Elastic net illustration

Paul DW Kirk

10/10/2021

An illustration of the differing ways in which LASSO and elastic net treat correlated predictors.

Introduction

We consider regression models of the form:

$$Y_i = \beta_0 + \sum_{l=1}^{p} \beta_l X_{il}. \tag{1}$$

The elastic net estimator puts a penalty of the form

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

on the parameter vector, where we have defined the vector β to exclude the intercept β_0.

Simulation setup

We simulate data vectors $X_i \in \mathbb{R}^p$ by sampling from a p-variate Gaussian distribution, $X_i \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$. Throughout, we take $p = 400$. We set the covariance matrix Σ to be equal to the identity, except that we define correlated sets of predictors $\{X_2, X_3\}, \{X_4, X_5, X_6\}$, and $\{X_7, X_8, X_9, X_{10}\}$ by setting

$$\Sigma_{2,3} = \Sigma_{4,5} = \Sigma_{4,6} = \Sigma_{5,6} = \Sigma_{7,8} = \Sigma_{7,9} = \Sigma_{7,10} = \Sigma_{8,9} = \Sigma_{8,10} = \Sigma_{9,10} = \rho \tag{3}$$

(together with their symmetric counterparts), so that the top left corner of the covariance matrix has the following form:

$$\Sigma = \begin{pmatrix} 1 & & & & & & & & & \\ & 1 & r & & & & & & & \\ & r & 1 & & & & & & & \\ & & & 1 & r & r & & & & \\ & & & r & 1 & r & & & & \\ & & & r & r & 1 & r & r & r & r \\ & & & & & & 1 & r & r & r \\ & & & & & & r & 1 & r & r \\ & & & & & & r & r & 1 & r \\ & & & & & & r & r & r & 1 \\ & & & & & & & & & \ddots \end{pmatrix}, \tag{4}$$

where omitted values are equal to zero. We may therefore control the strength of correlation between predictors in the same corrlated block by controlling ρ. We consider high-correlation settings, with $\rho \in \{0.85, 0.9, 0.95, 0.99\}$.

We initially generate an outcome y according to the following model:

$$Y_i = X_{i1} + (X_{i2} + X_{i3})/2 + (X_{i4} + X_{i5} + X_{i6})/3 + (X_{i7} + X_{i8} + X_{i9} + X_{i,10})/4 + \epsilon, \quad (5)$$

where $\epsilon \sim \mathcal{N}(0, 1)$.

`elastic_net`:

Results (1)

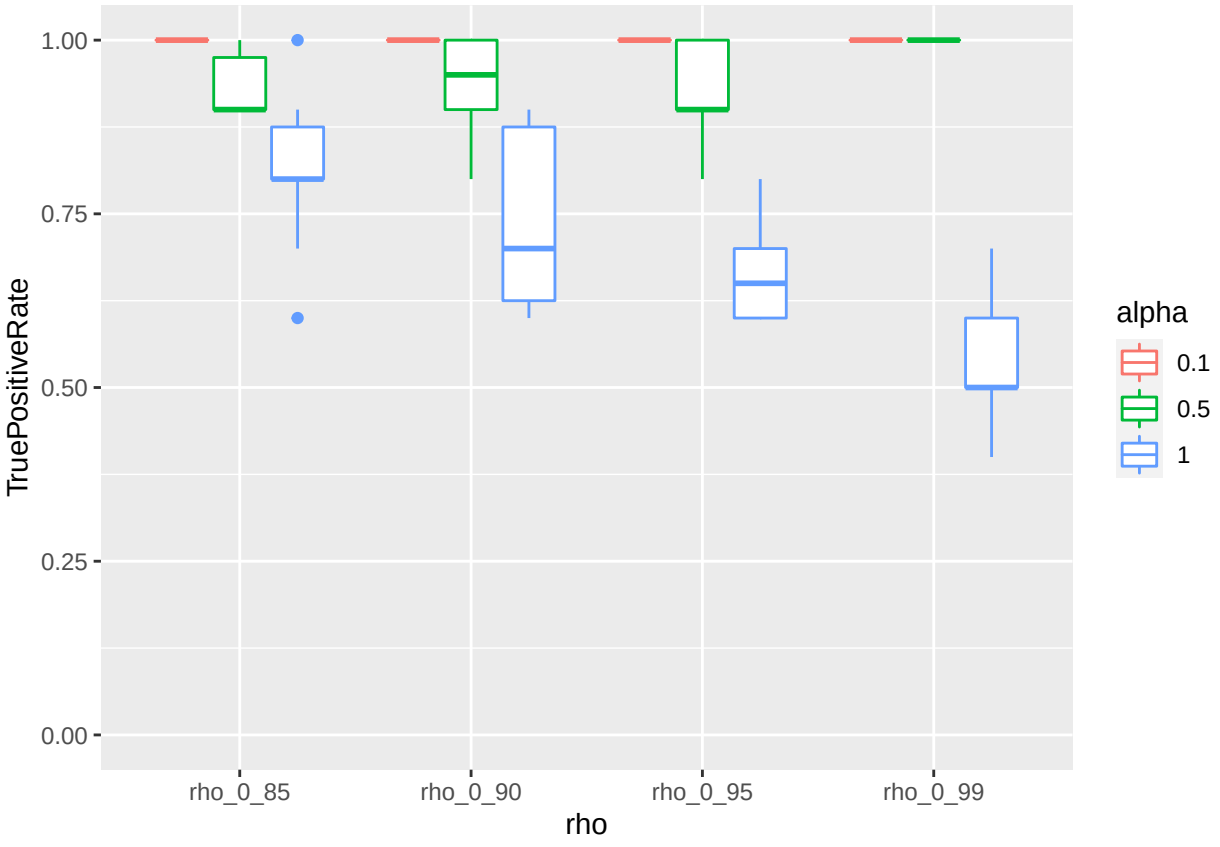
We train on 100 observations and predict for 100 out-of-sample observations. We consider $\rho \in \{0.85, 0.9, 0.95, 0.99\}$, and fix the elastic-net mixing parameter α to be one of $\alpha \in \{0.1, 0.5, 1\}$, where we note that $\alpha = 1$ corresponds to the LASSO penalty.

To assess the quality of predictions we calculate the sum-of-squares error (SSE) (shown for the out-of-sample predictions only).

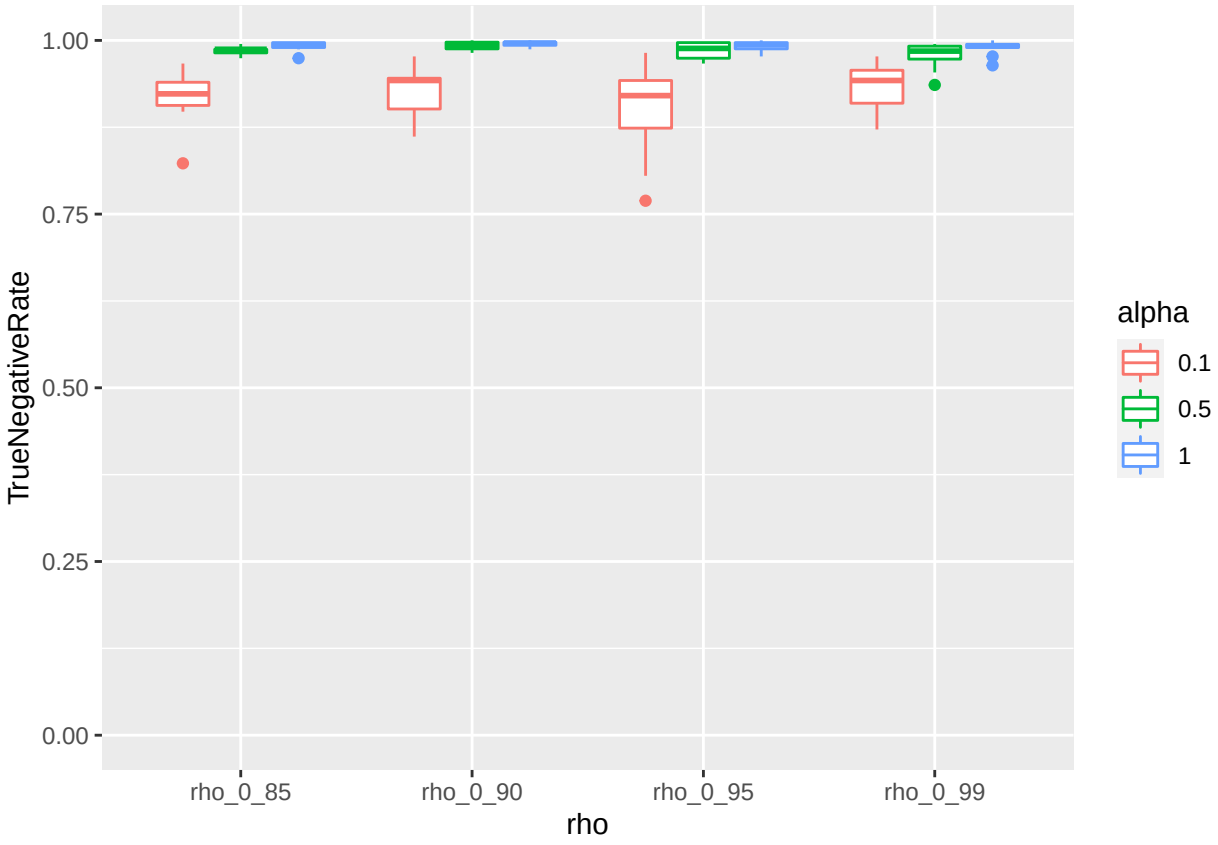
To assess the quality of the variable selection, we calculate the true positive, true negative, false positive and false negative selection rates, where – for example – the true positive rate is the proportion of selected variables that were correctly selected.

We repeat the above procedure for 10 simulated datasets, so that we can provide an assessment of the variability in the SSE, TPR, FPR, etc.

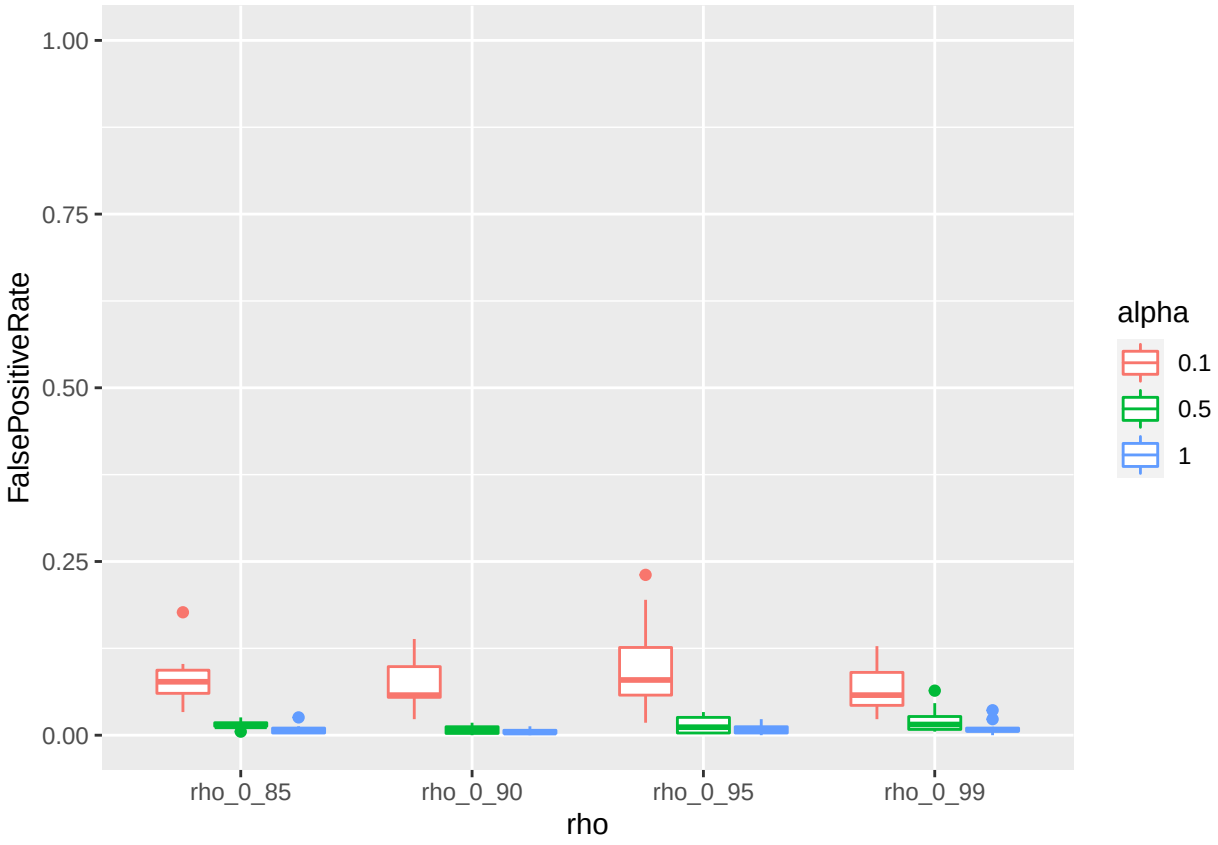
TPR



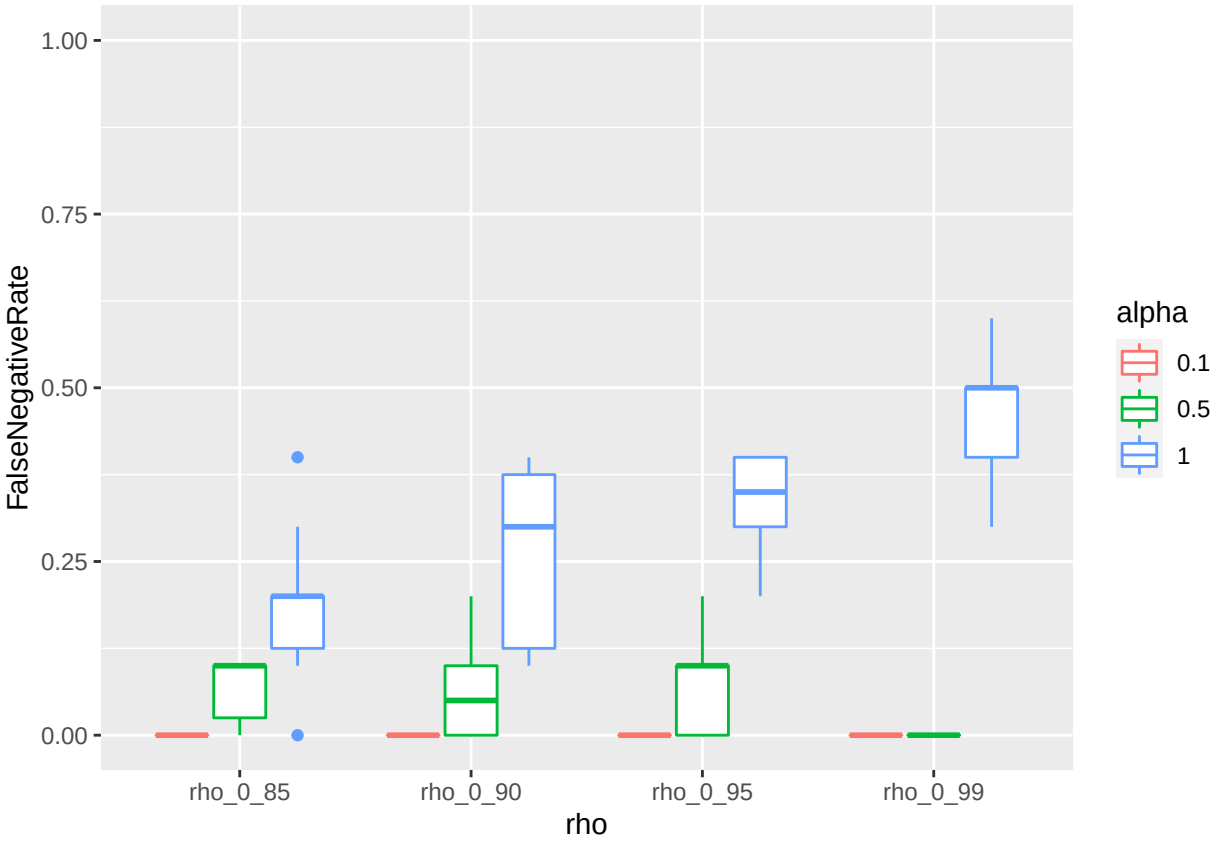
TNR


1.00
0.75
0.50
0.25
0.00
TrueNegativeRate
rho_0_85
rho_0_90
rho_0_95
rho_0_99
rho
alpha
0.1
0.5
1

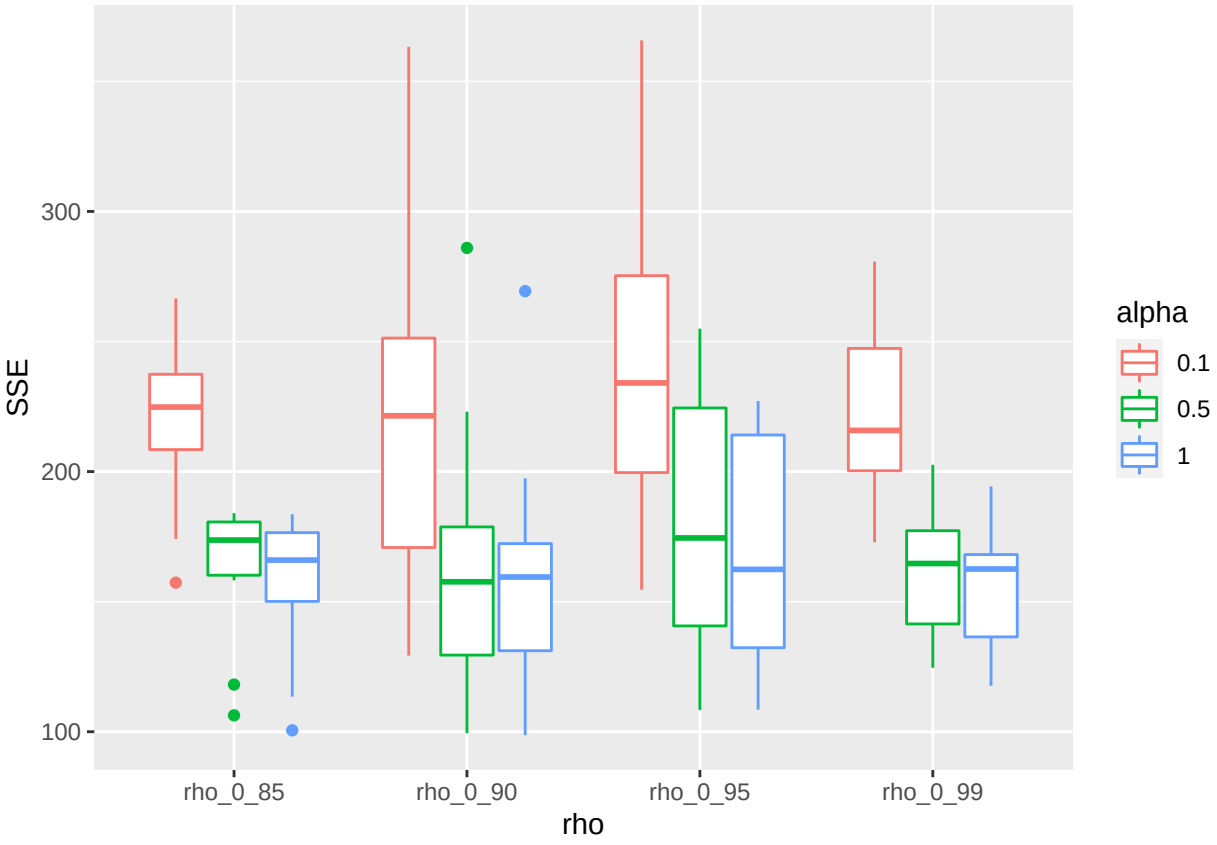
FPR


1.00
0.75
0.50
0.25
0.00
FalsePositiveRate
rho_0_85
rho_0_90
rho_0_95
rho_0_99
rho
alpha
0.1
0.5
1

FNR


1.00
0.75
0.50
0.25
0.00
FalseNegativeRate
rho_0_85
rho_0_90
rho_0_95
rho_0_99
rho
alpha
0.1
0.5
1

SSE



Summary

Overall, we can see that the elastic net with $\alpha = 0.1$ does a better job of identifying (i.e. not missing) the relevant predictors (high TPR), but that this comes at the cost of wrongly identifying some predictors as relevant (higher FPR than the LASSO, $\alpha = 1$, case). Conversely, the LASSO fails to identify some of the relevant predictors (lower TPR), but does not select any irrelevant predictors (lower FPR).

Nevertheless, the LASSO provides the best predictive performance on average (as quantified by the SSE). This is because the relevant predictors that the LASSO is failing to select are strongly correlated with relevant predictors that the LASSO does select. Hence the LASSO is providing a minimal model (smallest number of predictors) that provide the best predictive accuracy.

Whether we should prefer the LASSO or elastic net therefore depends on whether we are mainly interested in identifying a minimal model that provides good predictive performance, or if we care more about identifying all relevant predictors (potentially at the cost of including some irrelevant predictors among our selections). By tuning the α parameter, we can also find intermediate solutions between these two extremes.

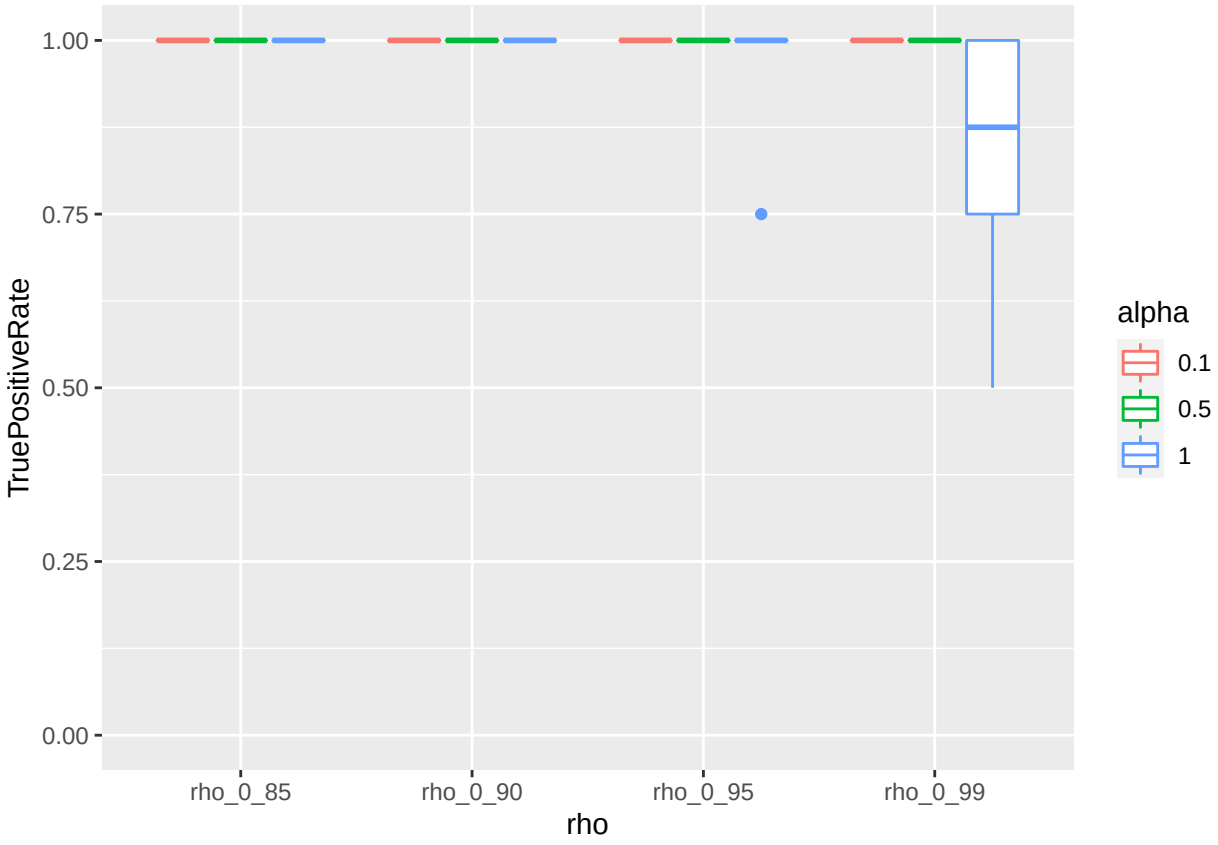
Results (2)

We now consider a second simulation scenario. Our predictors are simulated as before (i.e. we have the same correlation structure between the predictors), but we now generate the response as follows:

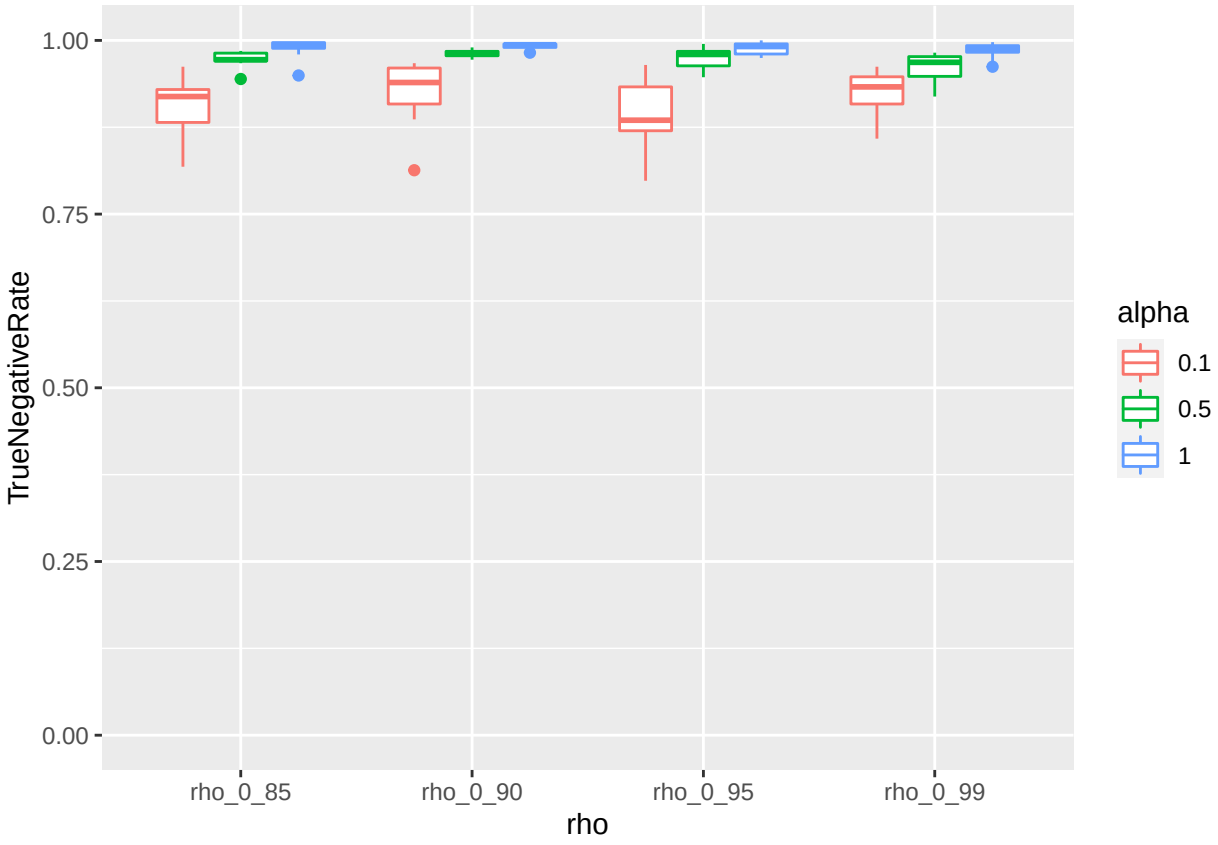
$$Y_i = X_{i1} + X_{i2} + X_{i4} + X_{i7} + \epsilon, \tag{6}$$

where $\epsilon \sim \mathcal{N}(0, 1)$. Thus, this time, we have only one member from each correlated set of predictors appearing in the model. We repeat the same analysis as previously.

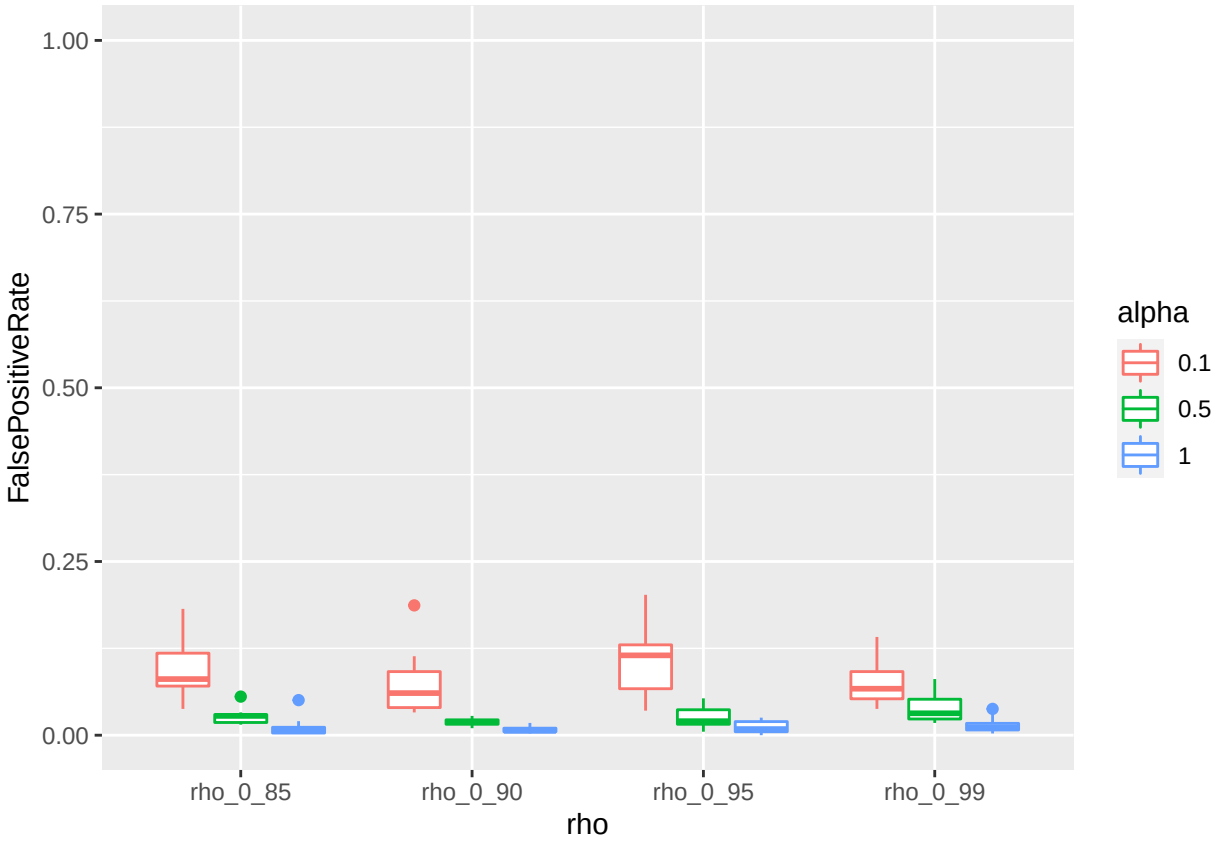
TPR


1.00
0.75
0.50
0.25
0.00
TruePositiveRate
rho_0_85
rho_0_90
rho_0_95
rho_0_99
rho
alpha
0.1
0.5
1

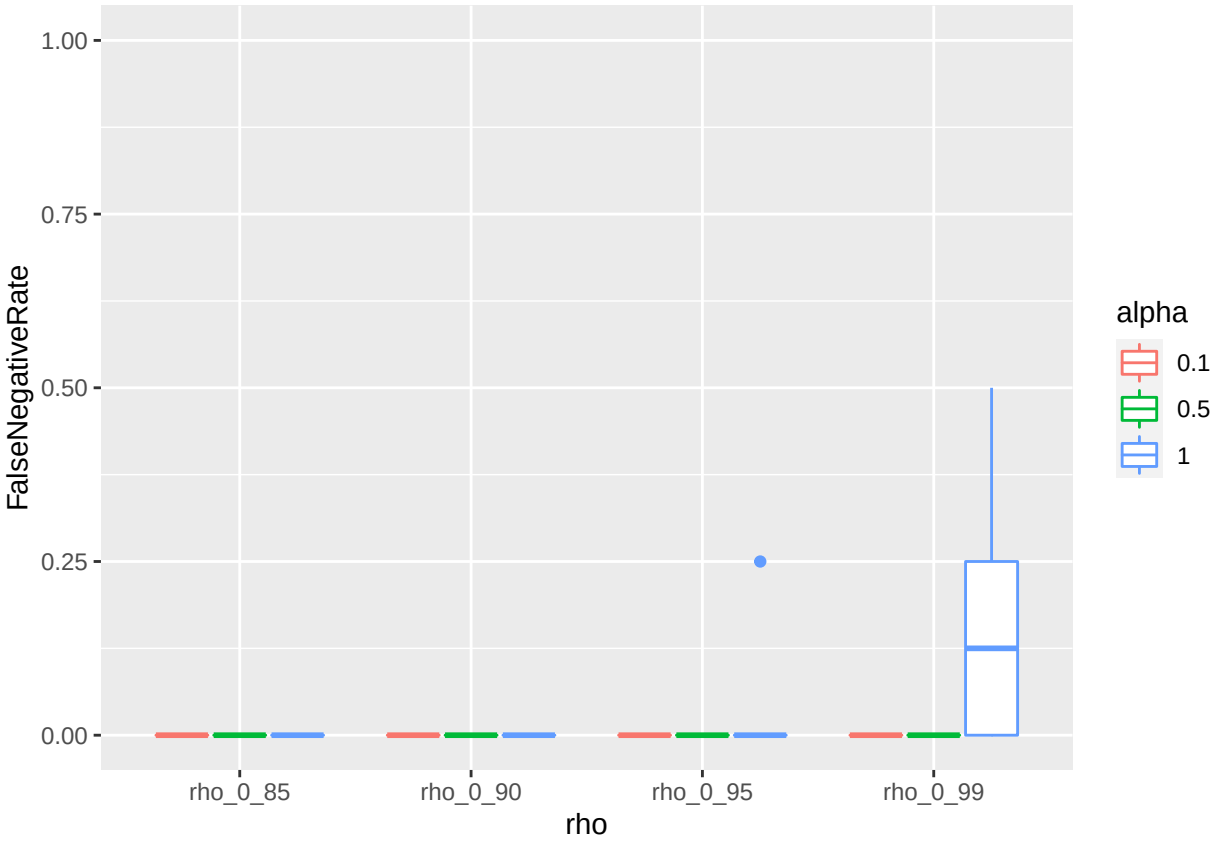
TNR


1.00
0.75
0.50
0.25
0.00
TrueNegativeRate
rho_0_85
rho_0_90
rho_0_95
rho_0_99
rho
alpha
0.1
0.5
1

FPR


1.00
0.75
0.50
0.25
0.00
FalsePositiveRate
rho_0_85
rho_0_90
rho_0_95
rho_0_99
rho
alpha
0.1
0.5
1

FNR



Summary

The results are similar to previously, but now the LASSO ($\alpha = 1$ case) seems to do better than before in terms of TPR. This is because of the LASSO's tendency to pick out just one representative from a set of correlated predictors – which, in this case, is the correct strategy. However, for particularly large ρ, we can see that this can result in irrelevant predictors being selected (in cases where the "wrong" representative is selected from the correlated set).