



Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II.

doi: 10.18637/jss.v000.i00

Imputation of Incomplete Multilevel Data with **mice**

Hanne Oberman
Utrecht University

Johanna Munoz Avila
University Medical Center Utrecht

Valentijn de Jong
University Medical Center Utrecht

Gerko Vink
Utrecht University

Thomas Debray
University Medical Center Utrecht

Abstract

Tutorial paper on imputing incomplete multilevel data with **mice**. Including methods for ignorable and non-ignorable missingness.

Keywords: missing data, multilevel, clustering, **mice**, R.

1. Introduction

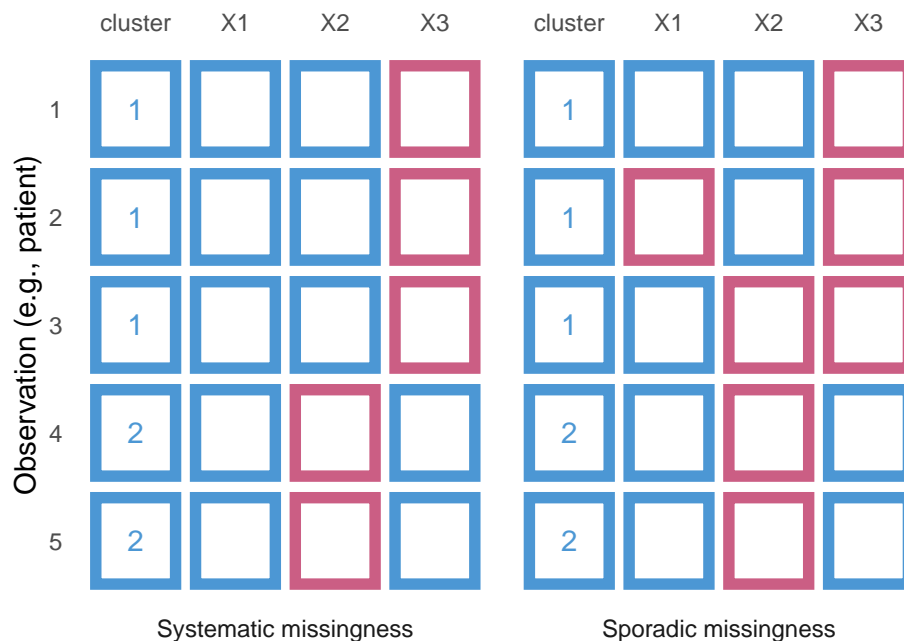
1.1. Multilevel data

- What is clustering/multilevel data? In this paper, we discuss grouped observations, not longitudinal data (within-patient clustering). -> ADD: timeseries also in Discussion section.
- What do we mean by clustering? In the medical field: Clustering by studies (IPDMA), hospitals in registries, multi-center studies etc. In other fields: e.g. official stats clustering at country-level, or social sciences clustering at school-level (related to the sampling design).
- What is heterogeneity? I.e. variability within studies vs. variability between studies

- What does multilevel data look like? ADD: figure to show difference between patient-level datapoints vs cluster-level datapoints. Maybe also add different data frame formats (or just explain in text that there's long and wide formats).
- What methods are required to analyze multilevel data? Add references, e.g. [Hox, Moerbeek, and van de Schoot \(b\)](#) and [de Jong, Moons, Eijkemans, Riley, and Debray](#). At least explain difference random effects for intercept term, predictor effects, and/or variance residual error.

1.2. Missing data

- Why/where does missingness occur in multilevel data? I.e., not only patient-level but also cluster-level.
- How can we categorize this? Systematic vs sporadic missingness, see [Resche-Rigon, White, Bartlett, Peters, Thompson, and Group](#). ADD: visualization of systematic vs sporadic missingness. Within systematic we have always missing (same value per cluster) and non-measured variables (may differ per patient). TODO: adjust md pattern to match text. -> syst may vary or same for all patients (observations/participants).



- What kinds of missingness are there? ADD: missingness mechanisms here. See e.g. [Yucel and Hox, van Buuren, and Jolani \(a\)](#).
- Why are standard (ad hoc) missing data methods not well suited?

- What types of multilevel methods are available? General overview of approaches, see [Audigier, White, Jolani, Debray, Quartagno, Carpenter, van Buuren, and Resche-Rigon](#) and [Grund, Lüdtke, and Robitzsch](#). E.g., imputation of study level versus patient-level covariates, and one-stage imputation versus two-stage imputation methods.
- Additional difficulty that is addressed in this tutorial: MNAR data.

1.3. Aim of this paper

- Provide practical guidelines with code snippets for imputation of incomplete multilevel data.
- We focus on the workflow for conditional modeling (not JOMO) in `mice`. Refer to other packages: `mitml`, `miceadds`, `mdmb`.
- Case study options: `metamisc::impact` (real IPD on traumatic brain injuries, without NAs), `mice::popularity` (simulated data on school kids, with MNAR/MAR mixture). TODO: Check example data Gelman.
- Introduce case study and set scope of this tutorial: We're providing an overview of implementations. It's up-to the reader to decide which strategy suits their data. So we won't go into detail for the different methods (and equations). This paper is just a software tutorial. We'll keep it practical. -> ADD: some kind of help function that suggests a suitable predictor matrix to the user, given a certain analysis model.

2. Workflows

We'll use the IMPACT data (`metamisc::impact`) and a MAR/MNAR version of the `mice::popmis` data (i.e., a variation on the Hox (2010) popularity data, where the missingness in the variables is either missing at random (MAR) or missing not at random (MNAR)). -> ask whether we can use the Heckman repo data or simulate data ourselves

Heckman options:

- `leiden85`
- `GJRM::hiv` (<https://rdr.io/github/egeminiani/GJRM/man/hiv.html>)
- `simulating`
- `IMPACT`

2.1. Case study I: IMPACT

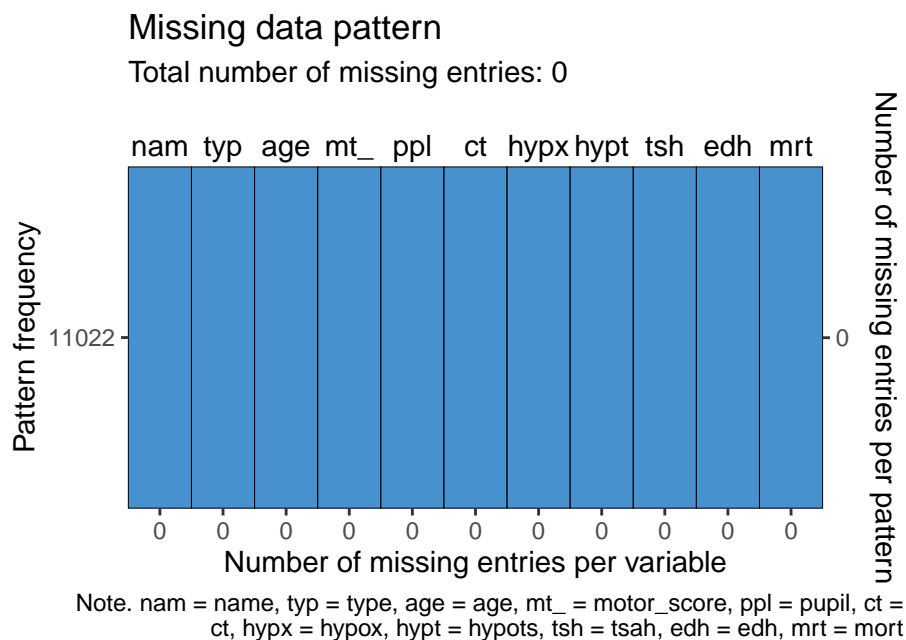
- What does the data look like? `impact` is traumatic brain injury data with patients clustered in studies, $n_{\text{participants}} = 11022$ and $n_{\text{clusters}} = 15$, on the following 11 variables:

- **name** Name of the study,
- **type** Type of study (RCT: randomized controlled trial, OBS: observational cohort),
- **age** Age of the patient,
- **motor_score** Glasgow Coma Scale motor score,
- **pupil** Pupillary reactivity,
- **ct** Marshall Computerized Tomography classification,
- **hypox** Hypoxia (0=no, 1=yes),
- **hypots** Hypotension (0=no, 1=yes),
- **tsah** Traumatic subarachnoid hemorrhage (0=no, 1=yes),
- **edh** Epidural hematoma (0=no, 1=yes),
- **mort** 6-month mortality (0=alive, 1=dead).

```

/\      /\
{ '---' }
{ 0    0 }
==> V <== No need for mice. This data set is completely observed.
\  \|/  /
 '-----'

```



-> Why are there no missings? According to the **vignette**, the data is already imputed (Steyerberg et al, 2008).

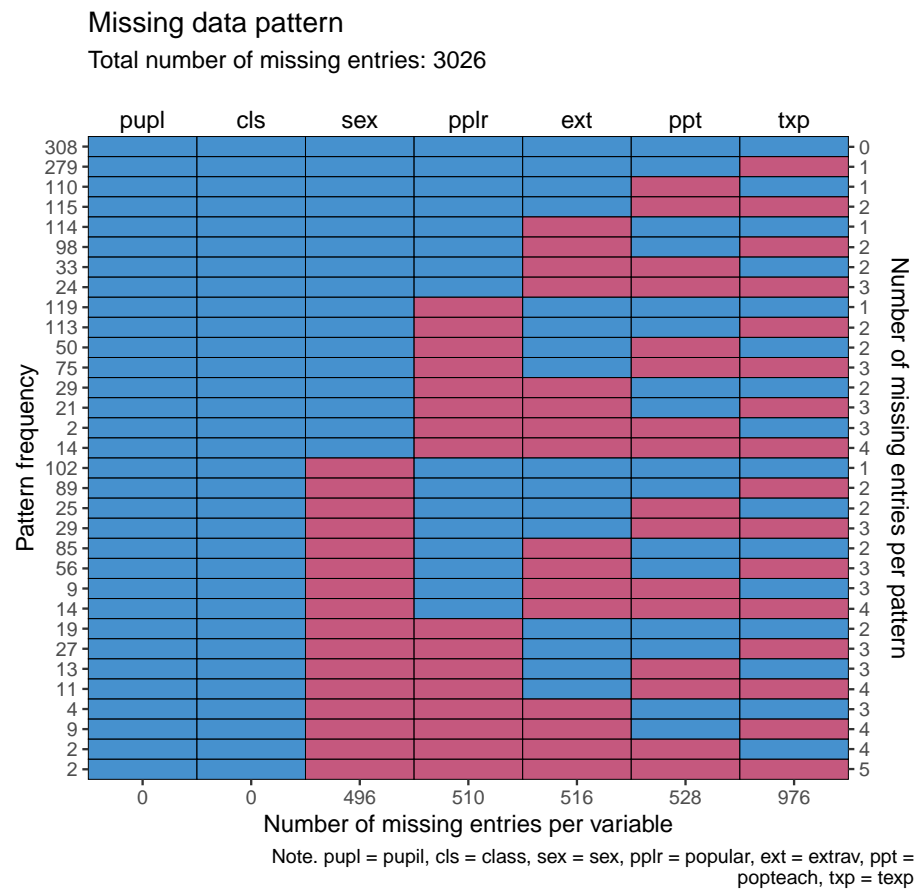
- MAR miss varying by cluster. Obs data patt differ per cluster. E.g., in cluster 1 miss depends on age but not in cluster two. Split the dataframe and run **ampute()** on each cluster. -> TODO: also make MNAR missingness to heckman model. Maybe based on **ct**

variable? Inclusion-selection variable. -> otherwise: use `leiden85` data on blood pressure with MNAR. Then run cox regression like the boshuizen article but with living situation as clusters. -> TODO: get analyses from https://www.gerkovink.com/mimp/Contents/Exercises/Day%203%20-%20Wednesday/Sensitivity_analysis/Sensitivity_analysis.html.

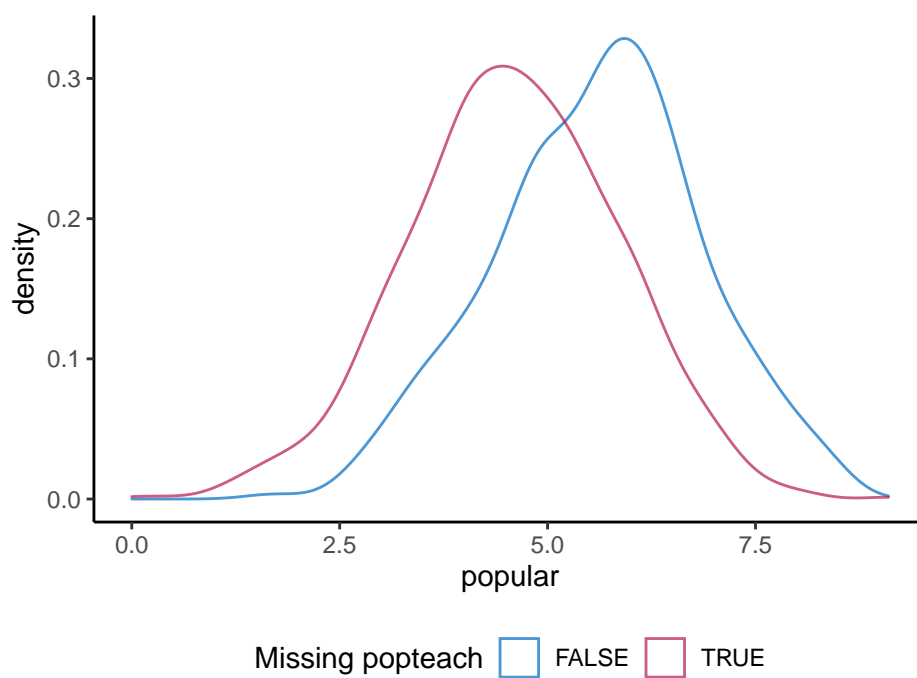
- ADD: `multilevel_ampute()` wrapper function in `mice`.

2.2. Case Study II: Popularity

- What does the data look like? `popNCR` is a simulated dataset with pupils clustered in classes, $n_{\text{participants}} = 2000$, $n_{\text{clusters}} = 100$, on 7 variables:
 - `pupil` Pupil number within class,
 - `class` Class number,
 - `extrav` Pupil extraversion,
 - `sex` Pupil gender,
 - `teexp` Teacher experience (years),
 - `popular` Pupil popularity,
 - `popteach` Teacher popularity.
- What are the ICCs? For `popular` the ICC is 0.33. For `popteach` it is 0.31. It would be wise to use multilevel modeling.
- What does the missingness look like? Induced MAR/MNAR missingness. Missing data pattern:



- Does the missing data of **popular** depend on **popteach**? Does the missingness in teacher popularity depend on pupil popularity? -> Check this by making a histogram of **popteach** separately for the pupils with known popularity and missing popularity, and the other way around.



- We do see that the distribution for the missing `popular` is further to the right than the distribution for observed `popular`. This would indicate a right-tailed MAR missingness. In fact this is exactly what happens, because we created the missingness in these data ourselves. But we made it observable by examining the relations between the missingness in `popular` and the observed data in `popteach`. There is also a dependency

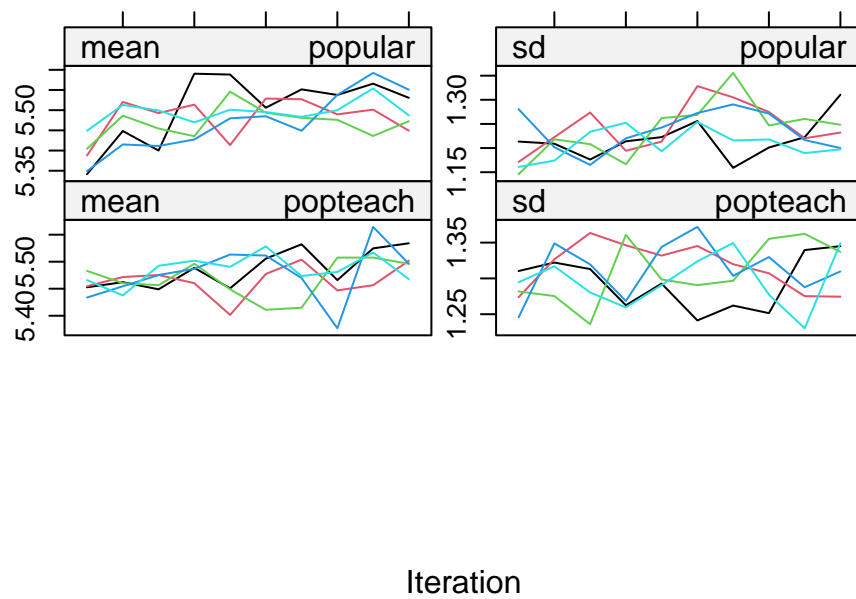
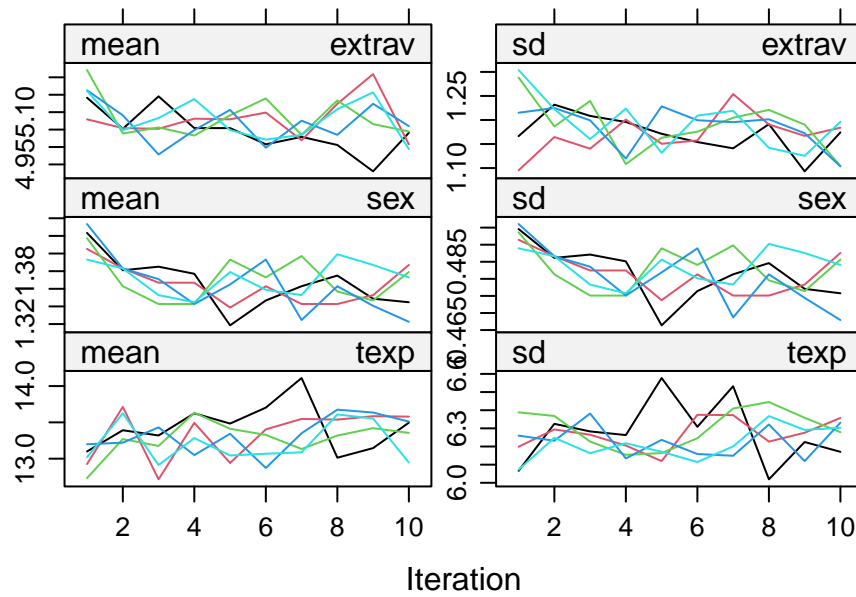
between the missingness in teacher popularity and pupil popularity. The relation seems to be right-tailed as well.

- We can impute the missingness the ‘standard’ way, ignoring the multilevel structure of the data. This is surely invalid, given the high ICCs, but we’ll do it anyways.
- We’ll use predictive mean matching to impute the continuous variables (some appear to be somewhat ordinal), and logistic regression to impute the binary variable `sex`. We do not use the observation identifier `pupil` or cluster identifier `class` as predictors to impute other variables.

```
R> # dry run to get imputation parameters
R> ini <- mice(pop, maxit = 0)
R>
R> # extract predictor matrix and adjust
R> pred <- ini$pred
R> pred[, c("class", "pupil")] <- 0
R> pred
```

	pupil	class	extrav	sex	texp	popular	popteach
pupil	0	0	1	1	1	1	1
class	0	0	1	1	1	1	1
extrav	0	0	0	1	1	1	1
sex	0	0	1	0	1	1	1
texp	0	0	1	1	0	1	1
popular	0	0	1	1	1	0	1
popteach	0	0	1	1	1	1	0

```
R> # impute the data, ignoring the cluster structure
R> imp_ignored <- mice(pop, maxit = 10, pred = pred, print = FALSE)
R>
R> # check convergence of the imputation model
R> plot(imp_ignored)
```

```
R> # compare descriptives before and after imputation
R> psych::describe(pop)[, c("n", "mean", "median", "min", "max", "sd")]
```

	n	mean	median	min	max	sd
pupil	2000	10.65	11.0	1	26.0	5.97

```

class*  2000 50.37   51.0   1 100.0 29.08
extrav  1484 5.31    5.0   1  10.0  1.29
sex*    1504 1.56    2.0   1   2.0  0.50
texp    1024 11.80   12.0   2  25.0  6.26
popular 1490 4.83    4.8   0   9.1  1.34
popteach 1472 4.83   5.0   1  10.0  1.36

```

```
R> psych::describe(mice::complete(imp_ignored))[, c("n", "mean", "median", "min", "max", "sd", "skewness", "kurtosis")]
```

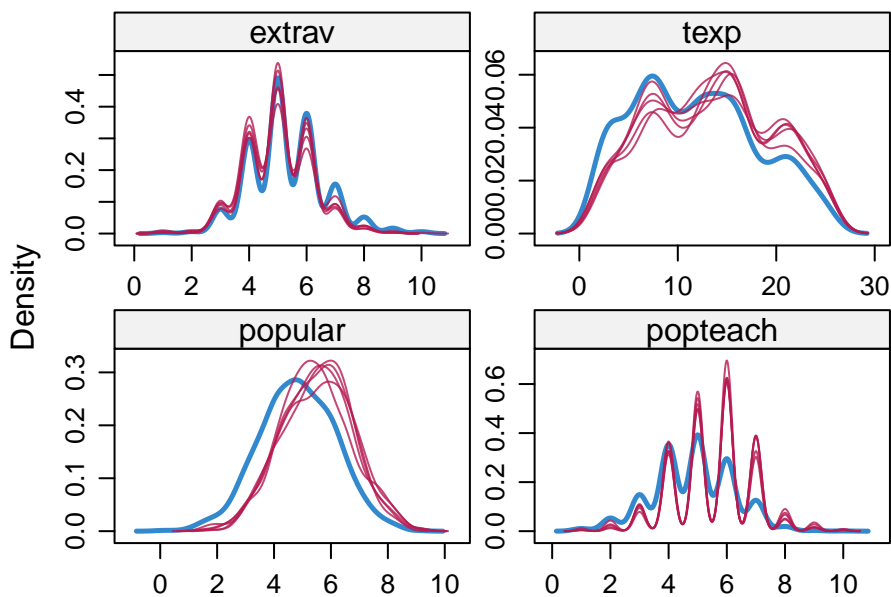
```

      n mean median min  max  sd
pupil 2000 10.65    11   1 26.0 5.97
class* 2000 50.37    51   1 100.0 29.08
extrav 2000  5.24     5   1  10.0  1.26
sex*    2000  1.51     2   1   2.0  0.50
texp    2000 12.63    13   2  25.0  6.27
popular 2000  5.01     5   0   9.1  1.37
popteach 2000  5.02     5   1  10.0  1.39

```

```
R> # further inspection of the imputations
```

```
R> densityplot(imp_ignored)
```



```
R> # compare ICCs before and after imputation
```

```

R> ICCs <- data.frame(
+   vars = c("popular", "popteach", "texp"),
+   incomplete = c(multilevel::ICC1(aov(popular ~ class, pop)),
+                  multilevel::ICC1(aov(popteach ~ class, pop)),

```

```
+           multilevel::ICC1(aov(texp ~ class, pop))),
+   ignored = c(multilevel::ICC1(aov(popular ~ class, complete(imp_ignored))),
+               multilevel::ICC1(aov(popteach ~ class, complete(imp_ignored))),
+               multilevel::ICC1(aov(texp ~ class, complete(imp_ignored))))
+ )
R> ICCs
```

```
      vars incomplete  ignored
1 popular  0.3280070 0.2743092
2 popteach 0.3138658 0.2479930
3      texp  1.0000000 0.4510366
```

- As the original ICCs show, 100% of the variance in `texp` can be attributed to the clustering variable `class`. This tells us that the multilevel structure of the data should be taken into account. If we don't, we'll end up with incorrect imputations, biasing the effect of the clusters towards zero.
- We can also observe that the teacher experience increases slightly after imputation. This is due to the MNAR missingness in `texp`. Higher values for `texp` have a larger probability to be missing. This may not a problem, however, if at least one pupil in each class has teacher experience recorded, we can deductively impute the correct (i.e. true) value for every pupil in the class.
- We'll now use `class` as a predictor to impute all other variables.

```
R> # adjust the predictor matrix
R> pred <- ini$pred
R> pred[, "pupil"] <- 0
R> pred
```

	pupil	class	extrav	sex	texp	popular	popteach
pupil	0	1	1	1	1	1	1
class	0	0	1	1	1	1	1
extrav	0	1	0	1	1	1	1
sex	0	1	1	0	1	1	1
texp	0	1	1	1	0	1	1
popular	0	1	1	1	1	0	1
popteach	0	1	1	1	1	1	0

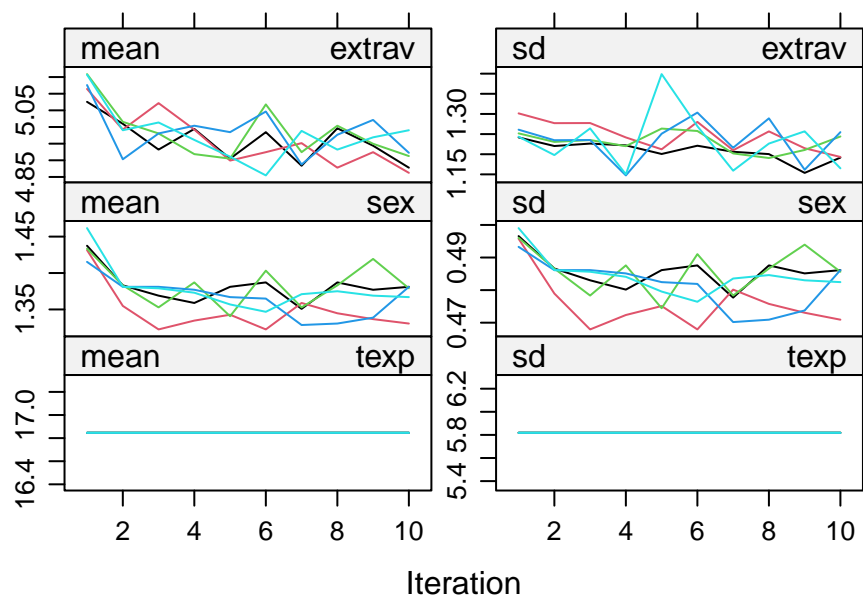
```
R> # impute the data, cluster as predictor
R> imp_predictor <- mice(pop, maxit = 10, pred = pred, print = FALSE)
```

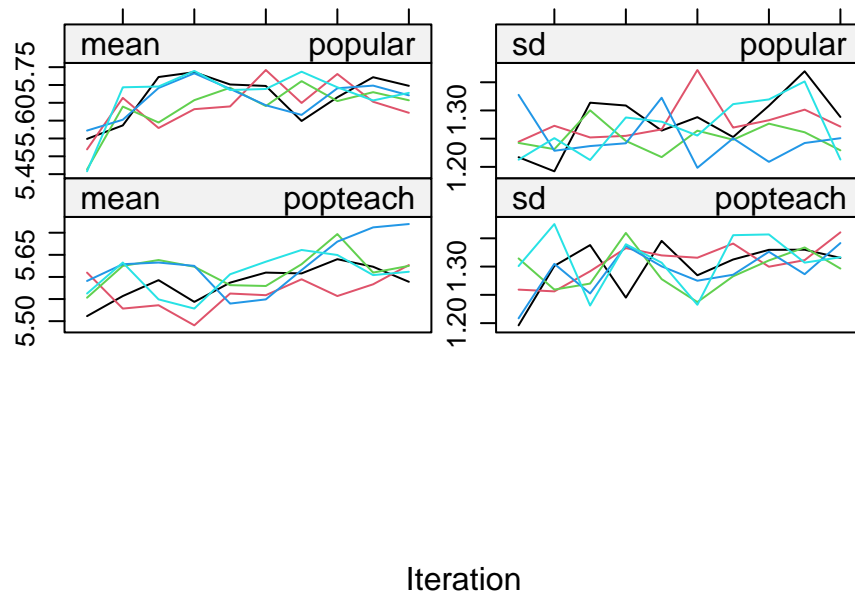
Warning: Number of logged events: 190

```
R> # check logged events
R> head(imp_predictor$loggedEvents)
```

	it	im	dep	meth	out
1	1	1	popular	pmm	texp
2	1	1	popteach	pmm	texp
3	1	2	popular	pmm	texp
4	1	2	popteach	pmm	texp
5	1	3	popular	pmm	texp
6	1	3	popteach	pmm	texp

```
R> ## "The mice() function detects multicollinearity, and solves the problem by removing c
R>
R> # check convergence of the imputation model
R> plot(imp_predictor)
```





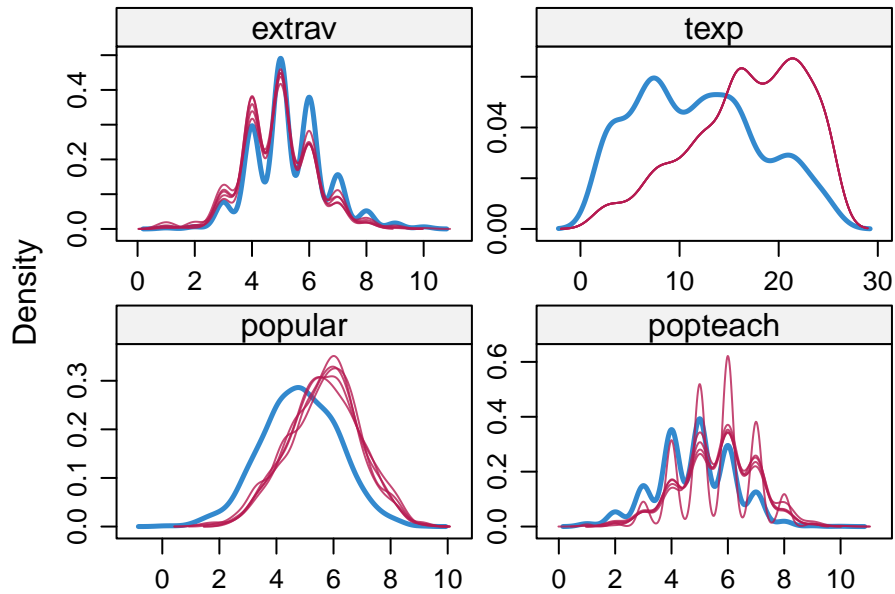
```
R> # compare descriptives before and after imputation
R> psych::describe(pop)[, c("n", "mean", "median", "min", "max", "sd")]
```

	n	mean	median	min	max	sd
pupil	2000	10.65	11.0	1	26.0	5.97
class*	2000	50.37	51.0	1	100.0	29.08
extrav	1484	5.31	5.0	1	10.0	1.29
sex*	1504	1.56	2.0	1	2.0	0.50
texp	1024	11.80	12.0	2	25.0	6.26
popular	1490	4.83	4.8	0	9.1	1.34
popteach	1472	4.83	5.0	1	10.0	1.36

```
R> psych::describe(mice::complete(imp_predictor))[, c("n", "mean", "median", "min", "max", "sd")]
```

	n	mean	median	min	max	sd
pupil	2000	10.65	11.0	1	26.0	5.97
class*	2000	50.37	51.0	1	100.0	29.08
extrav	2000	5.20	5.0	1	10.0	1.28
sex*	2000	1.52	2.0	1	2.0	0.50
texp	2000	14.26	15.0	2	25.0	6.55
popular	2000	5.05	5.1	0	9.1	1.38
popteach	2000	5.03	5.0	1	10.0	1.39

```
R> # further inspection of the imputations
R> densityplot(imp_predictor)
```



```
R> # compare ICCs before and after imputation
R> ICCs <- ICCs %>% cbind(
+   predictor = c(multilevel::ICC1(aov(popular ~ class, complete(imp_predictor)))
+                 multilevel::ICC1(aov(popteach ~ class, complete(imp_predictor)))
+                 multilevel::ICC1(aov(texp ~ class, complete(imp_predictor))))
+ )
R> ICCs
```

	vars	incomplete	ignored	predictor
1	popular	0.3280070	0.2743092	0.3694899
2	popteach	0.3138658	0.2479930	0.3335612
3	texp	1.0000000	0.4510366	1.0000000

- Now, we can clearly see that the imputed values of `texp` are higher than the observed values, which is in line with right-tailed MNAR.
- The ICCs are way more in line with the ICCs in the incomplete data. But this is a quick and dirty way of imputing multilevel data. We *should* be using a multilevel model.

2.3. Amputation

•

2.4. Modeling choices

- Which models will we discuss? We'll build the model to grow in complexity. The final model is the most complex but also the most versatile.
- Note on model complexity: Typically, we should at least use random intercepts, but often random slopes as well. Ideally we impute with random everything and heteroscedastic errors: most generic method (no worry about congeniality, but don't mention the term) -> Refer to other papers for background, we'll focus just on the software implementation of the situations mentioned there. Sometimes there's little reason to assume some variable is affected by heterogeneity. -> Refer to [Meng](#), an Audigier paper, and a paper by Grund on congeniality and random slopes.
- Step 0: As predictor + CCA to scare off users
- Step 1: Random intercepts
- Step 2: Random slopes
- Step 3: Residuals
- Heckman model for MNAR
- What do the different implementations look like? How to define the imputation model(s) in `mice`?

2.5. Step 0

- AKA multilevel imputation for dummies.
- Doesn't work for systematic missingness.

2.6. Step 1-3 + MNAR

- TODO: fill in.

2.7. Pooling

- Analysis of scientific interest.
- Pooling using `mitml`.
- Pooling 'regular' parameters vs more 'exotic' parameters (SE of residual errors, or autocorrelation)
- ADD: export `mids` objects to other packages like `lme4` or `coxme`?

3. Discussion

- JOMO in mice → on the side for now
- Additional levels of clustering
- Timeseries: and polynomial relationship in the clustering.

References

- Audigier V, White IR, Jolani S, Debray TPA, Quartagno M, Carpenter J, van Buuren S, Resche-Rigon M (????). “Multiple Imputation for Multilevel Data with Continuous and Binary Variables.” **33**(2), 160–183. ISSN 0883-4237, 2168-8745. doi:10.1214/18-STS646. 1702.00971, URL <https://projecteuclid.org/journals/statistical-science/volume-33/issue-2/Multiple-Imputation-for-Multilevel-Data-with-Continuous-and-Binary-Variables/10.1214/18-STS646.full>.
- de Jong VMT, Moons KGM, Eijkemans MJC, Riley RD, Debray TPA (????). “Developing More Generalizable Prediction Models from Pooled Studies and Large Clustered Data Sets.” **40**(15), 3533–3559. ISSN 1097-0258. doi:10.1002/sim.8981. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8981>.
- Grund S, Lüdtke O, Robitzsch A (????). “Multiple Imputation of Missing Data for Multilevel Models: Simulations and Recommendations.” **21**(1), 111–149. ISSN 1094-4281. doi:10.1177/1094428117703686. URL <https://doi.org/10.1177/1094428117703686>.
- Hox J, van Buuren S, Jolani S (????a). “Incomplete Multilevel Data: Problems and Solutions.” In J Harring, L Stapleton, S Beretvas (eds.), *Advances in Multilevel Modeling for Educational Research: Addressing Practical Issues Found in Real-World Applications*, CILVR Series on Latent Variable Methodology, pp. 39–62. Information Age Publishing Inc. ISBN 978-1-68123-328-4. URL [https://books.google.nl/books?id=HAcoDwAAQBAJ&pg=PR12&lpg=PR12&dq=Advances+in+Multilevel+Modeling+for+Educational+Research:+Addressing+Practical+Issues+Found+in+Real-World+Applications+\(CILVR+Series+on+Latent+Variable+Methodology\)&source=bl&ots=Jh5XSVCbSp&sig=ACfU3U3_f-ynwmemsOVqs8SLHQ53B98kdw&hl=en&sa=X&ved=2ahUKEwjT-T0mpzyAhWPC0wKHdqFDIIQ6AF6BAgCEAM](https://books.google.nl/books?id=HAcoDwAAQBAJ&pg=PR12&lpg=PR12&dq=Advances+in+Multilevel+Modeling+for+Educational+Research:+Addressing+Practical+Issues+Found+in+Real-World+Applications+(CILVR+Series+on+Latent+Variable+Methodology)&source=bl&ots=Jh5XSVCbSp&sig=ACfU3U3_f-ynwmemsOVqs8SLHQ53B98kdw&hl=en&sa=X&ved=2ahUKEwjT-T0mpzyAhWPC0wKHdqFDIIQ6AF6BAgCEAM).
- Hox JJ, Moerbeek M, van de Schoot R (????b). *Multilevel Analysis: Techniques and Applications, Third Edition*. Routledge. ISBN 978-1-317-30868-3. iLD_DwAAQBAJ.
- Meng XL (????). “Multiple-Imputation Inferences with Uncongenial Sources of Input.” **9**(4), 538–558. ISSN 0883-4237, 2168-8745. doi:10.1214/ss/1177010269. URL <https://projecteuclid.org/euclid.ss/1177010269>.

- Resche-Rigon M, White IR, Bartlett JW, Peters SA, Thompson SG, Group obotPIS (???). “Multiple Imputation for Handling Systematically Missing Confounders in Meta-Analysis of Individual Participant Data.” **32**(28), 4890–4905. ISSN 1097-0258. doi:10.1002/sim.5894. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5894>.
- Yucel RM (???). “Multiple Imputation Inference for Multivariate Multilevel Continuous Data with Ignorable Non-Response.” **366**(1874), 2389–2403. doi:10.1098/rsta.2008.0038. URL <https://royalsocietypublishing.org/doi/10.1098/rsta.2008.0038>.

Affiliation:

Hanne Oberman
Utrecht University
Padualaan 14
3584 CH Utrecht
E-mail: h.i.oberman@uu.nl
URL: <https://hanneoberman.github.io/>