27411 Chemometrics and Biological Data Analysis

Lecture 7: Ridge and Lasso Regression

Per Bruun Brockhoff

DTU Compute, Statistics and Data Analysis
Building 324, Room 220
Danish Technical University
2800 Lyngby – Danmark
e-mail: perbb@dtu.dk

DTU Compute
Department of Applied Mathematics and Computer Science

Overview

DTU Compute
Department of Applied Mathematics and Computer Science

Multivariate calibration

- Classical (chemical) calibration: Model x as function of y
- Predict y fro x by "going backwards".
- We are doing Inverse calibration: $y = f(\boldsymbol{x})$
- What is a good model: In theory: $f(\boldsymbol{x}) = \mathrm{E}(Y|\boldsymbol{X} = \boldsymbol{x})$.
- If $Y = \mathrm{E}(Y|\boldsymbol{X} = \boldsymbol{x}) + \boldsymbol{\varepsilon}$ and $\hat{f}(\boldsymbol{x})$ is some model estimator, expected prediction error:

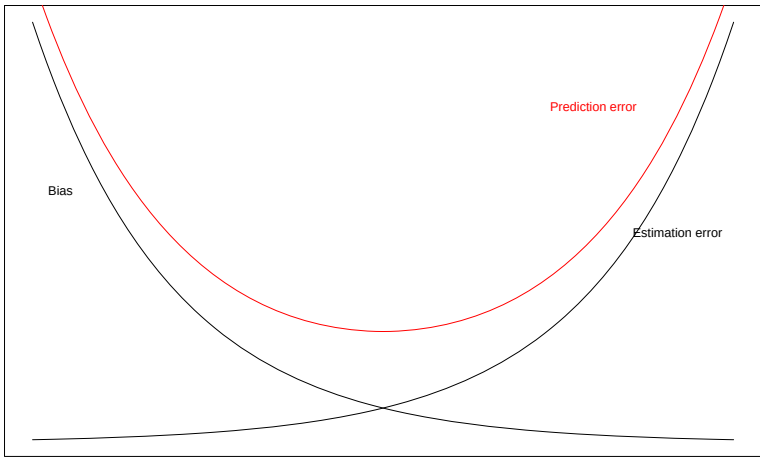
$$\mathrm{E}_{Y|X=\boldsymbol{x}}((y - \hat{f}(\boldsymbol{x}))^2) =$$

$$\mathrm{E}((y - \mathrm{E}(Y|X = \boldsymbol{x}))^2) + \mathrm{E}((f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}))^2) + (f(\boldsymbol{x}) - \mathrm{E}(Y|X = \boldsymbol{x}))^2$$

$$= \sigma^2 + \mathrm{Var}(\hat{f}(\boldsymbol{x})) + \mathrm{Bias}^2$$

DTU Compute
Department of Applied Mathematics and Computer Science

Find optimal predictor by CV



DTU Compute
Department of Applied Mathematics and Computer Science

Basic model: MLR

$$\mathrm{E}(y_i|x_{i1}, x_{i2}, \ldots, x_{ip}) = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}$$
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\varepsilon}$$

- Estimation: $\hat{\boldsymbol{b}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$
- Problem: The parameters are badly estimated!
- HIGH Uncertainty!
- HIGH value of one parameter can go together with a parameter of opposite sign - NOT changing the model

DTU Compute
Department of Applied Mathematics and Computer Science

Basic model: MLR

- Uncertainty:

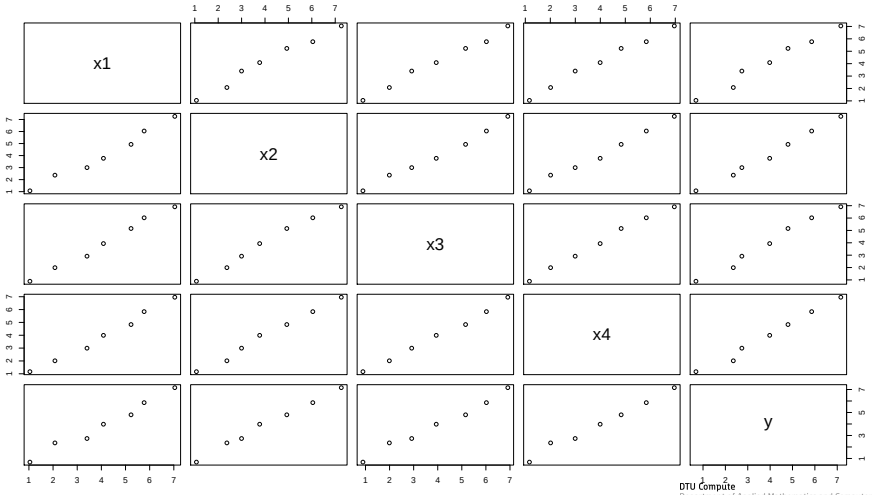
$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{b}}) &= \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \hat{\sigma}^2 \sum_{j=1}^{p} (\boldsymbol{v}_j \boldsymbol{v}_j')/d_j^2\end{aligned}$$

vv_j is the jth principal loading vector
d_j^2 is the jth principal eigenvalue of $\boldsymbol{X}'\boldsymbol{X}$
$(\mathrm{Var}(\boldsymbol{X}\boldsymbol{v}_j) = d_j^2/n)$

- SMALL principal directions will induce HIGH uncertainty!

DTU Compute
Department of Applied Mathematics and Computer Science

Example Data



DTU Compute
Department of Applied Mathematics and Computer Science

MLR results:

	Estimate	Std. Error	t	P-val
Intercept	-0.1668	0.4309	-0.387	0.736
x1	-0.8141	1.6888	-0.482	0.677
x2	-0.1027	0.8635	-0.119	0.916
x3	2.0695	1.4133	1.464	0.281
x4	-0.1354	0.6183	-0.219	0.847

b_3 becomes too large AND the others too small.

DTU Compute
Department of Applied Mathematics and Computer Science

Ridge regression

- Penalizes (numerically) large regression coefficients
- $\hat{\boldsymbol{b}}^{Ridge}$ minimizes (for a given s)

$$\sum_{i=1}^{n}\left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} b_j^2 \leq s$$

- $\hat{\boldsymbol{b}}^{Ridge}$ minimizes (for a given λ)

$$\sum_{i=1}^{n}\left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} b_j^2$$

DTU Compute
Department of Applied Mathematics and Computer Science

Lasso regression

- Penalizes (numerically) large regression coefficients
- $\hat{\boldsymbol{b}}^{Lasso}$ minimizes (for a given s)

$$\sum_{i=1}^{n}\left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |b|_j \leq s$$

- $\hat{\boldsymbol{b}}^{Lasso}$ minimizes (for a given λ)

$$\sum_{i=1}^{n}\left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} |b|_j$$

DTU Compute
Department of Applied Mathematics and Computer Science

Ridge regression: Solution

- Add λ to the diagonal of $\boldsymbol{X}'\boldsymbol{X}$:

$$\boldsymbol{b}^{Ridge} = (\boldsymbol{X}'\boldsymbol{X} + \lambda I)^{-1}\boldsymbol{X}'\boldsymbol{y}$$

- RR-predictor: $\hat{\boldsymbol{y}}^{Ridge} = \boldsymbol{X}\boldsymbol{b}^{Ridge}$
- Operational: Try different λs - choose the optimal by Cross Validation.

DTU Compute
Department of Applied Mathematics and Computer Science

MLR(LS), PCR and RR

$$\hat{\boldsymbol{y}}^{LS} = \boldsymbol{X}\boldsymbol{b}^{LS} = \sum_{j=1}^{p} \boldsymbol{u}_j \boldsymbol{u}_j' \boldsymbol{y}$$

$$\hat{\boldsymbol{y}}^{Ridge} = \boldsymbol{X}\boldsymbol{b}^{Ridge} = \sum_{j=1}^{p} \boldsymbol{u}_j \frac{d_j^2}{d_j^2 + \lambda} \boldsymbol{u}_j' \boldsymbol{y}$$

$$\hat{\boldsymbol{y}}^{PCR} = \boldsymbol{U}\boldsymbol{\theta}^{PCR} = \boldsymbol{X}\boldsymbol{b}^{PCR} = \sum_{j=1}^{A} \boldsymbol{u}_j \boldsymbol{u}_j' \boldsymbol{y}$$

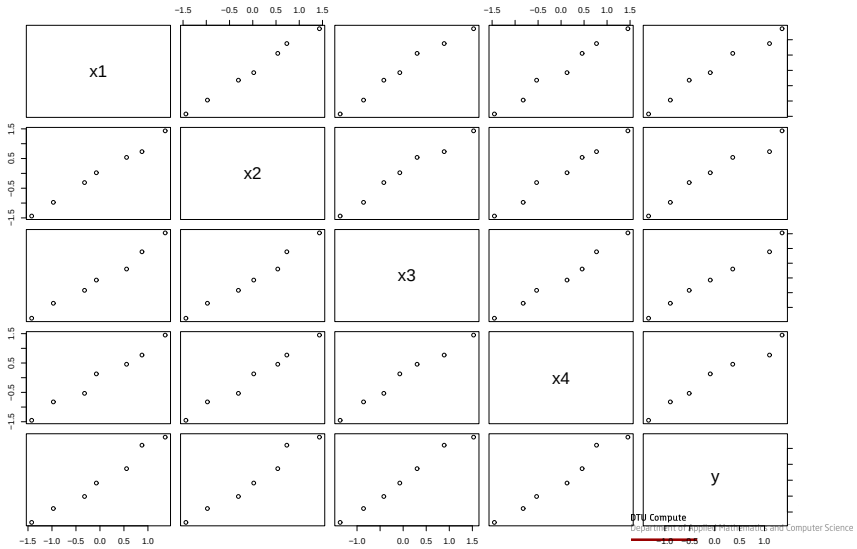
$\boldsymbol{u}_j$ is the jth principal score values

DTU Compute
Department of Applied Mathematics and Computer Science

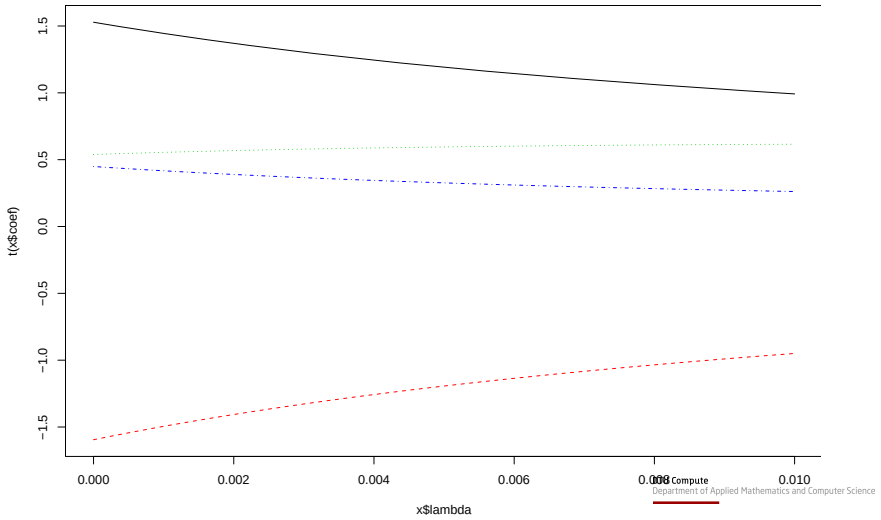
MLR(LS), PCR and RR/Lasso

- The LS-solution is a vector in space.
- Ridge regression *shrinks* the LS-solution
- Highest shrinkage in low-variance directions
- PCR also shrinks: No shrinkage in main directions, 100% shrinkage in low variance directions.
- (PLS is ALSO a shrinkage method - allthough "expansion" is also possible)
- Lasso also shrinks - some coefficients down to EAXCTLY zero
- Lasso is hence also an automated variable selection method
- Lasso is an L_1 method - VERY different from:
- The others are L_2 methods

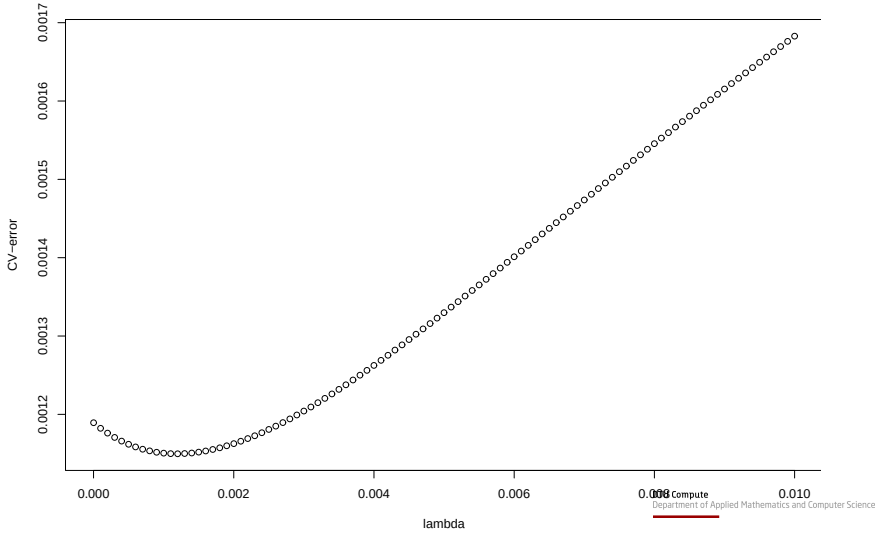
Example Data



Regression coefficients



Cross Validation



MLR and RR results:

	MLR Estimate	RR estimate ($\lambda = 0.0012$)
x1	1.651	1.543
x2	-1.723	-1.595
x3	0.5819	0.6023
x4	0.4848	0.4440

How to do it?(same as for PCR/PLS)

1. Explore data
2. Do modelling (choose number of components, consider variable selection)
3. Validate (residuals, outliers, influence etc) (not so straightforward)
4. Iterate e.g. on 2. and 3.
5. Interpret, conclude, report.
6. If relevant: predict future values.