Aalto-yliopisto
Perustieteiden
korkeakoulu

A Robust Optimisation Approach for Stable Linear Regression (thesis presentation)

Veera Wilkki

16.6.2023

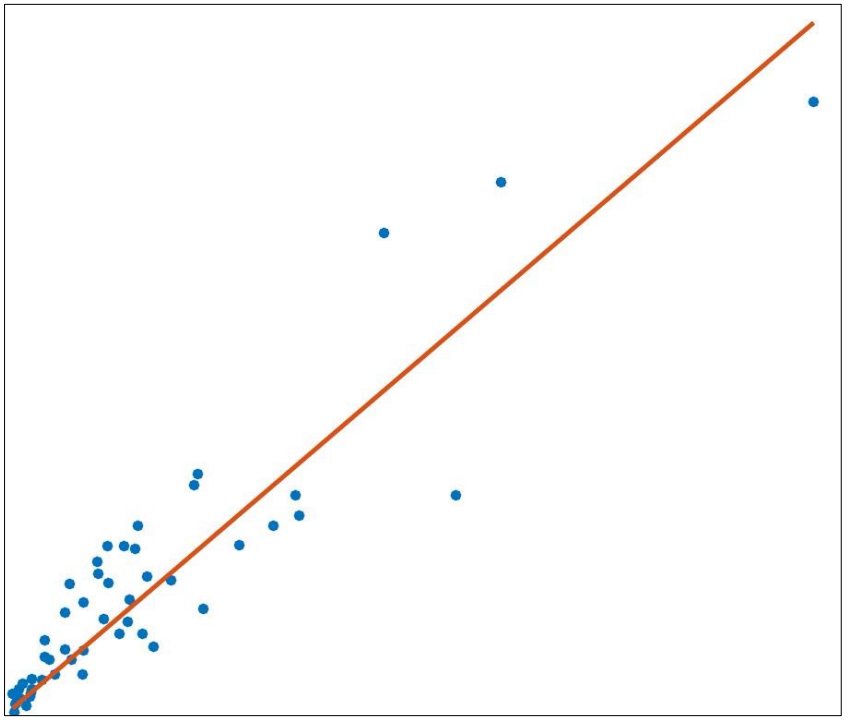
Instructor: *Paula Weller*

Supervisor: *Fabricio Oliveira*

Systeemianalyysin
laboratorio

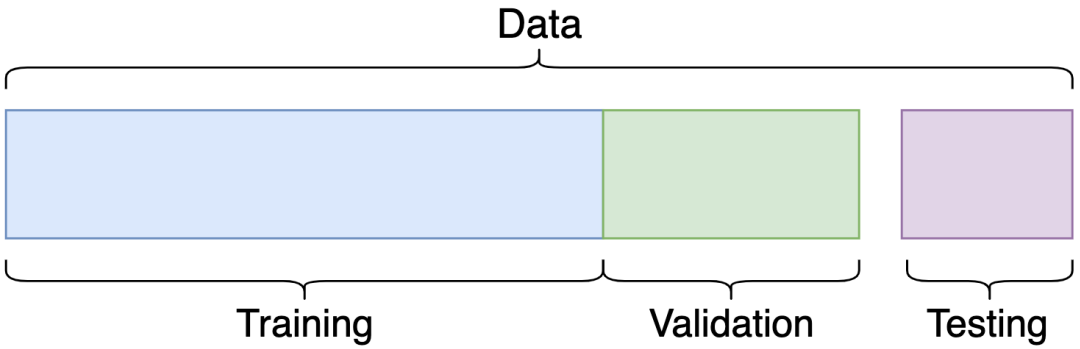
Background: training linear regression models

- Linear regression: method used to model the linear relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data



Background: training linear regression models

- Usually trained by assigning data to training and validation sets **randomly**
  1. Choose a random subset as the testing set
  2. Split the rest of the data into a training set and a validation set
  3. Test the final accuracy of the model using the held out testing set



Training linear regression models using robust optimisation

- Instead of splitting data randomly, this step can be integrated into the optimisation problem directly as presented in Bertsimas and Paskov (2020):

$$\min_{\beta} \max_{z \in \mathcal{Z}} \sum_{i=1}^{n} z_i |y_i - x_i^T \beta| + \lambda \sum_{i=1}^{p} \Gamma(\beta_i)$$

$$\text{with} \quad \mathcal{Z} = \left\{ z : \sum_{i=1}^{n} z_i = k, \quad z_i \in \{0, 1\} \right\}$$

where z_i is an indicator variable, indicating which point (x_i, y_i) belongs to the training set and which to the validation set, λ is a regularisation parameter, $\Gamma(\cdot)$ is the regularisation function and k represents the desired proportion between the size of the training and validation sets

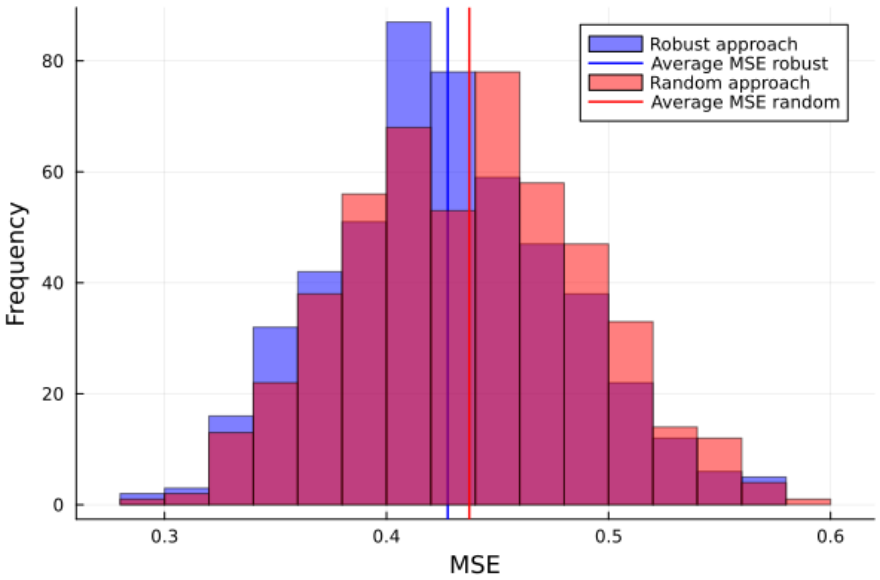
Experimental setup

- Aim is to examine predictive ability and runtime when training linear regression models using the robust approach vs. the regular randomised approach with two regularisation methods: Lasso and Ridge
- Predictive ability is examined through the mean squared error (MSE)
- Testing is conducted using three functions created in Julia:
  - The first function performs training using robust approach
  - The second function performs training using randomised approach
  - The third function uses one of the previous two to train regression models 500 times using a different testing set each time. The function then uses these testing sets to calculate MSE values.

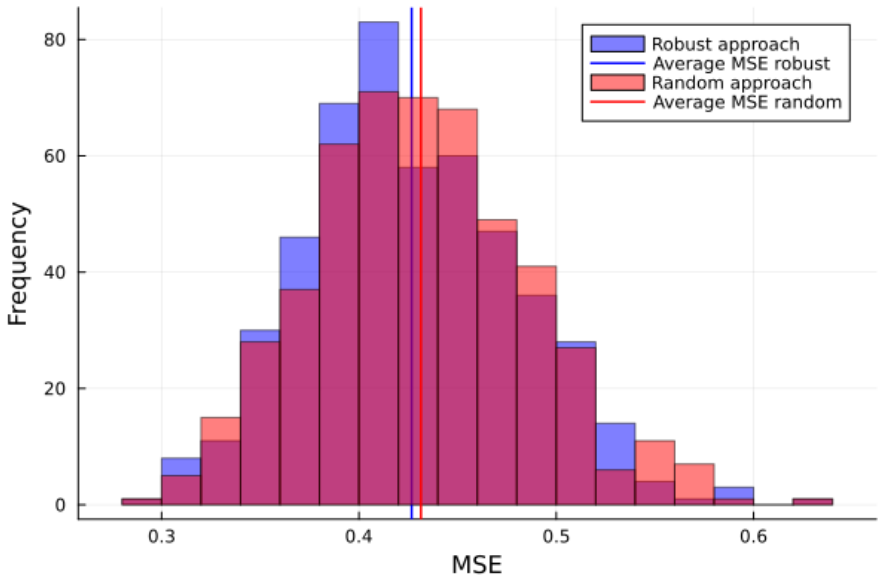
Datasets

- Tests are performed using two datasets from the UCI Machine Learning Repository:
  1. Red Wine Quality (1599 data points)
  2. White Wine Quality (4898 data points)
- Both have 12 variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and quality
- Aim is to use linear regression to predict the value of the last variable, i.e. the quality score of a wine (integer between 0 and 10), based on the remaining variables

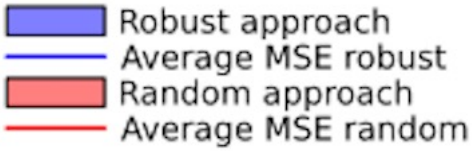
Results: predictive ability for red wine



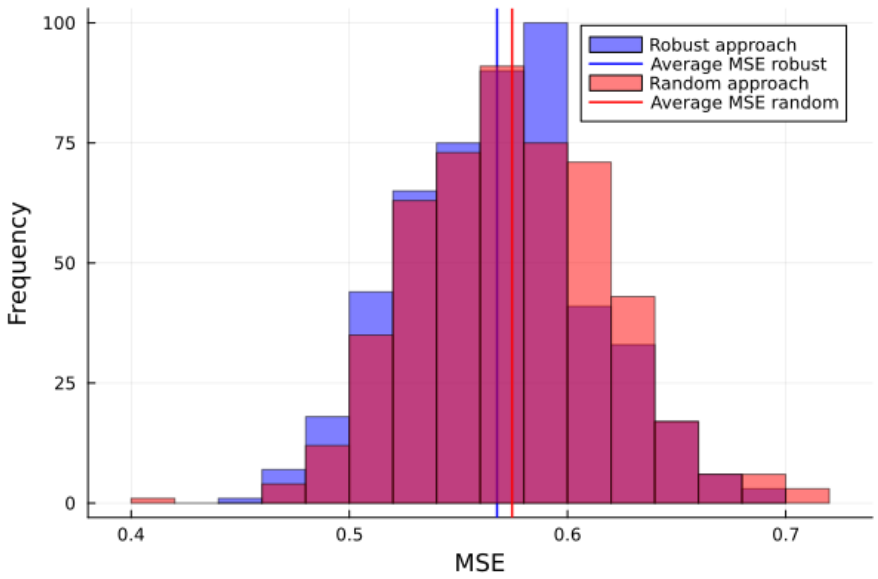
Lasso regularisation



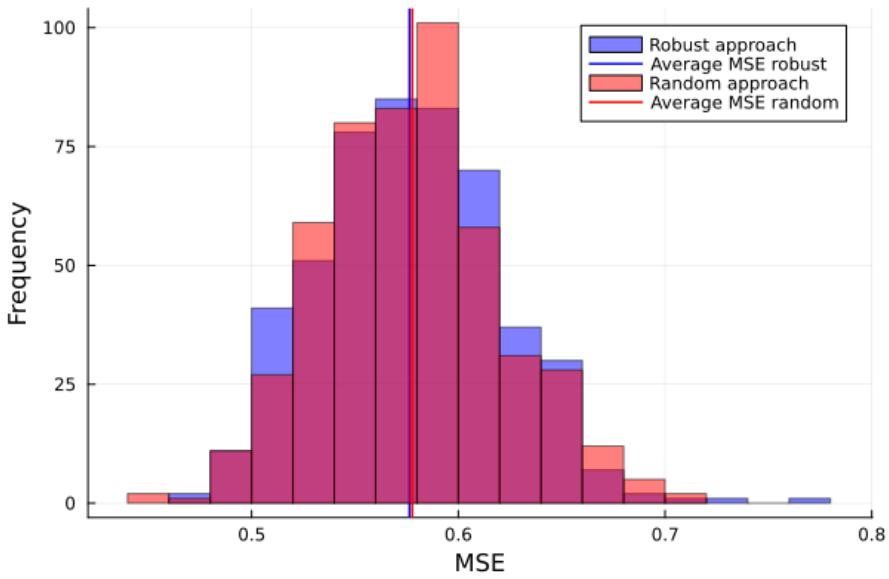
Ridge regularisation



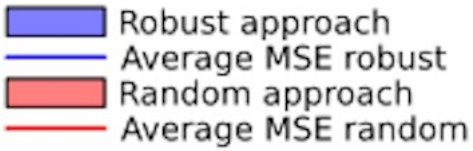
Results: predictive ability for white wine



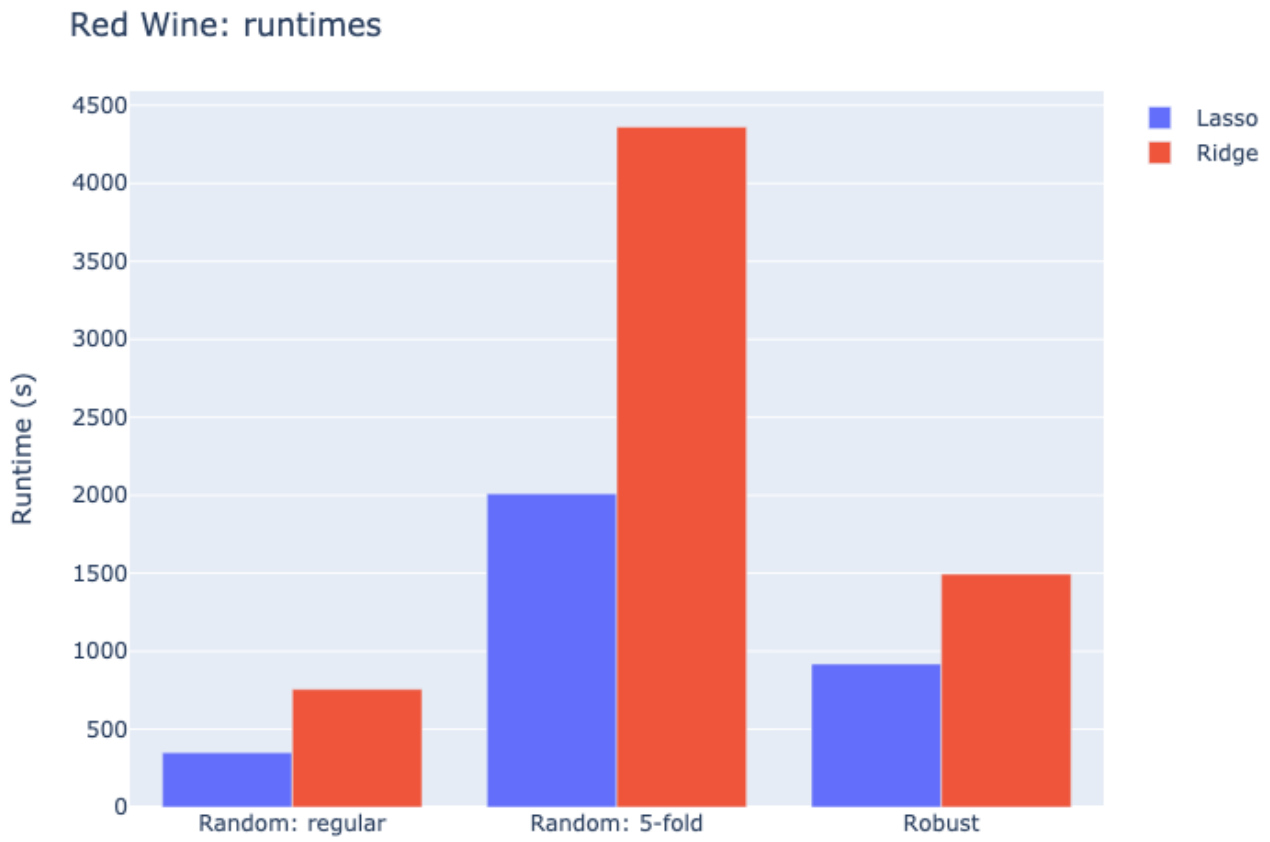
Lasso regularisation



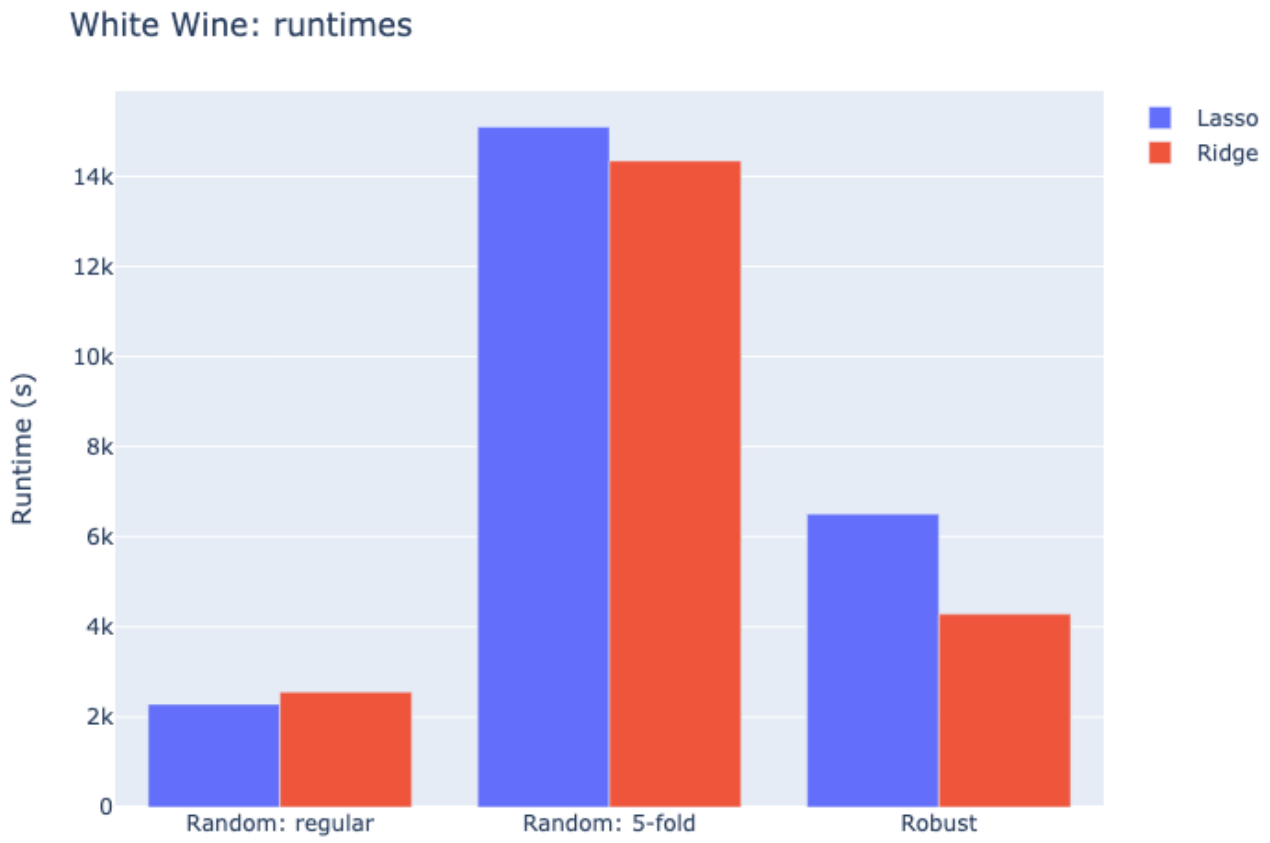
Ridge regularisation



Results: runtimes for red wine



Results: runtimes for white wine



Conclusions

- The results demonstrate that:
  - predictive ability is improved using the robust approach
  - using the robust approach leads to significantly shorter runtimes compared to the randomised approach with 5-fold cross validation
    - Very useful as training is not often performed without cross validation
- In addition to these benefits proven through testing, using the robust approach also allows for the identification of the ”hardest subpopulation” as previously explained
- This robust approach could possibly become the new default method for training linear regression models

References

Bertsimas, D., & Paskov, I. (2020). Stable regression: On the power of optimization over randomization in training regression problems. *The Journal of Machine Learning Research*, *21*(1), 9374-9398.