Predicting Missing Responses in the CORE Social-Emotional Learning Survey

Klint Kanopka, Ryan Eberhardt, Martin Mbuthia

Stanford University

Abstract

Missing data is a challenge in many areas of statistical analysis. Using a socioemotional learning survey with over 1,000,000 responses, we present a method for imputation of missing survey responses that requires no domain knowledge or theory of underlying construct relationships. This method uses a three-stage estimation process that outperforms state-of-the-art single imputation methods on our dataset.

Project Scope

- Item nonresponse complicates the analysis of survey and test data.
- To address this problem, we developed a method for imputing missing data using a CSP, clustering and neural networks.
- Model input is student item responses with missing data.
- Output is item responses with omitted item responses imputed.

Dataset

- Students were asked to respond to 25 questions.
- Questions mapped to 4 constructs: self management, growth mindset, self-efficacy, and social awareness.
- Dataset contains 1,027,710 student responses.
- 747,656 of these are complete responses. 27% of responses omitted one or more items.

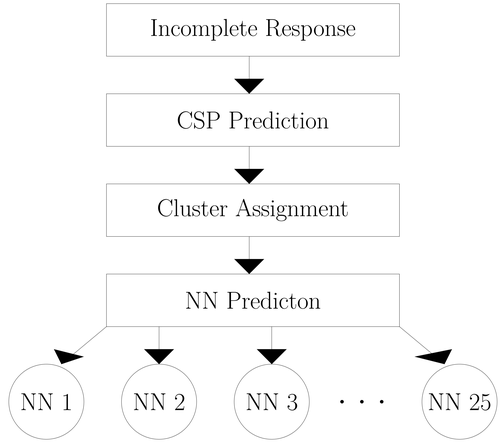
Evaluation

- Accuracy: Each value represents a qualitatively different response, so we want to get that "right."
- MSE: Possible responses are ordered, so if we can't be "right," we should be "close."

Benchmarks

- Baseline: Mean Imputation - Naive (but common) single imputation, missing values are replaced by within-item mean.
- Oracle: Iterated Ridge Regression Imputation - State of the art single imputation, missing value estimations are iteratively refined.

Approach



Predictions are made using the following three step process:

- The CSP is used to make baseline predictions of missing item responses.
- The now complete response is assigned to a cluster.
- The complete data and cluster assignment are fed to the neural networks to make final predictions of each originally omitted item.

Model Development

- Only complete item responses were used for development.
- A CSP was developed to do first stage imputation.
- Clustering was performed on the full set of complete item responses using k-means, with $k = 3$ clusters. The number of clusters was selected using the elbow method.
- 25 individual neural networks were trained, one per item. Their input is the response to the 24 other items and cluster membership. Output is an imputed item response.
- To simplify development, networks were constrained to share hyperparameters.
- Three hidden layers, 10 nodes per layer, softmax output layer and an Adam optimizer provided the best balance of performance to training time.

Results

We evaluated each model on a test set that had six randomly-selected values hidden from each survey response.

Imputation Model	**Accuracy**	**MSE**
Baseline: Mean Imputation	0.31	1.27
Oracle: Iterated Imputation	0.49	0.85
CSP Alone	0.49	1.29
Mean Imputation $\rightarrow$ Neural Networks	0.52	1.02
CSP $\rightarrow$ Neural Networks	0.53	1.09
CSP $\rightarrow$ Clustering $\rightarrow$ Neural Networks	0.53	1.08

Discussion

Our approach has a number of benefits:

- Accuracy - It outperforms both our baseline and oracle.
- Once the model is trained, imputation can be done for individual respondents or in batches.
- The framework is flexible. It does not require domain knowledge and can be adapted to any input format.
- Potential for further optimization - performance may improve with deeper neural networks or more expressive CSPs.
- By making specific predictions, output can be fed into other models.
- The output of the softmax layer could be used to weight values for multiple imputation tasks, though this requires further experimentation.

Despite this, there are drawbacks:

- Optimizing many parallel neural networks is a non-trivial task.
- Training the networks and K-means model is a computationally intensive up-front cost.
- Solving a CSP for each row of data incurs a heavy prediction-time cost.
- Each model is a bespoke construction for a particular survey, instrument, or application.

Acknowledgements

Thank you to Policy Analysis for California Education (PACE) at Stanford for providing us access to the CORE SEL dataset.