Summary of Missing values handling for machine learning portfolios

作者：Chen and McCoy (2024)　　阅读：王梦涵

2024 年 10 月 16 日

1. **What are the research questions?**
   - Explore the origins and imputation methods of missing values for ML portfolios.
2. **Why are the research questions interesting?**
   - A growing literature uses ML to combine numerous stock return predictors for asset pricing.
   - Buried in this literature is the problem of missing values
     - While imputing missing values seems dangerous, ML researchers often have no choice but to impute.
3. **What is the paper's contribution?**
   - Contribute to literature on missing values in cross-sectional predictor data.
     - Prior studies: advocate different imputation algorithms
     - Extension: 1) Take a more neutral approach and compare with EM imputation; 2) discuss the origins of missingness and exploring why simple mean imputation seems to perform well.
   - Contribute to literature using ML to combine numerous stock return predictors for asset pricing.
     - Prior studies: Complex imputation algorithm and complementary empirical results
     - Extension: Provide a rigorous justification for the ubiquitous use of mean imputation in ML papers.
   - Contribute to the ongoing debate about the factor structure of the cross-section of return
     - Complement the idea that 1) predictability is multidimensional for all stocks, and that 2) a moderately strong factor structure is possible for only large stocks.
4. **What hypotheses are tested in the paper? list them explicitly**
   - H1: Simple cross-sectional mean imputation handles missing values in ML portfolios better than EM imputation.

   (a) **Do these hypotheses follow from and answer the research questions?**
   Yes

   (b) **Do these hypotheses follow from theory or are they otherwise adequately developed? Please explain the logic of the hypotheses (use visualization if possible)**
   - Observed predictors provide little information about the missing predictors.
   - Sophisticated imputations may introduce estimation noise
5. **Sample: comment on the appropriateness of the sample selection procedures**
   - The selection of 159 predictors is reasonable.
6. **Dependent and independent variables: comment on the appropriateness of variable definition and measurement (focus on the key dependent variables and independent variables)**
   - Mean imputation: the imputation represent information about the factor itself.
7. **Regression/prediction model specification: comment on the appropriateness of the regression/prediction model specification**

- ML methods and strategies are classical and appropriate.

8. **What difficulties arise in drawing inferences from the empirical work**

    - The empirical word is rigorous.

9. **Describe at least one publishable and feasible extension of this research**

    - Missing value imputation in ML for other tasks (e.g., variable prediction)