Missing Value Imputation using Low-Rank and Low-Norm Models

Knowledge Lab Team Presentation

Nandana Sengupta
(co-authored with James Evans and Madeleine Udell)

November 30, 2015

Introduction

- Missing data arise in almost all empirical analysis.
- Distracts from main goal of study.
- Ad-hoc methods
  - Complete case analysis.
  - Available case analysis.
  - Mean Imputation.
- Concerns about validity of inferences.
- Types of Missing Data
  - Missing Completely at Random.
  - Missing at Random.
  - Missingness depends on unobservables.

Multiple Imputation

- Rubin (1976), Schafer (1998), Van Buuren et al (1999), King et al (2000, 2015)
- Idea: Analysis should reflect uncertainty inherent in imputation.
- Assumption: MAR
- 3 stage scheme
  - Imputation
  - Analysis
  - Combining Results
- Imputation Step:
  - Parametric Assumptions (like multivariate normality).
  - Iterative procedures used.

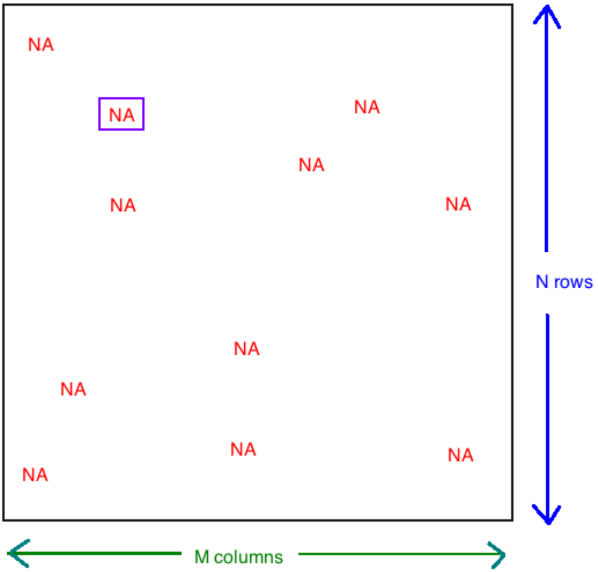
Multiple Imputation

- Two Standard Imputation Approaches:
  - MCMC mechanism: $(Y_{miss}^{(1)}, \theta^{(1)}), (Y_{miss}^{(2)}, \theta^{(2)}), \ldots$.
  - Chained Equations: iteratively fit univariate regression models.
- Analysis: perform as if full data is observed.
- Combining Results:
  - Point Estimate: $\overline{Q} = \frac{1}{m}\sum_{i=1}^{m} \widehat{Q_i}$; $\widehat{Q_i}$ = point estimate from imputation i.
  - Variance: $T = \overline{U} + \left(1 + \frac{1}{m}\right) B$; U = within imputation variance ; B = between imputation variance.
- 'R' Packages: Amelia, MICE, MI.

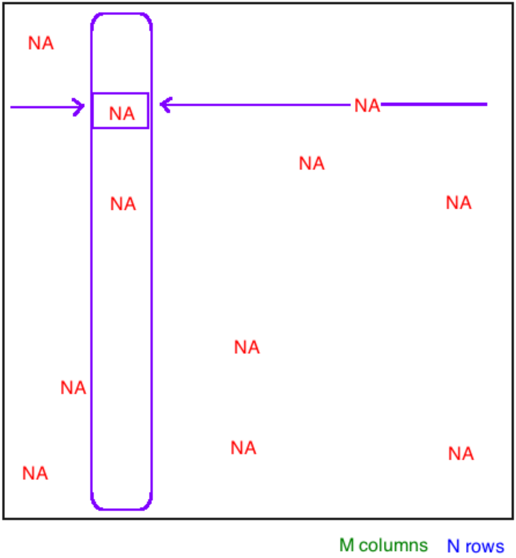
Low Norm and Low Rank Models

- Matrix Factorization approaches.
- Srebro (2004), Udell et al (2014).
- Approximate matrix A (dimension $m \times n$) by $X'Y$.
- minimize $\sum_{i,j} L_{i,j}(x_i y_j, a_{ij}) + \sum_{i=1}^{m} r_i(x_i) + \sum_{j=1}^{n} r_j(y_j)$.
  - L: Loss function (over columns).
  - $r(.)$: regularization functions.
  - X, Y initialization: SVD good starting point.
  - Low Norm Models: $r(x) = \gamma ||X^2||$.
  - Low Rank Models: $Rank(X'Y) \leq k$.
  - Low Rank, Low Norm Models: Both
  - k, γ chosen via crossvalidation.
- Julia Implementation: LowRankModels

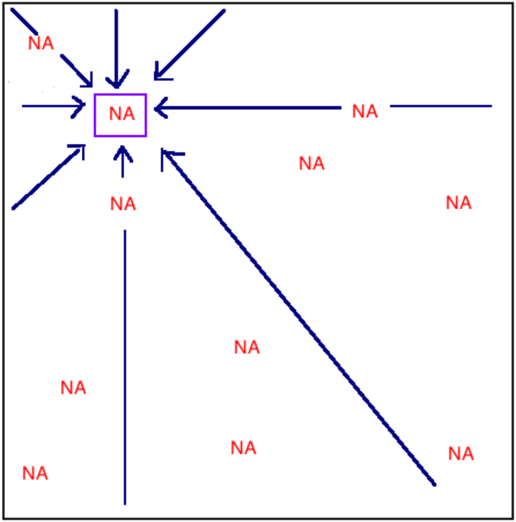
Dataset with missing values



MICE Single Imputation

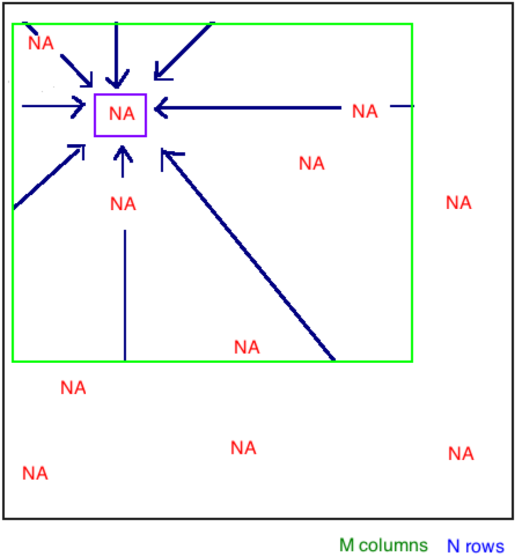


Low Norm Models



M columns N rows

Low Rank Models



Application 1: General Social Survey Data (GSS)

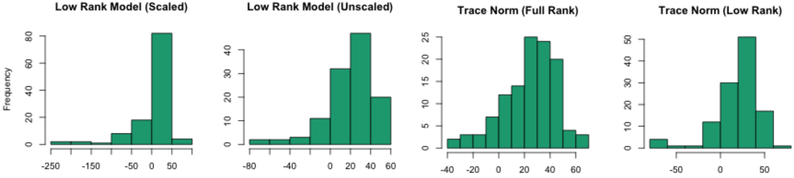
- Sociological survey: adults in randomly selected US households.
- Data on attitudes and demographic characteristics of adults.
- Subset of GSS 2014 data used for analysis
  - columns corresponding to identifying variables
  - columns with non-varying entries
  - $\geq 33\%$ missing entries
  - highly correlated columns ($\rho > 0.70$).
- Evaluation Strategy
  - 10% of observed data are randomly assumed missing ($N_{miss,ind}$)
  - Imputations using
    - Low Rank (Scaled), Low Rank (Unscaled), Trace Norm (Full Rank), Trace Norm (Low Rank), MICE.
  - Loss calculated over $N_{miss,ind}$ observations:
    - scale columns, quadratic loss over non-categorical columns, zero-one loss over categorical columns.

Results

- Overall Trace (Full Rank) had lowest loss, all Low Rank and Low Norm models outperformed MICE
- Column-wise: $\approx 80\%$ columns had lower loss compared to MICE
  - Summary Table

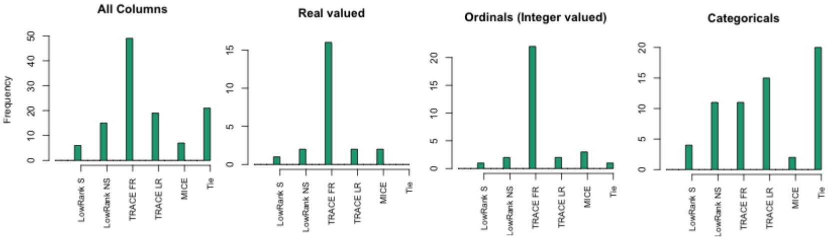
	LowRank (S)	LowRank (NS)	Trace (FR)	Trace (LR)	MICE
Scaled Loss/(10^3)	18.50	15.80	14.40	15.80	20.60
%age reduction over MICE	10.10 %	23.40 %	30.10 %	23.00 %	–
%age cols w/ lower loss	73.50 %	84.60 %	87.20 %	84.60 %	–

  - Columnwise percentage reduction in Loss over MICE

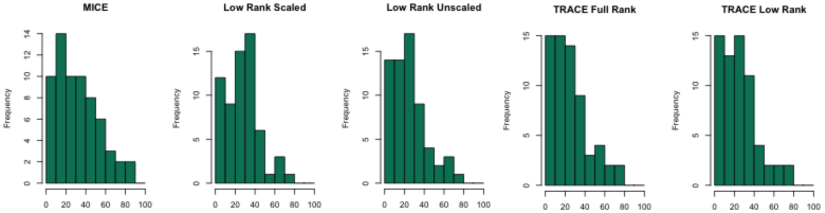


Results

- Method with lowest loss across columns



- Categorical columns misclassification by method



Next Steps

- Replicating missingness patterns before applying imputation techniques.
- Extending and applying to longitudnal survey data (e.g. National Longitudnal Survey of Youth).
- Applying to larger subsets of GSS data.
- Working with more advanced options of MICE and LowRankModels.
- Extending to Max and Frobenius norms.
- Extending Low Rank and Low Norm methods to Multiple Imputation setting.

Thank you!
(Comments and Suggestions Welcome)