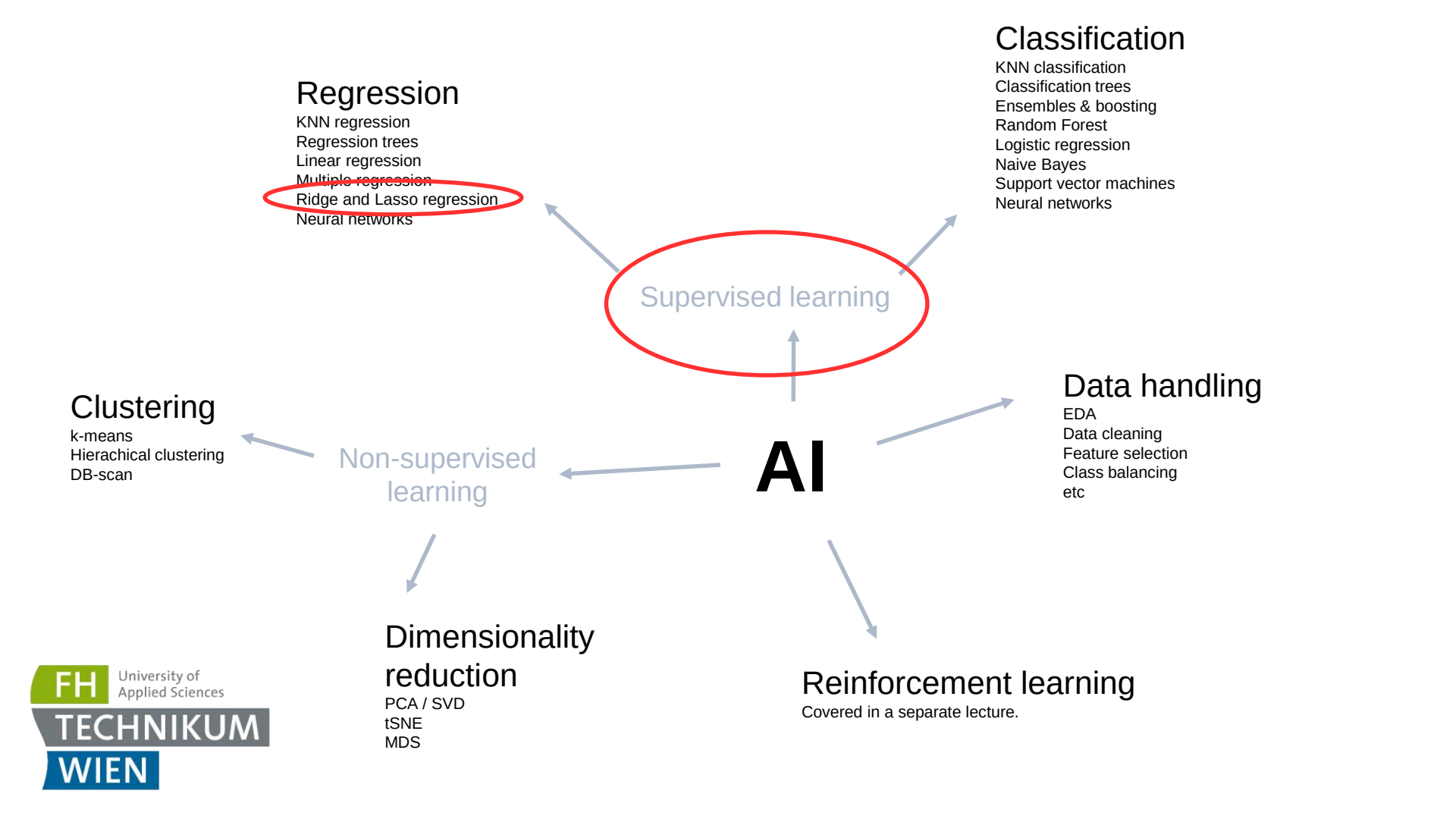
Regularization Methods for Regression (Ridge & Lasso)

ML2: AI Concepts and Algorithms (SS2025)
Faculty of Computer Science and Applied Mathematics
University of Applied Sciences Technikum Wien

FH University of Applied Sciences TECHNIKUM WIEN

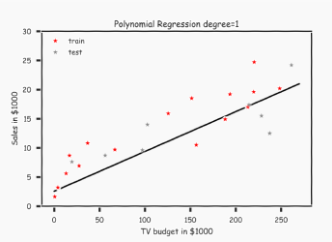
Lecturer: Rosana de Oliveira Gomes
Author: M. Blaickner, B. Knapp, S. Rezagholi, R.O. Gomes



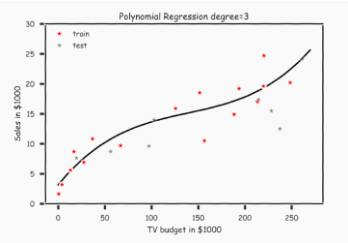
FH University of Applied Sciences TECHNIKUM WIEN

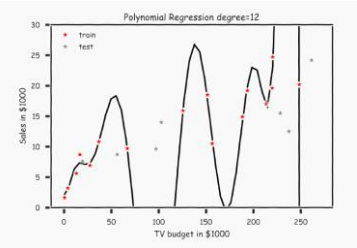
Recap: Regression

Multi-linear regression: Models the relationship between multiple predictors and a response variable

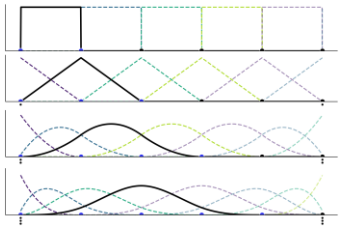


Polynomial regression: Captures nonlinear relationships by adding polynomial terms but risks overfitting





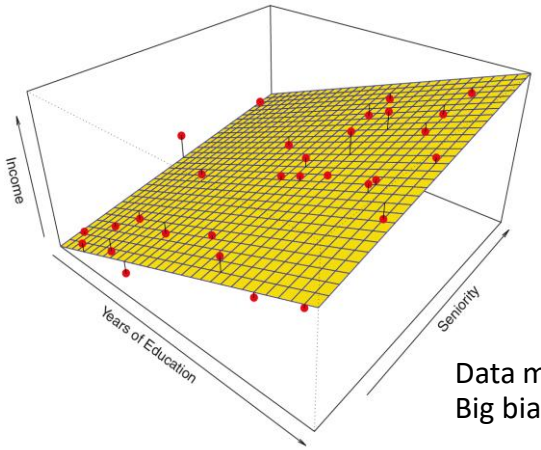
Spline regression: A more flexible approach to fitting nonlinear data but may still suffer from overfitting



How can we control overfitting while keeping a model flexible?

The Bias-Variance Trade-off

Bias: error due to simplifying assumption.
H*igh bias: underfitting*

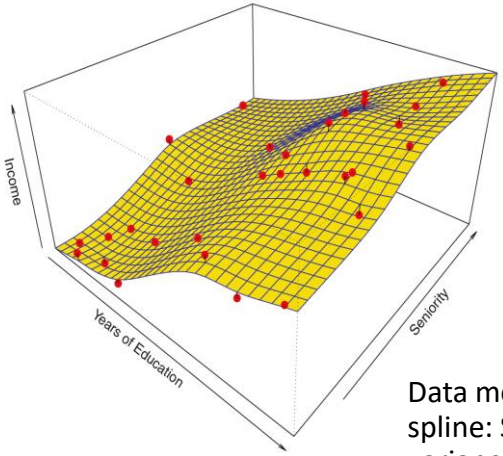


Data modeled by linear regression: Big bias, small variance.

(Modified from [1])

Variance: The amount by which a model would change given small changes in the training set.

High variance: model fails to generalize



Data modeled by thin-plate spline: Small bias, big variance. (Modified from [1])

What is the right balance between model complexity and generalization?

Flexibility vs. Interpretability

Simple Models

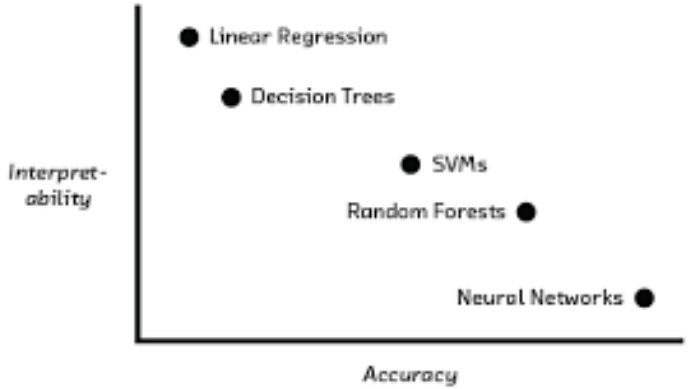
A model with **big bias** provides:

- **Low flexibility (accuracy)**,
- **high interpretability**.

Complex Models

A model with **large variance** provides:

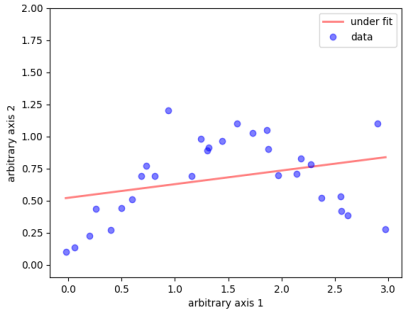
- **High flexibility (accuracy),**
- **low interpretability**.



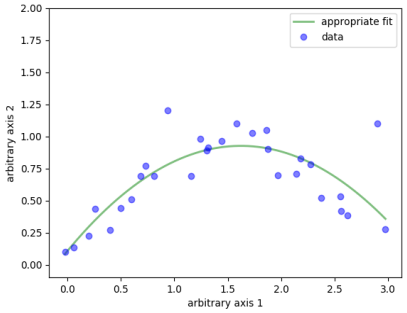
•**Regularization** forces models to be simpler, enhancing interpretability while maintaining predictive power.

Regularization: Intuition

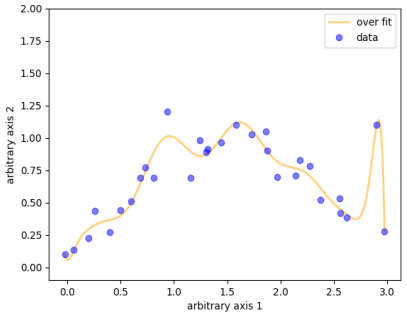
Model Complexity	Bias	Variance	Generalization
Too Simple (Underfitting)	**High**	Low	Poor
Optimal Model	Balanced	Balanced	Best
Too Complex (Overfitting)	Low	**High**	Poor



Too simple!



Optimal



Too complex!

FH University of Applied Sciences TECHNIKUM WIEN

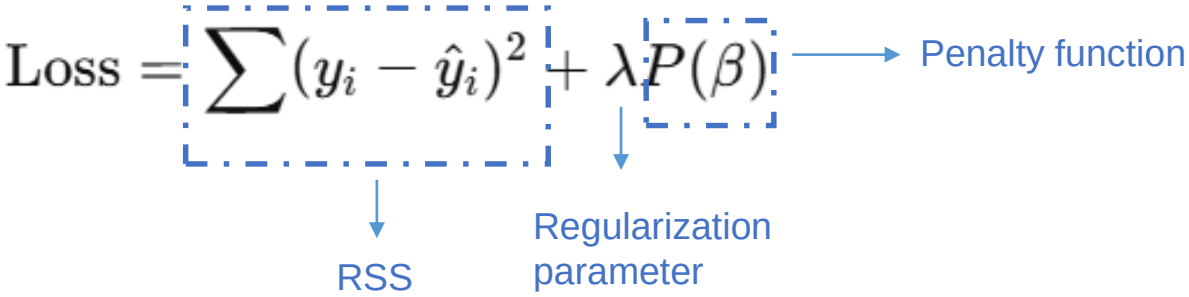
Regularization: Key Concept

Regularization helps control model complexity to improve predictive performance.

Methodology:

Modify the **loss function** to add a **penalty term** on coefficients

Shrinks coefficients → **Prevents large weights that cause overfitting**.



Constraining the estimated coefficients can reduce the variance at the cost of a (negligible) increase in bias.

FH University of Applied Sciences TECHNIKUM WIEN

Regularization

- **In multilinear models, regularization can help with colinearities in the feature variables!**
- **Model Interpretability: removing irrelevant variables** may increase accuracy and increases **variables contributions.**
- **Discourages Complexity: regularization** constrains the coefficients of a model, avoiding overfitting.
- **Shrinkage**: type of regularization which fits a model involving **all p predictors**, but the estimated coefficients are encouraged to be small relative to the least squares estimates.
- **Variable Selection:** depending on what type of shrinkage is performed, some of the **coefficients** may be effectively set to **zero**.
- The two common regularization methods for linear regression are **Ridge Regression** and **Lasso Regression**.

Ridge Regression

- **Ridge regression is related to L2-regularization**. In ridge regression one minimizes the following expression.

$$\mathrm{RSS} + \lambda \|\beta\|_2^2 = \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \geq 0$ is a **hyperparameter** to be determined.

Shrinkage penalty $\lambda \|\beta\|_2^2$: shrinks the parameters towards zero.

Small for $\beta_1, \ldots, \beta_p$ close to zero.

L2 Penalty

$$P(\beta) = \sum_{j=1}^{p} \beta_j^2$$

FH University of Applied Sciences TECHNIKUM WIEN

Ridge Regression

$$\mathrm{RSS} + \lambda \|\beta\|_2^2 = \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- If λ**=0**, the penalty term has **no effect** and ridge regression will reproduce the least squares estimates.
- As $\lambda \rightarrow \infty$ the impact of the shrinkage penalty grows and the absolute values of the estimated parameters **approach zero**.
- Ridge regression produces a **different set** of parameters for each value of λ

What is a **good value** for λ?

FH University of Applied Sciences TECHNIKUM WIEN

Ridge Regression: Hyperparameter λ

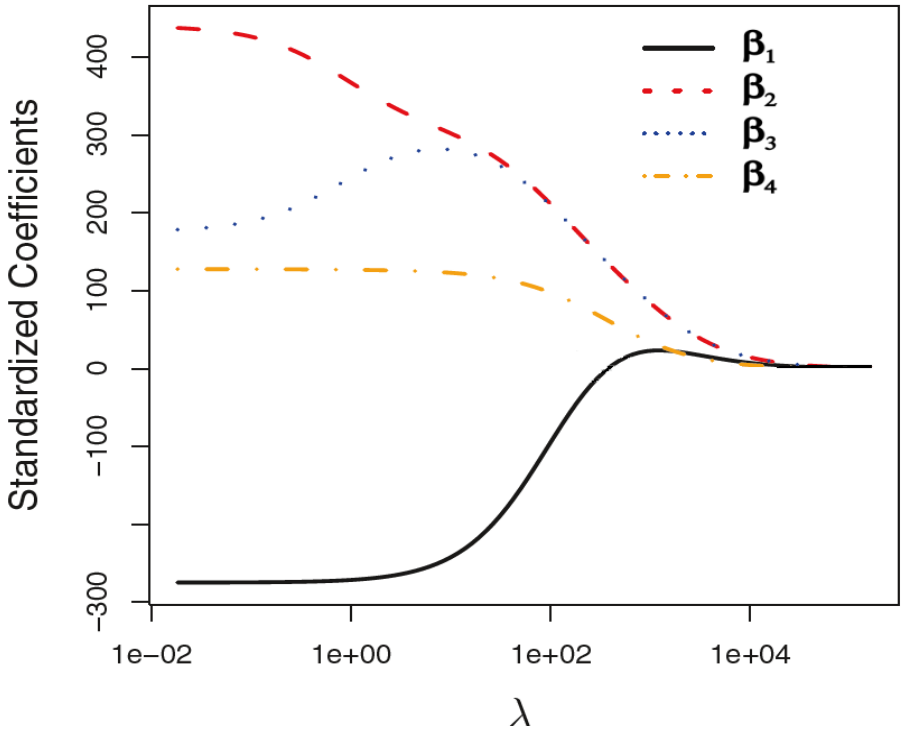
Ridge regression produces a different set of parameters for each value of lambda

On the very left-hand side:

- λ = 0 and coefficients are **as in least squares**.

On the right-hand side:

- λ is large and coefficients are zero. This is **the null model**.



Typical values range **from** 10^{-4} **to** 10^{4}

Modified from [1]

Ridge Regression: Conclusion

Use cases: When you want to regularize without removing variables

- when all features are relevant (ex: finance, physics, medicine, marketing)
- When working with high-dimensional data (many features, relatively few samples).
- When interpretability matters (all features are kept with smaller coefficients)

- **Trades off** a **small increase in bias** for a **large decrease in variance**.
- Ridge regression has the largest impact when the least squares estimates have high variance.
- Ridge regression is recommended if one suspects colinearity in the feature variables.

$$\hat{\beta} = \arg\min_{\beta} \left(\mathrm{RSS} + \lambda \|\beta\|_2^2 \right)$$

FH University of Applied Sciences TECHNIKUM WIEN

Quizz

You train a Ridge regression model and notice that the R-squared score on the test set decreases as λ increases. Which of the following is the most likely explanation?

(A) Increasing λ increases variance, leading to worse generalization.
(B) Increasing λ reduces model complexity, leading to underfitting.
(C) Increasing λ increases the number of features used, leading to overfitting.
(D) Increasing λ improves feature selection, which reduces multicollinearity.

Lasso Regression

- In Lasso regression one **minimizes** the following expression:

$$\mathrm{RSS} + \lambda \|\beta\|_1 = \mathrm{RSS} + \lambda \sum_{j=1}^{n} |\beta_j|$$

where $\lambda \geq 0$ is a **hyperparameter** to be determined.

This is Tikhonov regularization using the **L1-norm**.

Shrinkage penalty: more **pronounced tendency to shrink single parameters strongly** (often to practically **zero)** when λ is sufficiently large **(Variable Selection)**.

Leads to more interpretable models!

L1 Penalty

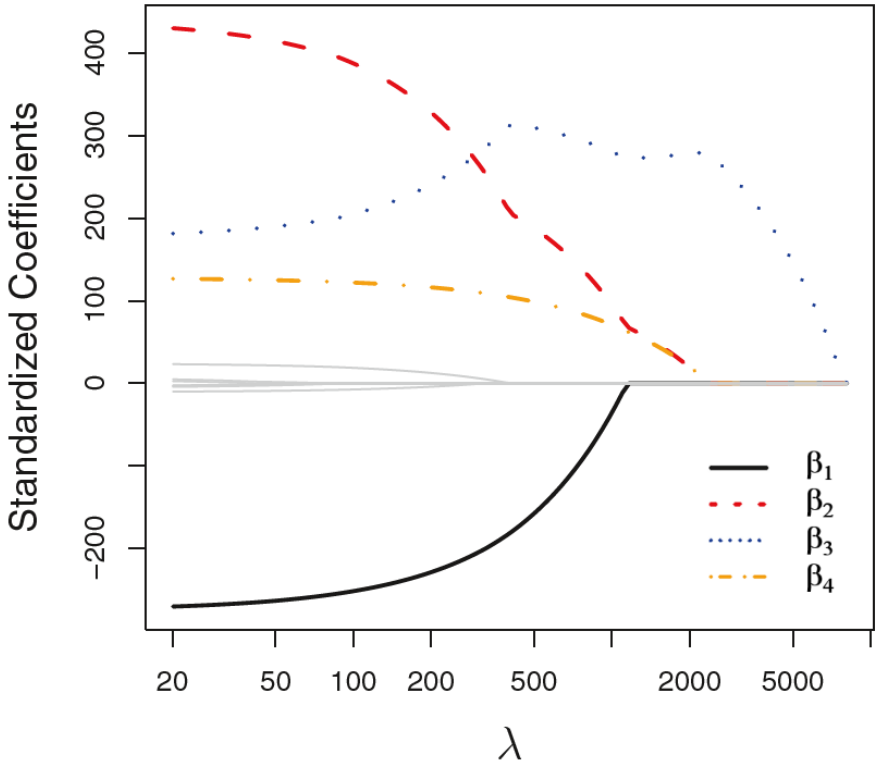
$$P(\beta) = \sum_{j=1}^{n} |\beta_j|$$

FH University of Applied Sciences TECHNIKUM WIEN

Lasso Regression: Hyperparameter λ

Lasso Regression produces a different set of parameters for each value of lambda

- If **λ=0** then coefficients are the same as in least squares.
- As λ grows coefficients converge to zero (**the null model**).
- For intermediate values of λ, lasso regression can produce a model **involving varying number** of variables.



Modified from [1]

Shrinkage in L1 and L2

Regularization: a coefficient value is reduced

L1 Penalty: the loss function is reduced by the same amount as the coefficient

L2 Regularization: the loss function gets reduced by the square of the change in the coefficient value.

Ex: If a coefficient is reduced from 1 to 0.5 it will reduce the loss function by 0.5 in L1, but by 0.25 in L2.

L2 (Ridge) regularization will generally cause the model to converge without shrinking coefficients all the way to zero.

Lasso Regression: Conclusion

Use cases: When you want to select relevant variables

- When data is high dimensional (ex: genomics, energies, etc)
- When feature selection is needed (sparse models)

- **Trades off** a **small increase in bias** for a **large decrease in variance (Regularization)**
- Lasso regression has the largest impact when working with small sample sizes (avoids overfitting)
- Lasso regression is recommended when variable selection is desired, e.g. when there are more features than observations.

$$\hat{\beta} = \arg\min_{\beta} \left(\mathrm{RSS} + \lambda \|\beta\|_1 \right)$$

FH University of Applied Sciences TECHNIKUM WIEN

Quizz

You are using Lasso regression on a dataset with 100 features, but only 10 are truly useful for predicting the target variable. However, you observe that the selected features change slightly each time you retrain the model. What is the most likely explanation?

(A) Lasso struggles with feature selection when features are highly correlated.
(B) The dataset size is too small, making Ridge regression a better choice.
(C) The regularization parameter (λ\lambdaλ) is too small, causing overfitting.
(D) Lasso is deterministic, so this should not happen unless the dataset changes.

FH University of Applied Sciences TECHNIKUM WIEN

Other Regularization Methods

Elastic Net Regression (Hybrid of Ridge & Lasso)
Combines **L1 (lasso)** and **L2 (ridge)** penalties.

$$\sum (y_i - \hat{y}_i)^2 + \lambda_1 \sum |\beta_j| + \lambda_2 \sum \beta_j^2$$

- Handles **multicollinearity** better than lasso alone.
- Still performs **feature selection** but avoids lasso's instability in high-dimensional data.
- Useful when there are **many correlated features and a sparse true signal**.

Used when: Lasso selects few features and/or Ridge retains many irrelevant features.

See also:
Group Lasso (structured data), Adaptive Lasso,
Bayesian Ridge/Lasso (probabilistic)
Dropout (CNNs) – comes up next class

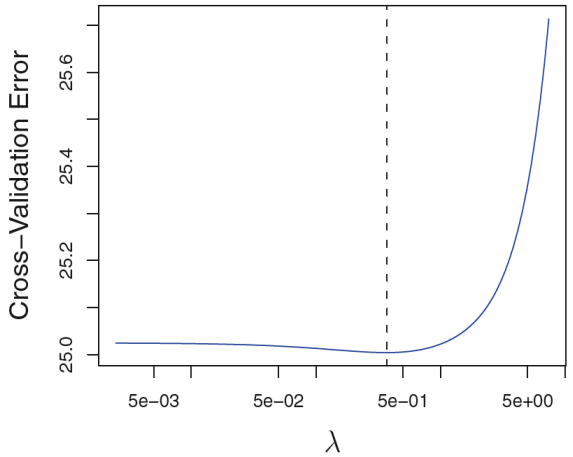
FH University of Applied Sciences TECHNIKUM WIEN

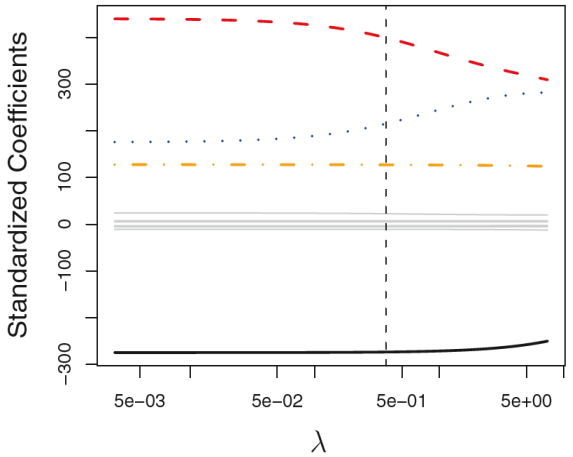
Selecting λ

- Choice of a grid of λ values and computation of cross-validation error for each value of λ.
- Selection of λ for which the cross-validation error is smallest.
- Refit using all of the available observations and the selected λ.

Left:

Cross-validation errors from applying ridge regression.





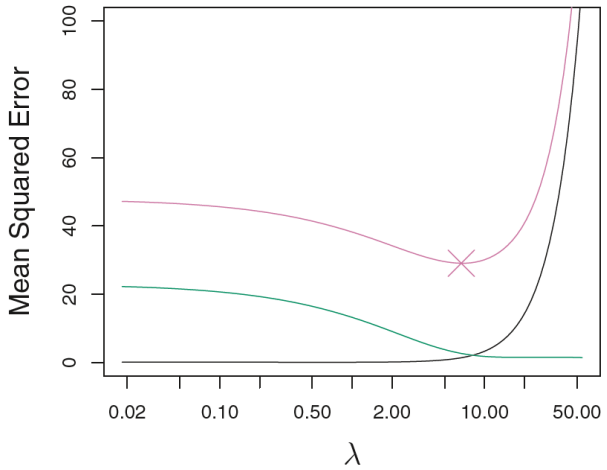
Right:

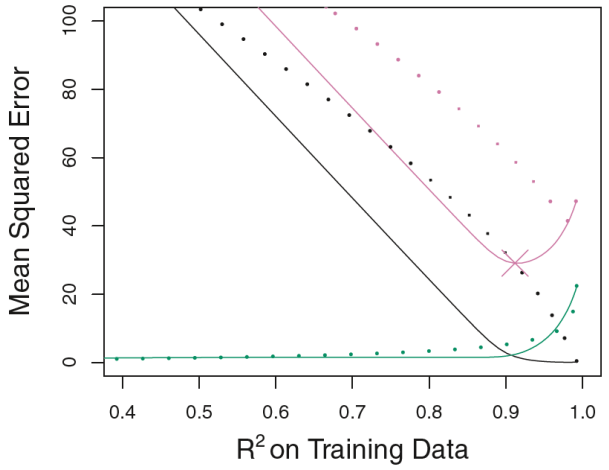
Coefficient estimates as a function of λ. The vertical dashed lines indicate the value of λ selected by cross-validation.
From [1].

Comparing Lasso and Ridge Regression

- Left: Plots of **in-sample error (black)**, **variance (green)**, and **test MSE (pink)** for lasso regression.
- Right: Comparison of squared bias, variance and test MSE between **lasso (solid)** and **ridge regression (dashed)** plotted against R^2 on training data.

The crosses indicate the lasso models for which MSE is smallest.





From [1]

Takeaway: Ridge and Lasso Regularization

Aspect	Ridge Regression (L2)	Lasso Regression (L1)
Effect on Coefficients	Shrinks all coefficients but keeps them	Shrinks some coefficients to exactly **zero**
Feature Selection?	❌ No	✅ Yes
Handles Collinearity?	✅ Distributes weights among correlated features	❌ Picks only one feature and ignores others
Computational Complexity	✅ Faster (solves closed-form solution)	❌ Slower (uses iterative optimization)
Best Used When	Features are **all relevant**, collinear data	**Sparse models**, many irrelevant features

The # of predictors that is related to the response is never known a priori.

Cross-validation can be used in order to determine which approach is better.

FH University of Applied Sciences TECHNIKUM WIEN

Quizz

Which of the following statements about Ridge and Lasso Regression is TRUE?

(A) Ridge regression can eliminate irrelevant features by setting some coefficients to exactly zero.
(B) Lasso regression always selects the correct set of relevant features if given enough training data.
(C) Increasing the regularization parameter λ in Ridge regression always improves the model's test performance.
(D) Lasso regression is more likely to perform well in high-dimensional settings where many features are irrelevant.

Assignment: Ridge and Lasso Regression

a) Explain ridge and lasso regression (1 page/slide each).

Use self-made images or even hand drawings (of which you take a photo).

Use self-written explanations.

Do not copy from the lecture slides or the internet (neither text nor images).

b) Implement a simple version of ridge and/or lasso regression.

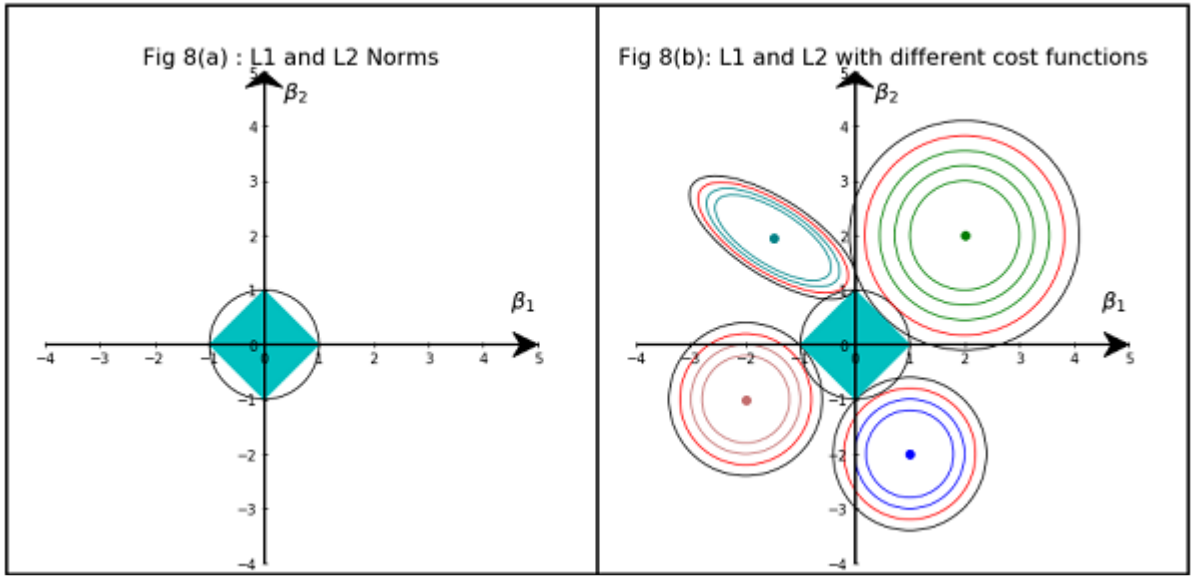
You can use a library for regression but add a penalty term to the RSS. Then try to change some coefficients. You do not need to find the optimum - just try some combinations of parameters to get a feeling of how these algorithms work and how they affect the outcome of least squares regression.

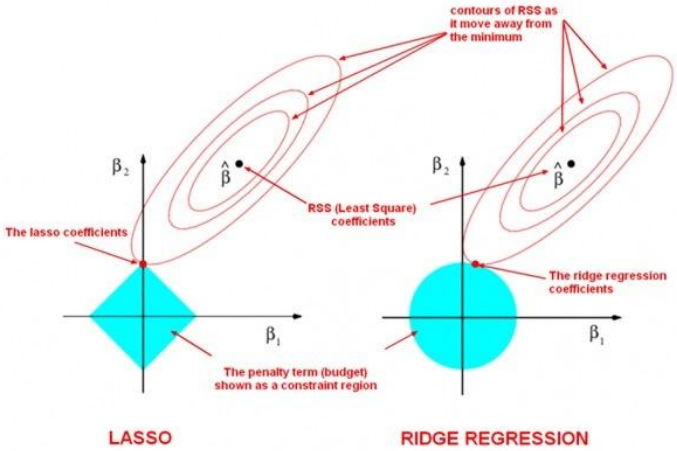
Run your code on a test set of your choice.

Also use sklearn.linear_model.Ridge on the same data.

References

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013.





L1 geometry (Lasso): Unit ball is square.
L2 geometry ("Ridge"): Unit ball is disc.

Look at the intersection where the gradient descent contours intersect with the lasso and ridge areas. Lasso has a greater propensity for intersection at the axes, that is to set coefficients to zero.

https://www.quora.com/Why-is-it-that-the-lasso-unlike-ridge-regression-results-in-coefficient-estimates-that-are-exactly-equal-to-zero
https://www.r-bloggers.com/2020/05/quick-tutorial-on-lasso-regression-with-example/