


REGRESSION ANALYSIS AND RESAMPLING METHODS

FYS-STK4155: PROJECT 1

Morten Ledum & Håkon Kristiansen

 github.com/mortele/FYS-STK4155

September 26, 2018

Abstract

We parameterize digital terrain data using linear regression analysis algorithms: Ordinary least squares (OLS), Ridge regression, and Lasso regression. The bootstrap resampling technique is used to gauge the bias and variance of the models. We use basis sets of homogeneous monomials in two variables, up to and including total degree 5. We find that xxxxxx.

For initial validation of our models, we employ the test function of R. Franke.¹

CONTENTS

I	Introduction	2
II	Theory	2
A	Linear regression	2
B	Ordinary least squares	2
	The design matrix	3
C	Polynomial basis sets	3
D	Ridge regression	3
	Singular-value decomposition	4
	Ridge regression using the SVD	4
E	Lasso regression	4
F	Principal components	5
G	Resampling and the <i>Bootstrap</i> method	6
III	Data sets	6
A	The Franke function	6
B	Terrain data	6
IV	Results and discussion	7
A	Verification of the models: The Franke function	7
B	Terrain data parametrization	7
V	Conclusion	7
	References	7

I. INTRODUCTION

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

II. THEORY

In the following we briefly introduce the theory underlying the technical aspects of the present work. We begin by considering linear regression in general, and the ordinary least squares (OLS) method.

A. Linear regression

In order to introduce the least squares methods, we consider a case in which p characteristics of n samples are measured. The outcome, or the *response*, is denoted \mathbf{y} : a vector of size n . The measured characteristics, denoted the predictors, are organized in a matrix \mathbf{X} of size $n \times p$. This is called the *design matrix*.

In regression analysis, we aim to explain the response in terms of the predictors, i.e. construct a function $\mathbf{y}(\mathbf{X})$. Assuming a linear relationship between \mathbf{X} and \mathbf{y} gives rise to *linear regression*, in which the response can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1)$$

where $\boldsymbol{\varepsilon}$ denotes the deviation of the linear model $\mathbf{X}\boldsymbol{\beta}$ and the response \mathbf{y} and $\boldsymbol{\beta}$ is a parameter vector containing the linear regression coefficients β_i . The β_i variables are the unknowns in the linear regression problem, and they represent the partial derivative of the *modelled* response w.r.t. the descriptors.

In any non-trivial case, the error terms ε_i in the error vector $\boldsymbol{\varepsilon}$ will be non-zero. In this case, we regard our linear ansatz as a *model* of

the true response, the observed values y_i . We denote our model by $\tilde{\mathbf{y}}$, and define

$$\begin{aligned} \tilde{\mathbf{y}} &= \mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y} - \boldsymbol{\varepsilon}. \end{aligned} \quad (2)$$

The objective of linear regression thus emerges: Determine $\boldsymbol{\beta}$ in such a way that $\boldsymbol{\varepsilon}$ is minimized, thus giving a best possible linear fit of the response (minimizing the deviation $|\mathbf{y} - \tilde{\mathbf{y}}|$).

B. Ordinary least squares

In order to *minimize* the error $\boldsymbol{\varepsilon}$, we must define exactly what that means. We require a functional expression—commonly referred to as the *cost function*—and a metric in which to calculate it's size. Choosing the Euclidean L^2 norm ($\|\mathbf{v}\|_2 = \sqrt{\sum_i v_i^2}$) and the absolute value of $\boldsymbol{\varepsilon}$ as the metric and cost function, respectively, leads to the *ordinary least squares* (OLS) method. Defining the cost function,

$$\begin{aligned} C(\boldsymbol{\beta}) &= \|\mathbf{y} - \tilde{\mathbf{y}}\|_2^2 \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \\ &= \sum_{i=1}^n \left| y_i - \beta_0 - \sum_{j=1}^p X_{ij}\beta_j \right|^2, \end{aligned} \quad (3)$$

we can formulate the OLS method as computing $\boldsymbol{\beta}_{\text{optimal}}$ by

$$\boldsymbol{\beta}_{\text{optimal}} = \arg \min_{\boldsymbol{\beta}} \{C(\boldsymbol{\beta})\}. \quad (4)$$

In order to find $\boldsymbol{\beta}_{\text{optimal}}$, we may simply differentiate $C(\boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}$ and enforce $\partial C(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = 0$. Following Hastie, Tibshirani & Friedman,² we find that

$$\begin{aligned} \frac{\partial C(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= -2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \stackrel{!}{=} 0 \\ \Rightarrow \mathbf{X}^T \mathbf{y} &= \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} \\ \boldsymbol{\beta}_{\text{optimal}} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}, \end{aligned} \quad (5)$$

where we have written $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ as $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. We note that even though $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a “large” matrix (assuming the number of observations $n \gg p$ the number of predictors per observation), the product $\mathbf{X}^T \mathbf{X} \in \mathbb{R}^{p \times p}$ is “small”. Thus explicitly inverting $\mathbf{X}^T \mathbf{X}$ is not a problem on a modern computer.

We note that the *model prediction* may now be calculated simply as $\tilde{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}_{\text{optimal}}$. This

represents the optimal linear model subject under the Euclidean norm of the cost function as given in Eq. (3). This method was first rigorously described by Legendre in 1805.³

The design matrix

The design matrix, \mathbf{X} , can in principle contain any set of linearly independent functions of the predictors¹. Every column in the design matrix corresponds to a mapping of the predictors, with elements $\mathbf{p}_i \mapsto \mathbf{X}_{ij}$. We will now consider an example of such a design matrix. We use two predictors—we will denote them x and y —with the response y . We introduce our model using the mappings $(x, y) \mapsto x$, $(x, y) \mapsto y$, and $(x, y) \mapsto xy$. Including also what is commonly referred to as the intercept, this gives rise to the design matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & y_1 & x_1 y_1 \\ 1 & x_2 & y_2 & x_2 y_2 \\ & & \vdots & \\ 1 & x_{n-1} & y_{n-1} & x_{n-1} y_{n-1} \\ 1 & x_n & y_n & x_n y_n \end{bmatrix}. \quad (6)$$

For inputs (x_i, y_i) our model $\tilde{\mathbf{y}}$ now returns

$$\tilde{\mathbf{y}}_i = \mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_i + \beta_2 y_i + \beta_3 x_i y_i, \quad (7)$$

where we used the shorthand notation \mathbf{x}_i^T to denote $\text{Row}_i(\mathbf{X})$.

Before we continue describing the Ridge and Lasso regression schemes, we briefly introduce the basis sets used in this project.

C. Polynomial basis sets

Throughout the present work we employ a basis set of homogeneous monomials². We will be working with 2D terrain data, and thus will need to consider monomials of up to and including two variables— x and y —in all possible homogeneous combinations. Disregarding the zero degree monomial, there are two possible such terms of degree up to and including one. These are simply x and y . Moving up to degree two, we must include x^2 , y^2 , and xy , for a total of five terms up to and including degree 2. Degree three adds an additional four terms: x^3 ,

$x^2 y$, xy^2 , and y^3 , and so on. In general, there are $n+1$ such terms for monomials of degree n , namely

$$x^n, x^{n-1}y, x^{n-2}y^2, \dots, xy^{n-1}, \text{ and } y^n.$$

The total basis sets of all such monomials of degree up to and including degree n — \mathcal{B}_n —thus contains

$$\text{size}(\mathcal{B}_n) \sum_{k=2}^{n+1} k = \frac{n(n+3)}{2} \quad (8)$$

terms.

D. Ridge regression

As mentioned in section B, defining exactly what we mean by *minimizing the error* requires a cost function and a metric. The previous choice of $C(\boldsymbol{\beta}) = \|\mathbf{y} - \tilde{\mathbf{y}}\|_2^2$ is obviously not the only possible one. In fact, a possibly superior method may be devised by keeping the Euclidean L^2 norm, but including a term in the cost function which penalizes large values of β_i . Such an approach was first used in statistics by Hoerl & Kennard,⁴ but was proposed already in the 1940s by Andrey Tikhonov.⁵ Taking the cost function to be

$$C_T(\boldsymbol{\beta}) = \|\mathbf{y} - \tilde{\mathbf{y}}\|_2^2 + \|\boldsymbol{\Gamma} \boldsymbol{\beta}\|_2^2 \quad (9)$$

gives rise to a regressions scheme known as Tikhonov regularization. The Tikhonov matrix $\boldsymbol{\Gamma}$ governs the form of the regularization term. A simple case of $\boldsymbol{\Gamma} = \sqrt{\lambda} \mathbf{1}$, favoring solutions with small (in the L^2 norm sense) values of the parameters $\boldsymbol{\beta}$, results in the *Ridge regression* of Hoerl & Kennard. The $\lambda \geq 0$ parameter here represents a tuneable penalty for large $\boldsymbol{\beta}$ values. We note that $\lambda = 0$ recovers the OLS method of section B.

Writing out the cost function of Ridge regression, we find that

$$\begin{aligned} C_R(\boldsymbol{\beta}) &= \|\mathbf{y} - \tilde{\mathbf{y}}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \\ &= \|\mathbf{y} - \mathbf{X} \boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \\ &= \sum_{i=1}^n \left| y_i - \beta_0 - \sum_{j=1}^p X_{ip} \beta_p \right|^2 + \lambda \sum_{j=1}^p \beta_j^2, \end{aligned} \quad (10)$$

¹We require linear independence to ensure the normal equations have a unique solution

²A homogeneous polynomial is a polynomial in which all terms have the same total degree, e.g. $xy + y^2 + x^2$ is a homogeneous polynomial, x is a homogeneous monomial, while $xy + x^3$ is *not*. Monomials are simply polynomials with only a single term.

where we have left out the intercept β_0 from the regularization term. This is done to ensure the solutions do not explicitly depend on the zero point chosen for \mathbf{y} , i.e. adding a constant to each response value y_i would not result in a simple shift of the predictions by the same amount.²

In the same way as before, we may differentiate the cost function and enforce $C_R(\beta) = 0$ in order to find the optimal β . We obtain

$$\begin{aligned}\frac{\partial C_R(\beta)}{\partial \beta} &= \frac{\partial}{\partial \beta} \left[(\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \mathbf{1}^T \beta \right] \\ &= -2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) + 2\lambda \mathbf{1} \stackrel{!}{=} 0 \\ \Rightarrow \mathbf{X}^T \mathbf{y} &= \mathbf{X}^T \mathbf{X} \beta + \lambda \mathbf{1} \beta \\ \beta_{\text{optimal}}^R &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{1})^{-1} \mathbf{X}^T \mathbf{y}.\end{aligned}\quad (11)$$

Singular-value decomposition

Before we continue, we introduce briefly the singular-value decomposition (SVD). Note carefully that *any* $m \times n$ matrix \mathbf{A} can be decomposed into a product like this.

Let \mathbf{A} be an $m \times n$ matrix with rank r . Then there exists an $m \times n$ matrix $\mathbf{\Sigma}$ in which the first r diagonal entries are the singular values of \mathbf{A} , $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$, and there exist an $m \times m$ orthogonal matrix \mathbf{U} and an $n \times n$ orthogonal matrix \mathbf{V} such that

$$\mathbf{A} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T. \quad (12)$$

All entries in $\mathbf{\Sigma}$ outside of the first r diagonal elements are zero. The singular values of \mathbf{A} denote the square roots of the eigenvalues of $\mathbf{A}^T \mathbf{A}$.⁶

The (first r columns of the) matrix \mathbf{U} contains the eigenvectors of $\mathbf{A}^T \mathbf{A}$ and represents an orthonormal basis for the column space of \mathbf{A} , $\text{Col } \mathbf{A}$. The (first r columns of the) matrix \mathbf{V} contains the eigenvectors of $\mathbf{A} \mathbf{A}^T$ and represents an orthonormal basis for the row space of \mathbf{A} , $\text{Row } \mathbf{A}$. The remaining $m - r$ and $n - r$ columns of \mathbf{U} and \mathbf{V} form orthonormal bases for $\text{Nul } \mathbf{A}$ and $\text{Nul } \mathbf{A}^T$.

Thus we can interpret the SVD loosely as finding the orthonormal bases of $\text{Col } \mathbf{A}$ and $\text{Row } \mathbf{A}$ such that application of \mathbf{A} maps $\mathbf{v}_i \mapsto \sigma_i \mathbf{u}_i$.

Ridge regression using the SVD

If we consider the SVD of \mathbf{X} , $\mathbf{X} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T$, and compute $\mathbf{X}^T \mathbf{X}$ we find that

$$\begin{aligned}\mathbf{X}^T \mathbf{X} &= (\mathbf{U} \mathbf{\Sigma} \mathbf{V}^T)^T (\mathbf{U} \mathbf{\Sigma} \mathbf{V}^T) \\ &= \mathbf{V} \mathbf{\Sigma}^T \mathbf{U}^T \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T \\ &= \mathbf{V} \mathbf{\Sigma}^T \mathbf{\Sigma} \mathbf{V}^T = \mathbf{V} \mathbf{\Sigma}^2 \mathbf{V}^T,\end{aligned}\quad (13)$$

where the orthogonality of \mathbf{U} made $\mathbf{U}^T \mathbf{U} = \mathbf{1}$. Inserting this into the expression for the optimal β_{optimal}^R (Eq. (11)) yields²

$$\begin{aligned}\beta_{\text{optimal}}^R &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{1})^{-1} \mathbf{X}^T \mathbf{y} \\ &= (\mathbf{V} \mathbf{\Sigma}^2 \mathbf{V}^T + \lambda \mathbf{1})^{-1} \mathbf{X}^T \mathbf{y} \\ &= (\mathbf{V} \mathbf{\Sigma}^2 \mathbf{V}^T + \lambda \mathbf{1} \mathbf{V} \mathbf{V}^T)^{-1} \mathbf{X}^T \mathbf{y} \\ &= (\mathbf{V} \mathbf{\Sigma}^2 \mathbf{V}^T + \lambda \mathbf{V} \mathbf{1} \mathbf{V}^T)^{-1} \mathbf{X}^T \mathbf{y},\end{aligned}$$

where we multiplied by $\mathbf{1} = \mathbf{V} \mathbf{V}^T$ (recall that $\mathbf{V}^{-1} = \mathbf{V}^T$ due to orthogonality) and used the fact that the identity matrix commutes with any other matrix. Furthermore, we find

$$\begin{aligned}\beta_{\text{optimal}}^R &= (\mathbf{V} [\mathbf{\Sigma}^2 + \lambda \mathbf{1}] \mathbf{V}^T)^{-1} \mathbf{X}^T \mathbf{y} \\ &= \mathbf{V} [\mathbf{\Sigma}^2 + \lambda \mathbf{1}]^{-1} \mathbf{V}^T \mathbf{X}^T \mathbf{y}\end{aligned}\quad (14)$$

$$\begin{aligned}&= \mathbf{V} [\mathbf{\Sigma}^2 + \lambda \mathbf{1}]^{-1} \mathbf{V}^T (\mathbf{U} \mathbf{\Sigma} \mathbf{V}^T)^T \mathbf{y} \\ &= \mathbf{V} [\mathbf{\Sigma}^2 + \lambda \mathbf{1}]^{-1} \mathbf{V}^T \mathbf{V} \mathbf{\Sigma}^T \mathbf{U}^T \mathbf{y} \\ &= \mathbf{V} [\mathbf{\Sigma}^2 + \lambda \mathbf{1}]^{-1} \mathbf{\Sigma} \mathbf{U}^T \mathbf{y},\end{aligned}\quad (15)$$

where we used that $\mathbf{\Sigma}$ is diagonal, so $\mathbf{\Sigma}^T = \mathbf{\Sigma}$. Note that since $(\mathbf{A} \mathbf{B})^{-1} = \mathbf{B}^{-1} \mathbf{A}^{-1}$, we can rewrite the inverse product

$$\begin{aligned}(\mathbf{V} [\mathbf{\Sigma}^2 + \lambda \mathbf{1}] \mathbf{V}^T)^{-1} &= (\mathbf{V}^T)^{-1} [\mathbf{\Sigma}^2 + \lambda \mathbf{1}]^{-1} \mathbf{V}^{-1} \\ &= \mathbf{V} [\mathbf{\Sigma}^2 + \lambda \mathbf{1}]^{-1} \mathbf{V}^T,\end{aligned}$$

as was done in Eq. (14). Since we are now taking the inverse of a diagonal matrix, we can simply write out the terms. Note that the inverse will itself be diagonal, and given by

$$\left([\mathbf{\Sigma}^2 + \lambda \mathbf{1}]^{-1} \right)_{ii} = \frac{1}{\sigma_{ii}^2 + \lambda}. \quad (16)$$

Rewriting the OLS scheme in terms of the SVD yields a very similar

$$\beta_{\text{optimal}} = \mathbf{V} (\mathbf{\Sigma}^2)^{-1} \mathbf{V}^T \mathbf{\Sigma}^T \mathbf{U}^T \mathbf{y}. \quad (17)$$

E. Lasso regression

As mentioned repeatedly, we are free to choose what we mean by *minimizing the error* w.r.t. what metric and what cost function we use.

Whereas the Ridge regression of section D used L^2 regularization by adding a $\lambda\|\boldsymbol{\beta}\|_2^2$ term to $C(\boldsymbol{\beta})$, we may instead try a L^1 regularization. This constitutes setting up the cost function as

$$\begin{aligned} C_L(\boldsymbol{\beta}) &= \|\mathbf{y} - \tilde{\mathbf{y}}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \\ &= \sum_{i=1}^n \left| y_i - \beta_0 - \sum_{j=1}^p X_{ip}\beta_p \right|^2 + \lambda \sum_{j=1}^p |\beta_j|^2. \end{aligned} \quad (18)$$

Originally popularized by Tibshirani,⁷ the *Lasso regression* has certain potential advantages over the OLS and Ridge regression schemes. Most notably, the Lasso can perform *variable selection*, i.e. some β_j s may be identically zero as a result of the minimization. The name Lasso is short for “least absolute shrinkage and selection operator”.

Computing the derivative of the Lasso cost function yields

$$\begin{aligned} \frac{\partial C_L(\boldsymbol{\beta})}{\partial \beta_j} &= \frac{\partial C_{OLS}(\boldsymbol{\beta})}{\partial \beta_j} + \lambda \sum_{j=1}^p \frac{\partial}{\partial \beta_j} |\beta_j| \\ &= \frac{\partial C_{OLS}(\boldsymbol{\beta})}{\partial \beta_j} + \lambda \frac{\beta_j}{\sqrt{\beta_j^2}} \\ &= \frac{\partial C_{OLS}(\boldsymbol{\beta})}{\partial \beta_j} + \lambda \operatorname{sgn}(\beta_j) \stackrel{!}{=} 0. \end{aligned} \quad (19)$$

In general, this can not be directly solved for $\boldsymbol{\beta}_{\text{optimal}}^L$ as in the case of OLS or the Ridge scheme. Under the assumption that \mathbf{X} is orthogonal, an explicit solution exists and is given by⁸

$$(\boldsymbol{\beta}_{\text{optimal}}^L(\lambda))_j = \operatorname{sgn}(\beta_j^{\text{OLS}}) (|\beta_j^{\text{OLS}}| - \lambda)_+, \quad (20)$$

where $(\cdot)_+$ represents the positive part of \cdot . In the general case, an iterative solver must be used to compute $\boldsymbol{\beta}_{\text{optimal}}^L$.

F. Principal components

The following section follows Hastie, Tibshirani & Friedman.²

Recall from section that the SVD matrix \mathbf{U} represents an orthonormal basis for the column space of the decomposed matrix \mathbf{A} . If we consider the prediction resulting from OLS, and

perform a SVD of \mathbf{X} we find that

$$\begin{aligned} \tilde{\mathbf{y}}_{\text{OLS}} &= \mathbf{X}\boldsymbol{\beta}_{\text{OLS}} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{X}[(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T)^T\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T]^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T[\mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^T]^{-1}\mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T\mathbf{y} \\ &= \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{V}[\boldsymbol{\Sigma}^2]^{-1}\mathbf{V}^T\mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T\mathbf{y} \\ &= \mathbf{U}\boldsymbol{\Sigma}[\boldsymbol{\Sigma}^2]^{-1}\mathbf{V}^T\mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T\mathbf{y} \\ &= \mathbf{U}\boldsymbol{\Sigma}[\boldsymbol{\Sigma}^2]^{-1}\boldsymbol{\Sigma}^T\mathbf{U}^T\mathbf{y} \\ &= \mathbf{U}\mathbf{U}^T\mathbf{y}. \end{aligned} \quad (21)$$

A similar derivation for $\tilde{\mathbf{y}}_R$ for the Ridge regression yields

$$\begin{aligned} \tilde{\mathbf{y}}_R &= \mathbf{X}\boldsymbol{\beta}_R \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{1})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{U}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^2 + \lambda\mathbf{1})^{-1}\boldsymbol{\Sigma}^T\mathbf{U}^T\mathbf{y} \\ &= \sum_{j=1}^p \mathbf{u}_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda} \mathbf{u}_j^T \mathbf{y}, \end{aligned} \quad (22)$$

where σ_j denotes the diagonal elements of the diagonal matrix $\boldsymbol{\Sigma}$, i.e. $(\boldsymbol{\Sigma})_{jj} = \sigma_j$. Note now that $\mathbf{U}^T\mathbf{y}$ are the coordinates of \mathbf{y} w.r.t. the orthonormal basis \mathbf{U} . Comparing Eqs. (21) and (22), we note that the coordinates of \mathbf{y} in both cases are computed in the orthonormal basis of $\text{Col } \mathbf{X}$ (as specified by the SVD matrix \mathbf{U}), but the Ridge scheme also *shrinks* the coordinates. The shrinkage is large whenever σ_j^2 is small. Recalling that $\mathbf{X}^T\mathbf{X} = \mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^T$ is an eigendecomposition of $\mathbf{X}^T\mathbf{X}$, with \mathbf{V} containing the eigenvectors, we denote these eigenvectors \mathbf{v}_j to be the *principal components* of \mathbf{X} . The eigenvalues contained in $\boldsymbol{\Sigma}^2$ are precisely the proportional factors, σ_j^2 , involved in the shrinkage.

The diagonalization of $\mathbf{X}^T\mathbf{X}$ constitutes a coordinate transform into a coordinate system in which $\mathbf{X}^T\mathbf{X}$ itself is diagonal. The matrix $\mathbf{X}^T\mathbf{X}$ is the covariance matrix (apart from a constant factor $1/N$). In the orthonormal basis of \mathbf{V} , the covariance matrix is diagonal and contains the variances σ_j^2 of these linear combinations of the columns of \mathbf{X} .

The first principal component direction has the property that $\mathbf{z}_1 = \mathbf{X}\mathbf{v}_1$ has the largest variance of all the linear combinations of the columns of \mathbf{X} . In general, the principal components are ordered such that $\text{Var } \mathbf{z}_1 \geq \text{Var } \mathbf{z}_2 \geq \dots \geq \text{Var } \mathbf{z}_N$.

In essence, the first principal component represent the direction in Col \mathbf{X} in which the variance is highest, the second principal component represents the corresponding direction in which the variance is highest, apart from the first, and so on. It is clear that the Ridge regression scheme simply rotates the OLS solution into the principal components of the design matrix, and then shrinks the coefficients corresponding to low-variance components.

G. Resampling and the *Bootstrap* method

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

III. DATA SETS

We are chiefly interested in parametrizing digital terrain data. However, in order to test and validate our implementation of the regression model and the resampling technique, we employ the Franke function¹ as a test case before considering real data.

A. The Franke function

The test function of Franke—originally developed to test and rate different surface interpolation techniques—is “a surface with a variety of behaviour” which consists of “two Gaussian peaks and a sharper Gaussian dip superimposed on a surface sloping towards the first quadrant.”¹ It is noted by Franke in the his original paper that the slope was introduced mainly as a visual aid and presumably had little impact on the actual interpolations performed.

³EarthExplorer website: <https://earthexplorer.usgs.gov/>.

More specifically, the Franke function $f_F(x, y)$ takes the full form

$$\begin{aligned} f_F(x, y) = & \frac{3}{4} \exp \left\{ \frac{-1}{4} \left[(9x - 2)^2 + (9y - 2)^2 \right] \right\} \\ & + \frac{3}{4} \exp \left\{ \frac{-1}{49} (9x + 1)^2 + \frac{1}{10} (9y + 1)^2 \right\} \\ & + \frac{1}{2} \exp \left\{ \frac{-1}{4} \left[(9x - 7)^2 + (9y - 3)^2 \right] \right\} \\ & - \frac{1}{5} \exp \left\{ \frac{-1}{4} \left[(9x + 4)^2 + (9y - 7)^2 \right] \right\}. \end{aligned} \quad (23)$$

A plot of the $f_F(x, y)$ surface can be seen in Fig. 1.

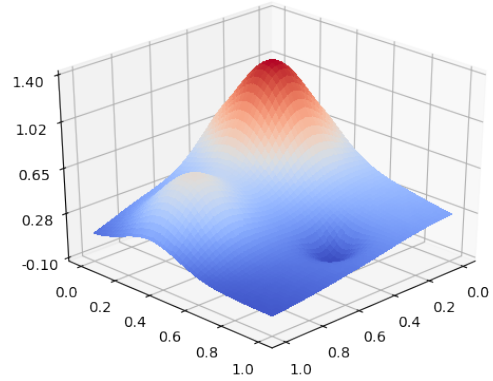


Figure 1: The Franke test function plotted for $0 \leq x, y \leq 1$.

B. Terrain data

The terrain data used is taken from the U.S. Department of the Interior U.S. Geological Survey’s (USGS) EarthExplorer³ website. The USGS stores data from the Shuttle Radar Topography Mission (SRTM) which maps the earth’s land surface topology with a resolution of 1 arc-second (about 30 m). We will use SRTM data taken from the EarthExplorer website as the basis for our terrain parametrization.

The specific terrain data we will use in the present project is taken from the Møsvatn Austfjell area in the municipality of Tinn in Telemark county, Norway. A visual representation of the data is shown in Fig. 2.

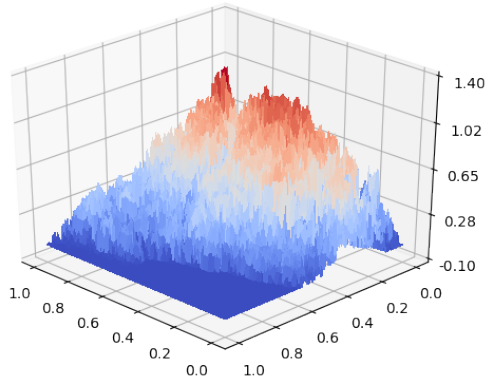


Figure 2: The terrain data in use in the present work, taken from the Møsvatn Austfjell area in the municipality of Tinn in Telemark county, Norway. Retrieved using the USGS EarthExplorer website. The height data is scaled to fit in $0 \leq z \leq 1$, and the reference zero point is set to zero.

IV. RESULTS AND DISCUSSION

A. Verification of the models: The Franke function

B. Terrain data parametrization

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis.

Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

V. CONCLUSION

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

REFERENCES

- [1] Richard Franke. *A critical comparison of some methods for interpolation of scattered data*. Tech. rep. Monterey, California: Naval Postgraduate School., 1979.
- [2] Trevor Hastie, Robert Tibshirani, and JH Friedman. *The elements of statistical learning: data mining, inference, and prediction*. 2009.
- [3] Adrien Marie Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805.
- [4] Arthur E Hoerl and Robert W Kennard. “Ridge regression: Biased estimation for nonorthogonal problems”. In: *Technometrics* 12.1 (1970), pp. 55–67.
- [5] Andrey Tikhonov. “On the stability of inverse problems”. In: *Doklady Akademii Nauk SSSR* 39.5 (1943), pp. 195–198.
- [6] David Lay. *Linear Algebra and Its Applications*. 4th ed. Pearson, 2012.
- [7] Robert Tibshirani. “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 00359246.
- [8] P. Mehta et al. “A high-bias, low-variance introduction to Machine Learning for physicists”. In: *ArXiv e-prints* (Mar. 2018). arXiv: [1803.08823](https://arxiv.org/abs/1803.08823) [[physics.comp-ph](https://arxiv.org/archive/physics)].

Table 1: Parameters β and their bootstrap computed variance $\sigma^2(\beta)$ for the OLS fits of the Franke function, shown in Fig. 3.

	$p = 2$		$p = 3$		$p = 4$		$p = 5$	
	β	$\sigma^2(\beta)$	β	$\sigma^2(\beta)$	β	$\sigma^2(\beta)$	β	$\sigma^2(\beta)$
1	1.17606	0.00007	0.99432	0.00014	0.60714	0.00006	0.37920	0.00052
x	-1.06533	0.00047	-0.60390	0.00314	4.23195	0.00478	8.05787	0.03989
y	-0.74196	0.00041	1.36710	0.00206	3.18895	0.00399	3.77385	0.03382
x^2	0.12032	0.00026	-1.40195	0.00885	-19.38216	0.05541	-35.01195	0.52045
xy	0.87501	0.00025	2.04306	0.00597	-2.28349	0.02825	-15.64235	0.50091
y^2	-0.36720	0.00029	-6.65928	0.00585	-12.53672	0.04434	-8.30869	0.44047
x^3			0.89177	0.00288	25.25628	0.10794	49.21854	1.61746
x^2y			0.36261	0.00219	8.10444	0.05690	45.87508	1.70397
xy^2			-1.50662	0.00188	1.40554	0.05242	21.25127	1.47419
y^3			4.69644	0.00207	12.61520	0.08471	-9.03064	1.50607
x^4					-10.84136	0.02548	-24.13509	1.25784
x^3y					-5.15114	0.01948	-54.81283	1.39124
x^2y^2					0.00764	0.01437	-8.23581	1.14134
xy^3					-1.90524	0.02013	-30.19186	1.12319
y^4					-3.51179	0.02032	30.20166	1.28260
x^5							1.53882	0.16041
x^4y							19.53504	0.18567
x^3y^2							10.83952	0.17047
x^2y^3							-5.29938	0.15935
xy^4							16.91243	0.15035
y^5							-16.84369	0.16920

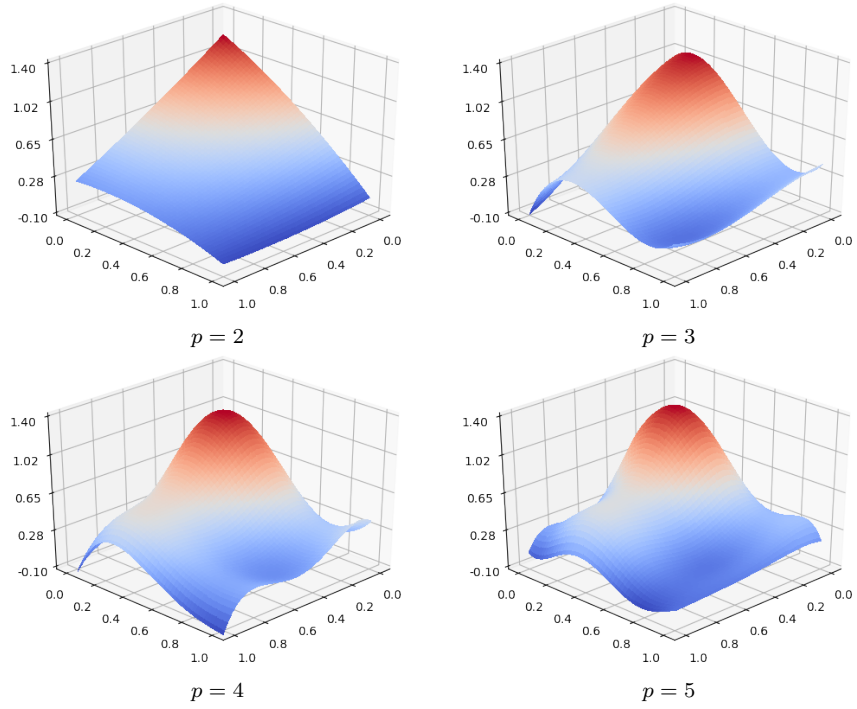


Figure 3: Ordinary least squares fits, using data from the Franke function with polynomials of degree 2, 3, 4, and 5. The p parameter indicates what order of polynomials are used. The target function of Franke can be seen in Fig. 1.

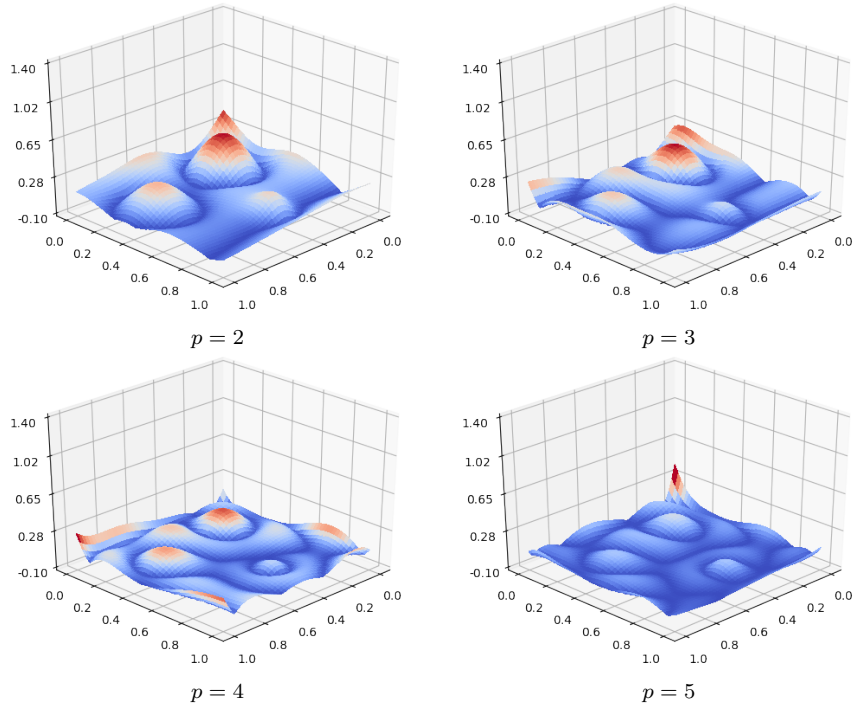


Figure 4: The absolute difference between the ordinary least squares fits of Fig. 3 and the true data, the Franke function. Polynomials of degree 2, 3, 4, and 5 have been used in the fitting. The p parameter indicates what order of polynomials are used.