UiO : Department of Physics
University of Oslo

Regression analysis and resampling methods

Erik Skaar
Sondre Torp
Mikael Kiste

Contents

Abstract

[compphys]

1 Introduction

2 Theory

2.1 Standard

$$\beta = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

2.2 Ridge

$$\beta = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

2.3 Lasso

$$\beta = \mathrm{argmin}_\beta \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}$$

2.4 k-fold and bootstrap

3 Method

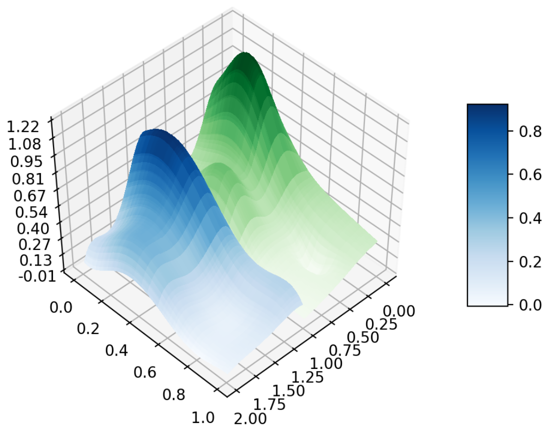
4 Implementation

The three different algorithms discussed in section xxx was implemented in our script. It is a few different versions, but the ëversion contains all you need. All the scripts discussed in this report can be found at our github.

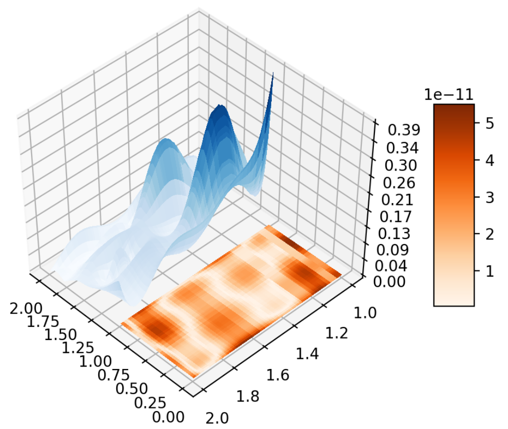
The program was tested on the Frank-function, see equation 1. With an known solution we did a k-fold test and an degree and λ/α test. Both tested was done with the script descriped earlier. The tables below shows the different results.

$$f(x,y) = \frac{3}{4}e^{\left(-\frac{(9x-2)^2}{4}-\frac{(9y-2)^2}{4}\right)} + \frac{3}{4}e^{\left(-\frac{(9x+1)^2}{49}-\frac{(9y+1)}{10}\right)} + \frac{1}{2}e^{\left(-\frac{(9x-7)^2}{4}-\frac{(9y-3)^2}{4}\right)} - \frac{1}{5}e^{\left(-(9x-4)^2-(9y-7)^2\right)} \tag{1}$$

4.1 Scikit vs. manually implementation



(a)



(b)

Figure 1: a)... b)...

4.2 Time evolution

Table 1: This tables shows how the MSE evoles for different degrees. Scikit OLS is to confirm that our implementation is not retarded. For lasso and ridge the λ/α was set to 1e-5. Also, if we go beyond fifth order the OLS solutions starts to crumble.

degree ↓	method →	OLS	SCIKIT	RIDGE	SCIKIT LASSO
2		0.01517	0.25830	0.00516	0.00543
$2_{relative}$		1.00	1.00	1.00	1.00
$3_{relative}$		2.42	1.58	2.45	2.38
$4_{relative}$		3.63	2.45	5.11	4.88
$5_{relative}$		4.98	3.61	8.77	8.31

4.3 Noise - MSE & R2 evolution

Table 2: This tables shows how the MSE evoles for different degrees. Scikit OLS is to confirm that our implementation is not retarded. For lasso and ridge the λ/α was set to 1e-5. Also, if we go beyond fifth order the OLS solutions starts to crumble.

Noise level ↓	method →	OLS	SCIKIT	RIDGE	SCIKIT LASSO
0		0.00127	0.00127	0.00514	0.00127
$0_{relative}$		1.00	1.00	1.00	1.00
$0.01_{relative}$		1.03	1.03	1.00	1.03
$0.2_{relative}$		12.84	12.84	3.68	12.84
$0.5_{relative}$		42.04	42.04	10.84	42.04

Table 3: This tables shows how the MSE evoles for different degrees. Scikit OLS is to confirm that our implementation is not retarded. For lasso and ridge the λ/α was set to 1e-5. Also, if we go beyond fifth order the OLS solutions starts to crumble.

Noise level ↓	method →	OLS	SCIKIT	RIDGE	SCIKIT LASSO
0		0.98	0.98	0.91	0.98
0.01		0.98	0.98	0.91	0.98
0.2		0.68	0.68	0.62	0.68
0.5		0.28	0.28	0.25	0.28

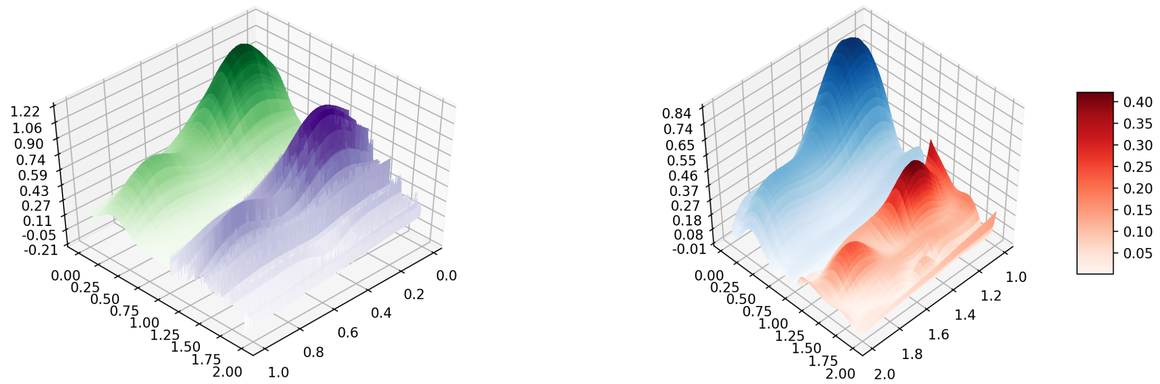
5 Result & Discussion

5.1 Ordinary least square, Ridge, and Lasso regression with resampling on the Franke function

In this subsection we will present Ordinary least square, Ridge and Lasso regression up to the fifth order, with a resampling technic, k-fold, on the Franke function. The Mean square error, (MSE), R^2 score function and the confidence intervall is also presented.

5.1.1 Ordinary least square

Here we present the OLS regression with up to a fifth order polynomial fit on Franke function, notice that in the plot we use a fifth order polynomial, the R^2 score and MSE according to order of the polynomial used for fitting of the data. And lastly a table containing the β values, the variance, and the confidence intervall according to the different polynomials
The confidence intervall of β through variance. The mean squared error(MSE) and the R^2 score function. Presenting the resampling of the data, where the data have been splitt into training and test data. Implementing k-folding and reevaluates MSE, R^2, on the test data. Finding Bias and variance with $Err(x_0) = irreducibleError + Bias^2 + variance$



(a) (b)

Figure 2: a) Franke function plottet in green, and the Frank function with noise plottet in purple. b) Our fifth order approximation of the Franke function plottet in blue. On the right we have the residuals, i.e. the error compared to the real function, and its relative size indicated by the red colour gradient.

Table 4: MSE and R^2 score for OLS by degree. These values are created by taking the average values over 100 different executions, with a noise level = 0.1 and a $\lambda/\alpha = 0.00001$

Degree	R2	MSE
2	0.71	0.01353
3	0.81	0.00916
4	0.86	0.00712
5	0.88	0.00585

Table 5: β, Var and Confidence intervall for OLS by degree of x and y.

x^iy^j	β	VAR	Confidence intervall
x^0y^0	0.259	0.001	[0.227, 0.291]
x^0y^1	4.117	0.108	[3.788, 4.446]
x^0y^2	-18.065	2.943	[-19.781, -16.349]
x^0y^3	29.764	15.934	[25.772, 33.756]
x^0y^4	-22.199	17.571	[-26.391, -18.007]
x^0y^5	6.361	2.637	[4.737, 7.985]
x^1y^0	5.277	0.074	[5.005, 5.549]
x^1y^1	-9.751	1.234	[-10.862, -8.64]
x^1y^2	13.591	6.186	[11.104, 16.078]
x^1y^3	-21.609	7.858	[-24.412, -18.806]
x^1y^4	12.635	1.628	[11.359, 13.911]
x^2y^0	-22.726	1.825	[-24.077, -21.375]
x^2y^1	28.964	5.900	[26.535, 31.393]
x^2y^2	-1.849	6.149	[-4.329, 0.631]
x^2y^3	-5.401	1.364	[-6.569, -4.233]
x^3y^0	30.056	10.123	[26.874, 33.238]
x^3y^1	-36.035	7.072	[-38.694, -33.376]
x^3y^2	6.742	1.441	[5.542, 7.942]
x^4y^0	-11.919	12.008	[-15.384, -8.454]
x^4y^1	12.813	1.444	[11.611, 14.015]
x^5y^0	-0.909	1.951	[-2.306, 0.488]

5.1.2 Ridge regression

Samme analyse (samme polynomiales og resampling teknikk) med forskjellige verdier av lambda. Se på avhengigheten av lambda imens du varierer mengden bråk(noise i Franke).

Table 6: MSE and R^2 score for Ridge by degree. These values are created by taking the average values over 100 different executions, with a noise level = 0.1 and a $\lambda = 0.00001$

Degree	R2	MSE
2	0.71	0.01353
3	0.80	0.00968
4	0.81	0.00928
5	0.81	0.00902

Table 7: MSE and R^2 score for OLS by degree. These values are created by taking the average values over 100 different executions, with a noise level = 0.1

λ/α	R2	MSE
0.0000001	0.81150	0.00587
0.0000100	0.80979	0.00605
0.0010000	0.74059	0.00592
0.1000000	-0.20901	0.00608
1.0000000	-1.85170	0.00613
2.0000000	-1.83823	0.00586
5.0000000	-1.79681	0.00601
10.0000000	-1.81950	0.00591

Table 8: β, Var and Confidence intervall for Ridge by degree of x and y.

x^iy^j	value	variance	confidence intervall
x^0y^0	0.384	0.001	x
x^0y^1	1.770	0.077	x
x^0y^2	-0.294	1.953	x
x^0y^3	-22.690	10.474	x
x^0y^4	41.542	12.033	x
x^0y^5	-20.675	1.949	x
x^1y^0	5.348	0.068	x
x^1y^1	-11.544	1.040	x
x^1y^2	18.567	4.891	x
x^1y^3	-27.573	6.071	x
x^1y^4	14.943	1.295	x
x^2y^0	-21.987	1.712	x
x^2y^1	32.153	5.221	x
x^2y^2	-5.093	5.179	x
x^2y^3	-3.142	1.161	x
x^3y^0	25.775	9.426	x
x^3y^1	-39.259	6.976	x
x^3y^2	6.673	1.214	x
x^4y^0	-4.833	11.244	x
x^4y^1	14.552	1.590	x
x^5y^0	-4.681	1.908	x

5.1.3 Lasso regression

Presenter det samme igjen, men med Lasso regression på Frank, Vurder de tre metodene opp mot hverandre

Table 9: MSE and R^2 score for Lasso by degree. These values are created by taking the average values over 100 different executions, with a noise level = 0.1 and a $\lambda/\alpha = 0.00001$

Degree	R2	MSE
2	0.71	0.01353
3	0.81	0.00916
4	0.86	0.00712
5	0.88	0.00585

Table 10: MSE and R^2 score for Lasso on Franke function by λ. These values are created by taking the average values over 100 different executions, with a noise level = 0.1

λ	R2	MSE
0.0000001	0.87497	0.00587
0.0000100	0.87517	0.00605
0.0010000	0.87723	0.00592
0.1000000	0.87242	0.00608
1.0000000	0.87708	0.00613
2.0000000	0.87833	0.00586
5.0000000	0.87572	0.00601
10.0000000	0.87528	0.00591

5.2 Ordinary least square, Ridge, and Lasso regression with resampling, now on real data

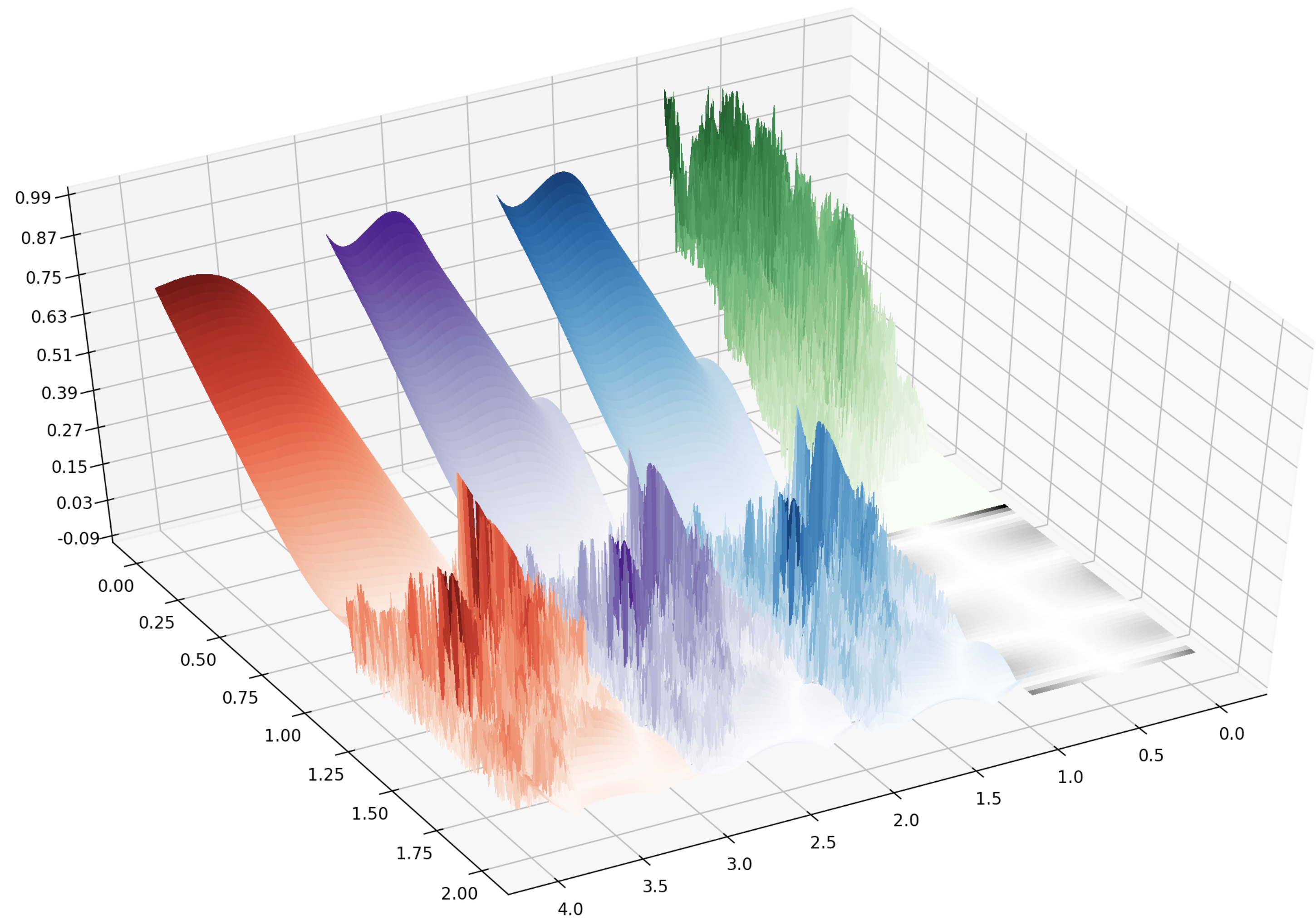


Figure 3: REAL DATAAAA.

Table 11: OLS on realdata. Can see that R^2 "går til helvete", when big alpha, so no point in plotting for higher alpha

λ	R2	MSE
0.0000001	0.823088	0.010720
0.0000100	0.823088	0.010720
0.0010000	0.823088	0.010720
0.1000000	0.823088	0.010720

Table 12: Lasso on realdata. Can see that R^2 "går til helvete", when big alpha, so no point in plotting for higher alpha SE LASSO R2

λ	R2	MSE
0.0000001	0.797486	0.012272
0.0000100	0.797470	0.012273
0.0010000	0.753280	0.014950
0.1000000	-0.165026	0.070596

Table 13: Ridge on realdata. MSE and R^2 score for Ridge on the real data by λ.

λ	R2	MSE
0.0000001	0.823088	0.010720
0.0000100	0.823088	0.010720
0.0010000	0.823027	0.010724
0.1000000	0.818116	0.011022

6 Conclusion

7 Appendix