Imputation under Missing at Random

How to Impute and How to Evaluate Imputations

Jeffrey Näf

March 14, 2024

Adding Parameters

- Assume p^* is parametrized by a vector θ and $\mathbb{P}(M = m \mid x)$ is parametrized by a vector ϕ.
- Rewrite the MAR definition slightly:

$$\mathbb{P}_\phi(M = m|x) = \mathbb{P}_\phi(M = m|o(x, m)) \text{ for all } m \in \mathcal{M} \text{ and } x. \tag{1}$$

  (just added ϕ)
- θ is the vector of interest, ϕ is a nuisance parameter

1. **Ignorability and Maximum Likelihood**

2. Uncertainty Adjustment

3. Conclusion

Rubin's Original Result

- Assume MAR holds + that "θ and ϕ are distinct"
- Should just mean that $\mathbb{P}_\phi(M = m|x)$ does not depend on θ. For every fixed ϕ we can choose θ freely and the other way around
- Now assume we want to find θ with Maximum Likelihood Estimation (MLE).
- Without missing values:

$$\hat{\theta} = \arg\max_{\theta} p^*_\theta(x).$$

Rubin's Original Result

- Consider the likelihood for a pattern $m \in \mathcal{M}$,

$$\begin{aligned}
\mathcal{L}(\theta; o(x,m)) &= \int p^*_{\theta,\phi}(x, M = m) do^c(x,m) \\
&= \int \mathbb{P}_\phi(M = m \mid x) p^*_\theta(x) do^c(x,m) \\
&= \mathbb{P}_\phi(M = m \mid o(x,m)) p^*_\theta(o(x,m)) \\
&= c(o(x,m)) p^*_\theta(o(x,m)).
\end{aligned}$$

$c(o(x,m))$ does not depend on θ. So for all $m \in \mathcal{M}$:

$$\hat{\theta} = \arg\max_\theta \mathcal{L}(\theta; o(x,m)) = \arg\max_\theta p^*_\theta(o(x,m)) \tag{2}$$

Rubin's Original Result

- This in particular means that for an i.i.d. sample:

$$
\begin{aligned}
\hat{\theta} &= \arg\max_{\theta} \prod_{i=1}^{n} \mathcal{L}(\theta; o(X_i, M_i)) \\
&= \arg\max_{\theta} \prod_{i=1}^{n} p^*_{\theta}(o(X_i, M_i))
\end{aligned}
$$

$\Longrightarrow$ The whole story with different patterns and different distribution from the first part doesn't really matter here(!)

Rubin's Original Result

- This in particular means that for an i.i.d. sample:

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \prod_{i=1}^{n} \mathcal{L}(\theta; o(X_i, M_i)) \\ &= \arg\max_{\theta} \prod_{i=1}^{n} p^*_\theta(o(X_i, M_i))\end{aligned}$$

$\Longrightarrow$ The whole story with different patterns and different distribution from the first part doesn't really matter here(!)
This is nice, but the optimization often gets quite complicated, and **EM algorithms** have to be employed.

1. Ignorability and Maximum Likelihood

2. Uncertainty Adjustment

3. Conclusion

Rubin's Rules

- If imputation is used, one has to be careful to not underestimate the variability of parameters
- Even if the values are now imputed, the fact that we didn't observe them introduces additional uncertainty

Rubin's Rules

- Assume we have generated *m* datasets

X_1	X_2	X_3	Y
3	20	10	s
-6	45	6	s
0	4	30	no s
-4	32	35	s
1	63	40	s
-2	15	12	no s

X_1	X_2	X_3	Y
-7	20	10	s
-6	45	9	s
0	12	30	no s
13	32	35	s
1	63	40	s
-2	10	12	no s

X_1	X_2	X_3	Y
7	20	10	s
-6	45	12	s
0	-5	30	no s
2	32	35	s
1	63	40	s
-2	20	12	no s

- Calculate $\hat{\theta}_j$ and $\hat{var}(\hat{\theta}_j)$, for $j = 1, \ldots, m$.
- Combine using Rubin's Rules:

$$\hat{\theta} = \frac{1}{m}\sum_j \hat{\theta}_j$$

$$T = \underbrace{\frac{1}{m}\sum_j \hat{var}(\hat{\theta}_j)}_{\text{within variance}} + \underbrace{\left(1 + \frac{1}{m}\right)\frac{1}{m-1}\sum_j (\hat{\theta}_j - \hat{\theta})^2}_{\text{between variance}}$$

Something we ignored

- Despite our careful analysis there is something we forgot
- For the above rules to be truly valid, an imputation should also include *model uncertainty*
- For instance in the Gaussian Imputation, the regression estimator $\hat{\beta}$ has some error in finite samples that should be accounted for.
- This might be a reason to use mice-rf over mice-cart, as mice-rf attempts to account for this by using different trees
- However this is somewhat negligible for larger sample sizes.

1. Ignorability and Maximum Likelihood

2. Uncertainty Adjustment

3. Conclusion

What Method to use?

- In terms of distributional distance, **mice-cart/RF/DRF** seem almost unbeatable for real data with a number of observations $\mathbf{n} > \mathbf{200}$
- For $\mathbf{n} \leq \mathbf{200}$, "**mice-pmm**" might even work a bit better

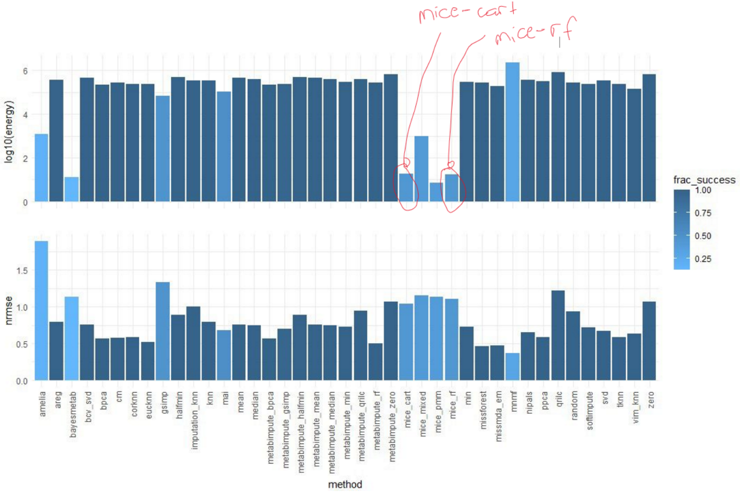

Figure: Preliminary Figure, showing the performance of a range of imputation methods for metabolomics data. The datasets and missing methods are described here

So which Imputation Method?

- In terms of distributional distance, **mice-cart/RF/DRF** seem almost unbeatable for real data with a number of observations $\mathbf{n} > \mathbf{200}$
- For $\mathbf{n} \leq \mathbf{200}$, "**mice-pmm**" might even work a bit better
- It could however be that these observations are a result of researchers not using realistic MAR assumptions (!)
- For instance, researchers often use the `ampute` function of the mice package.
- This function does not appear to induce heavy distributional shifts.

Some Helpful Links

- Rmisstastic
- Imputomics
- mice package vignette
- CRAN Task view on missing data

Bibliography