Athens Institute for Education and Research
ATINER

ATINER's Conference Paper Series
MAT2018-2661

Teaching Ridge Regression in Polynomial Data Fitting

Diarmuid O’Driscoll
Head of Department of Mathematics and Computer Studies
Mary Immaculate College, Limerick
Ireland

An Introduction to ATINER's Conference Paper Series

This paper should be cited as follows:

O'Driscoll, D. (2019). **"Teaching Ridge Regression in Polynomial Data Fitting",** Athens: ATINER'S Conference Paper Series, No: **MAT2018-2661.**

Athens Institute for Education and Research
8 Valaoritou Street, Kolonaki, 10671 Athens, Greece
Tel: + 30 210 3634210 Fax: + 30 210 3634209 Email: info@atiner.gr URL: www.atiner.gr
URL Conference Papers Series: www.atiner.gr/papers.htm
Printed in Athens, Greece by the Athens Institute for Education and Research. All rights reserved. Reproduction is allowed for non-commercial purposes if the source is fully acknowledged.
ISSN: 2241-2891
02/09/2019

Teaching Ridge Regression in Polynomial Data Fitting

Diarmuid O'Driscoll

Abstract

The standard linear regression model can be written as $Y=X\beta+\varepsilon$ with uncorrelated zero mean and homoscedastic errors. Here X is a full rank n x p matrix containing the explanatory variables and the response vector y is n x 1 consisting of the observed data. The Ordinary Least Squared (OLS) estimators are given by $\widehat{\beta_L} = (X'X)^{-1}X'y$ and the Gauss-Markov Theorem states that$\widehat{\beta_L}$ is the best linear unbiased estimator. However, the OLS solutions require that $(X'X)^{-1}$ be accurately computed. In most real life situations, for example in engineering, economics and medicine, data is often given in discrete values along a continuum and it is necessary to find estimates at points between the discrete values. In particular, the data may suggest that a polynomial best represents the general trend of the data. If we try to fit a polynomial of too high a degree to a data set, containing noise, using OLS, then $(X'X)^{-1}$ will be numerically difficult to calculate and can lead to very unstable solutions. This paper will use the surrogate estimators of Jensen and Ramirez (2008) to 'control' the complexity of the model, to reduce the size of the confidence intervals of the parameters and prevent the polynomial from fitting the noise in the data. As the models are nested, the F-test will be used to compare the models.

Keywords: Collinearity; Ill-conditioning; Surrogate estimators.

Introduction

It is often the case that in many of our institutions, especially the smaller ones, time factors often prevent us, as teachers, from providing applications of topics covered in modules at undergraduate level for students of mathematics and statistics. Fitting polynomials to noisy data affords teachers an opportunity to combine topics from Linear Algebra, Multivariable Calculus and Introductory Statistics modules and show students the latent problems that arise from ill conditioning and also offer solutions as how to best deal with these problems. The method of least squares is commonly used in polynomial fitting to data. The standard linear regression model can be written as $Y=X\beta+\varepsilon$ with uncorrelated zero mean and homoscedastic errors,$\boldsymbol{\varepsilon}\sim N(0,\sigma)$. Here X is a n x p matrix, which must be of full rank to implement ordinary least squares method, containing the explanatory variables and the response vector y is n x 1 consisting of the observed data. The Ordinary Least Squared (OLS) estimators are given by $\widehat{\beta_L} = (X'X)^{-1}X'y$and the Gauss-Markov Theorem states that $\widehat{\beta_L}$ is the best linear unbiased estimator. However, the OLS solutions require that $(X'X)^{-1}$ be accurately computed and the variance-covariance matrix is given by$\sigma^2(X'X)^{-1}$. Hence ill conditioning can result in very unstable solutions and high variances for the estimators from which it is impossible to determine worthwhile confidence intervals. This paper will show how ridge regression is one method to avoid overfitting of a data set by a polynomial of too high a degree.

Methodology

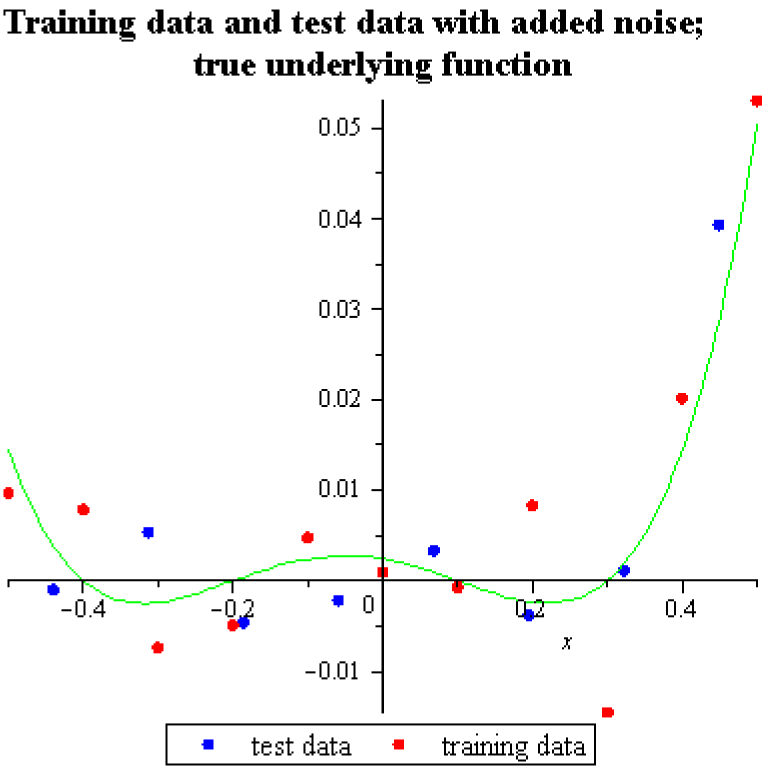
A set of eleven training data points and eight test data points were generated from the quartic polynomial

$$f(x) = x^4 + 0.2x^3 - 0.13x^2 - 0.014x + 0.0024$$

in the interval [-0.5,0.5] and noise is added to these data points from a normal distribution $N(0, 0.01)$.

The paper follows the educational idea of the inverted class, where you start with a finished product, such as the best fitting polynomial of degree 4 in Figure 1 and discuss the mathematics that are required to arrive at the final stage.

Figure 1. *Training and Test Points with Added Noise and Underlying Function*

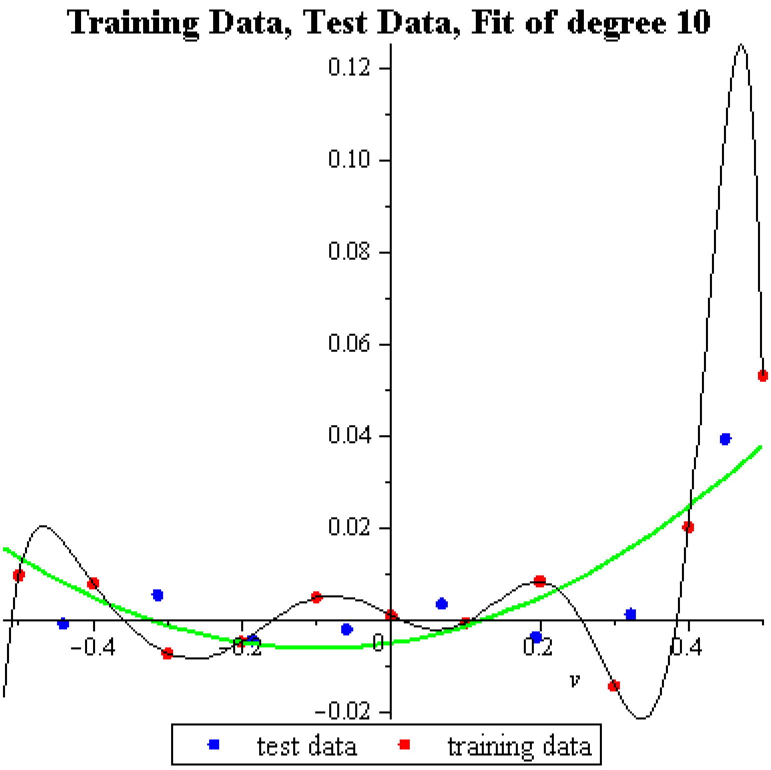


Polynomials of size 2, 4, 7, 9 and 10 are fitted to the training data set using OLS and the coefficients of these best fitting polynomials are displayed in Table 1.

Table 1. *OLS Estimators for Polynomial Coefficients of Different Degrees*

Est/Deg	*d=2*	*d=4*	*d=7*	*d=9*	*d=10*
$\widehat{\beta_0}$	-0.0054	0.0021	0.0057	-0.0004	0.0092
$\widehat{\beta_1}$	0.0242	-0.0159	0.0276	-0.0822	-0.0822
$\widehat{\beta_2}$	0.1239	-0.13764	-0.4171	0.5133	0.1157
$\widehat{\beta_3}$		0.2250	-0.7675	6.6169	6.6169
$\widehat{\beta_4}$		1.0426	4.1336	-16.9964	0.6882
$\widehat{\beta_5}$			5.8364	-121.9226	-121.9226
$\widehat{\beta_6}$			-8.1931	139.2335	-119.0019
$\widehat{\beta_7}$			-10.0372	771.7755	771.7755
$\widehat{\beta_8}$				-309.7198	1140.9142
$\widehat{\beta_9}$				-1527.6799	-1527.6799
$\widehat{\beta_{10}}$					-2702.1613

Figure 2. *Training Data Points, Test Data Points with Added Noise*



The *SSE* is first calculated for the training points and then for all data points (training and test) for the OLS estimators in Table 1 and the relative increase in the size of the errors is recorded in Table 2.

Table 2. *Relative Increase in SSE for Training Points and All Points*

Degree		*d=2*	*d=4*	*d=7*	*d=9*	*d=10*
Training	*SSE_T*	0.001252	0.000492	0.000339	0.000005	0.00000
All	*SSE_All*	0.004071	0.002257	0.000577	0.002975	0.007973
Rel_Incr		2.25	3.58	0.70	594	*N/A*

For *d* = 10, we note that the residual degrees of freedom are recorded as zero and over fitting has occurred, that is, the polynomial of degree 10 has fitted the data points (including noise) perfectly. However, due to the high oscillations between the training data points, it performs poorly in estimating the true underlying function as can be seen in Figures 1 and 2.

The estimators are very large and the polynomial behaves poorly at intermediate points and at the end points of the interval.

On the other hand, for *d* = 2, the fitted polynomial fails to capture the variation in the data and under fitting has occurred.

Ridge Regression

Ridge regression trades off some bias in the least squares estimators to gain a reduction in the variance of these estimators. Hoerl (1959) in Equation (1) limits the admissible size of the estimators to reduce the complexity of the model by adding the penalty term to the least squares problem as follows:

Minimize

$$||y - X\beta||^2 \text{ subject to } ||\beta||^2 = r^2, \qquad \text{Equation (1)}$$

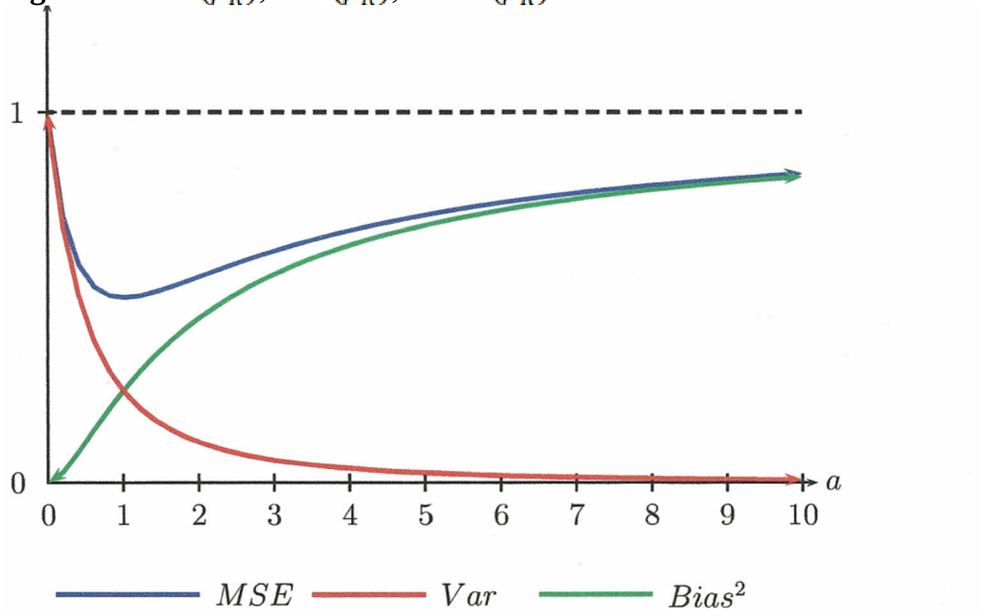
which is solved with Lagrange Multipliers and yields the biased solution

$$\widehat{\beta_R} = (X'X + \lambda I)^{-1}X'y. \qquad \text{Equation (2)}$$

Hoerl and Kennard (1970) proved that there always exist λ such that $MSE(\widehat{\beta_R}) < MSE(\widehat{\beta_L})$ and the main problem posed to researchers is to determine the optimal value of λ.

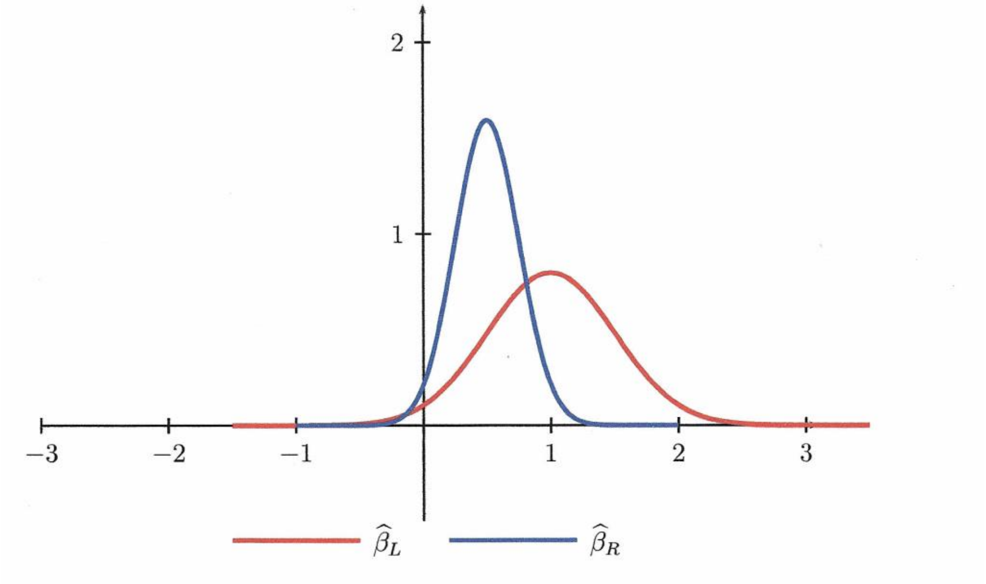
We illustrate the trade-off between bias and variance by the following example. If the unbiased OLS estimator $\widehat{\beta_L}$ is found to be $N(1,1)$ and the optimal choice for λ yields $\widehat{\beta_R} = 0.5\,\widehat{\beta_L}$, then the graphs of $MSE(\widehat{\beta_R})$, $Var(\widehat{\beta_R})$ and $Bias^2(\widehat{\beta_R})$ are illustrated in Figure 3 and the respective density functions are shown in Figure 4.

Figure 3. $MSE(\widehat{\beta_R}); Var(\widehat{\beta_R}); Bias^2(\widehat{\beta_R})$



Source: O'Driscoll and Ramirez (2016)

Figure 4. *Density Functions for* $\widehat{\beta_L}$ *and* $\widehat{\beta_R}$



Source: O'Driscoll and Ramirez (2016)

Writing $A_\lambda = X'X + \lambda I,$ the expected value and covariance of $\widehat{\beta_R}$ are

$E(\widehat{\beta_R}) = \beta - \lambda A_\lambda^{-1}\beta$ and $cov(\widehat{\beta_R}) = \sigma^2 A_\lambda^{-1} X'X A_\lambda^{-1}$.

As $\lambda \to 0,\ \widehat{\beta_R} \to \widehat{\beta_L}$ and as$\lambda \to \infty,\ \widehat{\beta_R} \to 0.$

Hoerl and Kennard (1970) established that the ridge estimators satisfy the *MSE Admissibility Condition* assuring an improvement in Mean Squared Error, $MSE(\widehat{\beta_R})$ for some$k \in (0,\infty)$. This result assures that for some positive value of *k*, the ridge model is an improved model.

However, Jensen and Ramirez (2010b) have shown the existence of cross-over values $k_0$for which, if $k > k_0$then $MSE(\widehat{\beta_R}) > MSE(\widehat{\beta_L})$.

The condition number of *X* is defined as $cn(X) = \dfrac{\max(eigenvalue(X'X))}{\min(eigenvalue(X'X))}$.

A high condition number indicates that the matrix is ill conditioned and suggests a high level of collinearity between the columns in the design matrix.

If we reduce the degree of the polynomial to six then the condition number of *X* is reduced to

$$cn(X) = 1.304 \times 10^7$$

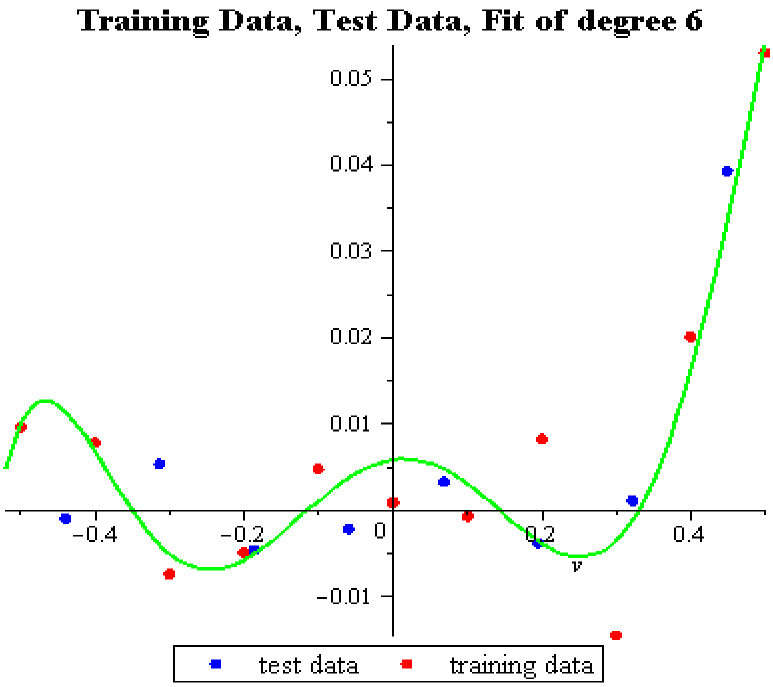
and the OLS estimator is

$$\widehat{\beta_l} = [0.0057, 0.0129, -0.4171, -0.2539, 4.1336, 1.5127, -8.1931].$$

However, due to the high condition number of X, the 95% confidence intervals for the 7 estimators are wide and each interval contains 0. In particular, the 95% confidence interval for $\widehat{\beta_6}$ is (-27.4741, 11.0879).

The sixth order polynomial is sketched in Figure 5 along with the training data points and test data points.

Figure 5. *OLS Polynomial of Degree 6*

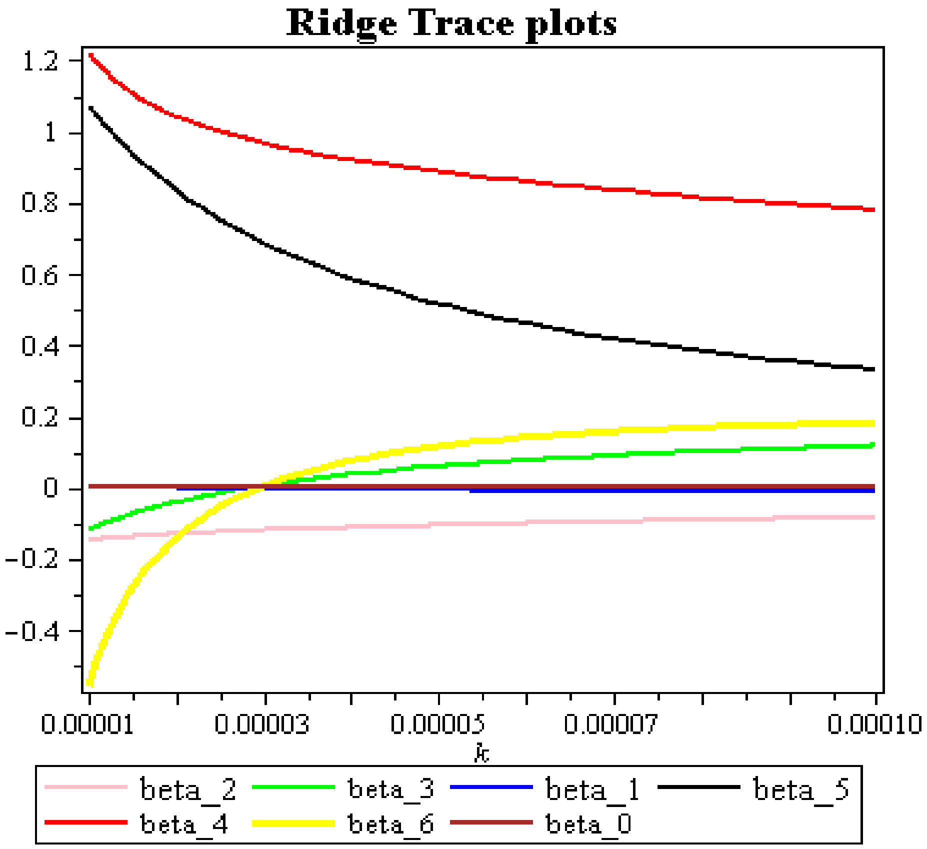


For a polynomial fit of degree 6, the *SSE* for the training data alone is 0.000345 which increases to 0.000668 when the test data are added to the fit, a relative increase of 0.94.

Ridge Regression and Penalty Term

To reduce the width of the confidence intervals we examine the ridge trace plots of the seven estimators. From Figure 6, we see that β_6 (yellow) and β_3 (green) change sign over the interval [0.00001, 0.00010], which indicates that the OLS solutions are unstable.

Figure 6. *Ridge Plots for the Seven Ridge Estimators*



From the ridge trace plots in Figure 5, we estimate that a steady state value for the λ parameter is approximately $\lambda = 0.00010$.

For this value of λ, the ridge estimator is given by

$$\widehat{\beta_R} = [0.0006, -0.0092, -0.0818,\ 0.1179, 0.7788, 0.3322, 0.1875]$$

and

$$\|\widehat{\beta_L}\| = 9.3135 \text{ is reduced to } \|\widehat{\beta_R}\| = 0.8785,$$

The instability of the solution vector combined with the fact that the 95% confidence interval for $\widehat{\beta_6}$ is (-27.4741, 11.0879) would suggest that the order of the polynomial might be reduced further.

Correlation Matrix

The correlation matrix is also a good indicator of collinearity between the columns of the design matrix. For this data set, the correlation matrix for the best fitting polynomial of degree 6 (excluding the constant column) is

$$\begin{pmatrix} 1.00 & 0.976 & 0.931 & 0.885 & 0.843 & 0.805 \\ & 1.00 & 0.987 & 0.962 & 0.933 & 0.904 \\ & & 1.00 & 0.993 & 0.977 & 0.958 \\ & & & 1.00 & 0.995 & 0.985 \\ & & & & 1.00 & 0.997 \\ & & & & & 1.00 \end{pmatrix}.$$

It is clear that there is a high correlation between each pair of columns, but it is not good practice to simply identify the pair of columns with the highest correlation and remove one of the columns. It is best to proceed and examine the ridge trace plots and the variance inflation factors of the design matrix.

Variance Inflation Factor

In general design matrices, the variance inflation factor for the k^{th} predictor is defined as

$$VIF_k = \frac{1}{1 - R_k^2}.$$

where R_k^2 is the R^2-value obtained by regressing the k^{th} predictor on the remaining predictors.

VIF is a measure of how much the variance of the estimated regression coefficient $\beta_k$is "inflated" by the existence of correlation among the predictor variables in the model. A VIF of 1 means that there is no correlation among the k^{th} predictor and the remaining predictor variables, and hence there is no inflation of the variance of β_k. There is no agreed acceptable value of VIF in the literature but values higher than 10 are considered to show that there is high collinearity in the design matrix.

Following O'Driscoll and Ramirez (2015), we view the design matrix $X = [X_{[p]}, x_p]$ with x_p the p^{th} column of X and $X_{[p]}$ the matrix formed by the remaining columns. The variance inflation factors measure the effect of adding column x_p to $X_{[p]}$. For notational convenience, we demonstrate with the last column p. An ideal column would be orthogonal to the previous columns with the entries in the off diagonal elements of the p^{th}row and p^{th} column of $X'X$ all zeros.

We denote M_p as the idealized moment matrix

$$M_p = \begin{bmatrix} X'_{[p]}X_{[p]} & 0_{p-1} \\ 0'_{p-1} & x'_p x_p \end{bmatrix}$$

from which it follows that

$$VIF\left(\widehat{\beta_p}\right) = \frac{\det(M_p)}{\det(X'X)}. \qquad \text{Equation (3)}$$

The variance inflation vector for the seven parameters is given by

$VIF = [5.23, 22.66, 203.69, 171.48,1009.52, 87.133, 391.61]$.

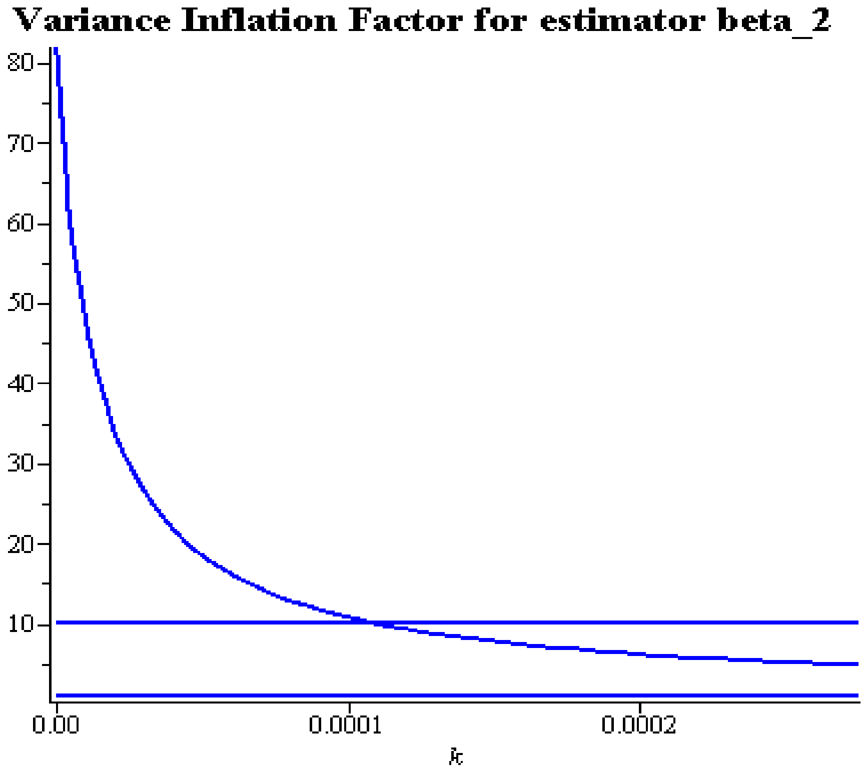
In the interval [0.00, 0.0001], the maximum variance inflation factor is given by 1009.52.As can be seen in Figure 7, by choosing $\lambda = 0.0001$, the VIF for $\widehat{\beta_2}$is reduced to 10.0 and reduced variance inflation vector becomes

$VIF = [3.35, 9.00, 10.00, 8.63,7.09, 2.85, 5.52]$

withassociated ridge estimator vector

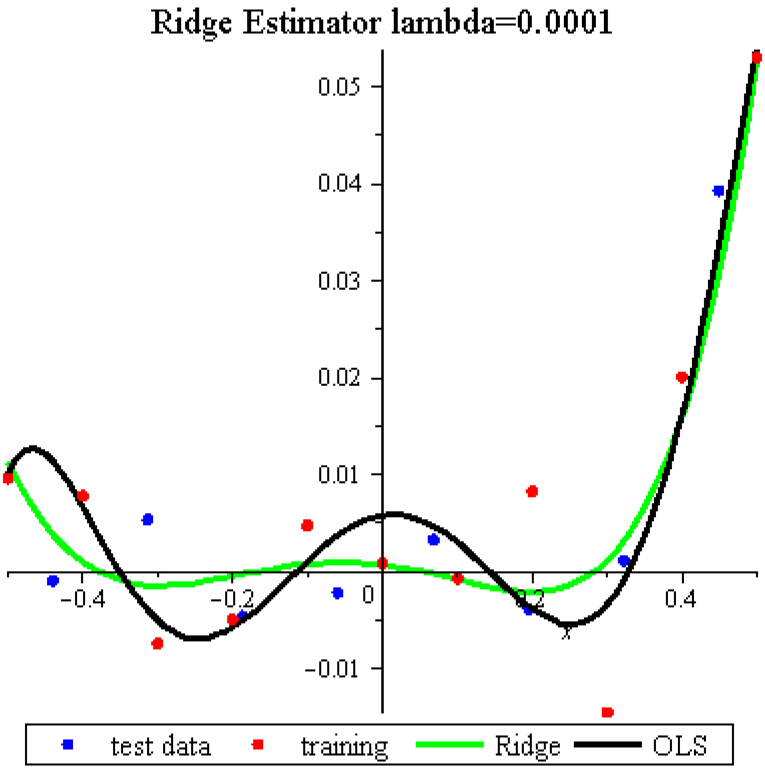
$\widehat{\beta_R} = [0.0006, -0.0092, -0.0818, 0.1179, 0.7788, 0.3322, 0.1848]$.

Figure 7. *Variance Inflation Factor for Beta_2*



The resultant polynomial is illustrated in Figure 8.

Figure 8. *OLS Polynomial of Degree 6 and Ridge Polynomial for* $\lambda = 0.0001$



Singular Value Decomposition and Surrogate Estimators

The singular value decomposition (SVD) of the $m \times p$ design matrix X can be written in the form
$X = UDV,$

where U is an $m \times m$ orthogonal matrix, V is a $p \times p$ orthogonal matrix and D is an $m \times p$ diagonal matrix with ordered diagonal entries $\sigma_n > \sigma_{n-1} > \sigma_{n-2} \dots > \sigma_2 > \sigma_1$, known as the singular values of X.

To alleviate the problems inherent with a singular value, say σ_p, whichis indicating collinearity in X, the surrogate estimators of Jensen and Ramirez modify the design matrix X on both sides of the OLS equation

$$X'X\beta = X'Y$$

by perturbing the singular values of X as

$\sigma_p \rightarrow \sqrt{\sigma_p^2 + \lambda}$
and thus moving the singular value away from zero.

The surrogate estimators greatly reduce the width of the confidence intervals for the estimators. Similar to ridge regression and writing $A_\lambda = X'X + \lambda I,$ the expected value and covariance of the surrogate estimator$\widehat{\beta_S}$ are

$$E(\widehat{\beta_S}) = A_\lambda^{-1} X_\lambda' X\beta - \beta \text{ and} cov(\widehat{\beta_S}) = \sigma^2 A_\lambda^{-1}.$$

Using the surrogate estimator for a polynomial fit of degree6 to the training data and with the same steady state value for $\lambda = 1.0 \times 10^{-4,}$ found in ridge regression,

$$\widehat{\beta_S} = [0.0016, -0.0027, -0.1310,\ 0.0067, 1.1651, 0.6873, -0.6043]$$

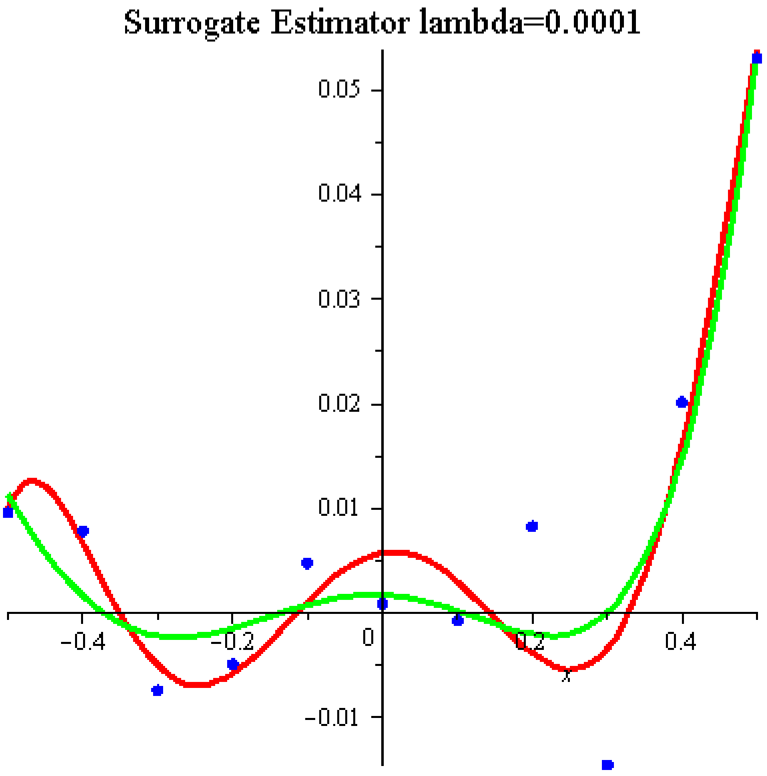
and

$$\|\widehat{\beta_L}\| = 9.3135 \text{ is reduced to } \|\widehat{\beta_R}\| = 1.4873.$$

The 95% confidence for estimator $\widehat{\beta_6}$ is further reduced from (-27.47, 11.08) to $(-2.70, 1.49)$ but still contains zero and there is not enough evidence to reject the null hypothesis that $\beta_6 = 0$.

The resultant polynomial is illustrated in Figure 9.

Figure 9. *Polynomial of Degree Six using Surrogate Estimator*



Based on the lack of evidence to reject the null hypothesis that$\beta_6 = 0$ and the sketches of the ridge and surrogate polynomials in Figures 8 and 9, our analysis suggests that a less complex modelof degree 4 may capture the true underlying function.

The condition number for the design matrix for a model of degree 4 reduces to

$$cn(X) = 2.86 \times 10^4$$

and the OLS estimator vector is

$$\widehat{\beta_L} = [0.0021, -0.0159, -0.1367, 0.2250,\ 1.0426]$$

with respective 95% confidence intervals of (-0.0092,0.0134), (-0.0641, 0.0323), (-0.3873, 0.1138), (-0.0243, 0.4744) and (0.0768, 2.0084) for each of the parameters.

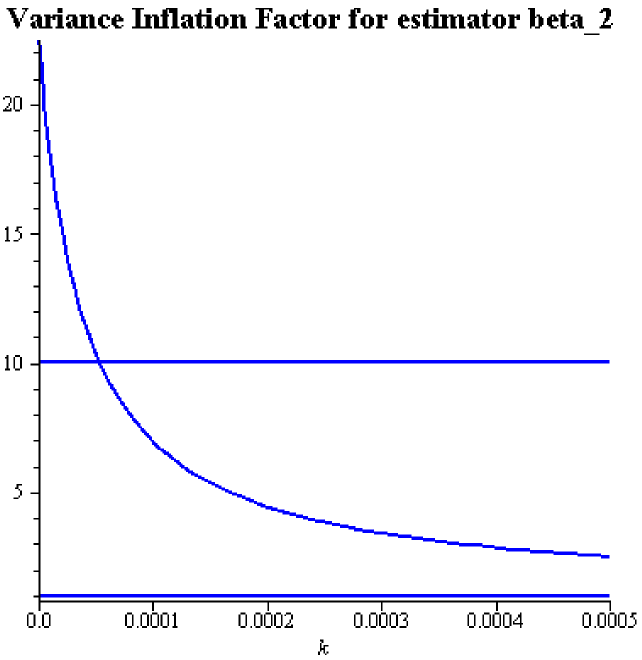
The variance inflation factor vector in this case is

$$VIF = [3.66, 6.64, 31.99, 6.64, 22.48].$$

To reduce the confidence limits further and to reduce the maximum variance inflation factor to 10, asteady state value for $\lambda = 5.0 \times 10^{-5}$. The associated reduced variance inflation factor vector is

$$VIF = [2.62, 5.83, 10.00, 5.83, 6.96].$$

Figure 10. *Variance Inflation Graph for* $\widehat{\beta_2}$



The surrogate estimator for the training data is

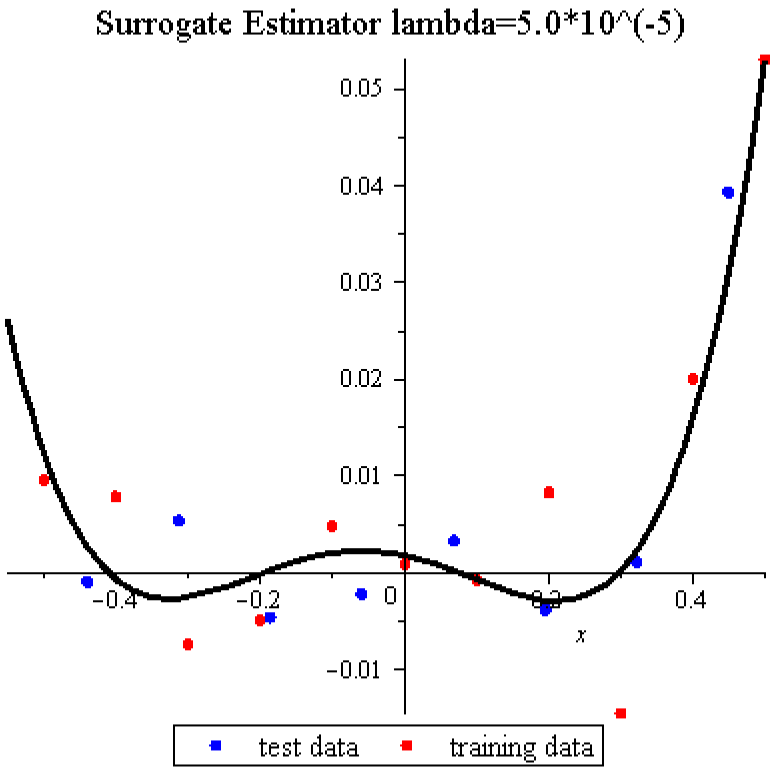
$$\widehat{\beta_S} = [0.0017, -0.0157, -0.1218,\ 0.2241, 0.9829]$$

with respective confidence intervals

(-0.0072, 0.0106), (-0.0533, 0.0218), (-0.2111, -0.0324), (0.0400, 0.4082) and (0.7234, 1.2424).

The polynomial of degree 4 associated with this surrogate estimator is shown in Figure 11.

Figure 11. *Polynomial of Degree 4 using Surrogate Estimator*



In this case, there is enough evidence to reject the null hypothesis that $\beta_4 = 0$. Using the *F* test with numerator*df* = 4 and denominator *df* = 6,

$$F = \frac{(SSE4 - SSE6)/(6-4)}{SSE6/(11-7)} = 1.0039,$$

we conclude that there is not enough evidence that either β_5 or β_6 differ from 0.

The SSE4 for the surrogate modelis 0.000518 and the SSE6for the OLS model of degree 6 is 0.000345.

Findings/Results

When there is collinearity in the design matrix in polynomial fit, the confidence intervals for the parameters are very wide and it is difficult to make worthwhile predictions from the estimators. The ridge and surrogate estimators help to reduce the width of the confidence intervals with the surrogate estimator performing better than the ridge estimator in estimating the underlying function.

As the ridge parameter λ increased, it was found that the norm of ridge estimator tended to zero more rapidly than the norm of the surrogate estimator.

For $\lambda = 5.0 \times 10^{-5}$,$\widehat{\beta_4} = 0.92$ in the case of ridge estimation while $\widehat{\beta_4} = 0.98$ in the case of surrogate, the true parameter being 1.0. The respective confidence intervals were (0.23,1.61) and (0.723,1.24). The coefficients for the true underlying function and the ridge parameters and surrogate parameters are shown in Table 3.

Table 3. *True Parameters; Ridge and Surrogate Parameters*

Parameters	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$
True	0.0024	-0.014	-0.13	0.2	1
OLS	0.0021	-0.0159	-0.1367	0.2250	1.0426
Ridge	0.0013	-0.0156	-0.1077	0.2250	0.9268
Surrogate	0.0017	-0.0157	-0.1218	0.2241	0.9829

Table 4. *95% Confidence Intervals for OLS, Ridge and Surrogate Parameters*

	OLS		Ridge		Surrogate	
	LC	UC	LC	UC	LC	UC
$\widehat{\beta_0}$	-0.01	0.01	-0.01	0.01	-0.01	0.01
$\widehat{\beta_1}$	-0.07	0.04	-0.07	0.04	-0.07	0.04
$\widehat{\beta_2}$	-0.42	0.15	-0.25	0.03	-0.32	0.07
$\widehat{\beta_3}$	-0.06	0.51	-0.04	0.49	-0.05	0.49
$\widehat{\beta_4}$	-0.05	2.13	0.45	1.41	0.26	1.71

The ridge graphs in Figure 6 also show how, in some cases, the OLS solutions are unstable, especially when the estimators change sign over a small interval.

Discussion

The purpose of this paper was to merge topics from different modules into a single topic. It follows the educational idea of the inverted class, where you start with a finished product, such as the best fitting polynomial of degree 4 and discuss the mathematics that are required to arrive at the final stage.

Conclusions

The surrogate estimators perform better than each of the OLS estimator and the ridge estimator in finding the polynomial of best fit to a given set of data. Since perturbation procedures are designed to improve the regression model, one would expect that$VIF(\widehat{\beta_R}) \rightarrow 1$ as $\lambda \rightarrow \infty$,but this is not always the case as shown by Jensen and Ramirez (2010a). However, in the case of the surrogate estimator, Jensen and Ramirez (2010a) proved that $VIF(\widehat{\beta_S}) \rightarrow 1$ as $\lambda \rightarrow \infty$,resulting inless collinearity between the surrogate estimators than exists between the OLSestimators.

References

Hoerl, A.E. 1959. *Optimum solution of Many Variable Equations.* Chemical Engineering Progress 58, 54-59.

Hoerl, A.E. and Kennard, R.W. 1970. *Ridge regression: biased estimation for nonorthogonal problems.* Technometrics, 12, 1 55-67.

Jensen, D.R. and Ramirez, D.E. 2008. *Anomalies in the foundations of ridge regression.* Int. Stat. Rev., 76, 89-105.

Jensen, D.R. and Ramirez, D.E. 2010a. *Surrogate models in ill-conditioned systems.* Journal of Statistical Planning and Inference, 140, 2069-2077.

Jensen, D.R. and Ramirez, D.E. 2010b. *Tracking MSE efficiencies in ridge regression.* Advances and Applications in Statistical Sciences, 1, 381-398.

O.Driscoll, D. and Ramirez, D. 2015. *Response surface design using the generalized variance inflation factors*. Cogent Mathematics, 2, 1-11.

O.Driscoll, D. and Ramirez, D. 2016. *Limitations of the Least Squares Estimators; A Teaching Perspective, ATINER's Conference Paper Series,* No: STA2016-2074 (2016).