

REGRESSION AND RESAMPLING TECHNIQUES
FYS-STK4155
ASSIGNMENT ONE

Ivar Haugaløkken Stangeby

October 6, 2019

Abstract

In this project, we implement and compare three different variants of linear regression. Namely, ordinary least squares (OLS), Ridge regression, and Lasso regression. We test the methods on terrain data taken from a region near Stavanger, Norway. Initial implementation tests are done on an artificial data-set based on noisy samples of the Franke function. Statistical quantities pertaining to the various model functions are computed and compared. We also give an introduction to the theory and motivation behind the three methods. Implementation details can be found on GitHub:

 github.com/qTipTip

Contents

Contents	2
1 Introduction	3
2 Regression Techniques	3
2.1 Ordinary Least Squares	5
2.2 Ridge Regression	8
2.3 Lasso Regression	8
3 Implementation and methodology	9
3.1 Model statistics and resampling methods	11
4 Numerical Results	13
4.1 Toy Data	14
4.2 Terrain Data	17
5 Conclusion	18
5.1 Further work	19

1 Introduction

Regression analysis is the act of estimating relationships among variables. In this project, we study various regression methods in more detail. In particular, we compare the *ordinary least squares* (OLS) method with the *Ridge regression* and *Lasso regression* techniques. As we shall see, these three methods are all variations over the same theme. We start by testing the methods on noisy data sampled from a function known as *Franke's* bivariate test function (visualized on the title page), which has been in widespread use in the evaluation of interpolation and data fitting techniques. Finally, we run regression on real terrain data, comparing the aforementioned methods.

2 Regression Techniques

In general, the goal of regression analysis is to fit a *model function* $f(\mathbf{x}, \beta)$ to a set of n data points $\mathbf{\Omega} = (\mathbf{x}_i, y_i)_{i=1}^n$. A simple example is that of a linear polynomial with two parameters:

$$f(\mathbf{x}, \beta) = \beta_0 + \beta_1 x. \quad (1)$$

The *model parameters* β are determined in order to minimize a suitable *cost function* $\mathcal{C}(\mathbf{\Omega}, \beta)$ which measures to which extent the model function manages to capture trends in the data $\mathbf{\Omega}$. It is the choice of cost function $\mathcal{C}(\mathbf{\Omega}, \beta)$ which distinguishes the three regression techniques, OLS, Lasso regression, and Ridge regression.

Assumption 1. It is often assumed a priori that the data is in fact generated from a noisy model, such that each y_i can be described as

$$y_i = f(\mathbf{x}_i) + \varepsilon_i \quad (2)$$

where each $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is normally distributed with zero mean and variance σ^2 . This assumption on the error gives rise to what is known as *general linear models*.

The design matrix. We are often interested in finding the best model function in a specific function space. Assuming this space has a basis $\mathcal{B} = \{\varphi_i\}_{i=1}^M$, we may write our model function in terms of the basis functions and the model parameters as

$$f(\mathbf{x}, \beta) := \sum_{i=1}^M \beta_i \varphi_i(\mathbf{x}). \quad (3)$$

As an example, if we were to use the space Π_2^2 of bi-quadratic polynomials, our basis functions would be

$$\mathcal{B} = \{1, x, x^2, y, y^2, xy\}. \quad (4)$$

With this formulation, we can represent the approximation $\hat{\mathbf{y}}$ to \mathbf{y} as a matrix product

$$\hat{\mathbf{y}} = \mathbf{X}\beta \quad (5)$$

where \mathbf{X} is the *design matrix* defined by components $X_{ij} = \varphi_j(\mathbf{x}_i)$, and $\beta = [\beta_1, \dots, \beta_M]$ is the model parameters.

Performance metrics. In order to evaluate how accurately the model function $f(\mathbf{X}, \beta)$ captures trends in the target data \mathbf{y} , a few standard performance metrics are used. Firstly, the *mean squared error*:

$$\text{MSE}(\mathbf{y}, \hat{\mathbf{y}}) := \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6)$$

which simply averages the squared error over all samples and estimates. Secondly, we have the *coefficient of determination* or *R²-score*:

$$R^2(\mathbf{y}, \hat{\mathbf{y}}) := 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \quad (7)$$

The *R²-score* measures much of the variation in \mathbf{y} which can be attributed to a simple linear relation between \mathbf{x} and \mathbf{y} . A value of one tells us that all the variation in the data can be attributed to an approximate linear relationship between the data \mathbf{x} and the response variable \mathbf{y} . A negative value tells us that the model performs worse than a horizontal line through the data.

2.1 Ordinary Least Squares

One common cost function is the one which involves the sum of squared residuals (or squared errors):

$$\mathcal{C}_{\text{OLS}}(\mathbf{\Omega}, \beta) = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 := \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (8)$$

where $\hat{y}_i := f(\mathbf{x}_i, \beta)$. The method involving the minimization of this specific cost function is known as *least squares*. With the matrix notation from above in mind, we can also write the cost function as

$$\mathcal{C}_{\text{OLS}}(\mathbf{\Omega}, \beta) = \frac{1}{n} (\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}}). \quad (9)$$

Optimizing the parameters in OLS.

Assume now that our model is of the form given in Equation (3) and that our cost function is the mean squared error defined in Equation (8). We are interested in finding the parameters β that minimize the cost function $\mathcal{C}(\mathbf{\Omega}, \beta)$. Since $\mathcal{C}(\mathbf{\Omega}, \beta)$ is convex, it suffices to differentiate with respect to β and equating to zero.

We have that

$$\frac{\partial \mathcal{C}_{\text{OLS}}(\mathbf{\Omega}, \beta)}{\partial \beta} = \mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta), \quad (10)$$

and equating this to zero yields the familiar *normal equations*

$$\mathbf{X}^T \mathbf{y} = \mathbf{X}^T \mathbf{X} \beta. \quad (11)$$

If the matrix $\mathbf{X}^T \mathbf{X}$ is invertible, we may obtain the solution by direct numerical matrix inversion. In this case, the optimal model parameters are found directly by

$$\beta_{\text{OLS}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (12)$$

However, these matrices may be ill-conditioned when the number of equations are very large, and it is therefore common to apply approximate solvers for the inverse, for instance using low-rank approximation based on SVD-decomposition, which we will briefly turn to in the following.

Singular value decomposition.

Recall that any matrix $\mathbf{A} \in \mathbb{C}^{n,m}$ can be decomposed as

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \quad (13)$$

where \mathbf{U} and \mathbf{V} are comprised of the eigenvectors of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ respectively. As these eigenvectors are orthonormal, it follows that both \mathbf{U} and \mathbf{V} are unitary matrices. Furthermore, $\mathbf{\Sigma}$ is a matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & 0 \\ 0 & 0 \end{bmatrix}, \quad (14)$$

where $\mathbf{\Sigma}_1$ is a square diagonal matrix of size $r \times r$ with the non-zero singular values of \mathbf{A} . The integer r is the *rank* of \mathbf{A} . As the matrix $\mathbf{\Sigma}$ is mostly containing zeros, the information stored in \mathbf{A} is attributed to only some parts of \mathbf{U} and \mathbf{V} . We can remove the redundant parts, and more compactly express \mathbf{A} as $\mathbf{A} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T$ without loss of accuracy in the decomposition. Here, \mathbf{U}_1 is $m \times r$ and \mathbf{V}_1 is $n \times r$.

Applying the singular value decomposition to the matrix $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ we can analyze the expression for the prediction $\hat{\mathbf{y}}$ in terms of the matrix $\mathbf{X}^T\mathbf{X}$. First of all, we have that

$$\begin{aligned} \mathbf{X}^T\mathbf{X} &= \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \\ &= \mathbf{V}\mathbf{D}\mathbf{V}^T, \end{aligned} \quad (15)$$

where we have used that $\mathbf{U}\mathbf{U}^T = \mathbf{I}$ and defined $\mathbf{D} = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$. Thus, plugging this into Equations (5) and (12), we obtain the expression

$$\hat{\mathbf{y}} = \mathbf{X}\beta = \mathbf{X}(\mathbf{V}\mathbf{D}\mathbf{V}^T)^{-1}\mathbf{X}^T\mathbf{y}. \quad (16)$$

Finally, substituting $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ and noting that diagonal matrices always commute, we end up with

$$\hat{\mathbf{y}} = \mathbf{U}\mathbf{U}^T\mathbf{y}. \quad (17)$$

The bias-variance tradeoff

Since we are attempting to estimate a true distribution based on finite samples, there are certain statistical properties it is worthwhile discussing. We follow the derivation from [Meh+19]. Recall Assumption 1 which states that our true data is generated from a noisy model

$$\mathbf{y} = f(\mathbf{x}) + \varepsilon \quad (18)$$

with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. We select a finite sample from the true data, which constitutes our finite data-set Ω . Furthermore, we have a model function $f(\Omega, \beta)$ which depends on the selected data-set and the model parameters β . Thus, our cost-function \mathcal{C} will depend on the specific sample taken from the true distribution. It is of interest to examine the expected out-of-sample error $\mathbb{E}_{\Omega, \varepsilon}[\mathcal{C}(\Omega, \beta)]$ of the model function over the sampled data set Ω and the noise ε . This can be decomposed into three distinct parts, *variance*, *bias*, and *noise*:

$$\mathbb{E}_{\Omega, \varepsilon}[\mathcal{C}(\Omega, \beta)] = \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i) - \mathbb{E}_{\Omega} [\hat{y}_i])^2 + \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\Omega} [(\hat{y}_i - \mathbb{E}_{\Omega} [\hat{y}_i])^2] + \sigma^2. \quad (19)$$

The bias-term measures the deviation of the expectation of the model function $f(\Omega, \beta)$ from the true value f , while the variance-term measures the fluctuations arising due to using different samples of the true data.

Now, note that the equation above suggests that we should minimize both the variance and the bias of the model, in order to obtain the smallest generalization error of the model function. The variance term will be larger for complex models, as performing a fit on one sampled data-set will not necessarily translate well to another sampled data-set. The bias term measures errors introduced by making certain assumptions about the relations between the covariates \mathbf{X} and the response \mathbf{y} . In our case, we are assuming a linear relationship (recall *linear regression*), and thus if the true underlying relation is non-linear, our bias will be high. More complex models will in general have lower bias.

This is known as the *bias-variance tradeoff* as having a highly complex model, we reduce the bias, yet increase the variance. Having a simple model (for instance a

linear model), we increase the bias, and reduce the variance.

2.2 Ridge Regression

When the dimensionality of the matrix \mathbf{X} is high, the problem of near-collinear columns often arise, as mentioned earlier. Thus, the model parameters β_{OLS} of the ordinary least squares regression cannot be determined. The proposed fix to this problem [Tik+95; HK] is to slightly perturb the matrix $\mathbf{X}^T \mathbf{X}$ by adding small positive values along the diagonal to ensure its invertability. This is known as *ridge regression* and yields a model parameter estimator

$$\beta_{\text{ridge}}(\lambda) = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \quad (20)$$

where $\lambda \geq 0$ is a hyper parameter. To see why the resulting modified matrix is invertible, one has to look at the singular values. It turns out that the model parameters $\beta_{\text{ridge}}(\lambda)$ directly minimize the *ridge loss function* defined as

$$\mathcal{C}_{\text{ridge}}(\mathbf{\Omega}, \beta; \lambda) := \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^p \beta_i^2 = \mathcal{C}(\mathbf{\Omega}, \beta) + \lambda \beta^T \beta, \quad (21)$$

which we recognize as standard penalized least squares.

Ridge regression carries several benefits. One of the more important is its resilience to overfitting. In a high-dimensional setting, if many of the data-features can be attributed to a few covariates, then the remaining covariates do nothing more than fit to the noise. Large values of β_i is then a good indication that your model might be overfitting as some covariates are given more weight than others. The Ridge regression solves this by directly penalizing large model parameters in the cost function.

2.3 Lasso Regression

One drawback of the OLS and Ridge regression approaches is that irrelevant features may still be present in the model. Lasso regression fixes this by — in

addition to penalizing large model parameters — also setting sufficiently small parameters to zero. The associated cost function reads

$$\mathcal{C}_{\text{lasso}}(\mathbf{\Omega}, \beta; \lambda) := \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^n |\beta_i|. \quad (22)$$

Again, $\lambda \leq 0$ is a model hyper-parameter. Note that the penalization-term now involves absolute values of the model parameters. As the Lasso estimator may set model parameters to zero, it is effectively producing a sparser model.

Attempting to differentiate the Lasso cost function with respect to the model parameters, we see that

$$\frac{\partial \mathcal{C}_{\text{lasso}}(\mathbf{\Omega}, \beta; \lambda)}{\partial \beta_i} = \frac{\partial \mathcal{C}_{\text{OLS}}(\mathbf{\Omega}, \beta)}{\partial \beta_i} + \lambda \text{sign}(\beta_i). \quad (23)$$

The analysis of the optimal Lasso parameters is made difficult due to the discontinuity of the sign function, and no closed form solution as in OLS and Ridge-regression can be found in the general case. Thus, numerical optimization techniques are often used, for instance gradient ascent/descent.

For further information on Ridge and Lasso regression, the excellent article “Lecture Notes on Ridge Regression” by van Wieringen [vWie15] is highly recommended.

3 Implementation and methodology

In order to verify the implementation, we test on sampled data from the Franke function [Fra79]. This bivariate function $F(x, y)$ is defined for $x, y \in [0, 1]$ as

follows:

$$\begin{aligned}
F(x, y) := & \frac{3}{4} \exp \left(-\frac{(9x-2)^2}{4} - \frac{(9y-2)^2}{4} \right) \\
& + \frac{3}{4} \exp \left(-\frac{(9x+1)^2}{49} - \frac{(9y+1)^2}{10} \right) \\
& + \frac{1}{2} \exp \left(-\frac{(9x-7)^2}{4} - \frac{(9y-3)^2}{4} \right) \\
& - \frac{1}{5} \exp \left(-(9x-4)^2 - (9y-7)^2 \right).
\end{aligned} \tag{24}$$

We are interested in fitting a bivariate polynomial surface to this data, by means of the regression techniques discussed in Chapter 2. That is, our model function $f(\mathbf{\Omega}, \beta)$ lies in the polynomial space Π_2^d of bi-degree (d, d) polynomials. We are mainly interested in bi-quintic polynomial surfaces, and will therefore restrict our attention to $d = 5$ in the numerical experiments.

We sample the function F both with and without noise at $N = 5000$ points in each parameter direction, yielding a matrix of function values

$$\begin{aligned}
\mathbf{Z} &= [F(x_i, y_j)]_{ij}, \\
\mathbf{Z}^* &= [F(x_i, y_j) + \alpha \varepsilon_{ij}]_{ij},
\end{aligned} \tag{25}$$

where each $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and α is the *signal-to-noise ratio*. The design matrix \mathbf{X} is computed using the POLYNOMIALFEATURES-method from the SKLEARN package. The dimension of the bi-quintic polynomial space is 21, yielding a design matrix of size $N^2 \times 21$.

For the real data set, we use the supplied TIFF data of an area close to Stavanger in Norway, as displayed in Figure 1. The data-set consists of approximately 1800×3600 data-points.

The ordinary least squares and the Ridge regression is performed by solving the corresponding normal equations. We use the Lasso regression available in SKLEARN package.

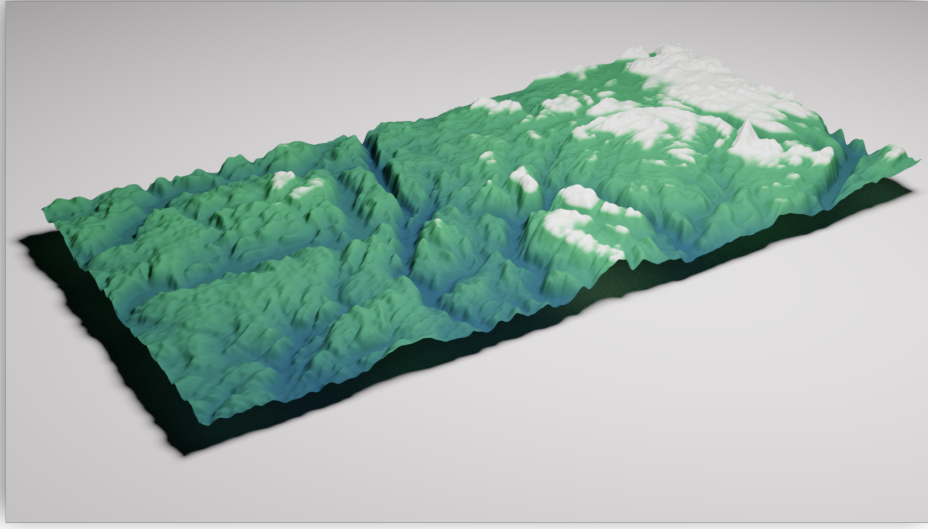


Figure 1: Real terrain data from a region close to Stavanger, Norway. For the regression tasks, we select data from small rectangular regions to keep the data-sizes manageable. The data is supplied in the file `SRTM_data_1_Norway.tif`.

3.1 Model statistics and resampling methods

We are interested in computing certain statistical quantities pertaining to our data and corresponding model function. Firstly, the *variance* of the model parameters β over all possible data sets tells us how sensitive β is to changes in the sampled data set. Using Equation (12) for the optimal OLS-parameters and setting $\mathbf{A} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$, we compute:

$$\text{Cov}[\beta_{\text{OLS}}] = \text{Cov}[\mathbf{A}\mathbf{y}] = \mathbf{A} \text{Cov}[\mathbf{y}] \mathbf{A}^\top = \mathbf{A} \mathbf{A}^\top \sigma^2 = (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2. \quad (26)$$

In this calculation, we have used Assumption 1 to deduce that $\text{Cov}[\mathbf{y}] = \text{diag}(\sigma)$, a diagonal matrix, which commutes with other matrices. It then follows that the estimated variance of β_{OLS} is given by

$$\text{Var}[\beta_{\text{OLS}}] = [\sigma^2 (\mathbf{X}^\top \mathbf{X})_{ii}^{-1}]_{i=1}^n. \quad (27)$$

Furthermore, by direct computation, we see that β_{OLS} is an *unbiased* estimator, as

$$\mathbb{E}[\beta_{\text{OLS}}] = \beta. \quad (28)$$

With both the expected value and the variance of β , we may formulate a confidence interval for β . Thus for an estimated parameter β_i we construct the 95%-confidence interval as

$$\begin{aligned} I_i &= (\mathbb{E}[\beta_i] - 1.96 \times \text{SD}[\beta_i], \mathbb{E}[\beta_i] + 1.96 \times \text{SD}[\beta_i]) \\ &= \left(\beta_i - 1.96 \times \sigma \sqrt{(\mathbf{X}^T \mathbf{X})_{ii}^{-1}}, \beta_i + 1.96 \times \sigma \sqrt{(\mathbf{X}^T \mathbf{X})_{ii}^{-1}} \right). \end{aligned} \quad (29)$$

Here $\text{SD}[\beta_i]$ denotes the standard deviation of β_i . The interpretation of this interval is “under repeated experiments, the number of computed confidence intervals that actually include the true parameter, will tend to 95%”.

We are also interested in computing the model bias and variance as discussed in Section 2.1. Numerically, we estimate these quantities by replacing the expectation values in Equation (19) by averages.

Cross-validation

In order to obtain more reasonable estimates for these statistical quantities, we perform so-called *resampling* of the data. There are several approaches to resampling, with two of the more common ones being *bootstrap* and *k-fold cross-validation*. Without going into details, the bootstrap method generally performs better with a significant data size, with cross-validation also performing well on smaller data sets. With this in mind, we will discuss cross-validation here.

Given a data-set $\mathbf{\Omega} = (\mathbf{x}_i, y_i)_{i=1}^n$, we divide it into two disjoint subset $\mathbf{\Omega}_{\text{train}}$ and $\mathbf{\Omega}_{\text{test}}$ with the purpose of optimizing our model by training on $\mathbf{\Omega}_{\text{train}}$ and evaluating model performance on the “unseen data” of $\mathbf{\Omega}_{\text{test}}$.

Cross-validation entails repeated such splitting of the data-set into test and training sets, evaluating model performance on each such split, and then finally aggregating the results to get a good estimate of model performance over the whole data-set. In *k-fold cross-validation*, the data is specifically subdivided into k equal parts, where each sub-set takes on the role of $\mathbf{\Omega}_{\text{test}}$ exactly once.

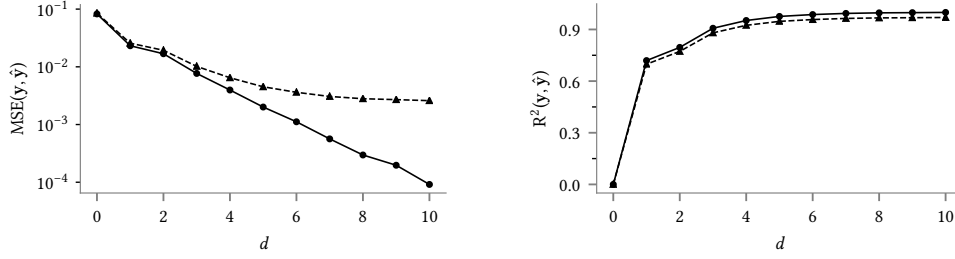


Figure 2: Ordinary least squares regression on samples from the Franke function. The mean squared error (left) and the R^2 -score (right) as a function of the polynomial degree d . The values are displayed both for the true sample data (solid) and the noisy samples (dashed). The noise is distributed as $\mathcal{N}(0, \frac{5}{100})$.

In our implementation we have made use of SKLEARN’s KFOLD method, with $k = 10$.

4 Numerical Results

In this section we present some of the results of the numerical simulations. We start simple by first considering the effects of model complexity — in terms of polynomial degree — using OLS regression on the Franke dataset. We assess the model accuracy by means of the mean squared error and the R^2 -score discussed previously. In order to increase the reliability of the statistical estimates, we employ k-fold cross validation.

We then take a closer look at Ridge and Lasso regression techniques, and find the optimal hyper-parameter λ . Again, we use the Franke dataset. In this context, we also see the effects of the bias-variance tradeoff.

Finally, we compare our methods on real world data, namely the terrain data in Figure 1.

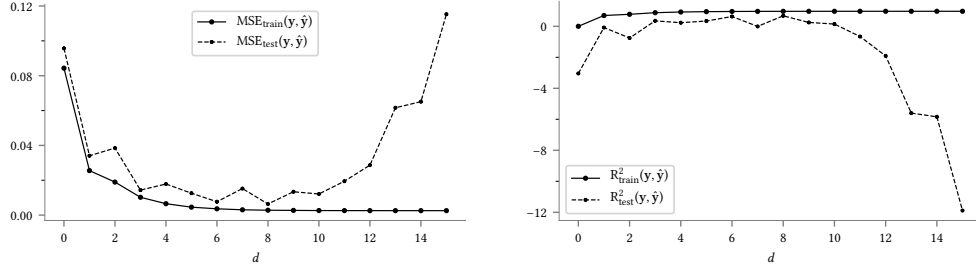


Figure 3: Ordinary least squares regression on samples from the Franke function with $N = 100$ points in each parameter direction, yielding a total data-size of $N^2 = 10000$ points. We have used K-fold cross validation with $K = 10$. The mean squared error (left) and the R^2 -score (right) as a function of the polynomial degree d . The values are displayed both for the training data (solid) and the test data (dashed). The noise is distributed as $\mathcal{N}(0, \frac{5}{100})$. Here we clearly see the effect of the bias-variance tradeoff. Note as the model complexity increases (higher polynomial degree), the model at some points overfits on the training data, yielding bad generalization scores, as signified by the sharp increase in MSE_{test} .

4.1 Toy Data

Ordinary Least Squares

We randomly sampled $N = 1000$ points in the unit-square and compute the corresponding N function values of the Franke function both with and without added noise. The noise was selected to be distributed as $\mathcal{N}(0, 1)$ with a signal-to-noise ratio $\alpha = 0.05$.

We then perform ordinary least squares, and compute the mean squared error as well as the R^2 -score for both the true and the noisy data. These are presented in Figure 2. With the Franke function being a weighted sum of exponentials, the MSE-error tends to zero as the polynomial degree increases. The added noise reduces the convergence rate, as is to be expected.

In addition to the evaluation metrics, we also compute the variance for each model parameter. These are displayed in Table 1 for each polynomial degree

d	2	3	4	5
1	5.315×10^{-5}	6.349×10^{-5}	7.072×10^{-5}	6.696×10^{-5}
x	4.673×10^{-4}	1.682×10^{-3}	4.398×10^{-3}	8.402×10^{-3}
y	4.673×10^{-4}	1.683×10^{-3}	4.396×10^{-3}	8.400×10^{-3}
x^2	3.686×10^{-4}	6.458×10^{-3}	4.694×10^{-2}	1.959×10^{-1}
xy	2.948×10^{-4}	4.106×10^{-3}	2.873×10^{-2}	1.195×10^{-1}
y^2	3.686×10^{-4}	6.460×10^{-3}	4.692×10^{-2}	1.959×10^{-1}
x^3		2.575×10^{-3}	8.868×10^{-2}	9.884×10^{-1}
x^2y		1.985×10^{-3}	5.199×10^{-2}	5.432×10^{-1}
xy^2		1.985×10^{-3}	5.199×10^{-2}	5.429×10^{-1}
y^3		2.575×10^{-3}	8.864×10^{-2}	9.884×10^{-1}
x^4			2.084×10^{-2}	1.078
x^3y			1.585×10^{-2}	6.287×10^{-1}
x^2y^2			1.528×10^{-2}	5.443×10^{-1}
xy^3			1.585×10^{-2}	6.285×10^{-1}
y^4			2.083×10^{-2}	1.078
x^5				1.658×10^{-1}
x^4y				1.253×10^{-1}
x^3y^2				1.193×10^{-1}
x^2y^3				1.192×10^{-1}
xy^4				1.253×10^{-1}
y^5				1.658×10^{-1}

Table 1: The variance of the model parameters β with respect to the polynomial degree d . The corresponding basis function is displayed to the left. Note in particular how the variance of the model parameters seems to increase with the polynomial degree, which is what is to be expected according to the *bias-variance tradeoff*.

from $d = 2$ up to and including $d = 5$.

Ridge and Lasso regression

We now incorporate Ridge and Lasso regression in our numerical experiments. Both of these regression techniques rely on a hyper-parameter λ which has to be tuned to the data at hand. By plotting the estimated bias and variance of the model as a function of the hyper-parameter λ we find the optimal parameter. The fits were performed on the same data as in the aforementioned OLS-test, and the results are presented in Figure 4. However, in this case we only use bi-quintic polynomial surfaces.

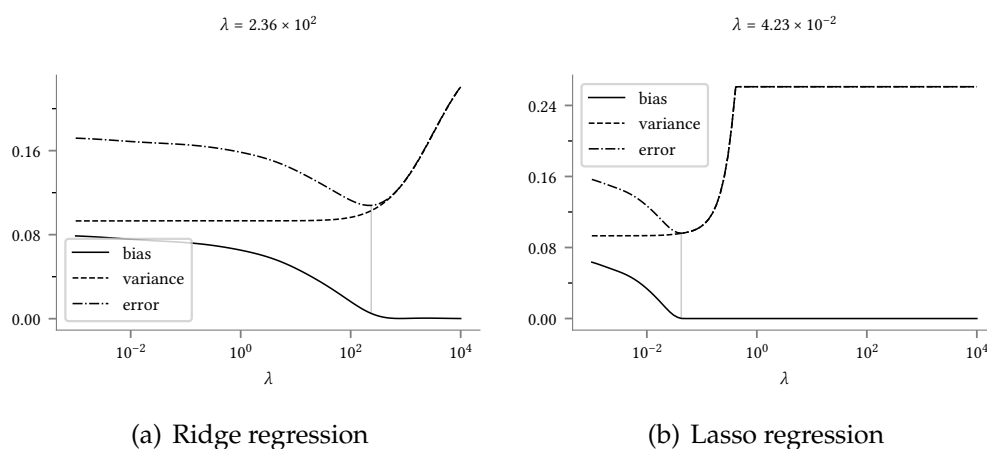


Figure 4: Finding the optimal hyper-parameter λ for both Ridge and Lasso regression. The fits were performed for bi-quintic polynomial surfaces with noisy data sampled from the Franke function. As we see, the variance is fairly constant, until it suddenly increases. The bias on the other hand converges to zero. The optimal lambda is indicated by a vertical line. Note how the error is constant after a certain value of λ . By looking at Figure 5, this is the point at which all Lasso-parameters are set to zero.

We also plot the model parameters β and their standard deviations as a function of λ . It is expected that the parameters β_{ridge} tends to zero as λ increases, while the parameters β_{lasso} may in addition be set equal to zero. This is verified by the plots in Figure 5.

In Table 2 we present the mean squared error and R^2 -score for fits performed with optimal hyper parameter λ as indicated in Figure 4 and compare them to

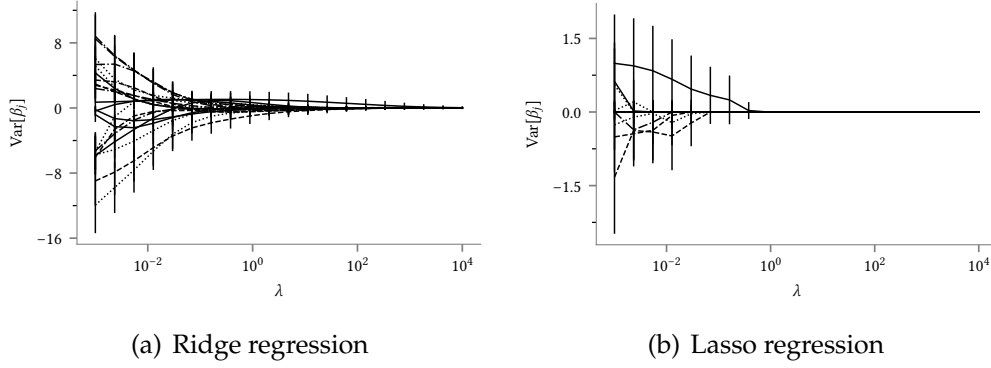


Figure 5: The model parameters β with associated standard deviations. We do not label each line to not clutter the plot, as its main purpose is to illustrate how Ridge and Lasso regression, respectively, penalizes model parameters. Note that for the Lasso regression, the model is quite sparse, and in particular, note how the λ where all parameters are zero correspond to the constant behaviour in Figure 4.

an OLS-fit.

Table 2: The MSE and R^2 score for the Franke function using bi-quintic polynomial surfaces and the optimal hyper parameters for Ridge and Lasso respectively. According to the reported metrics, the OLS-method performs significantly better than both Ridge and Lasso. The R^2 -scores reported for Ridge and Lasso seem unreasonably low compared to OLS, which may indicate an error in the implementation, or the authors lack of experience in the field.

	OLS	Ridge	Lasso
MSE	0.011	0.080	0.092
R^2	0.879	0.140	0.005

4.2 Terrain Data

When testing on real data, in order to keep the data-size manageable, we select a small patch from the large dataset. This patch is subsequently normalized in order to improve numerical properties of the methods. Unfortunately, due to time-constraints, and numerical bugs - we only present the MSE and R^2 scores for

the various methods with fixed hyper-parameters λ for Ridge and Lasso regression originally selected for the previous dataset. Ideally, hyper-parameter optimization as done in Figure 4 should be performed for this new data-set as well, and further statistical quantities reported.

The normalized patch is displayed in Figure 6, and the performance metrics in Figure 7.

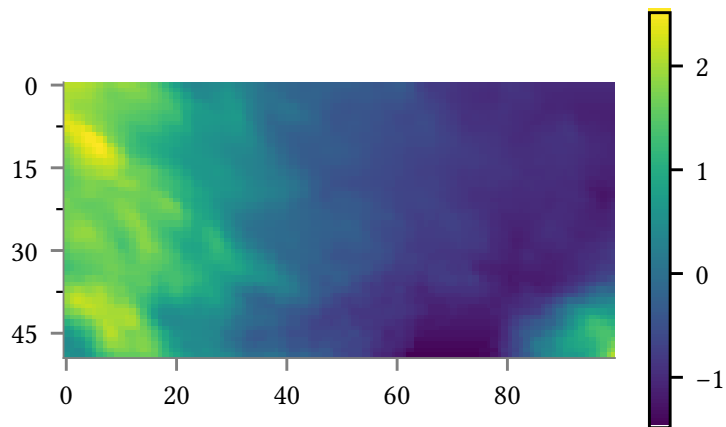


Figure 6: A small patch from the larger data-set displayed in Figure 1. The data has been normalized.

5 Conclusion

In this project we have studied three kinds of linear regression, namely Ordinary Least Squares, Ridge, and Lasso regression. We tested the methods on artificial data sampled from the Franke function, and later tested the methods on terrain data.

Due to some unfortunate implementation errors and time constraints, the study on the terrain data were rather uncomplete, however the one result we did

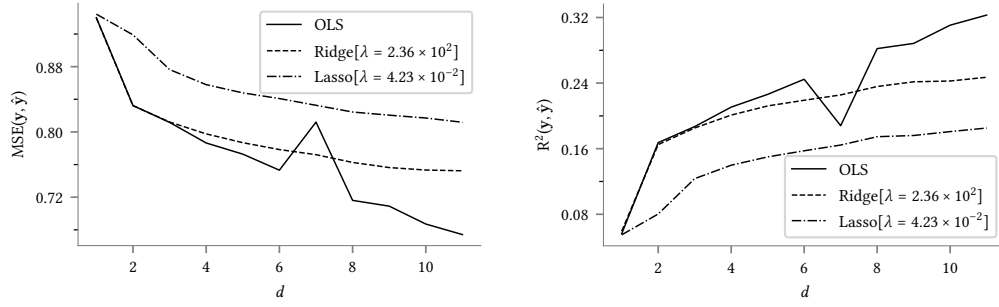


Figure 7: The performance metrics reported for polynomial fits to the small terrain patch. As we can see, the Ordinary Least Squares regression seems to be the better choice for terrain data with high local variance. It seems like both Ridge and Lasso struggles with finer features going missing, as they penalize large but “irrelevant” model parameters.

manage to get indicates that for data such as the terrain data presented here, the ordinary least squares method may be slightly better. All numerical methods were tested against SKLEARN’s methods, and the problems persisted, indicating rather that there is a more fundamental misunderstanding from the authors part.

5.1 Further work

In this project we have discussed three different methods of linear regression. In all three cases, we have fitted polynomial surfaces to the data sets. One drawback with using polynomial surfaces is that they are inherently global methods. Thus, in order to capture local trends in the data, a very complex model is required, which (in the polynomial case) means surfaces of higher polynomial degree. In recent years, several attempts at constructing efficient local methods have been proposed. Spline-based (piecewise polynomial) methods in particular have been important, and several approaches to locally refined spline spaces have started to gain momentum. We note in particular that of *hierarchical B-splines* (HB), *truncated hierarchical B-splines* (THB) and *locally refined B-splines* (LRB). These methods allow us to place more degrees of freedom in regions of high data-variance, thus locally increasing the model complexity.

A study on least-squares surfaces using LR B-splines for representing terrain data was presented in [SBD15]. Here, regions of high discrepancy between the true surface and the approximating surface was locally refined until an acceptable tolerance was reached.

It would be interesting to make a comparison of the three methods in such a simple setting as linear regression, as done in this project, taking into account adaptive refinement of the spline spaces with respect to the sampled data set.

References

- [Fra79] Richard Franke. *A Critical Comparison of Some Methods for Interpolation of Scattered Data*. NPS53-79-003. NAVAL POSTGRADUATE SCHOOL MONTEREY CA, Dec. 1979.
- [HK] Arthur E. Hoerl and Robert W. Kennard. *Ridge Regression: Biased Estimation for Nonorthogonal Problems: Technometrics: Vol 12, No 1*. URL: <https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634> (visited on 09/24/2019).
- [Meh+19] Pankaj Mehta et al. “A High-Bias, Low-Variance Introduction to Machine Learning for Physicists”. In: *Physics Reports* 810 (May 2019), pp. 1–124. ISSN: 03701573. arXiv: 1803.08823.
- [SBD15] Vibeke Skytt, Oliver Barrowclough, and Tor Dokken. “Locally Refined Spline Surfaces for Representation of Terrain Data”. In: *Computers & Graphics* 49 (June 1, 2015), pp. 58–68. ISSN: 0097-8493.
- [Tik+95] A. N. Tikhonov et al. *Numerical Methods for the Solution of Ill-Posed Problems*. Mathematics and Its Applications. Springer Netherlands, 1995. ISBN: 978-0-7923-3583-2.
- [vWie15] Wessel N. van Wieringen. “Lecture Notes on Ridge Regression”. In: (Sept. 30, 2015). arXiv: 1509.09169 [stat].