Multiple Imputation for Continuous and Ordinal Data: Comparing Joint Multivariate Normal and Conditional Approaches

Thu Nguyen

Stat 536E Final Project

April 20, 2023

Abstract

This report studies the performance of two popular approaches to multiple imputation (MI), namely joint multivariate normal (MVN) MI, in which the data is modeled as a jointly MVN distributed, and conditional MI, in which each variable is modeled conditionally on all of the other variables. As a consequence, implementing joint MVN MI requires an extra step of transforming discrete variables from continuous variables, which is often done via probabilistic models. In order to compare the relative performance, we simulate data and score each approach with two types of metrics: (1) the accuracy of the imputed values and (2) the accuracy of the coefficients and fitted values based on a model fitted to their completed datasets.

Building on top of the work of Kropko, Goodrich, Gelman, and Hill in [Kro+17], we extend the simulations on two dimensions: (1) the missingness mechanism to include missing not at random (MNAR) in additiona to missing at random (MAR) and (2) the miss rates in both the variable of interest and the explanatory variables. In consideration as continuous and ordinal (ordered-discrete) variables. We find that conditional MI tends to outperform in all metrics when the variable is continuous, whereas there does not appear a clearly dominating approach when the variable is ordinal.

The code, the simulated data, and the analysis for this report are available at https://github.com/ngthu003/stat536E_finalProject.

Contents

1 Introduction

1.1 Missing Data and Multiple Imputation (MI)

Missing data are common in multivariate or longitudinal datasets, as it is unlikely to have all data available for each individual on every variable (for both explanatory variables and target variables, the prediction of which is desired). Ignoring missing data can lead to biased results or complications in statistical analysis. Therefore, it is important to find appropriate methods to handle missing data in real data analysis.

First introduced in the 1980s by Rubin in [Rub87], *multiple imputation* (hereafter to be referred to as simply MI) has emerged as a principled tool to handle missing data thanks to extensive studies. MI is a two-phase procedure:

1. The imputer creates completed datasets by filling in missing values using an imputation model, which are now passed over to the analyst.

2. The analyst analyzes the completed datasets individually using standard statistical tools, the results of which are then combined to form an overall conclusion. The details of how the individual results are pooled together can be found in most textbooks on handling missing data with multiple imputation, for example [Buu18].

MI was originally developed for statistical agencies to provide imputed databases to distinct end-users, allowing for valid inference and a consistent set of imputations for all analyses. This approach also shifts the burden of handling missing data from the analysts to the imputers.

1.2 Reference Paper

Kropko, Goodrich, Gelman, and Hill study the problem of multiple imputation (MI) using different techniques across continuous and discrete variables in [Kro+17]. The techniques considered are *joint multivariate normal (MVN)* and *conditional* MI, tow of the most common MI approaches.

At a high level, joint MVN MI makes the assumption that the variables in the data are jointly distributed in a multivariate normal distribution and proceeds to estimate the common parameters mean and covariance. Meanwhile conditional MI is somewhat less restrictive in only modeling the variables individually, conditional on all of the other variables. The two approaches share a lot in common but also differ in some key ideas and their implications. More details on the two approaches can be found in section 2.

[Kro+17] compares the performance of joint MVN and conditional MI on simulated data and a real dataset. The authors observe two kinds of results: either (1) conditional MI outperforms joint MVN MI significantly or (2) the two approaches perform equally well. They note that there are no scenarios in which joint MVN MI outperforms conditional MI.

In light of the findings, the authors hypothesize two reasons for these results. Both of these results concern the assumptions explicitly made by joint MVN MI. The first reason is the joint multivariate normal (MVN) distribution assumption. If the underlying data distribution is indeed MVN then the joint MVN MI can be expected to perform well. However, if it is not it is likely that the performance may be considerably impacted. The second reason is how joint MVN MI handles discrete variables. Since the MVN distribution can only accept continuous variables, any and all discrete variables have to be transformed to continuous variables. This transformation may create additional complications for the joint MVN MI.

1.3 This Report

The report aims to study the two MI approaches above, joint MVN MI and conditional MI. We will review the designs of the two approaches and compare and contrast the two under different criteria such as the assumptions, the algorithms, and the implications. Building on top of [Kro+17] we will do an experiment on simulated data to empirically compare the performance of the two approaches and expand the scenarios along different dimensions, such as the missing mechanism and data miss rates.

The report is structured as follows. Section 2 explains what the approaches are and how they compare to each other. Section 3 provides the plan to how simulations are to be carried out and what metrics will be used to compare the performance of the two approaches. Section 4 summarizes the main results and the findings. Section 5 concludes with some final remarks.

2 Joint Multivariate Normal MI and Conditional MI

2.1 Joint Multivariate Normal MI

2.1.1 The Idea

Joint modeling is a common technique in statistical modeling. The main idea is that:

1. the observed data can be modeled with a multivariate distribution;
2. once specified, the parameters of such distribution are estimated;
3. imputed values are sampled from the distribution.

In theory, this distribution can be anything, though the most common choice is the multivariate normal distribution. Such choice gives the joint multivariate normal (MVN) approach.

This approach works especially well when everything aligns, i.e. the observed data indeed follow a multivariate normal (MVN) distribution. If not, the next best thing is if the distribution resembles MVN, in which case a variant of the EM (Expectation-Maximization) algorithm can be implemented to alternate between estimating the parameters (means, variances, and covariances) of the MVN distribution and sampling new imputed values [DLR77].

2.1.2 Algorithm

The key step in the joint MVN approach is estimating the parameters (μ, Σ) in the multivariate normal distribution. This is generally very difficult when the missing data is non-monotonic, which is usually the case in real-world data. [TW87] introduces the *data augmentation* algorithm to iteratively impute missing values and estimate the parameters. Algorithm 1 gives an overview of the joint MVN approach.

Algorithm 1 Joint MVN MI [Buu18]

Require: Observed data X, missing data Y
Require: Imputation model $\mathcal{N}(\mu, \Sigma)$
  Initialize $\theta^{(0)} = \left(\mu^{(0)}, \Sigma^{(0)}\right)$
  for $t = 1, \ldots, T$ **do**
    Impute $\hat{Y}^{(t)}$ and estimate $\theta^{(t)}$ with the data augmentation algorithm [TW87]
  end for
  Sample $Y \sim \mathcal{N}\left(\mu^{(T)}, \Sigma^{(T)}\right)$

2.1.3 Handling Discrete Variables

As a consequence of the MVN assumption, any and all discrete variables (binary, ordinal (ordered-), and nominal (unordered-categorical)) must be converted to continuous variables prior to the imputation process and back thereafter.

The simplest approach to handling these variable transformations is with rounding. Take for example the binary variables. They are usually initially represented as 0 or 1, in which case the imputed values are rounded to 0 if they are less than .5 and to 1 otherwise:

$$\hat{Y} = \begin{cases} 0, & \text{if } \hat{Y} \in (-\infty, .5) \\ 1, & \text{otherwise,} \end{cases}$$

Ordinal variables can be handled similarly, given their inherent ranking. While rounding can also be applied to nominal variables, it is less straightforward as the imputed values might not be indicative of the actual values.

The second approach is to draw samples from the variable-appropriate distributions parameterized by the continuous imputed values. They are *Bernoulli* for binary, *binomial* for ordinal, and *multinomial* for nominal variables [JKB11]. The `Amelia` package in `R` follows this approach to implement imputation.

The third approach is an extension of the first (rounding). The threshold is no longer fixed (e.g. .5 in the case of binary variables) but is treated as a hyper-parameter to be estimated [BBS07]. We note that this idea is similar to one of the many techniques in building classification models when the training data are highly imbalanced.

2.2 Conditional MI

2.2.1 The Idea

The idea behind conditional modeling is that the vast majority of statistical models are conditional, and thus it is advantageous to find the appropriate models given the specific features of each variable type. Once such models are specified, conditional MI attempts to estimate their parameters to then sample from them.

As such, the conditional MI approach avoids having to explicitly specify a joint distribution (which is replaced with a collection of univariate conditional distributions) and allows for a wider class of joint distributions, beyond just multivariate normal in the case of joint MVN MI above.

2.2.2 Algorithm

The conditional MI algorithm comprises of two nested loops. The inside loop is what characterizes the algorithm. It is a sequence of sampling through the collection of univariate conditional distributions. The outside loop is to sample to convergence. Algorithm 2 gives the detail of the conditional approach.

Algorithm 2 Conditional MI [Buu18]

Require: Observed data X, missing data Y
Require: Imputation model $\pi\left(Y_j|X, Y_{-j}, \theta_j\right)$ for each $j = 1, \ldots, p$
  Initialize $\hat{Y}_j^{(0)}$ as sampled from X_j for each $j = 1, \ldots, p$
  for $t = 1, \ldots, T$ **do**
    for $j = 1, \ldots, p$ **do**
      Define $\hat{Y}_{-j}^{(t)} = \begin{bmatrix} \hat{Y}_1^{(t)} \\ \hat{Y}_{j-1}^{(t)} \\ \hat{Y}_{j+1}^{(t)} \\ \hat{Y}_p^{(t)} \end{bmatrix}$ as the currently completed data without Y_j
      Sample the parameter $\hat{\theta}_j^{(t)} \sim \pi\left(\theta_j^{(t)}|X, \hat{Y}_{-j}^{(t)}\right)$
      Sample the imputed value $\hat{Y}^{(t)} \sim \pi\left(Y^{(t)}|X, \hat{Y}_{-j}^{(t)}, \hat{\theta}_j^{(t)}\right)$
    end for
  end for

2.2.3 Comparison to Gibbs Sampling

We note that while the conditional MI algorithm might resemble that of a Gibbs sampler at first glance, it is indeed not. The implication is that while the Gibbs sampler guarantees convergence, conditional MI does not always converge or might converge to a different distribution pending the imputation order of the variables [LYR12].

2.3 Relations between Joint MVN and Conditional MI

We note a special case where the two approaches are equivalent.

If (X, Y) indeed follows a multivariate normal distribution, then all imputation conditional densities $p(Y_j|X, Y_{-j})$ can be represented with linear regression models with a common fixed normally distributed error term. The reverse is also true. If $p(Y_j|X, Y_{-j})$ are all linear with a common fixed normal error term then the joint distribution is a multivariate normal. In this case, imputation with joint MVN MI is equivalent to that wth conditional MI.

2.4 Comparisons between Joint MVN and Conditional MI

The key advantages of the joint MVN MI approach are it's better understood theoretical results, that is if the imputation models are indeed multivariate normal. On the other hand, conditional MI allows for a wider class of joint distribution, besides joint multivaiate normal, and thus may be more suited if there are strong beliefs on the distributions of certain variables (e.g. categorical variables).

We note other similarities between the two approaches:

1. both make assumption and hence sample from their respective joint distributions (in the case of joint MVN, the distribution is explicitly MVN);
2. both require multiple imputed datasets to correctly model the imputation uncertainty (however empirically this does not need to be large, the default is 5)
3. neither are likely to be accurate estimators of the imputed values when the data are strongly NMAR (not missing at random)

3 Simulations

3.1 Objectives

Our reference paper [Kro+17] studies and compares the performance of joint multivariate normal and conditional MI approaches with simulated data over the different variable types (continuous, binary, ordinal, and nominal). In all four cases, Y denotes such variable of interest (one per case). The miss rate for Y is set at 25%. All of the other variables are continuous and collectively denoted by X. Some of these X are also missing at a (common) rate of 10%. The data are generated under the most favorable configurations: MAR (missing at random) and the initial distribution is indeed MVN.

Building on top of the findings in [Kro+17], this project aims to investigate the performance of joint multivariate normal and conditional MI approaches through two additional dimensions:

Dimension	in [Kro+17]	in this project
miss rate	on explanatory variables X at 10%	on X at $\{10\%, 20\%, 30\%\}$
	on variable of interest Y at 25%	on Y at $\{25\%, 50\%\}$
missing mechanism	strictly MAR	MAR and NMAR

Table 1: Dimensions to be expanded on in this project.

3.2 MI Algorithms

There are numerous packages to implement the different multiple imputation methods (besides joint multivariate normal and conditional MI). For our simulations, we will use the packages listed in table 2. We make a note on implementing the joint multivariate normal (MVN) MI method. We will do so with both the `Amelia` and `norm` packages. These two implement the same imputation method but differ in how they calculate the pooled variance. We will thus compare conditional MI as implemented in `mi` against both of them, in order to ensure that the comparison is indeed between the two methods and is not influenced by packages.

Approach	Package
complete case analysis	
joint MVN MI	`Amelia` and `norm`
conditional MI	`mi`

Table 2: The different MI approaches and their respective `R` packages to be experimented with in the simulations. Complete case analysis is added for additional perspectives.

3.3 Evaluation Metrics

A usual multiple imputation procedure involves two steps

1. impute the missing dataset into multiple complete datasets;
2. analyze the complete datasets independently with the standard techniques and then pool them together to create one final result.

In this project, the first step is done using the MI algorithms introduced in 3.2. The second step is to model the variable of interest Y in terms of the other variables X with a (Bayesian) generalized linear model

$$g(\mathbb{E}[Y|X]) = \beta_1 X_1 + \cdots + \beta_d + X_d$$

where $g(\cdot)$ is a link function.

Table 3 lists the notations for the quantities to be used in calculating the 4 evaluation metrics in table 4.

Notation	Meaning
N_M	number of Y observations with missing data
N	total number of observation
d	number of variables X
$\hat{Y}_i$	imputed value for Y_i if Y_i is missing
Y_i^*	fitted value of Y_i from the glm trained on the true data
$\hat{Y}_i^*$	fitted value of Y_i from the glm trained on the imputed data

Table 3: Notations for the quantities used in calculating the evaluation metrics in table 4.

We recall from section 2.1 that all discrete variables have to be transformed to continuous if the underlying joint distributions or conditional distributions are multivariate normal. All of the packages in section 3.2 will handle these transformations directly and return the corresponding choice probabilities. Choice probabilities may be used as an alternative to numeric imputed values in assessing the accuracy of the imputation.

Accuracy type	Variable type	Metric	Formula
imputed values	continuous	RMSE	$\sqrt{\frac{1}{N_M}\sum_i (Y_i - \hat{Y}_i)^2}$
		Bias	$\left\lvert \frac{1}{N_M}\sum_i (Y_i - \hat{Y}_i) \right\rvert$
	ordinal	RMSE	$\sqrt{\frac{1}{N_M}\sum_i \mathbb{I}(Y_i = \hat{Y}_i)^2}$
		Bias	$\left\lvert \frac{1}{N_M}\sum_i \mathbb{I}(Y_i = \hat{Y}_i) \right\rvert$
choice probabilities	ordinal	RMSE	$\sqrt{\frac{1}{N_M}\sum_i (p_{Y_i} = p_{\hat{Y}_i})^2}$
		Bias	$\left\lvert \frac{1}{N_M}\sum_i (p_{Y_i} - p_{\hat{Y}_i}) \right\rvert$
glm coefficients	continuous & ordinal	RMSE	$\sqrt{\frac{1}{d}\sum_j (\beta_j - \hat{\beta}_j)^2}$
		Bias	$\left\lvert \frac{1}{d}\sum_j (\beta_j - \hat{\beta}_j) \right\rvert$
fitted values	continuous & ordinal	RMSE	$\sqrt{\frac{1}{N}\sum_i (Y_i^* - \hat{Y}_i^*)^2}$
		Bias	$\left\lvert \frac{1}{N}\sum_i (Y_i^* - \hat{Y}_i^*) \right\rvert$

Table 4: Metrics to be used in evaluating and comparing the joint MVN and conditional MI approaches.

4 Results

We adopt the setting in [Kro+17] as the reference setting. Table 5 summarizes the setting.
For every desired dimension under study (introduced in section 3.1 we will run experiments with all but that dimension with the setting above and vary the values for that dimension.

We will present here the results from implementing the imputation method of PPD (predictive posterior distribution). the appendix 5 summarizes similar results but implemented under the PMM (predictive mean matching) method. PMM is an alternative to probabilistically drawing samples from the PPD.

Parameter	Value
sample size	$1,000$
number of fully observed explanatory variables X	5
number of partially observed explanatory variables X	3
miss rate of explanatory variables X	10%
miss rate of target variable Y	25%
number of categories in target variable Y (if applicable)	5
number of imputations	5
missing mechanism	missing at random (MAR)

Table 5: Summary of the default reference setting, adopted from [Kro+17].

Instead, PMM searches from a subset of *most similar* fully observed observations and draws samples from this. More details on PMM are available in [CG13].

4.1 MAR vs. MNAR

We consider the different missing mechanism a dataset can take. In real-world data, it is usually the best to hope for missing at random, which is also the mechanism that multiple imputation assumes. Realistically though, it is often the case of missing not at random:

- missing at random (MAR) is when the missingness does not depend on the missing values themselves (but may depend on other observed values)
- missing not at random (MNAR) is when the missingness depends on the missing values (and probably other observed values).

Figure 1 shows the comparison in the RMSE of the different accuracy types (listed in table 4) when the target variable Y is continuous. We note that the different missing mechanism may have an impact on the choice of MI approach. Under MAR, there appears no clearly better approach, whereas under MNAR, conditional MI seems a better fit, evidenced by the lower RMSE across all accuracy types.

Figure 2 shows the same results but when Y is ordinal. The trend is now in the opposite direction. Conditional MI works (significantly) better when the data is MAR, but no longer shows that advantage under MNAR.

4.2 Miss Rate in Y

We revert back to the default setting (with MAR as the missing mechanism) and experiment over different missing rate for the target variable Y.

Figure 3 shows the RMSE results when the miss rate is either 25% or 50% and that Y is continuous. Here joint MVN MI looks to be a better choice in both situations.

Figure 4 is when Y is instead ordinal. We again note the opposing trend. Conditional MI works (significantly) better in both of those scenarios.

4.3 Miss Rate in X

Lastly. we look at what impacts, if any, the rate of missing in the explanatory variables X can have.

Figure 5 shows the RMSE results when the miss rate for X is $10\%, 20\%$ or 30% and that Y is continuous. In this case, there does not appear a preferred approach since while one might work better given a certain miss rate, that same approach does not work as well given a different miss rate (e.g. conditional MI does worse under 10% but better under 20% miss rate).

Figure 6 is when Y is ordinal. Now we have a clearly dominating approach - conditional MI, whose performance ranges from a bit better to significantly better than joint MVN MI.

5 Conclusion

In general, section 4 show two kinds of results. The first result is that conditional MI significantly outperforms joint multivariate normal (MVN) MI, in particular if the variable of interest is ordinal. The second result is when all the assumptions of the joint MVN MI are in place (the joint distribution of the data is indeed MVN and there are no discrete variables) then joint MVN MI tends to perform better; however even in these cases the differences are not as significant as in the first result.

At a high level, these findings agree with that is reported in the reference paper [Kro+17], and give further evidence to the hypothesis that for joint MVN MI to perform best it's assumptions need to hold, either exactly or approximately highly.

However, we note that the findings here are only empirical. We have not proved, or are aware of mathematical proofs, that conditional MI will always outperform joint MVN MI, even if only in certain settings. Thus, it is encouraged that the investigators should understand their data at hand and their characteristics in order to choose an appropriate MI approach.

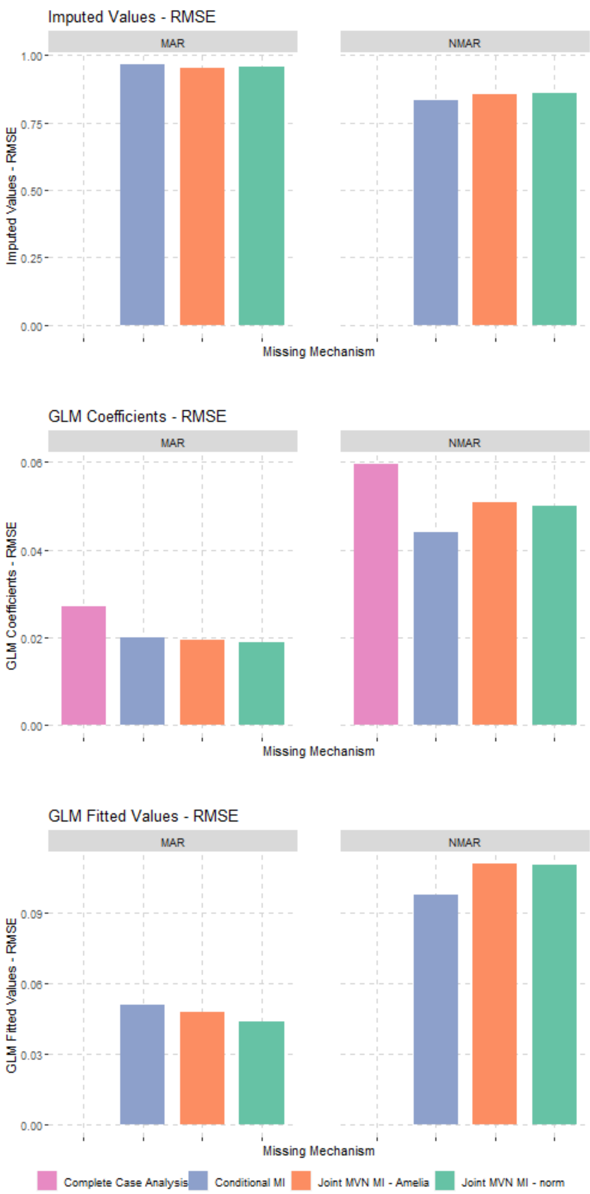

Figure 1: PPD imputation on Y being continuous. When the missingness is MAR, no approach complete dominates the other. However, conditional MI seems a better fit under MNAR.

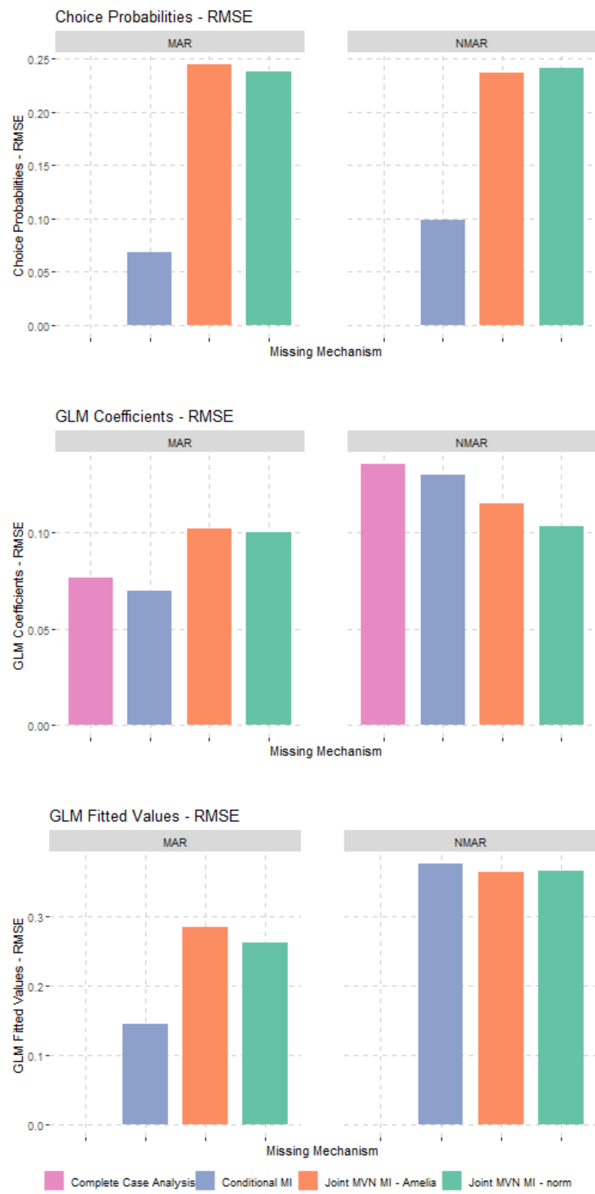

Figure 2: PPD imputation on Y being ordinal. The trend is opposite to that of continuous (left): conditional MI works much better under MAR but loses that advantage under MNAR.

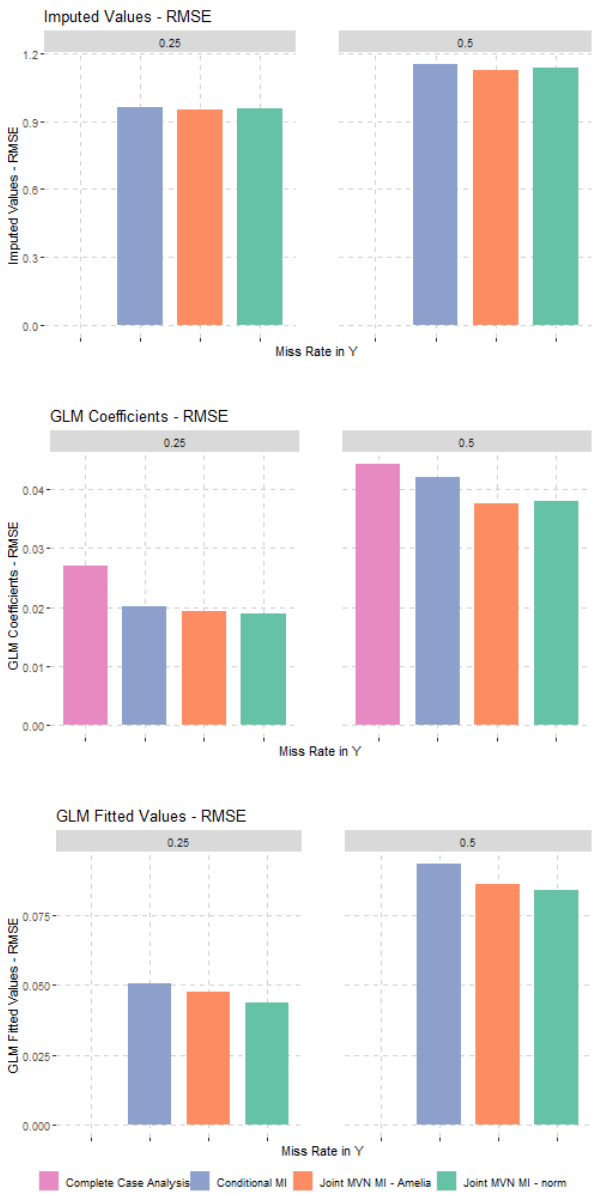

Figure 3: PPD imputation on Y being continuous. Joint MI seems a better fit under both scenarios of the miss rate on Y being 25% and 50%.

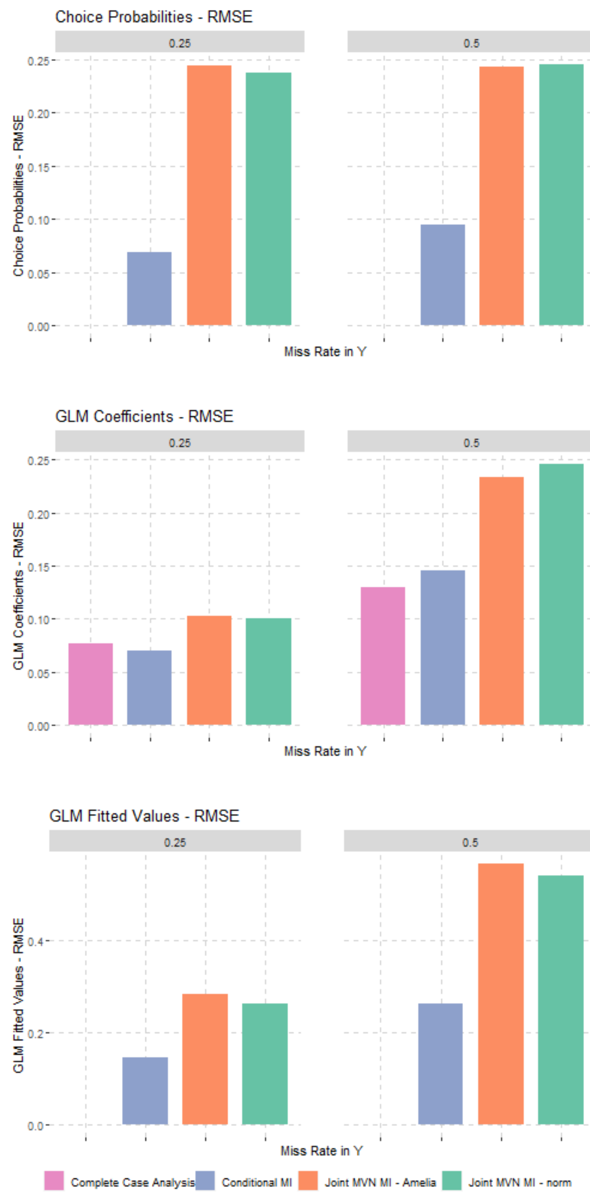

Figure 4: PPD imputation on Y being ordinal. The trend is opposite to that of continuous (left): conditional MI seems to fit the data much better.

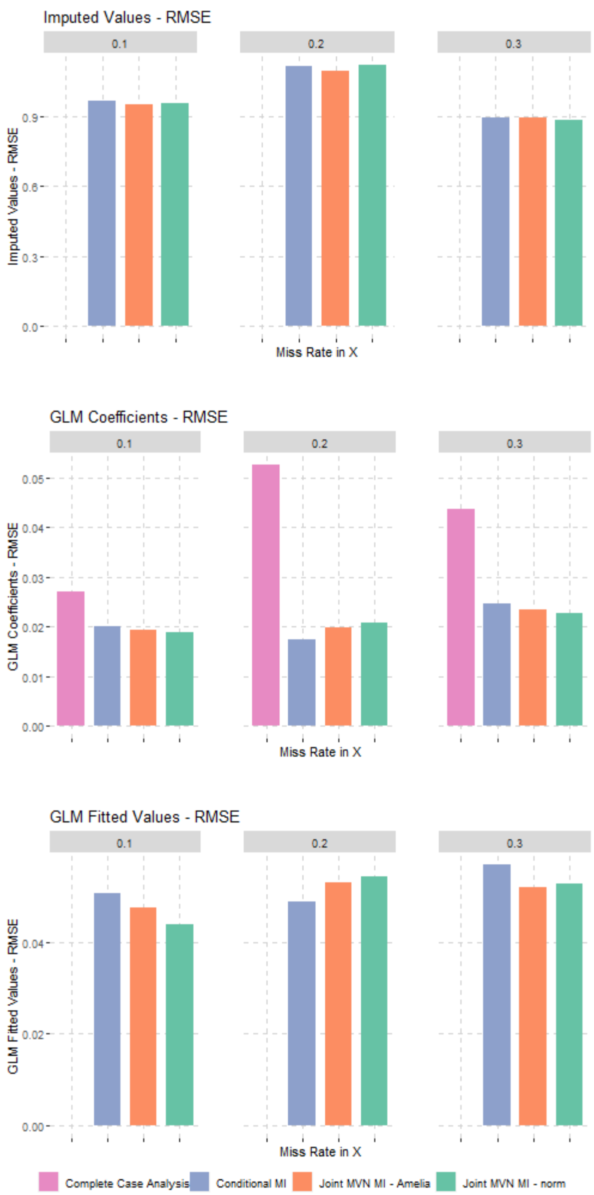

Figure 5: PPD imputation on Y being continuous. The results are more mixed as the miss rate on X varies.

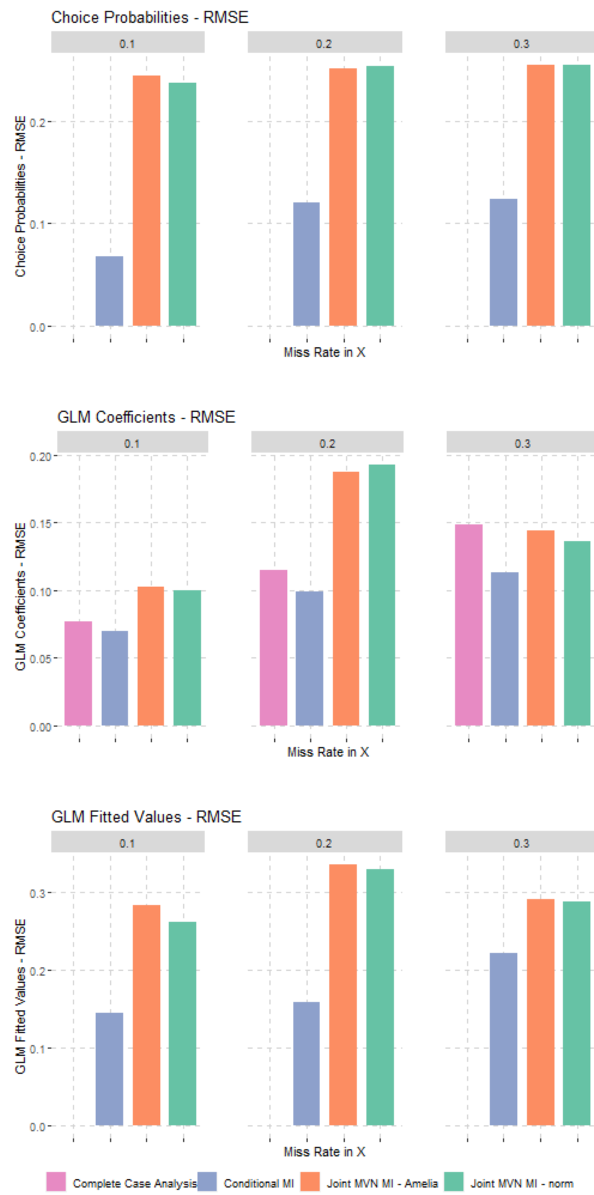

Figure 6: PPD imputation on Y being ordinal. Here there is a real winner - conditional MI, which dominates in all scenarios.

References

[Kro+17] J. Kropko et al. "Multiple Imputation for Continuous and Categorical Data: Comparing Joint Multivariate Normal and Conditional Approaches". In: *Political Analysis* 22.4 (2017), pp. 497–519. DOI: 10.1093/pan/mpu007.

[Rub87] D. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Wiley, 1987. ISBN: 9780471087052. URL: https://books.google.at/books?id=0KruAAAAMAAJ.

[Buu18] S. van Buuren. *Flexible Imputation of Missing Data, Second Edition*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press, 2018. ISBN: 9780429960345. URL: https://books.google.ca/books?id=RjZlDwAAQBAJ.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society: series B (methodological)* 39.1 (1977), pp. 1–22.

[TW87] M. A. Tanner and W. H. Wong. "The calculation of posterior distributions by data augmentation". In: *Journal of the American statistical Association* 82.398 (1987), pp. 528–540.

[JKB11] H. James, G. King, and M. Blackwell. "Amelia II: A Program for Missing Data Version 1.6. 2". In: *Available at [MATHEMATICAL LEFT ANGLE BRACKET] http://gking. harvard. edu/software/amelia-ii-program-missing-data/162¿. Accessed* 1 (2011).

[BBS07] C. A. Bernaards, T. R. Belin, and J. L. Schafer. "Robustness of a multivariate normal approximation for imputation of incomplete binary data". In: *Statistics in medicine* 26.6 (2007), pp. 1368–1382.

[LYR12] F. Li, Y. Yu, and D. B. Rubin. "Imputing missing data by fully conditional models: Some cautionary examples and guidelines". In: *Duke University Department of Statistical Science Discussion Paper* 11.14 (2012), pp. 1–35.

[CG13] S. J. Cranmer and J. Gill. "We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data". In: *British Journal of Political Science* 43.2 (2013), pp. 425–449. DOI: 10.1017/S0007123412000312.

Appendix

Section 4 summarizes the performance of joint multivariate normal MI versus conditional MI given the PPD (predictive posterior distribution) imputation method. An alternative to PPD is PMM (predictive mean matching). We present the results from PMM in this appendix.

1. *MAR vs. MNAR:* figures 7 and 8
    - When the target variable Y is continuous, no approach seems to work better than the other regardless of the missing mechanism. However, when Y is ordinal, conditional MI dominates joint MVN MI in both MAR and MNAR.
2. *Miss rate in Y:* figures 9 and 10
    - When Y is continuous, it is the same observation: no approach stands out. When Y is ordinal though, again conditional MI clearly dominates joint MVN MI for both miss rates in Y (10% and 25%).
3. *Miss rate in X:* figures 11 and 12
    - The trend is similar to that in miss rate in Y above.

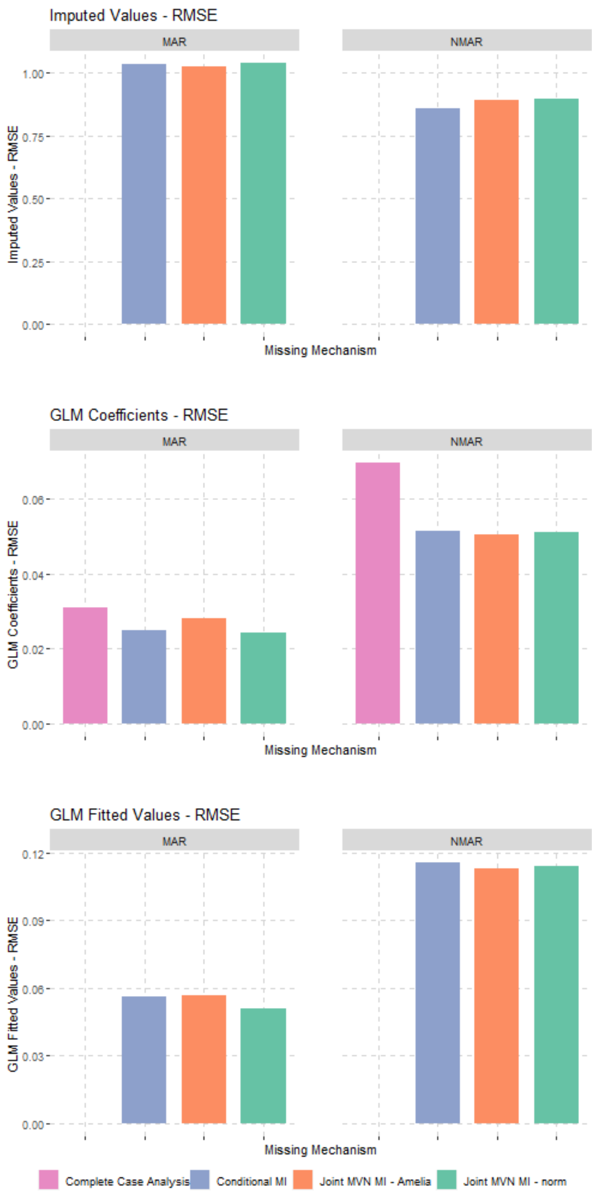

Figure 7: PMM imputation on Y being continuous. Neither approach seems to work better than the other in MAR or MNAR.

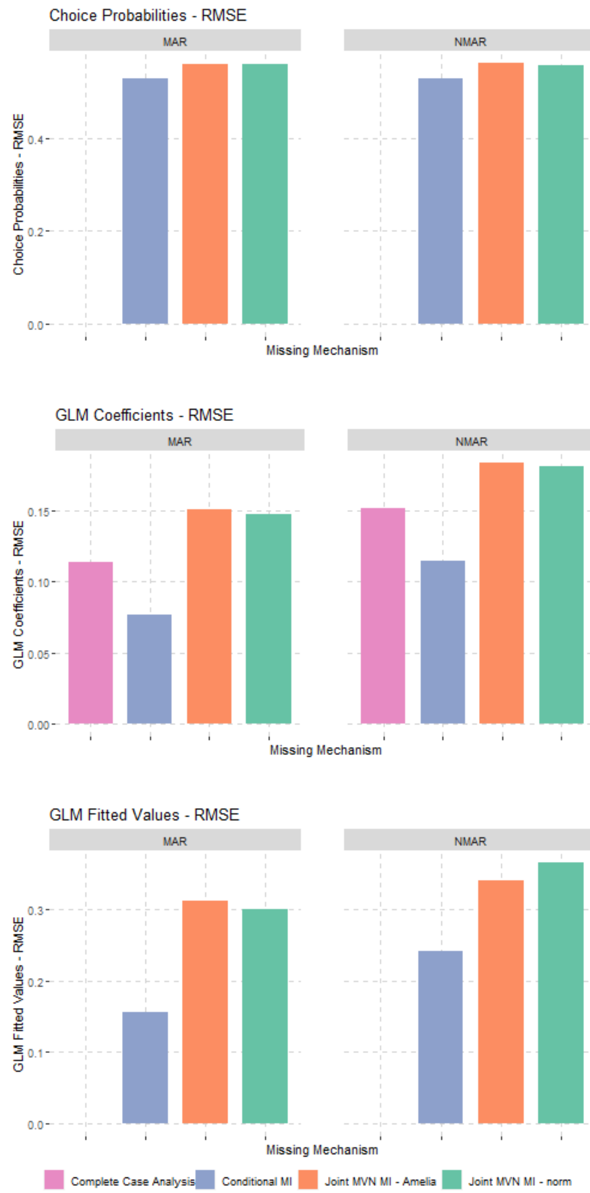

Figure 8: PMM imputation on Y being ordinal. Conditional MI dominates joint MVN MI in both MAR and MNAR.

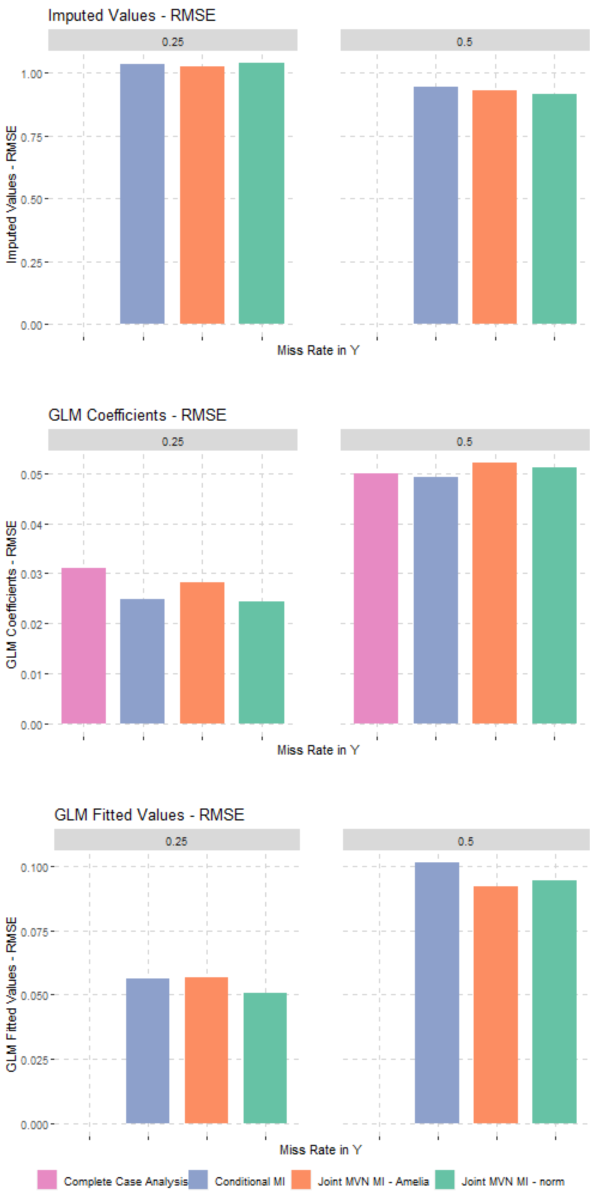

Figure 9: PMM imputation on Y being continuous. Neither approach seems to work better than the other.

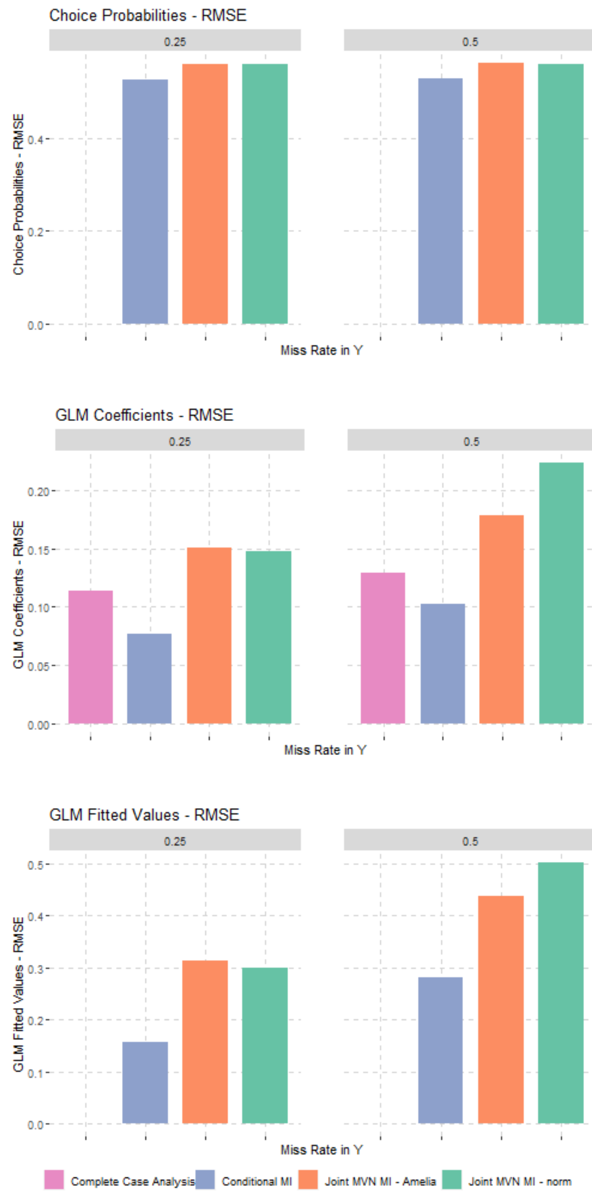

Figure 10: PMM imputation on Y being ordinal. Conditional MI dominates joint MVN MI in both scenarios.

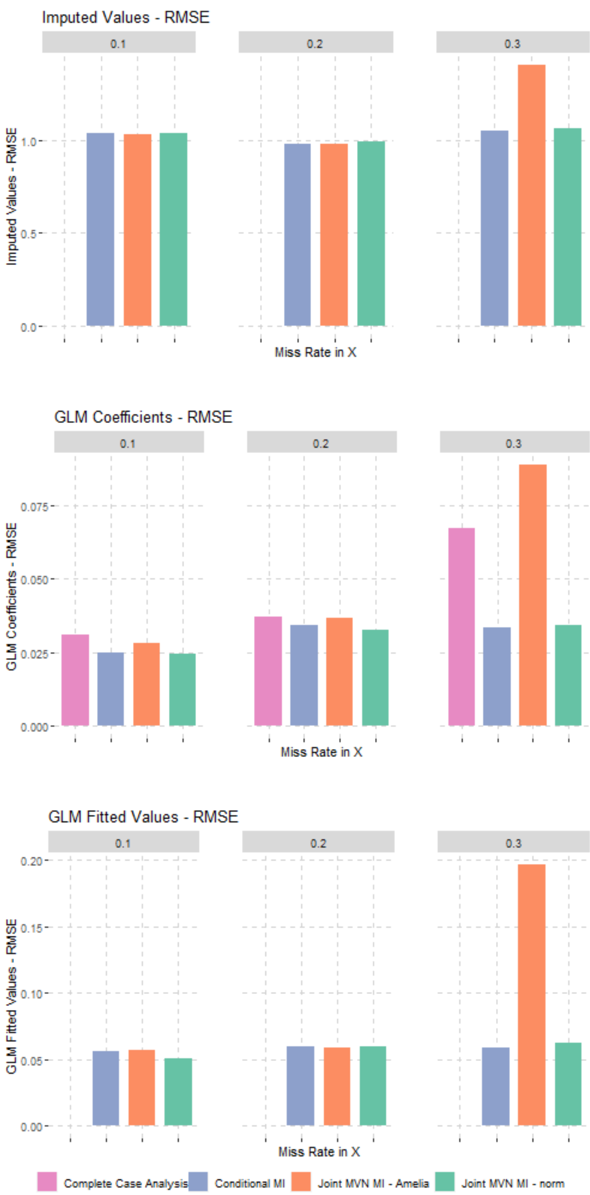

Figure 11: PMM imputation on Y being continuous. Neither approach seems to work better than the other.

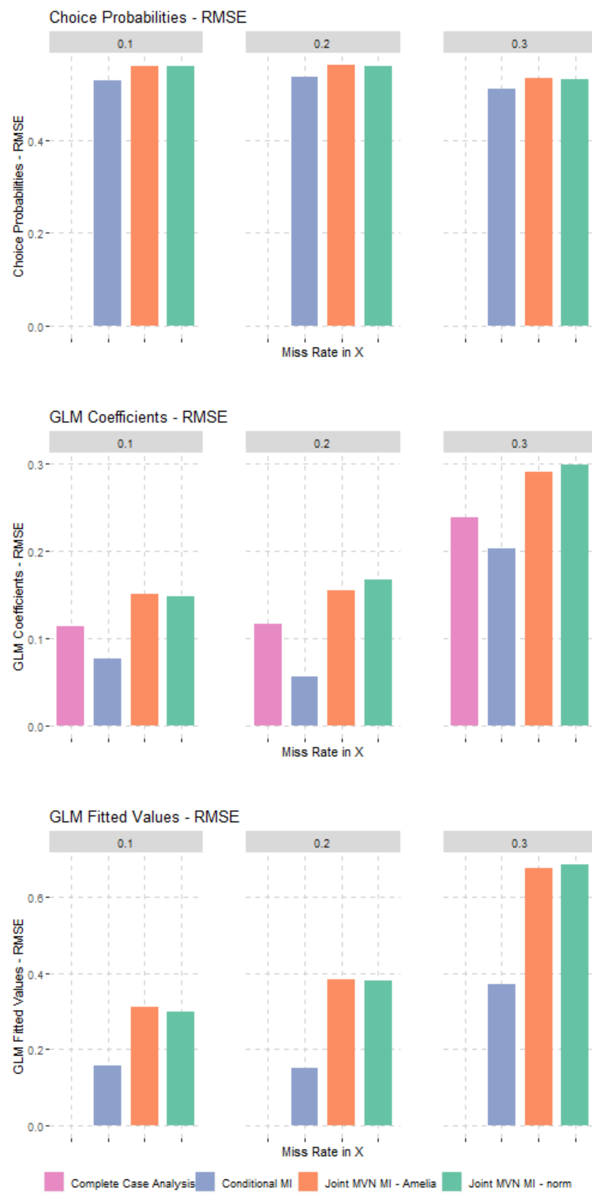

Figure 12: PMM imputation on Y being ordinal. Again, conditional MI dominates joint MVN MI in all scenarios.