete data likelihood, we in turn maximize the function we really want, namely the observed likelihood.

Another result is that if θ^{t+1} is chosen such that: 1. $\frac{\partial}{\partial \theta} Q(\theta \mid \theta^t) \mid_{\theta = \theta^{t+1}} = 0$

2. $\theta^{t+1} \to \theta^\star$
3. $f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta)$ is sufficiently smooth in θ

Then

$\frac{\partial}{\partial \theta} \ell(\theta \mid Y_{(0)} = \tilde{y}_{(0)}) \mid_{\theta = \theta^\star} = 0$

$$
\frac{\partial}{\partial \theta} \ell(\theta \mid Y_{(0)} = \tilde{y}_{(0)}) \mid_{\theta = \theta^\star} = \frac{\partial}{\partial \theta} Q(\theta \mid \theta^t) \mid_{\theta = \theta^\star} - \frac{\partial}{\partial \theta} H(\theta \mid \theta^t) \mid_{\theta = \theta^\star}
$$

The first quantity on the RHS is zero by the condition that we pick θ^{t+1} as that which leads to $\frac{\partial}{\partial\theta}Q(\theta \mid \theta^t) \mid_{\theta=\theta^\star}= 0$, so if $\theta^t \to \theta^\star$, we must have gotten there by setting these gradients equal to zero. The second quantity on the RHS is

$$\begin{aligned}\frac{\partial}{\partial\theta}H(\theta \mid \theta^t) \mid_{\theta=\theta^\star} &= \frac{\partial}{\partial\theta}\int_{\mathcal{Y}_{(1)}} \log f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta) f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta^t) dY_{(1)} \mid_{\theta=\theta^\star} \\ &= \int_{\mathcal{Y}_{(1)}} \frac{\frac{\partial}{\partial\theta} f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta) \mid_{\theta=\theta^\star}}{f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta^\star)} f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta^t) dY_{(1)} \mid_{\theta=\theta^\star}\end{aligned}$$

as $\theta^t \to \theta^\star$ the denominators cancel, leaving

$$\int_{\mathcal{Y}_{(1)}} \frac{\partial}{\partial\theta} f(Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta) \mid_{\theta=\theta^\star} dY_{(1)}$$

which, if we pull the derivative out of the integral again, equals zero because we're differentiating a constant.

Multivariate normal example again

In the multivariate normal example we know how to maximize the likelihood, but how do we do the conditional expectation?

$$\mathbb{E}\left[\ell_Y(\mu, \Sigma \mid Y) \mid Y_{(0)}, \mu^t, \Sigma^t\right] = \frac{1}{2}\log(\det\Sigma^{-1}) - \frac{1}{2}\mathrm{tr}\sum_i \mathbb{E}\left[\left((y_i - \mu)(y_i - \mu)^T \mid Y_{(0)}, \mu^t, \Sigma^t\right]\Sigma^{-1}\right)$$

If we expand out the cross product, we see we need

$$\mathbb{E}\left[(y_i - \mu)(y_i - \mu)^T \mid Y_{(0)}, \mu^t, \Sigma^t\right] = \mathbb{E}\left[y_i y_i^T \mid Y_{(0)}, \mu^t, \Sigma^t\right] - \mathbb{E}\left[\mu y_i^T - y_i^T\mu \mid Y_{(0)}, \mu^t, \Sigma^t\right] - \mu\mu^T$$

The first term is :

$$\mathbb{E}\left[y_i y_i^T \mid Y_{(0)}, \mu^t, \Sigma^t\right] = \mathrm{Cov}(y_i \mid Y_{(0)}, \mu^t, \Sigma^t) + \mathbb{E}\left[y_i \mid Y_{(0)}, \mu^t, \Sigma^t\right]\mathbb{E}\left[y_i \mid Y_{(0)}, \mu^t, \Sigma^t\right]^T$$

Plugging this back in above gives:

$$\mathrm{Cov}(y_i \mid Y_{(0)}, \mu^t, \Sigma^t) + \mathbb{E}\left[y_i \mid Y_{(0)}, \mu^t, \Sigma^t\right]\mathbb{E}\left[y_i \mid Y_{(0)}, \mu^t, \Sigma^t\right]^T - \mathbb{E}\left[\mu y_i^T - y_i^T\mu \mid Y_{(0)}, \mu^t, \Sigma^t\right] - \mu\mu^T$$

simplifying to

$$\mathrm{Cov}(y_i \mid Y_{(0)}, \mu^t, \Sigma^t) + (\mathbb{E}\left[y_i \mid Y_{(0)}, \mu^t, \Sigma^t\right] - \mu)(\mathbb{E}\left[y_i \mid Y_{(0)}, \mu^t, \Sigma^t\right] - \mu)^T$$

Pluggig this back in above, we get the expected log-likelihood

$$\mathbb{E}\left[\ell_Y(\mu, \Sigma \mid Y) \mid Y_{(0)}, \mu^t, \Sigma^t\right] = \frac{1}{2}\log(\det \Sigma^{-1})$$
$$-\frac{1}{2}\text{tr}\sum_i \left(\text{Cov}(y_i \mid Y_{(0)}, \mu^t, \Sigma^t) + (\mathbb{E}\left[y_i \mid Y_{(0)}, \mu^t, \Sigma^t\right] - \mu)(\mathbb{E}\left[y_i \mid Y_{(0)}, \mu^t, \Sigma^t\right] - \mu)^T\right)\Sigma^{-1}$$

This leads to the M step estimates:

$$\mu^{t+1} = \frac{1}{n}\sum_i \mathbb{E}\left[y_i \mid Y_{(0)}, \mu^t, \Sigma^t\right],$$

and for

$$\Sigma^{t+1} = \frac{1}{n}\sum_i \text{Cov}(y_i \mid y_{i(0)}, \mu^t, \Sigma^t) + (\mathbb{E}\left[y_i \mid y_{i(0)}, \mu^t, \Sigma^t\right] - \mu^{t+1})(\mathbb{E}\left[y_i \mid y_{i(0)}, \mu^t, \Sigma^t\right] - \mu^{t+1})^T$$

The key is that the conditional expectation and covariances for y_i are informed by the observed data.