Summary of *Missing values handling for machine learning portfolios*

Andrew Y. Chen, Jack McCoy(JFE, 2024)

2024.10.16 石宛青

1. **What are the research questions?**
   - How to handling missing values for machine learning portfolios?
2. **Why are the research questions interesting?**
   - literature applies methods from machine learning (ML) to asset pricing.
   - Faced many predictors, standard practice of dropping missing values stocks is untenable.
   - ML researchers often have no choice but to impute.
3. **What is the paper' s contribution?**
   - contribute on literature in missing values in cross-sectional predictor data.
     - **Prior**:
       * different imputation algorithm, though all avoid modeling the missingness process(Freyberger et al. ,2023; Bryzgalova et al.,2023)
     - **Extend**:
       * more neutral approach,compares textbook imputation methods with cross-sectional mean imputation used in asset pricing.
       * discussion of the origins of missingness for 159 predictors
   - contribute on ongoing debate about the factor structure of the cross-section of returns.
     - consistent with many dimensions in all stocks((Green et al.,2014s) and oderately strong factor in only large stocks(Green et al.,2017)
4. **What hypotheses are tested in the paper?**
   - simple cross-sectional mean imputation performs well compared to more sophisticated imputation methods, such as EM, for handling missing values for ML portfolios.

   Do these hypotheses follow from theory? Explain logic of the hypotheses.
   - observed data provide limited information about missing values,therefore,complex imputation techniques may introduce noise, leading to underperformance in return forecast
5. **Comment on the appropriateness of the sample,variable,model**
   - dataset includes a wide predictors (159) and covers long period (1985-2021) stock
6. **What difficulties arise in drawing inferences from the empirical work?**
   - Why not consider compare some time seires imputation algorithm?
7. **Describe at least one publishable and feasible extension of this research.**
   - consider the variations in correlation and missing patterns, examine different types of predictors (like market vs. financial) affect missing value imputation, and look at industry differences to improve imputation methods.