EDWARD P. FITTS DEPARTMENT OF INDUSTRIAL AND SYSTEMS ENGINEERING

NC STATE UNIVERSITY

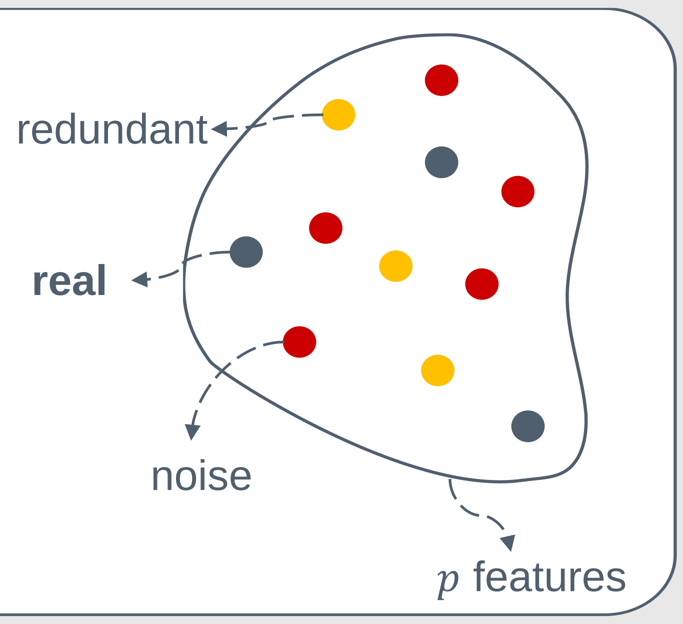
SIMULATION OPTIMIZATION BASED FEATURE SELECTION

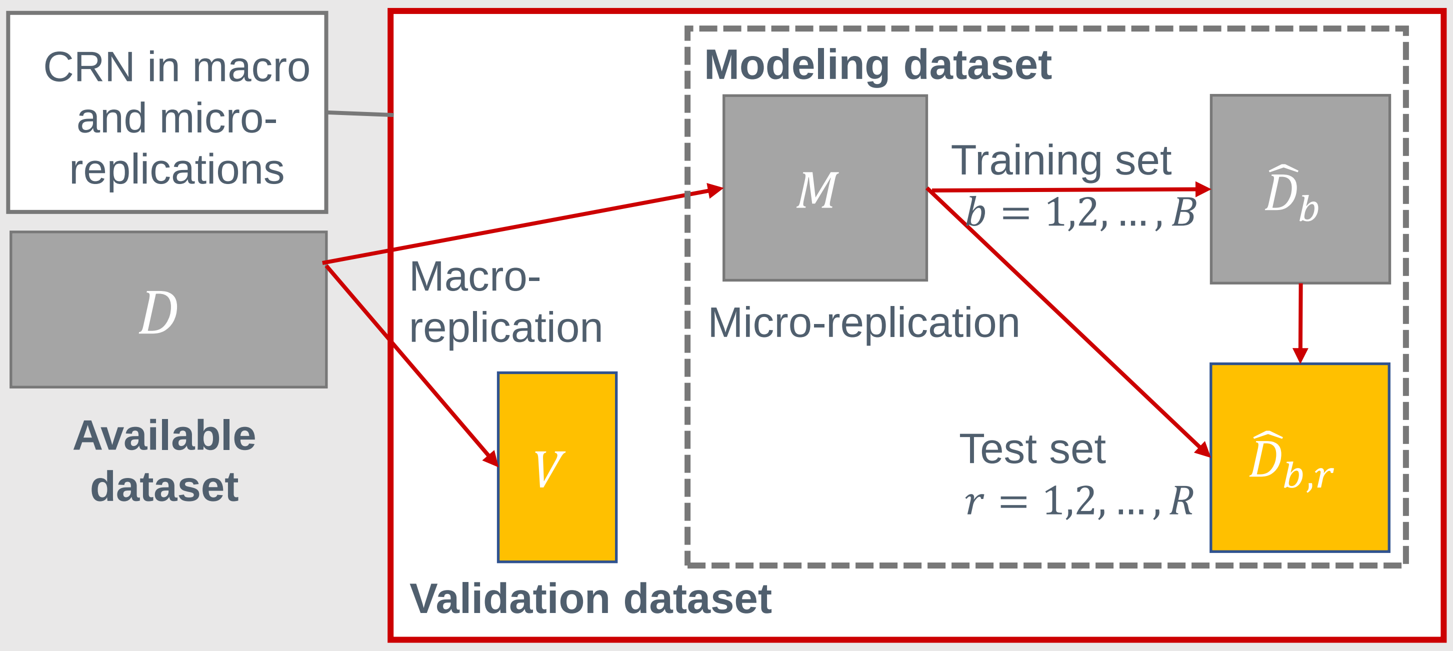
Kimia Vahdat (kvahdat@ncsu.edu) and Sara Shashaani, Ph.D. (sshasha2@ncsu.edu)

Why Feature Selection?

Our goal is to Find the true features to improve the model in:

- ✓ interpretability,
- ✓ prediction accuracy, and
- ✓ efficiency.





Finding the most informative features in the big data applications is a challenge that can improve predictions and interpretability of the underlying systems.

Due to the uncertainty in the data, we formulate this problem stochastically, which is generalizable for any learning algorithm of choice. The resulting feature subsets are more robust to the changes in the data and lead to better predictions in simulated and real datasets.

Problem Statement:

$$\min_{x\in\{0,1\}} f(\boldsymbol{x}) := \mathbb{E}_{D\sim P}\left[\mathbb{E}_{D_0\sim P_0}\left[Q_{D_0}(r_D(\boldsymbol{z},\boldsymbol{x}),\boldsymbol{y})\right]\right]$$

Out-of-sample $D_0 \sim P_0$

In-sample $D \sim P$

- P_0 and P are the unknown data distributions, which introduce the input uncertainty to the model.
- $D = \{< z_i, y_i >\}_i$ is the dataset on hand, where z_i is a p-dimensional variable.
- $\boldsymbol{x} = (x^1, \dots, x^p)$ $\boldsymbol{x_i} = \{0,1\}$
- $r_D(\boldsymbol{z},\boldsymbol{x})$ is a prediction model
- $Q_{D_0}(r_D(\boldsymbol{z},\boldsymbol{x}),\boldsymbol{y})$ is the deviation of predicted from observed

Datasets:

1 CMS dataset with binary response (**zero-inflated** with 9% non-zero), 19k instances and 380 features.

Prediction of unplanned admission due to heart failure, in the following month.

For patients >65 years old, who have:

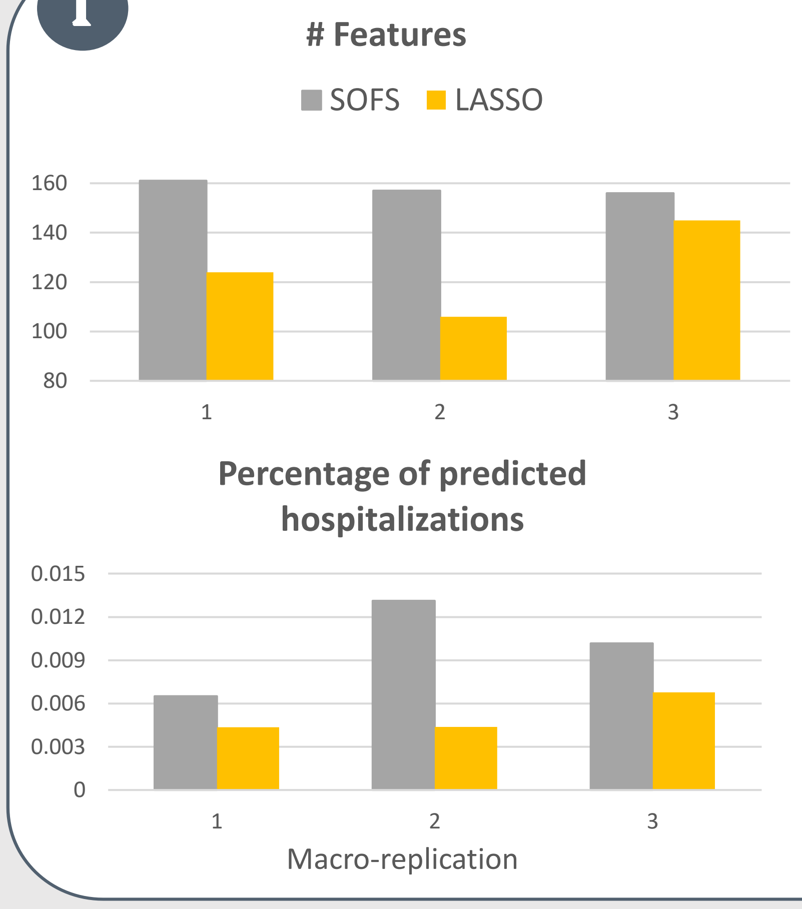
- no cancer history,
- more than 10 comorbidities,
- other minor conditions.

2 Simulated dataset with continuous response, 300 instances and 220 features; where real variables$\sim Gamma(2,2)$.

Feature type	Count
Real	12
Redundant	100
Noise	108

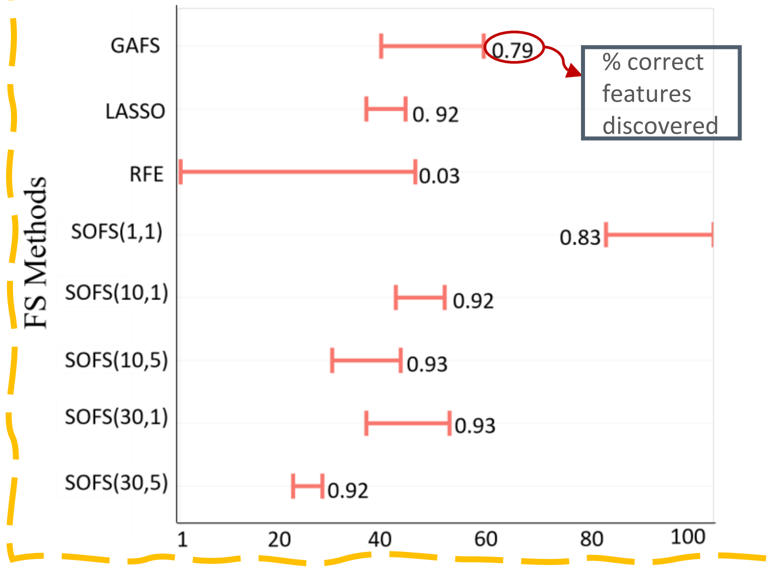
How does SOFS(B,R) perform?

1



2

The $[q_{0.1}, q_{0.9}]$ prediction interval for the size of the optimal feature set



✓ Better performance in: Accuracy and Sharpness

✗ Very time consuming

Conclusion

- Increased reliability in the results (more accuracy and less variability)
- Flexibility with any learning algorithm and response type
- Improved performance in identifying true features
- Surpassed all evaluated benchmarks

Future Work

- Employing adaptive sampling to reduce computation time
- Large sample sizes does not necessarily help
- More comparison on selected subsets
- Changing the learning algorithm

Thanks To Seth Guikema (University of Michigan) for initial discussions, ISE faculty, Dr. Mayorga and Ivy for their comments, and NC State Graduate Fellowship for funding support.