Applied Machine Learning in Health Sciences 2023
-
Regularization

Peter Mondrup Rasmussen

CFIN

pmr@cfin.au.dk

Regularization

Ridge regularization
Lasso regularization

Regularization

- When fitting models we often seek to maximize the agreement between the model's prediction and the observed data, e.g. by least squares estimation, maximum likelihood estimation.
- However, alternative fitting approaches may be preferred
  - ***Prediction accuracy*:** More constrained/biased models may be preferrable when fitting models on a limited amount of training data.
  - ***Model interpreation:*** Attempt to identify important features by e.g. *variable* selection or *feature selection.*
- ***Feature selection:*** Try to idenfity a subset of the the predictors that we believe to be relavant to the model.
- ***Regularization (Shrinkage)*:** Impose constraints on the estimated model parameters.
- ***Dimension reduction*:** *Project* predictors onto a subspace and fit the model based on the new represetation.

Regularization – Ridge penalty

- Linear regression uses the least squares fitting procedure to minimize

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 + \sum_{j=1}^{p}\beta_j x_j\right)^2$$

- In *ridge regression* we *constrain* or *regularize* the coefficient estimates towards zeros by finding parameter estimates $\hat{\beta}^R$ that minimize

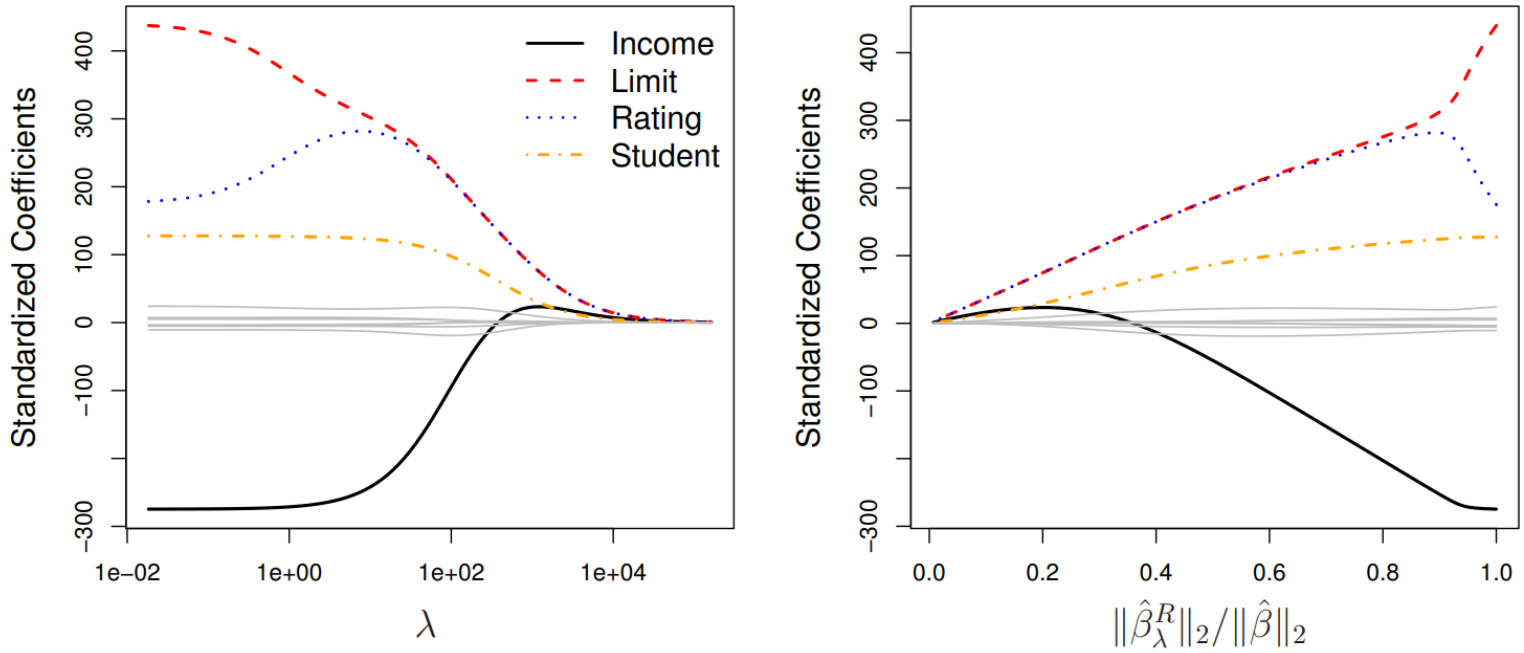
$$\sum_{i=1}^{n}\left(y_i - \beta_0 + \sum_{j=1}^{p}\beta_j x_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

- The extra term is a *regularization/penalty term*, and λ is a *regularization parameter* that needs to be selected. The objective to be minimized now consist of two (often competing) term. The first term quantifies the model's fit and the second term quantifies model flexibility/complexity.

Regularization – Ridge penalty

$$\sum_{i=1}^{n}\left(y_i - \beta_0 \sum_{j=1}^{p} \beta_j x_j\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Different fits are produced depending on the choice of λ.

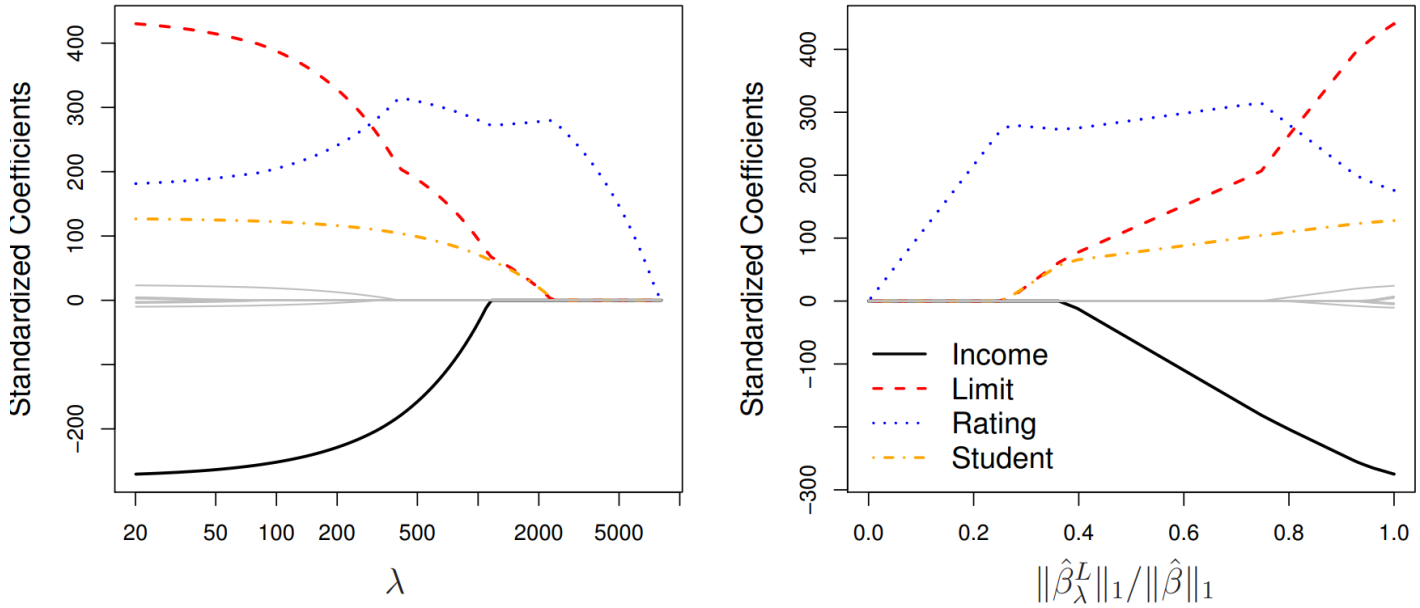


Regularization – Lasso penalty

- The Lasso penalty is similar to the ridge penalty

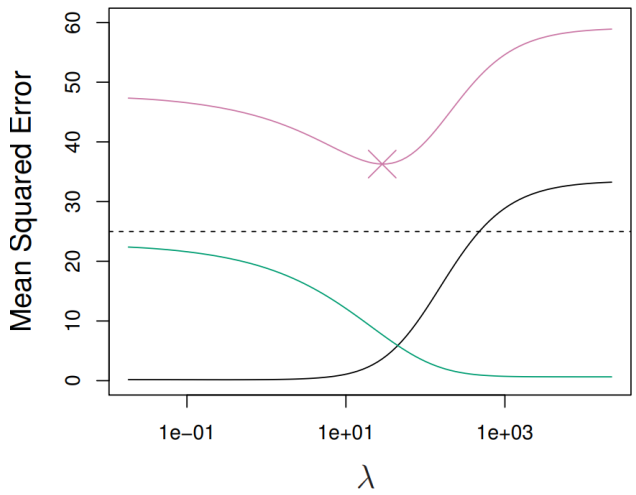
$$\sum_{i=1}^{n}\left(y_i - \beta_0 \sum_{j=1}^{p}\beta_j x_j\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j|$$

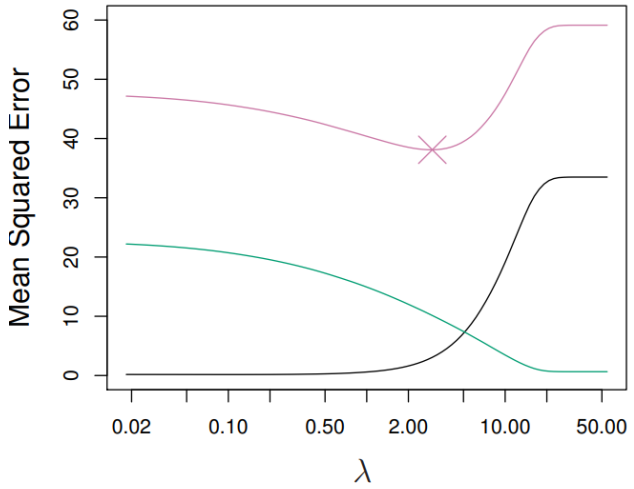
- However, the Lasso penalty performs quite differently and can provide *sparse solutions*, where some coefficients are exactely 0.



Regularization – Bias Variance Trade-Off

- Regularization may lead to increased generalization preformance.
- Regularization is a way to control the Bias-Variance Trade-Off. Increasing λ will increase bias but at the same time variance is reduced. In combination, this may result in a reduced test error.

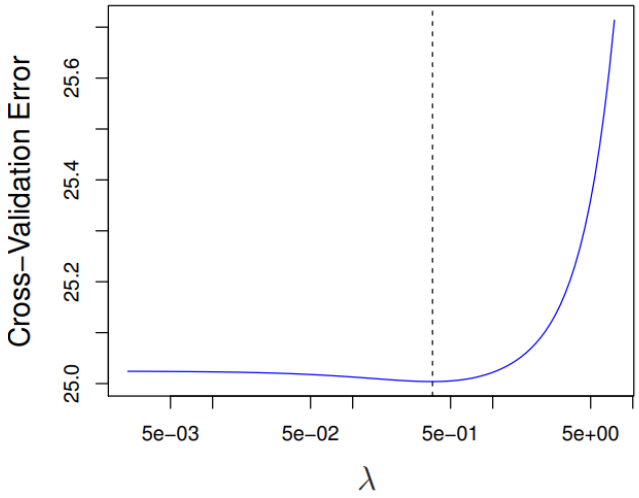


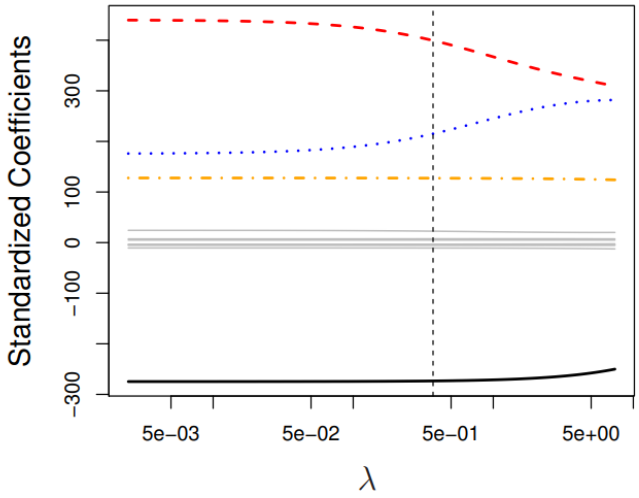


Lhs: Ridge penalty, rhs: Lasso penalty. Squared bias (black), variance (green,), test error(purple)

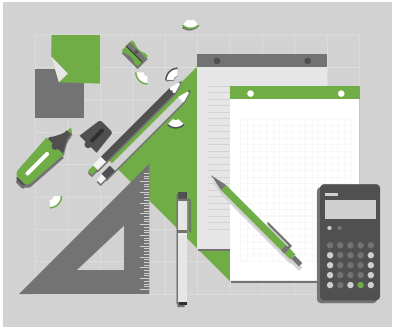
Regularization

- The regularization parameter λ needs to be selected. This is typically done by cross-validation.





- The relative scale of predictors has an influence on the coefficient estimates when using regularization. Common practice to *standardize* predictor variables before fitting regularized models.
- Ridge regularization and Lasso regularization can also be used in classification with, for example, the *logistic regression* model.



In-class exercises

6 Regularization

References

Figures from James et al. *An Introduction to Statistical Learning*, second edition, https://www.statlearning.com/resources-second-edition