# Penalized regression with prior research support: the multi-step elastic net

Elizabeth Chase and Philip Boonstra, University of Michigan Department of Biostatistics

MSSISS, April 3, 2018

## Aims and Goals

To develop a method that helps investigators identify clinically relevant predictors of disease.

Specifically, we want to modify the elastic net in order to promote:

1. Inclusion of well-established, proven covariates
2. Sensitivity to true, unestablished covariates
3. Specificity against untrue, unestablished covariates
4. Model efficiency and accuracy

## Introduction

Since Robert Tibshirani's development of the lasso in 1996, research on penalized regression has grown rapidly. The lasso is now used across a range of disciplines, and dozens of extensions to the lasso have been proposed. Rather than minimizing the $L_2$-loss, as is done in ordinary least squares, the lasso minimizes the $L_2$-loss plus the penalty:

$$p(\lambda) = \lambda \|\beta\|_1$$

where $\lambda$ is a tuning parameter that controls the severity of the penalty. This additional penalty makes it more difficult for variables to enter the model, thus promoting sparsity and efficiency.

But what if we already strongly believe that a variable should be included in the model? Examples of these variables include:

- Smoking history when estimating lung cancer risk
- LDL cholesterol for predicting heart disease
- BRCA1/2 status for breast cancer risk

It makes little sense to penalize these variables, and doing so could result in a distorted model that underestimates these known variables' true effect. Moreover, by using this extra information—the knowledge that some variables have already met our standard of credibility—we can hope to gain some extra efficiency and produce a better model.

## Method

Rather than work with the lasso itself, we use the elastic net, Zou and Hastie's 2005 extension to the lasso that uses a mixing parameter ($\alpha$) to blend ridge regression with lasso. This results in a more even blend of shrinkage (the ridge component) and selection (the lasso component). The elastic net minimizes the $L_2$-loss plus:

$$(1-\alpha)\tfrac{\lambda}{2}\|\beta\|_2^2 + \alpha\lambda\|\beta\|_1$$

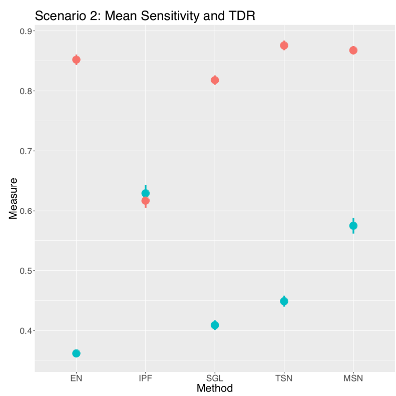
In our extension, we instead minimize with the penalty:

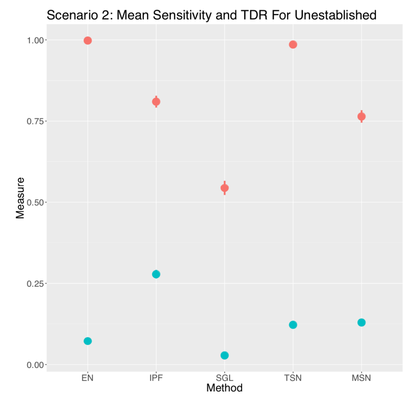
$$(1-\alpha)\tfrac{\lambda}{2}(\phi_1\|\beta_1\|_2^2 + \phi_2\|\beta_2\|_2^2) + \alpha\lambda(\phi_1\|\beta_1\|_1 + \phi_2\|\beta_2\|_1)$$

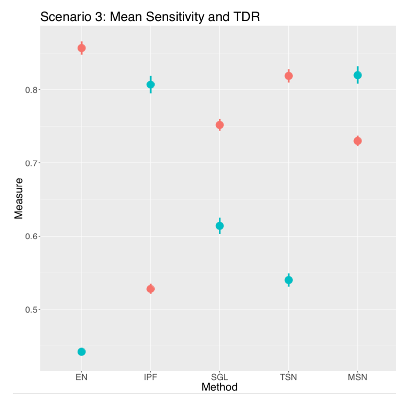
where $\beta_1$ denotes the established covariates, $\beta_2$ denotes the unestablished covariates, and $\phi_1,\phi_2$ determine their respective degree of penalization. We consider two ways of selecting $\phi_1,\phi_2$:

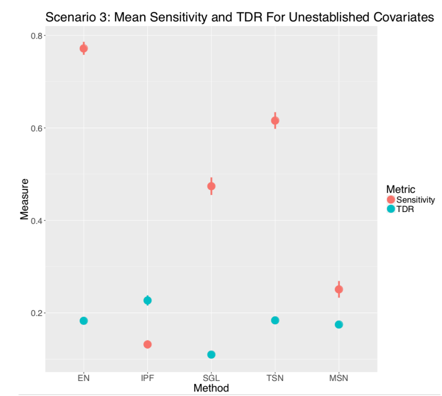
*Two-Step Net (TSN):* Use cross-validation to select an appropriate $\lambda$, $\alpha$ for 3 different penalty factors: $\phi_1$: {0, 0.5, 1} while $\phi_2$ is held at 1 — no penalization on the established covariates, half-penalization on the established covariates, or full penalization on the established covariates (note that this is the normal elastic net). Then choose the best of these 3 models. We use deviance as the selection criteria here (although other criteria could be used).

*Multi-Step Net (MSN):* Use cross-validation to select an appropriate $\lambda$, $\alpha$ for 5 combinations of $\phi_1,\phi_2$ : $\phi_1 = \{0, 0.5, 1\}$ while $\phi_2$ is held at 1 and $\phi_2 = \{2, \infty\}$ while $\phi_1$ is held at 1—these are the same three options as above, plus the options of double penalization on the unestablished covariates or infinite penalization on the unestablished covariates. Choose the best of the 5 models.




## Simulation Study

We considered three different combinations of established vs. unestablished covariates, conducted over a range of sample sizes. Here, we present the results from the smallest scenario:

| Scenario | Pe | Pu | Magnitude of established | Magnitude of unestablished |
|---|---|---|---|---|
| 1 | 10 | 30 | All 0.26 | All zero |
| 2 | 10 | 30 | All 0.2 | 1 $\beta = 0.6$ |
| 3 | 10 | 90 | All 0.25 | 5 $\beta = 0.05$ |

All scenarios had n=200 and correlation of 0.2. Cross-validation was conducted with 5 folds and 25 replicates. Each scenario was replicated 500 times. We compared our method to the original elastic net (EN), Boulesteix's IPF-Lasso (IPF), and Simon's sparse group lasso (SGL).

*Mean AUC (S.E.)*

| | Scenario 1 | Scenario 2 | Scenario 3 |
|---|---|---|---|
| EN | 0.765 (0.001) | 0.730 (0.001) | 0.706 (0.001) |
| IPF | 0.750 (0.002) | 0.726 (0.002) | 0.707 (0.002) |
| SGL | **0.775 (0.001)** | 0.729 (0.001) | **0.730 (0.001)** |
| TSN | 0.770 (0.001) | **0.731 (0.001)** | 0.718 (0.001) |
| MSN | 0.771 (0.001) | 0.728 (0.001) | 0.727 (0.001) |

In Scenarios 1 and 3, the MSN performs well, with AUC almost comparable to the SGL and with a much higher sensitivity and TDR. For Scenario 2, IPF performs better.

## Conclusions and Significance

We present a penalized regression method that incorporates prior research support. When the effect of unestablished covariates is small, our method performs better than other similar methods. However, when unestablished predictors have a strong effect, the IPF-lasso might be a better option. To our knowledge, ours is the first attempt at systematically modifying penalty factors within in an elastic net type penalty.

## References

Boulesteix, A; De Bin, R; Jiang, X; Fuchs, M. (2017). IPF-LASSO: Integrative L1-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. *Computational and Mathematical Methods in Medicine*.

Simon, N; Friedman, J; Hastie, T; Tibshirani, R. (2012). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics 22(2)*, 231-45.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *JRSS Series B 58(1)*, 267-88.

Zou, H; Hastie, T. (2005). Regularization and variable selection via the elastic net. *JRSS Series B 67*, 301-20.

## Acknowledgements and Contact Information

This material is based upon work supported by National Science Foundation Grant DGE-1256260 and National Institutes of Health Grant T32-CA083654.

For more information or to address concerns, please contact: Elizabeth Chase, ecchase@umich.edu.