ROSE-HULMAN
INSTITUTE OF TECHNOLOGY

Tuning Variable Selection via Noise When Prediction is Not the Primary Objective

Eric M. Reyes and Xiaomo Wang
Department of Mathematics, Rose-Hulman Institute of Technology

Abstract

Historically, variable selection algorithms are tuned to maximize the predictive ability of the model. In some applications, such as medical research, prediction is not the primary aim of the model development. Instead, the primary goal is to correctly identify variables on the causal pathway. Wu, Boos, and Stefanski (2007) developed an approach for tuning variable selection algorithms through the addition of pseudovariables. We extend their approach to select the tuning parameter to maximize the F-score, a measure of the quality of proper classification. We also present a version of the method which avoids the generation of pseudovariables in forward selection.

Variable Selection as Classification

To motivate our deviation from this approach, cast the variable selection problem as a classification problem.

		Variable Selection Algorithm (tuning parameter $= \alpha$)		
		Selected Variable	Excluded Variable	
Data Generating Process	Informative Variable	True Positives $I(\alpha)$	False Positives $k_I - I(\alpha)$	k_I
	Uninformative Variable	False Negatives $U(\alpha)$	True Negatives $k_U - U(\alpha)$	k_U
		$S(\alpha)$	$k_T - S(\alpha)$	$k_T = p$

$$\text{Sensitivity} = I(\alpha)/k_I \qquad \text{PPV} = I(\alpha)/S(\alpha)$$

$$F_1 = \frac{2\,(\text{PPV})\,(\text{Sensitivity})}{(\text{PPV}) + (\text{Sensitivity})} = \frac{2I(\alpha)}{k_I + I(\alpha) + U(\alpha)}$$

Goal: choose α to maximize $E\left[\frac{2I(\alpha)}{1 + k_I + S(\alpha)}\right]$

Tuning via Noise

Wu, Boos and Stefanski [1] propose augmenting the design matrix with pseudovariables $\mathbf{Z}$ constructed by permuting the rows of $\mathbf{X}$. By running variable selection on augmented design matrix, we "follow the noise" to inform tuning the selection procedure.

Primary Assumption: $E\left[U(\alpha)\right] = E\left[U_{p,b}(\alpha)\right] = E\left[U^*_{p,b}(\alpha)\right]\left(\frac{k_U}{k_p}\right)$

where the p denotes using the pseudovariables and b a bootstrap replication. Under this assumption, and considering that $k_U \approx k_T - S(\alpha)$ for good choices of α,

$$\widehat{F}_1(\alpha) = \frac{2\left[S(\alpha) - \bar{U}^*_p(\alpha)\frac{k_T - S(\alpha)}{k_p}\right]}{2S(\alpha) + 1}$$

where $\bar{U}^*_p(\alpha) = \sum_{b=1}^{B} U^*_{p,b}(\alpha)$

Simulation Study

For each of 1000 replications, we generated a response vector

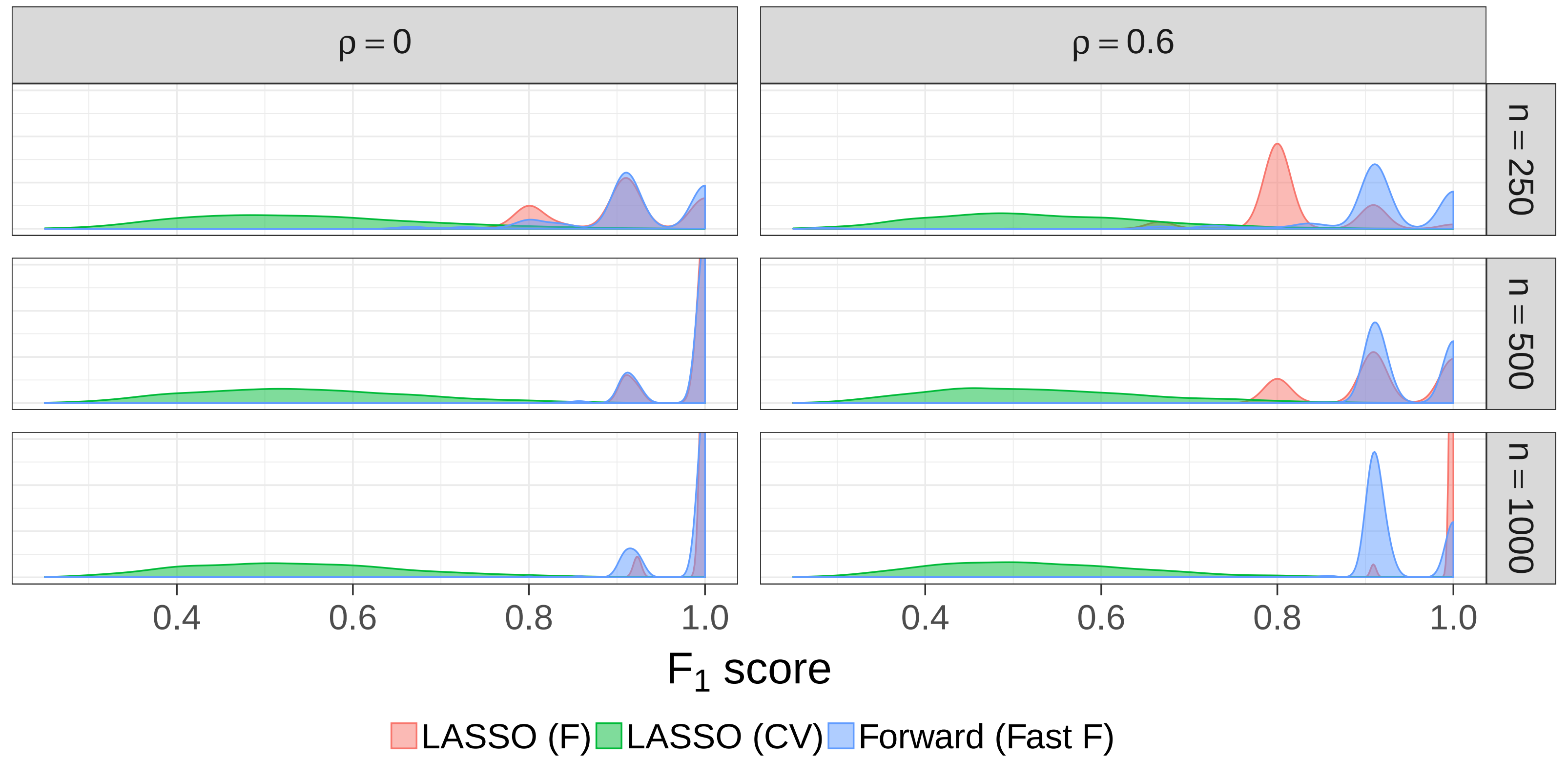
$$\mathbf{y} \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right)$$

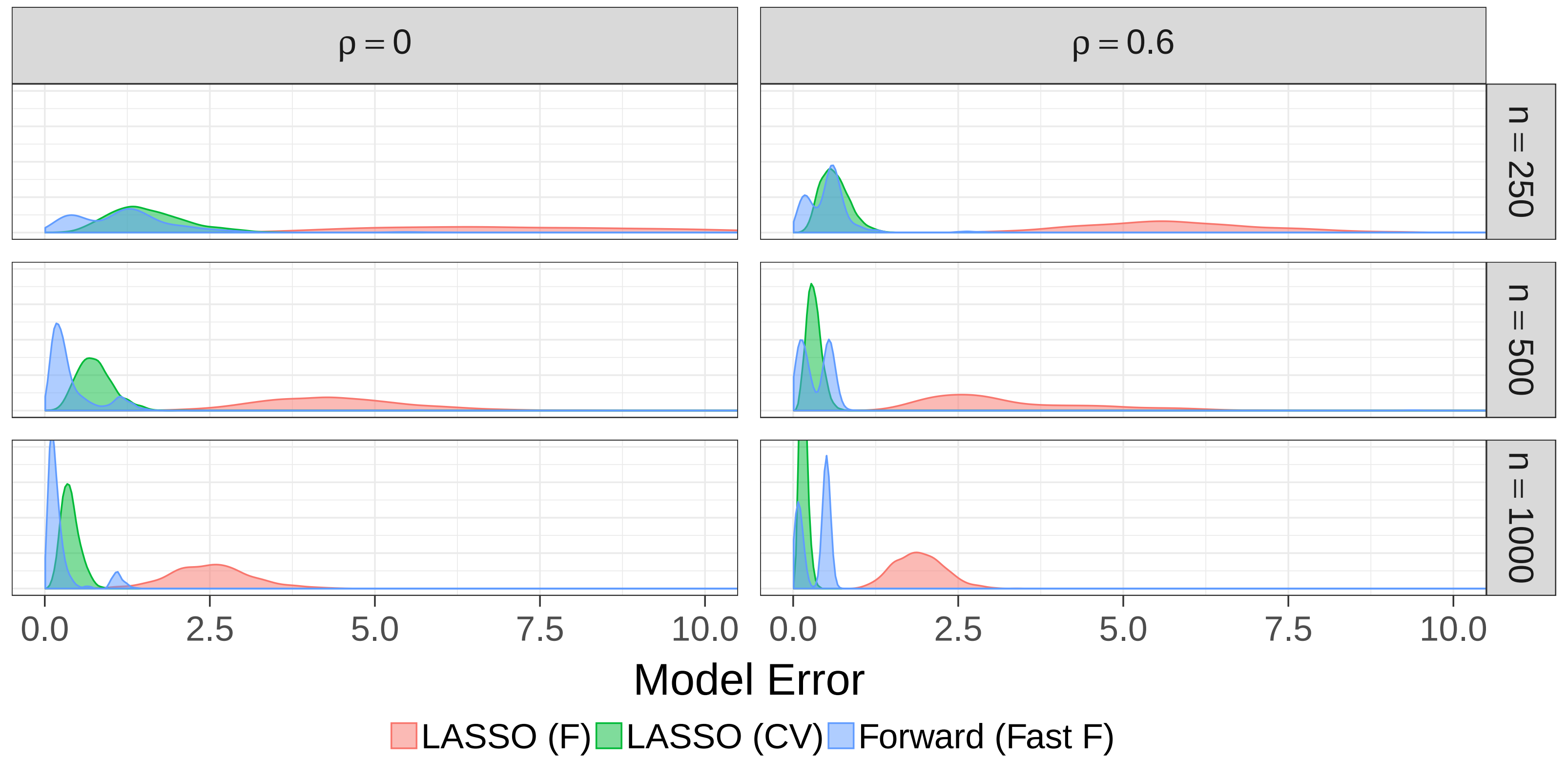
where $\mathbf{X}$ is an $(n \times 50)$ design matrix with $\mathbf{x}_i \sim N\left(\mathbf{0}, \boldsymbol{\Gamma}\right)$, $\boldsymbol{\Gamma}_{i,j} = \rho$ for all $i \neq j$ and $\boldsymbol{\Gamma}_{i,i} = 1$, and $\boldsymbol{\beta} = \left(-3, -2, -1, 1, 2, 3, \mathbf{0}^\top\right)^\top$. The variance σ^2 was chosen such that the theoretical $\mathrm{R}^2 = 0.6$; we considered $n = 250, 500, 1000$ and $\rho = 0, 0.6$. In all cases, $B = 500$ replications were used for the psuedovariable generation.

We considered three variable selection methods:

- LASSO (tuned via 10-fold cross-validation)
- LASSO (tuned to maximize F_1 via pseudovariables)
- Forward Selection (fast version of F_1 maximization)

Simulation Results





Conclusions

This general procedure for tuning a variable selection algorithm to optimize selection performance instead of predictive accuracy is competitive with existing variable selection procedures. Future work could suggest a better measure of selection performance or a faster version for penalty-based methods.

[1] Wu Y, Boos DD and Stefanski LA.
Controlling Variable Selection by the Addition of Pseudovariables.
Journal of the American Statistical Association, 477:235-243, 2007.

Joint Statistical Meetings (2017) https://reyesem.github.io reyesem@rose-hulman.edu