Unified Approach to Variable Selection in Missing Data via Least Squares Approximation

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Eric M. Reyes and Cody Roberts Department of Mathematics, Rose-Hulman Institute of Technology

Abstract

Ad hoc methods continue to be employed for performing variable selection in the presence of missing data. We evaluate the use of multiple imputation (MI) in conjunction with the least squares approximation (LSA) of Wang and Leng (*JASA*, 2007). The result is a general procedure for performing variable selection while accounting for data which are subject to covariates missing at random.

Step 1: Multiple Imputation

Our two-step approach begins with addressing the missing data; this is done through multiple imputation [1]. This is arguably the most common method for handling missing data in many fields and is now implemented in many major statistical analysis packages. This approach is illustrated in Figure 1 below.

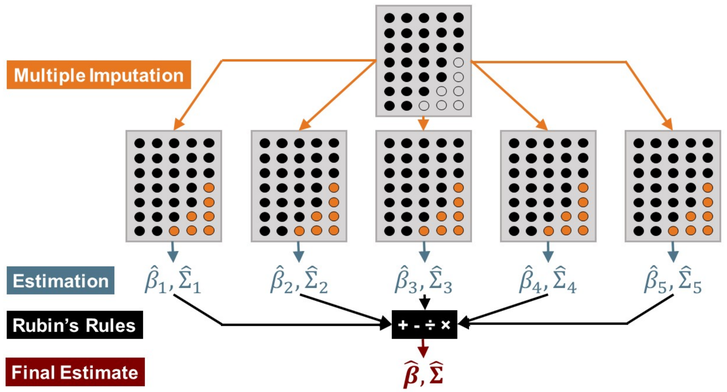

Figure 1: Illustration of the multiple imputation procedure.

- Assuming the models are correctly specified, the resulting estimates are consistent.

Step 2: Least Squares Approximation

Wang and Leng [2] proposed approximating the LASSO-type objective function with its asymptotic least squares equivalent. Let $\boldsymbol{\beta}$ be a parameter of interest and $\mathcal{L}_n(\boldsymbol{\beta})$ be the loss function such that $\widehat{\boldsymbol{\beta}} = \arg\min \mathcal{L}_n(\boldsymbol{\beta})$ yields an estimator with asymptotic covariance matrix $\boldsymbol{\Sigma}$. Then, the adaptive LASSO objective function

$$n^{-1}\mathcal{L}_n(\boldsymbol{\beta}) + \sum_{j=1}^{p} \lambda_j |\beta_j| \approx \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)^{\top} \widehat{\boldsymbol{\Sigma}}^{-1} \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right) + \sum_{j=1}^{p} \lambda_j |\beta_j|$$

- LSA requires only consistent estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$.
- The tuning parameter is chosen by BIC-type measure, resulting in the oracle property.
- We propose using the MI estimates in the LSA procedure to conduct variable selection in the presence of missing data.

Simulation Study

For each of 5000 replications, we generated a response vector

$$\mathbf{y} \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}\right)$$

where $\mathbf{X}$ is an $(n \times 20)$ design matrix with $\mathbf{x}_i \sim N(\mathbf{0}, \boldsymbol{\Gamma})$, $\boldsymbol{\Gamma}_{i,j} = \rho$ for all $i \neq j$ and $\boldsymbol{\Gamma}_{i,i} = 1$, and $\boldsymbol{\beta} = \left(4, 2, 1, 1, \mathbf{0}^{\top}\right)^{\top}$. The variance σ^2 was chosen such that the theoretical $R^2 = 0.6$; we considered $n = 100, 500$ and $\rho = 0, 0.3, 0.6$.

Missingness was introduced in x_2 and x_5 such that the i-th observation was missing if $R_i > 0$ where

$$R_i = \gamma_0 + y_i + x_{1,i} + x_{4,i} + x_{6,i} + x_{7,i} + \epsilon_i$$

where ϵ_i has a logistic distribution for which the scale parameter was chosen such that the theoretical $R^2 = 0.3$, and γ_0 was chosen such that $Pr\left(R_i > 0\right) = 0.1$ or 0.4.

Simulation Results

- Performance of assessed using the geometric mean G of sensitivity and specificity (higher is better).

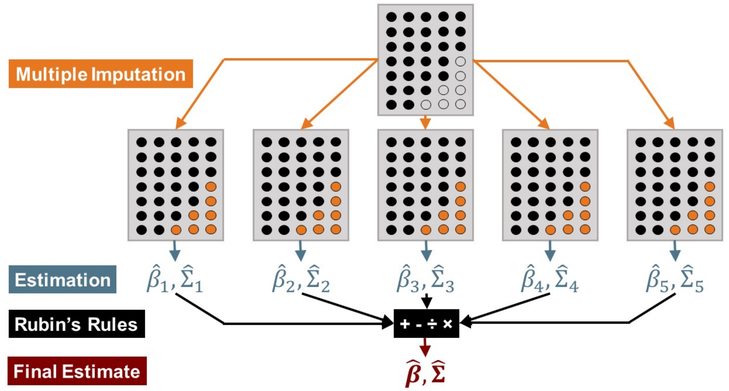

Figure 2: Ratio of G for each method to that from the full data.

Conclusions

The two-step MI-LSA procedure is a general method for variable selection when the covariates are subject to missingness, and it is superior to analysis on the complete cases alone. Future work is needed to evaluate its performance against other proposed methods for variable selection in the presence of missing data and to evaluate the performance of LSA when used in conjunction with inverse probability weighting.

[1] Rubin DB.
Multiple Imputation After 18+ Years.
Journal of the American Statistical Association, 91:473-489, 1996.

[2] Wang H and Leng C.
Unified LASSO Estimation by Least Squares Approximation.
Journal of the American Statistical Association, 102:1039-1048, 2007.