technische universität
dortmund

Prediction Rule Ensembles

Interpretable predictions in the presence of missing data

Philipp Doebler, Jakob Schwerter, Vincent Schroeder, Marjolein Fokkema

GEBF 2024, Potsdam, 18.03.2024

FAIR

Funded by:
Ministry of Culture and Science
of the State of
North Rhine-Westphalia

Acknowledgements

The project "From Prediction to Agile Interventions in the Social Sciences (FAIR)" is receiving funding from the programme "Profilbildung 2020", an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia. The sole responsibility for the content of this publication lies with the authors.

Outline

A brief introduction to prediction rule ensembles

Current challenges: Missing data and coarse data

- A simulation study

Recommendations

- Should we simplify predictors?
- How should one impute missing data?

Motivation

- With few predictors it's ...
  - ... hard to integrate different theoretical perspectives
    (Byrnes & Miller, 2007)
  - ... easy to miss factors relevant for personalized interventions
    (Bernacki et al, 2020; Winne, 2023)

Motivation

- With few predictors it's ...
  - ... hard to integrate different theoretical perspectives
    (Byrnes & Miller, 2007)
  - ... easy to miss factors relevant for personalized interventions
    (Bernacki et al, 2020; Winne, 2023)

- Empirical Example: Why are women less likely to pursue STEM-related degrees?
  - Multiple theories used to explain the research question
    - Expectancy-value theory (Eccles et al., 1983)
    - Theory of circumscription and compromise (Gottfredson, 2005)
    - Differential effects model (Parker et al., 2012, 2014)
    - Internal-external frames of reference model (Marsh, 1986)
    - Goal congruity research (Diekman et al., 2010, 2017)

Motivation

- With few predictors it's ...
  - ... hard to integrate different theoretical perspectives
    (Byrnes & Miller, 2007)
  - ... easy to miss factors relevant for personalized interventions
    (Bernacki et al, 2020; Winne, 2023)

- Empirical Example: Why are women less likely to pursue STEM-related degrees?
  - Multiple theories used to explain the research question
- **But:** How to integrate several complementary theoretical perspectives?
  - How to fit variables of multiple theories in one analysis?

Motivation

- With few predictors it's ...
  - ... hard to integrate different theoretical perspectives
    (Byrnes & Miller, 2007)
  - ... easy to miss factors relevant for personalized interventions
    (Bernacki et al, 2020; Winne, 2023)

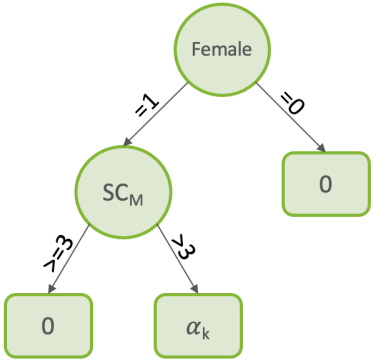
- Empirical Example: Why are women less likely to pursue STEM-related degrees?
  - Multiple theories used to explain the research question
- **But:** How to integrate several complementary theoretical perspectives?
  - How to fit variables of multiple theories in one analysis?

$\Rightarrow$ Prediction Rule Ensembles

- (ensembles of) trees help understand how variables (non-linearly) interact with each other

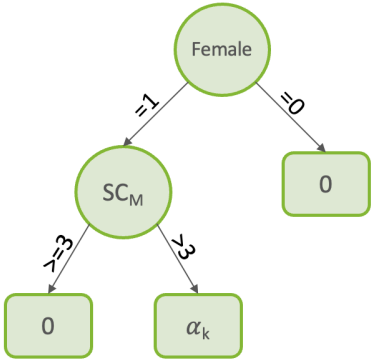
A brief introduction to prediction rule ensembles

Warm-up: A single prediction rule for STEM major prediction



- *If [condition], then [prediction].*
- one path through tree = one rule
- E.g.: If (Female = 1 & $SC_M > 3$), then add α_k to predicted value.
- at each node: a split of the sample (=binary decision)
- depth of tree can vary (here: 2)

Warm-up: A single prediction rule for STEM major prediction



- *If [condition], then [prediction].*
- one path through tree = one rule
- E.g.: If (Female = 1 & $SC_M > 3$), then add α_k to predicted value.
- at each node: a split of the sample (=binary decision)
- depth of tree can vary (here: 2)

- Each tree is highly interpretable.
- Handles numeric, categorical and count responses

- Key idea for Prediction Rule Ensembles (PREs): combine several trees
- STEM example: STEM major choice predicted by combinations of rules and linear terms (potentially a lot of rules!)

Prediction Rule Ensemble (Friedman & Popescu, 2008; Fokkema & Strobl, 2020)

$$STEM_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k r_k(x_i) + \sum_{j=1}^{J} \beta_j l_j(x_i), \text{ where } \alpha_0, \alpha_1, ..., \beta_1, \beta_2, ... \text{ minimize}$$

$$\underbrace{\sum_{i=1}^{N} L\left(y_i, \alpha_0 + \sum_{k=1}^{K} \alpha_k r_k(x_i) + \sum_{j=1}^{J} \beta_j l_j(x_{i,j})\right)}_{\text{loss = measure of discrepancy of predictions and observations}} + \underbrace{\lambda\left(\sum_{k=1}^{K} |\alpha_k| + \sum_{j=1}^{J} |\beta_j|\right)}_{\text{penalty for model complexity}}$$

Prediction Rule Ensemble (Friedman & Popescu, 2008; Fokkema & Strobl, 2020)

$$STEM_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k r_k(x_i) + \sum_{j=1}^{J} \beta_j l_j(x_i), \text{ where } \alpha_0, \alpha_1, ..., \beta_1, \beta_2, ... \text{ minimize}$$

$$\underbrace{\sum_{i=1}^{N} L\left(y_i, \alpha_0 + \sum_{k=1}^{K} \alpha_k r_k(x_i) + \sum_{j=1}^{J} \beta_j l_j(x_{i,j})\right)}_{\text{loss = measure of discrepancy of predictions and observations}} + \underbrace{\lambda\left(\sum_{k=1}^{K} |\alpha_k| + \sum_{j=1}^{J} |\beta_j|\right)}_{\text{penalty for model complexity}}$$

- $r_k(x_i)$: (set of) prediction rule(s) $\Rightarrow \alpha_k$: coefficient for a prediction rule
- $l_j(x_j)$: winsorized predictors $x_j \Rightarrow \beta_j$: coefficient for a linear term

Prediction Rule Ensemble (Friedman & Popescu, 2008; Fokkema & Strobl, 2020)

$$STEM_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k r_k(x_i) + \sum_{j=1}^{J} \beta_j l_j(x_i), \text{ where } \alpha_0, \alpha_1, ..., \beta_1, \beta_2, ... \text{ minimize}$$

$$\underbrace{\sum_{i=1}^{N} L\left(y_i, \alpha_0 + \sum_{k=1}^{K} \alpha_k r_k(x_i) + \sum_{j=1}^{J} \beta_j l_j(x_{i,j})\right)}_{\text{loss = measure of discrepancy of predictions and observations}} + \underbrace{\lambda\left(\sum_{k=1}^{K} |\alpha_k| + \sum_{j=1}^{J} |\beta_j|\right)}_{\text{penalty for model complexity}}$$

- Default of maximal tree depth: 3 $\Rightarrow$ at most 3-way interactions can be captured
- Unbiased conditional inference tree induction algorithm (Hothorn et al. 2006)
- RuleFit algorithm: (Relaxed) LASSO to obtain a sparse final rule ensemble to optimize interpretability and avoid overfitting (Friedman & Popescu, 2008; Fokkema & Strobl, 2020)

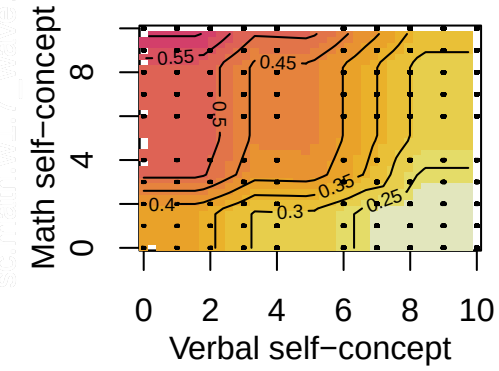
Example PRE results

Selected rule	Importance	Coefficient	SD
Female $= 1$ & math self-concept > 3	0.07	0.46	0.19

Example PRE results

Selected rule	Importance	Coefficient	SD
Female $= 1$ & math self-concept > 3	0.07	0.46	0.19
math self-concept > 3 & verbal self-concept < 4	0.11	0.44	0.25
$\vdots$	$\vdots$	$\vdots$	$\vdots$

- partial dependency plots aid interpretation in the presence of multiple rules and dependencies
- Convenient R package `pre` (Fokkema, 2020)



Current challenges: Missing data and coarse data

Current challenges

Question 1: What's the best way to handle missing data?

- Vanilla multiple imputation workflow challenging (many rules!)
- Strategy: Stack multiple imputed datasets (cf. Du et al., 2022)

? Which imputation method works best?

? Multiple imputations necessary, or is single imputation enough?

Current challenges

Question 1: What's the best way to handle missing data?

- Vanilla multiple imputation workflow challenging (many rules!)
- Strategy: Stack multiple imputed datasets (cf. Du et al., 2022)

? Which imputation method works best?

? Multiple imputations necessary, or is single imputation enough?

Question 2: Should one simplify variables for faster calculation and easier interpretation? Is coarser data better?

- each observed value is a potential split point
- A PRE might contain two very similar rules, e.g., $X_1 > 0.123$ and $X_1 > 0.110$.
- Rounding enforces simpler rules (e.g., $X_1 > 0.1$) and fewer potential rules

? How should one coarsen the data?

Current challenges

Question 1: What's the best way to handle missing data?

- Vanilla multiple imputation workflow challenging (many rules!)
- Strategy: Stack multiple imputed datasets (cf. Du et al., 2022)

? Which imputation method works best?

? Multiple imputations necessary, or is single imputation enough?

Question 2: Should one simplify variables for faster calculation and easier interpretation? Is coarser data better?

- each observed value is a potential split point
- A PRE might contain two very similar rules, e.g., $X_1 > 0.123$ and $X_1 > 0.110$.
- Rounding enforces simpler rules (e.g., $X_1 > 0.1$) and fewer potential rules

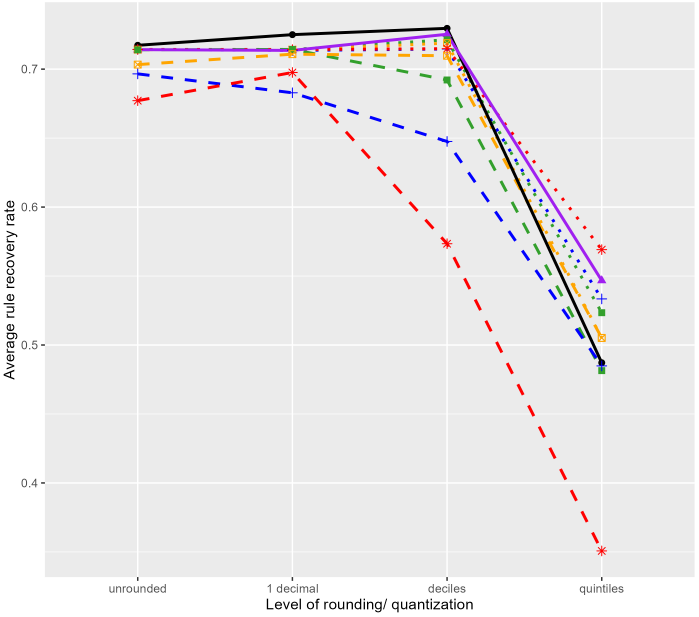
? How should one coarsen the data?

$\Rightarrow$ Simulation study for recommendations

Design (1,000 replications)

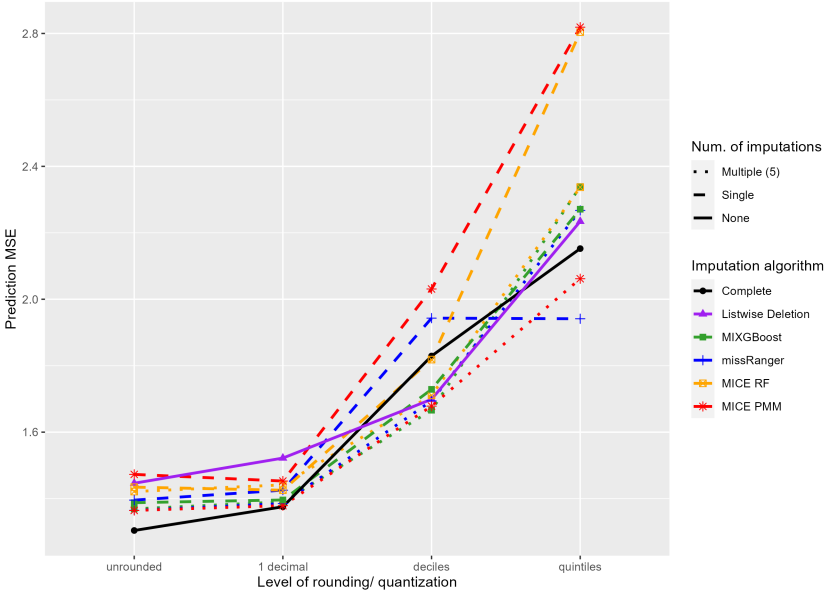
- 10 numerical predictor variables, standard normally distributed
- Fully crossed factors:
  - Sample size: N = 200 vs. N = 400
  - Missingness: 8% total missingness (random 40% of observations have 2 missings) vs. 48% (random 80% of observations have 6 missings)
  - Imputation methods: MICE PMM, MICE RF, missRanger, MIXGBoost
  - Multiple imputation (MI) vs. single imputation (SI)
  - Degree of simplification: Unrounded, 1 decimal rounding, deciles, quintiles
- 6 rules and one linear term generate the outcome variable

Rule recovery ($N = 400$, 8% missingness rate)



- Coarsening to quintiles decreases performance
- MI performs better than SI
- MICE with SI is worst
- Listwise deletion (surprisingly) well performing for rule recovery

Predictive Performance (MSE; $N = 400$, 8% missingness rate)



- Coarsening to quintiles or deciles decreases performance
- MI performs better than SI (again)
- Listwise deletion clearly subpar

Recommendations

Recommendations (Question 2)

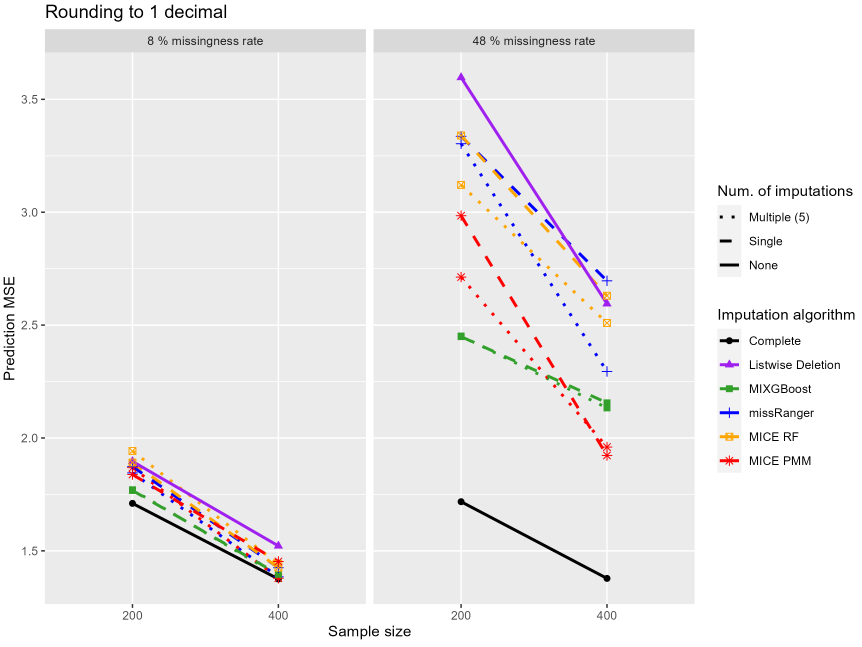
- Q: Should we simplify predictors by rounding or quantization?
- A: Yes!

Recommendations (Question 2)

- Q: Should we simplify predictors by rounding or quantization?
- A: Yes!

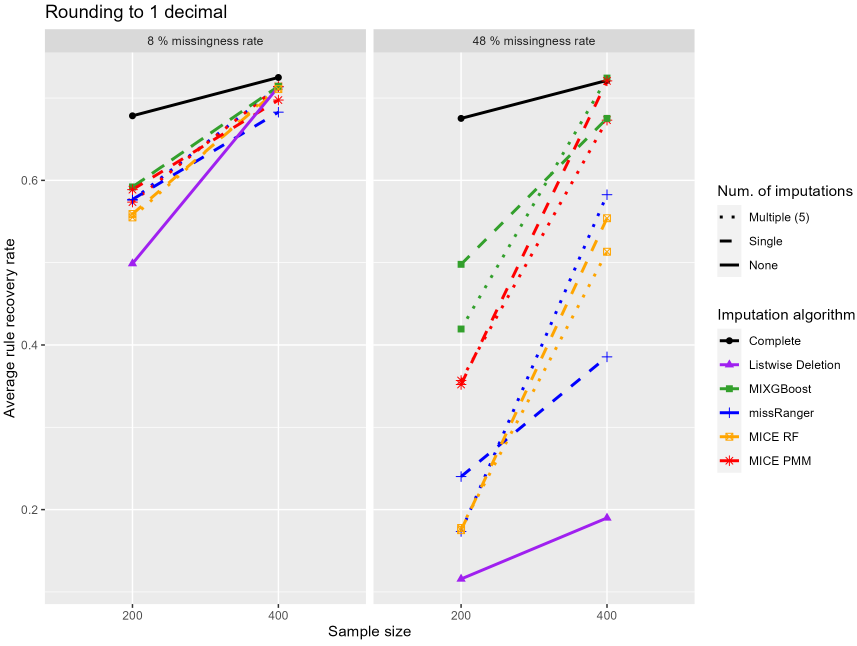
- Q: How should we coarsen the data?
- A: Rounding to first decimal (standard normal variables/z-scale) has only a modest decrease in performance but improved interpretability.
- A: Decile quantization is an option, giving up some predictive power for the sake of interpretability

Predictive Performance (MSE; rounding to 1 decimal)



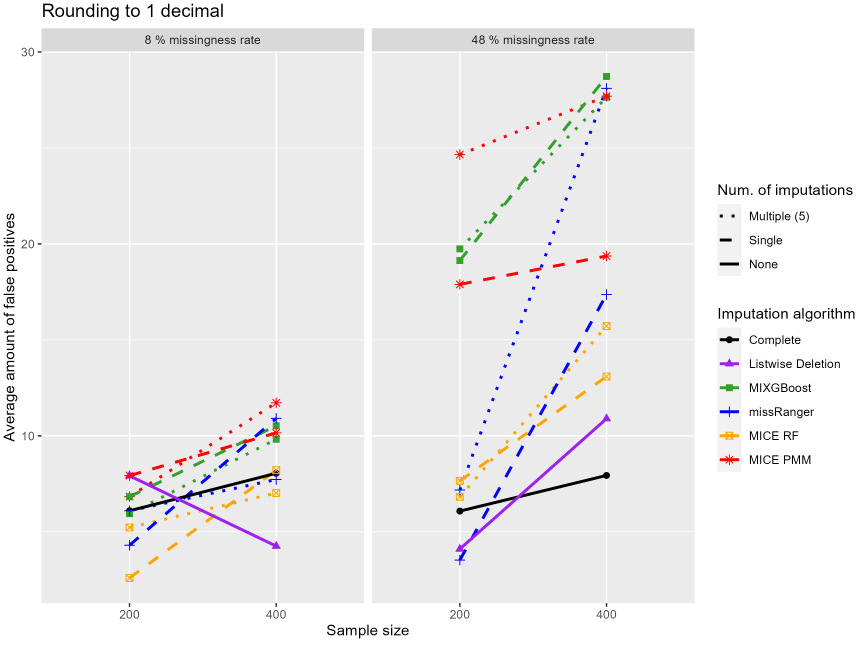
- MIXGBoost very competitive
- MICE PMM shines in larger samples
- LD not recommended

Rule recovery (rounding to 1 decimal)



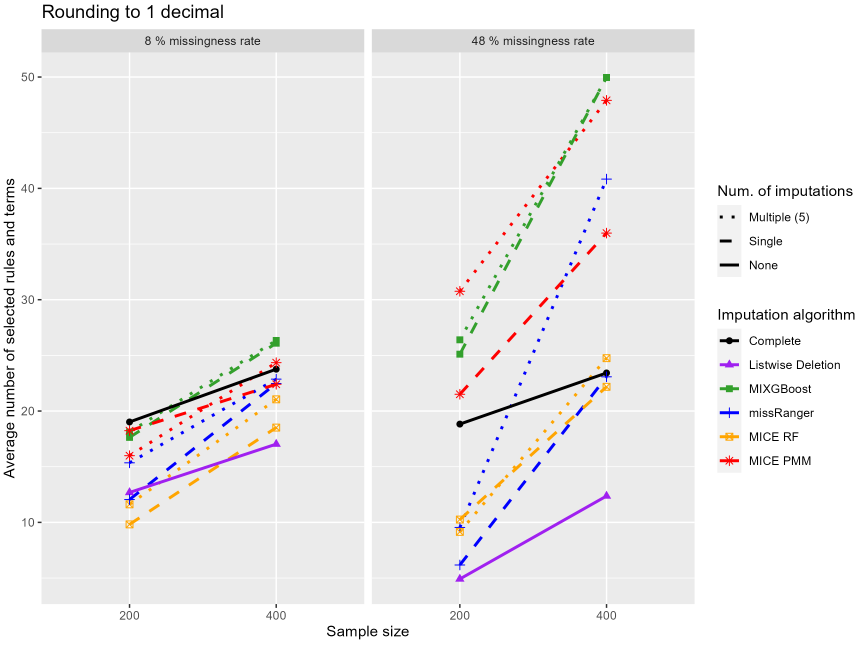
- only MIXGBoost with MI convincing in all scenarios
- MICE PMM needs a large sample when there is substantial missingness

False positives (rounding to 1 decimal)



- Good MSE and rule recovery come at the price of false positives!
- Is MICE RF a good enough compromise?

Model size (rounding to 1 decimal)



- Similarly, good MSE and rule recovery need larger, less interpretable models.

Recommendations (Question 1)

- listwise deletion only viable if missingess rate low and sample size large

Recommendations (Question 1)

- listwise deletion only viable if missingess rate low and sample size large

- MIXGBoost: strong predictive performance & good rule recovery but larger models with more false positives
- MICE PMM: large samples to be competitive
- missRanger/MICE Random Forest: smaller models, but lower prediction quality
- no clear favorite, but MIXGBoost a good default

Recommendations (Question 1)

- listwise deletion only viable if missingess rate low and sample size large

- MIXGBoost: strong predictive performance & good rule recovery but larger models with more false positives
- MICE PMM: large samples to be competitive
- missRanger/MICE Random Forest: smaller models, but lower prediction quality
- no clear favorite, but MIXGBoost a good default

- Single imputation a fallback option, if computational effort is an issue

Outlook

- Data reduction
  - Can we decrease model size?
  - Any kind of dimensionality reduction would need to keep non-linear interactions intact!
- Reduce tree complexity for interpretability
  - Constrain rule set?
  - With the right algorithms, prediction and interpretation not necessarily at odds (Rudin, 2019; Weihs & Buschfeld, 2021).
- missing completely at random (MCAR) vs. missing at random (MAR)
- R code will be available

References

- Bernacki, M. L., Vosicka, L., & Utz, J. C. (2020). **Can a brief, digital skill training intervention help undergraduates "learn to learn" and improve their stem achievement?** *Journal of Educational Psychology*, *112*(4), 765–781.
- Byrnes, J. P., & Miller, D. C. (2007). **The relative importance of predictors of math and science achievement: An opportunity–propensity analysis.** *Contemporary Educational Psychology*, *32*(4), 599–629.
- Diekman, A. B., Brown, E. R., Johnston, A. M., & Clark, E. K. (2010). **Seeking congruity between goals and roles: A new look at why women opt out of science, technology, engineering, and mathematics careers.** *Psychological science*, *21*(8), 1051–1057.
- Diekman, A. B., Steinberg, M., Brown, E. R., Belanger, A. L., & Clark, E. K. (2017). **A goal congruity model of role entry, engagement, and exit: Understanding communal goal processes in stem gender gaps.** *Personality and social psychology review*, *21*(2), 142–175.

Du, J., Boss, J., Han, P., Beesley, L. J., Kleinsasser, M., Goutman, S. A., Batterman, S., Feldman, E. L., & Mukherjee, B. (2022).**Variable selection with multiply-imputed datasets: Choosing between stacked and grouped methods.** *Journal of Computational and Graphical Statistics*, *31*(4), 1063–1075.

Eccles, J. S. (1983). **Expectancies, values, and academic behaviors.** In J. T. Spence (Ed.), *Achievement and achievement motives* (pp. 75–146). Freeman.

Fokkema, M. (2020).**Fitting prediction rule ensembles with R package pre.** *Journal of Statistical Software*, *92*(12), 1–30. https://doi.org/10.18637/jss.v092.i12

Fokkema, M., & Strobl, C. (2020).**Fitting prediction rule ensembles to psychological research data: An introduction and tutorial.** *Psychological Methods*, *25*(5), 636–652.

Friedman, J. H., & Popescu, B. E. (2008).**Predictive learning via rule ensembles.** *The Annals of Applied Statistics*, *2*(3), 916–954. https://doi.org/10.1214/07-AOAS148

Gottfredson, L. S. (1996). **Gottfredson's theory of circumscription and compromise.** In D. Brown & B. L. (Eds.), *Career choice and development* (3rd, pp. 179–232). Jossey-Bass.

Marsh, H. W. (1986).**Global self-esteem: Its relation to specific facets of self-concept and their importance..** *Journal of Personality and Social Psychology*, *51*(6), 1224–1236.

- Parker, P., Nagy, G., Trautwein, U., & Lüdtke, O. (2014). **Predicting career aspirations and university majors from academic ability and self-concept: A longitudinal application of the internal–external frame of reference model.** In I. Schoon & J. S. Eccles (Eds.). Cambridge University Press.
- Parker, P. D., Schoon, I., Tsai, Y.-M., Nagy, G., Trautwein, U., & Eccles, J. S. (2012).**Achievement, agency, gender, and socioeconomic background as predictors of postschool choices: A multicontext study.** *Developmental psychology*, *48*(6), 1629–1642.
- Rudin, C. (2019).**Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.** *Nature Machine Intelligence*, *1*(5), 206–215.
- Weihs, C., & Buschfeld, S. (2021).**Combining prediction and interpretation in decision trees (PrInDT) – a linguistic example.** *arXiv preprint arXiv:2103.02336*.
- Winne, P. H. (2023). **Roles for information in trace data used to model self-regulated learning.** In *Unobtrusive observations of learning in digital environments: Examining behavior, cognition, emotion, metacognition and social processes using learning analytics* (pp. 175–196). Springer.

Overview of Appendix

Appendix: Further details of the simulation design

Appendix: Plots for the effect of rounding/quantization

Appendix: Plots for the effect of sample size and missingness rate

Appendix: Overview of results

Appendix: Exemplary R Code

Appendix: Further details of the simulation design

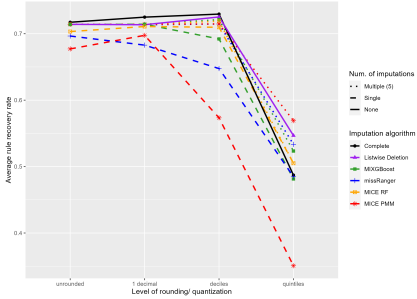
Simulation design: Dependent Variable generation

- Varying coefficients from 0.5 to 2
- Larger coefficients lead to stronger influence on target variable
- Varying cutoff values from -1 to 0.5
- Smaller cutoff values lead to more observations which fulfill the rule
- E.g. If $X_1 > 0$ & $X_2 > 0$, then Y gets increased by coefficient 1.5
- Predictive performance, recovery of the rules, coefficient bias, mean cutoff distance, false positive rate, model size
- E.g. model contains empirical rule: $X_1 > 0.1$ & $X_2 > 0.3$ with coefficient 1.4
  - Coefficient bias: 0.1
  - Mean cutoff distance: 0.2
- Mean cutoff distance is only calculated for recovered rules

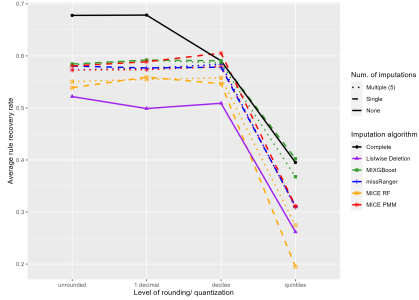
Appendix: Plots for the effect of rounding/quantization

Rule recovery

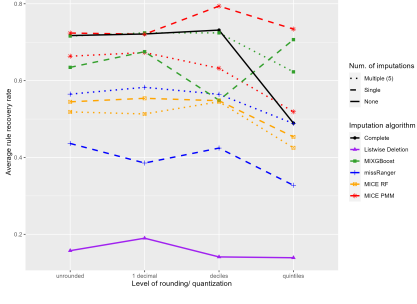
$N = 400$ & 8%



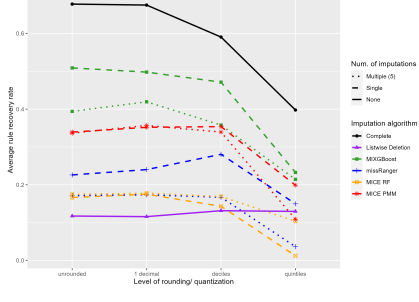
$N = 200$ & 8%



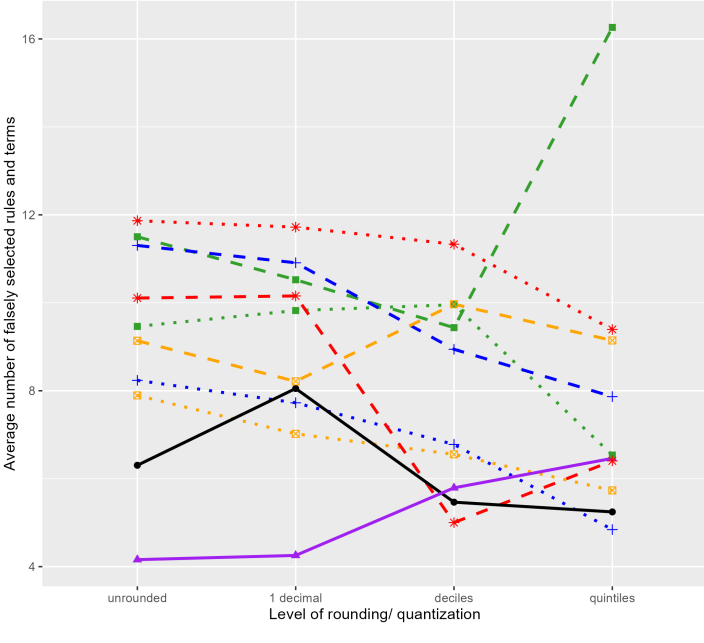
$N = 400$ & 48%



$N = 200$ & 48%



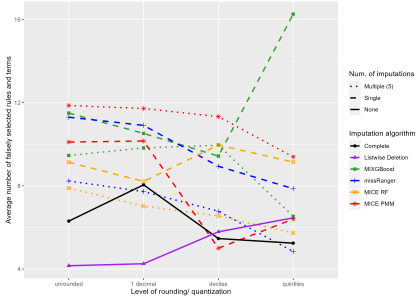
False positives ($N = 400$, 8% missingness rate)



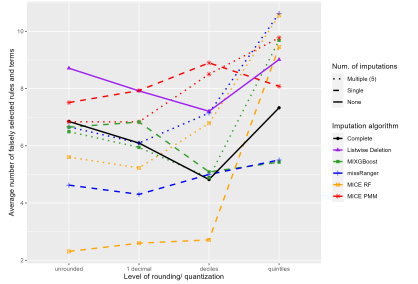
- MI performs better than SI except for MICE PMM
- MICE RF and missRanger MI closest to complete
- Listwise deletion (surprisingly) well performing (again)

False positives

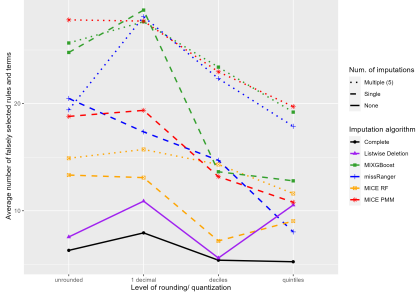
$N = 400$ & 8%



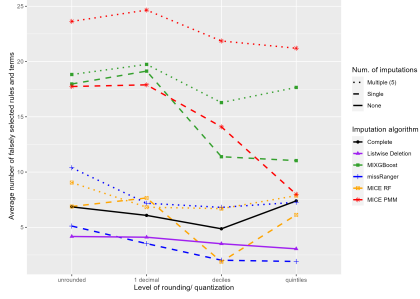
$N = 200$ & 8%



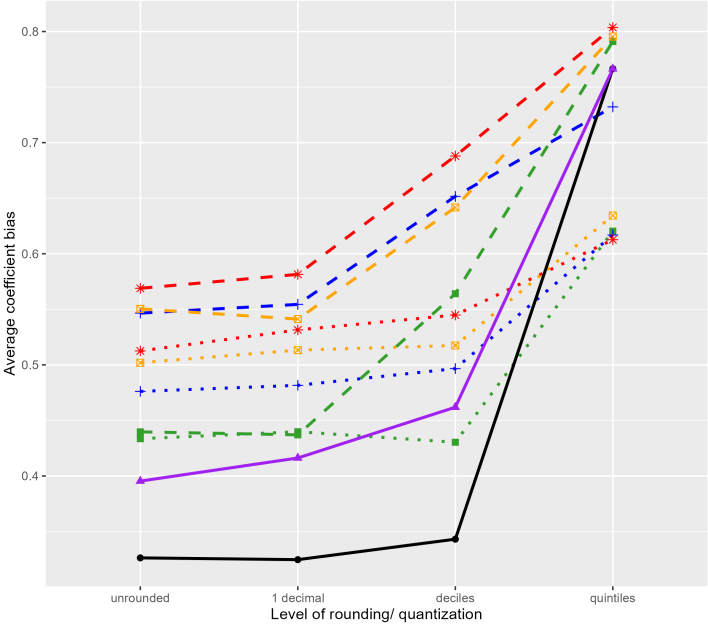
$N = 400$ & 48%



$N = 200$ & 48%



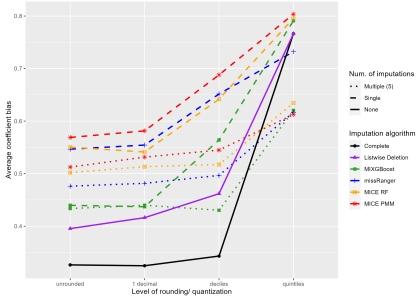
Results using $N = 400$, low missingness rate: coefficient bias



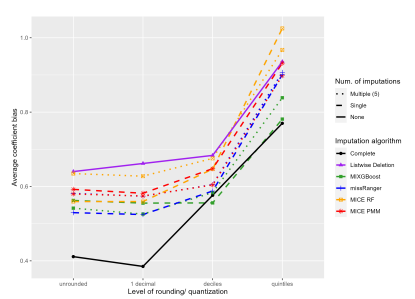
- Simplifying to quintiles decreases performance (again), but also deciles (again)
- For deciles, the loss of performance is much larger for SI than for MI
- MI performs better than SI (again)
- MIXGBoost and listwise deletion closest to complete

Bias

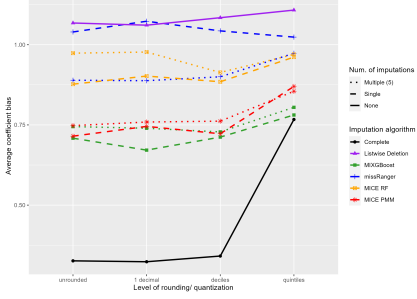
$N = 400$ & $40m$



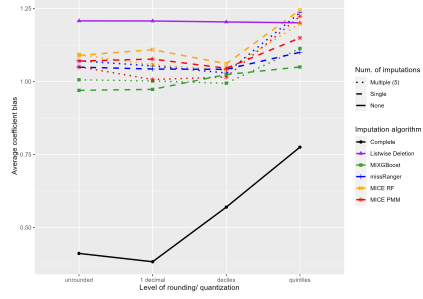
$N = 200$ & $40m$



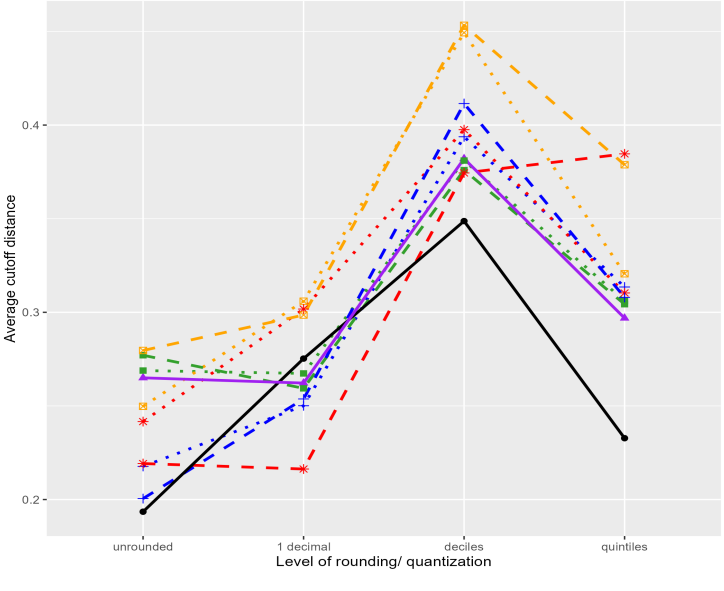
$N = 400$ & $80m$



$N = 200$ & $80m$



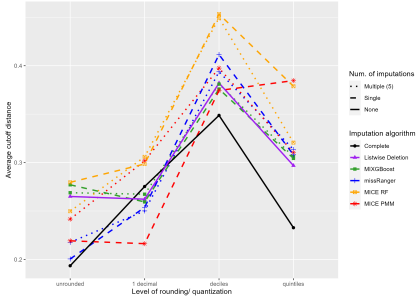
Results using $N = 400$, low missingness rate: cutoff distances



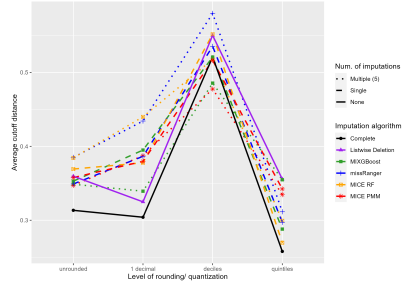
- Coarsening to deciles decreases performance (again)
- MI $\leqslant$ SI depends on imputation algorithm

Distance

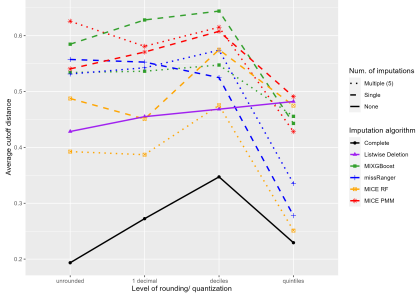
$N = 400$ & $40m$



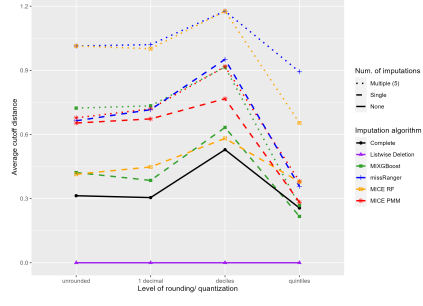
$N = 200$ & $40m$



$N = 400$ & $80m$

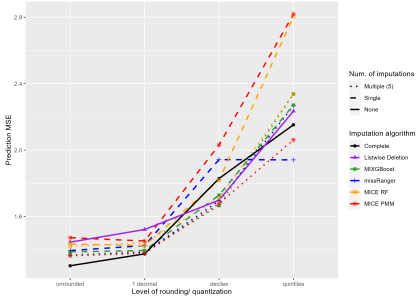


$N = 200$ & $80m$

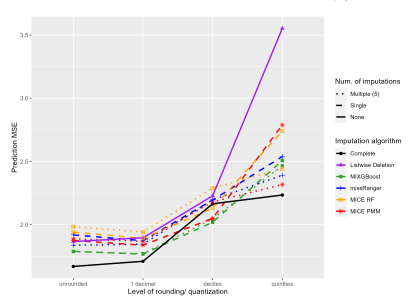


MSE

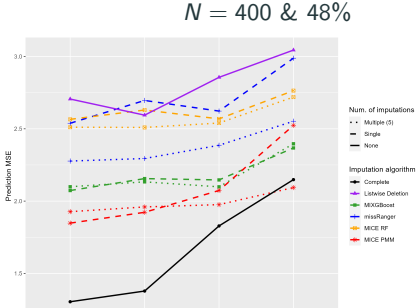
$N = 400$ & 8%



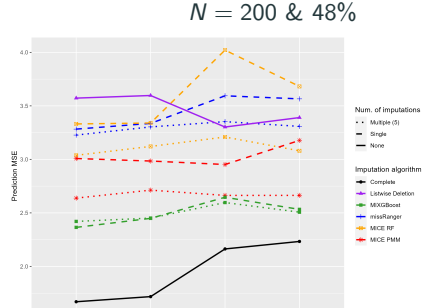
$N = 200$ & 8%



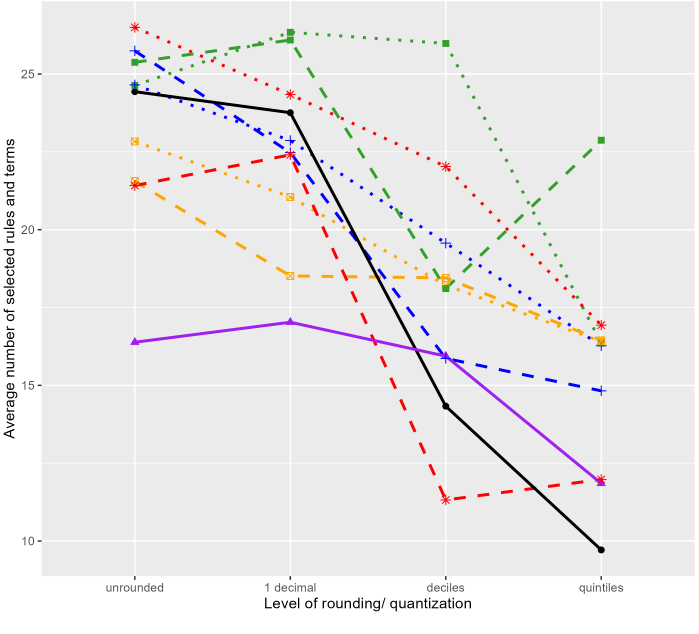
$N = 400$ & 48%



$N = 200$ & 48%



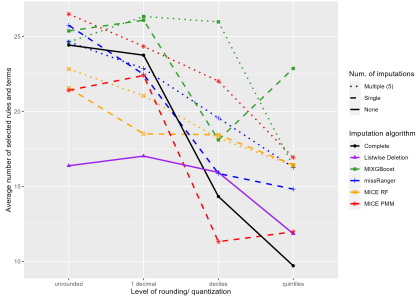
Results using $N = 400$, low missingness rate: model size



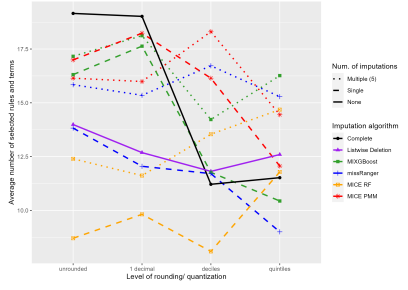
- Coarsening to deciles improves interpretability
- Single imputation leads to smaller models than multiple imputation

Size

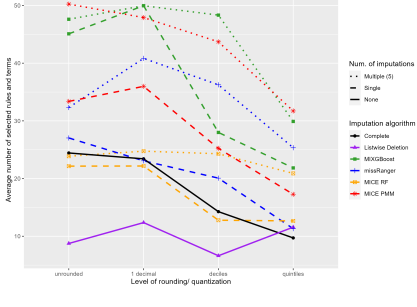
$N = 400$ & $40m$



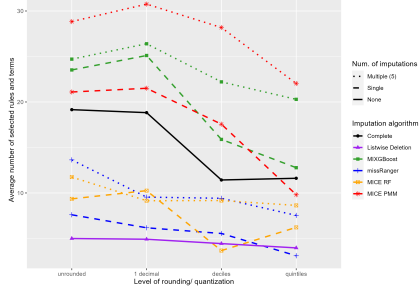
$N = 200$ & $40m$



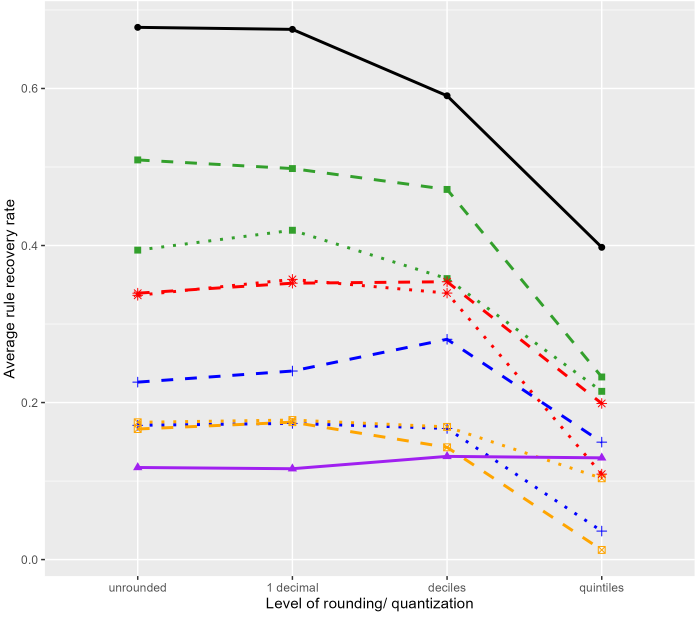
$N = 400$ & $80m$



$N = 200$ & $80m$

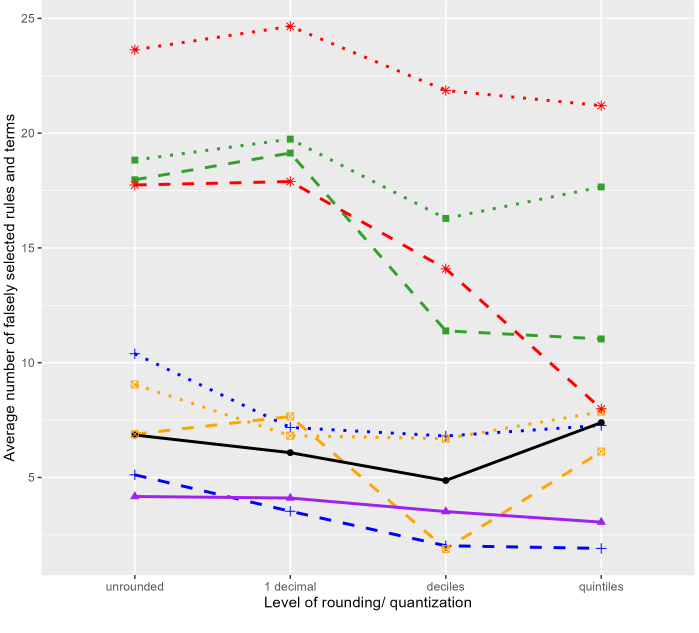


Results using $N = 200$, high missingness rate: rule recovery



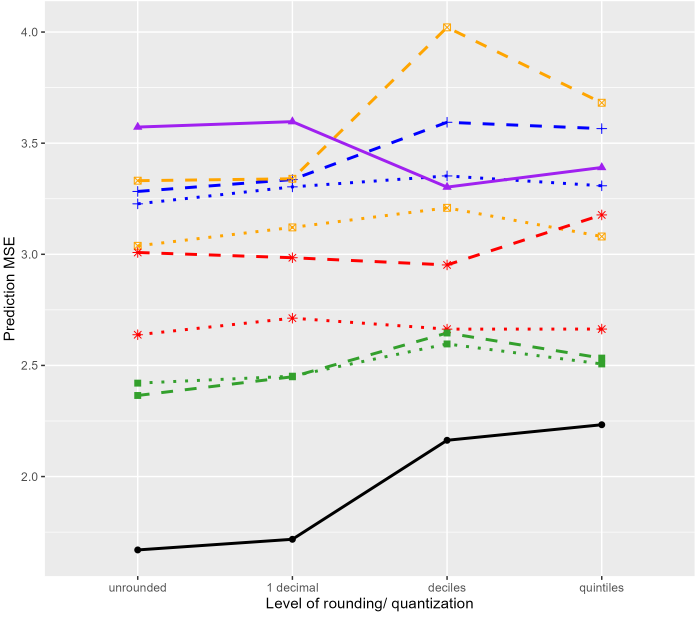
- Listwise deletion performs worst now
- MIXBoost performs better then other algorithms
- SI (surprisingly) performs better or equal then MI

Results using $N = 200$, high missingness rate: false positives



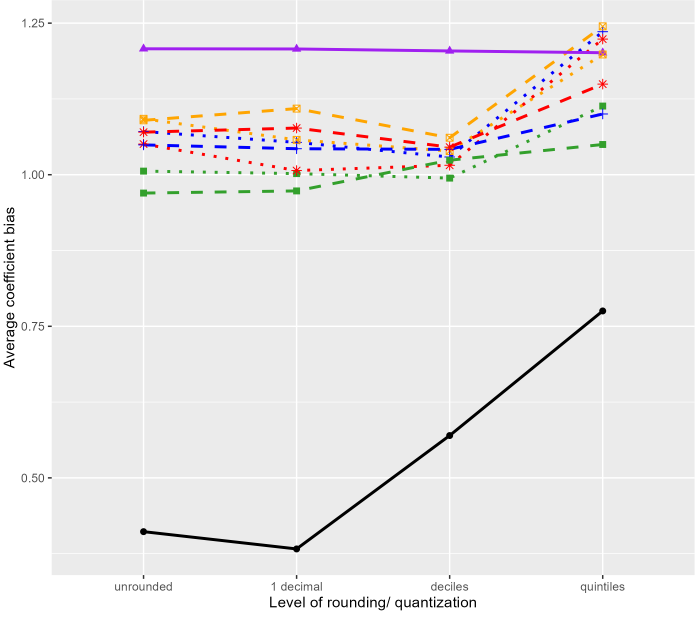
- missRanger, MICE RF, listwise deletion closest to complete
- MICE PMM and MIXGBoost lead to much higher amount of false positives
- SI performs better than MI
- Coarsing decreases false positives

Results using $N = 200$, high missingness rate: predictive performance



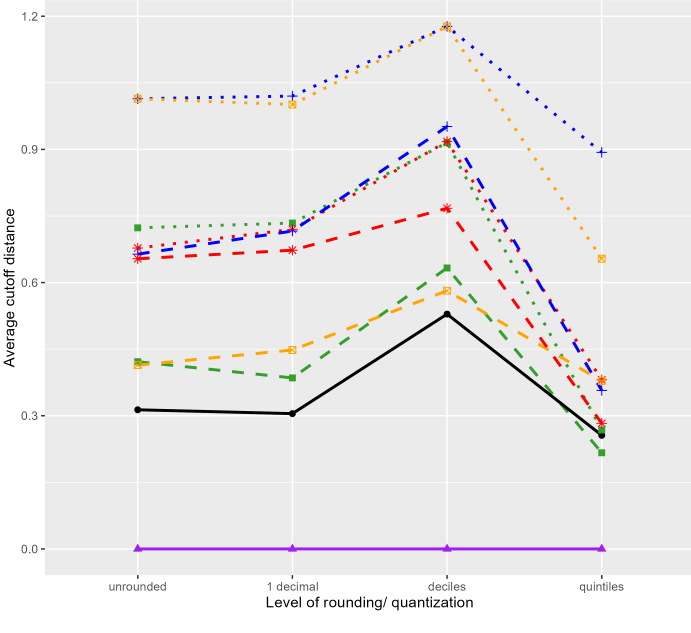
- MIXGBoost performs best
- Listwise deletion performs worst
- MI performs better than SI except for MIXGboost
- Much smaller differences between different levels of rounding

Results using $N = 200$, high missingness rate: coefficient bias



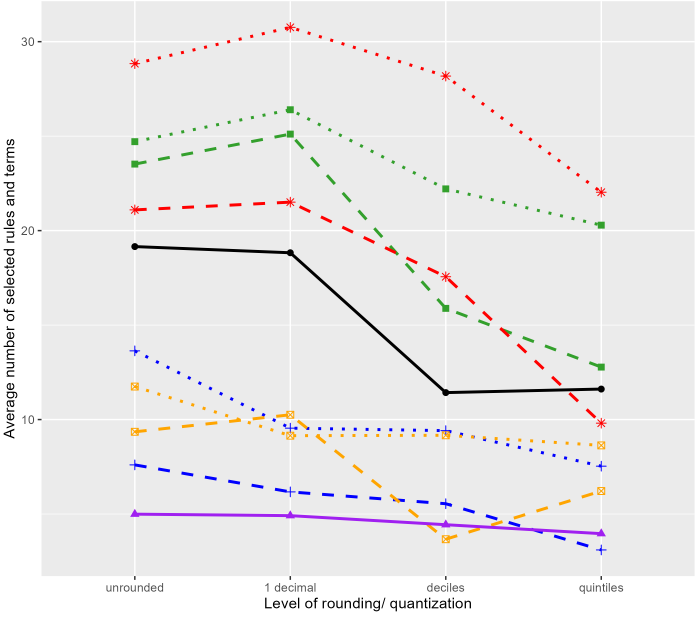
- Listwise deletion performs worst (again)
- MIXGboost performs best (again)
- MI $\lesseqgtr$ SI depends on the algorithm (again)

Results using $N = 200$, high missingness rate: cutoff distance



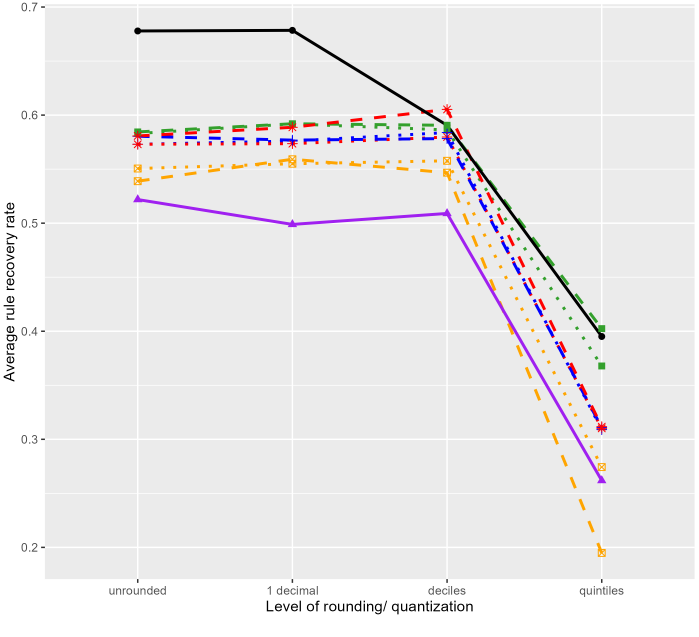
- Listwise deletion only recovered linear term, never any rules, therefore no cutoff distance

Results using $N = 200$, high missingness rate: model size



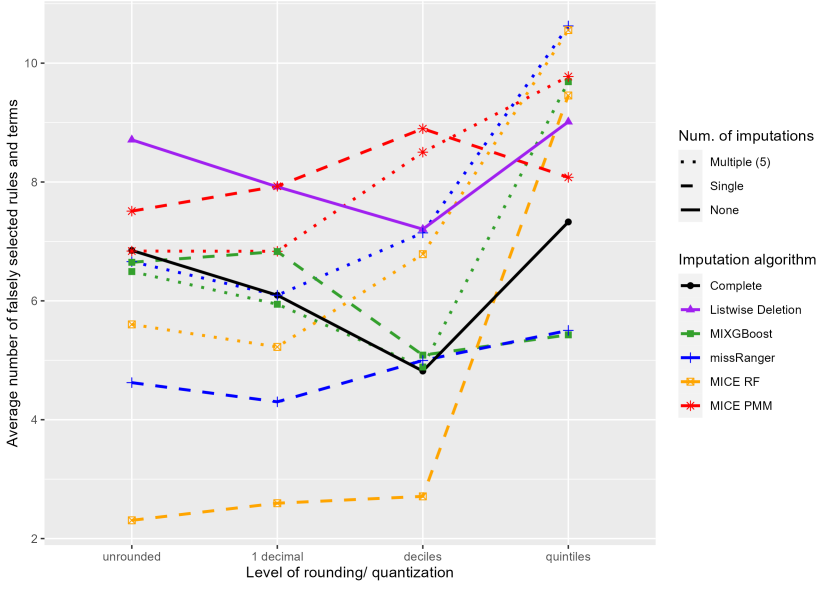
- Coarsing to deciles can decrease model size (again)
- SI leads to smaller models (again)
- MIXGBoost and MICE PMM lead to larger models than missRanger and MICE RF

Results using small sample size, low missingness rate: rule recovery



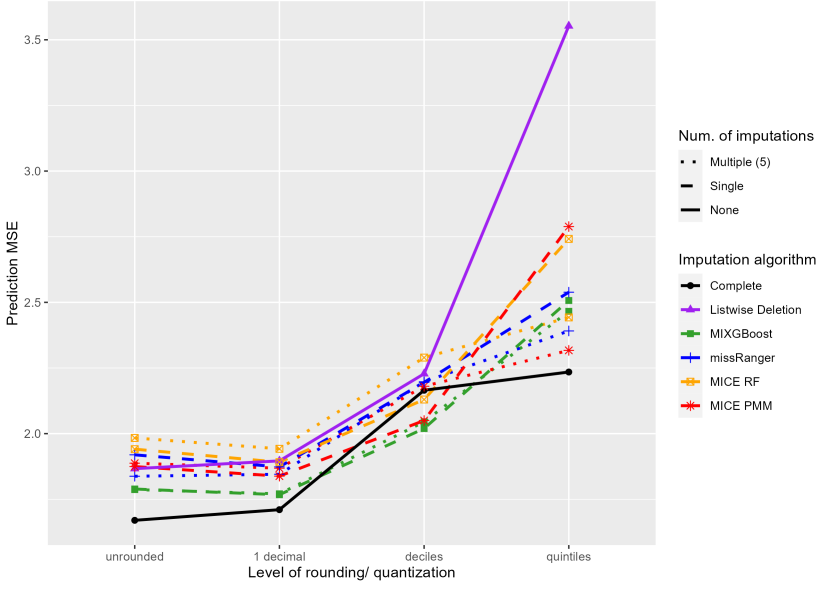
- Listwise deletion performs worse
- Coarsening to quintiles decreases performance (again)
- MICE RF performs worse than other imputation algorithms

Results using small sample size, low missingness rate: false positives



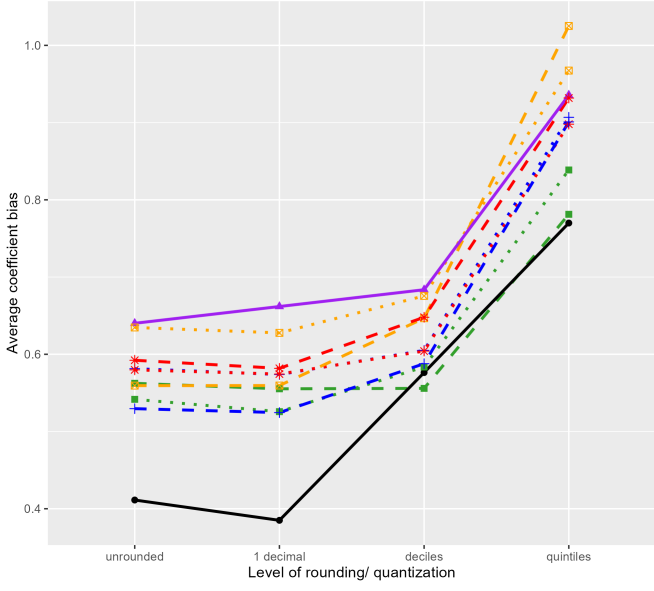
- Listwise deletion performs worse now
- MI closer to complete than SI

Results using small sample size, low missingness rate: predictive performance



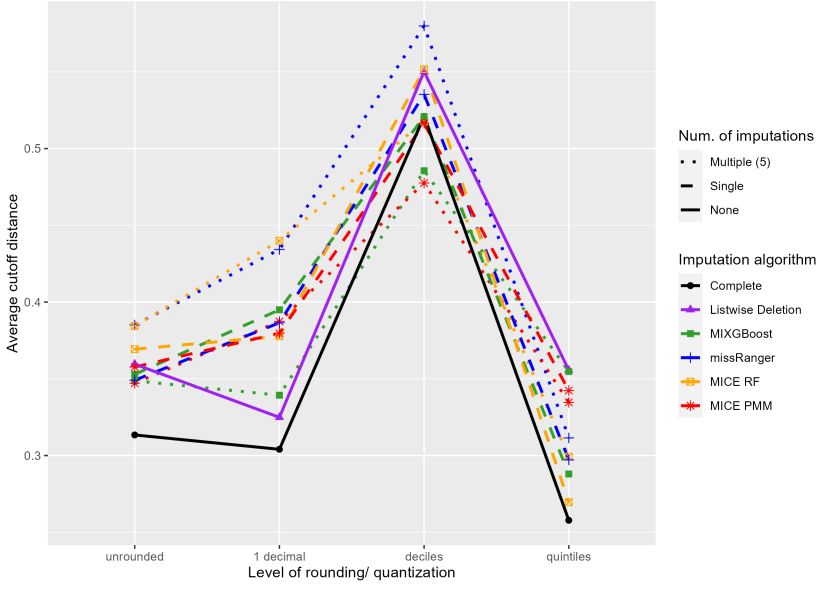
- Coarsening to quintiles or deciles decreases performance

Results using small sample size, low missingness rate: coefficient bias



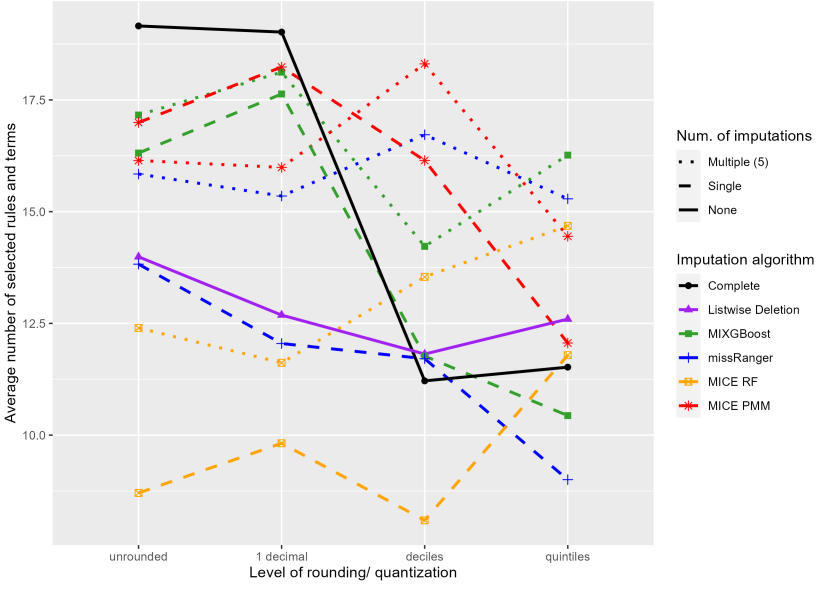
- Listwise deletion performs worst
- Coarsing to quintiles decreases performance
- MI $\lesseqgtr$ SI depends on the algorithm

Results using small sample size, low missingness rate: cutoff distance



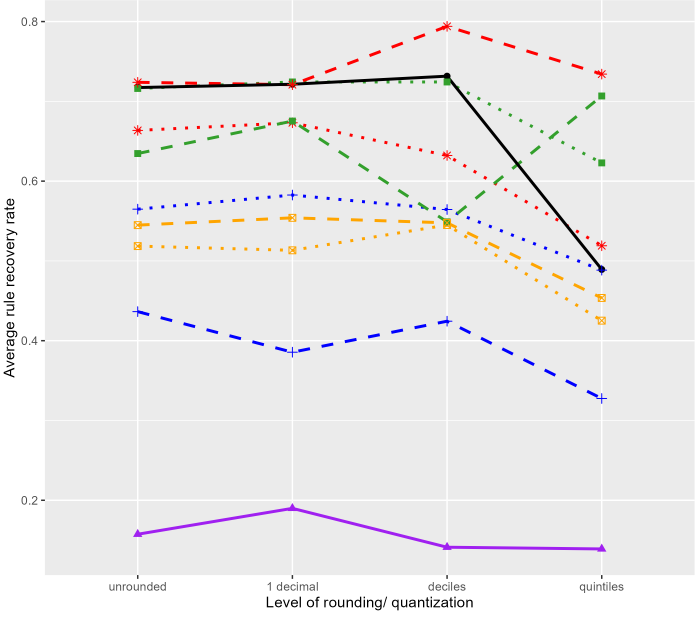
- Coarsening to deciles decreases performance (again)

Results using small sample size, low missingness rate: model size



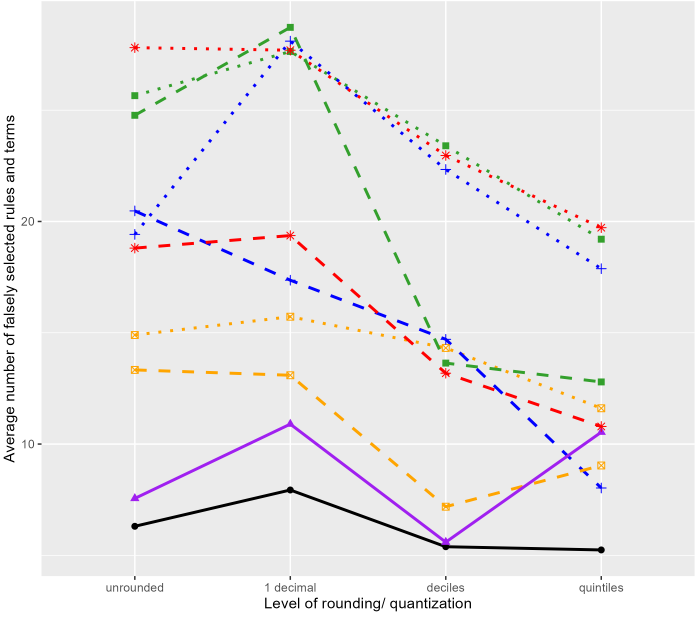
- SI leads to smaller models (again)
- MICE RF leads to the smallest models

Results using large sample size, high missingness rate: rule recovery



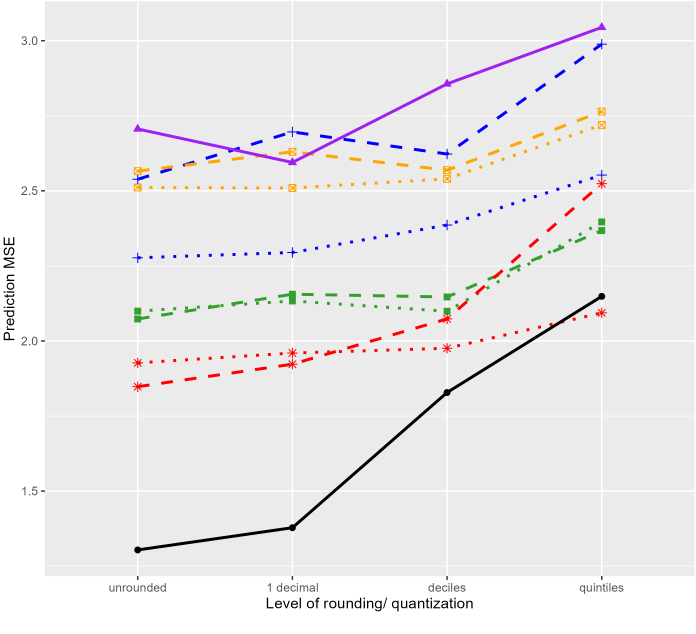
- Listwise deletion performs worst (again)
- MICE PMM and MIXBoost (MI) performs best
- MI performs better than SI except for MICE RF

Results using large sample size, high missingness rate: false positives



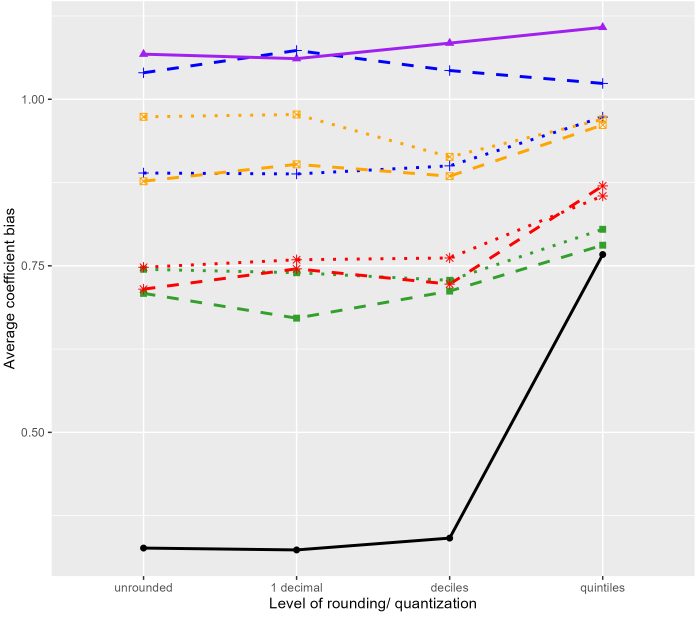
- Listwise deletion closest to complete
- MICE RF performs better than other imputation algorithms
- Coarsing to deciles can improve performance
- SI performs better than MI except for missRanger

Results using large sample size, high missingness rate: predictive performance



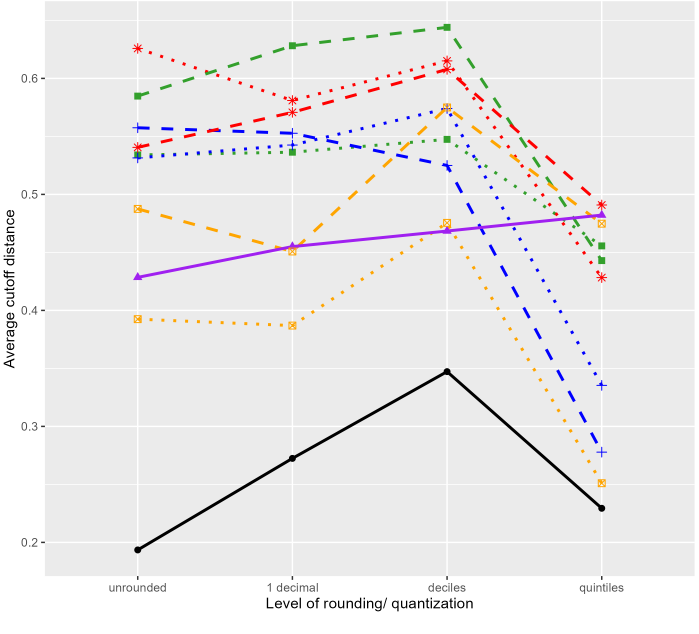
- Listwise deletion performs worst (again)
- MICE PMM performs best
- MI $\lesseqgtr$ SI depends on the algorithm

Results using large sample size, high missingness rate: coefficient bias



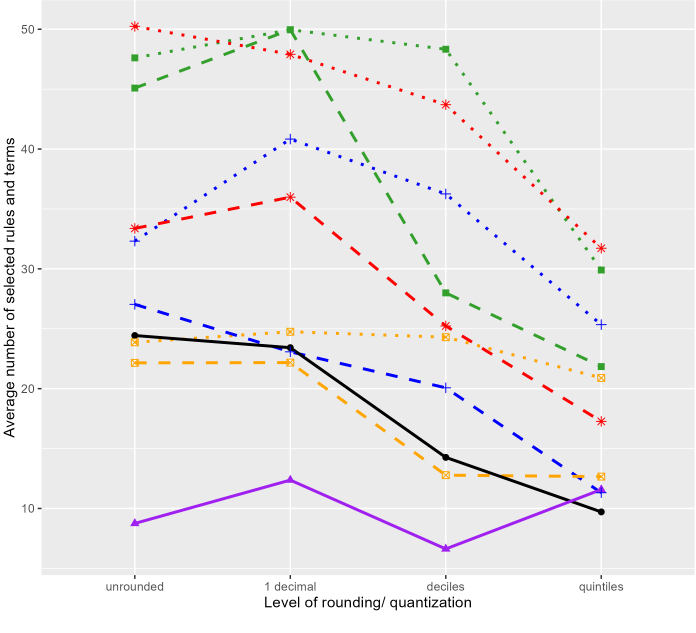
- Listwise deletion performs worst (again)
- MICE PMM and MIXGBoost (SI) performs best
- SI performs better then MI except for missRanger

Results using large sample size, high missingness rate: cutoff distance



- Coarsing to quintiles decreases cutoff distances
- MICE RF (MI) performs best

Results using large sample size, high missingness rate: model size

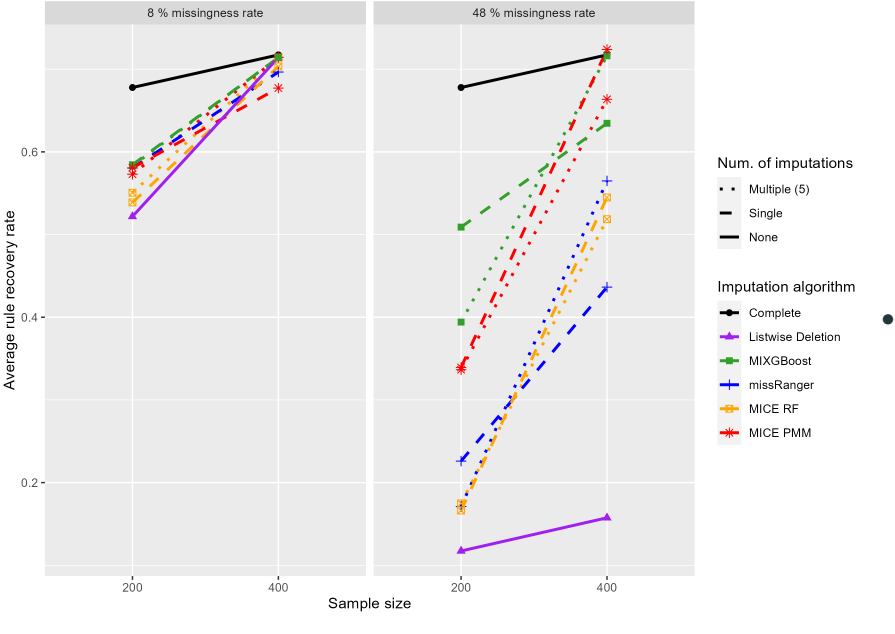


- Listwise deletion leads to the smallest models
- MICE RF (MI) closest to complete

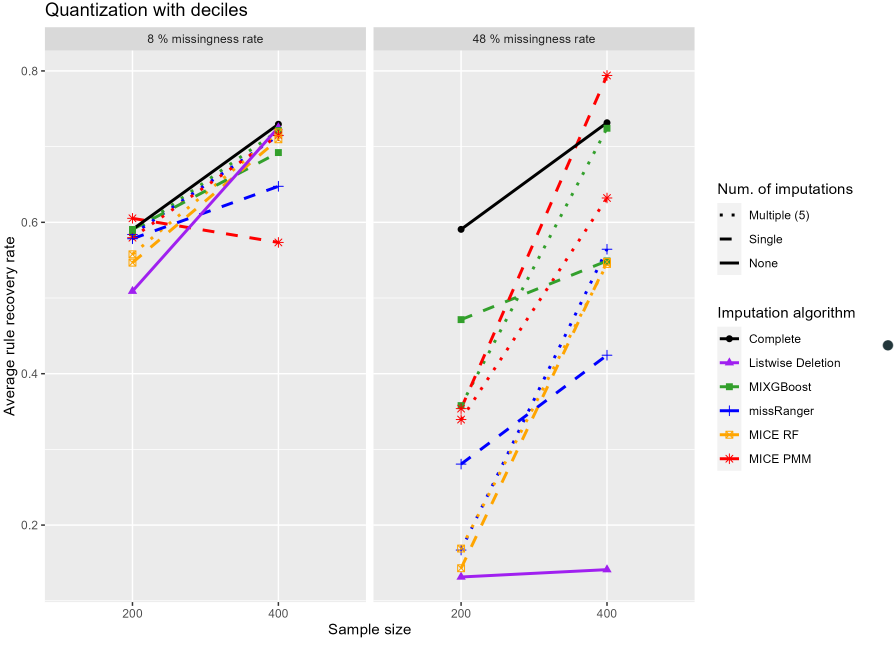
Appendix: Plots for the effect of sample size and missingness rate

Results

Unrounded

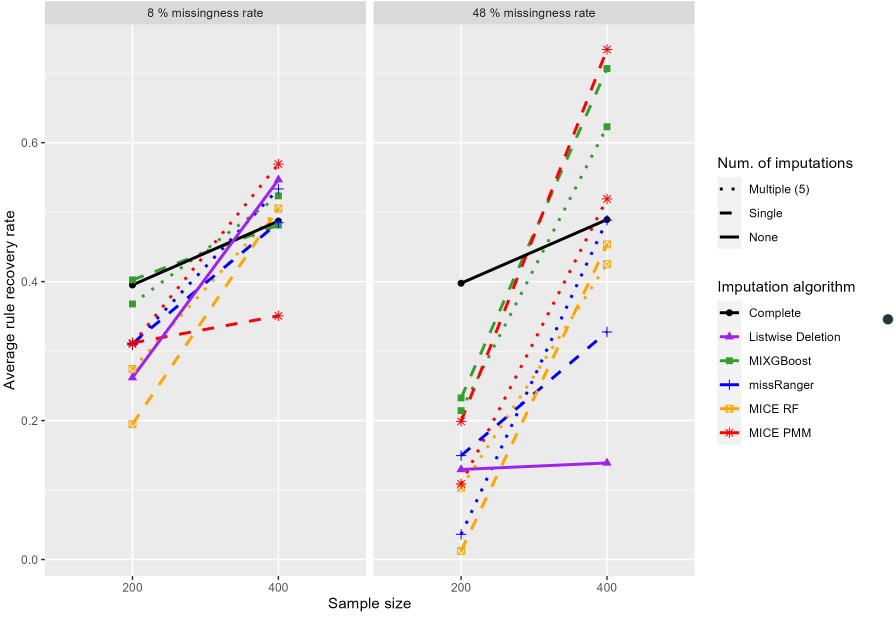


Results


Quantization with deciles
8 % missingness rate
48 % missingness rate
Average rule recovery rate
0.8
0.6
0.4
0.2
200
400
Sample size
Num. of imputations
Multiple (5)
Single
None
Imputation algorithm
Complete
Listwise Deletion
MIXGBoost
missRanger
MICE RF
MICE PMM

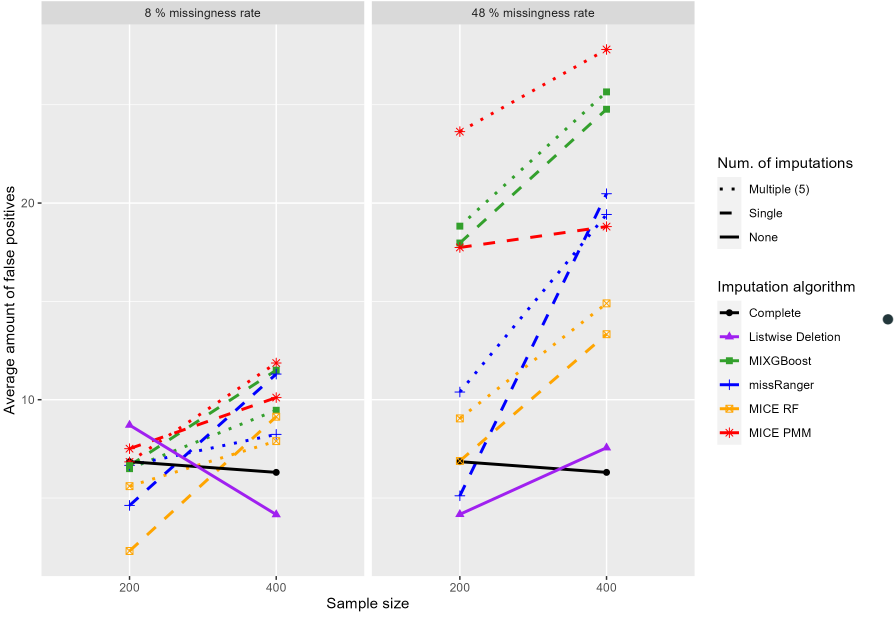
Results

Quantization with quintiles



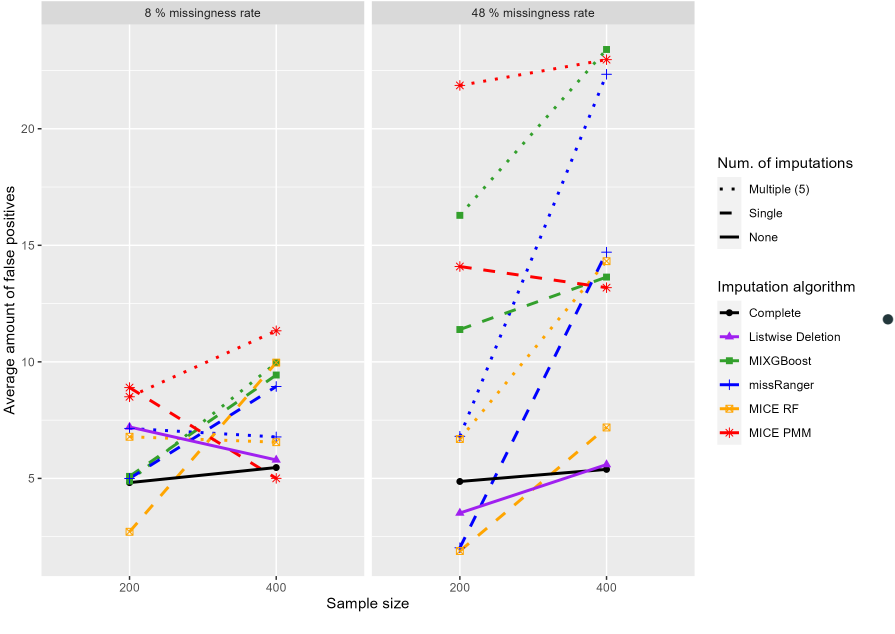
Results

Unrounded



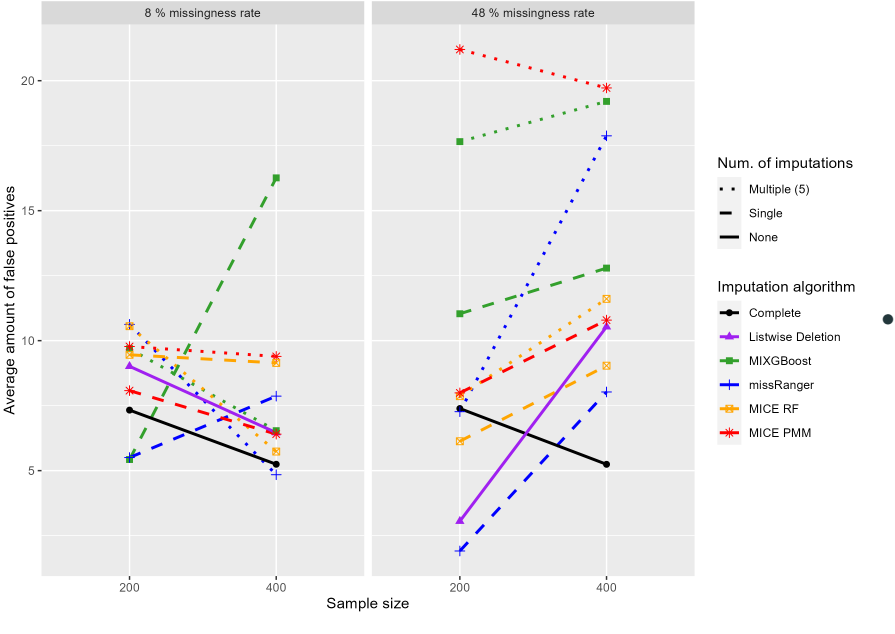
Results

Quantization with deciles



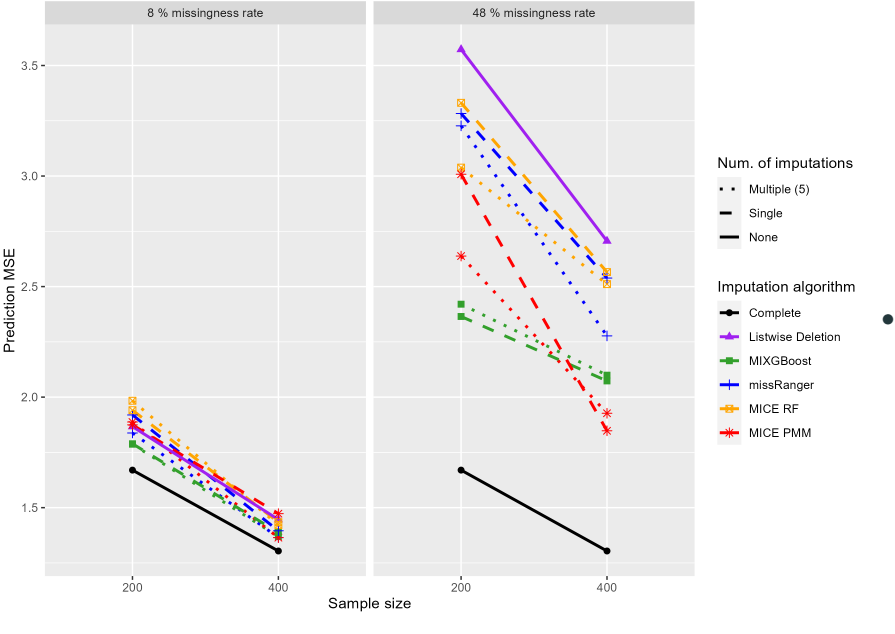
Results

Quantization with quintiles



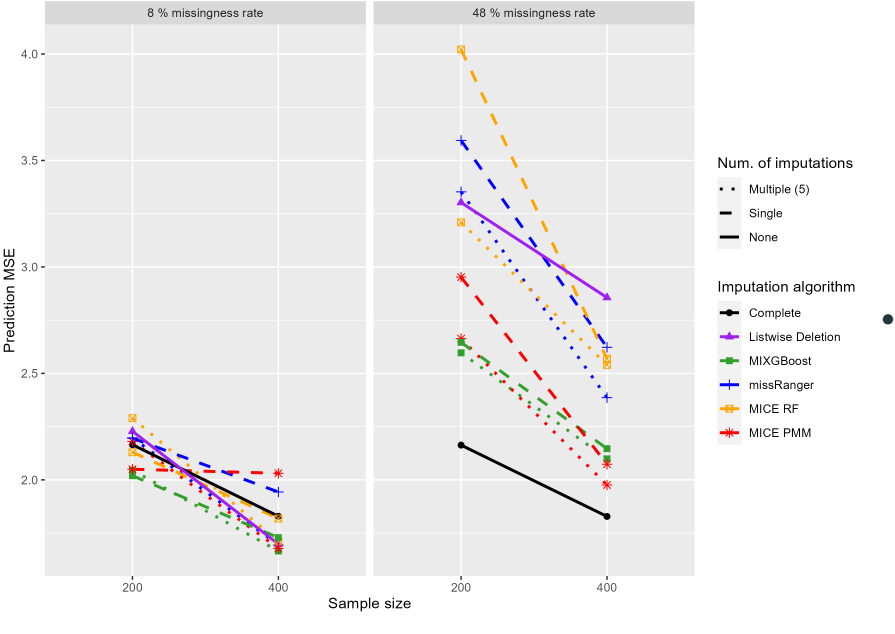
Results

Unrounded



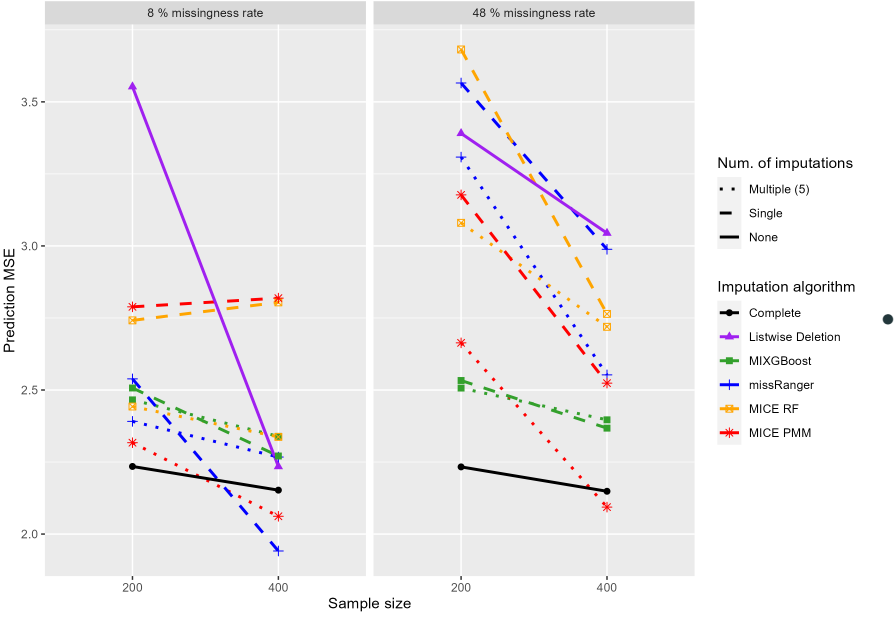
Results

Quantization with deciles



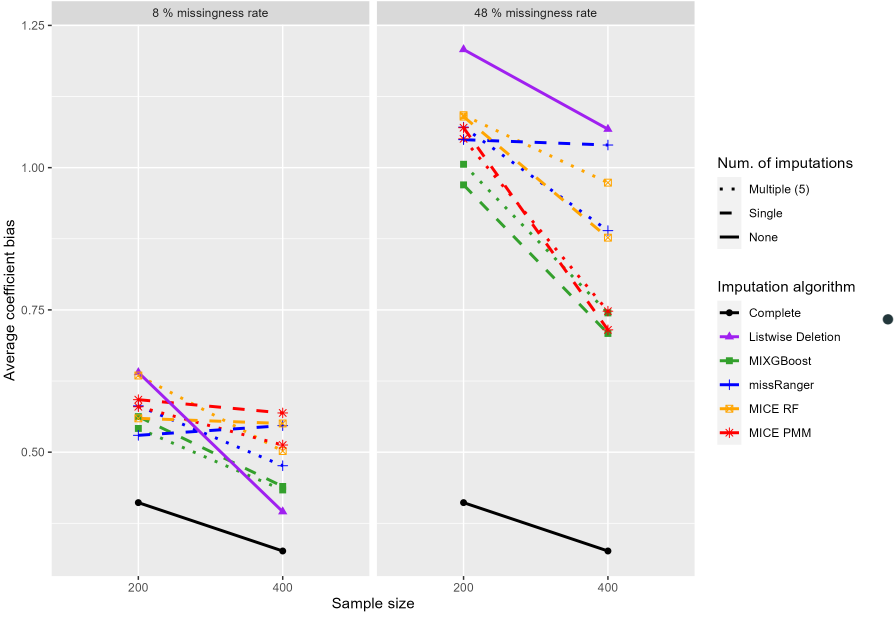
Results

Quantization with quintiles

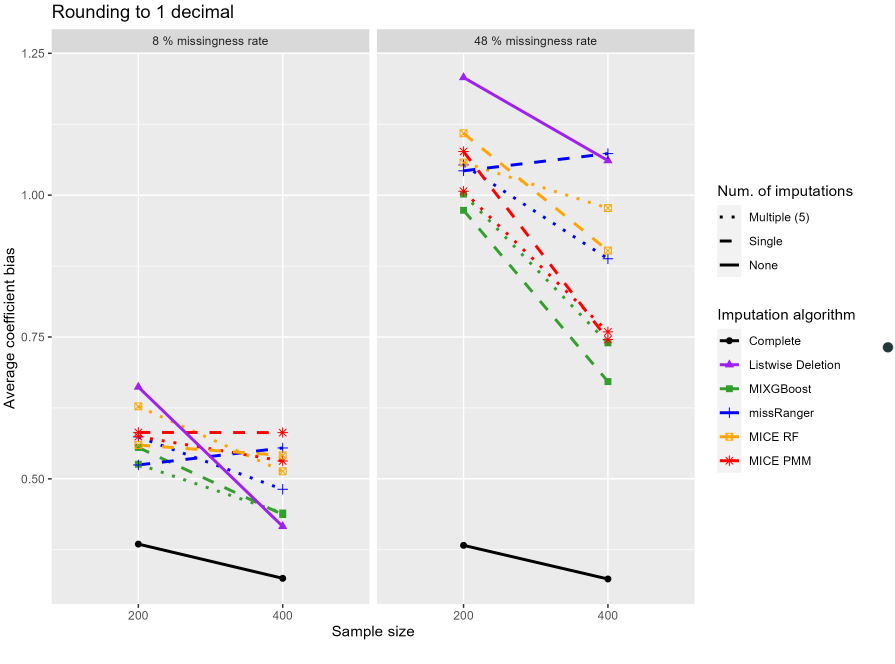


Results

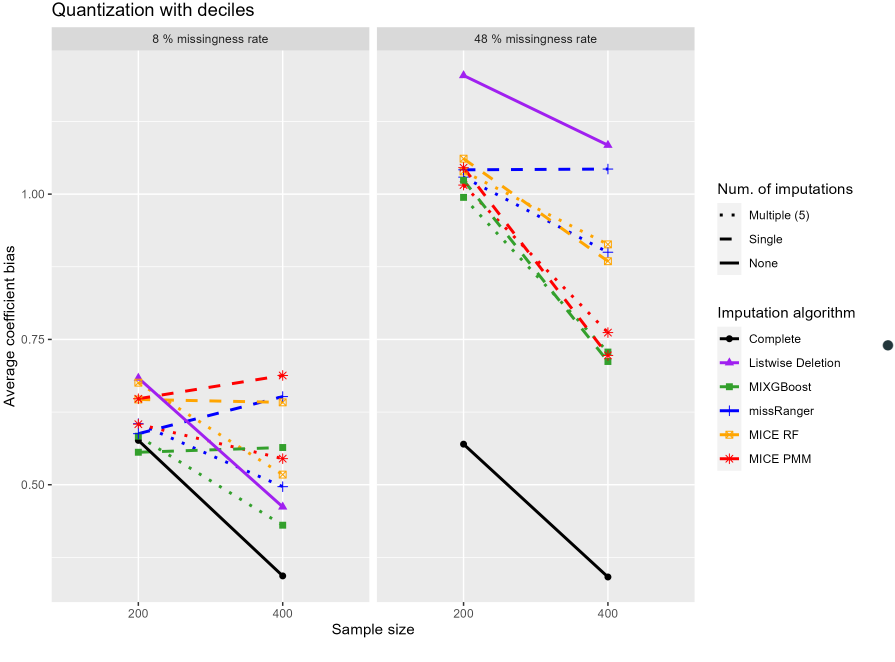
Unrounded



Results

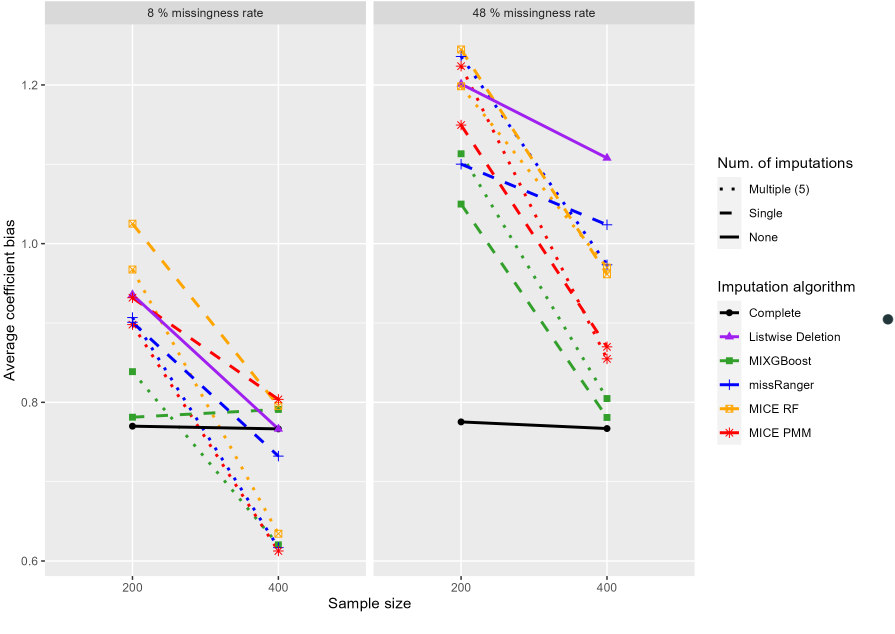

Rounding to 1 decimal
8 % missingness rate
48 % missingness rate
Average coefficient bias
1.25
1.00
0.75
0.50
200
400
Sample size
Num. of imputations
Multiple (5)
Single
None
Imputation algorithm
Complete
Listwise Deletion
MIXGBoost
missRanger
MICE RF
MICE PMM

Results


Quantization with deciles
8 % missingness rate
48 % missingness rate
Average coefficient bias
1.00
0.75
0.50
200
400
200
400
Sample size
Num. of imputations
Multiple (5)
Single
None
Imputation algorithm
Complete
Listwise Deletion
MIXGBoost
missRanger
MICE RF
MICE PMM

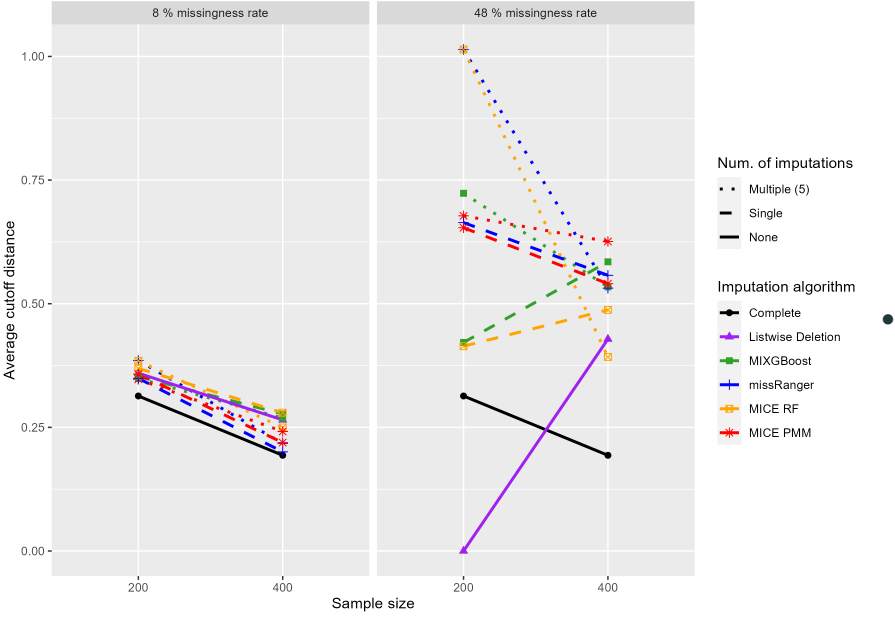
Results

Quantization with quintiles



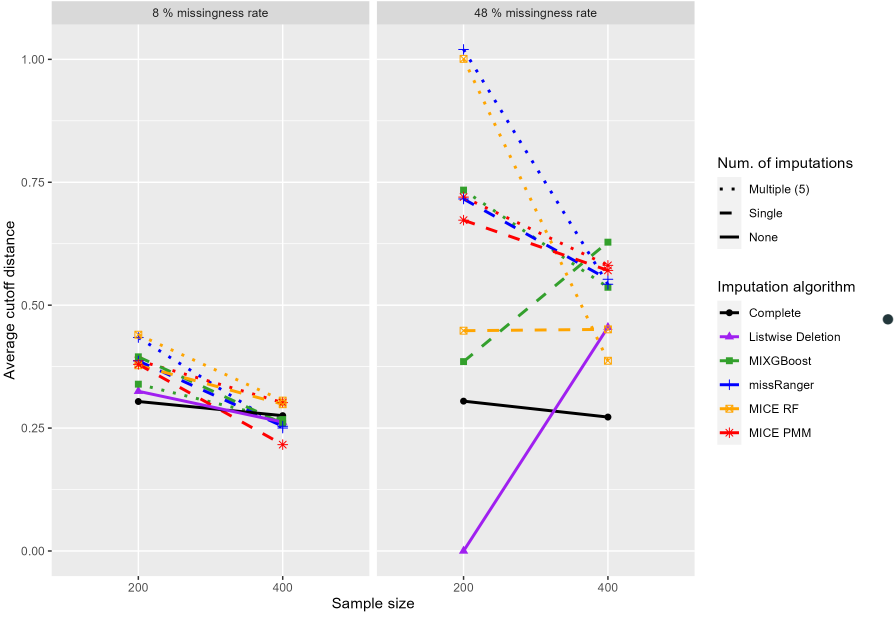
Results

Unrounded



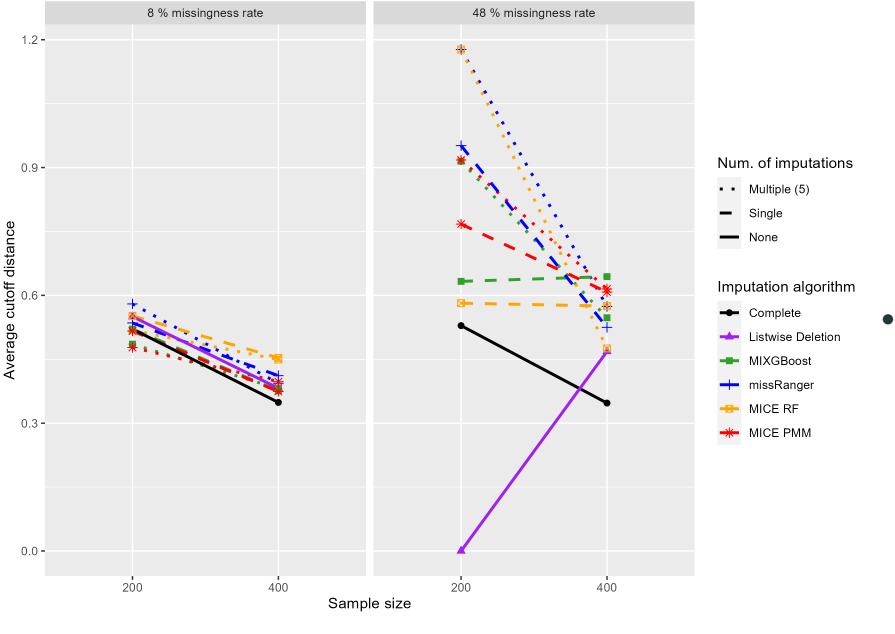
Results

Rounding to 1 decimal

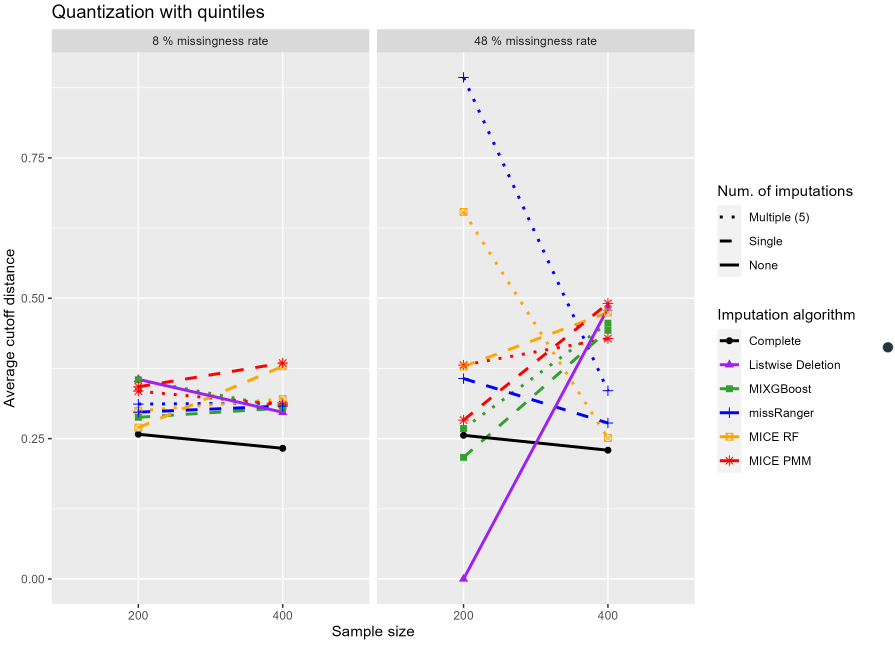


Results

Quantization with deciles

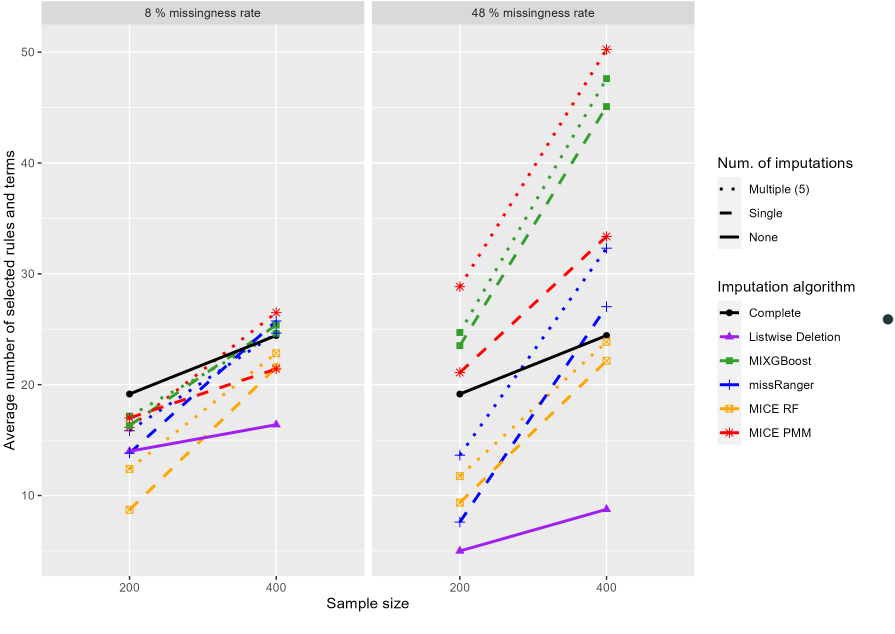


Results

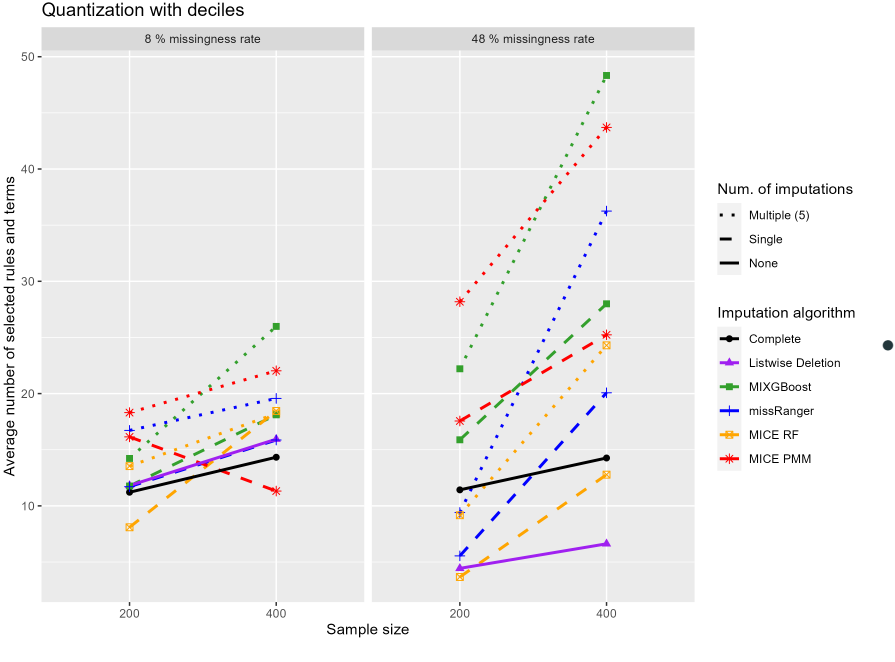

Quantization with quintiles
8 % missingness rate
48 % missingness rate
Average cutoff distance
0.00
0.25
0.50
0.75
200
400
Sample size
Num. of imputations
Multiple (5)
Single
None
Imputation algorithm
Complete
Listwise Deletion
MIXGBoost
missRanger
MICE RF
MICE PMM

Results

Unrounded

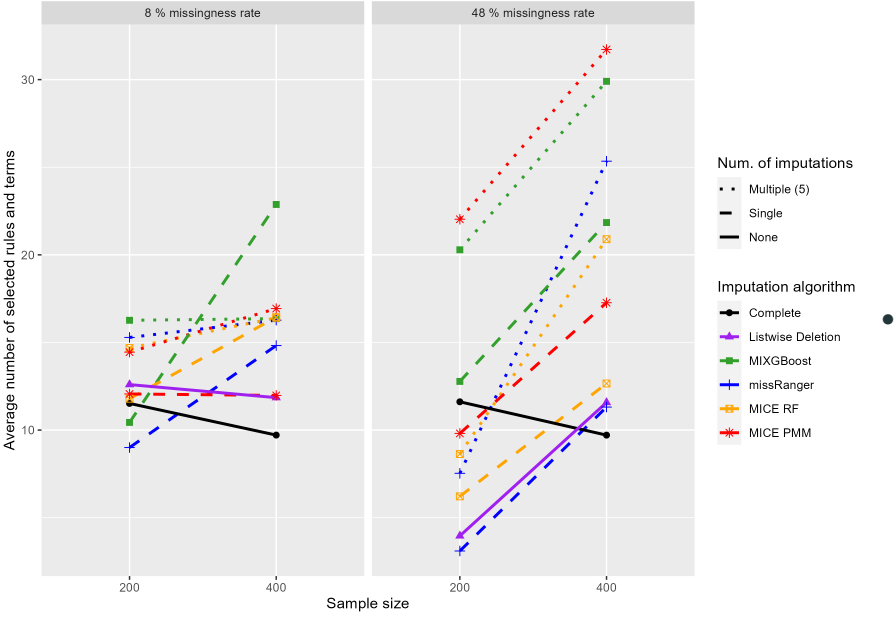


Results


Quantization with deciles
8 % missingness rate
48 % missingness rate
Average number of selected rules and terms
50
40
30
20
10
200
400
Sample size
Num. of imputations
Multiple (5)
Single
None
Imputation algorithm
Complete
Listwise Deletion
MIXGBoost
missRanger
MICE RF
MICE PMM

Results

Quantization with quintiles



Appendix: Overview of results

Summary: Other data settings

Metric	Small sample, low missing rate	Large sample, high missing rate
RR	LD worst, MIXBoost best, MI $\leqslant$ SI	LD worst, MICE PMM SI best, MI $\leqslant$ SI
FP	MI closest to complete	LD closest to complete
MSE	MIXBoost best, MI $\leqslant$ SI	MICE PMM best
Bias	missRanger SI best, MI $\leqslant$ SI	SI better than MI
CutoffD.	Decile coarsening worst	MICE RF performs best
Size	MICE RF smallest, complete largest	MICE RF closest to complete

RR: Rule recovery, FP: False positives, MSE: Predictive performance, Bias: Coefficient bias, CutoffD.: Cutoff distance, Size: Model Size

Appendix: Exemplary R Code

Exemplary R code I

iris_WithNA = cbind(mice::ampute(iris[, 1:4], prop = 0.4)$amp, iris$Species)

- MIXGB
  iris_imputed = mixgb(iris_WithNA, m = 5, pmm.k = 5)
- missRanger
  iris_imputed = replicate(5, missRanger(iris_WithNA, pmm.k = 5), simplify = FALSE)
- MICE RF
  imputed_data = mice(iris_WithNA, m = 5, method = 'rf')
  iris_imputed = lapply(1:5, function(i) complete(imputed_data, action = i))
- MICE PMM
  imputed_data = mice(iris_WithNA, m = 5, method = 'pmm')
  iris_imputed = lapply(1:5, function(i) complete(imputed_data, action = i))

Exemplary R code II

```
model_imputed = mi_pre(Sepal.Length ~ . , data = iris_imputed, family = 'gaussian',
maxdepth = 2)

model_listwiseDeletion = pre(Sepal.Length ~ . , data = iris_WithNA, family =
'gaussian', maxdepth = 2)

print(model_imputed)
print(model_listwiseDeletion)
importance(model_imputed)
importance(model_listwiseDeletion)
```