Missing the Point: Non-Convergence in Iterative Imputation Algorithms

Anonymous Authors[1]

Abstract

Iterative imputation is a popular tool to accommodate missing data. While it is widely accepted that valid inferences can be obtained with this technique, these inferences all rely on algorithmic convergence. There is no consensus on how to evaluate the convergence properties of the method. This paper provides insight into identifying non-convergence in iterative imputation algorithms. Our study found that—in the cases considered—inferential validity was achieved after five to ten iterations, much earlier than indicated by diagnostic methods. We conclude that it never hurts to iterate longer, but such calculations hardly bring added value.

1. Introduction

A popular method to accommodate missing data is to impute (i.e., 'fill in') the missing values in an incomplete dataset. It is widely accepted that imputation techniques such as multiple imputation (MI; Rubin, 1976) can yield statistically valid inferences. The validity of these inferences relies on the method through which imputations are obtained—often iterative algorithms. Iterative imputation algorithms are employed in e.g., `SPSS`, `Stata`, and the `R` packages `MI`, and `mice`. [Or use: Convergence of the algorithm used to solve the missing data problem is a topic that has not received much attention but is ever so important, as most imputation software packages draw inference from iterative imputation procedures.]

With iterative imputation, the validity of the inference depends on the state-space of the algorithm at the final iteration. This introduces a potential threat to the validity of the imputations: What if the algorithm has not converged? Are the imputations then to be trusted? And can we rely on the inference obtained on the completed data?

These remain open questions since the convergence properties of iterative imputation algorithms have not been systematically studied (Van Buuren, 2018).[1] There is no scientific consensus on how to evaluate the convergence of imputation algorithms (Zhu & Raghunathan, 2015; Takahashi, 2017). Moreover, the behavior of such algorithms under certain default imputation models (e.g., 'predictive mean matching') is an entirely open question (Murray, 2018). Therefore, algorithmic convergence should be monitored carefully.

The current practice of visually inspecting imputations may be undesirable for several reasons: 1) it may be challenging to the untrained eye, 2) only severely pathological cases of non-convergence may be diagnosed, and 3) there is not an objective measure that quantifies convergence (Van Buuren, 2018, § 6.5.2). On top of that, the recommended parameters to inspect may be insufficient. The parameters are scalar summaries of the state-space of the algorithm. They are either univariate, or depend on the substantive model of interest. Apparent convergence of a univariate parameter does not guarantee convergence of the multivariate state-space of the algorithm. And with a model-dependent statistic there is no guarantee that the algorithm is converged enough for other models of scientific interest. Additionally, model-dependency negates an important advantage of MI: solving the missing data problem separately from the scientific problem.

Therefore, we propose a novel parameter to inspect. We use this parameter to identify non-convergence with two quantitative diagnostic methods. We use model-based simulation to investigate the plausibility of identifiers for non-convergence for iterative imputation algorithms and provide guidelines for empirical research. For reasons of brevity, we only focus on the iterative imputation algorithm implemented in the

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]If the algorithm does not reach convergence, "the imputed datasets will not be truly independent and the variability among them may understate the true levels of missing-data uncertainty" (Schafer and Olsen, p. 556).

popular mice package in R (Van Buuren & Groothuis-Oudshoorn, 2011; R Core Team, 2020). [Provide notation here? E.g., the number of imputations, iterations, etc.]

2. Methods

2.1. Diagnostic Methods

We consider several sets of diagnostic methods to identify non-convergence in the iterative imputation algorithm. Each method is a combination of a parameter that summarizes the state-space of the algorithm, θ, and an identifier. As θs we use the default parameters of the MICE algorithm (i.e., the univariate θs chain means and chain variances), a model-dependent user-specified multivariate parameter Q, and a novel parameter: λ_1. λ_1 has the appealing property that it is not dependent on the model of scientific interest, yet still summarizes the multivariate state-space of the algorithm.

We define λ_1 as the first eigenvalue of the variance-covariance matrix of the completed data.[2] For each imputation ℓ, let $\lambda_{1,\ell} \geq \lambda_{2,\ell} \geq ... \geq \lambda_{j,\ell}$ be the eigenvalues of the variance-covariance matrix S_ℓ, where $\ell = 1, ..., m$. Since the eigenvalues of a variance-covariance matrix summarize the total set of covariances in the data, λ_1 captures the multivariate nature of the completed data without imposing a substantive model of interest.

The identifiers are as recommended by e.g. (Cowles & Carlin, 1996), namely the potential scale reduction factor $\widehat{R}$ and autocorrelation. $\widehat{R}$ is also known as the 'Gelman-Rubin statistic' after the original paper (Gelman & Rubin, 1992). We use a recently proposed adapted version of $\widehat{R}$ (Vehtari et al., 2019). We diagnose non-convergence if $\widehat{R}$-values exceed a certain diagnostic threshold. The thresholds under consideration are the traditional threshold $\widehat{R} = 1.2$ (Gelman & Rubin, 1992), the common $\widehat{R} = 1.1$ (Gelman et al., 2013), and the recent $\widehat{R} = 1.01$ (Vehtari et al., 2019).

We diagnose non-convergence with autocorrelation when the correlation between two subsequent θ-values within the same chain is significantly different from zero (Lynch, 2007; Hyndman & Athanasopoulos, 2018, p. 147, p. 42).

[2] By definition, this is equal to the variance of the first component of the principal component analylis (PCA) solution on the same data.

2.2. Simulation Study

We investigate non-convergence of the iterative imputation algorithm implemented in the `R` package `mice` through model-based simulation. The simulation set-up is summarized in pseudo-code (see Algorithm 1). The complete script and technical details are available from github.com/hanneoberman/MissingThePoint. [Define conditions here somewhere, and $p_{\text{inc}}$and T.]

Algorithm 1 Simulation set-up

```
Simulate data
repeat
  for p_inc = .05, .25, .50, .75, .95 (missingness conditions) do
    Create missingness
    for T = 1, 2, ..., 100 (early stopping conditions) do
      Impute missingness
      Perform analysis of scientific interest
      Compute non-convergence diagnostics
      Pool results across imputations
      Compute performance measures
    end for
  end for
  Combine outcomes of all conditions
until n_sim = 1000 (simulation repetitions)
Aggregate outcomes across simulation runs
```

The aim of the simulation study is twofold: 1) to determine the effect of early stopping and missingness conditions on the validity of statistical inferences, 2) to investigate the plausibility of identifiers for non-convergence.

We assess the validity of statistical inferences with bias, coverage rate and confidence interval width as performance measures (see @buur18 for definitions). And we evaluate the plausibility of the identifiers using the diagnostic thresholds for $\widehat{R}$ and autocorrelation (i.e., $\widehat{R} = 1.2, 1.1, 1.01$ and the critical value for AC = 0).

[Very incomplete from here!] We consider several quantities of scientific interest.

The statistical inference of interest is multivariate linear regression.

The data generating mechanism of the predictor space is a multivariate normal distribution

3. Results

Results are displayed in Figure 1 and 2. Simulation conditions with a high proportion of missing cases and a low number of iterations most often result in biased

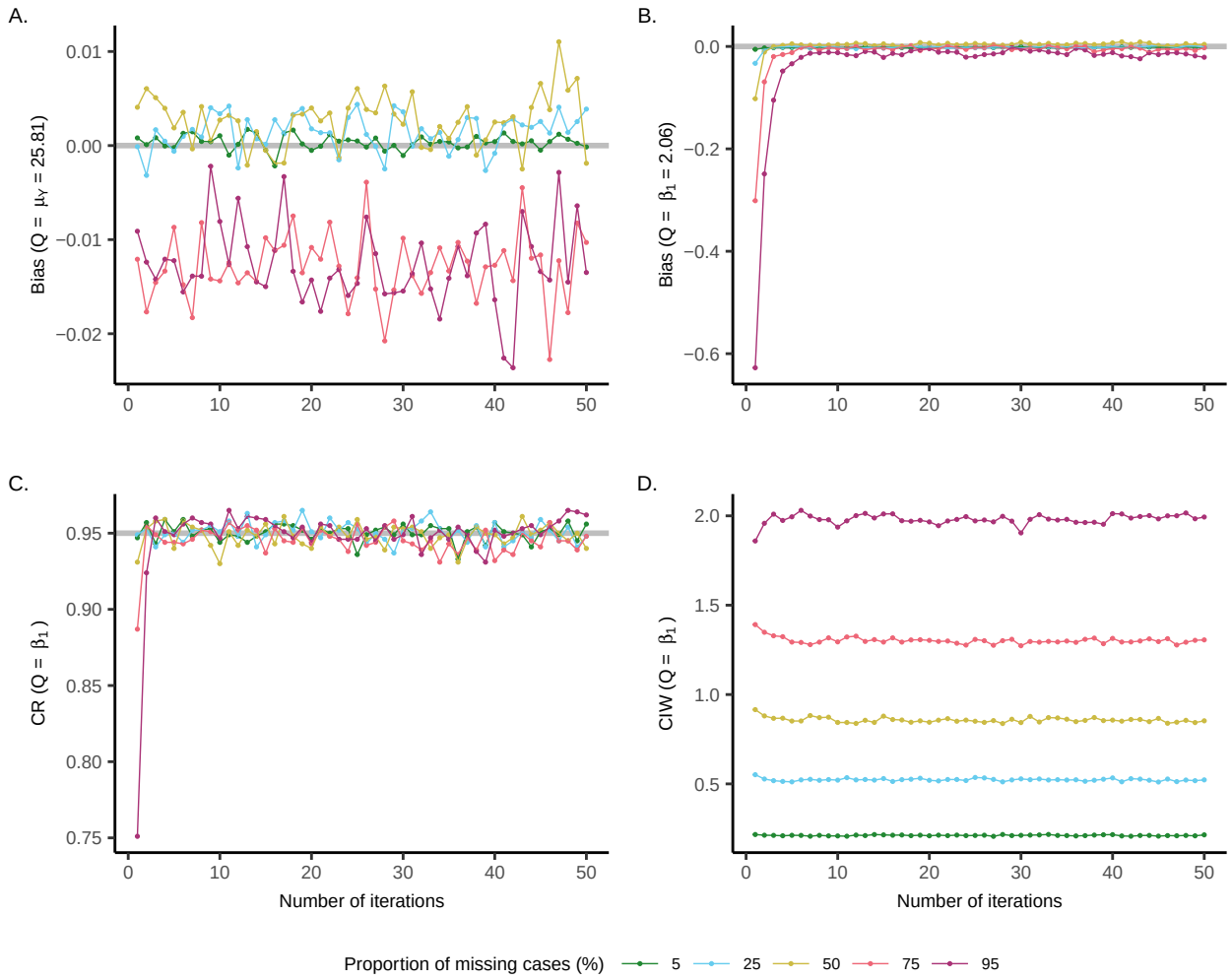

Figure 1. Impact of non-convergence on statistical inferences. Depicted are the bias, coverage rate (CR) and confidence interval width (CIW) of the worst-performing quantities of scientific interest Q in terms of bias. The gray lines represent the objectives: unbiased estimates with nominal coverage.

estimates and non-nominal coverages.

[Copy-paste from here!] These results demonstrate that the estimates of univariate scientific estimands Q are not impacted by early stopping of the MICE algorithm or the proportion of incomplete cases in y. Unbiased estimates may be obtained after just one iteration. Multivariate estimates, by contrast, are affected by both the number of iterations and the proportion of missing cases. Completely unbiased estimates are only obtained under low to moderate missingness ($p_{\text{inc}} \leq .50$), after at most three iterations. We observe approximately unbiased estimates after at most seven iterations (for any p_{inc} considered). This implies that the algorithm produces stable, non-improving estimates when $T \geq 7$.

Overall, the methods to diagnose non-convergence perform as expected: they indicate more signs of non-convergence in conditions with worse performance in terms of bias and confidence-validity. Under low to moderate missingness, it does not seem to matter on which θ the non-convergence diagnostics are applied. If we look at complete unbiasedness, however, we notice that univariate θs fail to diagnose the persistent bias in conditions where $p_{\text{inc}} \geq .75$.

The number of iterations necessary to obtain approximately unbiased, confidence-valid estimates corresponds to the $\widehat{R}$ threshold 1.2. The thresholds 1.1 and 1.01 seem too strict compared to the validity of the inferences. The threshold to diagnose significant AC-values does not appear to be appropriate in at least the current set-up, and perhaps in iterative imputation in general. A better heuristic to diagnose non-stationarity with AC may be through evaluation of the AC-values across iterations: If the values do not substantially decrease with T, approximate stationarity may be concluded.

4. Discussion

We have shown that iterative imputation algorithms can yield correct outcomes, even when a converged state has not yet formally been reached. Any further iterations would then burn computational resources without improving the statistical inferences. Our study found that—in the cases considered—inferential validity was achieved after five to ten iterations, much earlier than indicated by the $\widehat{R}$ and AC diagnostics. Of course, it never hurts to iterate longer, but such calculations hardly bring added value.

References

Cowles, M. K. and Carlin, B. P. Markov chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association, 91(434):883–904, 1996.

Gelman, A. and Rubin, D. B. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science, 7(4):457–472, November 1992. doi: 10.1214/ss/1177011136.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. Bayesian Data Analysis. CRC Press LLC, Philadelphia, PA, United States, 2013.

Hyndman, R. J. and Athanasopoulos, G. Forecasting: principles and practice. OTexts, 2018.

Lynch, S. M. Introduction to applied Bayesian statistics and estimation for social scientists. Springer Science & Business Media, 2007.

Murray, J. S. Multiple Imputation: A Review of Practical and Theoretical Findings. Statistical Science, 33(2):142–159, May 2018. doi: 10.1214/18-STS644.

R Core Team. R: A language and environment for statistical computing. Vienna, Austria, 2020.

Rubin, D. B. Inference and Missing Data. Biometrika, 63(3):581–592, 1976. doi: 10.2307/2335739.

Takahashi, M. Statistical Inference in Missing Data by MCMC and Non-MCMC Multiple Imputation Algorithms: Assessing the Effects of Between-Imputation Iterations. Data Science Journal, 16:37, July 2017. doi: 10.5334/dsj-2017-037.

Van Buuren, S. Flexible imputation of missing data. Chapman and Hall/CRC, 2018.

Van Buuren, S. and Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(1):1–67, December 2011. doi: 10.18637/jss.v045.i03.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. March 2019. URL http://arxiv.org/abs/1903.08008.

Zhu, J. and Raghunathan, T. E. Convergence Properties of a Sequential Regression Multiple Imputation Algorithm. Journal of the American Statistical Association, 110(511):1112–1124, July 2015. doi: 10.1080/01621459.2014.948117.

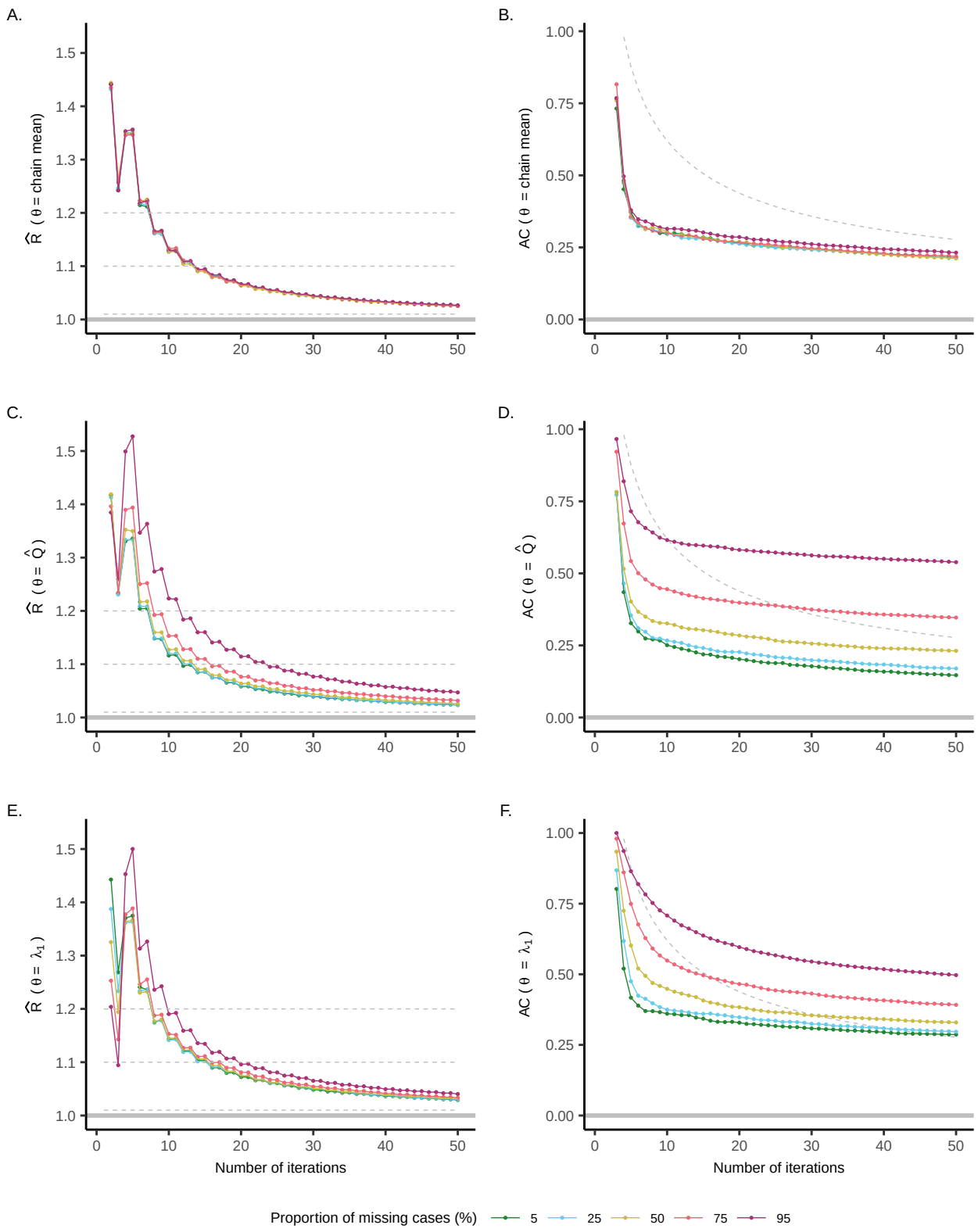

Figure 2. Non-convergence identified by diagnostic methods: $\widehat{R}$ and AC applied on several θs. The left-hand side of the figure contains $\widehat{R}$-values, and the right-hand side contains AC-values. Depicted in the rows are the scalar summaries θ: chain mean, chain variance, the quantity of scientific interest $\hat{Q}$, and the first eigenvalue of the variance-covariance matrix λ_1.