Imperial College London

Regularisation in statistics and Machine Learning

Guillem HURAULT

January 24, 2019

Theory

Background

Linear regression

$$Y = X\beta + \epsilon$$

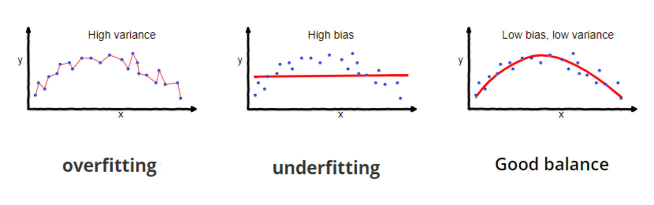
- Y the vector of response ($N \times 1$)
- X the design matrix ($N \times p$)
- β the vector of parameters ($p \times 1$)
- ϵ the vector of errors ($N \times 1$)

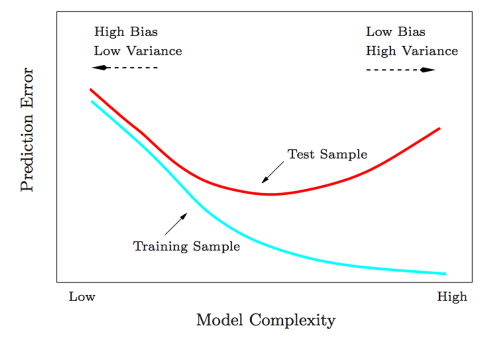
Ordinary Least Squares (OLS)

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} ||Y - X\beta||_2^2$$

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

Problem: overfitting





Solution: Penalise the coefficients

Ridge (L_2, Tikhonov) regularisation

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} ||Y - X\beta||^2 + \lambda_2 ||\beta||_2^2$$
$$\hat{\beta} = (X^t X + \lambda_2 I)^{-1} X^t Y$$

Equivalent to OLS with constraint $||\beta||_2^2 < t$

- Encourage grouping of highly correlated variables (multicollinearity)
- Strong shrinkage

Solution: Penalise the coefficients

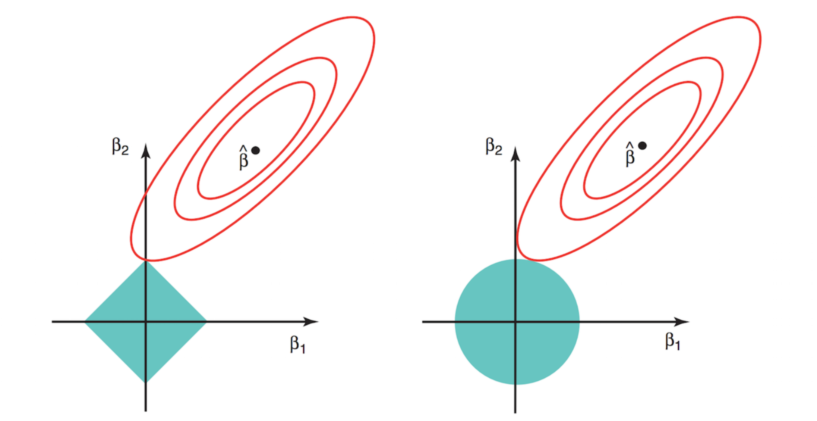
Lasso (L_1) regularisation

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \, ||Y - X\beta||^2 + \lambda_1 |\beta|$$

Equivalent to OLS with constraint $|\beta| < t$

- Encourage sparse model (set coefficients to 0)
- Smaller shrinkage compared to ridge
- Tends to select variable randomly in the presence of multicollinearity

Geometric interpretation

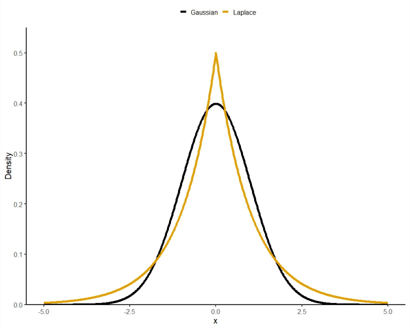


Bayesian interpretation

Bayes' theorem: $p(\theta|x) \quad = \quad \frac{p(x|\theta)\,p(\theta)}{p(x)} \propto \quad p(x|\theta)\,p(\theta)$

- $||Y - X\beta||^2 \propto$ Gaussian log-likelihood
- $\lambda_2||\beta||_2^2 \propto$ log prior of a Gaussian distribution: $\beta|\sigma^2 \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda_2})$
- $\lambda_1|\beta| \propto$ log prior of a Laplace distribution: $\beta|\sigma \sim \text{Laplace}(0, \frac{\sigma}{\lambda_1})$

Hence, $\hat{\beta}$ is the **Maximum A Posteriori** (MAP) estimate



Other methods

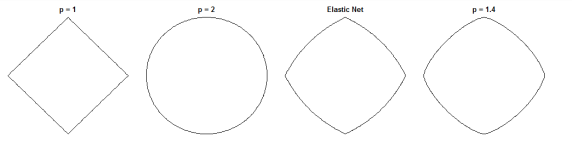
Elastic Net: Mixture of L_1 and L_2

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \, ||Y - X\beta||^2 + \lambda_2 ||\beta||_2^2 + \lambda_1 |\beta|$$

- Sparsity of the lasso
- Robust to multicollinearity as in ridge

Bridge penalty: L_p regularisation

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \, ||Y - X\beta||^2 + \lambda_p ||\beta||_p^p$$



Overshrinkage

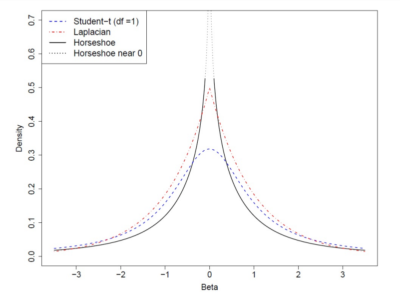
- Correcting for the double shrinkage in Elastic Net: $\hat{\beta}^{\text{new}} = (1 + \lambda_2)\hat{\beta}$
- Hybrid Lasso: Lasso followed by OLS
  1. Apply Lasso for variable selection
  2. Apply OLS on the subset of predictors selected by the Lasso
- Relaxed Lasso
  1. Apply Lasso for variable selection
  2. Apply Lasso on the subset of predictors selected by the Lasso
- Horseshoe

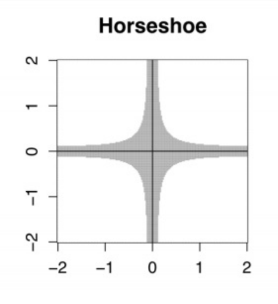
The Horseshoe

Bayesian linear regression $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$ with the horseshoe prior:

$$\begin{cases} \beta_i | \lambda_i, \tau & \sim \mathcal{N}(0, \lambda_i^2 \tau^2) \\ \lambda_i & \sim C^+(0,1) \end{cases}$$

- λ_i are the local shrinkage parameters
- τ is the global shrinkage parameter

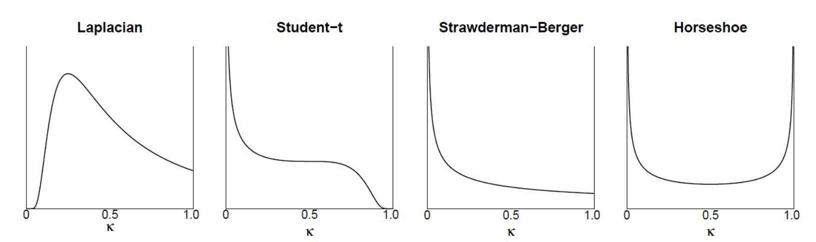




Shrinkage profile

$\kappa_i = \frac{1}{1+\lambda_i^2}$ is the random shrinkage coefficient

- $\kappa_i = 0$: no shrinkage (full signal)
- $\kappa_i = 1$: total shrinkage (no signal)



Regularised Horseshoe

- Set a prior for τ with a prior guess p_0 for the number of non-zero coefficients

$$\tau|\sigma \sim C^{+}(0, \frac{p_0}{D - p_0}\frac{\sigma}{\sqrt{N}})$$

- Specify the shrinkage with a prior guess s on the scale of the signal

$$\beta_i|\lambda_i, \tau, c \sim \mathcal{N}(0, \tilde{\lambda}_i^{2}\tau^2)$$

$$\tilde{\lambda}_i^{2} = \frac{c^2\lambda_i^2}{c^2 + \tau^2\lambda_i^2}$$

$$c \sim \text{Student-t}_{\nu}(0, s^2)$$

Case study

Presentation

Objective

Compare the different regularisation methods in terms of coefficient estimation and predictive power, by investigating:

- Different patterns of β
- Multicollinearity
- Different SNR

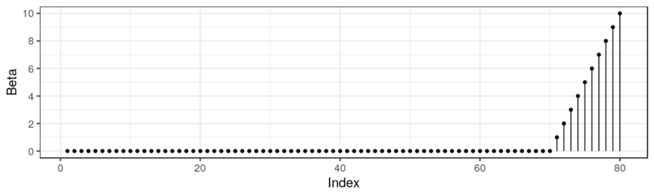
Toy data

$$Y \sim \mathcal{N}(X\beta, \sigma^2 I)$$

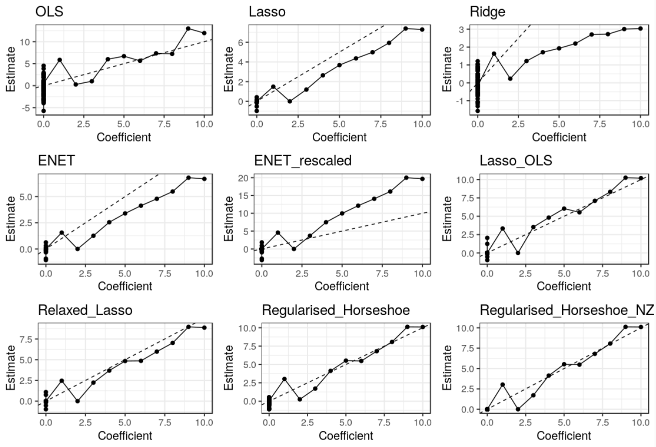
- $N_{\text{train}} = 100$ observations for training
- $N_{\text{test}} = 1000$ observations for testing
- $p = 80$ features

First condition

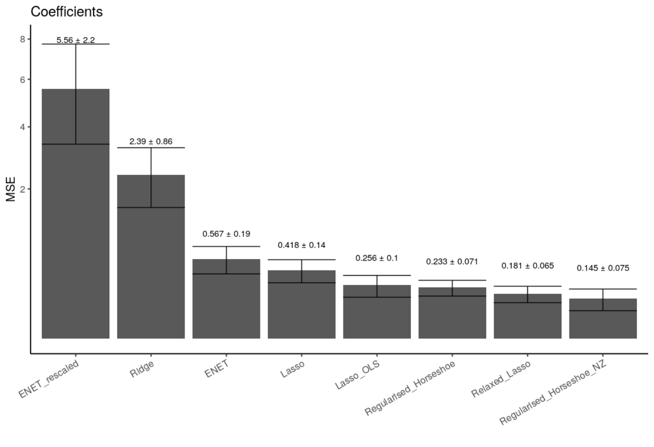
- No multicollinearity (e.g. principal components)
- $SNR = 2$



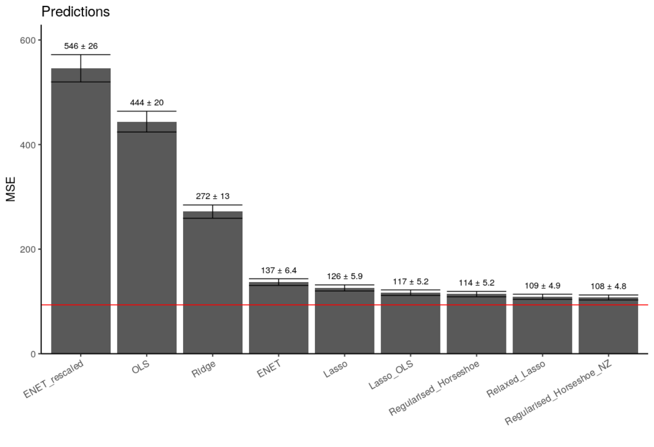
First condition



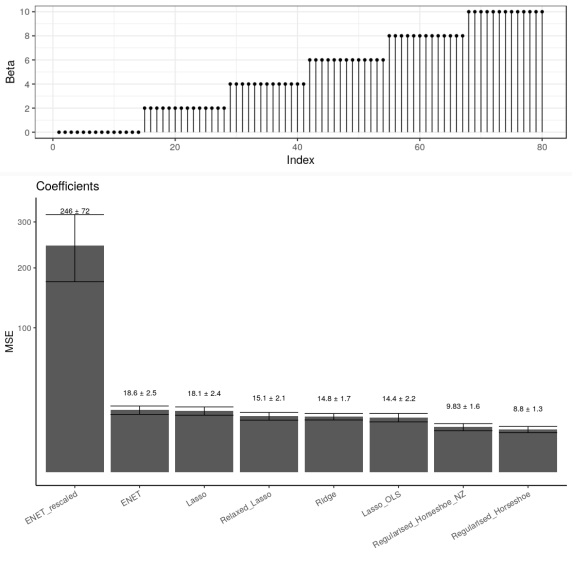
First condition



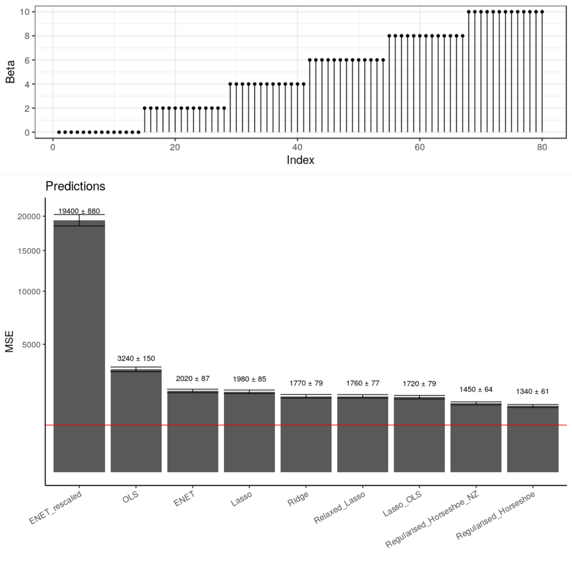
First condition



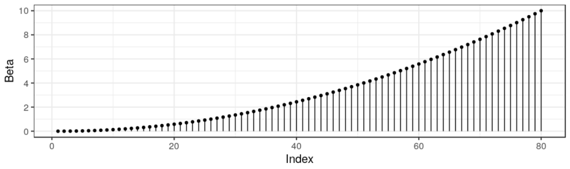
Changing the pattern of β

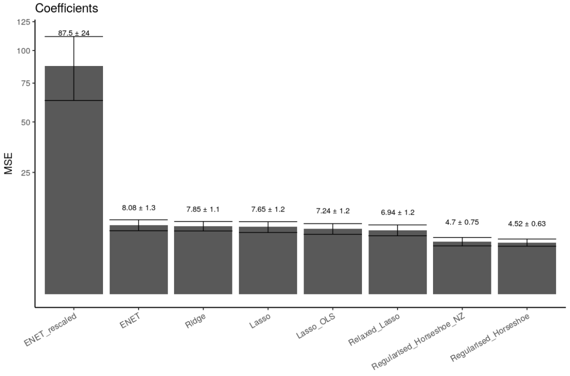


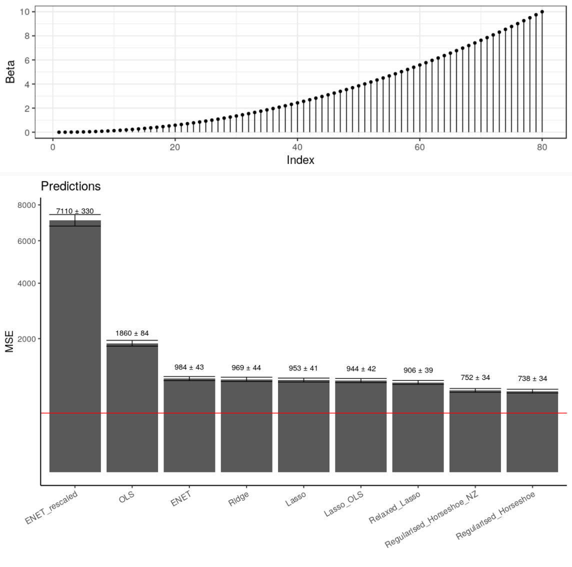
Changing the pattern of β



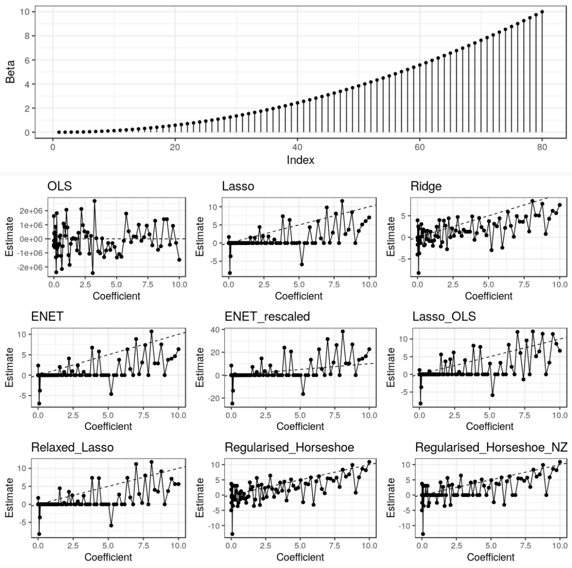
Changing the pattern of β



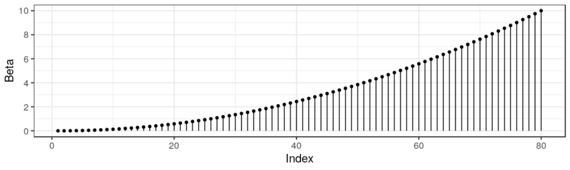


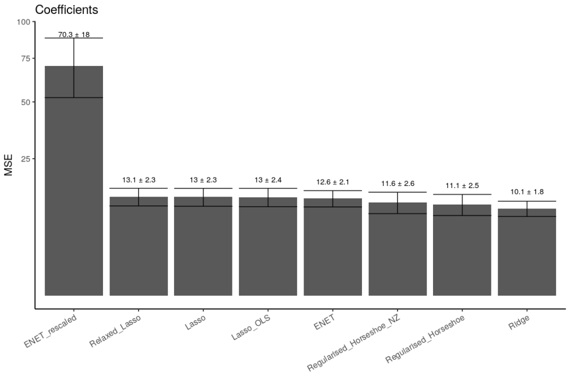

Beta
Index
Predictions
MSE
7110 ± 330
1860 ± 84
984 ± 43
969 ± 44
953 ± 41
944 ± 42
906 ± 39
752 ± 34
738 ± 34
ENET_rescaled
OLS
ENET
Ridge
Lasso
Lasso_OLS
Relaxed_Lasso
Regularised_Horseshoe_NZ
Regularised_Horseshoe

Multicollinearity in predictors

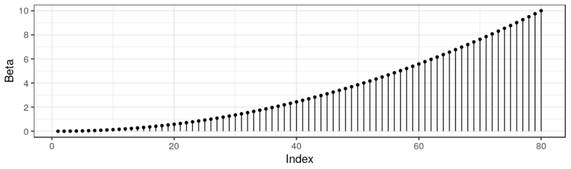


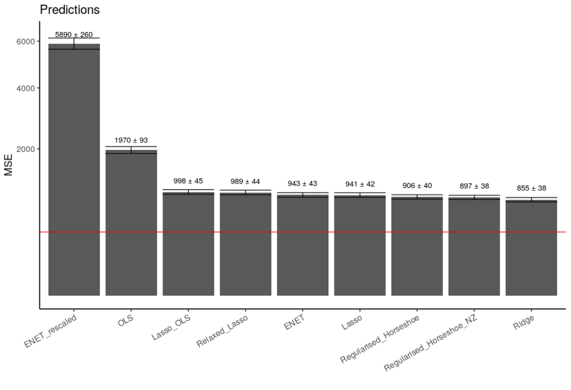
Multicollinearity in predictors



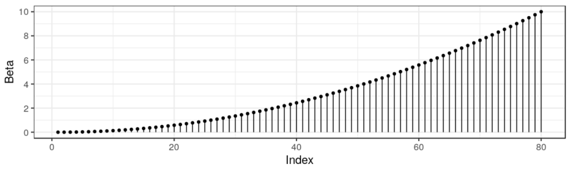


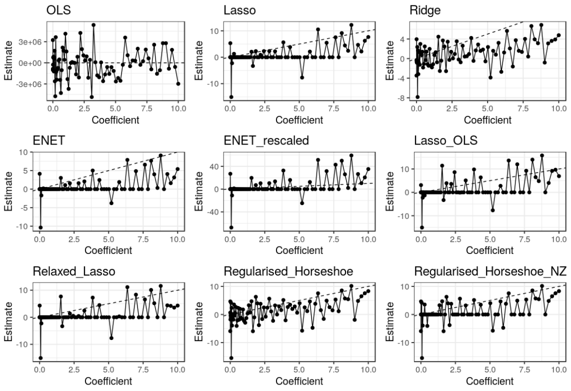
Multicollinearity in predictors



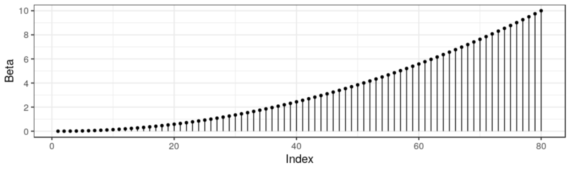


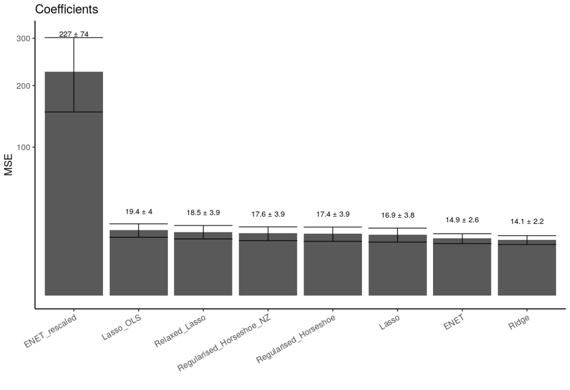
Multicollinearity, $SNR = 1$



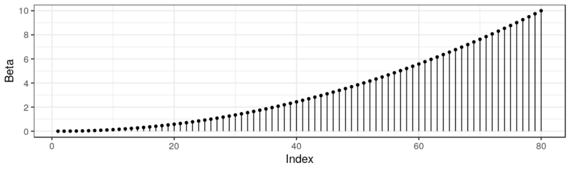


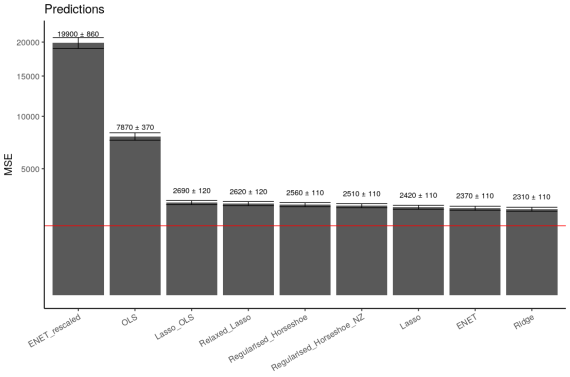
Multicollinearity, $SNR = 1$



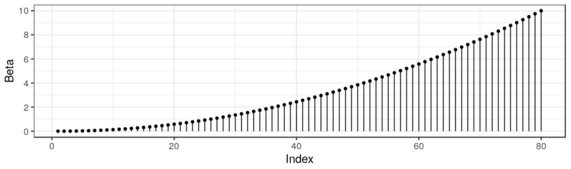


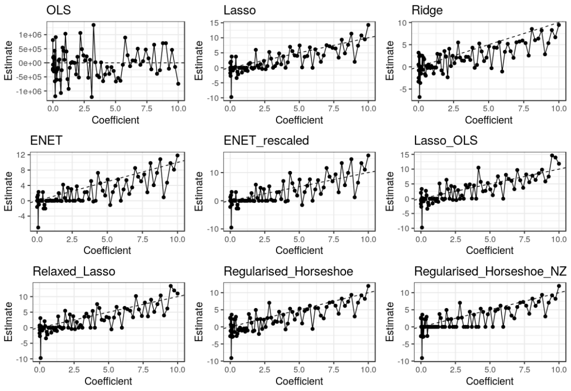
Multicollinearity, $SNR = 1$



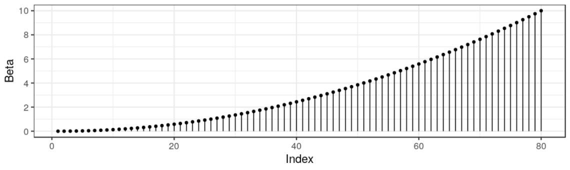


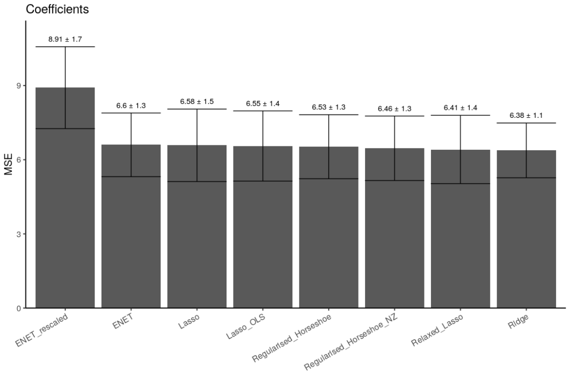
Multicollinearity, $SNR = 4$



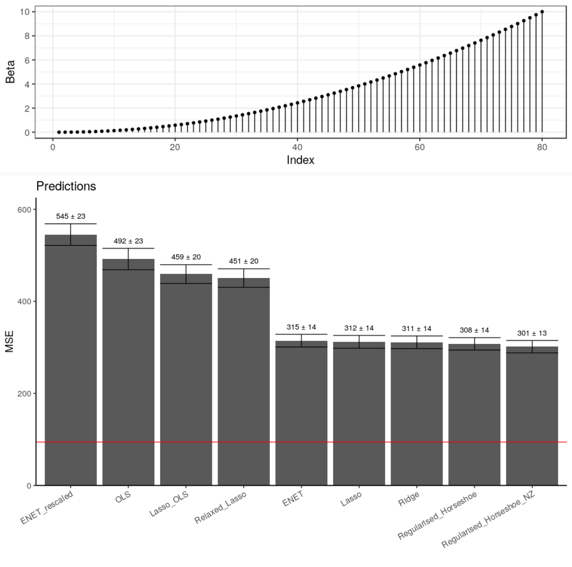


Multicollinearity, $SNR = 4$



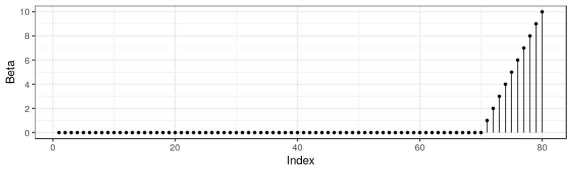


Multicollinearity, $SNR = 4$

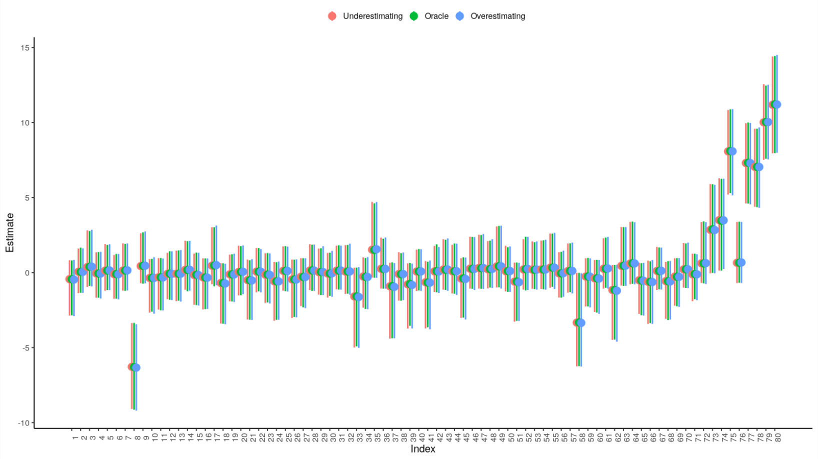


Prior number of relevant parameters for the horseshoe

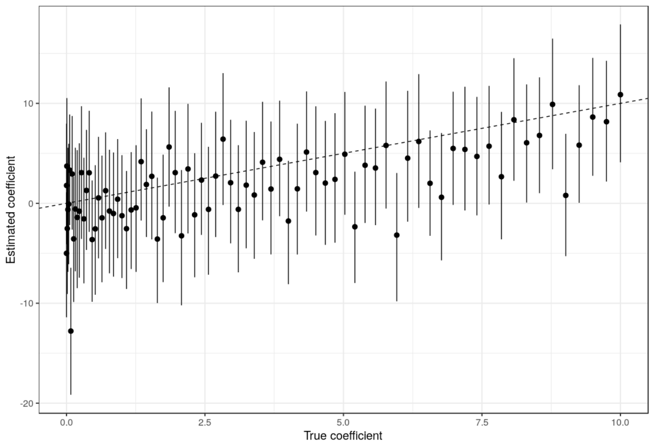
- So far, I assumed an oracle guess for the horseshoe: $p_0 = K$, the true number of non-zero features.
- What if the prior is wrong?
  - $p_0 = \frac{K}{2}$ (underestimating)
  - $p_0 = 2K$ (overestimating)
- Let's assume:
  - Multicollinearity
  - $SNR = 2$



Prior number of relevant parameters for the horseshoe



Uncertainty estimates for the horseshoe



Conclusion

Conclusion (1)

About the Lasso

- Hybrid Lasso or Relaxed Lasso outperforms simple lasso
- Lasso-based regularisation seems the best option when the underlying model is sparse...

Conclusion (1)

About the Lasso

- Hybrid Lasso or Relaxed Lasso outperforms simple lasso
- Lasso-based regularisation seems the best option when the underlying model is sparse...

- ... But the patterns of β shouldn't influence much the choice of regularisation
- Similarly, the SNR shouldn't influence much the choice of regularisation
- However, multicollinearity is important

Conclusion (2)

About Ridge/Elastic Net

- Ridge outperforms Lasso in the presence of multicollinearity
- Elastic Net seems like a good compromise between Lasso and Ridge
- Rescaling Elastic Net coefficient is usually a bad idea
- Relaxed/Hybrid Elastic Net ?

Conclusion (3)

About the horseshoe

- Horseshoe is in the top regardless of the situation
- A bad guess for the number of relevant parameters for the horseshoe has little effect
- Horseshoe can provide uncertainty estimates