

Lecture notes on ridge regression

Version 0.60, June 27, 2023.

arXiv:1509.09169v8 [stat.ME] 27 Jun 2023

Wessel N. van Wieringen^{1,2}

¹ Department of Epidemiology and Data Science,
Amsterdam Public Health research institute, Amsterdam UMC, location VUmc
P.O. Box 7057, 1007 MB Amsterdam, The Netherlands

² Department of Mathematics, Vrije Universiteit Amsterdam
De Boelelaan 1111, 1081 HV Amsterdam, The Netherlands
Email: w.vanwieringen@amsterdamumc.nl

License

This document is distributed under the Creative Commons Attribution-NonCommercial-ShareAlike license: <http://creativecommons.org/licenses/by-nc-sa/4.0/>



Disclaimer

This document is a collection of many well-known results on ridge regression. The current status of the document is ‘work-in-progress’ as it is incomplete (more results from literature will be included) and it may contain inconsistencies and errors. Hence, reading and believing at own risk. Finally, proper reference to the original source may sometimes be lacking. This is regrettable and these references – if ever known to the author – will be included in later versions.

Acknowledgements

Many people aided in various ways to the construction of these notes. Mark A. van de Wiel commented on various parts of Chapter 2. Jelle J. Goeman clarified some matters behind the method described in Section 6.4.3. Paul H.C. Eilers pointed to helpful references for Chapter 4 and provided parts of the code used in Section 4.3. Harry van Zanten commented on Chapters 1 and ??, while Stéphanie van de Pas made suggestions for the improvement of Chapters ?? and ?. Glenn Andrews thoroughly read Chapter 1 filtering out errors, unclarities, and inaccuracies.

Small typo's or minor errors, that have – hopefully – been corrected in the latest version, were pointed out by (among others): Rikkert Hindriks, Micah Blake McCurdy, José P. González-Brenes, Hassan Pazira, and numerous students from the *High-dimensional data analysis-* and *Statistics for high-dimensional data*-courses taught at Leiden University and the Vrije Universiteit Amsterdam, respectively.

Contents

1 Ridge regression	2
1.1 Linear regression	2
1.2 The ridge regression estimator	5
1.3 Eigenvalue shrinkage	9
1.3.1 Principal component regression	10
1.4 Moments	10
1.4.1 Expectation	11
1.4.2 Variance	12
1.4.3 Mean squared error	14
1.4.4 Debiasing	17
1.5 Constrained estimation	17
1.6 Degrees of freedom	21
1.7 Computationally efficient evaluation	21
1.8 Penalty parameter selection	22
1.8.1 Information criterion	23
1.8.2 Cross-validation	24
1.8.3 Generalized cross-validation	25
1.8.4 Randomness	25
1.9 Simulations	26
1.9.1 Role of the variance of the covariates	26
1.9.2 Ridge regression and collinearity	28
1.9.3 Variance inflation factor	30
1.10 Illustration	32
1.10.1 MCM7 expression regulation by microRNAs	33
1.11 Conclusion	37
1.12 Exercises	37
2 Bayesian regression	46
2.1 A minimum of prior knowledge on Bayesian statistics	46
2.2 Relation to ridge regression	47
2.3 Markov chain Monte Carlo	50
2.4 Empirical Bayes	55
2.5 Conclusion	56
2.6 Exercises	56
3 Generalizing ridge regression	59
3.1 Moments	64
3.2 The Bayesian connection	65
3.3 Application	65
3.4 Generalized ridge regression	68
3.5 Conclusion	69
3.6 Exercises	69
4 Mixed model	73
4.1 Link to ridge regression	78

4.2	REML consistency, high-dimensionally	79
4.3	Illustration: P-splines	81
5	Ridge logistic regression	85
5.1	Logistic regression	85
5.2	Separable and high-dimensional data	87
5.3	Ridge estimation	88
5.4	Moments	91
5.5	Constrained estimation	92
5.6	Degrees of freedom	92
5.7	The Bayesian connection	93
5.8	Computationally efficient evaluation	94
5.9	Penalty parameter selection	94
5.10	Generalizing ridge logistic regression	95
5.11	Application	96
5.12	Beyond logistic regression	98
5.13	Conclusion	99
5.14	Exercises	99
6	Lasso regression	103
6.1	Uniqueness	104
6.2	Analytic solutions	106
6.3	Sparcity	109
6.3.1	Maximum number of selected covariates	111
6.4	Estimation	112
6.4.1	Quadratic programming	112
6.4.2	Iterative ridge	113
6.4.3	Gradient ascent	115
6.4.4	Coordinate descent	116
6.5	Moments	117
6.6	Degrees of freedom	118
6.7	The Bayesian connection	118
6.8	Penalty parameter selection	120
6.8.1	Stability selection	120
6.9	Comparison to ridge	121
6.9.1	Linearity	121
6.9.2	Shrinkage	121
6.9.3	Simulation I: covariate selection	122
6.9.4	Simulation II: correlated covariates	123
6.9.5	Simulation III: sparsity	123
6.10	Application	126
6.11	Exercises	129
7	Generalizing lasso regression	133
7.1	Elastic net	133
7.2	Fused lasso	135
7.3	The (sparse) group lasso	136
7.4	Adaptive lasso	138
7.5	The ℓ_0 penalty	138
7.6	Conclusion	139
7.7	Exercises	139

1 Ridge regression

High-throughput techniques measure many characteristics of a single sample simultaneously. The number of characteristics p measured may easily exceed ten thousand. In most medical studies the number of samples n involved often falls behind the number of characteristics measured, i.e.: $p > n$. The resulting $(n \times p)$ -dimensional data matrix \mathbf{X} :

$$\mathbf{X} = (X_{*,1} | \dots | X_{*,p}) = \begin{pmatrix} X_{1,*} \\ \vdots \\ X_{n,*} \end{pmatrix} = \begin{pmatrix} X_{1,1} & \dots & X_{1,p} \\ \vdots & \ddots & \vdots \\ X_{n,1} & \dots & X_{n,p} \end{pmatrix}$$

from such a study contains a larger number of covariates than samples. When $p > n$ the data matrix \mathbf{X} is said to be *high-dimensional*, although no formal definition exists.

In this chapter we adopt the traditional statistical notation of the data matrix. An alternative notation would be \mathbf{X}^\top (rather than \mathbf{X}), which is employed in the field of (statistical) bioinformatics. In \mathbf{X}^\top the columns comprise the samples rather than the covariates. The case for the bioinformatics notation stems from practical arguments. A spreadsheet is designed to have more rows than columns. In case $p > n$ the traditional notation yields a spreadsheet with more columns than rows. When $p > 10000$ the conventional display is impractical. In this chapter we stick to the conventional statistical notation of the data matrix as all mathematical expressions involving \mathbf{X} are then in line with those of standard textbooks on regression.

The information contained in \mathbf{X} is often used to explain a particular property of the samples involved. In applications in molecular biology \mathbf{X} may contain microRNA expression data from which the expression levels of a gene are to be described. When the gene's expression levels are denoted by $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$, the aim is to find the linear relation $Y_i = \mathbf{X}_{i,*}\boldsymbol{\beta}$ from the data at hand by means of regression analysis. Regression is however frustrated by the high-dimensionality of \mathbf{X} (illustrated in Section 1.2 and at the end of Section 1.5). These notes discuss how regression may be modified to accommodate the high-dimensionality of \mathbf{X} . First, linear regression is recapitulated.

1.1 Linear regression

Consider an experiment in which p characteristics of n samples are measured. The data from this experiment are denoted \mathbf{X} , with \mathbf{X} as above. The matrix \mathbf{X} is called the *design matrix*. Additional information of the samples is available in the form of \mathbf{Y} (also as above). The variable \mathbf{Y} is generally referred to as the *response variable*. The aim of regression analysis is to explain \mathbf{Y} in terms of \mathbf{X} through a functional relationship like $Y_i = f(\mathbf{X}_{i,*})$. When no prior knowledge on the form of $f(\cdot)$ is available, it is common to assume a linear relationship between \mathbf{X} and \mathbf{Y} . This assumption gives rise to the *linear regression model*:

$$Y_i = \mathbf{X}_{i,*}\boldsymbol{\beta} + \varepsilon_i = \beta_1 X_{i,1} + \dots + \beta_p X_{i,p} + \varepsilon_i. \quad (1.1)$$

In model (1.1) $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ is the *regression parameter*. The parameter β_j , $j = 1, \dots, p$, represents the effect size of covariate j on the response. That is, for each unit change in covariate j (while keeping the other covariates fixed) the observed change in the response is equal to β_j . The second summand on the right-hand side of the model, ε_i , is referred to as the error. It represents the part of the response not explained by the functional part $\mathbf{X}_{i,*}\boldsymbol{\beta}$ of the model (1.1). In contrast to the functional part, which is considered to be systematic (i.e. non-random), the error is assumed to be random. Consequently, Y_i need not be equal $Y_{i'}$ for $i \neq i'$, even if $\mathbf{X}_{i,*} = \mathbf{X}_{i',*}$. To complete the formulation of model (1.1) we need to specify the probability distribution of ε_i . It is assumed that $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and the ε_i are independent, i.e.:

$$\text{Cov}(\varepsilon_i, \varepsilon_{i'}) = \begin{cases} \sigma^2 & \text{if } i = i', \\ 0 & \text{if } i \neq i'. \end{cases}$$

The randomness of ε_i implies that \mathbf{Y}_i is also a random variable. In particular, \mathbf{Y}_i is normally distributed, because $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\mathbf{X}_{i,*} \boldsymbol{\beta}$ is a non-random scalar. To specify the parameters of the distribution of \mathbf{Y}_i we need to calculate its first two moments. Its expectation equals:

$$\mathbb{E}(Y_i) = \mathbb{E}(\mathbf{X}_{i,*} \boldsymbol{\beta}) + \mathbb{E}(\varepsilon_i) = \mathbf{X}_{i,*} \boldsymbol{\beta},$$

while its variance is:

$$\begin{aligned} \text{Var}(Y_i) &= \mathbb{E}\{[Y_i - \mathbb{E}(Y_i)]^2\} &= \mathbb{E}(Y_i^2) - [\mathbb{E}(Y_i)]^2 \\ &= \mathbb{E}[(\mathbf{X}_{i,*} \boldsymbol{\beta})^2 + 2\varepsilon_i \mathbf{X}_{i,*} \boldsymbol{\beta} + \varepsilon_i^2] - (\mathbf{X}_{i,*} \boldsymbol{\beta})^2 &= \mathbb{E}(\varepsilon_i^2) \\ &= \text{Var}(\varepsilon_i) &= \sigma^2. \end{aligned}$$

Hence, $Y_i \sim \mathcal{N}(\mathbf{X}_{i,*} \boldsymbol{\beta}, \sigma^2)$. This formulation (in terms of the normal distribution) is equivalent to the formulation of model (1.1), as both capture the assumptions involved: the linearity of the functional part and the normality of the error.

Model (1.1) is often written in a more condensed matrix form:

$$\mathbf{Y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1.2)$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^\top$ and distributed as $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_p, \sigma^2 \mathbf{I}_{nn})$. As above model (1.2) can be expressed as a multivariate normal distribution: $\mathbf{Y} \sim \mathcal{N}(\mathbf{X} \boldsymbol{\beta}, \sigma^2 \mathbf{I}_{nn})$.

Model (1.2) is a so-called hierarchical model (not to be confused with the Bayesian meaning of this term). Here this terminology emphasizes that \mathbf{X} and \mathbf{Y} are not on a par, they play different roles in the model. The former is used to explain the latter. In model (1.1) \mathbf{X} is referred as the *explanatory* or *independent* variable, while the variable \mathbf{Y} is generally referred to as the *response* or *dependent* variable.

The covariates, the columns of \mathbf{X} , may themselves be random. To apply the linear model they are temporarily assumed fixed. The linear regression model is then to be interpreted as $\mathbf{Y} | \mathbf{X} \sim \mathcal{N}(\mathbf{X} \boldsymbol{\beta}, \sigma^2 \mathbf{I}_{nn})$

Example 1.1 (Methylation of a tumor-suppressor gene)

Consider a study which measures the gene expression levels of a tumor-suppressor genes (TSG) and two methylation markers (MM1 and MM2) on 67 samples. A methylation marker is a gene that promotes methylation. Methylation refers to attachment of a methyl group to a nucleotide of the DNA. In case this attachment takes place in or close by the promotor region of a gene, this complicates the transcription of the gene. Methylation may down-regulate a gene. This mechanism also works in the reverse direction: removal of methyl groups may up-regulate a gene. A tumor-suppressor gene is a gene that halts the progression of the cell towards a cancerous state.

The medical question associated with these data: do the expression levels methylation markers affect the expression levels of the tumor-suppressor gene? To answer this question we may formulate the following linear regression model:

$$Y_{i,\text{tsg}} = \beta_0 + \beta_{\text{mm1}} X_{i,\text{mm1}} + \beta_{\text{mm2}} X_{i,\text{mm2}} + \varepsilon_i,$$

with $i = 1, \dots, 67$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The interest focusses on β_{mm1} and β_{mm2} . A non-zero value of at least one of these two regression parameters indicates that there is a linear association between the expression levels of the tumor-suppressor gene and that of the methylation markers.

Prior knowledge from biology suggests that the β_{mm1} and β_{mm2} are both non-positive. High expression levels of the methylation markers lead to hyper-methylation, in turn inhibiting the transcription of the tumor-suppressor gene. Vice versa, low expression levels of MM1 and MM2 are (via hypo-methylation) associated with high expression levels of TSG. Hence, a negative concordant effect between MM1 and MM2 (on one side) and TSG (on the other) is expected. Of course, the methylation markers may affect expression levels of other genes that in turn regulate the tumor-suppressor gene. The regression parameters β_{mm1} and β_{mm2} then reflect the indirect effect of the methylation markers on the expression levels of the tumor suppressor gene. \square

The linear regression model (1.1) involves the unknown parameters: $\boldsymbol{\beta}$ and σ^2 , which need to be learned from the data. The parameters of the regression model, $\boldsymbol{\beta}$ and σ^2 are estimated by means of likelihood maximization. Recall that $Y_i \sim \mathcal{N}(\mathbf{X}_{i,*} \boldsymbol{\beta}, \sigma^2)$ with corresponding density: $f_{Y_i}(y_i) = (2\pi\sigma^2)^{-1/2} \exp[-(y_i - \mathbf{X}_{i,*} \boldsymbol{\beta})^2/2\sigma^2]$. The likelihood thus is:

$$L(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^n (\sqrt{2\pi\sigma^2})^{-1} \exp[-(Y_i - \mathbf{X}_{i,*} \boldsymbol{\beta})^2/2\sigma^2],$$

in which the independence of the observations has been used. Because of the strict monotonicity of the logarithm, the maximization of the likelihood coincides with the maximum of the logarithm of the likelihood (called the log-likelihood). Hence, to obtain maximum likelihood estimates of the parameter it is equivalent to find the maximum of the log-likelihood. The log-likelihood is:

$$\mathcal{L}(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2) = \log[L(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2)] = -n \log(\sqrt{2\pi}\sigma) - \frac{1}{2}\sigma^{-2} \sum_{i=1}^n (y_i - \mathbf{X}_{i,*}\boldsymbol{\beta})^2.$$

After noting that $\sum_{i=1}^n (Y_i - \mathbf{X}_{i,*}\boldsymbol{\beta})^2 = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$, the log-likelihood can be written as:

$$\mathcal{L}(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2) = -n \log(\sqrt{2\pi}\sigma) - \frac{1}{2}\sigma^{-2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

In order to find the maximum of the log-likelihood, take its derivative with respect to $\boldsymbol{\beta}$:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathcal{L}(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2) = -\frac{1}{2}\sigma^{-2} \frac{\partial}{\partial \boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \sigma^{-2} \mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Equate this derivative to zero gives the estimating equation for $\boldsymbol{\beta}$:

$$\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top \mathbf{Y}. \quad (1.3)$$

Equation (1.3) is called to the *normal equation*. Pre-multiplication of both sides of the normal equation by $(\mathbf{X}^\top \mathbf{X})^{-1}$ now yields the maximum likelihood estimator of the regression parameter: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$, in which it is assumed that $(\mathbf{X}^\top \mathbf{X})^{-1}$ is well-defined.

Along the same lines one obtains the maximum likelihood estimator of the residual variance. Take the partial derivative of the loglikelihood with respect to σ^2 :

$$\frac{\partial}{\partial \sigma} \mathcal{L}(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2) = -\sigma^{-1} + \sigma^{-3} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

Equate the right-hand side to zero and solve for σ^2 to find $\hat{\sigma}^2 = n^{-1} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$. In this expression $\boldsymbol{\beta}$ is unknown and the maximum likelihood estimate of $\boldsymbol{\beta}$ is plugged-in.

With explicit expressions of the maximum likelihood estimators at hand, we can study their properties. The expectation of the maximum likelihood estimator of the regression parameter $\boldsymbol{\beta}$ is:

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}] = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}(\mathbf{Y}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Hence, the maximum likelihood estimator of the regression coefficients is unbiased.

The variance of the maximum likelihood estimator of $\boldsymbol{\beta}$ is:

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}\{[\hat{\boldsymbol{\beta}} - \mathbb{E}(\hat{\boldsymbol{\beta}})][\hat{\boldsymbol{\beta}} - \mathbb{E}(\hat{\boldsymbol{\beta}})]^\top\} \\ &= \mathbb{E}\{[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} - \boldsymbol{\beta}][(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} - \boldsymbol{\beta}]^\top\} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}\{\mathbf{Y}\mathbf{Y}^\top\} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}^\top \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \{\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^\top \mathbf{X}^\top + \Sigma\} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}^\top \\ &= \boldsymbol{\beta}\boldsymbol{\beta}^\top + \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}^\top \\ &= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}, \end{aligned}$$

in which we have used that $\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top) = \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^\top \mathbf{X}^\top + \sigma^2 \mathbf{I}_{nn}$. From $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$, one obtains an estimate of the variance of the estimator of the j -th regression coefficient: $\hat{\sigma}^2(\hat{\beta}_j) = \hat{\sigma}^2[(\mathbf{X}^\top \mathbf{X})^{-1}]_{jj}$. This may be used to construct a confidence interval for the estimates or test the hypothesis $H_0 : \beta_j = 0$. In the latter display, $\hat{\sigma}^2$ should not be the maximum likelihood estimator, but is to be replaced by the residual sum-of-squares divided by $n - p$ rather than n . The *residual sum-of-squares* is defined as $\sum_{i=1}^n (Y_i - \mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})^2$.

The prediction of Y_i , denoted \hat{Y}_i , is the expected value of Y_i according the linear regression model (with its parameters replaced by their estimates). The prediction of Y_i thus equals $\mathbb{E}(Y_i; \hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \mathbf{X}_{i,*}\hat{\boldsymbol{\beta}}$. In matrix notation the prediction is:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} := \mathbf{H}\mathbf{Y},$$

where \mathbf{H} is the *hat matrix*, as it ‘puts the hat’ on \mathbf{Y} . Note that the hat matrix is a projection matrix, i.e. $\mathbf{H}^2 = \mathbf{H}$ for

$$\mathbf{H}^2 = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top.$$

Moreover, $\mathbf{H}^\top = \mathbf{H}$. Thus, the prediction $\widehat{\mathbf{Y}}$ is an orthogonal projection of \mathbf{Y} onto the space spanned by the columns of \mathbf{X} .

With $\widehat{\boldsymbol{\beta}}$ available, an estimate of the errors $\hat{\varepsilon}_i$, dubbed the *residuals* are obtained via:

$$\hat{\varepsilon} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y} = [\mathbf{I} - \mathbf{H}] \mathbf{Y}.$$

Thus, the residuals are a projection of \mathbf{Y} onto the orthogonal complement of the space spanned by the columns of \mathbf{X} . The residuals are to be used in diagnostics, e.g. checking of the normality assumption by means of a normal probability plot.

For more on the linear regression model confer the monograph of Draper and Smith (1998).

1.2 The ridge regression estimator

The ridge regression estimator, originally proposed to deal with collinearity, has seen a renewed interest since the advent of high-dimensional data. For, if the design matrix is high-dimensional, the covariates (the columns of \mathbf{X}) are super-collinear. Recall *collinearity* in regression analysis refers to the event of two (or multiple) covariates being strongly linearly related. Consequently, the space spanned by super-collinear covariates is a lower-dimensional subspace of the parameter space. The design matrix \mathbf{X} is then (close to) rank deficient and it is (almost) impossible to separate the contribution of the individual covariates. The uncertainty, with respect to the covariate responsible for the variation explained in \mathbf{Y} , is reflected in the fit of the linear regression model to data. Collinearity reveals itself in the fit through a large error of the regression parameters’ estimates corresponding to the collinear covariates and, consequently, usually accompanied by large values of the estimates.

Example 1.2 (Collinearity)

The flotillins (the FLOT-1 and FLOT-2 genes) have been observed to regulate the proto-oncogene ERBB2 *in vitro* (Pust *et al.*, 2013). One may wish to corroborate this *in vivo*. To this end we use gene expression data of a breast cancer study, available as a Bioconductor package: *breastCancerVDX*. From this study the expression levels of probes interrogating the FLOT-1 and ERBB2 genes are retrieved. For clarity of the illustration the FLOT-2 gene is ignored. After centering, the expression levels of the first ERBB2 probe are regressed on those of the four FLOT-1 probes. The R-code below carries out the data retrieval and analysis.

Listing 1.1 R code

```
# load packages
library(Biobase)
library(breastCancerVDX)

# load data into memory
data(vdx)

# ids of genes FLOT1
idFLOT1 <- which(fData(vdx)[,5] == 10211)

# ids of ERBB2
idERBB2 <- which(fData(vdx)[,5] == 2064)

# get expression levels of probes mapping to FLOT genes
X <- t(exprs(vdx)[idFLOT1,])
X <- sweep(X, 2, colMeans(X))

# get expression levels of probes mapping to FLOT genes
Y <- t(exprs(vdx)[idERBB2,])
Y <- sweep(Y, 2, colMeans(Y))
```

```
# regression analysis
summary(lm(formula = Y[,1] ~ X[,1] + X[,2] + X[,3] + X[,4]))

# correlation among the covariates
cor(X)
```

Prior to the regression analysis, we first assess whether there is collinearity among the FLOT-1 probes through evaluation of the correlation matrix. This reveals a strong correlation ($\hat{\rho} = 0.91$) between the second and third probe. All other cross-correlations do not exceed the 0.20 (in an absolute sense). Hence, there is strong collinearity among the columns of the design matrix in the to-be-performed regression analysis.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)		
(Intercept)	0.0000	0.0633	0.0000	1.0000		
X[, 1]	0.1641	0.0616	2.6637	0.0081 **		
X[, 2]	0.3203	0.3773	0.8490	0.3965		
X[, 3]	0.0393	0.2974	0.1321	0.8949		
X[, 4]	0.1117	0.0773	1.4444	0.1496		

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ''	1

Residual standard error: 1.175 on 339 degrees of freedom

Multiple R-squared: 0.04834, Adjusted R-squared: 0.03711

F-statistic: 4.305 on 4 and 339 DF, p-value: 0.002072

The output of the regression analysis above shows the first probe to be significantly associated to the expression levels of ERBB2. The collinearity of the second and third probe reveals itself in the standard errors of the effect size: for these probes the standard error is much larger than those of the other two probes. This reflects the uncertainty in the estimates. Regression analysis has difficulty to decide to which covariate the explained proportion of variation in the response should be attributed. The large standard error of these effect sizes propagates to the testing as the Wald test statistic is the ratio of the estimated effect size and its standard error. Collinear covariates are thus less likely to pass the significance threshold. \square

The case of two (or multiple) covariates being perfectly linearly dependent is referred to as *super-collinearity*. The rank of a high-dimensional design matrix is maximally equal to n : $\text{rank}(\mathbf{X}) \leq n$. Consequently, the dimension of subspace spanned by the columns of \mathbf{X} is smaller than or equal to n . As $p > n$, this implies that columns of \mathbf{X} are linearly dependent. Put differently, a high-dimensional \mathbf{X} suffers from super-collinearity.

Example 1.3 (Super-collinearity)

Consider the design matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 2 \\ 1 & 0 & 1 \\ 1 & 2 & -1 \\ 1 & 1 & 0 \end{pmatrix}$$

The columns of \mathbf{X} are linearly dependent: the first column is the row-wise sum of the other two columns. The rank (more precise, the column rank) of a matrix is the dimension of space spanned by the column vectors. Hence, the rank of \mathbf{X} is equal to the number of linearly independent columns: $\text{rank}(\mathbf{X}) = 2$. \square

Super-collinearity of an $n \times p$ -dimensional design matrix \mathbf{X} implies* that the rank of the $p \times p$ -dimensional matrix $\mathbf{X}^\top \mathbf{X}$ is smaller than p , and, consequently, it is singular. A square matrix that does not have an inverse is called *singular*. A square matrix \mathbf{A} is singular if and only if its determinant is zero: $\det(\mathbf{A}) = 0$.

Let the $p \times p$ -dimensional matrix \mathbf{A} is not only square, but also symmetric. Then, as $\det(\mathbf{A})$ is equal to the product of the eigenvalues ν_j of \mathbf{A} , the matrix \mathbf{A} is singular if one (or more) of the eigenvalues of \mathbf{A} is zero. To see this, consider the spectral decomposition of \mathbf{A} : $\mathbf{A} = \sum_{j=1}^p \nu_j \mathbf{v}_j \mathbf{v}_j^\top$, where \mathbf{v}_j is the eigenvector corresponding to ν_j . The inverse of \mathbf{A} can then be obtained through the spectral decomposition. It requires to take the reciprocal of the eigenvalues: $\mathbf{A}^{-1} = \sum_{j=1}^p \nu_j^{-1} \mathbf{v}_j \mathbf{v}_j^\top$. The right-hand side is undefined if $\nu_j = 0$ for any j .

*If the (column) rank of \mathbf{X} is smaller than p , there exists a non-trivial $\mathbf{v} \in \mathbb{R}^p$ such that $\mathbf{X}\mathbf{v} = \mathbf{0}_p$. Multiplication of this inequality by \mathbf{X}^\top yields $\mathbf{X}^\top \mathbf{X}\mathbf{v} = \mathbf{0}_p$. As $\mathbf{v} \neq \mathbf{0}_p$, this implies that $\mathbf{X}^\top \mathbf{X}$ is not invertible.

Applied to our high-dimensional setting, the columns of a high-dimensional design matrix \mathbf{X} are linearly dependent and this super-collinearity causes $\mathbf{X}^\top \mathbf{X}$ to be singular. Let us now recall the maximum likelihood estimator of the parameter of the linear regression model: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$. This estimator is only well-defined if $(\mathbf{X}^\top \mathbf{X})^{-1}$ exists. Hence, when \mathbf{X} is high-dimensional the regression parameter $\boldsymbol{\beta}$ cannot be estimated using the maximum likelihood procedure.

So far we only presented the practical consequence of high-dimensionality: the expression $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ cannot be evaluated numerically. But the problem arising from the high-dimensionality of the data is more fundamental. To appreciate this, consider the normal equations: $\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^\top \mathbf{Y}$. The matrix $\mathbf{X}^\top \mathbf{X}$ is of rank n , while $\boldsymbol{\beta}$ is a vector of length p . Hence, while there are p unknowns, the system of linear equations from which these are to be solved effectively comprises n degrees of freedom. If $p > n$, the vector $\boldsymbol{\beta}$ cannot uniquely be determined from this system of equations. To make this more specific let \mathcal{U} be the n -dimensional space spanned by the columns of \mathbf{X} and the $p - n$ -dimensional space \mathcal{V} be orthogonal complement of \mathcal{U} , i.e. $\mathcal{V} = \mathcal{U}^\perp$. Then, $\mathbf{X}\mathbf{v} = \mathbf{0}_p$ for all $\mathbf{v} \in \mathcal{V}$. So, \mathcal{V} is the non-trivial null space of \mathbf{X} . Consequently, as $\mathbf{X}^\top \mathbf{X}\mathbf{v} = \mathbf{X}^\top \mathbf{0}_p = \mathbf{0}_n$, the solution of the normal equations is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \mathbf{Y} + \mathbf{v} \quad \text{for all } \mathbf{v} \in \mathcal{V},$$

where \mathbf{A}^+ denotes the Moore-Penrose inverse of the matrix \mathbf{A} (adopting the notation of Harville, 2008). For a square symmetric matrix, the generalized inverse is defined as:

$$\mathbf{A}^+ = \sum_{j=1}^p \nu_j^{-1} \mathbb{1}_{\{\nu_j \neq 0\}} \mathbf{v}_j \mathbf{v}_j^\top,$$

where \mathbf{v}_j are the eigenvectors of \mathbf{A} (and are not – necessarily – an element of \mathcal{V}). The solution of the normal equations is thus only determined up to an element from a non-trivial space \mathcal{V} , and there is no unique estimator of the regression parameter.

To arrive at a unique regression estimator for studies with rank deficient design matrices, the minimum least squares estimator may be employed.

Definition 1.1 (Ishwaran and Rao, 2014)

The *minimum least squares estimator* of regression parameter minimizes the sum-of-squares criterion and is of minimum length. Formally, $\hat{\boldsymbol{\beta}}_{\text{MLS}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ such that $\|\hat{\boldsymbol{\beta}}_{\text{MLS}}\|_2^2 < \|\boldsymbol{\beta}\|_2^2$ for all $\boldsymbol{\beta}$ that minimize $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$.

If \mathbf{X} is of full rank, the minimum least squares regression estimator coincides with the least squares/maximum likelihood one as the latter is a unique minimizer of the sum-of-squares criterion and, thereby, automatically also the minimizer of minimum length. When \mathbf{X} is rank deficient, $\hat{\boldsymbol{\beta}}_{\text{MLS}} = (\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \mathbf{Y}$. To see this recall from above that $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ is minimized by $(\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \mathbf{Y} + \mathbf{v}$ for all $\mathbf{v} \in \mathcal{V}$. The length of these minimizers is:

$$\|(\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \mathbf{Y} + \mathbf{v}\|_2^2 = \|(\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \mathbf{Y}\|_2^2 + 2\mathbf{Y}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^+ \mathbf{v} + \|\mathbf{v}\|_2^2,$$

which, by the orthogonality of \mathcal{V} and the space spanned by the columns of \mathbf{X} , equals $\|(\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \mathbf{Y}\|_2^2 + \|\mathbf{v}\|_2^2$. Clearly, any nontrivial \mathbf{v} , i.e. $\mathbf{v} \neq \mathbf{0}_p$, results in $\|(\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \mathbf{Y}\|_2^2 + \|\mathbf{v}\|_2^2 > \|(\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \mathbf{Y}\|_2^2$ and, thus, $\hat{\boldsymbol{\beta}}_{\text{MLS}} = (\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \mathbf{Y}$.

An alternative (and related) estimator of the regression parameter $\boldsymbol{\beta}$ that avoids the use of the Moore-Penrose inverse and is able to deal with (super)-collinearity among the columns of the design matrix is the proposed ridge regression estimator by Hoerl and Kennard (1970). It essentially comprises of an ad-hoc fix to resolve the (almost) singularity of $\mathbf{X}^\top \mathbf{X}$. Hoerl and Kennard (1970) propose to simply replace $\mathbf{X}^\top \mathbf{X}$ by $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp}$ with $\lambda \in [0, \infty)$. The scalar λ is a tuning parameter, henceforth called the *penalty parameter* for reasons that will become clear later. The ad-hoc fix solves the singularity as it adds a positive matrix, $\lambda \mathbf{I}_{pp}$, to a positive semi-definite one, $\mathbf{X}^\top \mathbf{X}$, making the total a positive definite matrix (Lemma 14.2.4 of Harville, 2008), which is invertible.

Example 1.3 (Super-collinearity, continued)

Recall the super-collinear design matrix \mathbf{X} of Example 1.3. Then, for (say) $\lambda = 1$:

$$\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp} = \begin{pmatrix} 5 & 2 & 2 \\ 2 & 7 & -4 \\ 2 & -4 & 7 \end{pmatrix}.$$

The eigenvalues of this matrix are 11, 7, and 1. Hence, $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp}$ has no zero eigenvalue and its inverse is well-defined. \square

With the ad-hoc fix for the singularity of $\mathbf{X}^\top \mathbf{X}$ at hand, Hoerl and Kennard (1970) proceed to define the ridge regression estimator.

Definition 1.2

The *ridge regression estimator* of the regression parameter of the linear regression model is:

$$\hat{\beta}(\lambda) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y}, \quad (1.4)$$

for $\lambda \in [0, \infty)$.

For λ strictly positive (Question 1.16 discusses the consequences of a negative values of λ), the ridge regression estimator is a well-defined estimator, even if \mathbf{X} is high-dimensional. However, each choice of λ leads to a different ridge regression estimator. The set of all ridge regression estimates $\{\hat{\beta}(\lambda) : \lambda \in [0, \infty)\}$ is called the *solution path* or *regularization path* of the ridge regression estimator.

Example 1.3 (Super-collinearity, continued)

Recall the super-collinear design matrix \mathbf{X} of Example 1.3. Suppose that the corresponding response vector is $\mathbf{Y} = (1.3, -0.5, 2.6, 0.9)^\top$. The ridge regression estimates for, e.g. $\lambda = 1, 2$, and 10 are then: $\hat{\beta}(1) = (0.614, 0.548, 0.066)^\top$, $\hat{\beta}(2) = (0.537, 0.490, 0.048)^\top$, and $\hat{\beta}(10) = (0.269, 0.267, 0.002)^\top$. The full solution path of the ridge regression estimator is shown in the left-hand side panel of Figure 1.1.

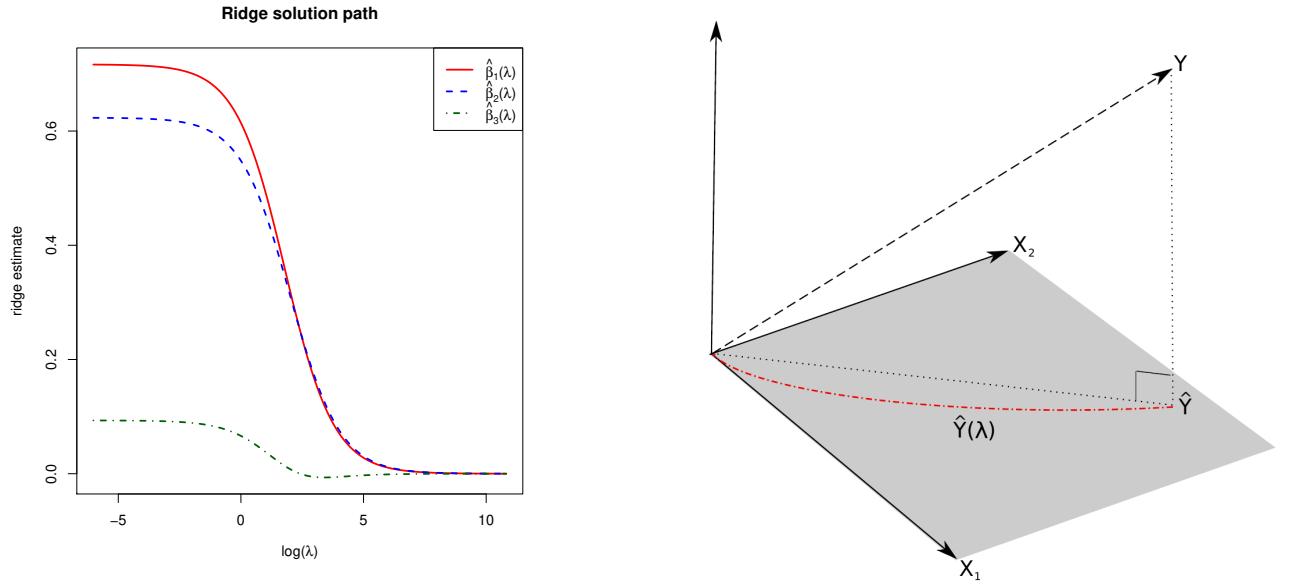


Figure 1.1: Left panel: the regularization path of the ridge regression estimator for the data of Example 1.3. Right panel: the ‘maximum likelihood fit’ \hat{Y} and the ‘ridge fit’ $\hat{Y}(\lambda)$ (the dashed-dotted red line) to the observation Y in the (hyper)plane spanned by the covariates.

Low-dimensionally, when $\mathbf{X}^\top \mathbf{X}$ is of full rank, the ridge regression estimator is linearly related to its maximum likelihood counterpart. To see this define the linear operator $\mathbf{W}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{X}$. The ridge regression estimator $\hat{\beta}(\lambda)$ can then be expressed as $\mathbf{W}_\lambda \hat{\beta}$ for:

$$\begin{aligned} \mathbf{W}_\lambda \hat{\beta} &= \mathbf{W}_\lambda (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \\ &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y} &= \hat{\beta}(\lambda). \end{aligned}$$

The linear operator \mathbf{W}_λ thus transforms the maximum likelihood estimator of the regression parameter into its ridge regularized counterpart. High-dimensionally, no such linear relation between the ridge and the minimum

least squares regression estimators exists (see Exercise 1.7).

With an estimate of the regression parameter β available, we can define the fit. For the ridge regression estimator the fit is defined analogous to the maximum likelihood case:

$$\hat{\mathbf{Y}}(\lambda) = \mathbf{X}\hat{\beta}(\lambda) = \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y} := \mathbf{H}(\lambda) \mathbf{Y}.$$

For the maximum likelihood regression estimator the fit could be understood as a projection of \mathbf{Y} onto the subspace spanned by the columns of \mathbf{X} . This is depicted in the right panel of Figure 1.1, where \hat{Y} is the projection of the observation Y onto the covariate space. The projected observation \hat{Y} is orthogonal to the residual $\varepsilon = Y - \hat{Y}$. This means the fit is the point in the covariate space closest to the observation. Put differently, the covariate space does not contain a point that is better (in some sense) in explaining the observation. Compare this to the ‘ridge fit’ which is plotted as a dashed-dotted red line in the right panel of Figure 1.1. The ‘ridge fit’ is a line, parameterized by $\{\lambda : \lambda \in \mathbb{R}_{\geq 0}\}$, where each point on this line matches to the corresponding intersection of the regularization path $\hat{\beta}(\lambda)$ and the vertical line $x = \lambda$. The ‘ridge fit’ $\hat{Y}(\lambda)$ runs from the maximum likelihood fit $\hat{Y} = \hat{Y}(0)$ to an intercept-only (if present) fit in which the covariates do not contribute to the explanation of the observation. From the figure it is obvious that for any $\lambda > 0$ the ‘ridge fit’ $\hat{Y}(\lambda)$ is not orthogonal to the observation Y . In other words, the ‘ridge residuals’ $Y - \hat{Y}(\lambda)$ are not orthogonal to the fit $\hat{Y}(\lambda)$ (confer Exercise 1.8b). Hence, the ad-hoc fix of the ridge regression estimator resolves the non-evaluation of the estimator in the face of super-collinearity but yields a ‘ridge fit’ that is not optimal in explaining the observation. Mathematically, this is due to the fact that the fit $\hat{Y}(\lambda)$ corresponding to the ridge regression estimator is not a projection of Y onto the covariate space (confer Exercise 1.8a).

1.3 Eigenvalue shrinkage

The effect of the ridge penalty is also studied from the perspective of singular values. Let the singular value decomposition of the $n \times p$ -dimensional design matrix \mathbf{X} be:

$$\mathbf{X} = \mathbf{U}_x \mathbf{D}_x \mathbf{V}_x^\top.$$

In the above \mathbf{D}_x an $n \times p$ -dimensional block matrix. Its upper left block is a $(\text{rank}(\mathbf{X}) \times \text{rank}(\mathbf{X}))$ -dimensional diagonal matrix with the singular values on the diagonal. The remaining blocks, zero if $p = n$ and \mathbf{X} is of full rank, one if $\text{rank}(\mathbf{X}) = n$ or $\text{rank}(\mathbf{X}) = p$, or three if $\text{rank}(\mathbf{X}) < \min\{n, p\}$, are of appropriate dimensions and comprise zeros only. The matrix \mathbf{U}_x an $n \times n$ -dimensional matrix with columns containing the left singular vectors (denoted \mathbf{u}_i), and \mathbf{V}_x a $p \times p$ -dimensional matrix with columns containing the right singular vectors (denoted \mathbf{v}_i). The columns of \mathbf{U}_x and \mathbf{V}_x are orthogonal: $\mathbf{U}_x^\top \mathbf{U}_x = \mathbf{I}_{nn} = \mathbf{U}_x \mathbf{U}_x^\top$ and $\mathbf{V}_x^\top \mathbf{V}_x = \mathbf{I}_{pp} = \mathbf{V}_x \mathbf{V}_x^\top$.

The maximum likelihood estimator, which is well-defined if $n > p$ and $\text{rank}(\mathbf{X}) = p$, can then be rewritten in terms of the SVD-matrices as:

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} &= (\mathbf{V}_x \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{U}_x \mathbf{D}_x \mathbf{V}_x^\top)^{-1} \mathbf{V}_x \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{Y} \\ &= (\mathbf{V}_x \mathbf{D}_x^\top \mathbf{D}_x \mathbf{V}_x^\top)^{-1} \mathbf{V}_x \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{Y} &= \mathbf{V}_x (\mathbf{D}_x^\top \mathbf{D}_x)^{-1} \mathbf{V}_x^\top \mathbf{V}_x \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{Y} \\ &= \mathbf{V}_x (\mathbf{D}_x^\top \mathbf{D}_x)^{-1} \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{Y}. \end{aligned}$$

The block structure of the design matrix implies that matrix $(\mathbf{D}_x^\top \mathbf{D}_x)^{-1} \mathbf{D}_x$ results in a $p \times n$ -dimensional matrix with the reciprocal of the nonzero singular values on the diagonal of the left $p \times p$ -dimensional upper left block. Similarly, the ridge regression estimator can be rewritten in terms of the SVD-matrices as:

$$\begin{aligned} \hat{\beta}(\lambda) &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y} \\ &= (\mathbf{V}_x \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{U}_x \mathbf{D}_x \mathbf{V}_x^\top + \lambda \mathbf{I}_{pp})^{-1} \mathbf{V}_x \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{Y} \\ &= (\mathbf{V}_x \mathbf{D}_x^\top \mathbf{D}_x \mathbf{V}_x^\top + \lambda \mathbf{V}_x \mathbf{V}_x^\top)^{-1} \mathbf{V}_x \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{Y} \\ &= \mathbf{V}_x (\mathbf{D}_x^\top \mathbf{D}_x + \lambda \mathbf{I}_{pp})^{-1} \mathbf{V}_x^\top \mathbf{V}_x \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{Y} \\ &= \mathbf{V}_x (\mathbf{D}_x^\top \mathbf{D}_x + \lambda \mathbf{I}_{pp})^{-1} \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{Y}. \end{aligned} \tag{1.5}$$

Combining the two results and writing $(\mathbf{D}_x)_{jj} = d_{x,jj}$ for the p nonzero singular values on the diagonal of the upper block of \mathbf{D}_x we have: $d_{x,jj}^{-1} \geq d_{x,jj} (d_{x,jj}^2 + \lambda)^{-1}$ for all $\lambda > 0$. Thus, the ridge penalty shrinks the singular values.

Return to the problem of the super-collinearity of \mathbf{X} in the high-dimensional setting ($p > n$). The super-collinearity implies the singularity of $\mathbf{X}^\top \mathbf{X}$ and prevents the calculation of the maximum likelihood estimator of the regression coefficients. However, $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp}$ is non-singular, with inverse: $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} = \sum_{j=1}^p (d_{x,jj}^2 + \lambda)^{-1} \mathbf{v}_j \mathbf{v}_j^\top$ where $d_{x,jj} = 0$ for $j > \text{rank}(\mathbf{X})$. The right-hand side is well-defined for $\lambda > 0$.

From the ‘spectral formulation’ of the ridge regression estimator (1.5) the λ -limits can be deduced. The lower λ -limit of the ridge regression estimator $\hat{\beta}(0_+) = \lim_{\lambda \downarrow 0} \hat{\beta}(\lambda)$ coincides with the minimum least squares estimator. This is immediate when \mathbf{X} is of full rank. In the high-dimensional situation, if the dimension p exceeds the sample size n , it follows from the limit:

$$\lim_{\lambda \downarrow 0} \frac{d_{x,jj}}{d_{x,jj}^2 + \lambda} = \begin{cases} d_{x,jj}^{-1} & \text{if } d_{x,jj} \neq 0 \\ 0 & \text{if } d_{x,jj} = 0 \end{cases}$$

Then, $\lim_{\lambda \downarrow 0} \hat{\beta}(\lambda) = \hat{\beta}_{\text{MLS}}$. Similarly, the upper λ -limit is evident from the fact that $\lim_{\lambda \rightarrow \infty} d_{x,jj} (d_{x,jj}^2 + \lambda)^{-1} = 0$, which implies $\lim_{\lambda \rightarrow \infty} \hat{\beta}(\lambda) = \mathbf{0}_p$. Hence, all regression coefficients are shrunk towards zero as the penalty parameter increases. This also holds for \mathbf{X} with $p > n$. Furthermore, this behaviour is not strictly monotone in λ : $\lambda_a > \lambda_b$ does not necessarily imply $|\hat{\beta}_j(\lambda_a)| < |\hat{\beta}_j(\lambda_b)|$. Upon close inspection this can be witnessed from the ridge solution path of β_3 in Figure 1.1.

1.3.1 Principal component regression

Principal component regression is a close relative to ridge regression that can also be applied in a high-dimensional context. Principal component regression explains the response not by the covariates themselves but by linear combinations of the covariates as defined by the principal components of \mathbf{X} . Let $\mathbf{U}_x \mathbf{D}_x \mathbf{V}_x^\top$ be the singular value decomposition of \mathbf{X} . The k_0 -th principal component of \mathbf{X} is then $\mathbf{X} \mathbf{v}_{k_0}$, henceforth denoted \mathbf{z}_i . Let \mathbf{Z}_k be the matrix of the first k principal components, i.e. $\mathbf{Z}_k = \mathbf{X} \mathbf{V}_{x,k}$ where $\mathbf{V}_{x,k}$ contains the first k right singular vectors as columns. Principal component regression then amounts to regressing the response \mathbf{Y} onto \mathbf{Z}_k , that is, it fits the model $\mathbf{Y} = \mathbf{Z}_k \gamma + \varepsilon$. The least squares estimator of γ then is (with some abuse of notation):

$$\begin{aligned} \hat{\gamma} &= (\mathbf{Z}_k^\top \mathbf{Z}_k)^{-1} \mathbf{Z}_k^\top \mathbf{Y} \\ &= (\mathbf{V}_{x,k}^\top \mathbf{X}^\top \mathbf{X} \mathbf{V}_{x,k})^{-1} \mathbf{V}_{x,k}^\top \mathbf{X}^\top \mathbf{Y} \\ &= (\mathbf{V}_{x,k}^\top \mathbf{V}_{x,k} \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{U}_x \mathbf{D}_x \mathbf{V}_{x,k}^\top \mathbf{V}_{x,k})^{-1} \mathbf{V}_{x,k}^\top \mathbf{V}_{x,k} \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{Y} \\ &= (\mathbf{I}_{kp} \mathbf{D}_x^\top \mathbf{D}_x \mathbf{I}_{pk})^{-1} \mathbf{I}_{kp} \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{Y} \\ &= (\mathbf{D}_{x,k}^\top \mathbf{D}_{x,k})^{-1} \mathbf{D}_{x,k}^\top \mathbf{U}_x^\top \mathbf{Y} \\ &= (\mathbf{D}_{x,k}^\top \mathbf{D}_{x,k})^{-1} \mathbf{D}_{x,k}^\top \mathbf{U}_x^\top \mathbf{Y}, \end{aligned}$$

where $\mathbf{D}_{x,k}$ is a submatrix of \mathbf{D}_x formed from \mathbf{D}_x by removal of the last $p - k$ columns. Similarly, \mathbf{I}_{kp} and \mathbf{I}_{pk} are obtained from \mathbf{I}_{pp} by removal of the last $p - k$ rows and columns, respectively. The principal component regression estimator of β then is $\hat{\beta}_{\text{per}} = \mathbf{V}_{x,k} (\mathbf{D}_{x,k}^\top \mathbf{D}_{x,k})^{-1} \mathbf{D}_{x,k}^\top \mathbf{U}_x^\top \mathbf{Y}$. When k is set equal to the column rank of \mathbf{X} , and thus to the rank of $\mathbf{X}^\top \mathbf{X}$, the principal component regression estimator $\hat{\beta}_{\text{per}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$, where \mathbf{A}^- denotes the Moore-Penrose inverse of matrix \mathbf{A} .

The relation between ridge and principal component regression becomes clear when their corresponding estimators are written in terms of the singular value decomposition of \mathbf{X} :

$$\begin{aligned} \hat{\beta}_{\text{per}} &= \mathbf{V}_{x,k} (\mathbf{I}_{kp} \mathbf{D}_x^\top \mathbf{D}_x \mathbf{I}_{pk})^{-1} \mathbf{I}_{kp} \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{Y}, \\ \hat{\beta}(\lambda) &= \mathbf{V}_x (\mathbf{D}_x^\top \mathbf{D}_x + \lambda \mathbf{I}_{pp})^{-1} \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{Y}. \end{aligned}$$

Both operate on the singular values of the design matrix. But where principal component regression thresholds the singular values of \mathbf{X} , ridge regression shrinks them (depending on their size). Hence, one applies a discrete map on the singular values while the other a continuous one.

1.4 Moments

The first two moments of the ridge regression estimator are derived. Next the performance of the ridge regression estimator is studied in terms of the mean squared error, which combines the first two moments.

1.4.1 Expectation

The left panel of Figure 1.1 shows ridge estimates of the regression parameters converging to zero as the penalty parameter tends to infinity. This behaviour of the ridge regression estimator does not depend on the specifics of the data set. To see this study the expectation of the ridge regression estimator:

$$\begin{aligned}\mathbb{E}[\hat{\beta}(\lambda)] &= \mathbb{E}[(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y}] = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbb{E}(\mathbf{Y}) \\ &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{X} \beta = \beta - \lambda (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \beta.\end{aligned}\quad (1.6)$$

Clearly, $\mathbb{E}[\hat{\beta}(\lambda)] \neq \beta$ for any $\lambda > 0$. Hence, the ridge regression estimator is biased.

Example 1.4 (Orthonormal design matrix)

Consider an orthonormal design matrix \mathbf{X} , i.e.: $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_{pp} = (\mathbf{X}^\top \mathbf{X})^{-1}$. The relation between the maximum likelihood and ridge regression estimator then is:

$$\begin{aligned}\hat{\beta}(\lambda) &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y} = (\mathbf{I}_{pp} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y} \\ &= (1 + \lambda)^{-1} \mathbf{I}_{pp} \mathbf{X}^\top \mathbf{Y} = (1 + \lambda)^{-1} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \\ &= (1 + \lambda)^{-1} \hat{\beta}.\end{aligned}$$

Hence, the ridge regression estimator scales the maximum likelihood estimator by a factor. When taking the expectation on both sides, it is evident that the ridge regression estimator is biased: $\mathbb{E}[\hat{\beta}(\lambda)] = \mathbb{E}[(1 + \lambda)^{-1} \hat{\beta}] = (1 + \lambda)^{-1} \mathbb{E}(\hat{\beta}) = (1 + \lambda)^{-1} \beta \neq \beta$. From this it also clear that the estimator, and thus its expectation, vanishes as $\lambda \rightarrow \infty$. \square

The bias of the ridge regression estimator may be decomposed into two parts, one attributable to the penalization and another to the high-dimensionality of the study design.

Proposition 1.1 (after Shao and Deng, 2012)

The bias of the ridge regression estimator can be decomposed as $\mathbb{E}[\hat{\beta}(\lambda) - \beta] = \mathbf{P}_x \mathbb{E}[\hat{\beta}(\lambda) - \beta] + (\mathbf{P}_x - \mathbf{I}_{pp})\beta$.

Proof. To arrive at the bias decomposition, we assume $p > n$ and define the projection matrix, i.e. a matrix \mathbf{P} such that $\mathbf{P} = \mathbf{P}^2$, that projects the parameter space \mathbb{R}^p onto the subspace $\mathcal{R}(\mathbf{X}) \subset \mathbb{R}^p$ spanned by the rows of the design matrix \mathbf{X} , denoted \mathbf{P}_x . It is given by: $\mathbf{P}_x = \mathbf{X}^\top (\mathbf{X} \mathbf{X}^\top)^+ \mathbf{X}$, where the rows of \mathbf{X} are linearly independent and $(\mathbf{X} \mathbf{X}^\top)^+$ is the Moore-Penrose inverse of $\mathbf{X} \mathbf{X}^\top$. The ridge regression estimator lives in the subspace defined by the projection \mathbf{P}_x of \mathbb{R}^p onto $\mathcal{R}(\mathbf{X})$. To verify this, consider the singular value decomposition $\mathbf{X} = \mathbf{U}_x \mathbf{D}_x \mathbf{V}_x^\top$ (with matrices defined as before) and note that:

$$\begin{aligned}\mathbf{P}_x &= \mathbf{X}^\top (\mathbf{X} \mathbf{X}^\top)^+ \mathbf{X} = \mathbf{V}_x \mathbf{D}_x^\top \mathbf{U}_x^\top (\mathbf{U}_x \mathbf{D}_x \mathbf{V}_x^\top \mathbf{V}_x \mathbf{D}_x^\top \mathbf{U}_x^\top)^+ \mathbf{U}_x \mathbf{D}_x \mathbf{V}_x^\top \\ &= \mathbf{V}_x \mathbf{D}_x^\top (\mathbf{D}_x \mathbf{D}_x^\top)^+ \mathbf{D}_x \mathbf{V}_x^\top = \mathbf{V}_x \mathbf{I}_{pn} \mathbf{I}_{np} \mathbf{V}_x^\top.\end{aligned}$$

The identity above does not hold if the rows \mathbf{X} harbor linear dependency. For instance, if the study contains a replicate corresponding to a duplicated row. This does not hamper the envisioned bias decomposition as the definition of the projection matrix can be modified, e.g. without one of instances of the duplicated row, such that the identity in the display above holds and it still projects onto $\mathcal{R}(\mathbf{X})$. With the projection matrix at hand, we note that

$$\begin{aligned}\mathbf{P}_x \hat{\beta}(\lambda) &= \mathbf{V}_x \mathbf{I}_{pn} \mathbf{I}_{np} \mathbf{V}_x^\top \mathbf{V}_x (\mathbf{D}_x^\top \mathbf{D}_x + \lambda \mathbf{I}_{pp})^{-1} \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{Y} \\ &= \mathbf{V}_x (\mathbf{D}_x^\top \mathbf{D}_x + \lambda \mathbf{I}_{pp})^{-1} \mathbf{I}_{pn} \mathbf{I}_{np} \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{Y} = \hat{\beta}(\lambda).\end{aligned}$$

The ridge regression estimator is thus unaffected by the projection, as $\mathbf{P}_x \hat{\beta}(\lambda) = \hat{\beta}(\lambda)$, and it must therefore already be an element of the projected subspace $\mathcal{R}(\mathbf{X})$. The bias can now be decomposed as:

$$\mathbb{E}[\hat{\beta}(\lambda) - \beta] = \mathbb{E}[\hat{\beta}(\lambda) - \mathbf{P}_x \beta + \mathbf{P}_x \beta - \beta] = \mathbf{P}_x \mathbb{E}[\hat{\beta}(\lambda) - \beta] + (\mathbf{P}_x - \mathbf{I}_{pp})\beta,$$

which concludes the proof. \blacksquare

The first summand of Proposition 1.1's bias decomposition represents the bias of the ridge regression estimator to the projection of the true parameter value, whereas the second summand is the bias introduced by the high-dimensionality of the study design. Either if *i*) \mathbf{X} is of full row rank (i.e. the study design in low-dimensional and $\mathbf{P}_x = \mathbf{I}_{pp}$) or if *ii*) the true regression parameter β is an element the projected subspace (i.e. $\beta = \mathbf{P}_x \beta \in \mathcal{R}(\mathbf{X})$), the second summand of the bias will vanish.

1.4.2 Variance

The second moment of the ridge regression estimator is straightforwardly obtained when exploiting its linearly relation with the maximum likelihood regression estimator. Then,

$$\begin{aligned}\text{Var}[\hat{\beta}(\lambda)] &= \text{Var}(\mathbf{W}_\lambda \hat{\beta}) = \mathbf{W}_\lambda \text{Var}(\hat{\beta}) \mathbf{W}_\lambda^\top \\ &= \sigma^2 \mathbf{W}_\lambda (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{W}_\lambda^\top = \sigma^2 (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{X} [(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}]^\top,\end{aligned}$$

in which we have used $\text{Var}(\mathbf{A}\mathbf{Y}) = \mathbf{A}\text{Var}(\mathbf{Y})\mathbf{A}^\top$ for a non-random matrix \mathbf{A} , the fact that \mathbf{W}_λ is non-random, and $\text{Var}[\hat{\beta}] = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

We characterize the behavior of the variance of the ridge regression estimator. We first observe that, similar to the expectation, it vanishes as λ tends to infinity:

$$\lim_{\lambda \rightarrow \infty} \text{Var}[\hat{\beta}(\lambda)] = \lim_{\lambda \rightarrow \infty} \sigma^2 \mathbf{W}_\lambda (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{W}_\lambda^\top = \mathbf{0}_{pp}.$$

Hence, the variance of the ridge regression coefficient estimates decreases towards zero as the penalty parameter becomes large. This is illustrated in the right panel of Figure 1.1 for the data of Example 1.3. Secondly, the variance of the maximum likelihood estimator is ‘larger’ than that of the ridge regression estimator (Proposition 1.2).

Proposition 1.2

The variance of the maximum likelihood regression estimator exceeds (in the positive definite ordering sense) that of the ridge regression estimator, $\text{Var}(\hat{\beta}) \succeq \text{Var}[\hat{\beta}(\lambda)]$, with the inequality being strict if $\lambda > 0$.

Proof. Use the analytic expression of the variance of the ridge regression estimator to study its difference to that of the maximum likelihood regression estimator:

$$\begin{aligned}\text{Var}[\hat{\beta}] - \text{Var}[\hat{\beta}(\lambda)] &= \sigma^2 [(\mathbf{X}^\top \mathbf{X})^{-1} - \mathbf{W}_\lambda (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{W}_\lambda^\top] \\ &= \sigma^2 \mathbf{W}_\lambda \{ [\mathbf{I} + \lambda (\mathbf{X}^\top \mathbf{X})^{-1}] (\mathbf{X}^\top \mathbf{X})^{-1} [\mathbf{I} + \lambda (\mathbf{X}^\top \mathbf{X})^{-1}]^\top - (\mathbf{X}^\top \mathbf{X})^{-1} \} \mathbf{W}_\lambda^\top \\ &= \sigma^2 \mathbf{W}_\lambda [2\lambda (\mathbf{X}^\top \mathbf{X})^{-2} + \lambda^2 (\mathbf{X}^\top \mathbf{X})^{-3}] \mathbf{W}_\lambda^\top \\ &= \sigma^2 (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} [2\lambda \mathbf{I}_{pp} + \lambda^2 (\mathbf{X}^\top \mathbf{X})^{-1}] [(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}]^\top.\end{aligned}$$

This difference is non-negative definite as each component in the matrix product is non-negative definite, and even positive definite if either $\lambda > 0$ or $\text{rank}(\mathbf{X}^\top \mathbf{X}) = p$. ■

The variance inequality of Proposition 1.2 can be interpreted in terms of the stochastic behaviour of both involved estimates. This is illustrated by the next example.

Example 1.5 (Variance comparison)

Consider the design matrix:

$$\mathbf{X}^\top = \begin{pmatrix} -1 & 0 & 2 & 1 \\ 2 & 1 & -1 & 0 \end{pmatrix}.$$

he variances of the maximum likelihood and ridge (with $\lambda = 1$) estimates of the regression coefficients then are:

$$\text{Var}(\hat{\beta}) = \sigma^2 \begin{pmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{pmatrix} \quad \text{and} \quad \text{Var}[\hat{\beta}(\lambda)] = \sigma^2 \begin{pmatrix} 0.1524 & 0.0698 \\ 0.0698 & 0.1524 \end{pmatrix}.$$

These variances can be used to construct levels sets of the distribution of the estimates. The level sets that contain 50%, 75% and 95% of the distribution of the maximum likelihood and ridge regression estimates are plotted in Figure 1.2. In line with inequality of Proposition 1.2 the level sets of the ridge regression estimate are smaller than that of the maximum likelihood one. The former thus varies less. □

Example 1.4 (Orthonormal design matrix, continued)

Assume the design matrix \mathbf{X} is orthonormal. Then, $\text{Var}(\hat{\beta}) = \sigma^2 \mathbf{I}_{pp}$ and

$$\text{Var}[\hat{\beta}(\lambda)] = \sigma^2 \mathbf{W}_\lambda (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{W}_\lambda^\top = \sigma^2 (\mathbf{I}_{pp} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{I}_{pp} [(\mathbf{I}_{pp} + \lambda \mathbf{I}_{pp})^{-1}]^\top = \sigma^2 (1 + \lambda)^{-2} \mathbf{I}_{pp}.$$

As the penalty parameter λ is non-negative the former exceeds the latter. In particular, the expression after the utmost right equality sign vanishes as $\lambda \rightarrow \infty$. □

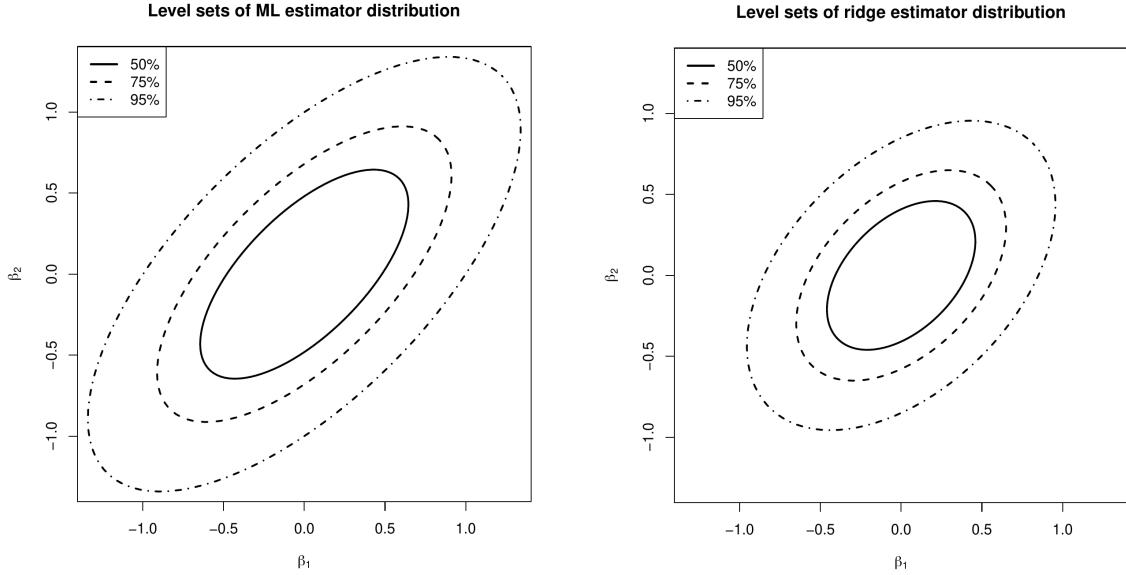


Figure 1.2: Level sets of the distribution of the maximum likelihood (left panel) and ridge (right panel) regression estimators.

The variance of the ridge regression estimator may be decomposed in the same way as its bias (cf. the end of Section 1.4.1). There is, however, no contribution of the high-dimensionality of the study design as that is non-random and, consequently, exhibits no variation. Hence, the variance only relates to the variation in the projected subspace $\mathcal{R}(\mathbf{X})$ as is obvious from:

$$\text{Var}[\hat{\boldsymbol{\beta}}(\lambda)] = \text{Var}[\mathbf{P}_x \hat{\boldsymbol{\beta}}(\lambda)] = \mathbf{P}_x \text{Var}[\hat{\boldsymbol{\beta}}(\lambda)] \mathbf{P}_x^\top = \text{Var}[\hat{\boldsymbol{\beta}}(\lambda)].$$

Perhaps this is seen more clearly when writing the variance of the ridge regression estimator in terms of the matrices that constitute the singular value decomposition of \mathbf{X} :

$$\text{Var}[\hat{\boldsymbol{\beta}}(\lambda)] = \mathbf{V}_x (\mathbf{D}_x^\top \mathbf{D}_x + \lambda \mathbf{I}_{pp})^{-1} \mathbf{D}_x^\top \mathbf{D}_x (\mathbf{D}_x^\top \mathbf{D}_x + \lambda \mathbf{I}_{pp})^{-1} \mathbf{V}_x^\top.$$

High-dimensionally, $(\mathbf{D}_x^\top \mathbf{D}_x)_{jj} = 0$ for $j = n+1, \dots, p$. And if $(\mathbf{D}_x^\top \mathbf{D}_x)_{jj} = 0$, so is $[(\mathbf{D}_x^\top \mathbf{D}_x + \lambda \mathbf{I}_{pp})^{-1} \mathbf{D}_x^\top \mathbf{D}_x (\mathbf{D}_x^\top \mathbf{D}_x + \lambda \mathbf{I}_{pp})^{-1}]_{jj} = 0$. Hence, the variance is determined by the first n columns of \mathbf{V}_x . When $n < p$, the variance is then to interpreted as the spread of the ridge regression estimator (with the same choice of λ) when the study is repeated with exactly the same design matrix such that the resulting estimator is confined to the same subspace $\mathcal{R}(\mathbf{X})$. The following R-script illustrates this by an arbitrary data example (plot not shown):

Listing 1.2 R code

```
# set parameters
X      <- matrix(rnorm(2), nrow=1)
betas <- matrix(c(2, -1), ncol=1)
lambda <- 1

# generate multiple ridge regression estimators with a fixed design matrix
bHats <- numeric()
for (k in 1:1000){
  Y      <- matrix(X %*% betas + rnorm(1), ncol=1)
  bHats <- rbind(bHats, t(solve(t(X) %*% X + lambda * diag(2)) %*% t(X) %*% Y))
}

# plot the ridge regression estimators
plot(bHats, xlab=expression(paste(hat(beta)[1], "(", lambda, ")")), 
     ylab=expression(paste(hat(beta)[2], "(", lambda, ")")), pch=20)
```

The full distribution of the ridge regression estimator is now known. The estimator, $\hat{\boldsymbol{\beta}}(\lambda) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y}$

is a linear estimator, linear in \mathbf{Y} . As \mathbf{Y} is normally distributed, so is $\hat{\beta}(\lambda)$. Moreover, the normal distribution is fully characterized by its first two moments, which are available. Hence:

$$\hat{\beta}(\lambda) \sim \mathcal{N}\{(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{X} \beta, \sigma^2 (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{X} [(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}]^\top\}.$$

Given λ and \mathbf{X} , the random behavior of the estimator is thus known. In particular, when $n < p$, the variance is semi-positive definite and this p -variate normal distribution is degenerate, i.e. there is no probability mass outside $\mathcal{R}(\mathbf{X})$ the subspace of \mathbb{R}^p spanned by the rows of the \mathbf{X} .

1.4.3 Mean squared error

Previously, we motivated the ridge regression estimator as an ad hoc solution to collinearity. An alternative motivation comes from studying the Mean Squared Error (MSE) of the ridge regression estimator: for a suitable choice of λ the ridge regression estimator may outperform the ML regression estimator in terms of the MSE. Before we prove this, we first derive the MSE of the ridge regression estimator and quote some auxiliary results. Note that, as the ridge regression estimator is compared to its ML counterpart, throughout this subsection $n > p$ is assumed to warrant the uniqueness of the latter.

Recall that (in general) for any estimator of a parameter θ :

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2.$$

This measure of the quality of the estimator thus has a convenient decomposition. The lemma below provides the MSE of the ridge regression estimator.

Lemma 1.1

The mean squared error of the ridge regression estimator is:

$$\text{MSE}[\hat{\beta}(\lambda)] = \sigma^2 \text{tr}[\mathbf{W}_\lambda (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{W}_\lambda^\top] + \beta^\top (\mathbf{W}_\lambda - \mathbf{I}_{pp})^\top (\mathbf{W}_\lambda - \mathbf{I}_{pp}) \beta, \quad (1.7)$$

where \mathbf{W}_λ is as defined in Section 1.2.

Proof. Straightforward linear algebra and the expectation calculus yields:

$$\begin{aligned} \text{MSE}[\hat{\beta}(\lambda)] &= \mathbb{E}[(\mathbf{W}_\lambda \hat{\beta} - \beta)^\top (\mathbf{W}_\lambda \hat{\beta} - \beta)] \\ &= \mathbb{E}(\hat{\beta}^\top \mathbf{W}_\lambda^\top \mathbf{W}_\lambda \hat{\beta}) - \mathbb{E}(\beta^\top \mathbf{W}_\lambda \hat{\beta}) - \mathbb{E}(\hat{\beta}^\top \mathbf{W}_\lambda^\top \beta) + \mathbb{E}(\beta^\top \beta) \\ &= \mathbb{E}(\hat{\beta}^\top \mathbf{W}_\lambda^\top \mathbf{W}_\lambda \hat{\beta}) - \mathbb{E}(\beta^\top \mathbf{W}_\lambda^\top \mathbf{W}_\lambda \hat{\beta}) - \mathbb{E}(\hat{\beta}^\top \mathbf{W}_\lambda^\top \mathbf{W}_\lambda \beta) + \mathbb{E}(\beta^\top \mathbf{W}_\lambda^\top \mathbf{W}_\lambda \beta) \\ &\quad - \mathbb{E}(\beta^\top \mathbf{W}_\lambda^\top \mathbf{W}_\lambda \beta) + \mathbb{E}(\beta^\top \mathbf{W}_\lambda^\top \mathbf{W}_\lambda \hat{\beta}) + \mathbb{E}(\hat{\beta}^\top \mathbf{W}_\lambda^\top \mathbf{W}_\lambda \beta) \\ &\quad - \mathbb{E}(\beta^\top \mathbf{W}_\lambda \hat{\beta}) - \mathbb{E}(\hat{\beta}^\top \mathbf{W}_\lambda^\top \beta) + \mathbb{E}(\beta^\top \beta) \\ &= \mathbb{E}[(\hat{\beta} - \beta)^\top \mathbf{W}_\lambda^\top \mathbf{W}_\lambda (\hat{\beta} - \beta)] \\ &\quad - \beta^\top \mathbf{W}_\lambda^\top \mathbf{W}_\lambda \beta + \beta^\top \mathbf{W}_\lambda^\top \mathbf{W}_\lambda \beta + \beta^\top \mathbf{W}_\lambda^\top \mathbf{W}_\lambda \beta \\ &\quad - \beta^\top \mathbf{W}_\lambda \beta - \beta^\top \mathbf{W}_\lambda^\top \beta + \beta^\top \beta \\ &= \mathbb{E}[(\hat{\beta} - \beta)^\top \mathbf{W}_\lambda^\top \mathbf{W}_\lambda (\hat{\beta} - \beta)] + \beta^\top (\mathbf{W}_\lambda - \mathbf{I}_{pp})^\top (\mathbf{W}_\lambda - \mathbf{I}_{pp}) \beta \\ &= \sigma^2 \text{tr}[\mathbf{W}_\lambda (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{W}_\lambda^\top] + \beta^\top (\mathbf{W}_\lambda - \mathbf{I}_{pp})^\top (\mathbf{W}_\lambda - \mathbf{I}_{pp}) \beta. \end{aligned}$$

In the last step we have used $\hat{\beta} \sim \mathcal{N}[\beta, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}]$ and the expectation of the quadratic form of a multivariate random variable $\varepsilon \sim \mathcal{N}(\mu_\varepsilon, \Sigma_\varepsilon)$ that for a nonrandom symmetric positive definite matrix Λ is (cf. Mathai and Provost 1992) $\mathbb{E}[\varepsilon^\top \Lambda \varepsilon] = \text{tr}(\Lambda \Sigma_\varepsilon) + \mu_\varepsilon^\top \Lambda \mu_\varepsilon$, of course replacing ε by $\hat{\beta}$ in this expectation. ■

The first summand in the Lemma 1.1's expression of $\text{MSE}[\hat{\beta}(\lambda)]$ represents the sum of the variances of the ridge regression estimator, while the second summand can be thought of the “squared bias” of the ridge regression estimator. In particular, $\lim_{\lambda \rightarrow \infty} \text{MSE}[\hat{\beta}(\lambda)] = \beta^\top \beta$, which is the squared biased for an estimator that equals zero (as does the ridge regression estimator in the limit).

Example 1.6 (Orthonormal design matrix, continued)

Assume the design matrix \mathbf{X} is orthonormal. Then, $\text{MSE}[\hat{\beta}] = p \sigma^2$ and

$$\text{MSE}[\hat{\beta}(\lambda)] = p \sigma^2 (1 + \lambda)^{-2} + \lambda^2 (1 + \lambda)^{-2} \beta^\top \beta.$$

The latter achieves its minimum at: $\lambda = p \sigma^2 / \beta^\top \beta$. □

The following theorem and proposition are required for the proof of the main result.

Theorem 1.1 (*Theorem 1 of Theobald, 1974*)

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be (different) estimators of θ with second order moments:

$$\mathbf{M}_k = \mathbb{E}[(\hat{\theta}_k - \theta)(\hat{\theta}_k - \theta)^\top] \quad \text{for } k = 1, 2,$$

and

$$\text{MSE}(\hat{\theta}_k) = \mathbb{E}[(\hat{\theta}_k - \theta)^\top \mathbf{A}(\hat{\theta}_k - \theta)] \quad \text{for } k = 1, 2,$$

where $\mathbf{A} \succeq 0$. Then, $\mathbf{M}_1 - \mathbf{M}_2 \succeq 0$ if and only if $\text{MSE}(\hat{\theta}_1) - \text{MSE}(\hat{\theta}_2) \geq 0$ for all $\mathbf{A} \succeq 0$.

Proposition 1.3 (*Farebrother, 1976*)

Let \mathbf{A} be a $p \times p$ -dimensional, positive definite matrix, \mathbf{b} be a nonzero p dimensional vector, and $c \in \mathbb{R}_+$. Then, $c\mathbf{A} - \mathbf{b}\mathbf{b}^\top \succ 0$ if and only if $\mathbf{b}^\top \mathbf{A}^{-1} \mathbf{b} < c$.

We are now ready to proof the main result, formalized as Theorem 1.2, that for some λ the ridge regression estimator yields a lower MSE than the ML regression estimator. Question 1.12 provides a simpler (?) but more limited proof of this result.

Theorem 1.2 (*Theorem 2 of Theobald, 1974*)

There exists $\lambda > 0$ such that $\text{MSE}[\hat{\beta}(\lambda)] < \text{MSE}[\hat{\beta}(0)] = \text{MSE}(\hat{\beta})$.

Proof. The second order moment matrix of the ridge regression estimator is:

$$\begin{aligned} \mathbf{M}(\lambda) &:= \mathbb{E}[(\hat{\beta}(\lambda) - \beta)(\hat{\beta}(\lambda) - \beta)^\top] \\ &= \mathbb{E}\{\hat{\beta}(\lambda)[\hat{\beta}(\lambda)^\top]\} - \mathbb{E}[\hat{\beta}(\lambda)]\{\mathbb{E}[\hat{\beta}(\lambda)]\}^\top + \mathbb{E}[\hat{\beta}(\lambda) - \beta]\{\mathbb{E}[\hat{\beta}(\lambda) - \beta]\}^\top \\ &= \text{Var}[\hat{\beta}(\lambda)] + \mathbb{E}[\hat{\beta}(\lambda) - \beta]\{\mathbb{E}[\hat{\beta}(\lambda) - \beta]\}^\top. \end{aligned}$$

Then:

$$\begin{aligned} \mathbf{M}(0) - \mathbf{M}(\lambda) &= \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1} - \sigma^2 \mathbf{W}_\lambda (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{W}_\lambda^\top - (\mathbf{W}_\lambda - \mathbf{I}_{pp}) \beta \beta^\top (\mathbf{W}_\lambda - \mathbf{I}_{pp})^\top \\ &= \sigma^2 \mathbf{W}_\lambda (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp}) (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp}) (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{W}_\lambda^\top \\ &\quad - \sigma^2 \mathbf{W}_\lambda (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{W}_\lambda^\top - (\mathbf{W}_\lambda - \mathbf{I}_{pp}) \beta \beta^\top (\mathbf{W}_\lambda - \mathbf{I}_{pp})^\top \\ &= \sigma^2 \mathbf{W}_\lambda [2\lambda(\mathbf{X}^\top \mathbf{X})^{-2} + \lambda^2(\mathbf{X}^\top \mathbf{X})^{-3}] \mathbf{W}_\lambda^\top \\ &\quad - \lambda^2(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \beta \beta^\top [(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}]^\top \\ &= \sigma^2 (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} [2\lambda \mathbf{I}_{pp} + \lambda^2 (\mathbf{X}^\top \mathbf{X})^{-1}] [(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}]^\top \\ &\quad - \lambda^2 (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \beta \beta^\top [(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}]^\top \\ &= \lambda (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} [2\sigma^2 \mathbf{I}_{pp} + \lambda \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} - \lambda \beta \beta^\top] [(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}]^\top. \end{aligned}$$

This is positive definite if and only if $2\sigma^2 \mathbf{I}_{pp} + \lambda \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} - \lambda \beta \beta^\top \succ 0$. Hereto it suffices to show that $2\sigma^2 \mathbf{I}_{pp} - \lambda \beta \beta^\top \succ 0$. By Proposition 1.3 this holds for λ such that $2\sigma^2(\beta^\top \beta)^{-1} > \lambda$. For these λ , we thus have $\mathbf{M}(0) - \mathbf{M}(\lambda)$. Application of Theorem 1.1 now concludes the proof. ■

This result of Theobald (1974) is generalized by Farebrother (1976) to the class of design matrices \mathbf{X} with $\text{rank}(\mathbf{X}) < p$.

Theorem 1.2 can be used to illustrate that the ridge regression estimator strikes a balance between the bias and variance. This is illustrated in the left panel of Figure 1.3. For small λ , the variance of the ridge regression estimator dominates the MSE. This may be understood when realizing that in this domain of λ the ridge regression estimator is close to the unbiased maximum likelihood regression estimator. For large λ , the variance vanishes and the bias dominates the MSE. For small enough values of λ , the decrease in variance of the ridge regression estimator exceeds the increase in its bias. As the MSE is the sum of these two, the MSE first decreases as λ moves away from zero. In particular, as $\lambda = 0$ corresponds to the ML regression estimator, the ridge regression estimator yields a lower MSE for these values of λ . In the right panel of Figure 1.3 $\text{MSE}[\hat{\beta}(\lambda)] < \text{MSE}[\hat{\beta}(0)]$ for $\lambda < 7$ (roughly) and the ridge regression estimator outperforms its maximum likelihood counterpart.

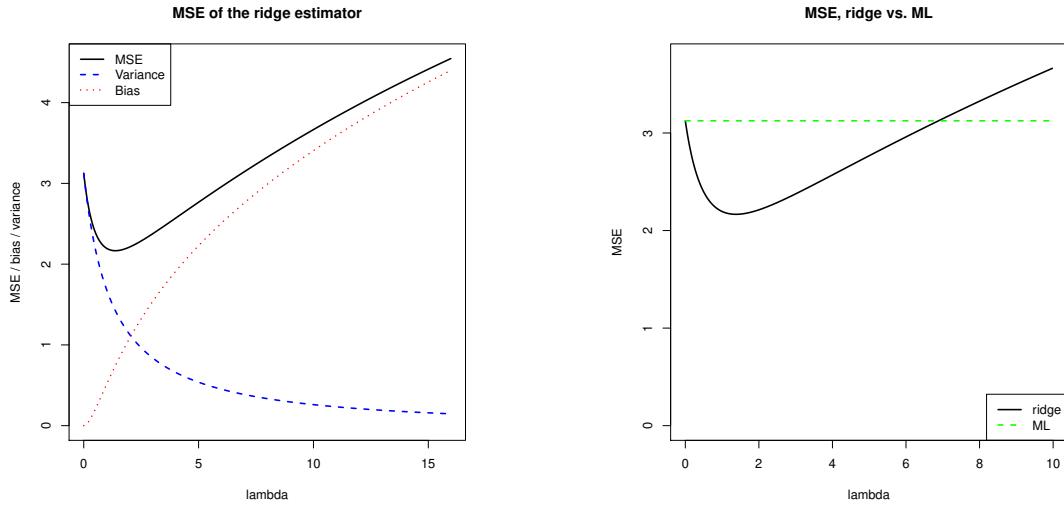


Figure 1.3: Left panel: mean squared error, and its ‘bias’ and ‘variance’ parts, of the ridge regression estimator (for artificial data). Right panel: mean squared error of the ridge and ML estimator of the regression coefficient vector (for the same artificial data).

Besides another motivation behind the ridge regression estimator, the use of Theorem 1.2 is limited. The optimal choice of λ depends on the quantities β and σ^2 . These are unknown in practice. Then, the penalty parameter is chosen in a data-driven fashion (see e.g. Section 1.8.2 and various other places).

Theorem 1.2 may be of limited practical use, it does give insight in when the ridge regression estimator may be preferred over its ML counterpart. Ideally, the range of penalty parameters for which the ridge regression estimator outperforms – in the MSE sense – the ML regression estimator is as large as possible. The factors that influence the size of this range may be deduced from the optimal penalty $\lambda_{\text{opt}} = \sigma^2(\beta^\top \beta/p)^{-1}$ found under the assumption of an orthonormal \mathbf{X} (see Example 1.6). But also from the bound on the penalty parameter, $\lambda_{\max} = 2\sigma^2(\beta^\top \beta)^{-1}$ such that $\text{MSE}[\hat{\beta}(\lambda)] < \text{MSE}[\hat{\beta}(0)]$ for all $\lambda \in (0, \lambda_{\max})$, derived in the proof of Theorem 1.2. Firstly, an increase of the error variance σ^2 yields a larger λ_{opt} and λ_{\max} . Put differently, more noisy data benefits the ridge regression estimator. Secondly, λ_{opt} and λ_{\max} also become larger when their denominators decreases. The denominator $\beta^\top \beta/p$ may be viewed as an estimator of the ‘signal’ variance ‘ σ_β^2 ’. A quick conclusion would be that ridge regression profits from less signal. But more can be learned from the denominator. Contrast the two regression parameters $\beta_{\text{unif}} = \mathbf{1}_p$ and β_{sparse} which comprises of only zeros except the first element which equals p , i.e. $\beta_{\text{sparse}} = (p, 0, \dots, 0)^\top$. Then, the β_{unif} and β_{sparse} have comparable signal in the sense that $\sum_{j=1}^p \beta_j = p$. The denominator of λ_{opt} corresponding both parameters equals 1 and p , respectively. This suggests that ridge regression will perform better in the former case where the regression parameter is not dominated by a few elements but rather all contribute comparably to the explanation of the variation in the response. Of course, more factors contribute. For instance, collinearity among the columns of \mathbf{X} , which gave rise to ridge regression in the first place.

The choice of the penalty parameter on the basis of the mean squared error strikes a balance between the bias and variance, a so-called *bias-variance trade-off*. In the extremes, the model has either a high variance but low bias and will overfit the data. Or, the model exhibits little variance but has a high bias and as a result underfits the data. Ideally, the balance between bias and variance results in a model that neither over- nor underfits. Apart from the regularization parameter, the bias-variance trade-off is affected by other means. For instance, the addition of more covariates (or transformations thereof) to the model is likely to result in a lower bias but also in a high variance, while the employment of a simpler model has the opposite effect. Alternative, a larger sample size typically reduces the variance.

Remark 1.1

Theorem 1.2 can also be used to conclude on the biasedness of the ridge regression estimator. The Gauss-Markov theorem (Rao, 1973) states (under some assumptions) that the ML regression estimator is the best linear unbiased

estimator (BLUE) with the smallest MSE. As the ridge regression estimator is a linear estimator and outperforms (in terms of MSE) this ML estimator, it must be biased (for it would otherwise refute the Gauss-Markov theorem).

1.4.4 Debiasing

Despite a potential superior performance in the MSE sense of the ridge regression estimator, it remains biased. This bias hampers the direct application of most machinery presented in the statistical literature as that is geared towards unbiased estimators. Unbiasedness facilitates proper inference and confidence interval construction. For instance, an estimator with a large bias and small variance may yield a confidence interval that does not contain the true parameter value. Hence, several proposals to de-bias the ridge regression estimator have been presented.

A straightforward approach would be to correct for the bias of the estimator (1.6), using the ‘non-debiased’ ridge regression estimator as an estimator for the regression parameter. This debiased ridge regression estimator is exactly what Zhang and Politis (2022) propose:

$$\hat{\beta}_d(\lambda) = \hat{\beta}(\lambda) + \lambda(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \hat{\beta}(\lambda).$$

To study the effect of this debiasing, note that

$$\hat{\beta}_d(\lambda) - \beta = -\lambda^2 (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-2} \beta + (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} [\mathbf{I}_{pp} + \lambda (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}] \mathbf{X}^\top \varepsilon,$$

where we have *i*) substituted the analytic expression of the non-debiased ridge regression estimator, *ii*) substituted $\mathbf{X}\beta + \varepsilon$ for \mathbf{Y} by virtue of the linear model, and *iii*) manipulated the resulting expression using straightforward linear algebra. From the expression of the preceding display, we directly obtain the bias, $\mathbb{E}[\hat{\beta}_d(\lambda)] - \beta = -\lambda^2 (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-2} \beta$, and variance,

$$\begin{aligned} \text{Var}[\hat{\beta}_d(\lambda)] &= \sigma^2 [(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} + \lambda (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-2}] \\ &\quad \times \mathbf{X}^\top \mathbf{X} [(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} + \lambda (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-2}]. \end{aligned}$$

From these expressions, it is – as Zhang and Politis (2022) point out – clear that, if $\lambda d_{x,n}^{-2}$ tends to zero as $n \rightarrow \infty$, the bias of debiased estimator is much smaller than that of its non-debiased counterpart, while the difference in their variances becomes negligible. It is not immediate how this condition on $\lambda d_{x,n}^{-2}$ translates practically to finite sample sizes in terms of a class of design matrices and a domain of the regularization parameter.

An alternative debiased ridge regression estimator is proposed in Bühlmann (2013). It starts from the bias decomposition of Proposition (1.1) into a part attributable to the regularization and one to the high-dimensionality. Bühlmann (2013) assumes a relatively small penalty parameter such that effectively the regularization bias, $\mathbf{P}_x \{\mathbb{E}[\hat{\beta}(\lambda)] - \beta\}$, is smaller than the standard error and, thereby, not of primary concern. We are then left to reduce the bias introduced by the high-dimensionality, $(\mathbf{P}_x - \mathbf{I}_{pp})\beta$, called *projection bias* in Bühlmann (2013). Hereto Bühlmann (2013) assumes the existence of alternative but accurate estimator $\hat{\beta}^{(\text{init})}$. Under (strong?) assumptions, the lasso regression estimator (Chapter 6) may serve as such. With this alternative, accurate estimator at hand, the projection bias is eliminated by replacing β by $\hat{\beta}^{(\text{init})}$ and subtracting the projection bias term from the non-debiased ridge regression estimator.

1.5 Constrained estimation

The ad-hoc fix of Hoerl and Kennard (1970) to super-collinearity of the design matrix (and, consequently the singularity of the matrix $\mathbf{X}^\top \mathbf{X}$) has been motivated post-hoc. The ridge regression estimator minimizes the *ridge loss function*, which is defined as:

$$\mathcal{L}_{\text{ridge}}(\beta; \lambda) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 = \sum_{i=1}^n (Y_i - \mathbf{X}_{i*}\beta)^2 + \lambda \sum_{j=1}^p \beta_j^2. \quad (1.8)$$

This loss function is the traditional sum-of-squares augmented with a *penalty*. The particular form of the penalty, $\lambda \|\beta\|_2^2$ is referred to as the *ridge penalty* or *ridge regularization term* and λ as the *penalty parameter* or *regularization parameter*. For $\lambda = 0$, minimization of the ridge loss function yields the ML estimator (if it exists). For any $\lambda > 0$, the ridge penalty contributes to the loss function, affecting its minimum and its location. The minimum of the sum-of-squares is well-known. The minimum of the ridge penalty is attained at $\beta = \mathbf{0}_p$ whenever $\lambda > 0$. The β that minimizes $\mathcal{L}_{\text{ridge}}(\beta; \lambda)$ then balances the sum-of-squares and the penalty. The effect of the penalty in this balancing act is to shrink the regression coefficients towards zero, its minimum. In particular, the larger λ , the

larger the contribution of the penalty to the loss function, the stronger the tendency to shrink non-zero regression coefficients to zero (and decrease the contribution of the penalty to the loss function). This motivates the name ‘penalty’ as non-zero elements of β increase (or penalize) the loss function.

To verify that the ridge regression estimator indeed minimizes the ridge loss function, proceed as usual. Take the derivative with respect to β :

$$\frac{\partial}{\partial \beta} \mathcal{L}_{\text{ridge}}(\beta; \lambda) = -2\mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\beta) + 2\lambda \mathbf{I}_{pp}\beta = -2\mathbf{X}^\top \mathbf{Y} + 2(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})\beta.$$

Equate the derivative to zero and solve for β . This yields the ridge regression estimator.

The ridge regression estimator is thus a stationary point of the ridge loss function. A stationary point corresponds to a minimum if the Hessian matrix with second order partial derivatives is positive definite. The Hessian of the ridge loss function is

$$\frac{\partial^2}{\partial \beta \partial \beta^\top} \mathcal{L}_{\text{ridge}}(\beta; \lambda) = 2(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp}).$$

This Hessian is the sum of the (semi-)positive definite matrix $\mathbf{X}^\top \mathbf{X}$ and the positive definite matrix $\lambda \mathbf{I}_{pp}$. Lemma 14.2.4 of Harville (2008) then states that the sum of these matrices is itself a positive definite matrix. Hence, the Hessian is positive definite and the ridge loss function has a stationary point at the ridge regression estimator, which is a minimum.

The ridge regression estimator minimizes the ridge loss function. It remains to verify that it is a global minimum. To this end we introduce the concept of a convex function. As a prerequisite, a set $\mathcal{S} \subset \mathbb{R}^p$ is called *convex* if for all $\beta_1, \beta_2 \in \mathcal{S}$ their weighted average $\beta_\theta = (1 - \theta)\beta_1 + \theta\beta_2$ for all $\theta \in [0, 1]$ is itself an element of \mathcal{S} , thus $\beta_\theta \in \mathcal{S}$. If for all $\theta \in (0, 1)$, the weighted average β_θ is inside \mathcal{S} and not on its boundary, the set is called *strictly convex*. Examples of (strictly) convex and nonconvex sets are depicted in Figure 1.4. A function $f(\cdot)$ is (*strictly*) *convex* if the set $\{y : y \geq f(\beta) \text{ for all } \beta \in \mathcal{S} \text{ for any convex } \mathcal{S}\}$, called the epigraph of $f(\cdot)$, is (strictly) convex. Examples of (strictly) convex and nonconvex functions are depicted in Figure 1.4. The ridge loss function is the sum of two parabola’s: one is at least convex and the other strictly convex in β . The sum of a convex and strictly convex function is itself strictly convex (see Lemma 9.4.2 of Fletcher 2008). The ridge loss function is thus strictly convex. Theorem 9.4.1 of Fletcher 2008 then warrants, by the strict convexity of the ridge loss function, that the ridge regression estimator is a global minimum.

From the ridge loss function the limiting behavior of the variance of the ridge regression estimator can be understood. The ridge penalty with its minimum $\beta = \mathbf{0}_p$ does not involve data and, consequently, the variance of its minimum equals zero. With the ridge regression estimator being a compromise between the maximum likelihood estimator and the minimum of the penalty, so is its variance a compromise of their variances. As λ tends to infinity, the ridge regression estimator and its variance converge to the minimizer of the loss function and the variance of the minimizer, respectively. Hence, in the limit (large λ) the variance of the ridge regression estimator vanishes. Understandably, as the penalty now fully dominates the loss function and, consequently, it does no longer involve data (i.e. randomness).

The loss function of the ridge regression estimator facilitates another view on the estimator. Hereto now define the ridge regression estimator as:

$$\hat{\beta}(\lambda) = \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2. \quad (1.9)$$

This minimization problem can be reformulated into the following constrained optimization problem:

$$\hat{\beta}(\lambda) = \arg \min_{\{\beta \in \mathbb{R}^p : \|\beta\|_2^2 \leq c\}} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2, \quad (1.10)$$

for some suitable $c > 0$. The constrained optimization problem (1.10) can be solved by means of the Karush-Kuhn-Tucker (KKT) multiplier method (Fletcher, 2008), which minimizes a function subject to inequality constraints. The KKT multiplier method states that, under some regularity conditions (all met here), there exists a constant $\nu \geq 0$, called the *multiplier*, such that the solution $\hat{\beta}(\nu)$ of the constrained minimization problem (1.10) satisfies the so-called KKT conditions. The first KKT condition (referred to as the stationarity condition) demands that the gradient (with respect to β) of the Lagrangian associated with the minimization problem equals zero at the solution $\hat{\beta}(\nu)$. The Lagrangian for problem (1.10) is:

$$\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \nu(\|\beta\|_2^2 - c).$$

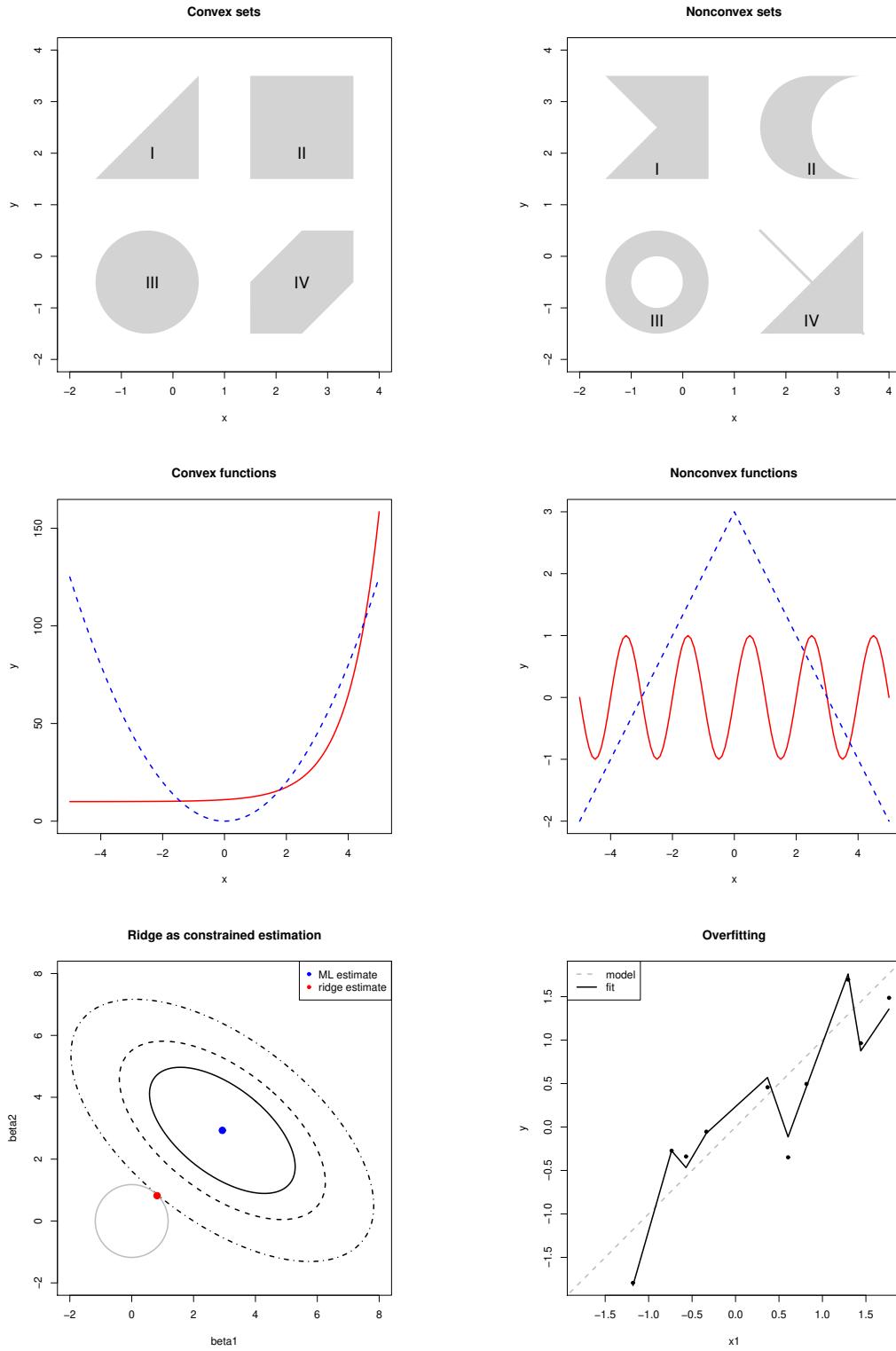


Figure 1.4: Top panels show examples of convex (left) and nonconvex (right) sets. Middle panels show examples of convex (left) and nonconvex (right) functions. The left bottom panel illustrates the ridge estimation as a constrained estimation problem. The ellipses represent the contours of the ML loss function, with the blue dot at the center the ML estimate. The circle is the ridge parameter constraint. The red dot is the ridge estimate. It is at the intersection of the ridge constraint and the smallest contour with a non-empty intersection with the constraint. The right bottom panel shows the data corresponding to Example 1.7. The grey line represents the ‘true’ relationship, while the black line the fitted one.

The second KKT condition (the complementarity condition) requires that $\nu(\|\hat{\beta}(\nu)\|_2^2 - c) = 0$. If $\nu = \lambda$ and $c = \|\hat{\beta}(\lambda)\|_2^2$, the ridge regression estimator $\beta(\lambda)$ satisfies both KKT conditions. Hence, both problems have the same solution if $c = \|\hat{\beta}(\lambda)\|_2^2$.

The constrained estimation interpretation of the ridge regression estimator is illustrated in the left bottom panel of Figure 1.4. It shows the level sets of the sum-of-squares criterion and centered around zero the circular ridge parameter constraint, parametrized by $\{\beta : \|\beta\|_2^2 = \beta_1^2 + \beta_2^2 \leq c\}$ for some $c > 0$. The ridge regression estimator is then the point where the sum-of-squares' smallest level set hits the constraint. Exactly at that point the sum-of-squares is minimized over those β 's that live on or inside the constraint. In the high-dimensional setting the ellipsoidal level sets are degenerated. For instance, in the 2-dimensional case of the left bottom panel of Figure 1.4, the ellipsoids would then be lines but the geometric interpretation is unaltered.

The ridge regression estimator is always to be found on the boundary of the ridge parameter constraint and is never an interior point. To see this, assume, for simplicity, that the $\mathbf{X}^\top \mathbf{X}$ matrix is of full rank. The radius of the ridge parameter constraint can then be bounded as follows:

$$\|\hat{\beta}(\lambda)\|_2^2 = \mathbf{Y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-2} \mathbf{X}^\top \mathbf{Y} \leq \mathbf{Y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-2} \mathbf{X}^\top \mathbf{Y} = \|\hat{\beta}^m\|_2^2.$$

The inequality in the display above follows from *i*) $\mathbf{X}^\top \mathbf{X} \succ 0$ and $\lambda \mathbf{I}_{pp} \succ 0$, *ii*) $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp} \succ \mathbf{X}^\top \mathbf{X}$ (due to Lemma 14.2.4 of Harville, 2008), and *iii*) $(\mathbf{X}^\top \mathbf{X})^{-2} \succ (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-2}$ (inferring Corollary 7.7.4 of Horn and Johnson, 2012). The ridge regression estimator is thus always on the boundary or in a circular constraint centered around the origin with a radius that is smaller than the size of the maximum likelihood estimator. Moreover, the constrained estimation formulation of the ridge regression estimator then implies that the latter must be on the boundary of the constraint.

The size of the spherical ridge parameter constraint shrinks monotonously as λ increases, and eventually, in the $\lambda \rightarrow \infty$ -limit collapses to zero (as is formalized by Proposition 1.4).

Proposition 1.4

The squared norm of the ridge regression estimator satisfies:

- i*) $d\|\hat{\beta}(\lambda)\|_2^2/d\lambda < 0$ for $\lambda > 0$,
- ii*) $\lim_{\lambda \rightarrow \infty} \|\hat{\beta}(\lambda)\|_2^2 = 0$.

Proof. For part *i*) we need to verify that $\|\hat{\beta}(\lambda)\|_2^2 > \|\hat{\beta}(\lambda + h)\|_2^2$ for $h > 0$. Hereto substitute the singular value decomposition of the design matrix, $\mathbf{X} = \mathbf{U}_x \mathbf{D}_x \mathbf{V}_x^\top$, into the display above. Then, after a little algebra, take the derivative with respect to λ , and obtain:

$$\frac{d}{d\lambda} \|\hat{\beta}(\lambda)\|_2^2 = -2 \sum_{i=1}^n d_{x,i}^2 (d_{x,i}^2 + \lambda)^{-3} (\mathbf{Y}^\top \mathbf{U}_{x,*i})^2.$$

This is negative for all $\lambda > 0$. Indeed, the parameter constraint thus becomes smaller and smaller as λ increases, and so does the size of the estimator.

Part *ii*) follows from $\lim_{\lambda \rightarrow \infty} \hat{\beta}(\lambda) = \mathbf{0}_p$, which has been concluded previously by other means. ■

The relevance of viewing the ridge regression estimator as the solution to a constrained estimation problem becomes obvious when considering a typical threat to high-dimensional data analysis: overfitting. *Overfitting* refers to the phenomenon of modelling the noise rather than the signal. In case the true model is parsimonious (few covariates driving the response) and data on many covariates are available, it is likely that a linear combination of all covariates yields a higher likelihood than a combination of the few that are actually related to the response. As only the few covariates related to the response contain the signal, the model involving all covariates then cannot but explain more than the signal alone: it also models the error. Hence, it overfits the data. In high-dimensional settings overfitting is a real threat. The number of explanatory variables exceeds the number of observations. It is thus possible to form a linear combination of the covariates that perfectly explains the response, including the noise.

Large estimates of regression coefficients are often an indication of overfitting. Augmentation of the estimation procedure with a constraint on the regression coefficients is a simple remedy to large parameter estimates. As a consequence it decreases the probability of overfitting. Overfitting is illustrated in the next example.

Example 1.7 (Overfitting)

Consider an artificial data set comprising of ten observations on a response Y_i and nine covariates $X_{i,j}$. All covariate data are sampled from the standard normal distribution: $X_{i,j} \sim \mathcal{N}(0, 1)$. The response is generated by $Y_i = X_{i,1} + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \frac{1}{4})$. Hence, only the first covariate contributes to the response.

The regression model $Y_i = \sum_{j=1}^9 X_{i,j}\beta_j + \varepsilon_i$ is fitted to the artificial data using R. This yields the regression parameter estimates:

$$\hat{\beta}^\top = (0.048, -2.386, -5.528, 6.243, -4.819, 0.760, -3.345, -4.748, 2.136).$$

As $\beta^\top = (1, 0, \dots, 0)$, many regression coefficient are clearly over-estimated.

The fitted values $\hat{Y}_i = \mathbf{X}_i\hat{\beta}$ are plotted against the values of the first covariates in the right bottom panel of Figure 1.4. As a reference the line $x = y$ is added, which represents the ‘true’ model. The fitted model follows the ‘true’ relationship. But it also captures the deviations from this line that represent the errors. \square

1.6 Degrees of freedom

The degrees of freedom consumed by the ridge regression estimator, which may aid in the choice of the value of the penalty parameter, is derived. A formal definition of the degrees of freedom, which can among others be found in Efron (1986), is

$$df(\lambda) = \sum_{i=1}^n [\text{Var}(Y_i)]^{-1} \text{Cov}(\hat{Y}_i, Y_i). \quad (1.11)$$

It represents the effective number of parameters used by the estimator. It can also be viewed as the amount of self-explanation by the observations of the fit. Recall from ordinary regression that $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \mathbf{H}\mathbf{Y}$ where \mathbf{H} is the hat matrix. Application of the defintion (1.11) then yields the degrees of freedom used by the maximum likelihood regression estimator and equals $\text{tr}(\mathbf{H})$, the trace of \mathbf{H} . In particular, if \mathbf{X} is of full rank, i.e. $\text{rank}(\mathbf{X}) = p$, then $\text{tr}(\mathbf{H}) = p$.

We adopt the degrees of freedom definition (1.11) for the ridge regression estimator. We then find

$$\begin{aligned} df(\lambda) &= \sum_{i=1}^n [\text{Var}(Y_i)]^{-1} \text{Cov}(\hat{Y}_i, Y_i) \\ &= \sigma^{-2} \sum_{i=1}^n \text{Cov}[\mathbf{X}_{i,*}\hat{\beta}(\lambda), Y_i] \\ &= \sigma^{-2} \sum_{i=1}^n \text{Cov}[\mathbf{X}_{i,*}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X} \mathbf{Y}, Y_i] \\ &= \sigma^{-2} \sum_{i=1}^n \mathbf{X}_{i,*}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X} \text{Cov}(\mathbf{Y}, Y_i) \\ &= \text{tr}[\mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top] \\ &= \sum_{j=1}^p (\mathbf{D}_x^\top \mathbf{D}_x)_{jj} [(\mathbf{D}_x^\top \mathbf{D}_x)_{jj} + \lambda]^{-1}, \end{aligned}$$

where we have used the independence among the observations. High-dimensionally, the sum on the right-hand side of the last line of the display above may be limited to n . The degrees of freedom consumed by the ridge regression estimator is monotone decreasing in λ . In particular, $\lim_{\lambda \rightarrow \infty} df(\lambda) = 0$. That is, in the limit no information from \mathbf{X} is used. Indeed, $\hat{\beta}(\lambda)$ is forced to equal $\mathbf{0}_p$ which is not derived from data. Finally, from the derivation in the display above we may deduce a definition of the ‘ridge hat matrix’: $\mathbf{H}(\lambda) := \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top$. We can then write, in analogy to the ordinary regression case, $df(\lambda) = \text{tr}[\mathbf{H}(\lambda)]$.

1.7 Computationally efficient evaluation

In the high-dimensional setting the number of covariates p is large compared to the number of samples n . In a microarray experiment $p = 40000$ and $n = 100$ is not uncommon. To perform ridge regression in this context, the following expression needs to be evaluated numerically: $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y}$. For $p = 40000$ this requires the inversion of a 40000×40000 dimensional matrix. This is not feasible on most desktop computers.

There, however, is a workaround to the computational burden. Revisit the singular value decomposition of $\mathbf{X} = \mathbf{U}_x \mathbf{D}_x \mathbf{V}_x^\top$. Drop from \mathbf{D}_x and \mathbf{V}_x the columns that correspond to zero singular values. The resulting \mathbf{D}_x and \mathbf{V}_x are $n \times n$ and $p \times n$ -dimensional, respectively. Note that dropping these columns has no effect on the matrix factorization of \mathbf{X} , i.e. still $\mathbf{X} = \mathbf{U}_x \mathbf{D}_x \mathbf{V}_x^\top$ but with the last two matrices in this decomposition defined differently from the traditional singular value decomposition. Now write $\mathbf{R}_x = \mathbf{U}_x \mathbf{D}_x$. As both \mathbf{U}_x and \mathbf{D}_x are $n \times n$ -dimensional matrices, so is \mathbf{R}_x . Consequently, \mathbf{X} is now decomposed as $\mathbf{X} = \mathbf{R}_x \mathbf{V}_x^\top$. The ridge regression

estimator can be rewritten in terms of \mathbf{R}_x and \mathbf{V}_x :

$$\begin{aligned}\hat{\beta}(\lambda) &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y} \\ &= (\mathbf{V}_x \mathbf{R}_x^\top \mathbf{R}_x \mathbf{V}_x^\top + \lambda \mathbf{V}_x \mathbf{V}_x^\top)^{-1} \mathbf{V}_x \mathbf{R}_x^\top \mathbf{Y} = \mathbf{V}_x (\mathbf{R}_x^\top \mathbf{R}_x + \lambda \mathbf{I}_{nn})^{-1} \mathbf{V}_x^\top \mathbf{V}_x \mathbf{R}_x^\top \mathbf{Y} \\ &= \mathbf{V}_x (\mathbf{R}_x^\top \mathbf{R}_x + \lambda \mathbf{I}_{nn})^{-1} \mathbf{R}_x^\top \mathbf{Y}.\end{aligned}$$

Hence, the reformulated ridge regressin estimator involves the inversion of an $n \times n$ -dimensional matrix. With $n = 100$ this is feasible on most standard computers.

Hastie and Tibshirani (2004) point out that, with the SVD-trick above, the number of computation operations reduces from $\mathcal{O}(p^3)$ to $\mathcal{O}(pn^2)$. In addition, they point out that this computational short-cut can be used in combination with other loss functions, for instance that of standard generalized linear models (see Chapter 5). This computation is illustrated in Figure 1.5, which shows the substantial gain in computation time of the evaluation of the ridge regression estimator using the efficient over the naive implementation against the dimension p . Details of this figure are provided in Question 1.24.

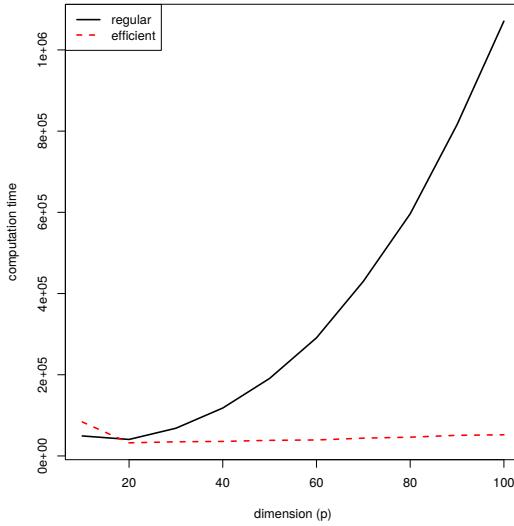


Figure 1.5: Computation time of the evaluation of the ridge regression estimator using the naive and efficient implementation against the dimension p .

The inversion of the $p \times p$ -dimensional matrix can be avoided in an other way. Hereto one needs the Woodbury identity. Let \mathbf{A} , \mathbf{U} and \mathbf{V} be $p \times p$ -, $p \times n$ - and $n \times p$ -dimensional matrices, respectively. The (simplified form of the) Woodbury identity then is:

$$(\mathbf{A} + \mathbf{UV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{U} (\mathbf{I}_{nn} + \mathbf{V} \mathbf{A}^{-1} \mathbf{U})^{-1} \mathbf{V} \mathbf{A}^{-1}.$$

Application of the Woodbury identity to the matrix inverse in the ridge estimator of the regression parameter gives:

$$\begin{aligned}\hat{\beta}(\lambda) &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y} = [\lambda^{-1} \mathbf{I}_{pp} - \lambda^{-2} \mathbf{X}^\top (\mathbf{I}_{nn} + \lambda^{-1} \mathbf{X} \mathbf{X}^\top)^{-1} \mathbf{X}] \mathbf{X}^\top \mathbf{Y} \\ &= \lambda^{-1} \mathbf{X}^\top (\mathbf{I}_{nn} + \lambda^{-1} \mathbf{X} \mathbf{X}^\top)^{-1} \mathbf{Y}.\end{aligned}\tag{1.12}$$

The inversion of the $p \times p$ -dimensional matrix $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp}$ is thus replaced by that of the $n \times n$ -dimensional matrix $\mathbf{I}_{nn} + \lambda^{-1} \mathbf{X} \mathbf{X}^\top$. In addition, this expression of the ridge regression estimator avoids the singular value decomposition of \mathbf{X} , which may in some cases introduce additional numerical errors (e.g. at the level of machine precision).

1.8 Penalty parameter selection

Throughout the introduction of the ridge regression estimator and the subsequent discussion of its properties, we considered the penalty parameter known or ‘given’. In practice, it is unknown and the user needs to make an informed decision on its value. We present several strategies to facilitate such a decision. Prior to that, we discuss

some sanity requirements one may wish to impose on the ridge regression estimator. Those requirements do not yield a specific choice of the penalty parameter but they specify a domain of sensible penalty parameter values.

The evaluation of the ridge regression estimator may be subject to numerical inaccuracy, which ideally is avoided. This numerical inaccuracy results from the ill-conditionedness of $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp}$. A matrix is *ill-conditioned* if its condition number is high. The *condition number* of a square positive definite matrix \mathbf{A} is the ratio of its largest and smallest eigenvalue. If the smallest eigenvalue is zero, the condition number is undefined and so is \mathbf{A}^{-1} . Furthermore, a high condition number is indicative of the loss (on a log-scale) in numerical accuracy of the evaluation of \mathbf{A}^{-1} . To ensure the numerical accurate evaluation of the ridge regression estimator, the choice of the penalty parameter is thus to be restricted to a subset of the positive real numbers such that it yields a well-conditioned matrix $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp}$. Clearly, the penalty parameter λ should then not be too close to zero. There is however no consensus on the criteria on the condition number for a matrix to be well-defined. This depends among others on how much numerical inaccuracy is tolerated by the context. Of course, as pointed out in Section 1.7, inversion of the $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp}$ matrix can be circumvented. One then still needs to ensure the well-conditionedness of $\lambda^{-1} \mathbf{X} \mathbf{X}^\top + \mathbf{I}_{nn}$, which too results in a lower bound for the penalty parameter. Practically, following Peeters *et al.* (2019) (who do so in a different context), we suggest to generate a conditional number plot. It plots (on some convenient scale) the condition number of the matrix $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp}$ against the penalty parameter λ . From this plot, we identify the domain of the penalty parameter value associated with well-conditionedness. To guide in this choice, Peeters *et al.* (2019) overlay this plot with a curve indicative of the numerical inaccuracy.

Traditional text books on regression analysis suggest, in order to prevent over-fitting, to limit the number of covariates of the model. This ought to leave enough degrees of freedom to estimate the error and facilitate proper inference. While ridge regression is commonly used for prediction purposes and inference need not be the objective, over-fitting is certainly to be avoided. This can be achieved by limiting the degrees of freedom spent on the estimation of the regression parameter (Harrell, 2001). We thus follow Saleh *et al.* (2019), who illustrate this in the ridge regression context, and use the degrees of freedom to bound the search domain of the penalty parameter. This requires the specification of a maximum degrees of freedom, denoted by ν with $0 \leq \nu \leq n$, one wishes to spend on the construction of the ridge regression estimator. Now choose λ such that $\text{df}[\hat{\beta}(\lambda)] \leq \nu$. To find the bound on the penalty parameter, note that

$$\begin{aligned}\text{df}[\hat{\beta}(\lambda)] &= \text{tr}[\mathbf{H}(\lambda)] = \sum_{j=1}^{\min\{n,p\}} (\mathbf{D}_x^\top \mathbf{D}_x)_{jj} [(\mathbf{D}_x^\top \mathbf{D}_x)_{jj} + \lambda]^{-1} \\ &\leq \min\{n, p\} [d_1(\mathbf{X})]^2 \{[d_1(\mathbf{X})]^2 + \lambda\}^{-1}.\end{aligned}$$

Then, if $\lambda \geq \max\{0, \nu^{-1}[d_1(\mathbf{X})]^2(\min\{n, p\} - \nu)\}$, the degrees of freedom consumed by the ridge regression estimator is smaller than ν . It remains to choose the maximum degrees of freedom ν to be spent on the estimator. This may be determined by the context. If that does not resolve the matter, Harrell (2001) suggests as a rule of thumb to choose ν as a fraction of the sample size but provides no recommendation on the size of this fraction. As a last resort, we can base this choice of ν on the degrees of freedom necessary to obtain a reliable estimate of the error variance, which suggests a conservative upperbound of $\max\{0, n - 30\}$ on ν .

1.8.1 Information criterion

A popular strategy is to choose a penalty parameter that yields a good but parsimonious model. Information criteria measure the balance between model fit and model complexity. Here we present the Akaike's information criterion (AIC, Akaike, 1974), but many other criteria have been presented in the literature (see, e.g. Schwarz, 1978). The AIC measures model fit by the loglikelihood and model complexity as measured by the number of parameters used by the model. The number of model parameters in regular regression simply corresponds to the number of covariates in the model. Or, by the degrees of freedom consumed by the model, which is equivalent to the trace of the hat matrix. For ridge regression it thus seems natural to define model complexity analogously by the trace of the ridge hat matrix. This yields the AIC for the linear regression model with ridge regression estimates:

$$\begin{aligned}\text{AIC}(\lambda) &= 2(\# \text{estimated model parameters}) - 2 \log\{L[\mathbf{Y}, \mathbf{X}; \hat{\beta}(\lambda), \hat{\sigma}^2(\lambda)]\} \\ &= 2 \text{tr}[\mathbf{H}(\lambda)] - 2 \log\{L[\mathbf{Y}, \mathbf{X}; \hat{\beta}(\lambda), \hat{\sigma}^2(\lambda)]\} \\ &= 2 \sum_{j=1}^p (\mathbf{D}_x^\top \mathbf{D}_x)_{jj} [(\mathbf{D}_x^\top \mathbf{D}_x)_{jj} + \lambda]^{-1} \\ &\quad + 2n \log[\sqrt{2\pi} \hat{\sigma}(\lambda)] + \hat{\sigma}^{-2}(\lambda) \sum_{i=1}^n [y_i - \mathbf{X}_{i,*} \hat{\beta}(\lambda)]^2. \\ &= 2 \sum_{j=1}^p (\mathbf{D}_x^\top \mathbf{D}_x)_{jj} [(\mathbf{D}_x^\top \mathbf{D}_x)_{jj} + \lambda]^{-1} + 2n \log[\sqrt{2\pi} \hat{\sigma}(\lambda)] + n,\end{aligned}$$

as $\hat{\sigma}^{-2}(\lambda) = n^{-1} \|\mathbf{Y} - \mathbf{X}\hat{\beta}(\lambda)\|_2^2$. The value of λ which minimizes $\text{AIC}(\lambda)$ corresponds to the ‘optimal’ balance of model complexity and overfitting.

Although information criteria are widely used to guide the choice of the penalty parameter, we comment on their use within the context of ridge regression. Information criteria guide the decision process when having to decide among various different models. Different models use different sets of explanatory variables to explain the behaviour of the response variable. In that sense, the use of information criteria for the deciding on the ridge penalty parameter may be considered inappropriate: ridge regression uses the same set of explanatory variables irrespective of the value of the penalty parameter. Moreover, often ridge regression is employed to predict a response and not to provide an insightful explanatory model. The latter need not yield the best predictions. Finally, empirically we observed that the AIC may have its optimum, not inside but, at the boundaries of the domain of the ridge penalty parameter.

1.8.2 Cross-validation

Instead of choosing the penalty parameter to balance model fit with model complexity, cross-validation requires it (i.e. the penalty parameter) to yield a model with good prediction performance. Commonly, this performance is evaluated on novel data. Novel data need not be easy to come by and one has to make do with the data at hand. The setting of ‘original’ and novel data is then mimicked by sample splitting: the data set is divided into two (groups of) samples. One of these two data sets, called the *training set*, plays the role of ‘original’ data on which the model is built. The second of these data sets, called the *test set*, plays the role of the ‘novel’ data and is used to evaluate the prediction performance (often operationalized as the loglikelihood or the prediction error) of the model built on the training data set. This procedure (model building and prediction evaluation on training and test set, respectively) is done for a collection of possible penalty parameter choices. The penalty parameter that yields the model with the best prediction performance is to be preferred. The thus obtained performance evaluation depends on the actual split of the data set. To remove this dependence, the data set is split many times into a training and test set. Per split the model parameters are estimated for all choices of λ using the training data and estimated parameters are evaluated on the corresponding test set. The penalty parameter, that on average over the test sets performs best (in some sense), is then selected.

When the repetitive splitting of the data set is done randomly, samples may accidentally end up in a vast majority of the splits in either training or test set. Such samples may have an unbalanced influence on either model building or prediction evaluation. To avoid this k -fold cross-validation structures the data splitting. The samples are divided into k more or less equally sized exhaustive and mutually exclusive subsets. In turn (at each split) one of these subsets plays the role of the test set while the union of the remaining subsets constitutes the training set. Such a splitting warrants a balanced representation of each sample in both training and test set over the splits. Still the division into the k subsets involves a degree of randomness. This may be fully excluded when choosing $k = n$. This particular case is referred to as leave-one-out cross-validation (LOOCV). For illustration purposes the LOOCV procedure is detailed fully below:

- 0) Define a range of interest for the penalty parameter.
- 1) Divide the data set into training and test set comprising samples $\{1, \dots, n\} \setminus i$ and $\{i\}$, respectively.
- 2) Fit the linear regression model by means of ridge estimation for each λ in the grid using the training set. This yields:

$$\hat{\beta}_{-i}(\lambda) = (\mathbf{X}_{-i,*}^\top \mathbf{X}_{-i,*} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}_{-i,*}^\top \mathbf{Y}_{-i},$$

where $\mathbf{X}_{-i,*}$ and \mathbf{Y}_{-i} are the design matrix and response vector with the i -th row and element, respectively, excluded. The corresponding estimate of the error variance $\hat{\sigma}_{-i}^2(\lambda)$.

- 3) Evaluate the prediction performance of these models on the test set by $\log\{L[Y_i, \mathbf{X}_{i,*}; \hat{\beta}_{-i}(\lambda), \hat{\sigma}_{-i}^2(\lambda)]\}$. Or, by the prediction error $|Y_i - \mathbf{X}_{i,*}\hat{\beta}_{-i}(\lambda)|$, possibly squared.
- 4) Repeat steps 1) to 3) such that each sample plays the role of the test set once.
- 5) Average the prediction performances of the test sets at each grid point of the penalty parameter:

$$n^{-1} \sum_{i=1}^n \log\{L[Y_i, \mathbf{X}_{i,*}; \hat{\beta}_{-i}(\lambda), \hat{\sigma}_{-i}^2(\lambda)]\}.$$

The quantity above is called the *cross-validated loglikelihood*. It is an estimate of the prediction performance of the model corresponding to this value of the penalty parameter on novel data.

- 6) The value of the penalty parameter that maximizes the cross-validated loglikelihood is the value of choice.

The procedure is straightforwardly adopted to k -fold cross-validation, a different criterion, and different estimators.

In the LOOCV procedure above resampling can be avoided when the predictive performance is measured by Allen's PRESS (Predicted Residual Error Sum of Squares) statistic (Allen, 1974). For then, the LOOCV predictive performance can be expressed analytically in terms of the known quantities derived from the design matrix and response (as pointed out but not detailed in Golub *et al.* 1979). Define the optimal penalty parameter to minimize Allen's PRESS statistic:

$$\lambda_{\text{opt}} = \arg \min_{\lambda > 0} n^{-1} \sum_{i=1}^n [Y_i - \mathbf{X}_{i,*}\hat{\boldsymbol{\beta}}_{-i}(\lambda)]^2 = \arg \min_{\lambda > 0} n^{-1} \|\mathbf{B}(\lambda)[\mathbf{I}_{nn} - \mathbf{H}(\lambda)]\mathbf{Y}\|_2^2,$$

where $\mathbf{B}(\lambda)$ is diagonal with $[\mathbf{B}(\lambda)]_{ii} = [1 - \mathbf{H}_{ii}(\lambda)]^{-1}$. The second equality in the preceding display is elaborated in Exercise 1.25. Its main takeaway is that the predictive performance for a given λ can be assessed directly from the ridge hat matrix and the response vector without the recalculation of the n leave-one-out ridge regression estimators. Computationally, this is a considerable gain.

No such analytic expression of the cross-validated loss as above exists for general K -fold cross-validation, but considerable computational gain can nonetheless be achieved (van de Wiel *et al.*, 2021). This exploits the fact that the ridge regression estimator appears in Allen's PRESS statistics – and the likelihood – only in combination with the design matrix, together forming the linear predictor. There is thus no need to evaluate the estimator itself when interest is only in its predictive performance. Then, if $\mathcal{G}_1, \dots, \mathcal{G}_K \subset \{1, \dots, n\}$ are the mutually exclusive and exhaustive K -fold sample index sets, the linear predictor for the k -th fold can be expressed as:

$$\mathbf{X}_{\mathcal{G}_k,*}\hat{\boldsymbol{\beta}}_{-\mathcal{G}_k}(\lambda) = \lambda^{-1}\mathbf{X}_{\mathcal{G}_k,*}\mathbf{X}_{-\mathcal{G}_k,*}^\top(\lambda^{-1}\mathbf{X}_{-\mathcal{G}_k,*}\mathbf{X}_{-\mathcal{G}_k,*}^\top + \mathbf{I}_{n-|\mathcal{G}_k|,n-|\mathcal{G}_k|})^{-1}\mathbf{Y}_{-\mathcal{G}_k,*}, \quad (1.13)$$

where we have used the Woodbury identity again. Finally, for each fold the computationally most demanding matrices of this expression, $\mathbf{X}_{\mathcal{G}_k,*}\mathbf{X}_{-\mathcal{G}_k,*}^\top$ and $\mathbf{X}_{-\mathcal{G}_k,*}\mathbf{X}_{-\mathcal{G}_k,*}^\top$, are both submatrices of $\mathbf{X}\mathbf{X}^\top$. If the latter matrix is evaluated prior to the cross-validation loop, all calculations inside the loop involve only matrices of dimensions $n \times n$, maximally, and obtained from $\mathbf{X}\mathbf{X}^\top$ by subsetting.

1.8.3 Generalized cross-validation

Generalized cross-validation is another method to guide the choice of the penalty parameter. It is like cross-validation but with a different criterion to evaluate the performance of the ridge regression estimator on novel data. This criterion, denoted GCV(λ) (where GCV is an acronym of Generalized Cross-Validation), is an approximation to Allen's PRESS statistic. In the previous subsection this statistic was reformulated as:

$$n^{-1} \sum_{i=1}^n [Y_i - \mathbf{X}_{i,*}\hat{\boldsymbol{\beta}}_{-i}(\lambda)]^2 = n^{-1} \sum_{i=1}^n [1 - \mathbf{H}_{ii}(\lambda)]^{-2}[Y_i - \mathbf{X}_{i,*}^\top\hat{\boldsymbol{\beta}}(\lambda)]^2.$$

The identity $\text{tr}[\mathbf{H}(\lambda)] = \sum_{i=1}^n [\mathbf{H}(\lambda)]_{ii}$ suggests $[\mathbf{H}(\lambda)]_{ii} \approx n^{-1}\text{tr}[\mathbf{H}(\lambda)]$. The approximation thus proposed by Golub *et al.* (1979), which they endow with a ‘weighted version of Allen's PRESS statistic’-interpretation, is:

$$\begin{aligned} \text{GCV}(\lambda) &= n^{-1} \sum_{i=1}^n \{1 - n^{-1}\text{tr}[\mathbf{H}(\lambda)]\}^{-2}[Y_i - \mathbf{X}_{i,*}^\top\hat{\boldsymbol{\beta}}(\lambda)]^2 \\ &= n^{-1}\{1 - n^{-1}\text{tr}[\mathbf{H}(\lambda)]\}^{-2}\|[\mathbf{I}_{nn} - \mathbf{H}(\lambda)]\mathbf{Y}\|_2^2. \end{aligned}$$

The need for this approximation is pointed out by Golub *et al.* (1979) by example through an ‘extreme’ case where the minimization of Allen's PRESS statistic fails to produce a well-defined choice of the penalty parameter λ . This ‘extreme’ case requires a (unit) diagonal design matrix \mathbf{X} . Straightforward (linear) algebraic manipulations of Allen's PRESS statistic then yield:

$$n^{-1} \sum_{i=1}^n [1 - \mathbf{H}_{ii}(\lambda)]^{-2}[Y_i - \mathbf{X}_{i,*}^\top\hat{\boldsymbol{\beta}}(\lambda)]^2 = n^{-1} \sum_{i=1}^n Y_i^2,$$

which indeed has no unique minimizer in λ . Additionally, the GCV(λ) criterion may in some cases be preferred computationally when it is easier to evaluate $\text{tr}[\mathbf{H}(\lambda)]$ (e.g. from the singular values) than the individual diagonal elements of $\mathbf{H}(\lambda)$.

1.8.4 Randomness

The discussed procedures for penalty parameter selection all depend on the data at hand. As a result, so does the selected penalty parameter. It should therefore be considered a statistic, a quantity calculated from data. In case

of K -fold cross-validation, the random formation of the splits adds another layer of randomness. Irrespectively, a statistic exhibits randomness. This randomness propagates into the ridge regression estimator. The distributional properties of the ridge regression estimator derived in Section 1.4 are thus conditional on the penalty parameter. Those of the unconditional distribution of the ridge regression estimator, now denoted $\hat{\beta}(\hat{\lambda})$ with $\hat{\lambda}$ indicating

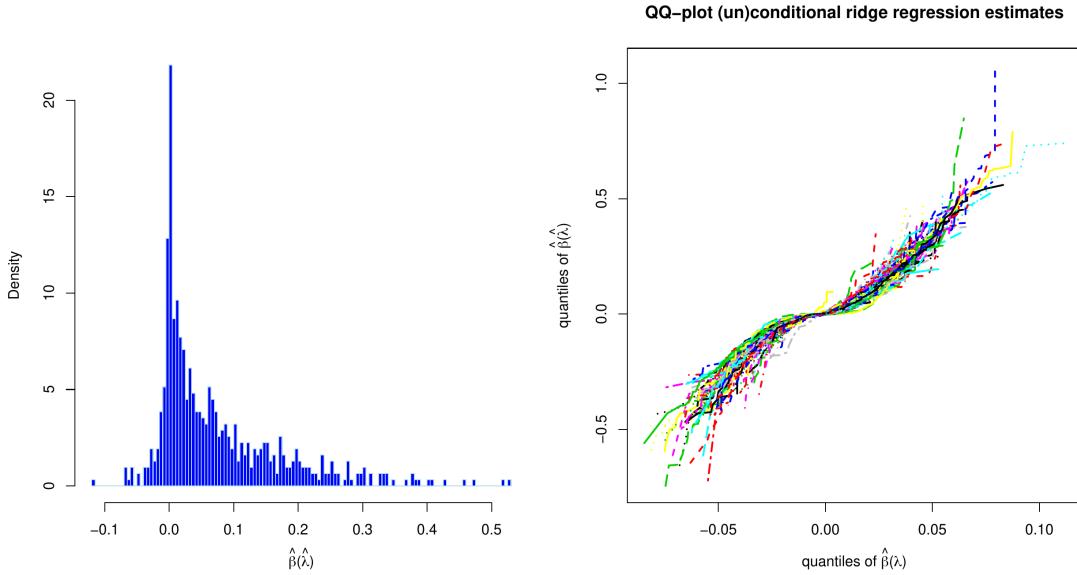


Figure 1.6: Left panel: histogram of estimates of an element of $\hat{\beta}(\hat{\lambda})$ obtained from thousand bootstrap samples with re-tuned penalty parameters. Right panel: qq-plot of the estimates estimates of an element of $\hat{\beta}(\hat{\lambda})$ obtained from thousand bootstrap samples with re-tuned vs. fixed penalty parameters.

that the penalty depends on the data (and possibly the particulars of the splits), may be rather different. Analytic finite sample results appear to be unavailable. An impression of the distribution of $\hat{\beta}(\hat{\lambda})$ can be obtained through simulation. Hereto we have first drawn a data set from the linear regression model with dimension $p = 100$, sample size $n = 25$, a standard normally distribution error, the rows of the design matrix sampled from a zero-centered multivariate normal distribution with a uniformly correlated but equivariant covariance matrix, and a regression parameter with elements equidistantly distributed over the interval $[-2\frac{1}{2}, 2\frac{1}{2}]$. From this data set, we then generate thousand nonparametric bootstrapped data sets. For each bootstrapped data set, we select the penalty parameter by means of $K = 10$ -fold cross-validation and evaluate the ridge regression estimator using the selected penalty parameter. The left panel of Figure 1.6 shows the histogram of the thus acquired estimates of an arbitrary element of the regression parameter. The shape of the distribution clearly deviates from the normal one of the conditional ridge regression estimator. In the right panel of Figure 1.6, we have plotted element-wise quantiles of the conditional vs. unconditional ridge regression estimators. It reveals that not only the shape of the distribution, but also its moments are affected by the randomness of the penalty parameter.

1.9 Simulations

Simulations are presented that illustrate properties of the ridge regression estimator not discussed explicitly in the previous sections of this chapter.

1.9.1 Role of the variance of the covariates

In many applications of high-dimensional data the covariates are standardized prior to the execution of the ridge regression. Before we discuss whether this is appropriate, we first illustrate the effect of ridge penalization on covariates with distinct variances using simulated data.

The simulation involves one response to be (ridge) regressed on fifty covariates. Data (with $n = 1000$) for the covariates, denoted \mathbf{X} , are drawn from a multivariate normal distribution: $\mathbf{X} \sim \mathcal{N}(\mathbf{0}_{50}, \Sigma)$ with Σ diagonal and $(\Sigma)_{jj} = j/10$. From this the response is generated through $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\beta = \mathbf{1}_{50}$ and $\varepsilon \sim \mathcal{N}(\mathbf{0}_{50}, \mathbf{I}_{50 \times 50})$.

With the simulated data at hand the ridge regression estimators of β are evaluated for a large grid of the penalty parameter λ . The resulting ridge regularization paths of the regression coefficients are plotted (Figure 1.7). All

paths start close to one and vanish as $\lambda \rightarrow \infty$. However, ridge regularization paths of regression coefficients corresponding to covariates with a large variance dominate those with a low variance.

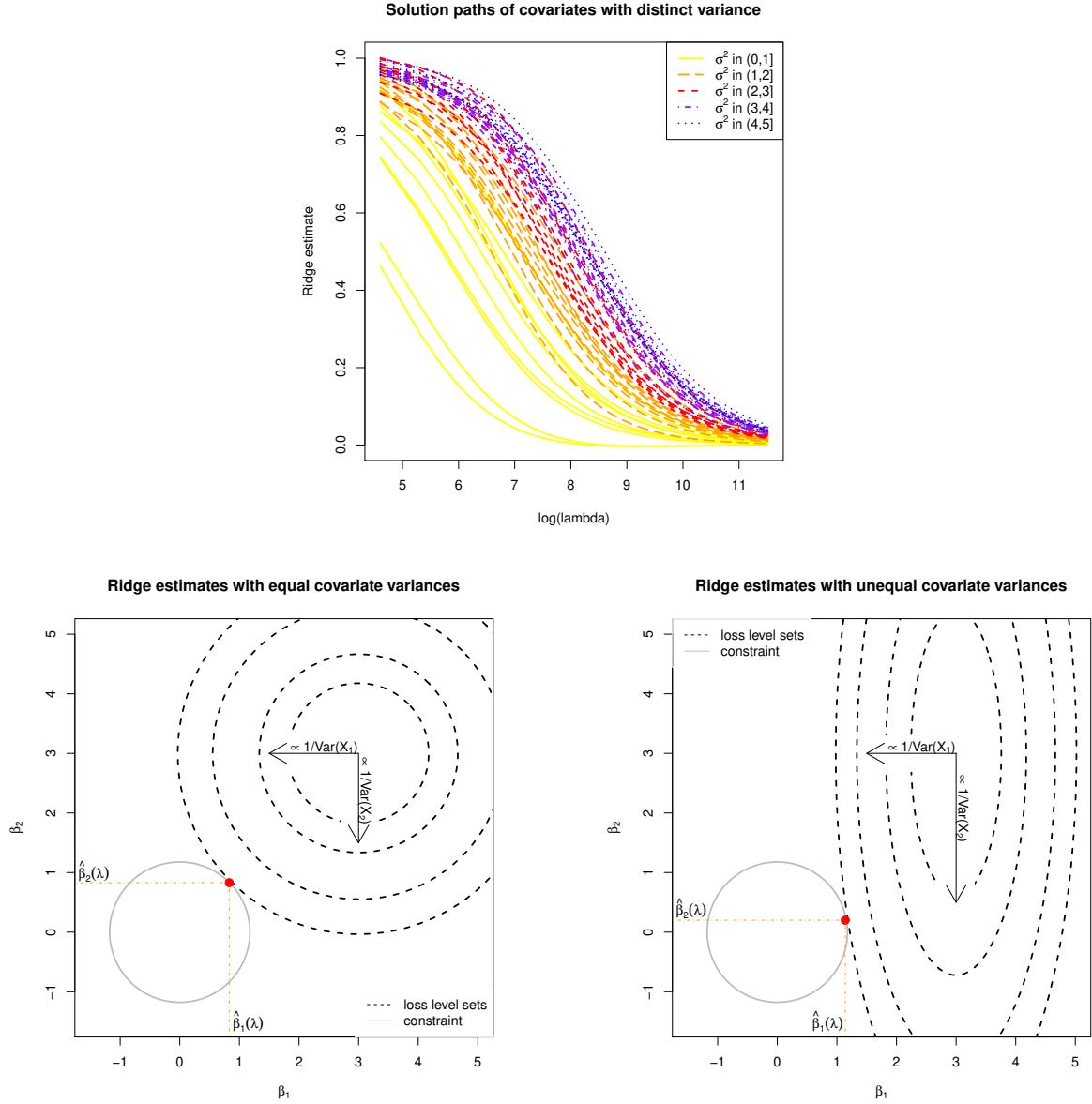


Figure 1.7: Top panel: Ridge regularization paths for coefficients of the 50 uncorrelated covariates with distinct variances. Color and line type indicated the grouping of the covariates by their variance. Bottom panels: Graphical illustration of the effect of a covariate's variance on the ridge regression estimator. The grey circle depicts the ridge parameter constraint. The dashed black ellipsoids are the level sets of the least squares loss function. The red dot is the ridge regression estimate. Left and right panels represent the cases with equal and unequal, respectively, variances of the covariates.

Ridge regression's preference of covariates with a large variance can intuitively be understood as follows. First note that the ridge regression estimator now can be written as:

$$\begin{aligned}\boldsymbol{\beta}(\lambda) &= [\text{Var}(\mathbf{X}) + \lambda \mathbf{I}_{50 \times 50}]^{-1} \text{Cov}(\mathbf{X}, \mathbf{Y}) \\ &= (\boldsymbol{\Sigma} + \lambda \mathbf{I}_{50 \times 50})^{-1} \boldsymbol{\Sigma} [\text{Var}(\mathbf{X})]^{-1} \text{Cov}(\mathbf{X}, \mathbf{Y}) \\ &= (\boldsymbol{\Sigma} + \lambda \mathbf{I}_{50 \times 50})^{-1} \boldsymbol{\Sigma} \boldsymbol{\beta}.\end{aligned}$$

Plug in the employed parametrization of $\boldsymbol{\Sigma}$, which gives: $[\boldsymbol{\beta}(\lambda)]_j = j(j + 50\lambda)^{-1} (\boldsymbol{\beta})_j$. Hence, the larger the

covariate's variance (corresponding to the larger j), the larger its ridge regression coefficient estimate. Ridge regression thus prefers, among a set of covariates with comparable effect sizes, those with larger variances.

The reformulation of ridge penalized estimation as a constrained estimation problem offers a geometrical interpretation of this phenomenon. Let $p = 2$ and the design matrix \mathbf{X} be orthogonal, while both covariates contribute equally to the response. Contrast the cases with $\text{Var}(X_1) \approx \text{Var}(X_2)$ and $\text{Var}(X_1) \gg \text{Var}(X_2)$. The level sets of the least squares loss function associated with the former case are circular, while that of the latter are strongly ellipsoidal (see Figure 1.7). The diameters along the principal axes (that – due to the orthogonality of \mathbf{X} – are parallel to that of the β_1 - and β_2 -axes) of both circle and ellipsoid are reciprocals of the variance of the covariates. When the variances of both covariates are equal, the level sets of the loss function expand equally fast along both axes. With the two covariates having the same regression coefficient, the point of these level sets closest to the parameter constraint is to be found on the line $\beta_1 = \beta_2$ (Figure 1.7, left panel). Consequently, the ridge regression estimator satisfies $\hat{\beta}_1(\lambda) \approx \hat{\beta}_2(\lambda)$. With unequal variances between the covariates, the ellipsoidal level sets of the loss function have diameters of rather different sizes. In particular, along the β_1 -axis it is narrow (as $\text{Var}(X_1)$ is large), and – vice versa – wide along the β_2 -axis. Consequently, the point of these level sets closest to the circular parameter constraint will be closer to the β_1 - than to the β_2 -axis (Figure 1.7, left panel). For the ridge estimator of the regression parameter this implies $0 \ll \hat{\beta}_1(\lambda) < 1$ and $0 < \hat{\beta}_2(\lambda) \ll 1$. Hence, the covariate with a larger variance yields the larger ridge regression estimator.

Should one thus standardize the covariates prior to ridge regression analysis? When dealing with gene expression data from microarrays, the data have been subjected to a series of pre-processing steps (e.g. quality control, background correction, within- and between-normalization). The purpose of these steps is to make the expression levels of genes comparable both within and between hybridizations. The preprocessing should thus be considered an inherent part of the measurement. As such, it is to be done independently of whatever down-stream analysis is to follow and further tinkering with the data is preferably to be avoided (as it may mess up the ‘comparable-ness’ of the expression levels as achieved by the preprocessing). For other data types different considerations may apply.

Among the considerations to decide on standardization of the covariates, one should also include the fact that ridge regression estimates prior and posterior to scaling do not simply differ by a factor. To see this assume that the covariates have been centered. Scaling of the covariates amounts to post-multiplication of the design matrix by a $p \times p$ -dimensional diagonal matrix \mathbf{A} with the reciprocals of the covariates' scale estimates on its diagonal (Sardy, 2008). Hence, the ridge regression estimator (for the rescaled data) is then given by:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\mathbf{A}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2.$$

Apply the change-of-variable $\boldsymbol{\gamma} = \mathbf{A}\boldsymbol{\beta}$ and obtain:

$$\min_{\boldsymbol{\gamma}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\gamma}\|_2^2 + \lambda\|\mathbf{A}^{-1}\boldsymbol{\gamma}\|_2^2 = \min_{\boldsymbol{\gamma}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\gamma}\|_2^2 + \sum_{j=1}^p \lambda[(\mathbf{A})_{jj}]^{-2}\gamma_j^2.$$

Effectively, the scaling is equivalent to covariate-wise penalization (see Chapter 3 for more on this). The ‘scaled’ ridge regression estimator may then be derived along the same lines as before in Section 1.5:

$$\hat{\boldsymbol{\beta}}^{(\text{scaled})}(\lambda) = \mathbf{A}^{-1}\hat{\boldsymbol{\gamma}}(\lambda) = \mathbf{A}^{-1}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{A}^{-2})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

In general, this is unequal to the ridge regression estimator without the rescaling of the columns of the design matrix. Moreover, it should be clear that $\hat{\boldsymbol{\beta}}^{(\text{scaled})}(\lambda) \neq \mathbf{A}\hat{\boldsymbol{\beta}}(\lambda)$.

1.9.2 Ridge regression and collinearity

Initially, ridge regression was motivated as an ad-hoc fix of (super)-collinear covariates in order to obtain a well-defined estimator. We now study the effect of this ad-hoc fix on the regression coefficient estimates of collinear covariates. In particular, their ridge regularization paths are contrasted to those of ‘non-collinear’ covariates.

To this end, we consider a simulation in which one response is regressed on 50 covariates. The data of these covariates, stored in a design matrix denoted \mathbf{X} , are sampled from a multivariate normal distribution, with mean zero and a 5×5 block covariance matrix:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0}_{10 \times 10} & \mathbf{0}_{10 \times 10} & \mathbf{0}_{10 \times 10} & \mathbf{0}_{10 \times 10} \\ \mathbf{0}_{10 \times 10} & \boldsymbol{\Sigma}_{22} & \mathbf{0}_{10 \times 10} & \mathbf{0}_{10 \times 10} & \mathbf{0}_{10 \times 10} \\ \mathbf{0}_{10 \times 10} & \mathbf{0}_{10 \times 10} & \boldsymbol{\Sigma}_{33} & \mathbf{0}_{10 \times 10} & \mathbf{0}_{10 \times 10} \\ \mathbf{0}_{10 \times 10} & \mathbf{0}_{10 \times 10} & \mathbf{0}_{10 \times 10} & \boldsymbol{\Sigma}_{44} & \mathbf{0}_{10 \times 10} \\ \mathbf{0}_{10 \times 10} & \mathbf{0}_{10 \times 10} & \mathbf{0}_{10 \times 10} & \mathbf{0}_{10 \times 10} & \boldsymbol{\Sigma}_{55} \end{pmatrix}$$

with

$$\Sigma_{kk} = \frac{1}{5}(k-1)\mathbf{1}_{10 \times 10} + \frac{1}{5}(6-k)\mathbf{I}_{10 \times 10}.$$

The data of the response variable \mathbf{Y} are then obtained through: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_{nn})$ and $\beta = \mathbf{1}_{50}$. Hence, all covariates contribute equally to the response. Would the columns of \mathbf{X} be orthogonal, little difference in the ridge estimates of the regression coefficients is expected.

The results of this simulation study with sample size $n = 1000$ are presented in Figure 1.8. All 50 regularization paths start close to one as λ is small and converge to zero as $\lambda \rightarrow \infty$. But the paths of covariates of the same block of the covariance matrix Σ quickly group, with those corresponding to a block with larger off-diagonal elements above those with smaller ones. Thus, ridge regression prefers (i.e. shrinks less) coefficient estimates of strongly positively correlated covariates.

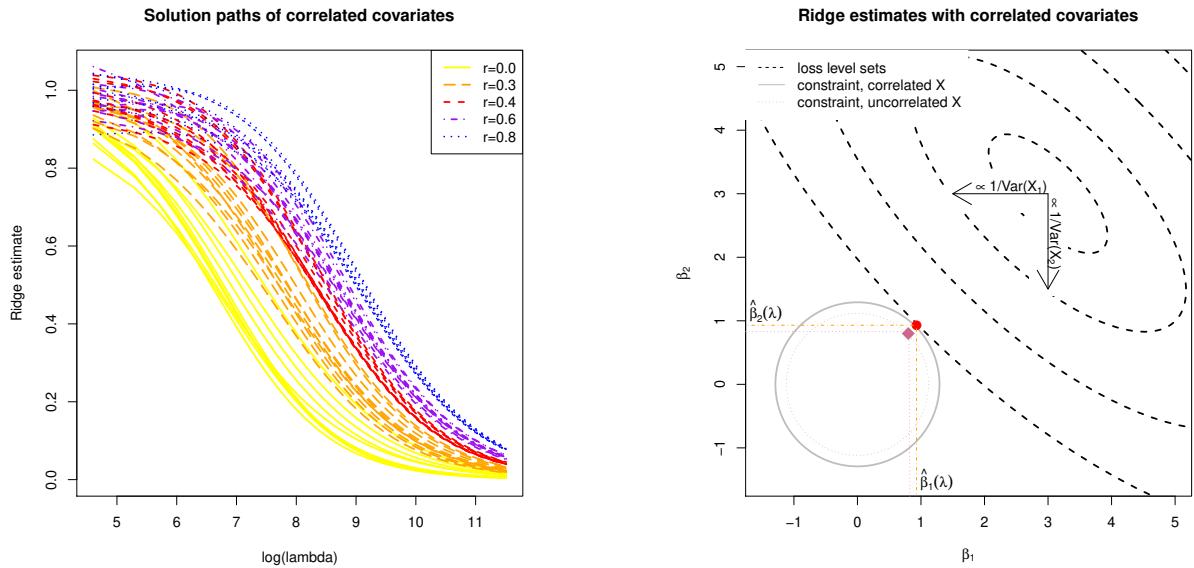


Figure 1.8: Left panel: Ridge regularization paths for coefficients of the 50 covariates, with various degree of collinearity but equal variance. Color and line type correspond to the five blocks of the covariate matrix Σ . Right panel: Graphical illustration of the effect of the collinearity among covariates on the ridge regression estimator. The solid and dotted grey circles depict the ridge parameter constraint for the collinear and orthogonal cases, respectively. The dashed black ellipsoids are the level sets of the sum-of-squares squares loss function. The red dot and violet diamond are the ridge regression for the positive collinear and orthogonal case, respectively.

Intuitive understanding of the observed behaviour may be obtained from the $p = 2$ case. Let U, V and ε be independent random variables with zero mean. Define $X_1 = U + V$, $X_2 = U - V$, and $Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ with β_1 and β_2 constants. Hence, $\mathbb{E}(Y) = 0$. Then:

$$Y = (\beta_1 + \beta_2)U + (\beta_1 - \beta_2)V + \varepsilon = \gamma_u U + \gamma_v V + \varepsilon$$

and $\text{Cor}(X_1, X_2) = [\text{Var}(U) - \text{Var}(V)]/[\text{Var}(U) + \text{Var}(V)]$. The random variables X_1 and X_2 are strongly positively correlated if $\text{Var}(U) \gg \text{Var}(V)$. The ridge regression estimator associated with regression of Y on U and V is:

$$\gamma(\lambda) = \begin{pmatrix} \text{Var}(U) + \lambda & 0 \\ 0 & \text{Var}(V) + \lambda \end{pmatrix}^{-1} \begin{pmatrix} \text{Cov}(U, Y) \\ \text{Cov}(V, Y) \end{pmatrix}.$$

For large enough λ

$$\gamma(\lambda) \approx \frac{1}{\lambda} \begin{pmatrix} \text{Var}(U) & 0 \\ 0 & \text{Var}(V) \end{pmatrix} \begin{pmatrix} \beta_1 + \beta_2 \\ \beta_1 - \beta_2 \end{pmatrix}.$$

If $\text{Var}(U) \gg \text{Var}(V)$ and $\beta_1 \approx \beta_2$, the ridge estimate of γ_v vanishes for large λ . Hence, ridge regression prefers positively covariates with similar effect sizes.

This phenomenon can also be explained geometrically. For the illustration consider ridge estimation with $\lambda = 1$ of the linear model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\beta = (3, 3)^\top$, $\varepsilon \sim \mathcal{N}(\mathbf{0}_2, \mathbf{I}_{22})$ and the columns of \mathbf{X} strongly and positively collinear. The level sets of the sum-of-squares loss, $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$, are plotted in the right panel of Figure 1.8. Recall that the ridge regression estimate is found by looking for the smallest loss level set that hits the ridge constraint. The sought-for estimate is then the point of intersection between this level set and the constraint, and – for the case at hand – is on the $x = y$ -line. This is no different from the case with orthogonal \mathbf{X} columns. Yet their estimates differ, even though the same λ is applied. The difference is due to the fact that the radius of the ridge constraint depends on λ , \mathbf{X} and \mathbf{Y} . This is immediate from the fact that the radius of the constraint equals $\|\hat{\beta}(\lambda)\|_2^2$ (see Section 1.5). To study the effect of \mathbf{X} on the radius, we remove its dependence on \mathbf{Y} by considering its expectation, which is:

$$\begin{aligned}\mathbb{E}[\|\hat{\beta}(\lambda)\|_2^2] &= \mathbb{E}\{[(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}(\mathbf{X}^\top \mathbf{X}) \hat{\beta}]^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}(\mathbf{X}^\top \mathbf{X}) \hat{\beta}\} \\ &= \mathbb{E}[\mathbf{Y}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-2} \mathbf{X}^\top \mathbf{Y}] \\ &= \sigma^2 \text{tr}\{\mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-2} \mathbf{X}^\top\} + \beta^\top \mathbf{X}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-2} \mathbf{X}^\top \mathbf{X} \beta.\end{aligned}$$

In the last step we have used $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}_{pp})$ and the expectation of the quadratic form of a multivariate random variable $\varepsilon \sim \mathcal{N}(\mu_\varepsilon, \Sigma_\varepsilon)$ is $\mathbb{E}[\varepsilon^\top \Lambda \varepsilon] = \text{tr}(\Lambda \Sigma_\varepsilon) + \mu_\varepsilon^\top \Lambda \mu_\varepsilon$ (cf. Mathai and Provost, 1992). The expression for the expectation of the radius of the ridge constraint can now be evaluated for the orthogonal \mathbf{X} and the strongly, positively collinear \mathbf{X} . It turns out that the latter is larger than the former. This results in a larger ridge constraint. For the larger ridge constraint there is a smaller level set that hits it first. The point of intersection, still on the $x = y$ -line, is now thus closer to β and further from the origin (cf. right panel of Figure 1.8). The resulting estimate is thus larger than that from the orthogonal case.

The above needs some attenuation. Among others it depends on: *i*) the number of covariates in each block, *ii*) the size of the effects, i.e. regression coefficients of each covariate, and *iii*) the degree of collinearity. Possibly, there are more factors influencing the behaviour of the ridge regression estimator presented in this subsection.

This behaviour of ridge regression is to be understood if a certain (say) clinical outcome is to be predicted from (say) gene expression data. Genes work in concert to fulfil a certain function in the cell. Consequently, one expects their expression levels to be correlated. Indeed, gene expression studies exhibit many co-expressed genes, that is, genes with correlating transcript levels. But also in most other applications with many explanatory variables collinearity will be omnipresent and similar issues are to be considered.

1.9.3 Variance inflation factor

The ridge regression estimator was introduced to resolve the undefinedness of its maximum likelihood counterpart in the face of (super)collinearity among the explanatory variables. The effect of collinearity on the uncertainty of estimates is often quantified by the Variance Inflation Factor (VIF). The VIF measures the change in the variance of the estimate due to the collinearity. Here we investigate how penalization affects the VIF. This requires a definition of the VIF of the ridge regression estimator.

The VIF of the maximum likelihood estimator of the j -th element of the regression parameter is defined as a factor in the following factorization of the variance of $\hat{\beta}_j$:

$$\begin{aligned}\text{Var}(\hat{\beta}_j) &= n^{-1} \sigma^2 [(\mathbf{X}^\top \mathbf{X})^{-1}]_{jj} \\ &\quad n^{-1} \sigma^2 [\text{Var}(X_{i,j} | X_{i,1}, \dots, X_{i,j-1}, X_{i,j+1}, \dots, X_{i,p})]^{-1} \\ &= \frac{n^{-1} \sigma^2}{\text{Var}(X_{i,j})} \cdot \frac{\text{Var}(X_{i,j})}{\text{Var}(X_{i,j} | X_{i,1}, \dots, X_{i,j-1}, X_{i,j+1}, \dots, X_{i,p})} \\ &:= n^{-1} \sigma^2 [\text{Var}(X_{i,j})]^{-1} \text{VIF}(\hat{\beta}_j),\end{aligned}$$

in which it assumed that the $X_{i,j}$'s are random and – using the column-wise zero ‘centered-ness’ of \mathbf{X} – that $n^{-1} \mathbf{X}^\top \mathbf{X}$ is estimator of their covariance matrix. Moreover, the identity used to arrive at the second line of the display, $[(\mathbf{X}^\top \mathbf{X})^{-1}]_{jj} = [\text{Var}(X_{i,j} | X_{i,1}, \dots, X_{i,j-1}, X_{i,j+1}, \dots, X_{i,p})]^{-1}$, originates from Corollary 5.8.1 of Whittaker (1990). Thus, $\text{Var}(\hat{\beta}_j)$ factorizes into $\sigma^2 [\text{Var}(X_{i,j})]^{-1}$ and the variance inflation factor $\text{VIF}(\hat{\beta}_j)$. When the j -th covariate is orthogonal to the other, i.e. there is no collinearity, then the VIF’s denominator, $\text{Var}(X_{i,j} | X_{i,1}, \dots, X_{i,j-1}, X_{i,j+1}, \dots, X_{i,p})$, equals $\text{Var}(X_{i,j})$. Consequently, $\text{VIF}(\hat{\beta}_j) = 1$. When there is collinearity among the covariates $\text{Var}(X_{i,j} | X_{i,1}, \dots, X_{i,j-1}, X_{i,j+1}, \dots, X_{i,p}) < \text{Var}(X_{i,j})$ and $\text{VIF}(\hat{\beta}_j) > 1$. The VIF then inflates the variance of the estimator of β_j under orthogonality – hence, the name – by a factor attributable to the collinearity.

The definition of the VIF needs modification to apply to the ridge regression estimator. In Marquardt (1970) the ‘ridge VIF’ is defined analogously to the above definition of the VIF of the maximum likelihood regression estimator as:

$$\begin{aligned}\text{Var}[\hat{\beta}_j(\lambda)] &= \sigma^2[(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}]_{jj} \\ &= \frac{n^{-1} \sigma^2}{\text{Var}(X_{i,j})} \cdot n \text{Var}(X_{i,j})[(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}]_{jj} \\ &:= n^{-1} \sigma^2 [\text{Var}(X_{i,j})]^{-1} \text{VIF}[\hat{\beta}_j(\lambda)],\end{aligned}$$

where the factorization is forced in line with that of the ‘maximum likelihood VIF’ but lacks a similar interpretation. When \mathbf{X} is orthogonal, $\text{VIF}[\hat{\beta}_j(\lambda)] = [\text{Var}(X_{i,j})]^2 [\text{Var}(X_{i,j}) + \lambda]^{-2} < 1$ for $\lambda > 0$. Penalization then deflates the VIF.

An alternative definition of the ‘ridge VIF’ presented by García *et al.* (2015) for the ‘ $p = 2$ ’-case, which they motivate from counterintuitive behaviour observed in the ‘ridge VIF’ defined by Marquardt (1970), adopts the

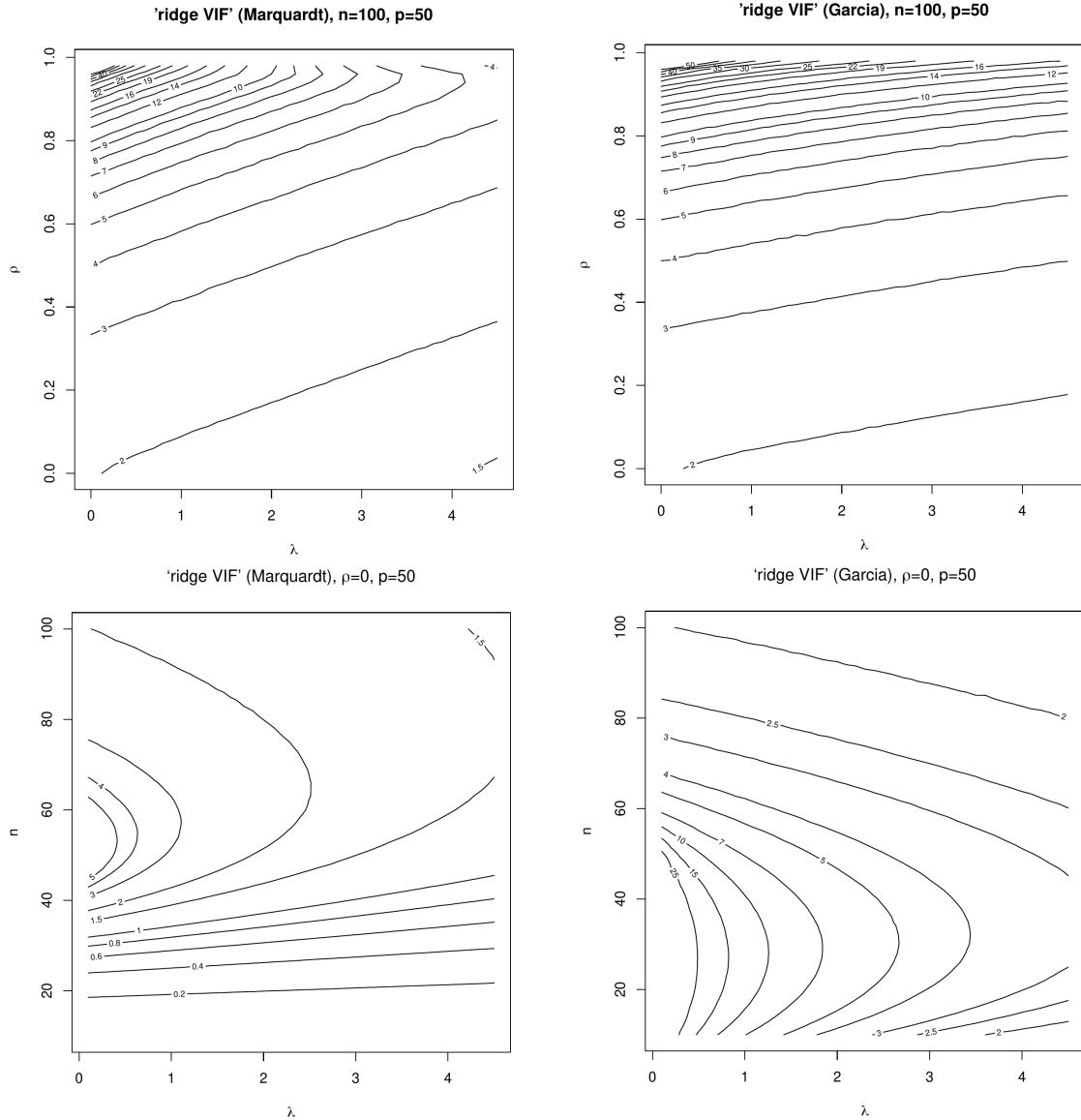


Figure 1.9: Contour plots of ‘Marquardt VIFs’ and ‘Garcia VIFs’, left and right columns, respectively. The top panels show these VIFs against the degree of penalization (x -axis) and collinearity (y -axis) for a fixed sample size ($n = 100$) and dimension ($p = 50$). The bottom panels show these VIFs against the degree of penalization (x -axis) and the sample size (y -axis) for a fixed dimension ($p = 50$).

‘maximum likelihood VIF’ definition but derives the ridge regression estimator from augmented data to comply with the maximum likelihood approach. This requires to augment the response vector with p zeros, i.e. $\mathbf{Y}_{\text{aug}} = (\mathbf{Y}^\top, \mathbf{0}_p^\top)^\top$ and the design matrix with p rows as $\mathbf{X}_{\text{aug}} = (\mathbf{X}^\top, \sqrt{\lambda}\mathbf{I}_{pp})^\top$. The ridge regression estimator can then be written as $\hat{\beta}(\lambda) = (\mathbf{X}_{\text{aug}}^\top \mathbf{X}_{\text{aug}})^{-1} \mathbf{X}_{\text{aug}}^\top \mathbf{Y}_{\text{aug}}$ (see Exercise 1.6). This reformulation of the ridge regression estimator in the form of its maximum likelihood counterpart suggests the adoption of latter’s VIF definition for the former. However, in the ‘maximum likelihood VIF’ the design matrix is assumed to be zero centered column-wise. Within the augmented data formulation this may be achieved by the inclusion of a column of ones in \mathbf{X}_{aug} representing the intercept. The inclusion of an intercept, however, requires a modification of the estimators of $\text{Var}(X_{i,j})$ and $\text{Var}(X_{i,j} | X_{i,1}, \dots, X_{i,j-1}, X_{i,j+1}, \dots, X_{i,p})$. The former is readily obtained, while the latter is given by the reciprocals of the 2nd to the $(p+1)$ -th diagonal elements of the inverse of:

$$\begin{pmatrix} \mathbf{1}_n & \mathbf{X} \\ \mathbf{1}_p & \sqrt{\lambda}\mathbf{I}_{pp} \end{pmatrix}^\top \begin{pmatrix} \mathbf{1}_n & \mathbf{X} \\ \mathbf{1}_p & \sqrt{\lambda}\mathbf{I}_{pp} \end{pmatrix} = \begin{pmatrix} n+p & \sqrt{\lambda}\mathbf{1}_p^\top \\ \sqrt{\lambda}\mathbf{1}_p & \mathbf{X}^\top \mathbf{X} + \lambda\mathbf{I}_{pp} \end{pmatrix}.$$

The lower right block of this inverse is obtained using well-known linear algebra results *i*) the analytic expression of the inverse of a 2×2 -partitioned block matrix and *ii*) the inverse of the sum of an invertible and a rank one matrix (given by the Sherman-Morrison formula, see Corollary 18.2.10, Harville, 2008). It equals:

$$(\mathbf{X}^\top \mathbf{X} + \lambda\mathbf{I}_{pp})^{-1} - (1+c)^{-1}(n+p)^{-1}\lambda(\mathbf{X}^\top \mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{1}_p\mathbf{1}_p^\top(\mathbf{X}^\top \mathbf{X} + \lambda\mathbf{I}_{pp})^{-1},$$

where $c = (n+p)^{-1}\lambda\mathbf{1}_p^\top(\mathbf{X}^\top \mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{1}_p = (n+p)^{-1}\lambda\sum_{j,j'=1}^p[(\mathbf{X}^\top \mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}]_{j,j'}$. Substitution of these expressions in the ratio of $\text{Var}(X_{i,j})$ and $\text{Var}(X_{i,j} | X_{i,1}, \dots, X_{i,j-1}, X_{i,j+1}, \dots, X_{i,p})$ in the above yields the alternative VIF of García *et al.* (2015).

The effect of collinearity on the variance of the ridge regression estimator and the influence of penalization on this effect is studied in two small simulation studies. In the first study the rows of the design matrix $\mathbf{X}_{i,*}^\top$ are sampled from $\mathcal{N}(\mathbf{0}_p, \Sigma)$ with $\Sigma = (1-\rho)\mathbf{I}_{pp} + \rho\mathbf{1}_p\mathbf{1}_p^\top$ fixing the dimension at $p = 50$ and the sample size at $n = 100$. The correlation coefficient ρ is varied over the unit interval $[0, 1]$, representing various levels of collinearity. The columns of the thus drawn \mathbf{X} are zero centered. The response is then formed in accordance with the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\beta = \mathbf{1}_p$ and the error drawn from $\mathcal{N}(\mathbf{0}_n, \frac{1}{4}\mathbf{I}_{nn})$. The effect of penalization is studied by varying λ . For each (ρ, λ) -combination data, \mathbf{X} and \mathbf{Y} , are sampled thousand times and both ‘ridge VIFs’, for discriminative purposes referred as the Marquardt and Garcia VIFs (after the first author of the proposing papers), are calculated and averaged. Figure 1.9 shows the contour plots of the averaged Marquardt and Garcia VIF against ρ and λ . At $(\rho, \lambda) \approx (0, 0)$ both VIFs are close to two, which – although the set-up is not high-dimensional in the strict ‘ $p > n$ ’-sense – is due to the strenuous p/n -ratio. Nonetheless, as expected an increase in collinearity results in an increase in the VIF (irrespective of the type). Moreover, indeed some counterintuitive behaviour is observed in the Marquardt VIF: at (say) a value of $\lambda = 3$ the VIF reaches a maximum for $\rho \approx 0.95$ and declines for larger degrees of collinearity. This is left without further interpretation as interest is not in deciding on the most appropriate operationalization among both VIFs of the ridge regression estimator, but is primarily in the effect of penalization on the VIF. In this respect both VIFs exhibit the same behaviour: a monotone decrease of both VIFs in λ .

In the second simulation the effect of ‘spurious collinearity’ introduced by the varying the sample size on the ridge regression estimator is studied. The settings are identical to that of the first simulation but with $\rho = 0$ and a sample size n that varies from ten to hundred. With $p = 50$ in particular the lower sample sizes will exhibit high degrees of collinearity as the sample correlation coefficient has been seen to inflate then. The resulting contour plots (Figure 1.9) now present the VIFs against n and λ . Although some counterintuitive behaviour can be seen around $n = p$ and small λ , the point to be noted and relevant here – as interest is in the VIF’s behaviour for data sets with a fixed sample size and covariates drawn without collinearity – is the monotone decrease of both VIFs in λ at any sample size.

In summary, the penalization does not remove collinearity but, irrespective of the choice of VIF, it reduces the effect of collinearity on the variance of the ridge regression estimator (as measured by the VIFs above). This led García *et al.* (2015) – although admittedly their focus appears to be low-dimensionally – to suggest that the VIFs may guide the choice the penalty parameter: choose λ such that the variance of the estimator is increased at most by a user-specified factor.

1.10 Illustration

The application of ridge regression to actual data aims to illustrate its use in practice.

1.10.1 MCM7 expression regulation by microRNAs

Recently, a new class of RNA was discovered, referred to as microRNA. MicroRNAs are non-coding, single stranded RNAs of approximately 22 nucleotides. Like mRNAs, microRNAs are encoded in and transcribed from the DNA. MicroRNAs play an important role in the regulatory mechanism of the cell. MicroRNAs down-regulate gene expression by either of two post-transcriptional mechanisms: mRNA cleavage or transcriptional repression. This depends on the degree of complementarity between the microRNA and the target. Perfect or nearly perfect complementarity of the mRNA to the microRNA will lead to cleavage and degradation of the target mRNA. Imperfect complementarity will repress the productive translation and reduction in protein levels without affecting the mRNA levels. A single microRNA can bind to and regulate many different mRNA targets. Conversely, several microRNAs can bind to and cooperatively control a single mRNA target (Bartel, 2004; Esquela-Kerscher and Slack, 2006; Kim and Nam, 2006).

In this illustration we wish to confirm the regulation of mRNA expression by microRNAs in an independent data set. We cherry pick an arbitrary finding from literature reported in Ambs *et al.* (2008), which focusses on the microRNA regulation of the MCM7 gene in prostate cancer. The MCM7 gene is involved in DNA replication (Tye, 1999), a cellular process often derailed in cancer. Furthermore, MCM7 interacts with the tumor-suppressor gene RB1 (Sterner *et al.*, 1998). Several studies indeed confirm the involvement of MCM7 in prostate cancer (Padmanabhan *et al.*, 2004). And recently, it has been reported that in prostate cancer MCM7 may be regulated by microRNAs (Ambs *et al.*, 2008).

We here assess whether the MCM7 down-regulation by microRNAs can be observed in a data set other than the one upon which the microRNA-regulation of MCM7 claim has been based. To this end we download from the Gene Expression Omnibus (GEO) a prostate cancer data set (presented by Wang *et al.*, 2009). This data set (with GEO identifier: GSE20161) has both mRNA and microRNA profiles for all samples available. The preprocessed (as detailed in Wang *et al.*, 2009) data are downloaded and require only minor further manipulations to suit our purpose. These manipulations comprise *i*) averaging of duplicated profiles of several samples, *ii*) gene- and mir-wise zero-centering of the expression data, *iii*) averaging the expression levels of the probes that interrogate MCM7. Eventually, this leaves 90 profiles each comprising of 735 microRNA expression measurements.

Listing 1.3 R code

```
# load libraries
library(GEOquery)
library(RmiR.hsa)
library(penalized)

# extract data
slh      <- getGEO("GSE20161", GSEMatrix=TRUE)
GEData   <- slh[1][[1]]
MIRdata <- slh[2][[1]]

# average duplicate profiles
Yge     <- numeric()
Xmir   <- numeric()
for (sName in 1:90){
  Yge   <- cbind(Yge, apply(exprs(GEData)[,sName,drop=FALSE], 1, mean))
  Xmir <- cbind(Xmir, apply(exprs(MIRdata)[,sName,drop=FALSE], 1, mean))
}
colnames(Yge) <- paste("S", 1:90, sep="")
colnames(Xmir) <- paste("S", 1:90, sep="")

# extact mRNA expression of the MCM7N tumor suppressor gene
entrezID <- c("4176")
geneName <- "MCM7"
Y         <- Yge[which(fData(GEData)[,10] == entrezID),]

# average gene expression levels over probes
Y <- apply(Y, 2, mean)

# mir-wise centering mir expression data
X <- t(sweep(Xmir, 1, rowMeans(Xmir)))
```

```
# generate cross-validated likelihood profile
profL2(Y, penalized=X, minlambda2=1, maxlambda2=20000, plot=TRUE)

# decide on the optimal penalty value directly
optLambda <- optL2(Y, penalized=X)$lambda

# obtain the ridge regression estimates
ridgeFit <- penalized(Y, penalized=X, lambda2=optLambda)

# plot them as histogram
hist(coef(ridgeFit, "penalized"), n=50, col="blue", border="lightblue",
      xlab="ridge_regression_estimates_with_optimal_lambda",
      main="Histogram_of_ridge_estimates")

# linear prediction from ridge
Yhat <- predict(ridgeFit, X)[,1]
plot(Y ~ Yhat, pch=20, xlab="pred._MCM7_expression",
      ylab="obs._MCM7_expression")
```

With this prostate data set at hand we now investigate whether MCM7 is regulated by microRNAs. Hereto we fit a linear regression model regressing the expression levels of MCM7 onto those of the microRNAs. As the number of microRNAs exceeds the number of samples, ordinary least squares fails and we resort to the ridge estimator of the regression coefficients. First, an informed choice of the penalty parameter is made through maximization of the LOOCV log-likelihood, resulting in $\lambda_{\text{opt}} = 1812.826$. Having decided on the value of the to-be-employed penalty parameter, the ridge regression estimator can now readily be evaluated. The thus fitted model allows for the evaluation of microRNA-regulation of MCM7. E.g., by the proportion of variation of the MCM7 expression levels by the microRNAs as expressed in coefficient of determination: $R^2 = 0.4492$. Alternatively, but closely related, observed expression levels may be related to the linear predictor of the MCM7 expression levels: $\hat{\mathbf{Y}}(\lambda_{\text{opt}}) = \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda_{\text{opt}})$. The Spearman correlation of response and predictor equals 0.6295. A visual inspection is provided by the left panel of Figure 1.10. Note the difference in scale of the x - and y -axes. This is due to the fact that the regression coefficients have been estimated in penalized fashion, consequently shrinking estimates of the regression coefficients towards zero leading to small estimates and in turn compressing the range of the linear prediction. The above suggests there is indeed association between the microRNA expression levels and those of MCM7.

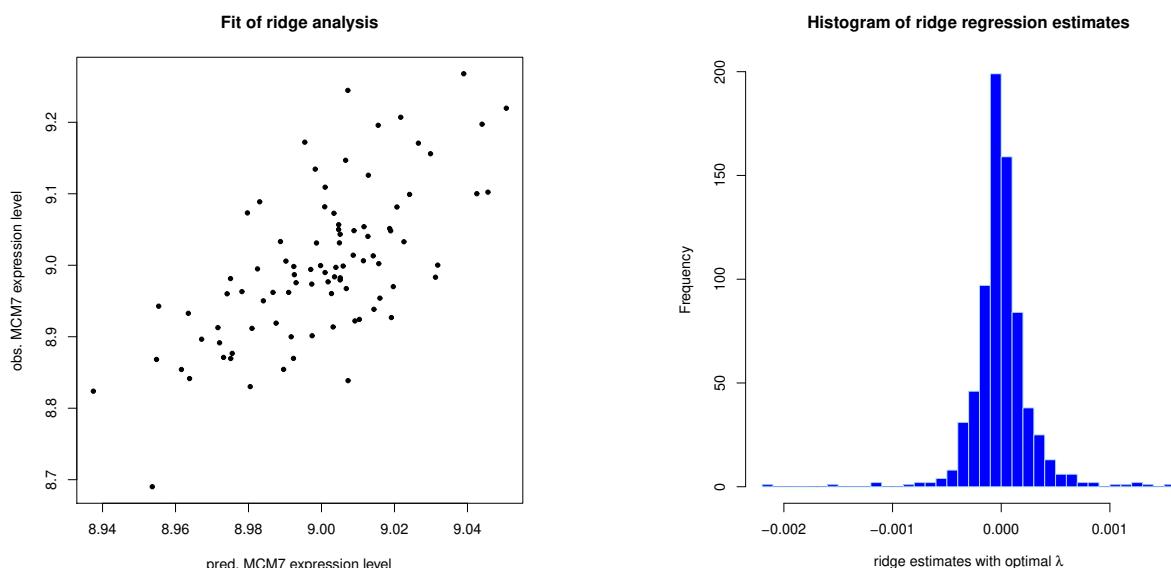


Figure 1.10: Left panel: Observed vs. (ridge) fitted MCM7 expression values. Right panel: Histogram of the ridge regression coefficient estimates.

The overall aim of this illustration was to assess whether microRNA-regulation of MCM7 could also be ob-

served in this prostate cancer data set. In this endeavour the dogma (stating this regulation should be negative) has nowhere been used. A first simple assessment of the validity of this dogma studies the signs of the estimated regression coefficients. The ridge regression estimate has 394 out of the 735 microRNA probes with a negative coefficient. Hence, a small majority has a sign in line with the ‘microRNA \downarrow mRNA’ dogma. When, in addition, taking the size of these coefficients into account (Figure 1.10, right panel), the negative regression coefficient estimates do not substantially differ from their positive counterparts (as can be witnessed from their almost symmetrical distribution around zero). Hence, the value of the ‘microRNA \downarrow mRNA’ dogma is not confirmed by this ridge regression analysis of the MCM7-regulation by microRNAs. Nor is it refuted.

The implementation of ridge regression in the penalized-package offers the possibility to fully obey the dogma on negative regulation of mRNA expression by microRNAs. This requires all regression coefficients to be negative. Incorporation of the requirement into the ridge estimation augments the constrained estimation problem with an additional constraint:

$$\hat{\beta}(\lambda) = \arg \min_{\|\beta\|_2^2 \leq c(\lambda); \beta_j \leq 0 \text{ for all } j} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2.$$

With the additional non-positivity constraint on the parameters, there is no explicit solution for the estimator. The ridge estimate of the regression parameters is then found by numerical optimization using e.g. the Newton-Raphson algorithm or a gradient descent approach. The next listing gives the R-code for ridge estimation with the non-positivity constraint of the linear regression model.

Listing 1.4 R code

```
# decide on the optimal penalty value with sign constraint on parameters
optLambda <- optL2(Y, penalized=-X, positive=rep(TRUE, ncol(X))) $lambda

# obtain the ridge regression estimates
ridgeFit <- penalized(Y, penalized=-X, lambda2=optLambda,
                        positive=rep(TRUE, ncol(X)))

# linear prediction from ridge
Yhat <- predict(ridgeFit, -X) [,1]
plot(Y ~ Yhat, pch=20, xlab="predicted_MCM7_expression_level",
      ylab="observed_MCM7_expression_level")
cor(Y, Yhat, m="s")
summary(lm(Y ~ Yhat)) [8]
```

The linear regression model linking MCM7 expression to that of the microRNAs is fitted by ridge regression while simultaneously obeying the ‘negative regulation of mRNA by microRNA’-dogma to the prostate cancer data. In the resulting model 401 out of 735 microRNA probes have a nonzero (and negative) coefficient. There is a large overlap in microRNAs with a negative coefficient between those from this and the previous fit. The models are also compared in terms of their fit to the data. The Spearman rank correlation coefficient between response and predictor for the model without positive regression coefficients equals 0.679 and its coefficient of determination 0.524 (confer the left panel of 1.11 for a visualization). This is a slight improvement upon the unconstrained ridge estimated model. The improvement may be small but it should be kept in mind that the number of parameters used by both models is 401 (for the model without positive regression coefficients) vs. 735. Hence, with close to half the number of parameters the dogma-obeying model gives a somewhat better description of the data. This may suggest that there is some value in the dogma as inclusion of this prior information leads to a more parsimonious model without any loss in fit.

The dogma-obeying model selects 401 microRNAs that aid in the explanation of the variation in the gene expression levels of MCM7. There is an active field of research, called *target prediction*, trying to identify which microRNAs target the mRNA of which genes. Within R there is a collection of packages that provide the target prediction of known microRNAs. The packages differ on the method (e.g. experimental or sequence comparison) that has been used to arrive at the prediction. These target predictions may be used to evaluate the value of the found 401 microRNAs. Ideally, there would be a substantial amount of overlap. The R-script that loads the target predictions and does the comparison is below.

Listing 1.5 R code

```
# extract mir names and their (hypothesized) mrna target
mir2target <- numeric()
mirPredProgram <- c("targetscan", "miranda", "mirbase", "pictar", "mirtarget2")
```

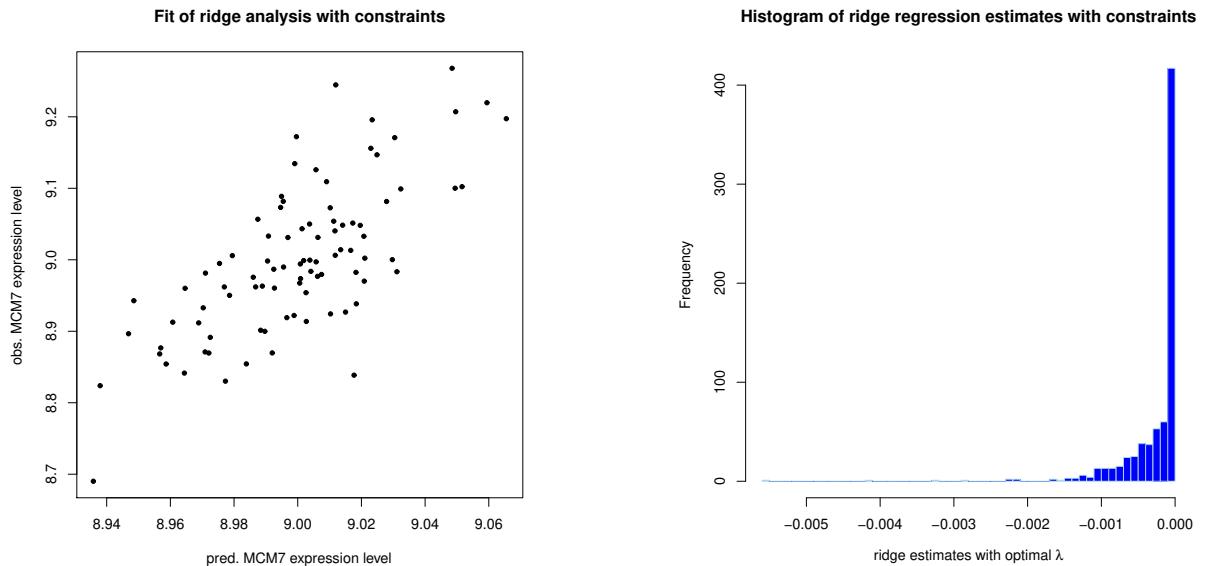


Figure 1.11: Left panel: Observed vs. (ridge) fitted MCM7 expression values (with the non-positive constraint on the parameters in place). Right panel: Histogram of the ridge regression coefficient estimates (from the non-negativity constrained analysis).

```

for (program in mirPredProgram) {
  slh      <- dbReadTable(RmiR.hsa_dbconn(), program)
  slh      <- cbind(program, slh[,1:2])
  colnames(slh) <- c("method", "mir", "target")
  mir2target    <- rbind(mir2target, slh)
}

mir2target <- unique(mir2target)
mir2target <- mir2target [which(mir2target [,3] == entrezID),]
uniqMirs   <- tolower(unique(mir2target [,2])) 

# extract names of mir-probe on array
arrayMirs <- tolower(levels(fData(MIRdata) [,3]) [fData(MIRdata) [,3]]) 

# which mir-probes are predicted to down-regulate MCM7
selMirs <- intersect(arrayMirs, uniqMirs)
ids     <- which(arrayMirs %in% selMirs)

# which ridge estimates are non-zero
nonzeroBetas <- (coef(ridgeFit, "penalized") != 0)

# which mirs are predicted to
nonzeroPred    <- 0 * betas
nonzeroPred[ids] <- 1

# contingency table and chi-square test
table(nonzeroBetas, nonzeroPred)
chisq.test(table(nonzeroBetas, nonzeroPred))

```

With knowledge available on each microRNA whether it is predicted (by at least one target prediction package) to be a potential target of MCM7, it may be cross-tabulated against its corresponding regression coefficient estimate in the dogma-obeying model being equal to zero or not. Table 1.1 contains the result. Somewhat superfluous considering the data, we may test whether the targets of MCM7 are overrepresented in the group of strictly negatively estimated regression coefficients. The corresponding chi-squared test (with Yates' continuity correction) yields the test statistic $\chi^2 = 0.0478$ with a p -value equal to 0.827. Hence, there is no enrichment among the 401 microRNAs of those that have been predicted to target MCM7. This may seem worrisome. However, the

	$\hat{\beta}_j = 0$	$\hat{\beta}_j < 0$
microRNA not target	323	390
microRNA target	11	11

Table 1.1: Cross-tabulation of the microRNAs being a potential target of MCM7 vs. the value of its regression coefficient in the dogma-obeying model.

microRNAs have been selected for their predictive power of the expression levels of MCM7. Variable selection has not been a criterion (although the sign constraint implies selection). Moreover, criticism on the value of the microRNA target prediction has been accumulating in recent years (REF).

1.11 Conclusion

We discussed ridge regression as a modification of linear regression to overcome the empirical non-identifiability of the latter when confronted with high-dimensional data. The means to this end was the addition of a (ridge) penalty to the sum-of-squares loss function of the linear regression model, which turned out to be equivalent to constraining the parameter domain. This warranted the identification of the regression coefficients, but came at the cost of introducing bias in the estimates. Several properties of ridge regression like moments and its MSE have been reviewed. Finally, its behaviour and use have been illustrated in simulation and omics data.

1.12 Exercises

Question 1.1

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma_\varepsilon^2 \mathbf{I}_{nn})$. This model (without intercept) is fitted to data using the ridge regression estimator $\hat{\boldsymbol{\beta}}(\lambda) = \arg \min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2$ with $\lambda > 0$. The data are:

$$\mathbf{X}^\top = \begin{pmatrix} -1 & 1 & 1 & -1 \end{pmatrix} \text{ and } \mathbf{Y}^\top = \begin{pmatrix} -1.5 & 2.9 & -3.5 & 0.7 \end{pmatrix}.$$

- a) Evaluate the maximum likelihood/ordinary least squares estimator of the regression parameter, i.e. $\hat{\boldsymbol{\beta}}(\lambda)$ for $\lambda = 0$.
- b) Evaluate the ridge regression estimator for $\lambda = 1$.
- c) Verify that the ridge regression estimator $\hat{\boldsymbol{\beta}}(\lambda)$ shrinks as λ increases. Hereto combine your answer to parts a) and b) and the evaluation of the ridge regression estimator for $\lambda = 10$ and $\lambda = 1000$. Is the order of the employed choices of λ and the absolute value of the corresponding estimates concordant?

Question 1.2 [†]

Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The data on the covariate and response are: $\mathbf{X}^\top = (X_1, X_2, \dots, X_8)^\top = (-2, -1, -1, -1, 0, 1, 2, 2)^\top$ and $\mathbf{Y}^\top = (Y_1, Y_2, \dots, Y_8)^\top = (35, 40, 36, 38, 40, 43, 45, 43)^\top$, with corresponding elements in the same order.

- a) Find the ridge regression estimator for the data above for a general value of λ .
- b) Evaluate the fit, i.e. $\hat{Y}_i(\lambda)$ for $\lambda = 10$. Would you judge the fit as good? If not, what is the most striking feature that you find unsatisfactory?
- c) Now zero center the covariate and response data, denote it by \tilde{X}_i and \tilde{Y}_i , and evaluate the ridge regression estimator of $\tilde{Y}_i = \beta_1 \tilde{X}_i + \varepsilon_i$ at $\lambda = 4$. Verify that in terms of original data the resulting predictor now is: $\hat{Y}_i(\lambda) = 40 + 1.75 X_i$.

Note that the employed estimate in the predictor found in part c) is effectively a combination of a maximum likelihood and ridge regression one for intercept and slope, respectively. Put differently, only the slope has been shrunken.

Question 1.3

Consider the simple linear regression model $Y_i = \beta_0 + X_i \boldsymbol{\beta} + \varepsilon_i$ for $i = 1, \dots, n$ and with $\varepsilon_i \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$. The model comprises a single covariate and an intercept. Response and covariate data are: $\{(y_i, x_i)\}_{i=1}^4 =$

[†]This exercise is inspired by one from Draper and Smith (1998)

$\{(1.4, 0.0), (1.4, -2.0), (0.8, 0.0), (0.4, 2.0)\}$. Find the value of λ that yields the ridge regression estimate (with an unregularized/unpenalized intercept as is done in part c) of Question 1.2) equal to $(1, -\frac{1}{8})^\top$.

Question 1.4

Plot the regularization path of the ridge regression estimator over the range $\lambda \in (0, 20.000]$ using the data of Example 1.2.

Question 1.5

Consider the ridge regression estimator $\hat{\beta}(\lambda) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y}$. Show that, for λ large enough, $\text{sign}[\hat{\beta}(\lambda)] = \text{sign}(\mathbf{X}^\top \mathbf{Y})$. Hint: If \mathbf{A} is a nonsingular matrix and the largest (in an absolute sense) singular value of $\mathbf{B}\mathbf{A}^{-1}$ is smaller than one, then $(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1} \sum_{k=1}^{\infty} (-1)^k (\mathbf{B}\mathbf{A}^{-1})^k$.

Question 1.6[‡]

Show that the ridge regression estimator can be obtained by ordinary least squares regression on an augmented data set. Hereto augment the matrix \mathbf{X} with p additional rows $\sqrt{\lambda} \mathbf{I}_{pp}$, and augment the response vector \mathbf{Y} with p zeros.

Question 1.7

Recall the definitions of $\hat{\beta}_{\text{MLS}}$, \mathbf{W}_λ and $\hat{\beta}(\lambda)$ from Section 1.2. Show that, unlike the linear relations between the ridge and maximum likelihood estimators if $\mathbf{X}^\top \mathbf{X}$ is of full rank, $\hat{\beta}(\lambda) \neq \mathbf{W}_\lambda \hat{\beta}_{\text{MLS}}$ high-dimensionally.

Question 1.8

The coefficients β of a linear regression model, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, are estimated by $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$. The associated fitted values then given by $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \mathbf{H}\mathbf{Y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ referred to as the hat matrix. The hat matrix \mathbf{H} is a projection matrix as it satisfies $\mathbf{H} = \mathbf{H}^2$. Hence, linear regression projects the response \mathbf{Y} onto the vector space spanned by the columns of \mathbf{Y} . Consequently, the residuals $\hat{\varepsilon}$ and $\hat{\mathbf{Y}}$ are orthogonal. Now consider the ridge estimator of the regression coefficients: $\hat{\beta}(\lambda) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y}$. Let $\hat{\mathbf{Y}}(\lambda) = \mathbf{X}\hat{\beta}(\lambda)$ be the vector of associated fitted values.

- a) Show that the ridge hat matrix $\mathbf{H}(\lambda) = \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top$, associated with ridge regression, is not a projection matrix (for any $\lambda > 0$), i.e. $\mathbf{H}(\lambda) \neq [\mathbf{H}(\lambda)]^2$.
- b) Show that for any $\lambda > 0$ the ‘ridge fit’ $\hat{\mathbf{Y}}(\lambda)$ is not orthogonal to the associated ‘ridge residuals’ $\hat{\varepsilon}(\lambda)$, defined as $\varepsilon(\lambda) = \mathbf{Y} - \mathbf{X}\hat{\beta}(\lambda)$.

Question 1.9

Consider the standard linear regression model $Y_i = \mathbf{X}_{i,*}\beta + \varepsilon_i$ for $i = 1, \dots, n$ and with the $\varepsilon_i \sim_{i.i.d} \mathcal{N}(0, \sigma^2)$. Suppose the parameter β is estimated by the ridge regression estimator $\hat{\beta}(\lambda) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y}$.

- a) The vector of ‘ridge residuals’, defined as $\varepsilon(\lambda) = \mathbf{Y} - \mathbf{X}\hat{\beta}(\lambda)$, are normally distributed. Why?
- b) Show that $\mathbb{E}[\varepsilon(\lambda)] = [\mathbf{I}_{nn} - \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top] \mathbf{X}\beta$.
- c) Show that $\text{Var}[\varepsilon(\lambda)] = \sigma^2 [\mathbf{I}_{nn} - \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top]^2$.
- d) Could the normal probability plot, i.e. a qq-plot with the quantiles of standard normal distribution plotted against those of the ridge residuals, be used to assess the normality of the latter? Motivate.

Question 1.10

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$. This model is fitted to data, $\mathbf{X}_{1,*} = (4, -2)$ and $Y_1 = (10)$, using the ridge regression estimator $\hat{\beta}(\lambda) = (\mathbf{X}_{1,*}^\top \mathbf{X}_{1,*} + \lambda \mathbf{I}_{22})^{-1} \mathbf{X}_{1,*}^\top Y_1$. Throughout use $\lambda = 5$

- a) Evaluate the ridge regression estimator.
- b) Suppose $\beta = (1, -1)^\top$. Evaluate the bias of the ridge regression estimator.
- c) Decompose the bias into a component due to the regularization and one attributable to the high-dimensionality of the study.
- d) Would β have equalled $(2, -1)^\top$, the bias’ component due to the high-dimensionality vanishes. Explain why.

[‡]This exercise is freely rendered from Hastie *et al.* (2009), but can be found in many other places. The original source is unknown to the author.

Question 1.11 (Numerical inaccuracy)

The linear regression model, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$, is fitted by to the data with the following response, design matrix, and relevant summary statistics:

$$\mathbf{X} = \begin{pmatrix} 0.3 & -0.7 \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} 0.2 \end{pmatrix}, \mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 0.09 & -0.21 \\ -0.21 & 0.49 \end{pmatrix}, \text{ and } \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} 0.06 \\ -0.14 \end{pmatrix}.$$

Hence, $p = 2$ and $n = 1$. The fitting uses the ridge regression estimator.

- a) Section 1.4.1 states that the regularization path of the ridge regression estimator, i.e. $\{\hat{\beta}(\lambda) : \lambda > 0\}$, is confined to a line in \mathbb{R}^2 . Give the details of this line and draw it in the (β_1, β_2) -plane.
- b) Verify numerically, for a set of penalty parameter values, whether the corresponding estimates $\hat{\beta}(\lambda)$ are indeed confined to the line found in part a). Do this by plotting the estimates in the (β_1, β_2) -plane (along with the line found in part a). In this use the following set of λ 's:

Listing 1.6 R code

```
lambdas <- exp(seq(log(10^(-15)), log(1), length.out=100))
```

- c) Part b) reveals that, for small values of λ , the estimates fall outside the line found in part a). Using the theory outlined in Section 1.4.1, the estimates can be decomposed into a part that falls on this line and a part that is orthogonal to it. The latter is given by $(\mathbf{I}_{22} - \mathbf{P}_x)\hat{\beta}(\lambda)$ where \mathbf{P}_x is the projection matrix onto the space spanned by the columns of \mathbf{X} . Evaluate the projection matrix \mathbf{P}_x .
- d) Numerical inaccuracy, resulting from the ill-conditionedness of $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{22}$, causes $(\mathbf{I}_{22} - \mathbf{P}_x)\hat{\beta}(\lambda) \neq \mathbf{0}_2$. Verify that the observed non-null $(\mathbf{I}_{22} - \mathbf{P}_x)\hat{\beta}(\lambda)$ are indeed due to numerical inaccuracy. Hereto generate a log-log plot of the condition number of $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{22}$ vs. the $\|(\mathbf{I}_{22} - \mathbf{P}_x)\hat{\beta}(\lambda)\|_2$ for the provided set of λ 's.

Question 1.12

Provide an alternative proof of Theorem 1.2 that states the existence of a positive value of the penalty parameter for which the ridge regression estimator has a superior MSE compared to that of its maximum likelihood counterpart.

- a) Show that the derivative of the MSE with respect to the penalty parameter is negative at zero. In this use the following results from matrix calculus:

$$\frac{d}{d\lambda} \text{tr}[\mathbf{A}(\lambda)] = \text{tr}\left[\frac{d}{d\lambda} \mathbf{A}(\lambda)\right], \quad \frac{d}{d\lambda} (\mathbf{A} + \lambda \mathbf{B})^{-1} = -(\mathbf{A} + \lambda \mathbf{B})^{-1} \mathbf{B} (\mathbf{A} + \lambda \mathbf{B})^{-1},$$

and the chain rule

$$\frac{d}{d\lambda} \mathbf{A}(\lambda) \mathbf{B}(\lambda) = \left[\frac{d}{d\lambda} \mathbf{A}(\lambda) \right] \mathbf{B}(\lambda) + \mathbf{A}(\lambda) \left[\frac{d}{d\lambda} \mathbf{B}(\lambda) \right],$$

where $\mathbf{A}(\lambda)$ and $\mathbf{B}(\lambda)$ are square, symmetric matrices parameterized by the scalar λ .

- b) Use Von Neumann's trace inequality to show that $\text{MSE}[\hat{\beta}(\lambda)] < \text{MSE}[\hat{\beta}(0)]$ for $0 < \lambda < \sigma^2 \|\beta\|_2^{-2}$. Von Neumann's trace inequality states that, for $p \times p$ -dimensional matrices \mathbf{A} and \mathbf{B} with singular values $d_{a,1} \geq d_{a,2} \geq \dots \geq d_{a,p}$ and $d_{b,1} \geq d_{b,2} \geq \dots \geq d_{b,p}$ respectively, $|\text{tr}(\mathbf{AB})| \leq \sum_{j=1}^p d_{a,j} d_{b,j}$.

Note: the proof in the lecture notes is a stronger one, as it provides a (larger) interval on the penalty parameter where the MSE of the ridge regression estimator is superior to that of the maximum likelihood one.

Question 1.13

Recall that there exists $\lambda > 0$ such that $\text{MSE}(\hat{\beta}) > \text{MSE}[\hat{\beta}(\lambda)]$. Then, show that, for that λ , the mean squared error of the linear predictor satisfies $\text{MSE}(\hat{\mathbf{Y}}) = \text{MSE}(\mathbf{X}\hat{\beta}) \geq \text{MSE}[\mathbf{X}\hat{\beta}(\lambda)]$. Hint: If \mathbf{A} and \mathbf{B} are nonnegative definite square matrices, then $\text{tr}(\mathbf{AB}) \geq 0$.

Question 1.14 (Constrained estimation)

The ridge regression estimator can be viewed as the solution of a constraint estimation problem. Consider the two panels of Figure 1.12. Both show the ridge parameter constraint represented by the grey circle around the origin. The left panel also contains an ellipsoid representing a level set of the sum-of-squares centered around the maximum likelihood regression estimate, while the right panel contains a line formed by the solutions of the normal equations. Both panels have four red dots on the boundary of the ridge parameter constraint. Identify for both cases, which of these four dots is (closest to) the ridge regression estimate?

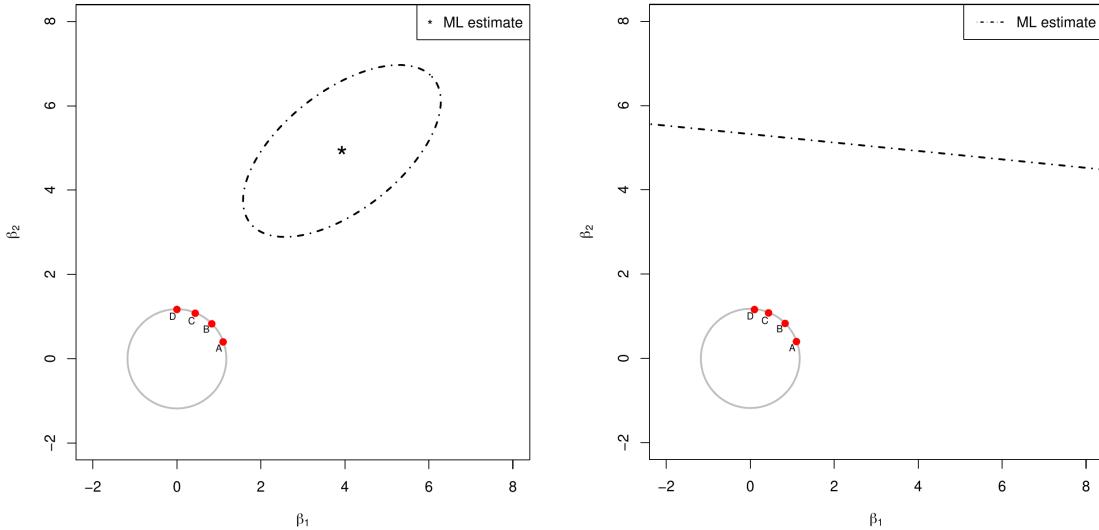


Figure 1.12: Figure related to Exercise 1.14.

Question 1.15

Consider the regularization paths of the elements of the ridge regression estimator of the linear regression model with two covariates. Could it be that, for $\lambda' > \lambda > 0$, we find the estimates

- a) $\hat{\beta}(\lambda) = (2, 1)^\top$ and $\hat{\beta}(\lambda') = (1\frac{1}{2}, 1\frac{1}{10})^\top$?
- b) $\hat{\beta}(\lambda') = (2, 1)^\top$ and $\hat{\beta}(\lambda) = (1\frac{1}{2}, 1\frac{1}{10})^\top$?
- c) $\hat{\beta}(\lambda') = (2, 0)^\top$ and $\hat{\beta}(\lambda) = (1\frac{1}{2}, 0)^\top$?

Question 1.16 (Negative penalty parameter)

Consider fitting the linear regression model, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$, to data by means of the ridge regression estimator. This estimator involves the penalty parameter which is said to be positive. It has been suggested, by among others Hua and Gunst (1983), to extend the range of the penalty parameter to the whole set of real numbers. That is, also tolerating negative values. Let's investigate the consequences of allowing negative values of the penalty parameter. Hereto use in the remainder the following numerical values for the design matrix, response, and corresponding summary statistics:

$$\mathbf{X} = \begin{pmatrix} 5 & 4 \\ -3 & -4 \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} 3 \\ -4 \end{pmatrix}, \quad \mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 34 & 32 \\ 32 & 32 \end{pmatrix}, \quad \text{and} \quad \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} 27 \\ 28 \end{pmatrix}.$$

- a) For which $\lambda < 0$ is the ridge regression estimator $\hat{\beta}(\lambda) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{22})^{-1} \mathbf{X}^\top \mathbf{Y}$ well-defined?
- b) Now consider the ridge regression estimator to be defined via the ridge loss function, i.e.

$$\hat{\beta}(\lambda) = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2.$$

Let $\lambda = -20$. Plot the level sets of this loss function, and add a point with the corresponding ridge regression estimate $\hat{\beta}(-20)$.

- c) Verify that the ridge regression estimate $\hat{\beta}(-20)$ is a saddle point of the ridge loss function, as can also be seen from the contour plot generated in part b). Hereto study the eigenvalues of its Hessian matrix. Moreover, specify the range of negative penalty parameters for which the ridge loss function is convex (and does have a unique well-defined minimum).
- d) Find the minimum of the ridge loss function.

Question 1.17

Consider the standard linear regression model $Y_i = \mathbf{X}_{i,*}\boldsymbol{\beta} + \varepsilon_i$ for $i = 1, \dots, n$ and with $\varepsilon_i \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$. Consider the following two ridge regression estimators of the regression parameter of this model, defined as:

$$\arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n (Y_i - \mathbf{X}_{i,*}\boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \quad \text{and} \quad \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n (Y_i - \mathbf{X}_{i,*}\boldsymbol{\beta})^2 + n\lambda \|\boldsymbol{\beta}\|_2^2.$$

Which do you prefer? Motivate.

Question 1.18

Consider the standard linear regression model $Y_i = \mathbf{X}_{i,*}\beta + \varepsilon_i$ for $i = 1, \dots, n$ and with the ε_i i.i.d. normally distributed with zero mean and a common variance. The rows of the design matrix \mathbf{X} have two elements, and neither column represents the intercept, but $\mathbf{X}_{*,1} = \mathbf{X}_{*,2}$.

- Suppose an estimator of the regression parameter β of this model is obtained through the minimization of the sum-of-squares augmented with a ridge penalty, $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$, in which $\lambda > 0$ is the penalty parameter. The minimizer is called the ridge regression estimator and is denoted by $\hat{\beta}(\lambda)$. Show that $[\hat{\beta}(\lambda)]_1 = [\hat{\beta}(\lambda)]_2$ for all $\lambda > 0$.
- The covariates are now related as $\mathbf{X}_{*,1} = -2\mathbf{X}_{*,2}$. Data on the response and the covariates are:

$$\{(y_i, x_{i,1}, x_{i,2})\}_{i=1}^6 = \{(1.5, 1.0, -0.5), (1.9, -2.0, 1.0), (-1.6, 1.0, -0.5), (0.8, 4.0, -2.0), (0.9, 2.0, -1.0), (-0.5, 4.0, -2.0)\}.$$

Evaluate the ridge regression estimator for these data with $\lambda = 1$.

- The data are as in part b). Show $\hat{\beta}(\lambda + \delta) = (52.5 + \lambda)(52.5 + \lambda + \delta)^{-1}\hat{\beta}(\lambda)$ for a fixed λ and any $\delta > 0$. That is, given the ridge regression estimator evaluated for a particular value of the penalty parameter λ , the remaining regularization path $\{\hat{\beta}(\lambda + \delta)\}_{\delta \geq 0}$ is known analytically. Hint: Use the singular value decomposition of the design matrix \mathbf{X} and the fact that its largest singular value equals $\sqrt{52.5}$.
- The data are as in part b). Consider the model $Y_i = X_{i,1}\gamma + \varepsilon_i$. The parameter γ is estimated through minimization of $\sum_{i=1}^6(Y_i - X_{i,1}\gamma)^2 + \lambda_\gamma\gamma^2$. The perfectly linear relation of the covariates suggests that the regularization paths of the linear predictors $X_{i,1}\hat{\gamma}(\lambda_\gamma)$ and $\mathbf{X}_{i,*}\hat{\beta}(\lambda)$ overlap. Find the functional relationship $\lambda_\gamma = f(\lambda)$ such that the resulting linear predictor $X_{i,1}\hat{\gamma}(\lambda_\gamma)$ indeed coincides with that obtained from the estimate evaluated in part b) of this exercise, i.e. $\mathbf{X}\hat{\beta}(\lambda)$.

Question 1.19

Consider the standard linear regression model $Y_i = \mathbf{X}_{i,*}\beta + \varepsilon_i$ for $i = 1, \dots, n$ and with the ε_i i.i.d. normally distributed with zero mean and a common variance. Moreover, $\mathbf{X}_{*,j} = \mathbf{X}_{*,j'}$ for all $j, j' = 1, \dots, p$ and $\sum_{i=1}^n X_{i,j}^2 = 1$. Show that the ridge regression estimator, defined as $\hat{\beta}(\lambda) = \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$ for $\lambda > 0$, equals:

$$\hat{\beta}(\lambda) = b(\lambda + p)^{-1}\mathbf{1}_p,$$

where $b = \mathbf{X}_{*,1}^\top \mathbf{Y}$. Hint: you may want to use the Sherman-Morrison formula. Let \mathbf{A} and \mathbf{B} be symmetric matrices of the same dimension, with \mathbf{A} invertible and \mathbf{B} of rank one. Moreover, define $g = \text{tr}(\mathbf{A}^{-1}\mathbf{B})$. Then: $(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1} - (1 + g)^{-1}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$.

Question 1.20

Consider the standard linear regression model $Y_i = \mathbf{X}_{i,*}\beta + \varepsilon_i$ for $i = 1, \dots, n$ and with the ε_i i.i.d. normally distributed with zero mean and a common but unknown variance. Information on the response, design matrix and relevant summary statistics are:

$$\mathbf{X}^\top = \begin{pmatrix} 2 & 1 & -2 \end{pmatrix}, \mathbf{Y}^\top = \begin{pmatrix} -1 & -1 & 1 \end{pmatrix}, \mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 9 \end{pmatrix}, \text{ and } \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} -5 \end{pmatrix},$$

from which the sample size and dimension of the covariate space are immediate.

- Evaluate the ridge regression estimator $\hat{\beta}(\lambda)$ with $\lambda = 1$.
- Evaluate the variance of the ridge regression estimator, i.e. $\widehat{\text{Var}}[\hat{\beta}(\lambda)]$, for $\lambda = 1$. In this the error variance σ^2 is estimated by $n^{-1}\|\mathbf{Y} - \mathbf{X}\hat{\beta}(\lambda)\|_2^2$.
- Recall that the ridge regression estimator $\hat{\beta}(\lambda)$ is normally distributed. Consider the interval

$$\mathcal{C} = (\hat{\beta}(\lambda) - 2\{\widehat{\text{Var}}[\hat{\beta}(\lambda)]\}^{1/2}, \hat{\beta}(\lambda) + 2\{\widehat{\text{Var}}[\hat{\beta}(\lambda)]\}^{1/2}).$$

Is this a genuine (approximate) 95% confidence interval for β ? If so, motivate. If not, what is the interpretation of this interval?

- Suppose the design matrix is augmented with an extra column identical to the first one. Moreover, assume λ to be fixed. Is the estimate of the error variance unaffected, or not? Motivate.

Question 1.21

Consider the standard linear regression model $Y_i = \mathbf{X}_{i,*}\beta + \varepsilon_i$ for $i = 1, \dots, n$ and with $\varepsilon_i \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$. The ridge regression estimator of β is denoted by $\hat{\beta}(\lambda)$ for $\lambda > 0$.

a) Show:

$$\text{tr}\{\text{Var}[\hat{\mathbf{Y}}(\lambda)]\} = \sigma^2 \sum_{j=1}^p (\mathbf{D}_x)_{jj}^4 [(\mathbf{D}_x)_{jj}^2 + \lambda]^{-2},$$

where $\hat{\mathbf{Y}}(\lambda) = \mathbf{X}\hat{\beta}(\lambda)$ and \mathbf{D}_x is the diagonal matrix containing the singular values of \mathbf{X} on its diagonal.
b) The coefficient of determination is defined as:

$$R^2 = [\text{Var}(\mathbf{Y}) - \text{Var}(\hat{\mathbf{Y}})]/[\text{Var}(\mathbf{Y})] = [\text{Var}(\mathbf{Y} - \hat{\mathbf{Y}})]/[\text{Var}(\mathbf{Y})],$$

where $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ with $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$. Show that the second equality does not hold when $\hat{\mathbf{Y}}$ is now replaced by the ridge regression predictor defined as $\hat{\mathbf{Y}}(\lambda) = \mathbf{H}(\lambda)\mathbf{Y}$ where $\mathbf{H}(\lambda) = \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top$. Hint: Use the fact that $\mathbf{H}(\lambda)$ is not a projection matrix, i.e. $\mathbf{H}(\lambda) \neq [\mathbf{H}(\lambda)]^2$.

Question 1.22

Fit the linear regression model, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with the usual assumptions, by means of the ridge regression estimator $\hat{\beta}(\lambda)$. Let $df(\lambda)$ denote the degrees of freedom consumed by the ridge regression estimator. Show that $\lim_{\lambda \rightarrow \infty} [n - df(\lambda)]^{-1} \|\mathbf{Y} - \mathbf{X}\hat{\beta}(\lambda)\|_2^2 = n^{-1} \|\mathbf{Y}\|_2^2$, the maximum likelihood estimator of the variance in the response.

Question 1.23

The linear regression model, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$, is fitted to the data with the following design matrix, response and relevant summary statistics:

$$\mathbf{X} = \begin{pmatrix} 2 & 1 \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} 0.5 \end{pmatrix}, \mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}, \text{ and } \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}.$$

Hence, $p = 2$ and $n = 1$.

a) Verify that the singular value decomposition of \mathbf{X} is

$$\mathbf{X} = \mathbf{U}_x \mathbf{D}_x \mathbf{V}_x^\top := \begin{pmatrix} -1 \end{pmatrix} \begin{pmatrix} \sqrt{5} & 0 \end{pmatrix} \begin{pmatrix} -2/\sqrt{5} & -1/\sqrt{5} \\ 1/\sqrt{5} & -2/\sqrt{5} \end{pmatrix}.$$

b) Evaluate the minimum least squares estimator of regression parameter.

c) Consider the ridge regression estimator. Suppose the parameter constraint induced by its ridge penalty has a radius equalling $\frac{1}{4}\sqrt{1/5}$. What is the value of its penalty parameter?

d) How many degrees of freedom are consumed by the ridge regression estimator for the penalty parameter found in part b)?

Question 1.24 (Computationally efficient evaluation)

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, without intercept and $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$, to explain the variation in the response \mathbf{Y} by a linear combination of the columns of the design matrix \mathbf{X} . The linear regression model is fitted by means of ridge estimation. The estimator is evaluated directly from its regular expression and a computationally efficient one:

$$\begin{aligned} \hat{\beta}_{\text{reg}}(\lambda) &= (\lambda \mathbf{I}_{pp} + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}, \\ \hat{\beta}_{\text{eff}}(\lambda) &= \lambda^{-1} \mathbf{X}^\top \mathbf{Y} - \lambda^{-2} \mathbf{X}^\top (\mathbf{I}_{nn} + \lambda^{-1} \mathbf{X} \mathbf{X}^\top)^{-1} \mathbf{X} \mathbf{X}^\top \mathbf{Y}, \end{aligned}$$

respectively. In the remainder study the computational gain of the latter. Hereto carry out the following instructions:

- a) Load the R-package `microbenchmark` (Mersmann, 2014).
- b) Generate data. In this fix the sample size at $n = 10$, and let the dimension range from $p = 10, 20, 30, \dots, 100$. Sample the elements of the response and the ten design matrices from the standard normal distribution.
- c) Verify the superior computation time of the latter by means of the `microbenchmark`-function with default settings. Throughout use the first design matrix and $\lambda = 1$. Write the output of the `microbenchmark`-function to an R-object. It will be a `data.frame` with two slots `expr` and `time` that contain the function calls and the corresponding computation times, respectively. Each call has by default been evaluated a hundred times in random order. Summary statistics of these individual computation times are given when printing the object on the screen.

- d) Use the `crossprod`- and `tcrossprod`-functions to improve the computation times of the evaluation of both $\hat{\beta}_{\text{reg}}(\lambda)$ and $\hat{\beta}_{\text{eff}}(\lambda)$ as much as possible.
- e) Use the `microbenchmark`-function to evaluate the (average) computation time both $\hat{\beta}_{\text{reg}}(\lambda)$ and $\hat{\beta}_{\text{eff}}(\lambda)$ on all ten data sets, i.e. defined by the ten design matrices with different dimensions.
- f) Plot, for both $\hat{\beta}_{\text{reg}}(\lambda)$ and $\hat{\beta}_{\text{eff}}(\lambda)$, the (average) computation time (on the y -axis) against the dimension of the ten data sets. Conclude on the computation gain from the plot.

Question 1.25 (Computationally efficient LOOCV)

Consider the ridge regression estimator $\hat{\beta}(\lambda)$ of the linear regression model parameter β . Its penalty parameter λ may be chosen as the minimizer of Allen's PRESS statistic, i.e.: $\lambda_{\text{opt}} = \arg \min_{\lambda > 0} n^{-1} \sum_{i=1}^n [Y_i - \mathbf{X}_{i,*} \hat{\beta}_{-i}(\lambda)]^2$, with the LOOCV ridge regression estimator $\hat{\beta}_{-i}(\lambda) = (\mathbf{X}_{-i,*}^\top \mathbf{X}_{-i,*} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}_{-i,*}^\top \mathbf{Y}_{-i}$. This is computationally demanding as it involves n evaluations of $\hat{\beta}_{-i}(\lambda)$, which can be circumvented by rewriting Allen's PRESS statistics. Hereto:

- a) Use the Woodbury matrix identity to verify:

$$(\mathbf{X}_{-i,*}^\top \mathbf{X}_{-i,*} + \lambda \mathbf{I}_{pp})^{-1} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} + (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}_{i,*}^\top [1 - \mathbf{H}_{ii}(\lambda)]^{-1} \mathbf{X}_{i,*} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1},$$

in which $\mathbf{H}_{ii}(\lambda) = \mathbf{X}_{i,*} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}_{i,*}^\top$.

- b) Rewrite the LOOCV ridge regression estimator to:

$$\hat{\beta}_{-i}(\lambda) = \hat{\beta}(\lambda) - (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}_{i,*}^\top [1 - \mathbf{H}_{ii}(\lambda)]^{-1} [Y_i - \mathbf{X}_{i,*} \hat{\beta}(\lambda)].$$

In this use part a) and the identity $\mathbf{X}_{-i}^\top \mathbf{Y}_{-i} = \mathbf{X}^\top \mathbf{Y} - \mathbf{X}_{i,*}^\top Y_i$.

- c) Reformulate, using part b), the prediction error as $Y_i - \mathbf{X}_{i,*} \hat{\beta}_{-i}(\lambda) = [1 - \mathbf{H}_{ii}(\lambda)]^{-1} [Y_i - \mathbf{X}_{i,*}^\top \hat{\beta}(\lambda)]$ and express Allen's PRESS statistic as:

$$n^{-1} \sum_{i=1}^n [Y_i - \mathbf{X}_{i,*} \hat{\beta}_{-i}(\lambda)]^2 = n^{-1} \|\mathbf{B}(\lambda) [\mathbf{I}_{nn} - \mathbf{H}(\lambda)] \mathbf{Y}\|_F^2,$$

where $\mathbf{B}(\lambda)$ is diagonal with $[\mathbf{B}(\lambda)]_{ii} = [1 - \mathbf{H}_{ii}(\lambda)]^{-1}$.

Question 1.26 (Computationally efficient K -fold cross-validation)

Verify the expression of the k -fold linear predictor (1.13).

Question 1.27 (Optimal penalty parameter)

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\beta \in \mathbb{R}^p$, $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$,

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}, \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} -3 \\ 5 \end{pmatrix}.$$

Find the penalty parameter:

- a) that minimizes the Mean Squared Error (MSE) of the ridge regression estimator for this data set.
b) by means of leave-one-out cross-validation, minimizing Allen's PRESS statistic.
c) that minimizes the Akaike's information criterion.

Question 1.28 (LOOCV)

The linear regression model, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$ is fitted by means of the ridge regression estimator. The design matrix and response are:

$$\mathbf{X} = \begin{pmatrix} 2 & -1 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} 1 \\ \frac{1}{2} \end{pmatrix}.$$

The penalty parameter is chosen as the minimizer of the leave-one-out cross-validated squared error of the prediction (i.e. Allen's PRESS statistic). Show that $\lambda = \infty$.

Question 1.29

Consider fitting the linear regression model, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with design matrix \mathbf{X} , $\beta \in \mathbb{R}^p$, and $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$,

by means of the ridge regression estimator, $\hat{\beta}(\lambda) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y}$, for the below provided choices of \mathbf{X} and \mathbf{Y} :

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix}, \quad \mathbf{H}(\lambda) = \frac{1}{(\lambda+1)(\lambda+3)} \begin{pmatrix} 2+\lambda & -1 & -(1+\lambda) \\ -1 & 2+\lambda & -(1+\lambda) \\ -(1+\lambda) & -(1+\lambda) & 2(1+\lambda) \end{pmatrix},$$

and $\mathbf{Y} = (-2, 1, 2)^\top$. The above display also contains the ridge hat matrix, which comes in handy for the answer. Prior to fitting, choose the penalty parameter through generalized cross-validation. Show that then $\lambda_{\text{opt}} \in (0, \frac{6}{53})$, and that this interval indeed contains a single minimum, which is indeed the global minimum of $\text{GCV}(\lambda)$ on the positive real line.

Question 1.30 (Penalty parameter selection)

PSA (Prostate-Specific Antigen) is a prognostic indicator of prostate cancer. Low and high PSA values indicate low and high risk, respectively. PSA interacts with the VEGF pathway. In cancer the VEGF pathway aids in the process of angiogenesis, i.e. the formation of blood vessels in solid tumors. Assume the aforementioned interaction can – at least partially – be captured by a linear relationship between PSA and the constituents of the VEGF pathway. Use the prostate cancer data of Ross-Adams *et al.* (2015) to estimate this linear relationship using the ridge regression estimator. The following R-script downloads and prepares the data.

Listing 1.7 R code

```
# load the necessary libraries
library(Biobase)
library(prostateCancerStockholm)
library(penalized)
library(KEGG.db)

# load data
data(stockholm)

# prepare PSA data
psa <- pData(stockholm) [, 9]
psa <- log(as.numeric(levels(psa)[psa]))

# prepare VEGF pathway data
X <- exprs(stockholm)
kegg2vegf <- as.list(KEGGPATHID2EXTID)
entrezIDvegf <- as.numeric(unlist(kegg2vegf[names(kegg2vegf) == "hsa04370"]))
entrezIDx <- as.numeric(fData(stockholm) [, 10])
idx2VEGF <- match(entrezIDvegf, entrezIDx)
entrezIDvegf <- entrezIDvegf[-which(is.na(idx2VEGF))]
idx2VEGF <- match(entrezIDvegf, entrezIDx)
X <- t(X[idx2VEGF, ])
gSymbols <- fData(stockholm)[idx2VEGF, 13]
gSymbols <- levels(gSymbols)[gSymbols]
colnames(X) <- gSymbols

# remove samples with missing outcome
X <- X[-which(is.na(psa)), ]
X <- sweep(X, 2, apply(X, 2, mean))
X <- sweep(X, 2, apply(X, 2, sd), "/")
psa <- psa[-which(is.na(psa))]
psa <- psa - mean(psa)
```

- a) Find the ridge penalty parameter by means of AIC minimization. *Hint:* the likelihood can be obtained from the penFit-object that is created by the penalized-function of the R-package penalized.
- b) Find the ridge penalty parameter by means of leave-one-out cross-validation, as implemented by the optL2-function provided by the R-package penalized.
- c) Find the ridge penalty parameter by means of leave-one-out cross-validation using Allen's PRESS statistic as performance measure (see Section 1.8.2).

- d) Discuss the reasons for the different values of the ridge penalty parameter obtained in parts a), b), and c). Also investigate the consequences of these values on the corresponding regression estimates.

Question 1.31 (Covariate rescaling)

The variance of the covariates influences the amount of shrinkage of the regression estimator induced by ridge regularization. Some deal with this through rescaling of the covariates to have a common unit variance. This is discussed at the end of Section 1.9.1. Investigate this numerically using the data of the microRNA-mRNA illustration discussed in Section 1.10.

- a) Load the data by running the first R-script of Section 1.10.
- b) Fit the linear regression model by means of the ridge regression estimator with $\lambda = 1$ using both the scaled and unscaled covariates. Compare the order of the coefficients between the both estimates as well as their corresponding linear predictors. Does the top 50 largest (in an absolute size) coefficients differ much between the two estimates?
- c) Repeat part b), now with the penalty parameter chosen by means of LOOCV.

2 Bayesian regression

The ridge regression estimator is equivalent to a Bayesian regression estimator. On one hand this equivalence provides another interpretation of the ridge regression estimator. But it also shows that within the Bayesian framework the high-dimensionality of the data need not frustrate the numerical evaluation of the estimator. In addition, the framework provides ways to quantify the consequences of high-dimensionality on the uncertainty of the estimator. Within this chapter, focus is on the equivalence of Bayesian and ridge regression. In this particular case, the connection is immediate from the analytic formulation of the Bayesian regression estimator. After this has been presented, it is shown how this estimator may also be obtained by means of sampling. The relevance of the sampling for the evaluation of the estimator and its uncertainty becomes apparent in subsequent chapters where we discuss other penalized estimators for which the connection cannot be captured in analytic form.

2.1 A minimum of prior knowledge on Bayesian statistics

The schism between Bayesian and frequentist statistics centers on the interpretation of the concept of probability. A frequentist views the probability of an event as the limiting frequency of observing the event among a large number of trials. In contrast, a Bayesian considers it to be a measure of believe (in the occurrence) of the event. The difference between the two interpretations becomes clear when considering events that can occur only once (or a small number of times). A Bayesian would happily discuss the probability of this (possibly hypothetical) event happening, which would be meaningless to a frequentist.

What are the consequences of this schism for the current purpose of estimating a regression model? Exponents from both paradigms assume a statistical model, e.g. here the linear regression model, to describe the data generating mechanism. But a frequentist treats the parameters as platonic quantities for which a true value exists that is to be estimated from the data. A Bayesian, however, first formalizes his/her current believe/knowledge on the parameters in the form of probability distributions. This is referred to as the prior – to the experiment – distribution. The parameters are thus random variables and their distributions are to be interpreted as reflecting the (relative) likelihood of the parameters' values. From the Bayesian point it then makes sense to talk about the random behaviour of the parameter, e.g., what is probability of that the parameter is larger than a particular value. In the frequentist context one may ask this of the estimator, but not of the parameter. The parameter either *is* or *is not* larger than this value as a platonic quantity is not governed by a chance process. Then, having specified his/her believe on the parameters, a Bayesian uses the data to update this belief of the parameters of the statistical model, which again is a distribution and called the posterior – to the experiment – distribution.

To illustrate this process of updating assume that the data $Y_1, \dots, Y_n \sim \{0, 1\}$ are assumed to be independently and identically drawn from a Bernouilli distribution with parameter $\theta = P(Y_i = 1)$. The likelihood of these data for a given choice of the parameter θ then is: $L(\mathbf{Y} = \mathbf{y}; \theta) = P(\mathbf{Y} = \mathbf{y} | \theta) = P(Y_1 = y_1, \dots, Y_n = y_n | \theta) = \prod_{i=1}^n P(Y_i = y_i | \theta) = \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i}$. A Bayesian now needs to specify the prior distribution, denoted $\pi_\theta(\cdot)$, on the parameter θ . A common choice in this case would be a beta-distribution: $\theta \sim \mathcal{B}(\alpha, \beta)$. The posterior distribution is now arrived at by the use of Bayes' rule:

$$\pi(\theta | \mathbf{Y} = \mathbf{y}) = \frac{P(\mathbf{Y} = \mathbf{y} | \theta) \pi(\theta)}{P(\mathbf{Y} = \mathbf{y})} = \frac{P(\mathbf{Y} = \mathbf{y} | \theta) \pi(\theta)}{\int_{0,1} P(\mathbf{Y} = \mathbf{y} | \theta) \pi(\theta) d\theta}.$$

The posterior distribution thus contains all knowledge on the parameter. As the denominator – referred to as the distribution's normalizing constant – in the expression of the posterior distribution does not involve the parameter θ , one often writes $\pi(\theta | \mathbf{Y} = \mathbf{y}) \propto P(\mathbf{Y} = \mathbf{y} | \theta) \pi(\theta)$, thus specifying the (relative) density of the posterior distribution of θ as 'likelihood \times prior'.

While the posterior distribution is all one wishes to know, often a point estimate is required. A point estimate of θ can be obtained from the posterior by taking the mean or the mode. Formally, the Bayesian point estimator of

a parameter θ is defined as the estimator that minimizes the Bayes risk over a prior distribution of the parameter θ . The Bayes risk is defined as $\int_{\theta} \mathbb{E}[(\hat{\theta} - \theta)^2] \pi_{\theta}(\theta; \alpha) d\theta$, where $\pi_{\theta}(\theta; \alpha)$ is the prior distribution of θ with hyperparameter α . It is thus a weighted average of the Mean Squared Error, with weights specified through the prior. The Bayes risk is minimized by the posterior mean:

$$\mathbb{E}_{\theta}(\theta | \text{data}) = \frac{\int_{\theta} \theta P(\mathbf{Y} = \mathbf{y} | \theta) \pi(\theta) d\theta}{\int_{\theta} P(\mathbf{Y} = \mathbf{y} | \theta) \pi(\theta) d\theta}.$$

(cf., e.g., Bijma *et al.*, 2017). The Bayesian point estimator of θ yields the smallest possible expected MSE, under the assumption of the employed prior. This estimator thus depends on the likelihood *and* the prior, and a different prior yields a different estimator.

A point estimator is preferably accompanied by a quantification of its uncertainty. Within the Bayesian context this is done by so-called credible intervals or regions, the Bayesian equivalent of confidence intervals or regions. A Bayesian credible interval encompass a certain percentage of the probability mass of the posterior. For instance, a subset of the parameter space \mathcal{C} forms an $100(1 - \alpha)\%$ credible interval if $\int_{\mathcal{C}} \pi(\theta | \text{data}) d\theta = 1 - \alpha$.

In the example above it is important to note the role of the prior in the updating of the knowledge on θ . First, when the prior distribution is uniform on the unit interval, the mode of the posterior coincides with the maximum likelihood estimator. Furthermore, when n is large the influence of the prior in the posterior is negligible. It is therefore that for large sample sizes frequentist and Bayesian analyses tend to produce similar results. However, when n is small, the prior's contribution to the posterior distribution cannot be neglected. The choice of the prior is then crucial as it determines the shape of the posterior distribution and, consequently, the posterior estimates. It is this strong dependence of the posterior on the arbitrary (?) choice of the prior that causes unease with a frequentist (leading some to accuse Bayesians of subjectivity). In the high-dimensional setting the sample size is usually small, especially in relation to the parameter dimension, and the choice of the prior distribution then matters. Interest here is not in resolving the frequentist's unease, but in identifying or illustrating the effect of the choice of the prior distribution on the parameter estimates.

2.2 Relation to ridge regression

Ridge regression has a close connection to Bayesian linear regression. Bayesian linear regression assumes the parameters β and σ^2 to be the random variables, while at the same time considering \mathbf{X} and \mathbf{Y} as fixed. Within the regression context, the commonly chosen priors of β and σ^2 are $\beta | \sigma^2 \sim \mathcal{N}(\mathbf{0}_p, \sigma^2 \lambda^{-1} \mathbf{I}_{pp})$ and $\sigma^2 \sim \mathcal{IG}(\alpha_0, \beta_0)$, where \mathcal{IG} denotes the inverse Gamma distribution with shape parameter α_0 and scale parameter β_0 . The penalty parameter can be interpreted as the (scaled) precision of the prior of β , determining how informative the prior should be. A smaller penalty (i.e. precision) corresponds to a wider prior, and a larger penalty to a more informative, concentrated prior (Figure 2.1).

The joint posterior distribution of β and σ^2 is, under the assumptions of the (likelihood of) the linear regression model and the priors above:

$$\begin{aligned} f_{\beta, \sigma^2}(\beta, \sigma^2 | \mathbf{Y}, \mathbf{X}) &\propto f_Y(\mathbf{Y} | \mathbf{X}, \beta, \sigma^2) f_{\beta}(\beta | \sigma^2) f_{\sigma}(\sigma^2) \\ &\propto \sigma^{-n} \exp[-\frac{1}{2}\sigma^{-2}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)] \\ &\quad \times \sigma^{-p} \exp(-\frac{1}{2}\sigma^{-2}\lambda\beta^T\beta) \times (\sigma^2)^{-\alpha_0-1} \exp(-\beta_0\sigma^{-2}). \end{aligned}$$

The posterior distribution can be expressed as a multivariate normal distribution. Hereto group the terms from the exponential functions that involve β and manipulate as follows:

$$\begin{aligned} &(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda\beta^T\beta \\ &= \mathbf{Y}^T\mathbf{Y} - 2\beta^T\mathbf{X}^T\mathbf{Y} + \beta^T\mathbf{X}^T\mathbf{X}\beta + \lambda\beta^T\beta \\ &= \mathbf{Y}^T\mathbf{Y} - 2\beta^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T\mathbf{Y} + \beta^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})\beta \\ &= \mathbf{Y}^T\mathbf{Y} - \beta^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})\hat{\beta}(\lambda) - [\hat{\beta}(\lambda)]^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})\beta + \beta^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})\beta \\ &= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T\mathbf{Y} + [\beta - \hat{\beta}(\lambda)]^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})[\beta - \hat{\beta}(\lambda)]. \end{aligned}$$

Using this result, the posterior distribution can be rewritten to:

$$f_{\beta, \sigma^2}(\beta, \sigma^2 | \mathbf{Y}, \mathbf{X}) \propto g_{\beta}(\beta | \sigma^2, \mathbf{Y}, \mathbf{X}) g_{\sigma^2}(\sigma^2 | \mathbf{Y}, \mathbf{X})$$

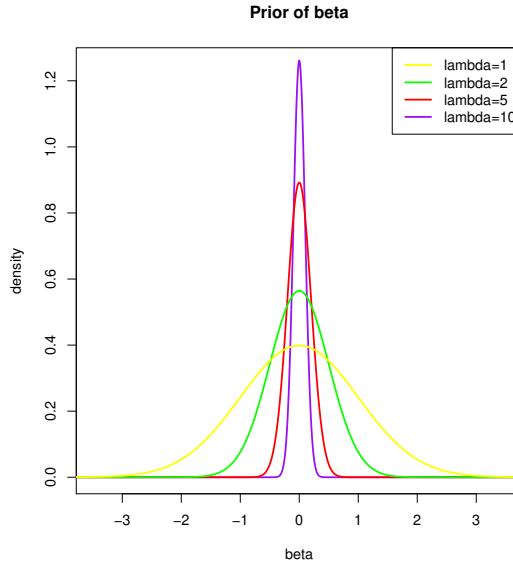


Figure 2.1: Conjugate prior of the regression parameter β for various choices of λ , the penalty parameters c.q. precision.

with

$$g_{\beta}(\beta | \sigma^2, \mathbf{Y}, \mathbf{X}) \propto \exp \left\{ -\frac{1}{2}\sigma^{-2} [\beta - \hat{\beta}(\lambda)]^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp}) [\beta - \hat{\beta}(\lambda)] \right\}.$$

Clearly, after having recognized the form of a multivariate normal density in the expression of the preceding display, the conditional posterior mean of β is $\mathbb{E}(\beta | \sigma^2, \mathbf{Y}, \mathbf{X}) = \hat{\beta}(\lambda)$. Hence, the ridge regression estimator can be viewed as the Bayesian posterior mean estimator of β when imposing a Gaussian prior on the regression parameter.

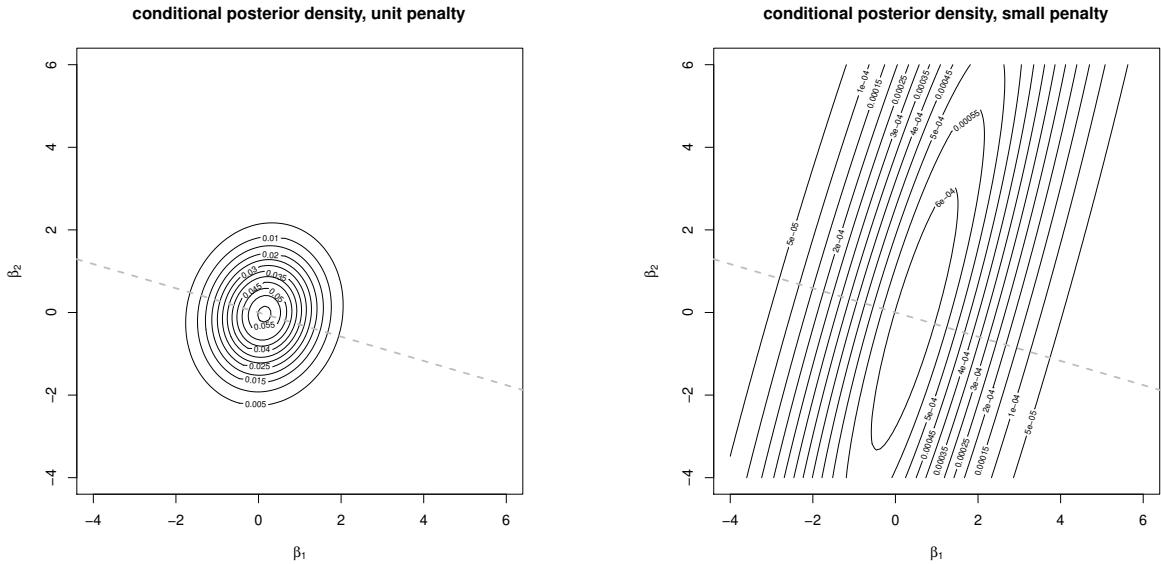


Figure 2.2: Level sets of the (unnormalized) conditional posterior of the regression parameter β . The grey dashed line depicts the support of the degenerated normal distribution of the frequentist's ridge regression estimate.

The conditional posterior distribution of β , $g_{\beta}(\beta | \sigma^2, \mathbf{Y}, \mathbf{X})$, related to a high-dimensional situation ($n = 1$, $p = 2$) is depicted in Figure 2.2 for two choices of the penalty parameter λ . Put differently, for two different priors.

In one, the left panel, λ is arbitrarily set equal to one. The choice of the employed λ , i.e. $\lambda = 1$, is irrelevant, the resulting posterior distribution only serves as a reference. This choice results in a posterior distribution that is concentrated around the Bayesian point estimate, which coincides with the ridge regression estimate. The almost spherical level sets around the point estimate may be interpreted as credible intervals for β . The grey dashed line, spanned by the row of the design matrix, represents the support of the degenerated normal distribution of the frequentist's ridge regression estimate (cf. end of Section 1.4.2). The contrast between the degenerated frequentist and well-defined Bayesian normal distribution illustrates that – for a suitable choice of the prior – within the Bayesian context high-dimensionality need not be an issue with respect to the evaluation of the posterior. In the right panel of Figure 2.2 the penalty parameter λ is set equal to a very small value, i.e. $0 < \lambda \ll 1$. This represents a very imprecise or uninformative prior. It results in a posterior distribution with level sets that are far from spherical and are very stretched in the direction orthogonal to the subspace spanned by the row of the design matrix and indicates in which direction there is most uncertainty with respect to β . In particular, when $\lambda \downarrow 0$, the level sets loose their ellipsoid form and the posterior collapses to the degenerated normal distribution of the frequentist's ridge regression estimate. Finally, in combination with the left-hand plot this illustrates the large effect of the prior in the high-dimensional context.

With little extra work we may also obtain the conditional posterior of σ^2 from the joint posterior distribution:

$$g_{\sigma^2}(\sigma^2 | \mathbf{Y}, \mathbf{X}) \propto (\sigma^2)^{-(n/2+\alpha_0+1)} \exp[-\frac{1}{2}\sigma^{-2}(\|\mathbf{Y}\|_2^2 - \lambda\|\hat{\beta}(\lambda)\|_2^2 - \|\mathbf{X}\hat{\beta}(\lambda)\|_2^2 + 2\beta_0)],$$

in which one can recognize the shape of an inverse gamma distribution. The relevance of this conditional distribution will be made clear in Section 2.3.

Sampling from the conditional posterior of β becomes horrifically slow in high dimensions. This is overcome by the computationally efficient sampling scheme proposed by Bhattacharya *et al.* (2016). To draw a sample from the multivariate normal distribution with a structured mean vector and covariance matrix,

$$\mathcal{N}[(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y}, \sigma^2 (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}]. \quad (2.1)$$

The pseudo-code of the algorithm comprises

input : design matrix \mathbf{X} , response vector \mathbf{Y} , variance σ^2 , and penalty parameter λ .
output: a draw from distribution (2.1).

- 1 Sample $\mathbf{u} \sim \mathcal{N}(\mathbf{0}_p, \sigma^2 \lambda^{-1} \mathbf{I}_{pp})$ and $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_{nn})$ independently.
- 2 Set $\mathbf{v} = \sigma^{-1} \mathbf{X} \mathbf{u} + \boldsymbol{\delta}$.
- 3 Solve $(\lambda^{-1} \mathbf{X} \mathbf{X}^\top + \mathbf{I}_{nn}) \mathbf{w} = \sigma^{-1} \mathbf{Y} - \mathbf{v}$ for \mathbf{w} .
- 4 Set $\beta = \mathbf{u} + \sigma \lambda^{-1} \mathbf{X}^\top \mathbf{w}$.

Algorithm 1: High-dimensional efficient sampling of distribution (2.1).

The next proposition warrants that the Algorithm 1 indeed samples from distribution (2.1).

Proposition 2.1 (*Proposition 1 of Bhattacharya et al., 2016*)

If β is sampled in accordance with Algorithm 1. Then, β is distributed as (2.1).

Proof. It is immediate that $\mathbf{v} \sim \mathcal{N}(\mathbf{0}_n, \lambda^{-1} \mathbf{X} \mathbf{X}^\top + \mathbf{I}_{nn})$. Furthermore, as

$$\beta = \mathbf{u} + \sigma \lambda^{-1} \mathbf{X}^\top (\lambda^{-1} \mathbf{X} \mathbf{X}^\top + \mathbf{I}_{nn})^{-1} (\sigma^{-1} \mathbf{Y} - \mathbf{v}),$$

β is normally distributed. Its mean is $\mathbb{E}(\beta) = \lambda^{-1} \mathbf{X}^\top (\lambda^{-1} \mathbf{X} \mathbf{X}^\top + \mathbf{I}_{nn})^{-1} \mathbf{Y}$, which is the expression of the ridge regression estimator (1.12). The variance is

$$\begin{aligned} \text{Var}(\beta) &= \text{Var}[\mathbf{u} - \sigma \lambda^{-1} \mathbf{X}^\top (\lambda^{-1} \mathbf{X} \mathbf{X}^\top + \mathbf{I}_{nn})^{-1} \mathbf{v}] \\ &= \sigma^2 \lambda^{-1} \mathbf{I}_{pp} - \sigma^2 \lambda^{-2} \mathbf{X}^\top (\lambda^{-1} \mathbf{X} \mathbf{X}^\top + \mathbf{I}_{nn})^{-1} \mathbf{X}^\top \\ &= \sigma^2 (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}, \end{aligned}$$

where we have used the rules of variance calculus, the independence between \mathbf{u} and \mathbf{v} , and the Woodbury matrix identity. ■

Algorithm 1 is valid for any (n, p) -combination, but most computational speed is gained if $p \gg n$.

In general, any penalized estimator has a Bayesian equivalent. That is generally not the posterior mean as for the ridge regression estimator, which is more a coincide due to the normal form of its likelihood and conjugate prior. A penalized estimator always coincides with the Bayesian MAP (Mode A Posteriori) estimator. Then, the MAP estimates estimates a parameter θ by the mode, i.e. the maximum, of the posterior density. To see the equivalence of both estimators we derive the MAP estimator. The latter is defined as:

$$\hat{\theta}_{\text{map}} = \arg \max_{\theta} \pi(\theta | \mathbf{Y} = \mathbf{y}) = \arg \max_{\theta} \frac{P(\mathbf{Y} = \mathbf{y} | \theta) \pi(\theta)}{\int P(\mathbf{Y} = \mathbf{y} | \theta) \pi(\theta) d\theta}.$$

To find the maximum of the posterior density take the logarithm (which does not change the location of the maximum) and drop terms that do not involve θ . The MAP estimator then is:

$$\hat{\theta}_{\text{map}} = \arg \max_{\theta} \log[P(\mathbf{Y} = \mathbf{y} | \theta)] + \log[\pi(\theta)].$$

The first summand on the right-hand side is the loglikelihood, while the second is the logarithm of the prior density. The latter, in case of a normal prior, is proportional to $\sigma^{-2} \lambda \|\beta\|_2^2$. This is – up to some factors – exactly the loss function of the ridge regression estimator. If the quadratic ridge penalty is replaced by different one, it is easy to see what form the prior density should have in order for both estimators – the penalized and MAP – to coincide.

2.3 Markov chain Monte Carlo

The ridge regression estimator was shown to coincide with the posterior mean of the regression parameter when it is endowed with a normal prior. This particular choice of the prior yielded a closed form expression of the posterior, and its mean. Prior distributions that result in well-characterized posterior ones are referred to as *conjugate*. E.g., for the standard linear regression model a normal prior is conjugate. Conjugate priors, however, are the exception rather than the rule. In general, an arbitrary prior distribution will not result in a posterior distribution that is familiar (amongst others due to the fact that its normalizing constant is not analytically evaluable). For instance, would a wildly different prior for the regression parameter be chosen, its posterior is unlikely to be analytically known. Then, although analytically unknown, such posteriors can be investigated *in silico* by means of the Markov chain Monte Carlo (MCMC) method. In this subsection the MCMC method is explained and illustrated – despite its conjugacy – on the linear regression model with a normal prior.

Markov chain Monte Carlo is a general purpose technique for sampling from complex probabilistic models/distributions. It draws a sample from a Markov chain that has been constructed such that its stationary distribution equals the desired distribution (read: the analytically untractable posterior distribution). The chain is, after an initial number of iterations (called the *burn-in* period), believed to have converged and reached stationarity. A sample from the stationary chain is then representative of the desired distribution. Statistics derived from this sample can be used to characterize the desired distribution. Hence, estimation is reduced to setting up and running a Markov process on a computer. For more on the MCMC, see Chib and Greenberg (1995) and Geyer (2011) and the references therein.

Recall Monte Carlo integration. Assume the analytical evaluation of $\mathbb{E}[h(\mathbf{Y})] = \int h(\mathbf{y}) \pi(\mathbf{y}) d\mathbf{y}$ is impossible. This quantity can nonetheless be estimated when samples from $\pi(\cdot)$ can be drawn. For, if $\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(T)} \sim \pi(\mathbf{y})$, the integral may be estimated by: $\widehat{\mathbb{E}}[h(\mathbf{Y})] = T^{-1} \sum_{t=1}^T h(\mathbf{Y}^{(t)})$. By the law of large numbers this is a consistent estimator, i.e. if $T \rightarrow \infty$ the estimator converges (in probability) to its true value. Thus, irrespective of the specifics of the posterior distribution $\pi(\cdot)$, if one is able to sample from it, its moments (or other quantities) can be estimated. Or, by the same principle it may be used in the Bayesian regression framework of Section 2.2 to sample from the marginal posterior of β ,

$$f_{\beta}(\beta | \mathbf{Y}, \mathbf{X}) = \int_0^{\infty} f_{\beta, \sigma^2}(\beta, \sigma^2 | \mathbf{Y}, \mathbf{X}) d\sigma^2 = \int_0^{\infty} g_{\beta}(\beta | \sigma^2, \mathbf{Y}, \mathbf{X}) g_{\sigma^2}(\sigma^2 | \mathbf{Y}, \mathbf{X}) d\sigma^2,$$

given that it of σ^2 is known. By drawing a sample $\sigma^{2,(1)}, \dots, \sigma^{2,(T)}$ of $g_{\sigma^2}(\sigma^2 | \mathbf{Y}, \mathbf{X})$ and, subsequently, a sample $\{\beta^{(t)}\}_{t=1}^T$ from $g_{\beta}(\beta | \sigma^{2,(t)}, \mathbf{Y}, \mathbf{X})$ for $t = 1, \dots, T$, the parameter σ^2 has effectively been integrated out and the resulting sample of $\beta^{(1)}, \beta^{(2)}, \dots, \beta^{(T)}$ is from $f_{\beta}(\beta | \mathbf{Y}, \mathbf{X})$.

It is clear from the above that the ability to sample from the posterior distribution is key to the estimation of the parameters. But this sampling is hampered by knowledge of the normalizing constant of the posterior.

This is overcome by the *acceptance-rejection sampler*. This sampler generates independent draws from a target density $\pi(\mathbf{y}) = f(\mathbf{y})/K$, where $f(\mathbf{y})$ is an unnormalized density and K its (unknown) normalizing constant. The sampler relies on the existence of a density $h(\mathbf{y})$ that dominates $f(\mathbf{y})$, i.e., $f(\mathbf{y}) \leq c h(\mathbf{y})$ for some known $c \in \mathbb{R}_{>0}$, from which it is possible to simulate. Then, in order to draw a \mathbf{Y} from the posterior $\pi(\cdot)$ run Algorithm 2:

```

input : densities  $f(\cdot)$  and  $h(\cdot)$ ;  

         constant  $c > 0$ ;  

output: a draw from  $\pi(\cdot)$ .
```

- 1 Generate \mathbf{Y} from $h(\cdot)$.
- 2 Generate U from the uniform distribution $\mathcal{U}[0, 1]$.
- 3 **if** $U \leq f(\mathbf{Y}) / [c \times h(\mathbf{Y})]$ **then**
- 4 | Return \mathbf{Y} .
- 5 **else**
- 6 | Go back to line 1.
- 7 **end**

Algorithm 2: Acceptance-rejection sampling.

A proof that this indeed yields a random sample from $\pi(\cdot)$ is given in Flury (1990). For this method to be computationally efficient, c is best set equal to $\sup_{\mathbf{y}}\{f(\mathbf{y})/h(\mathbf{y})\}$. The R-script below illustrates this sampler for the unnormalized density $f(y) = \cos^2(2\pi y)$ on the unit interval. As $\max_{y \in [0, 1]} \cos^2(2\pi y) = 1$, the density $f(y)$ is dominated by the density function of the uniform distribution $\mathcal{U}[0, 1]$ and, hence, the script uses $h(y) = 1$ for all $y \in [0, 1]$.

Listing 2.1 R code

```

# define unnormalized target density
cos2density <- function(x) { (cos(2*pi*x))^2 }

# define acceptance-rejection sampler
acSampler <- function(n) {
  draw <- numeric()
  for(i in 1:n) {
    accept <- FALSE
    while (!accept) {
      Y <- runif(1)
      if (runif(1) <= cos2density(Y)) {
        accept <- TRUE
        draw <- c(draw, Y)
      }
    }
  }
  return(draw)
}

# verify by eye-ballng
hist(acSampler(100000), n=100)

```

In practice, it may not be possible to find a density $h(x)$ that dominates the unnormalized target density $f(\mathbf{y})$ over the whole domain. Or, the constant c is too large, yielding a rather small acceptance probability, which would make the acceptance-rejection sampler impractically slow. A different sampler is then required.

The Metropolis-Hastings sampler overcomes the problems of the acceptance-rejection sampler and generates a sample from a target distribution that is known up to its normalizing constant. Hereto it constructs a Markov chain that converges to the target distribution. A Markov chain is a sequence of random variables $\{\mathbf{Y}_t\}_{t=1}^\infty$ with $\mathbf{Y}_t \in \mathbb{R}^p$ for all t that exhibit a simple dependence relationship among subsequent random variables in the sequence. Here that simple relationship refers to the fact/assumption that the distribution of \mathbf{Y}_{t+1} only depends on \mathbf{Y}_t and not on the part of the sequence preceding it. The Markov chain's random walk is usually initiated by a single draw from some distribution, resulting in \mathbf{Y}_1 . From then on the evolution of the Markov chain is described by the *transition kernel*. The transition kernel is a the conditional density $g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{Y}_{t+1} = \mathbf{y}_a | \mathbf{Y}_t = \mathbf{y}_b)$ that specifies the distribution of the random variable at the next instance given the realization of the current one. Under

some conditions (akin to aperiodicity and irreducibility for Markov chains in discrete time and with a discrete state space), the influence of the initiation washes out and the random walk converges. Not to a specific value, but to a distribution. This is called the stationary distribution, denoted by $\varphi(\mathbf{y})$, and $\mathbf{Y}_t \sim \varphi(\mathbf{y})$ for large enough t . The stationary distribution satisfies:

$$\varphi(\mathbf{y}_a) = \int_{\mathbb{R}^p} g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{Y}_{t+1} = \mathbf{y}_a | \mathbf{Y}_t = \mathbf{y}_b) \varphi(\mathbf{y}_b) d\mathbf{y}_b. \quad (2.2)$$

That is, the distribution of \mathbf{Y}_{t+1} , obtained by marginalization over \mathbf{Y}_t , coincides with that of \mathbf{Y}_t . Put differently, the mixing by the transition kernel does not affect the distribution of individual random variables of the chain. To verify that a particular distribution $\varphi(\mathbf{y})$ is the stationary one it is sufficient to verify it satisfies the detailed balance equations:

$$g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{y}_b | \mathbf{y}_a) \varphi(\mathbf{y}_a) = g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{y}_a | \mathbf{y}_b) \varphi(\mathbf{y}_b),$$

for all choices of $\mathbf{y}_a, \mathbf{y}_b \in \mathbb{R}^p$. If a Markov chain satisfies this detailed balance equations, it is said to be *reversible*. Reversibility means so much as that, from the realizations, the direction of the chain cannot be discerned as: $f_{\mathbf{Y}_t, \mathbf{Y}_{t+1}}(\mathbf{y}_a, \mathbf{y}_b) = f_{\mathbf{Y}_t, \mathbf{Y}_{t+1}}(\mathbf{y}_b, \mathbf{y}_a)$, that is, the probability of starting in state \mathbf{y}_a and finishing in state \mathbf{y}_b equals that of starting in \mathbf{y}_b and finishing in \mathbf{y}_a . The sufficiency of this condition is evident after its integration on both sides with respect to \mathbf{y}_b , from which condition (2.2) follows.

MCMC assumes the stationary distribution of the Markov chain to be known up to a scalar – this is the target density from which is to be sampled – but the transition kernel is unknown. This poses a problem as the transition kernel is required to produce a sample from the target density. An arbitrary kernel is unlikely to satisfy the detailed balance equations with the desired distribution as its stationary one. If indeed it does not, then:

$$g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{y}_b | \mathbf{y}_a) \varphi(\mathbf{y}_a) > g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{y}_a | \mathbf{y}_b) \varphi(\mathbf{y}_b),$$

for some \mathbf{y}_a and \mathbf{y}_b . This may (loosely) be interpreted as that the process moves from \mathbf{y}_a to \mathbf{y}_b too often and from \mathbf{y}_b to \mathbf{y}_a too rarely. To correct this the probability of moving from \mathbf{y}_b to \mathbf{y}_b is introduced to reduce the number of moves from \mathbf{y}_a to \mathbf{y}_b . As a consequence not always a new value is generated, the algorithm may decide to stay in \mathbf{y}_a (as opposed to acceptance-rejection sampling). To this end a kernel is constructed and comprises compound transitions that propose a possible next state which is simultaneously judged to be acceptable or not. Hereto take an arbitrary *candidate-generating* density function $g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{y}_{t+1} | \mathbf{y}_t)$. Given the current state $\mathbf{Y}_t = \mathbf{y}_t$, this kernel produces a suggestion \mathbf{y}_{t+1} for the next point in the Markov chain. This suggestion may take the random walk too far afield from the desired and known stationary distribution. If so, the point is to be rejected in favour of the current state, until a new suggestion is produced that is satisfactory close to or representative of the stationary distribution. Let the probability of an acceptable suggestion \mathbf{y}_{t+1} , given the current state $\mathbf{Y}_t = \mathbf{y}_t$, be denoted by $\mathcal{A}(\mathbf{y}_{t+1} | \mathbf{y}_t)$. The thus constructed transition kernel then formalizes to:

$$h_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{y}_a | \mathbf{y}_b) = g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{y}_a | \mathbf{y}_b) \mathcal{A}(\mathbf{y}_a | \mathbf{y}_b) + r(\mathbf{y}_a) \delta_{\mathbf{y}_b}(\mathbf{y}_a),$$

where $\mathcal{A}(\mathbf{y}_a | \mathbf{y}_b)$ is the acceptance probability of the suggestion \mathbf{y}_a given the current state \mathbf{y}_b and is defined as:

$$\mathcal{A} = \min \left\{ 1, \frac{\varphi(\mathbf{y}_a)}{\varphi(\mathbf{y}_b)} \frac{g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{y}_b | \mathbf{y}_a)}{g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{y}_a | \mathbf{y}_b)} \right\}.$$

Furthermore, $\delta_{\mathbf{y}_b}(\cdot)$ is the Dirac delta function and $r(\mathbf{y}_a) = 1 - \int_{\mathbb{R}^p} g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{y}_a | \mathbf{y}_b) \mathcal{A}(\mathbf{y}_a | \mathbf{y}_b) d\mathbf{y}_a$, which is associated with the rejection. The transition kernel $h_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\cdot | \cdot)$ equipped with the above specified acceptance probability is referred to as the Metropolis-Hastings kernel. Note that, although the normalizing constant of the stationary distribution may be unknown, the acceptance probability can be evaluated as this constant cancels out in the $\varphi(\mathbf{y}_a) [\varphi(\mathbf{y}_b)]^{-1}$ term.

It rests now to verify that $\varphi(\cdot)$ is the stationary distribution of the thus defined Markov process. Hereto first the reversibility of the proposed kernel is assessed. The contribution of the second summand of the kernel to the detailed balance equations, $r(\mathbf{y}_a) \delta_{\mathbf{y}_b}(\mathbf{y}_a) \varphi(\mathbf{y}_b)$, exists only if $\mathbf{y}_a = \mathbf{y}_b$. Thus, if it contributes to the detailed balance equations, it does so equally on both sides of the equation. It is now left to assess whether:

$$g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{y}_a | \mathbf{y}_b) \mathcal{A}(\mathbf{y}_a | \mathbf{y}_b) \varphi(\mathbf{y}_b) = g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{y}_b | \mathbf{y}_a) \mathcal{A}(\mathbf{y}_b | \mathbf{y}_a) \varphi(\mathbf{y}_a).$$

To verify this equality use the definition of the acceptance probability from which we note that if $\mathcal{A}(\mathbf{y}_a | \mathbf{y}_b) \leq 1$ (and thus $\mathcal{A}(\mathbf{y}_a | \mathbf{y}_b) = [\varphi(\mathbf{y}_a) g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{y}_b | \mathbf{y}_a)][\varphi(\mathbf{y}_b) g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{y}_a | \mathbf{y}_b)]^{-1}$), then $\mathcal{A}(\mathbf{y}_b | \mathbf{y}_a) = 1$ (and vice

versa). In either case, the equality in the preceding display holds, and thereby, together with the observation regarding the second summand, so do the detailed balance equations for Metropolis-Hastings kernel. Then, $\varphi(\cdot)$ is indeed the stationary distribution of the constructed transition kernel $h_{\mathbf{Y}_{t+1} \mid \mathbf{Y}_t}(\mathbf{y}_a \mid \mathbf{y}_b)$ as can be verified from Condition (2.2):

$$\begin{aligned} \int_{\mathbb{R}^p} h_{\mathbf{Y}_{t+1} \mid \mathbf{Y}_t}(\mathbf{y}_a \mid \mathbf{y}_b) \varphi(\mathbf{y}_b) d\mathbf{y}_b &= \int_{\mathbb{R}^p} h_{\mathbf{Y}_{t+1} \mid \mathbf{Y}_t}(\mathbf{y}_b \mid \mathbf{y}_a) \varphi(\mathbf{y}_a) d\mathbf{y}_b \\ &= \int_{\mathbb{R}^p} g_{\mathbf{Y}_{t+1} \mid \mathbf{Y}_t}(\mathbf{y}_b \mid \mathbf{y}_a) \mathcal{A}(\mathbf{y}_b \mid \mathbf{y}_a) \varphi(\mathbf{y}_a) d\mathbf{y}_b + \int_{\mathbb{R}^p} r(\mathbf{y}_b) \delta_{\mathbf{y}_a}(\mathbf{y}_b) \varphi(\mathbf{y}_a) d\mathbf{y}_b \\ &= \varphi(\mathbf{y}_a) \int_{\mathbb{R}^p} g_{\mathbf{Y}_{t+1} \mid \mathbf{Y}_t}(\mathbf{y}_b \mid \mathbf{y}_a) \mathcal{A}(\mathbf{y}_b \mid \mathbf{y}_a) d\mathbf{y}_b + \varphi(\mathbf{y}_a) \int_{\mathbb{R}^p} r(\mathbf{y}_b) \delta_{\mathbf{y}_a}(\mathbf{y}_b) d\mathbf{y}_b \\ &= \varphi(\mathbf{y}_a)[1 - r(\mathbf{y}_a)] + r(\mathbf{y}_a) \varphi(\mathbf{y}_a) = \varphi(\mathbf{y}_a), \end{aligned}$$

in which the reversibility of $h_{\mathbf{Y}_{t+1} \mid \mathbf{Y}_t}(\cdot, \cdot)$, the definition of $r(\cdot)$, and the properties of the Dirac delta function have been used.

The Metropolis-Hastings algorithm is now described by:

```
input : densities  $f(\cdot)$  and  $h(\cdot)$ ;  
constant  $c > 0$ ;  
output: A draw from  $\pi(\cdot)$ .  
1 Choose starting value  $x^{(0)}$ ..  
2 Generate  $y$  from  $q(x^{(j)}, \cdot)$  and  $U$  from  $\mathcal{U}[0, 1]$ .  
3 if  $U \leq \alpha(X^{(j)}, y)$  then  
4   | set  $x^{(j+1)} = y$ .  
5 else  
6   | set  $x^{(j+1)} = x^{(j)}$ .  
7 end  
8 Return the values  $\{x^{(1)}, x^{(2)}, x^{(3)}, \dots, \}$ 
```

Algorithm 3: Metropolis-Hastings algorithm.

The convergence rate of this sequence is subject of on-going research.

The following R-script shows how the Metropolis-Hastings sampler can be used to from a mixture of two normals $f(y) = \theta \phi_{\mu=2, \sigma^2=0.5}(y) + (1 - \theta) \phi_{\mu=-2, \sigma^2=2}(y)$ with $\theta = 14$. The sampler assumes this distribution is known up to its normalizing constant and uses its unnormalized density $\exp[-(y - 2)^2] + 3 \exp[-(y + 2)^2/4]$. The sampler employs the Cauchy distribution centered at the current state as transition kernel from a candidate for the next state is drawn. The chain is initiated arbitrarily with $Y^{(1)} = 1$.

Listing 2.2 R code

```
# define unnormalized mixture of two normals
mixDensity <- function(x) {
  # unnormalized mixture of two normals
  exp(-(x-2)^2) + 3*exp(-(x+2)^2/4)
}

# define the MCMC sampler
mcmcSampler <- function(n, initDraw) {
  draw <- initDraw
  for (i in 1:(n-1)) {
    # sample a candidate
    candidate <- rcauchy(1, draw[i])

    # calculate acceptance probability:
    alpha <- min(1, mixDensity(candidate) / mixDensity(draw[i]) *
      dcauchy(draw[i], candidate) / dcauchy(candidate, draw[i]))
```

```

# accept/reject candidate
if (runif(1) < alpha) {
  draw <- c(draw, candidate)
} else {
  draw <- c(draw, draw[i])
}
}
return(draw)
}

# verify by eye-ballng
Y <- mcmcSampler(100000, 1)
hist(Y[-c(1:1000)], n=100)

```

The histogram (not shown) shows that the sampler, although using a unimodal kernel, yields a sample from the bimodal mixture distribution.

The *Gibbs sampler* is a particular version of the MCMC algorithm. The Gibbs sampler enhances convergence to the stationary distribution (i.e. the posterior distribution) of the Markov chain. It requires, however, the full conditionals of all random variables (here: model parameters), i.e. the conditional distributions of one random variable given all others, to be known analytically. Let the random vector \mathbf{Y} for simplicity – more refined partitions possible – be partitioned as $(\mathbf{Y}_a, \mathbf{Y}_b)$ with subscripts a and b now referring to index sets that partition the random vector \mathbf{Y} (instead of the previously employed meaning of referring to two – possibly different – elements from the state space). The Gibbs sampler thus requires that both $f_{\mathbf{Y}_a | \mathbf{Y}_b}(\mathbf{y}_a, \mathbf{y}_b)$ and $f_{\mathbf{Y}_b | \mathbf{Y}_a}(\mathbf{y}_a, \mathbf{y}_b)$ are known. Being a specific form of the MCMC algorithm the Gibbs sampler seeks to draw $\mathbf{Y}_{t+1} = (\mathbf{Y}_{a,t+1}, \mathbf{Y}_{b,t+1})$ given the current state $\mathbf{Y}_t = (\mathbf{Y}_{a,t}, \mathbf{Y}_{b,t})$. It draws, however, only a new instance for a single element of the partition (e.g. $\mathbf{Y}_{a,t+1}$) keeping the remainder of the partition (e.g. $\mathbf{Y}_{b,t}$) temporarily fixed. Hereto define the transition kernel:

$$g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{Y}_{t+1} = \mathbf{y}_{t+1} | \mathbf{Y}_t = \mathbf{y}_t) = \begin{cases} f_{\mathbf{Y}_{a,t+1} | \mathbf{Y}_{b,t+1}}(\mathbf{y}_{a,t+1}, \mathbf{y}_{b,t+1}) & \text{if } \mathbf{y}_{b,t+1} = \mathbf{y}_{b,t}, \\ 0 & \text{otherwise.} \end{cases}$$

Using the definition of the conditional density the acceptance probability for the $t + 1$ -th proposal of the subvector \mathbf{Y}_a then is:

$$\begin{aligned} \mathcal{A}(\mathbf{y}_{t+1} | \mathbf{y}_t) &= \min \left\{ 1, \frac{\varphi(\mathbf{y}_{t+1}) g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{y}_t, \mathbf{y}_{t+1})}{\varphi(\mathbf{y}_t) g_{\mathbf{Y}_{t+1} | \mathbf{Y}_t}(\mathbf{y}_{t+1}, \mathbf{y}_t)} \right\} \\ &= \min \left\{ 1, \frac{\varphi(\mathbf{y}_{t+1})}{\varphi(\mathbf{y}_t)} \frac{f_{\mathbf{Y}_{a,t+1} | \mathbf{Y}_{b,t+1}}(\mathbf{y}_{a,t}, \mathbf{y}_{b,t})}{f_{\mathbf{Y}_{a,t+1} | \mathbf{Y}_{b,t+1}}(\mathbf{y}_{a,t+1}, \mathbf{y}_{b,t+1})} \right\} \\ &= \min \left\{ 1, \frac{\varphi(\mathbf{y}_{t+1})}{\varphi(\mathbf{y}_t)} \frac{f_{\mathbf{Y}_{a,t+1}, \mathbf{Y}_{b,t+1}}(\mathbf{y}_{a,t}, \mathbf{y}_{b,t}) / f_{\mathbf{Y}_{b,t+1}}(\mathbf{y}_{b,t})}{f_{\mathbf{Y}_{a,t+1}, \mathbf{Y}_{b,t+1}}(\mathbf{y}_{a,t+1}, \mathbf{y}_{b,t+1}) / f_{\mathbf{Y}_{b,t+1}}(\mathbf{y}_{b,t+1})} \right\} \\ &= \min \left\{ 1, \frac{f_{\mathbf{Y}_{b,t+1}}(\mathbf{y}_{b,t+1})}{f_{\mathbf{Y}_{b,t+1}}(\mathbf{y}_{b,t})} \right\} = 1. \end{aligned}$$

The acceptance probability of each proposal is thus one (which contributes to the enhanced convergence of the Gibbs sampler to the joint posterior). Having drawn an acceptable proposal for this element of the partition, the Gibbs sampler then draws a new instance for the next element of the partition, i.e. now \mathbf{Y}_b , keeping \mathbf{Y}_a fixed. This process of subsequently sampling each partition element is repeated until enough samples have been drawn.

To illustrate the Gibbs sampler revisit Bayesian regression. In Section 2.2 the full conditional distributions of β and σ^2 were derived. The Gibbs sampler now draws in alternating fashion from these conditional distributions (see Algorithm 4 for its pseudo-code).

```

input : sample size  $T$   

length of burn in period  $T_{\text{burn-in}}$   

thinning factor  $f_{\text{thinning}}$   

data  $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ ;  

conditional distributions  $f_{\sigma^2 | \beta}(\sigma^2, \beta; \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n)$  and  $f_{\beta | \sigma^2}(\beta, \sigma^2; \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n)$ .  

output: draws from the joint posterior  $\pi(\beta, \sigma^2 | \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n)$ .  

1 initialize  $(\beta_0, \sigma_0^2)$ .  

2 for  $t = 1$  to  $T_{\text{burn-in}} + T f_{\text{thinning}}$  do  

3   draw  $\beta_t$  from conditional distribution  $f_{\beta | \sigma^2}(\beta, \sigma_t^2; \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n)$ ,  

4   draw  $\sigma_t^2$  from conditional distribution  $f_{\sigma^2 | \beta}(\sigma^2, \beta_t; \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n)$ .  

5 end  

6 Remove the first  $T_{\text{burn-in}}$  draws (representing the burn-in phase).  

7 Select every  $f_{\text{thinning}}$ -th sample (thinning).

```

Algorithm 4: Pseudocode of the Gibbs sampler of the joint posterior of the Bayesian regression parameters.

2.4 Empirical Bayes

Empirical Bayes (EB) is a branch of Bayesian statistics that meets the subjectivity criticism of frequentists. Instead of fully specifying the prior distribution empirical Bayesians identify only its form. The hyper parameters of this prior distribution are left unspecified and are to be found empirically. In practice, these hyper-parameters are estimated from the data at hand. However, the thus estimated hyper parameters are used to obtain the Bayes estimator of the model parameters. As such the data are then used multiple times. This is usually considered an inappropriate practice but is deemed acceptable when the number of model parameters is large in comparison to the number of hyper parameters. Then, the data are not used twice but ‘ $1 + \varepsilon$ ’-times (i.e. once-and-a-little-bit) and only little information from the data is spent on the estimation of the hyper parameters. Having obtained an estimate of the hyper parameters, the prior is fully known, and the posterior distribution (and summary statistics thereof) are readily obtained by Bayes’ formula.

The most commonly used procedure for the estimation of the hyper parameters is marginal likelihood maximization, which is a maximum likelihood-type procedure. But the likelihood cannot directly be maximized with respect to the hyper parameters as it contains the model parameters that are assumed to be random within the Bayesian framework. This may be circumvented by choosing a specific value for the model parameter but would render the resulting hyper parameter estimates dependent on this choice. Instead of maximization with the model parameter set to a particular value one would preferably maximize over all possible realizations. The latter is achieved by marginalization with respect to the random model parameter, in which the (prior) distribution of the model parameter is taken into account. This amounts to integrating out the model parameter from the posterior distribution, i.e. $\int P(\mathbf{Y} = \mathbf{y} | \theta) \pi(\theta) d\theta$, resulting in the so-called marginal posterior. After marginalization the specifics of the model parameter have been discarded and the marginal posterior is a function of the observed data and the hyper parameters. The estimator of the hyper parameter is now defined as the maximizer of this marginal posterior.

To illustrate the estimation of hyper parameters of the Bayesian linear regression model through marginal likelihood maximization assume the regression parameter β and the error variance σ^2 to be endowed with conjugate priors: $\beta | \sigma^2 \sim \mathcal{N}(\mathbf{0}_p, \sigma^2 \lambda^{-1} \mathbf{I}_{pp})$ and $\sigma^2 \sim \mathcal{IG}(\alpha_0, \beta_0)$. Three hyper parameters are thus to be estimated: the shape and scale parameters, α_0 and β_0 , of the inverse gamma distribution and the λ parameter related to the variance of the regression coefficients. Straightforward application of the outlined marginal likelihood principle does, however, not work here. The joint prior, $\pi(\beta | \sigma^2) \pi(\sigma^2)$, is too flexible and does not yield sensible estimates of the hyper parameters. As interest is primarily in λ , this is resolved by setting the hyper parameters of $\pi(\sigma^2)$ such that the resulting prior is uninformative, i.e. as objectively as possible. This is operationalized as a very flat distribution. Then, with the particular choices of α_0 and β_0 that produce an uninformative prior for σ^2 ,

the empirical Bayes estimate of λ is:

$$\begin{aligned}
\hat{\lambda}_{eb} &= \arg \max_{\lambda} \int_0^{\infty} \int_{\mathbb{R}^p} f_{\beta, \sigma^2}(\beta, \sigma^2 | \mathbf{Y}, \mathbf{X}) d\beta d\sigma^2 \\
&= \arg \max_{\lambda} \int_0^{\infty} \int_{\mathbb{R}^p} \sigma^{-n} \exp[-\frac{1}{2}\sigma^{-2}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)] \\
&\quad \times \sigma^{-p} \exp(-\frac{1}{2}\sigma^{-2}\lambda\beta^T\beta) \times (\sigma^2)^{-\alpha_0-1} \exp(-\beta_0\sigma^{-2}) d\beta d\sigma^2 \\
&= \arg \max_{\lambda} \int_0^{\infty} \int_{\mathbb{R}^p} \sigma^{-n} \exp\{-\frac{1}{2}\sigma^{-2}[\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T\mathbf{Y}]\} \\
&\quad \times \sigma^{-p} \exp\{-\frac{1}{2}\sigma^{-2}[\beta - \hat{\beta}(\lambda)]^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})[\beta - \hat{\beta}(\lambda)]\} \\
&\quad \times (\sigma^2)^{-\alpha_0-1} \exp(-\beta_0\sigma^{-2}) d\beta d\sigma^2 \\
&= \arg \max_{\lambda} \int_0^{\infty} \sigma^{-n} \exp\{-\frac{1}{2}\sigma^{-2}[\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T\mathbf{Y}]\} \\
&\quad \times |\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp}|^{-1/2} (\sigma^2)^{-\alpha_0-1} \exp(-\beta_0\sigma^{-2}) d\sigma^2 \\
&= \arg \max_{\lambda} |\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp}|^{-1/2} \int_0^{\infty} \exp(-\sigma^{-2}\{\beta_0 + \frac{1}{2}[\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T\mathbf{Y}]\}) \\
&\quad \times (\sigma^2)^{-\alpha_0-n/2-1} d\sigma^2 \\
&= \arg \max_{\lambda} |\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp}|^{-1/2} b_1^{-\alpha_0-n/2},
\end{aligned}$$

where the factors not involving λ have been dropped throughout and $b_1 = \beta_0 + \frac{1}{2}[\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T\mathbf{Y}]$. Prior to the maximization of the marginal likelihood the logarithm is taken. That changes the maximum, but not its location, and yields an expression that is simpler to maximize. With the empirical Bayes estimate $\hat{\lambda}_{eb}$ at hand, the Bayesian estimate of the regression parameter β is $\hat{\beta}_{eb} = (\mathbf{X}^T\mathbf{X} + \hat{\lambda}_{eb}\mathbf{I}_{pp})^{-1}\mathbf{X}^T\mathbf{Y}$. Finally, the particular choice of the hyper parameters of the prior on σ^2 is not too relevant. Most values of α_0 and β_0 that correspond to a rather flat inverse gamma distribution yield resulting point estimates $\hat{\beta}_{eb}$ that do not differ too much numerically.

2.5 Conclusion

Bayesian regression was introduced and shown to be closely connected to ridge regression. Under a conjugate Gaussian prior on the regression parameter the Bayesian regression estimator coincides with the ridge regression estimator, which endows the ridge penalty with the interpretation of this prior. While an analytic expression of these estimators is available, a substantial part of this chapter was dedicated to evaluation of the estimator through resampling. The use of this resampling will be evident when other penalties and non-conjugate priors will be studied (cf. Exercise ?? and Sections 5.7 and 6.7). Finally, another informative procedure, empirical Bayes, to choose the penalty parameter was presented.

2.6 Exercises

Question 2.1

Consider the linear regression model $Y_i = X_i\beta + \varepsilon_i$ with the ε_i i.i.d. following a standard normal law $\mathcal{N}(0, 1)$. Data on the response and covariate are available: $\{(y_i, x_i)\}_{i=1}^8 = \{(-5, -2), (0, -1), (-4, -1), (-2, -1), (0, 0), (3, 1), (5, 2), (3, 2)\}$.

- a) Assume a zero-centered normal prior on β . What variance, i.e. which $\sigma_{\beta}^2 \in \mathbb{R}_{>0}$, of this prior yields a mean posterior $\mathbb{E}(\beta | \{(y_i, x_i)\}_{i=1}^8, \sigma_{\beta}^2)$ equal to 1.4?
- b) Assume a non-zero centered normal prior. What (mean, variance)-combinations for the prior will yield a mean posterior estimate $\hat{\beta} = 2$?

Question 2.2

Consider the Bayesian linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_{nn})$ and priors $\beta | \sigma^2 \sim \mathcal{N}(\mathbf{0}_p, \sigma_{\beta}^2\mathbf{I}_{pp})$ and $\sigma^2 \sim \mathcal{IG}(a_0, b_0)$ where $\sigma_{\beta}^2 = c\sigma^2$ for some $c > 0$ and a_0 and b_0 are the shape and scale

parameters, respectively, of the inverse Gamma distribution. This model is fitted to data from a study where the response is explained by a single covariate, and henceforth β is replaced by β , with the following relevant summary statistics: $\mathbf{X}^\top \mathbf{X} = 2$ and $\mathbf{X}^\top \mathbf{Y} = 5$.

- Suppose $\mathbb{E}(\beta | \sigma^2 = 1, c, \mathbf{X}, \mathbf{Y}) = 2$. What amount of regularization should be used such that the ridge regression estimate $\hat{\beta}(\lambda_2)$ coincides with the aforementioned posterior (conditional) mean?
- Give the (posterior) distribution of $\beta | \{\sigma^2 = 2, c = 2, \mathbf{X}, \mathbf{Y}\}$.
- Discuss how a different prior of σ^2 affects the correspondence between $\mathbb{E}(\beta | \sigma^2, c, \mathbf{X}, \mathbf{Y})$ and the ridge regression estimator.

Question 2.3

Revisit the microRNA data example of Section 1.10. Use the empirical Bayes procedure outlined in Section 2.4 to estimate the penalty parameter. Compare this to the one obtained via cross-validation. In particular, compare the resulting point estimates of the regression parameter.

Question 2.4

Revisit question 1.16. From a Bayesian perspective, is the suggestion of a negative ridge penalty parameter sensible?

Question 2.5

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\mathbf{X} \in \mathcal{M}^{n,p}$, $\beta \in \mathbb{R}^p$, and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{nn})$. Assume the β_j are independently and identically distributed with a generalized normal prior distribution. The latter has density function $f(x; \mu, \alpha_1, \alpha_2) = \alpha_2 [2\alpha_1 \Gamma(\alpha_2^{-1})]^{-1} \exp[-(|x - \mu|/\alpha_1)^{\alpha_2}]$ with location parameter μ , scale parameter α_1 and shape parameter α_2 . For which choice of these hyperparameter does the MAP estimator coincide with the bridge regression one? The *bridge regression estimator* of β is defined as:

$$\hat{\beta}(\lambda_\gamma) = \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_\gamma \sum_{j=1}^p |\beta_j|^\gamma,$$

with penalty parameter λ_b and shrinkage parameter $\gamma > 0$.

Question 2.6

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ without intercept and $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$, to explain the variation in the response \mathbf{Y} by a linear combination of the columns of the design matrix \mathbf{X} . Execute the R-code below to sample data from this model.

Listing 2.3 R code

```
# set dimension and sample size
n <- 50
p <- 100

# create covariates
X <- matrix(rnorm(n*p), ncol=p)

# create coefficients and draw response
betas <- matrix(seq(-2, 2, length.out=p), ncol=1)
betas <- (betas - min(betas)+0.001)
betas <- qnorm(betas / (max(betas)+0.001)) / 5
Y <- X %*% betas + rnorm(n, sd=1/4)
```

With these sampled data, find a Bayesian estimate, the posterior mean, of the regression parameter. Hereto adopt the following priors for the parameters: $\beta | \sigma^2 \sim \mathcal{N}(\mathbf{0}_p, \sigma^2 \lambda \mathbf{I}_{pp})$ with $\lambda = 1$ and $\sigma^2 \sim \mathcal{IG}(a_0, b_0)$ with $a_0 = 1 = b_0$.

- Build a Gibbs sampler (Algorithm 3 in the lecture notes). Initiate the sampler with $\beta_0 = \mathbf{0}_p$ and $\sigma_0^2 = 1$. The sampler then iterates between drawing new β and σ^2 , respectively, from their conditional posteriors:

$$\begin{aligned} \beta | \sigma_t^2, \mathbf{Y}, \mathbf{X} &\sim \mathcal{N}[(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y}, \sigma_t^2 (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}], \\ \sigma^2 | \mathbf{Y}, \mathbf{X} &\sim \mathcal{IG}(a, b), \end{aligned}$$

with shape parameter $a = a_0 + \frac{1}{2}(n + p)$ and rate parameter $b = b_0 + \frac{1}{2}\|\mathbf{Y}\|_2^2 - \frac{1}{2}\|\mathbf{X}\hat{\beta}(\lambda)\|_2^2 - \frac{1}{2}\lambda\|\hat{\beta}(\lambda)\|_2^2$ and in which t represents the index for the iteration. To sample from both distributions use the `rmvnorm`- and the `rivgamma`-functions.

- b) Use the Gibbs sampler to draw 10100 times from both conditional posteriors in alternating fashion. To remove the dependency on the choice of the initiation σ_0^2 throw away the first 100 (i.e. the burn-in period) draws for both parameters. To reduce the dependency between subsequent draws, apply thinning: keep each 10-th draw after the burn-in period. After both operations (removal of the burning-in period and the thinning) only those draws corresponding to iterations 110, 120, ..., 10100 are preserved. Those are considered a representative sample from the joint posterior of β and σ^2 .
- c) Calculate the lower bound of the 95% credible interval, containing the central $(100 - \alpha)\%$ with $\alpha = 0.05$ of the posterior probability mass, of the second element of the regression parameter. Use the quantile-function for the credible interval construction.
- d) Investigate the dependency among subsequent draws, i.e. without the thinning, of σ^2 from the Gibbs sampler. In this contrast the case with $\lambda = 1$ to that with $\lambda = 10^6$. Is there a difference in their 1st order dependencies? If so, explain this difference. If not, explain the absence thereof.

3 Generalizing ridge regression

The exposé on ridge regression may be generalized in many ways. Among others different generalized linear models may be considered (confer Section 5.3). In this section we stick to the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with the usual assumptions. But we now fit it in weighted fashion – to accommodate the ridge estimation of the logistic regression model (see Chapter 5) – and generalize the common, spherical penalty.

Our generalization of the ridge loss function, a weighted least squares criterion augmented with a generalized ridge penalty, is:

$$(\mathbf{Y} - \mathbf{X}\beta)^T \mathbf{W} (\mathbf{Y} - \mathbf{X}\beta) + (\beta - \beta_0)^T \Delta (\beta - \beta_0). \quad (3.1)$$

In the above display \mathbf{W} is a $n \times n$ -dimensional, diagonal matrix with $(\mathbf{W})_{ii} \in [0, 1]$ representing the weight of the i -th observation. The minimizer of loss function (3.2) is our generalization of the ridge regression estimator.

The generalized ridge penalty in loss function (3.2) is now a quadratic form with penalty parameter Δ , a $p \times p$ -dimensional, positive definite, symmetric matrix. When $\Delta = \lambda \mathbf{I}_{pp}$, one regains the spherical penalty of ‘regular ridge regression’. This penalty shrinks each element of the regression parameter β equally along the unit vectors \mathbf{e}_j . Generalizing Δ to the class of symmetric, positive definite matrices \mathcal{S}_{++} allows for i) different penalization per regression parameter, and ii) joint (or correlated) shrinkage among the elements of β . The penalty parameter Δ determines the speed and direction of shrinkage. The p -dimensional column vector β_0 is a user-specified, non-random target towards which β is shrunken as the penalty parameter increases. When recasting generalized ridge estimation as a constrained estimation problem, the implications of the penalty may be visualized (Figure 3.1, left panel). The generalized ridge penalty is a quadratic form centered around β_0 . In Figure 3.1 the parameter constraint clearly is ellipsoidal (and not spherical). Moreover, the center of this ellipsoid is not at zero.

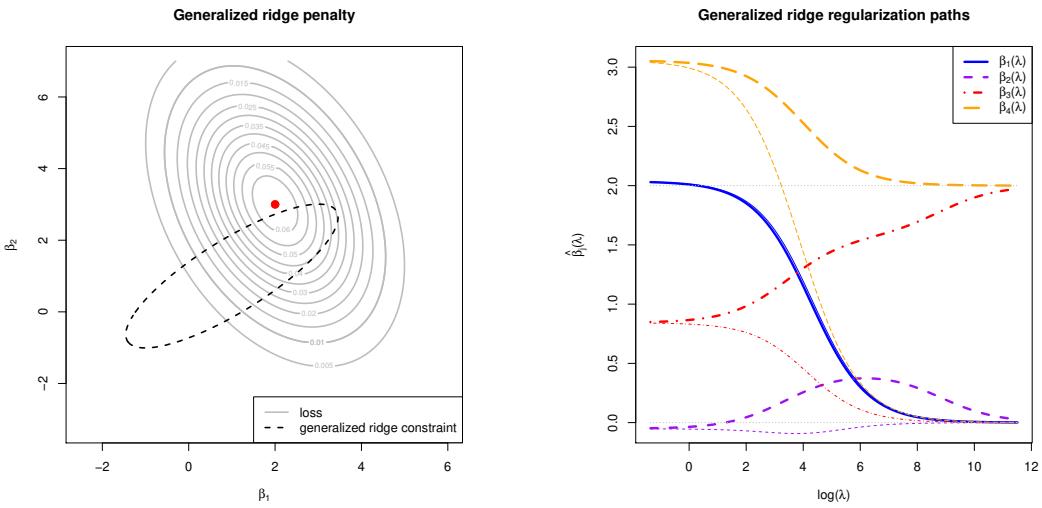


Figure 3.1: Left panel: the contours of the likelihood (grey solid ellipsoids) and the parameter constraint implied by the generalized penalty (black dashed ellipsoids) Right panel: generalized (fat coloured lines) and ‘regular’ (thin coloured lines) regularization paths of four regression coefficients. The dotted grey (straight) lines indicated the targets towards the generalized ridge penalty shrinks regression coefficient estimates.

The addition of the generalized ridge penalty to the sum-of-squares ensures the existence of a unique regression estimator in the face of super-collinearity. The generalized penalty is a non-degenerated quadratic form in β due

to the positive definiteness of the matrix Δ . As it is non-degenerate, it is strictly convex. Consequently, the generalized ridge regression loss function (3.1), being the sum of a convex and strictly convex function, is also strictly convex. This warrants the existence of a unique global minimum and, thereby, a unique estimator.

There is an analytic expression for the optimum of the generalized ridge loss function (3.1). To see this, obtain the estimating equation of β through equating its derivative with respect to β to zero:

$$2\mathbf{X}^\top \mathbf{WY} - 2\mathbf{X}^\top \mathbf{WX}\beta - 2\Delta\beta + 2\Delta\beta_0 = \mathbf{0}_p.$$

This is solved by:

$$\hat{\beta}(\Delta) = (\mathbf{X}^\top \mathbf{WX} + \Delta)^{-1}(\mathbf{X}^\top \mathbf{WY} + \Delta\beta_0). \quad (3.2)$$

Clearly, this reduces to the ‘regular’ ridge regression estimator by setting $\mathbf{W} = \mathbf{I}_{nn}$, $\beta_0 = \mathbf{0}_p$, and $\Delta = \lambda \mathbf{I}_{pp}$. The effects of the generalized ridge penalty on the coefficients of corresponding estimator can be seen in the regularization paths of the estimator’s coefficients. Figure 3.1 (right panel) contains an example of the regularization paths for coefficients of a linear regression model with four explanatory variables. Most striking is the limiting behaviour of the estimates of β_3 and β_4 for large values of the penalty parameter λ : they converge to a non-zero value (as was specified by a nonzero β_0). More subtle is the (temporary) convergence of the regularization paths of the estimates of β_2 and β_3 . That of β_2 is pulled away from zero (its true value and approximately its unpenalized estimate) towards the estimate of β_3 . In the regularization path of β_3 this can be observed in a delayed convergence to its nonzero target value (for comparison consider that of β_4). For reference the corresponding regularization paths of the ‘regular’ ridge estimates (as thinner lines of the same colour) are included in Figure 3.1.

One particular version of the generalized ridge penalty deserves special attention. It distinguishes between penalized and unpenalized covariates. The latter comprises explanatory variables that ought to be in the model, and are not to be shrunken. This aims to keep the bias of their estimates to a minimum and make them comparable to these reported in the literature. For instance, in the statistical analysis of clinical trials factors like ‘age’ and ‘gender’ are known to associated with the response, and any model without them would not be acceptable to the field. Additionally, the patients of the clinical trial may have been molecularly characterized at baseline. The resulting high-dimensional molecular information may then be included in the model, alongside the aforementioned factors. The effect of the high-dimensional molecular covariates can then only be estimated in penalized fashion, while that of ‘age’ and ‘gender’ are preferably estimated in an unpenalized manner.

To present our estimator with penalized and unpenalized covariates, we introduce separate notation for them. This modifies the model to $\mathbf{Y} = \mathbf{U}\gamma + \mathbf{X}\beta + \varepsilon$, where \mathbf{U} is a full rank $n \times q$ -dimensional design matrix of the q unpenalized covariates and γ the associated regression parameter. The ridge regression estimator is then generalized to:

$$\hat{\gamma}(\Delta, \beta_0), \hat{\beta}(\Delta, \beta_0) = \arg \min_{\gamma \in \mathbb{R}^q, \beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{U}\gamma - \mathbf{X}\beta\|_2^2 + (\beta - \beta_0)^\top \Delta(\beta - \beta_0).$$

Note that the estimator of regression parameter γ of the unpenalized covariates \mathbf{U} too depends on the penalty parameters. This is due to the fact that it is jointly estimated with the regression parameter β of their penalized counterparts \mathbf{X} . This bears consequence for the bias of the estimate of γ (see Exercise ??). Again an analytic expression of the estimator exists:

$$\begin{pmatrix} \hat{\gamma}(\Delta, \beta_0) \\ \hat{\beta}(\Delta, \beta_0) \end{pmatrix} = \begin{pmatrix} \mathbf{U}^\top \mathbf{U} & \mathbf{U}^\top \mathbf{X} \\ \mathbf{X}^\top \mathbf{U} & \mathbf{X}^\top \mathbf{X} + \Delta \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{U}^\top \mathbf{Y} \\ \mathbf{X}^\top \mathbf{Y} + \Delta\beta_0 \end{pmatrix}.$$

This expression can be ‘simplified’, as is done in (Lettink *et al.*, 2023), to:

$$\begin{aligned} \hat{\gamma}(\Delta, \beta_0) &= \{\mathbf{U}^\top [\mathbf{X}\Delta^{-1}\mathbf{X}^\top + \mathbf{I}_{nn}]^{-1}\mathbf{U}\}^{-1}\mathbf{U}^\top [\mathbf{X}\Delta^{-1}\mathbf{X}^\top + \mathbf{I}_{nn}]^{-1}(\mathbf{Y} - \mathbf{X}\beta_0), \\ \hat{\beta}(\Delta, \beta_0) &= \beta_0 + (\mathbf{X}^\top \mathbf{X} + \Delta)^{-1}\mathbf{X}^\top [\mathbf{Y} - \mathbf{U}\hat{\gamma}(\Delta, \beta_0) - \mathbf{X}\beta_0], \end{aligned}$$

which is obtained by means of the Woodbury matrix identity and the analytic expression of the inverse of a 2×2 block matrix.

Example 3.1 (Fused ridge estimation)

An example of a generalized ridge penalty is the *fused ridge penalty* (as introduced by Goeman, 2008). Consider the standard linear model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$. The fused ridge regression estimator of β then minimizes:

$$\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=2}^p (\beta_j - \beta_{j-1})^2. \quad (3.3)$$

The penalty in the loss function above can be written as a generalized ridge penalty:

$$\lambda \sum_{j=2}^p (\beta_j - \beta_{j-1})^2 = \boldsymbol{\beta}^\top \begin{pmatrix} \lambda & -\lambda & 0 & \dots & \dots & 0 \\ -\lambda & 2\lambda & -\lambda & \ddots & & \vdots \\ 0 & -\lambda & 2\lambda & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & -\lambda \\ 0 & \dots & \dots & 0 & -\lambda & \lambda \end{pmatrix} \boldsymbol{\beta}.$$

The matrix Δ employed above is semi-positive definite and therefore the loss function (3.3) need not be strictly convex. Hence, often a regular ridge penalty $\|\boldsymbol{\beta}\|_2^2$ is added (with its own penalty parameter).

To illustrate the effect of the fused ridge penalty on the estimation of the linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, let $\beta_j = \phi_{0,1}(z_j)$ with $z_j = -30 + \frac{6}{50}j$ for $j = 1, \dots, 500$. Sample the elements of the design matrix \mathbf{X} and those of the error vector $\boldsymbol{\varepsilon}$ from the standard normal distribution, then form the response \mathbf{Y} from the linear model. The regression parameter is estimated through fused ridge loss minimization with $\lambda = 1000$. The estimate is shown in the left panel of Figure 3.2 (red line). For reference the figure includes the true $\boldsymbol{\beta}$ (black line) and the ‘regular ridge’ estimate with $\lambda = 1$ (blue line). Clearly, the fused ridge regression estimate yields a nice smooth vector of $\boldsymbol{\beta}$ estimates.

The fused ridge penalty employed in the fused ridge loss function (3.3) shrinks one-dimensionally. It shrinks, when viewing the β_j as equidistantly distributed on a line ordered by their position in the regression parameter, contiguous elements towards one other. Should we expect the covariates’ effects to be similar, i.e. of equal size and sign, depending on their spatial proximity on a two-dimensional grid, this may be incorporated in the two-dimensional fused ridge regression estimator (Lettink *et al.*, 2023). \square

Example 3.2 (A ‘ridge to homogeneity’)

A different form of fusion proposed by Anatolyev (2020) shrinks (a subset of) the regression parameters to a common value. In the work of Anatolyev (2020) this common value is the mean of these regression parameters. This mean is also learned from data, alongside to individual elements of regression parameter. Hereto the sum-of-squares is augmented by the following penalty:

$$\lambda \boldsymbol{\beta}^\top (\mathbf{I}_{pp} - p^{-1} \mathbf{1}_{pp}) \boldsymbol{\beta} = \lambda \sum_{j=1}^p [\beta_j - p^{-1} \sum_{j'=1}^p \beta_{j'}]^2. \quad (3.4)$$

Straightforward application of this penalty within the context of the linear regression model may seem farfetched. That is, why would a common effect of all covariates be desirable? Indeed, the motivation of Anatolyev (2020) stems from a different context. Translated to the present standard linear regression model context, that motivation could be thought of – loosely – as an application of the penalty (3.4) to subsets of the regression parameter. Situations arise where many covariates are only slightly different operationalizations of the same trait. For instance, in brain image data the image itself is often summarized in many statistics virtually measuring the same thing. In extremis, those could be the mean, median and trimmed mean of the intensity of a certain region of the image. It would be ridiculous to assume these three summary statistics to have a wildly different effect, and shrinkage to a common value seems sensible practice. Finally, note that the above employed penalty matrix, which we write as $\lambda \Delta := \lambda (\mathbf{I}_{pp} - p^{-1} \mathbf{1}_{pp})$, is nonnegative definite. Hence, in the face of (super-) collinearity, an additional penalty term is required for a well-defined estimator.

The homogeneity ridge regression estimator can – but is left as an exercise – be reformulated using straightforward linear algebra as

$$\hat{\boldsymbol{\beta}}(\lambda) = a \mathbf{1}_p + (\Delta \mathbf{X}^\top \mathbf{X} \Delta + \lambda \Delta)^+ \mathbf{X}^\top (\mathbf{Y} - a \mathbf{X} \mathbf{1}_p).$$

with $a = [\mathbf{Y}^\top (\lambda \mathbf{I}_{nn} + \mathbf{X} \Delta \mathbf{X}^\top)^{-1} \mathbf{X} \mathbf{1}_p] [\mathbf{1}_p^\top \mathbf{X}^\top (\lambda \mathbf{I}_{nn} + \mathbf{X} \Delta \mathbf{X}^\top)^{-1} \mathbf{X} \mathbf{1}_p]^{-1}$. It follows that $\lim_{\lambda \rightarrow \infty} \hat{\boldsymbol{\beta}}(\lambda) = a \mathbf{1}_p$. That is, all elements are shrunken to the same value as the amount of regularization increases.

We illustrate the effect of the ‘homogeneity’ penalty (3.4). We sample from the linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\beta} = \mathbf{1}_{50}$ and both the elements of the design matrix as well as the error sampled from the standard normal distribution. We then evaluate the regularization path of the estimator (3.2) with Δ a diagonal 2×2 block matrix. The first diagonal block $\Delta_{11} = \lambda (\mathbf{I}_{10,10} - \frac{1}{10} \mathbf{1}_{10,10})$ and the second $\Delta_{22} = \lambda \mathbf{I}_{40,40}$. The resulting regularization paths are shown in the top right panel of Figure 3.2. \square

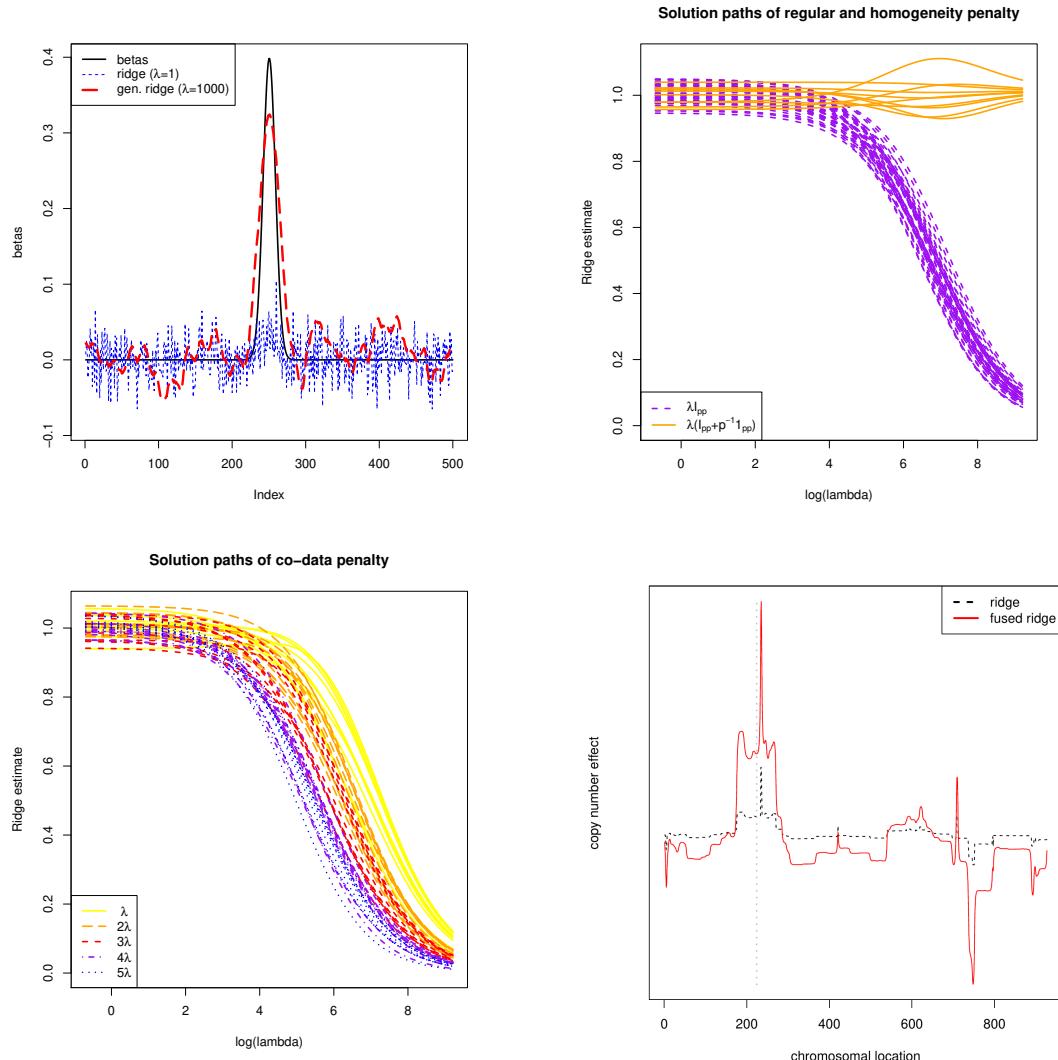


Figure 3.2: Top left panel: illustration of the fused ridge estimator (in simulation). The true parameter β and its ridge and fused ridge estimates against their spatial order. Top right panel: regularization path of the ridge regression estimator with a ‘homogeneity’ penalty matrix. Bottom left panel: regularization path of the ridge regression estimator with a ‘co-data’ penalty matrix. Bottom right panel: Ridge vs. fused ridge estimates of the DNA copy effect on KRAS expression levels. The dashed, grey vertical bar indicates the location of the KRAS gene.

Example 3.3 (Co-data)

Another utilization of the generalized ridge regression estimator (3.2) can be found in applications where groups of covariates are deemed to be differentially important for the explanation of the response. Van De Wiel *et al.* (2016) suggests to form these groups on the basis of *co-data*. The co-data concept refers to auxiliary data on the covariates not necessary directly related to but possibly implicitly informative for the to-be-estimated model. For instance, for each covariate a marginal p -value or effect estimate from a different study may be available. The covariate groups may then be formed on the basis of the size of this statistic. Alternatively, co-data may comprise some form of biological annotation, e.g. pathway membership of genes. Irrespectively, a shared group membership of covariates is indicative of an equal (implicit) relevance for model fitting. The possible importance differences among the covariate groups are accommodated through the augmentation of the loss by a sum of group-wise regular ridge penalties. That is, the elements of the regression parameter that correspond to the covariates of the same group are penalized by the sum of the square of these elements multiplied by a common, group-specific penalty parameter. To formalize this, let $\mathcal{J}_1, \dots, \mathcal{J}_K$ be K mutually exclusive and exhaustive subsets of the

covariate index set $\{1, \dots, p\}$, i.e. $\mathcal{J}_k \cap \mathcal{J}_{k'} = \emptyset$ for all $k \neq k'$ and $\cup_{k=1}^K \mathcal{J}_k = \{1, \dots, p\}$. The estimator then is:

$$\hat{\beta}(\lambda_1, \dots, \lambda_K) = \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \sum_{k=1}^K \lambda_k \sum_{j \in \mathcal{J}_k} \beta_j^2, \quad (3.5)$$

To express to above in terms the estimator 3.2, it involves $\mathbf{W} = \mathbf{I}_{nn}$, $\beta_0 = \mathbf{0}_p$, and a diagonal Δ . A group's importance is reflected in the size of the penalty parameter λ_k , relative to the other penalty parameters. The larger a group's penalty parameter, the smaller the ridge constraint of the corresponding elements of the regression parameter. And a small constraint allows these elements little room to vary, and the corresponding covariates little flexibility to accommodate the variation in the response. For instance, suppose the group index is concordant with the hypothesized importance of the covariates for the linear model. Ideally, the λ_k are then reciprocally concordant to the group index. Of course, the penalty parameters need not adhere to this concordance, as they are selected by means of data. For more on their selection in see Van De Wiel *et al.* (2016). Clearly, the first regression coefficients of the first ten covariates are shrunken towards a common value, while the others are all shrunken towards zero.

We illustrate the effect of the ‘co-data’ penalty on the same data as that used to illustrate the effect of the ‘homogeneity’ penalty (3.4). Now we employ a diagonal 5×5 block penalty matrix Δ , with blocks $\Delta_{11} = \lambda \mathbf{I}_{10,10}$, $\Delta_{12} = 2\lambda \mathbf{I}_{10,10}$, $\Delta_{13} = 3\lambda \mathbf{I}_{10,10}$, $\Delta_{14} = 4\lambda \mathbf{I}_{10,10}$, and $\Delta_{15} = 5\lambda \mathbf{I}_{10,10}$. The regularization paths of the estimator (3.5) are shown in the bottom left panel of Figure 3.2. Clearly, the regression coefficient estimates are shrunken less if they belong to a group with high important (i.e. lower penalty). \square

Example 3.4 (Updating)

An example that illustrates the use of the target of the generalized ridge penalty can be found in sequential learning (as is pointed out in van Wieringen and Binder, 2022). Sequential learning takes the lessons learned, i.e. estimates from previous studies into the same phenomenon, into account when updating these estimates from a data of a novel such study. The setting may be formalized by assuming

- A1) an infinite sequence of data sets $\{\mathbf{Y}^{(t)}, \mathbf{X}^{(t)}\}_{t=1}^\infty$ sampled from the linear regression model $\mathbf{Y}^{(t)} = \mathbf{X}^{(t)}\beta + \varepsilon^{(t)}$ with $\varepsilon^{(t)} \sim \mathcal{N}(\mathbf{0}_{n_t}, \sigma^2 \mathbf{I}_{n_t, n_t})$ for all $t = 1, \dots, T$ and $\varepsilon^{(t)} \perp \varepsilon^{(t')}$ for $t \neq t'$.

Effectively, the linear regression model is the same for all data sets as its parameters β and σ^2 are common to all. The number of covariates is fixed across data sets but the covariates' settings, i.e. their values, may differ among the data sets. Similarly, the sample sizes $n_t \in \mathbb{N}$ need not be common to all data sets.

We now estimate the regression parameter from each data set by the generalized ridge regression estimator using the estimate of the preceding data set as the target. Formally,

$$\hat{\beta}^{(t)}(\lambda_t) = \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{Y}^{(t)} - \mathbf{X}^{(t)}\beta\|_2^2 + \lambda_t \|\beta - \hat{\beta}^{(t-1)}(\lambda_{t-1})\|_2^2 \quad (3.6)$$

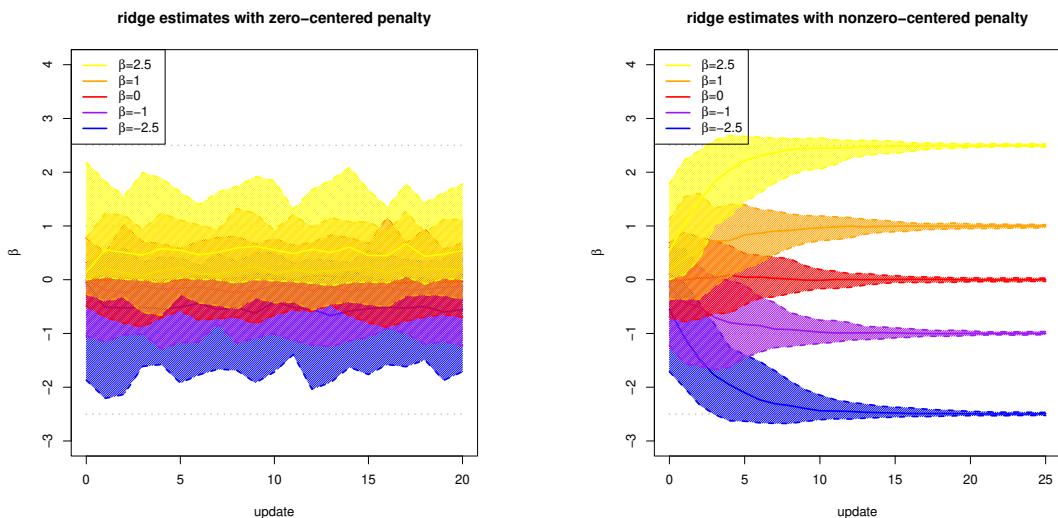


Figure 3.3: The top panels show the (5%, 95%-quantile intervals of the traditional (left) and updated (right) ridge estimates of β_j with $j \in \{0, 30, 50, 70, 100\}$ plotted against t . The solid, colored line inside these intervals is the corresponding 50% quantile. The dotted, grey lines are the true values of the β_j 's.

for $t = 1, \dots$. Effectively, with the arrival of novel data we update our current estimator of the regression parameter $\hat{\beta}^{(t-1)}(\lambda_{t-1})$, resulting in the updated one $\hat{\beta}^{(t)}(\lambda_t)$. We thus obtain a sequence of estimators $\{\hat{\beta}^{(t)}(\lambda_t)\}_{t=1}^{\infty}$, which is initiated by any nonrandom target, e.g. the null vector. Over time we expect this sequence of estimators to incorporate the lessons from the past.

Let us investigate by simulation whether the updated ridge regression estimator (3.6) indeed accumulates knowledge on the regression parameter. Hereto we draw a sequence of data sets from the linear regression model $\mathbf{Y}_t = \mathbf{X}_t\beta + \varepsilon_t$ with $t = 1, \dots, 25$. The elements of the design matrices are sampled from the standard normal distribution, the elements of β chosen as $\beta_j = (j - 50)/20$ for $j = 0, 1, 2, \dots, 100$, while $\varepsilon_t \sim \mathcal{N}(\mathbf{0}_n, 0.04\mathbf{I}_{nn})$. The regression parameter is estimated from each data set using both the regular and updated ridge estimators, in which the latter uses its previous estimate as target. The penalty parameter of both estimators are determined by cross-validation. We repeat all this a hundred times. The results, by the 5%, 50% and 95% quantiles, of the regular and updated ridge regression estimates of β_j with $j \in \{0, 30, 50, 70, 100\}$ are plotted against t (Figure 3.3). The quantiles of the regular ridge regression estimates of the selected elements of β are virtually constant over times (left panel of Figure 3.3). In contrast, the quantiles of the update ridge regression estimates (right panel of Figure 3.3) clearly improve as t increases. The improvement is two-fold: *i*) they become less biased, and *ii*) the distance between the 5% and 95% quantiles vanishes. The quantiles' behavior indicates that the updating does lead to an accumulation of knowledge.

In Figure 3.3 the 5%, 50% and 95% quantiles of the traditional and updated ridge estimates of β_j with $j \in \{0, 30, 50, 70, 100\}$ are plotted against t . These quantiles of the traditional ridge estimates of these elements of β are constant over t (left panel of Figure 3.3). Those of the update ridge estimates (right panel of Figure 3.3) clearly improve as t increases. The improvement is two-fold: *i*) they become less biased, and *ii*) the distance between the 5% and 95% quantiles vanishes.

The simulation results can be underpinned theoretically. Theorem 3.1 states (under conditions) the consistency in t of the sequence for the estimation of β and the associated linear predictor.

Theorem 3.1 (Theorem 2 of van Wieringen and Binder, 2022)

Adopt assumption *A1* and let $\{\hat{\beta}_t(\lambda_t)\}_{t=1}^{\infty}$ the corresponding sequence of update ridge regression estimators (3.6). Furthermore,

- i*) initiate the estimator sequence by any nonrandom target.
- ii*) let $\cap_{t=T}^{\infty} \text{null}(\mathbf{X}_t) = \mathbf{0}_p$ for sufficiently large $T \in \mathbb{N}$.
- iii*) choose the regularization scheme $\{\lambda_t\}_{t=1}^{\infty}$ such that $\lim_{t \rightarrow \infty} \sigma_{\varepsilon}^2 p d_1^2(\mathbf{X}_t) \lambda_t^{-2} = 0$ with $d_1(\mathbf{X}_t)$ the largest singular value of \mathbf{X}_t .

Then, for every $c > 0$:

$$\begin{aligned} \lim_{t \rightarrow \infty} P[\|\hat{\beta}_t(\lambda_t) - \beta\| \geq c | \{\lambda_{\tau}\}_{\tau=1}^t] &\rightarrow 0, \\ \lim_{t \rightarrow \infty} P[\|\mathbf{X}_{\text{new}}\hat{\beta}_t(\lambda_t) - \mathbf{X}_{\text{new}}\beta\| \geq c | \{\lambda_{\tau}\}_{\tau=1}^t] &\rightarrow 0, \end{aligned}$$

where \mathbf{X}_{new} is the design matrix of a novel study.

The updating procedure above is a frequentist analogue of Bayesian updating (Berger, 2013). From the Bayesian perspective, the ridge penalty corresponds to a normal prior on the regression parameter (see Chapter 2). With this normal prior, the posterior of β is also normal. In turn, this normal posterior serves as (normal) prior in the next update of β . The prior for the $t + 1$ -th update then becomes $\beta_{t+1} | \sigma^2 \sim \mathcal{N}[\beta_t(\lambda_t, \beta_{t-1}), \sigma^2 \lambda_{t+1}^{-1} \mathbf{I}_{pp}]$, which yields a posterior mean $\mathbb{E}[\beta_{t+1} | \mathbf{Y}_{t+1}, \mathbf{X}_{t+1}, \sigma^2, \beta_t(\lambda_t, \beta_{t-1})]$ coinciding with the frequentist estimator $\hat{\beta}_{t+1}(\lambda_{t+1}, \beta_t)$. \square

3.1 Moments

The expectation and variance of $\hat{\beta}(\Delta)$ are obtained through application of the same matrix algebra and expectation and covariance rules used in the derivation of their counterparts of the ‘regular’ ridge regression estimator. This leads to:

$$\begin{aligned} \mathbb{E}[\hat{\beta}(\Delta, \beta_0)] &= (\mathbf{X}^T \mathbf{W} \mathbf{X} + \Delta)^{-1} (\mathbf{X}^T \mathbf{W} \mathbf{X} \beta + \Delta \beta_0), \\ \text{Var}[\hat{\beta}(\Delta, \beta_0)] &= \sigma^2 (\mathbf{X}^T \mathbf{W} \mathbf{X} + \Delta)^{-1} \mathbf{X}^T \mathbf{W}^2 \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X} + \Delta)^{-1}, \end{aligned}$$

where σ^2 is the error variance. From these expressions similar limiting behaviour as for the ‘regular’ ridge regression case can be deduced. To this end let $\mathbf{V}_{\delta} \mathbf{D}_{\delta} \mathbf{V}_{\delta}^T$ be the eigendecomposition of Δ and $d_{\delta,j} = (\mathbf{D}_{\delta})_{jj}$.

Furthermore, define (with some abuse of notation) $\lim_{\Delta \rightarrow \infty}$ as the limit of all $d_{\delta,j}$ simultaneously tending to infinity. Then, $\lim_{\Delta \rightarrow \infty} \mathbb{E}[\hat{\beta}(\Delta, \beta_0)] = \beta_0$ and $\lim_{\Delta \rightarrow \infty} \text{Var}[\hat{\beta}(\Delta, \beta_0)] = \mathbf{0}_{pp}$.

Example 3.5

Let \mathbf{X} be an $n \times p$ -dimensional, orthonormal design matrix with $p \geq 2$. Contrast the regular and generalized ridge regression estimator, the latter with $\mathbf{W} = \mathbf{I}_{pp}$, $\beta_0 = \mathbf{0}_p$ and $\Delta = \lambda \mathbf{R}$ where $\mathbf{R} = (1 - \rho)\mathbf{I}_{pp} + \rho \mathbf{1}_{pp} \mathbf{1}_{pp}^\top$ for $\rho \in ((-p+1)^{-1}, 1)$. For $\rho = 0$ the two estimators coincide. The variance of the generalized ridge regression estimator then is $\text{Var}[\hat{\beta}(\Delta, \beta_0 = \mathbf{0}_p)] = (\mathbf{I}_{pp} + \Delta)^{-2}$. The efficiency of this estimator, measured by its generalized variance, is:

$$\det\{\text{Var}[\hat{\beta}(\Delta, \beta_0 = \mathbf{0}_p)]\} = \{[1 + \lambda + (p-1)\rho](1 + \lambda - \rho)^{p-1}\}^{-2}.$$

This efficiency attains its minimum at $\rho = 0$. In the present case, the regular ridge regression estimator is thus more efficient than its generalized counterpart. \square

Example 3.6 (MSE with perfect target)

Set $\beta_0 = \beta$, i.e. the target is equal to the true value of the regression parameter. Then:

$$\mathbb{E}[\hat{\beta}(\Delta, \beta_0 = \beta)] = (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \Delta)^{-1} (\mathbf{X}^\top \mathbf{W} \mathbf{X} \beta + \Delta \beta) = \beta.$$

Hence, irrespective of the choice of Δ , the generalized ridge is then unbiased. Thus:

$$\begin{aligned} \text{MSE}[\hat{\beta}(\Delta, \beta_0 = \beta)] &= \text{tr}\{\text{Var}[\hat{\beta}(\Delta, \beta_0 = \beta)]\} \\ &= \text{tr}[\sigma^2 (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \Delta)^{-1} \mathbf{X}^\top \mathbf{W}^2 \mathbf{X} (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \Delta)^{-1}] \\ &= \sigma^2 \text{tr}[\mathbf{X}^\top \mathbf{W}^2 \mathbf{X} (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \Delta)^{-2}]. \end{aligned}$$

When $\Delta = \lambda \mathbf{I}_{pp}$, this MSE is smaller than that of the ML regression estimator, irrespective of the choice of λ . \square

3.2 The Bayesian connection

This generalized ridge regression estimator can, like the regular ridge regression estimator, be viewed as a Bayesian estimator. It requires to replace the conjugate prior on β by a more general normal law, $\beta \sim \mathcal{N}(\beta_0, \sigma^2 \Delta^{-1})$, but retains the inverse gamma prior on σ^2 . The joint posterior distribution of β and σ^2 is then obtained analogously to the derivation of posterior with a standard normal prior on β as presented in Chapter 2 (the details are left as Exercise 3.5):

$$\begin{aligned} f_{\beta, \sigma^2}(\beta, \sigma^2 | \mathbf{Y}, \mathbf{X}) &= f_Y(\mathbf{Y} | \mathbf{X}, \beta, \sigma^2) f_{\beta|\sigma^2}(\beta | \sigma^2) f_{\sigma^2}(\sigma^2) \\ &\propto g_{\beta|\sigma^2, \mathbf{Y}, \mathbf{X}}(\beta | \sigma^2, \mathbf{Y}, \mathbf{X}) g_{\sigma^2|\mathbf{Y}, \mathbf{X}}(\sigma^2 | \mathbf{Y}, \mathbf{X}) \end{aligned}$$

with

$$g_{\beta|\sigma^2, \mathbf{Y}, \mathbf{X}}(\beta | \sigma^2, \mathbf{Y}, \mathbf{X}) \propto \exp\left\{-\frac{1}{2}\sigma^{-2} [\beta - \hat{\beta}(\Delta)]^\top (\mathbf{X}^\top \mathbf{X} + \Delta) [\beta - \hat{\beta}(\Delta)]\right\}.$$

This implies $\mathbb{E}(\beta | \sigma^2, \mathbf{Y}, \mathbf{X}) = \hat{\beta}(\Delta, \beta_0)$. Hence, the generalized ridge regression estimator too can be viewed as the Bayesian posterior mean estimator of β when imposing a multivariate Gaussian prior on the regression parameter.

The Bayesian formulation of our generalization of ridge regression provides additional intuition of the penalty parameters β_0 and Δ . For instance, a better initial guess, i.e. β_0 , yields a better posterior mean and mode. Or, less uncertainty in the prior, i.e. a smaller (in the positive definite ordering sense) variance Δ , yields a more concentrated posterior. These claims are underpinned in Question 3.6. The intuition they provide guides in the choice of penalty parameters β_0 and Δ .

3.3 Application

An illustration involving omics data can be found in the explanation of a gene's expression levels in terms of its DNA copy number. The latter is simply the number of gene copies encoded in the DNA. For instance, for most genes on the autosomal chromosomes the DNA copy number is two, as there is a single gene copy on

each chromosome and autosomal chromosomes come in pairs. Alternatively, in males the copy number is one for genes that map to the X or Y chromosome, while in females it is zero for genes on the Y chromosome. In cancer the DNA replication process has often been compromised leading to a (partially) reshuffled and aberrated DNA. Consequently, the cancer cell may exhibit gene copy numbers well over a hundred for classic oncogenes. A faulted replication process does – of course – not nicely follow the boundaries of gene encoding regions. This causes contiguous genes to commonly share aberrated copy numbers. With genes being transcribed from the DNA and a higher DNA copy number implying an enlarged availability of the gene's template, the latter is expected to lead to elevated expression levels. Intuitively, one expects this effect to be localized (a so-called *cis*-effect), but some suggest that aberrations elsewhere in the DNA may directly affect the expression levels of distant genes (referred to as a *trans*-effect). Figure 3.4 shows a cartoon of the *cis*- and *trans*-effect of DNA copy number on transcription.

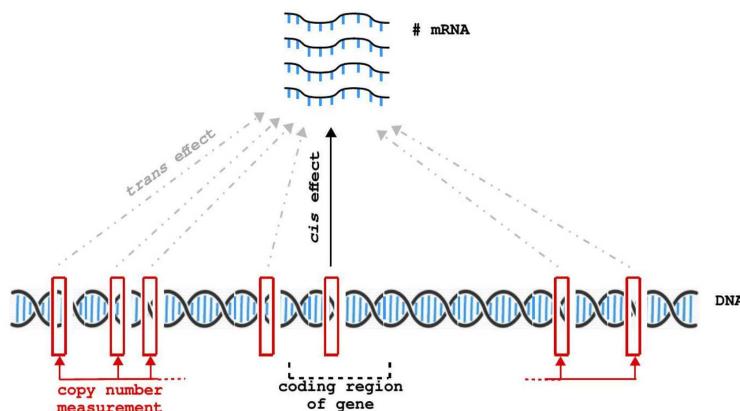


Figure 3.4: Illustration of the *cis*- and *trans*-effect of DNA copy number on gene expression levels.

The *cis*- and *trans*-effects of DNA copy aberrations on the expression levels of the KRAS oncogene in colorectal cancer are investigated. Data of both molecular levels from the TCGA (The Cancer Genome Atlas) repository are downloaded (Cancer Genome Atlas Network, 2012). The gene expression data are limited to that of KRAS, while for the DNA copy number data only that of chromosome 12, which harbors KRAS, is retained. This leaves genomic profiles of 195 samples comprising 927 aberrations. Both molecular data types are zero centered feature-wise. Moreover, the data are limited to ten – conveniently chosen? – samples. The KRAS expression levels are explained by the DNA copy number aberrations through the linear regression model. The model is fitted by means of ridge regression, with $\lambda\Delta$ and $\Delta = \mathbf{I}_{pp}$ and a single-banded Δ with unit diagonal and the elements of the first off-diagonal equal to the arbitrary value of -0.4 . The latter choice appeals to the spatial structure of the genome and encourages similar regression estimates for contiguous DNA copy numbers. The penalty parameter is chosen by means of leave-one-out cross-validation using the squared error loss.

Listing 3.1 R code

```
# load libraries
library(cgdsr)
library(biomaRt)
library(Matrix)

# get list of human genes
ensembl <- useMart("ensembl", dataset="hsapiens_gene_ensembl")
geneList <- getBM(attributes=c("ensembl_gene_id", "external_gene_name",
                             "entrezgene_trans_name", "chromosome_name",
                             "start_position", "end_position"), mart=ensembl)

# remove all gene without entrezID
geneList <- geneList[!is.na(geneList[,3]),]

# remove all genes not mapping to chr 12
geneList <- geneList[which(geneList[,4] %in% c(12)),]
geneList <- geneList[,-1]
geneList <- unique(geneList)
```

```

geneList <- geneList[order(geneList[,3], geneList[,4], geneList[,5]),]

# create CGDS object
mycgds   <- CGDS("http://www.cbioportal.org/public-portal/")

# get available case lists (collection of samples) for a given cancer study
mycancerstudy <- getStudies(mycgds)[37,1]
mycaselist    <- getCaseLists(mycgds, mycancerstudy)[1,1]

# get available genetic profiles
mrnaProf      <- getGeneticProfiles(mycgds, mycancerstudy)[c(4),1]
cnProf        <- getGeneticProfiles(mycgds, mycancerstudy)[c(6),1]

# get data slices for a specified list of genes, genetic profile and case list
cnData     <- numeric()
geData     <- numeric()
geneInfo  <- numeric()
for (j in 1:nrow(geneList)){
  geneName <- as.character(geneList[j,1])
  geneData <- getProfileData(mycgds, geneName, c(cnProf, mrnaProf), mycaselist)
  if (dim(geneData)[2] == 2 & dim(geneData)[1] > 0){
    cnData <- cbind(cnData, geneData[,1])
    geData <- cbind(geData, geneData[,2])
    geneInfo <- rbind(geneInfo, geneList[j,])
  }
}
colnames(cnData) <- rownames(geneData)
colnames(geData) <- rownames(geneData)

# preprocess data
Y <- geData[, match("KRAS", geneInfo[,1]), drop=FALSE]
Y <- Y - mean(Y)
X <- sweep(cnData, 2, apply(cnData, 2, mean))

# subset data
idSample <- c(50, 58, 61, 75, 66, 22, 67, 69, 44, 20)
Y         <- Y[idSample]
X         <- X[idSample,]

# generate banded penalty matrix
diags <- list(rep(1, ncol(X)), rep(-0.4, ncol(X)-1))
Delta <- as.matrix(bandSparse(ncol(X), k=-c(0:1), diag=c(diags), symm=TRUE))

# define loss function
CVloss <- function(lambda, X, Y, Delta){
  loss <- 0
  for (i in 1:nrow(X)){
    betasLoo <- solve(crossprod(X[-i,]) + lambda * Delta) %*%
      crossprod(X[-i,], Y[-i])
    loss <- loss + as.numeric((Y[i] - X[i,,drop=FALSE] %*% betasLoo)^2)
  }
  return(loss)
}

# optimize penalty parameter
limitsL <- c(10^(-10), 10^(10))
optLr  <- optimize(CVloss, limitsL, X=X, Y=Y, Delta=diag(ncol(X)))$minimum
optLgr <- optimize(CVloss, limitsL, X=X, Y=Y, Delta=Delta)$minimum

# evaluate (generalized) ridge estimators
betasGr <- solve(crossprod(X) + optLgr * Delta)           %*% crossprod(X, Y)

```

```

betasR <- solve(crossprod(X) + optLr * diag(ncol(X))) %*% crossprod(X, Y)

# plot estimates vs. location
ylims <- c(min(betasR, betasGr), max(betasR, betasGr))
plot(betasR, type="l", ylim=ylims, ylab="copy_number_effect",
      lty=2, yaxt="n", xlab="chromosomal_location")
lines(betasGr, lty=1, col="red")
lines(seq(ylims[1], ylims[2], length.out=50) ~
      rep(match("KRAS", geneInfo[,1]), 50), col="grey", lwd=2, lty=3)
legend("topright", c("ridge", "fused_ridge"), lwd=2,
       col=c("black", "red"), lty=c(2, 1))

```

The right panel of Figure 3.2 shows the ridge regression estimate with both choices of Δ and optimal penalty parameters plotted against the chromosomal order. The location of KRAS is indicated by a vertical dashed bar. The ordinary ridge regression estimates show a minor peak at the location of KRAS but is otherwise more or less flat. In the generalized ridge estimates the peak at KRAS is emphasized. Moreover, the region close to KRAS exhibits clearly elevated estimates, suggesting co-aberrated DNA. For the remainder the generalized ridge estimates portray a flat surface, with the exception of a single downward spike away from KRAS. Such negative effects are biologically nonsensible (more gene templates leading to reduced expression levels?). On the whole the generalized ridge estimates point towards the *cis*-effect as the dominant genomic regulation mechanism of KRAS expression. The isolated spike may suggest the presence of a *trans*-effect, but its sign is biological nonsensible and the spike is fully absent in the ordinary ridge estimates. This leads us to ignore the possibility of a genomic *trans*-effect on KRAS expression levels in colorectal cancer.

The sample selection demands justification. It yields a clear illustrateable difference between the ordinary and fused ridge estimates. When all samples are left in, the *cis*-effect is clearly present, discernable from both estimates that yield a virtually similar profile.

3.4 Generalized ridge regression

What is generally referred to as ‘generalized ridge regression’ (cf. Hoerl and Kennard, 1970; Hemmerle, 1975) is the particular case of loss function (3.1) in which $\mathbf{W} = \mathbf{I}_{nn}$, $\beta_0 = \mathbf{0}_p$, and $\Delta = \mathbf{V}_x \Lambda \mathbf{V}_x^\top$, where \mathbf{V}_x is obtained from the singular value decomposition of \mathbf{X} (i.e., $\mathbf{X} = \mathbf{U}_x \mathbf{D}_x \mathbf{V}_x^\top$ with its constituents endowed with the usual interpretation) and Λ a positive definite diagonal matrix. This gives the estimator:

$$\begin{aligned}\hat{\beta}(\Lambda) &= (\mathbf{X}^\top \mathbf{X} + \Delta)^{-1} \mathbf{X}^\top \mathbf{Y} \\ &= (\mathbf{V}_x \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{U}_x \mathbf{D}_x \mathbf{V}_x^\top + \mathbf{V}_x \Lambda \mathbf{V}_x^\top)^{-1} \mathbf{V}_x \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{Y} \\ &= \mathbf{V}_x (\mathbf{D}_x^\top \mathbf{D}_x + \Lambda)^{-1} \mathbf{D}_x^\top \mathbf{U}_x^\top \mathbf{Y}.\end{aligned}$$

From this last expression it becomes clear how this estimator generalizes the ‘regular ridge estimator’. The latter shrinks all eigenvalues, irrespectively of their size, in the same manner through a common penalty parameter. The ‘generalized ridge estimator’, through differing penalty parameters (i.e. the diagonal elements of Λ), shrinks them individually.

The generalized ridge estimator coincides with the Bayesian linear regression estimator with the normal prior $\mathcal{N}[\mathbf{0}_p, (\mathbf{V}_x \Lambda \mathbf{V}_x^\top)^{-1}]$ on the regression parameter β (and preserving the inverse gamma prior on the error variance). Assume \mathbf{X} to be of full column rank and choose $\Lambda = g^{-1} \mathbf{D}_x^\top \mathbf{D}_x$ with g a positive scalar. The prior on β then – assuming $(\mathbf{X}^\top \mathbf{X})^{-1}$ exists – reduces to Zellner’s g -prior: $\beta \sim \mathcal{N}[\mathbf{0}_p, g(\mathbf{X}^\top \mathbf{X})^{-1}]$ (Zellner, 1986). The corresponding estimator of the regression coefficient is: $\hat{\beta}(g) = g(1+g)^{-1} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$, which is proportional to the unpenalized ordinary least squares estimator of β .

For convenience of notation in the analysis of the generalized ridge estimator the linear regression model is usually rewritten as:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon = \mathbf{X}\mathbf{V}_x \mathbf{V}_x^\top \beta + \varepsilon = \tilde{\mathbf{X}}\alpha + \varepsilon,$$

with $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{V}_x = \mathbf{U}_x \mathbf{D}_x$ (and thus $\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} = \mathbf{D}_x^\top \mathbf{D}_x$) and $\alpha = \mathbf{V}_x^\top \beta$ with loss function $(\mathbf{Y} - \tilde{\mathbf{X}}\alpha)^\top (\mathbf{Y} - \tilde{\mathbf{X}}\alpha) + \alpha^\top \Lambda \alpha$. In the notation above the generalized ridge estimator is then:

$$\hat{\alpha}(\Lambda) = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \Lambda)^{-1} \tilde{\mathbf{X}}^\top \mathbf{Y} = (\mathbf{D}_x^\top \mathbf{D}_x + \Lambda)^{-1} \tilde{\mathbf{X}}^\top \mathbf{Y},$$

from which one obtains $\hat{\beta}(\Lambda) = \mathbf{V}_x \hat{\alpha}(\Lambda)$. Using $\mathbb{E}[\hat{\alpha}(\Lambda)] = (\mathbf{D}_x^\top \mathbf{D}_x + \Lambda)^{-1} \mathbf{D}_x^\top \mathbf{D}_x \alpha$ and $\text{Var}[\hat{\alpha}(\Lambda)] = \sigma^2 (\mathbf{D}_x^\top \mathbf{D}_x + \Lambda)^{-1} \mathbf{D}_x^\top \mathbf{D}_x (\mathbf{D}_x^\top \mathbf{D}_x + \Lambda)^{-1}$, the MSE for the generalized ridge estimator can be written as:

$$\text{MSE}[\hat{\alpha}(\Lambda)] = \sum_{j=1}^p (\sigma^2 d_{x,j}^2 + \alpha_j^2 \lambda_j^2) (d_{x,j}^2 + \lambda_j)^{-2},$$

where $d_{x,j} = (\mathbf{D}_x)_{jj}$ and $\lambda_j = (\Lambda)_{jj}$. Having α and σ^2 available, it is easily seen (equate the derivative w.r.t. λ_j to zero and solve) that the MSE of $\hat{\alpha}(\Lambda)$ is minimized by $\lambda_j = \sigma^2 / \alpha_j^2$ for all j . With α and σ^2 unknown, Hoerl and Kennard (1970) suggest an iterative procedure to estimate the λ_j 's. Initiate the procedure with the OLS estimates of α and σ^2 , followed by sequentially updating the λ_j 's and the estimates of α and σ^2 . An analytic expression of the limit of this procedure exists (Hemmerle, 1975). This limit, however, still depends on the observed \mathbf{Y} and as such it does not necessarily yield the minimal attainable value of the MSE. This limit may nonetheless still yield a potential gain in MSE. This is investigated in Lawless (1981). Under a variety of cases it seems to indeed outperform the OLS estimator, but there are exceptions.

A variation on this theme is presented by Guilkey and Murphy (1975) and dubbed “directed” ridge regression. Directed ridge regression only applies the above ‘generalized shrinkage’ in those eigenvector directions that have a corresponding small(er) – than some user-defined cut-off – eigenvalue. This intends to keep the bias low and yield good (or supposedly better) performance.

3.5 Conclusion

A note of caution to conclude. The generalized ridge penalty is extremely flexible. It can incorporate any prior knowledge on the parameter values (through specification of β_0) and the relations among these parameters (via Δ). While a pilot study or literature may provide a suggestion for β_0 , it is less obvious how to choose an informative Δ , although a spatial structure is a nice exception. In general, exact knowledge on the parameters should not be incorporated implicitly via the penalty (read: prior) but preferably be used explicitly in the model – the likelihood – itself. Though this may be the viewpoint of a prudent frequentist and a subjective Bayesian might disagree.

3.6 Exercises

Question 3.1

Consider a pathway comprising of three genes called A , B , and C . Let random variables $Y_{i,a}$, $Y_{i,b}$, and $Y_{i,c}$ be the random variable representing the expression of levels of genes A , B , and C in sample i . Hundred realizations, i.e. $i = 1, \dots, n$, of $Y_{i,a}$, $Y_{i,b}$, and $Y_{i,c}$ are available from an observational study. In order to assess how the expression levels of gene A are affected by that of genes B and C a medical researcher fits the

$$Y_{i,a} = \beta_b Y_{i,b} + \beta_c Y_{i,c} + \varepsilon_i,$$

with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. This model is fitted by means of ridge regression, but with a separate penalty parameter, λ_b and λ_c , for the two regression coefficients, β_b and β_c , respectively.

- a) Write down the ridge penalized loss function employed by the researcher.
- b) Does a different choice of penalty parameter for the second regression coefficient affect the estimation of the first regression coefficient? Motivate your answer.
- c) The researcher decides that the second covariate $Y_{i,c}$ is irrelevant. Instead of removing the covariate from model, the researcher decides to set $\lambda_c = \infty$. Show that this results in the same ridge estimate for β_b as when fitting (again by means of ridge regression) the model without the second covariate.

Question 3.2

Consider the linear regression model $Y_i = \beta_1 X_{i,1} + \beta_2 X_{i,2} + \varepsilon_i$ for $i = 1, 2, 3$. Information on the response, design matrix and relevant summary statistics are:

$$\mathbf{X}^\top = \begin{pmatrix} -1 & 0 & 2 \\ 1 & -2 & 1 \end{pmatrix}, \mathbf{Y}^\top = \begin{pmatrix} -1 & -1 & 2 \end{pmatrix}, \mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 5 & 1 \\ 1 & 6 \end{pmatrix}, \text{ and } \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} 5 \\ 3 \end{pmatrix}.$$

- a) Evaluate the fused ridge regression estimator with $\lambda = 2$.
- b) Draw the contour of the parameter constraint induced by the fused ridge penalty.
- c) Verify that for $p = 2$ the ridge homogeneity penalty is equivalent to the fused ridge penalty.

Question 3.3

Consider the linear regression model $Y_i = \beta_1 X_{i,1} + \beta_2 X_{i,2} + \varepsilon_i$ for $i = 1, \dots, n$. Suppose estimates of the regression parameters (β_1, β_2) of this model are obtained through the minimization of the sum-of-squares augmented with a ridge-type penalty:

$$\sum_{i=1}^n (Y_i - \beta_1 X_{i,1} - \beta_2 X_{i,2})^2 + \lambda(\beta_1^2 + \beta_2^2 + 2\nu\beta_1\beta_2),$$

with penalty parameters $\lambda \in \mathbb{R}_{>0}$ and $\nu \in (-1, 1)$.

- a) Recall the equivalence between constrained and penalized estimation (cf. Section 1.5). Sketch (for both $\nu = 0$ and $\nu = 0.9$) the shape of the parameter constraint induced by the penalty above and describe in words the qualitative difference between both shapes.
- b) When $\nu = -1$ and $\lambda \rightarrow \infty$ the estimates of β_1 and β_2 (resulting from minimization of the penalized loss function above) converge towards each other: $\lim_{\lambda \rightarrow \infty} \hat{\beta}_1(\lambda, -1) = \lim_{\lambda \rightarrow \infty} \hat{\beta}_2(\lambda, -1)$. Motivated by this observation a data scientist incorporates the equality constraint $\beta_1 = \beta = \beta_2$ explicitly into the model, and s/he estimates the ‘joint regression parameter’ β through the minimization (with respect to β) of:

$$\sum_{i=1}^n (Y_i - \beta X_{i,1} - \beta X_{i,2})^2 + \delta\beta^2,$$

with penalty parameter $\delta \in \mathbb{R}_{>0}$. The data scientist is surprised to find that resulting estimate $\hat{\beta}(\delta)$ does not have the same limiting (in the penalty parameter) behavior as the $\hat{\beta}_1(\lambda, -1)$, i.e. $\lim_{\delta \rightarrow \infty} \hat{\beta}(\delta) \neq \lim_{\lambda \rightarrow \infty} \hat{\beta}_1(\lambda, -1)$. Explain the misconception of the data scientist.

- c) Assume that i) $n \gg 2$, ii) the unpenalized estimates $(\hat{\beta}_1(0, 0), \hat{\beta}_2(0, 0))^\top$ equal $(-2, 2)$, and iii) that the two covariates X_1 and X_2 are zero-centered, have equal variance, and are strongly negatively correlated. Consider $(\hat{\beta}_1(\lambda, \nu), \hat{\beta}_2(\lambda, \nu))^\top$ for both $\nu = -0.9$ and $\nu = 0.9$. For which value of ν do you expect the sum of the absolute value of the estimates to be largest? Hint: Distinguish between small and large values of λ and think geometrically!

Question 3.4

Show that the generalized ridge regression estimator, $\hat{\beta}(\Delta) = (\mathbf{X}^\top \mathbf{X} + \Delta)^{-1} \mathbf{X}^\top \mathbf{Y}$, too (as in Question 1.6) can be obtained by ordinary least squares regression on an augmented data set. Hereto consider the Cholesky decomposition of the penalty matrix: $\Delta = \mathbf{L}^\top \mathbf{L}$. Now augment the matrix \mathbf{X} with p additional rows comprising the matrix \mathbf{L} , and augment the response vector \mathbf{Y} with p zeros.

Question 3.5

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_p, \sigma^2 \mathbf{I}_{pp})$. Assume $\beta \sim \mathcal{N}(\beta_0, \sigma^2 \Delta^{-1})$ with $\beta_0 \in \mathbb{R}^p$ and $\Delta \succ 0$ and a gamma prior on the error variance. Verify (i.e., work out the details of the derivation) that the posterior mean coincides with the generalized ridge estimator defined as:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X} + \Delta)^{-1} (\mathbf{X}^\top \mathbf{Y} + \Delta \beta_0).$$

Question 3.6

Consider the Bayesian linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$, a multivariate normal law as conditional prior distribution on the regression parameter: $\beta | \sigma^2 \sim \mathcal{N}(\beta_0, \sigma^2 \Delta^{-1})$, and an inverse gamma prior on the error variance $\sigma^2 \sim \mathcal{IG}(\gamma, \delta)$. The consequences of various choices for the hyper parameters of the prior distribution on β are studied.

- a) Consider the following conditional prior distributions on the regression parameters $\beta | \sigma^2 \sim \mathcal{N}(\beta_0, \sigma^2 \Delta_a^{-1})$ and $\beta | \sigma^2 \sim \mathcal{N}(\beta_0, \sigma^2 \Delta_b^{-1})$ with precision matrices $\Delta_a, \Delta_b \in \mathcal{S}_{++}^p$ such that $\Delta_a \succeq \Delta_b$, i.e. $\Delta_a = \Delta_b + \mathbf{D}$ for some positive semi-definite symmetric matrix of appropriate dimensions. Verify:

$$\text{Var}(\beta | \sigma^2, \mathbf{Y}, \mathbf{X}, \beta_0, \Delta_a) \preceq \text{Var}(\beta | \sigma^2, \mathbf{Y}, \mathbf{X}, \beta_0, \Delta_b),$$

i.e. the smaller (in the positive definite ordering) the variance of the prior the smaller that of the posterior.

- b) In the remainder of this exercise assume $\Delta_a = \Delta = \Delta_b$. Let β_t be the ‘true’ or ‘ideal’ value of the regression parameter, that has been used in the generation of the data, and show that a better initial guess yields a better posterior probability at β_t . That is, take two prior mean parameters $\beta_0 = \beta_0^{(a)}$ and $\beta_0 = \beta_0^{(b)}$ such that the former is closer to β_t than the latter. Here close is defined in terms of the Mahalanobis distance, which for, e.g. β_t and $\beta_0^{(a)}$ is defined as $d_M(\beta_t, \beta_0^{(a)}; \Sigma) = [(\beta_t - \beta_0^{(a)})^\top \Sigma^{-1} (\beta_t - \beta_0^{(a)})]^{1/2}$ with positive definite covariance matrix Σ with $\Sigma = \sigma^2 \Delta^{-1}$. Show that the posterior density $\pi_{\beta | \sigma^2}(\beta | \sigma^2, \mathbf{Y}, \mathbf{X}, \beta_0^{(a)}, \Delta)$ is larger at $\beta = \beta_t$ than with the other prior mean parameter.

- c) Adopt the assumptions of part b) and show that a better initial guess yields a better posterior mean. That is, show

$$d_M[\beta_t, \mathbb{E}(\beta | \sigma^2, \mathbf{Y}, \mathbf{X}, \beta_0^{(a)}, \Delta); \Sigma] \leq d_M[\beta_t, \mathbb{E}(\beta | \sigma^2, \mathbf{Y}, \mathbf{X}, \beta_0^{(b)}, \Delta); \Sigma],$$

now with $\Sigma = \sigma^2(\mathbf{X}^\top \mathbf{X} + \Delta)^{-1}$.

Question 3.7

The ridge penalty may be interpreted as a multivariate normal prior on the regression coefficients: $\beta \sim \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I}_{pp})$. Different priors may be considered. In case the covariates are spatially related in some sense (e.g. genetically), it may of interest to assume a first-order autoregressive prior: $\beta \sim \mathcal{N}(\mathbf{0}, \lambda^{-1}\Sigma_a)$, in which Σ_a is a $(p \times p)$ -dimensional correlation matrix with $(\Sigma_a)_{j_1, j_2} = \rho^{|j_1 - j_2|}$ for some correlation coefficient $\rho \in [0, 1)$. Hence,

$$\Sigma_a = \begin{pmatrix} 1 & \rho & \dots & \rho^{p-1} \\ \rho & 1 & \dots & \rho^{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \dots & 1 \end{pmatrix}.$$

- a) The penalized loss function associated with this AR(1) prior is:

$$\mathcal{L}(\beta; \lambda, \Sigma_a) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\beta^\top \Sigma_a^{-1}\beta.$$

Find the minimizer of this loss function.

- b) What is the effect of ρ on the ridge estimates? Contrast this to the effect of λ . Illustrate this on (simulated) data.
c) Instead of an AR(1) prior assume a prior with a uniform correlation between the elements of β . That is, replace Σ_a by Σ_u , given by $\Sigma_u = (1 - \rho)\mathbf{I}_{pp} + \rho\mathbf{1}_{pp}\mathbf{1}_{pp}^\top$. Investigate (again on data) the effect of changing from the AR(1) to the uniform prior on the ridge regression estimates.

Question 3.8

Consider the standard linear regression model $Y_i = \mathbf{X}_{i,*}\beta + \varepsilon_i$ for $i = 1, \dots, n$. Suppose estimates of the regression parameters β of this model are obtained through the minimization of the sum-of-squares augmented with a ridge-type penalty:

$$\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda[(1 - \alpha)\|\beta - \beta_{t,a}\|_2^2 + \alpha\|\beta - \beta_{t,b}\|_2^2],$$

for known $\alpha \in [0, 1]$, nonrandom p -dimensional target vectors $\beta_{t,a}$ and $\beta_{t,b}$ with $\beta_{t,a} \neq \beta_{t,b}$, and penalty parameter $\lambda > 0$. Here $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ and \mathbf{X} is $n \times p$ matrix with the n row-vectors $\mathbf{X}_{i,*}$ stacked.

- a) When $p > n$ the sum-of-squares part does not have a unique minimum. Does the above employed penalty warrant a unique minimum for the loss function above (i.e., sum-of-squares plus penalty)? Motivate your answer.
b) Could it be that for intermediate values of α , i.e. $0 < \alpha < 1$, the loss function assumes smaller values than for the boundary values $\alpha = 0$ and $\alpha = 1$? Motivate your answer.
c) Draw the parameter constraint induced by this penalty for $\alpha = 0, 0.5$ and 1 when $p = 2$
d) Derive the estimator of β , defined as the minimum of the loss function, explicitly.
e) Discuss the behaviour of the estimator $\alpha = 0, 0.5$ and 1 for $\lambda \rightarrow \infty$.

Question 3.9

Revisit Exercise 1.3. There the standard linear regression model $Y_i = \mathbf{X}_{i,*}\beta + \varepsilon_i$ for $i = 1, \dots, n$ and with $\varepsilon_i \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$ is considered. The model comprises a single covariate and an intercept. Response and covariate data are: $\{(y_i, x_{i,1})\}_{i=1}^4 = \{(1.4, 0.0), (1.4, -2.0), (0.8, 0.0), (0.4, 2.0)\}$.

- a) Evaluate the generalized ridge regression estimator of β with target $\beta_0 = \mathbf{0}_2$ and penalty matrix Δ given by $(\Delta)_{11} = \lambda = (\Delta)_{22}$ and $(\Delta)_{12} = \frac{1}{2}\lambda = (\Delta)_{21}$ in which $\lambda = 8$.
b) A data scientist wishes to leave the intercept unpenalized. Hereto s/he sets in part a) $(\Delta)_{11} = 0$. Why does the resulting estimate not coincide with the answer to Exercise 1.3? Motivate.

Question 3.10

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma_\varepsilon^2 \mathbf{I}_{nn})$. This model (without intercept) is

fitted to data using the generalized regression estimator $\hat{\beta}(\lambda) = \arg \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\beta^\top \Delta\beta$ with $\lambda > 0$. The penalty matrix Δ and the data are:

$$\Delta = \begin{pmatrix} 4 & c \\ c & 1 \end{pmatrix}, \mathbf{X} = \begin{pmatrix} -2 & 1 \end{pmatrix}, \text{ and } \mathbf{Y} = \begin{pmatrix} -2 \end{pmatrix}.$$

for $c \in [-2, 2]$.

- a) Draw the parameter constraint on β induced by the generalized ridge penalty for $c = 0$ and $c = 2$. Motivate the form of the drawn constraints.
- b) Evaluate the generalized ridge regression estimator $\hat{\beta}(\lambda, \Delta)$ for $c = 2$.
- c) Why is the generalized ridge regression estimator $\hat{\beta}(\lambda, \Delta)$ well-defined for $c = 2$ but not for $c = -2$? Motivate.
- d) Set $\lambda = 1$ and $c = 0$. Suppose $\beta = (2, 1)^\top$. Is the bias of the generalized ridge regression estimator larger or smaller than that of its regular counterpart (i.e. with $\Delta = \mathbf{I}_{22}$)? Motivate.

Question 3.11

Consider the linear regression model: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$. Let $\hat{\beta}(\lambda) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y}$ be the ridge regression estimator with penalty parameter λ . The shrinkage of the ridge regression estimator propagates to the scale of the ‘ridge prediction’ $\mathbf{X}\hat{\beta}(\lambda)$. To correct (a bit) for the shrinkage, de Vlaming and Groenen (2015) propose the alternative ridge regression estimator: $\hat{\beta}(\alpha) = [(1 - \alpha)\mathbf{X}^\top \mathbf{X} + \alpha \mathbf{I}_{pp}]^{-1} \mathbf{X}^\top \mathbf{Y}$ with shrinkage parameter $\alpha \in [0, 1]$.

- a) Let $\alpha = \lambda(1 + \lambda)^{-1}$. Show that $\hat{\beta}(\alpha) = (1 + \lambda)\hat{\beta}(\lambda)$ with $\hat{\beta}(\lambda)$ as in the introduction above.
- b) Use part a) and the parametrization of α provided there to show that the some shrinkage has been undone. That is, show: $\text{Var}[\mathbf{X}\hat{\beta}(\lambda)] < \text{Var}[\mathbf{X}\hat{\beta}(\alpha)]$ for any $\lambda > 0$.
- c) Use the singular value decomposition of \mathbf{X} to show that $\lim_{\alpha \downarrow 0} \hat{\beta}(\alpha) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ (should it exist) and $\lim_{\alpha \uparrow 1} \hat{\beta}(\alpha) = \mathbf{X}^\top \mathbf{Y}$.
- d) Derive the expectation, variance and mean squared error of $\hat{\beta}(\alpha)$.
- e) Temporarily assume that $p = 1$ and let $\mathbf{X}^\top \mathbf{X} = c$ for some $c > 0$. Then, $\text{MSE}[\hat{\beta}(\alpha)] = (c - 1)^2 \beta^2 + \sigma^2 c[(1 - \alpha)c + \alpha]^{-2}$. Does there exist an $\alpha \in (0, 1)$ such that the mean squared error of $\hat{\beta}(\alpha)$ is smaller than that of its maximum likelihood counterpart? Motivate.
- f) Now assume $p > 1$ and an orthonormal design matrix. Specify the regularization path of the alternative ridge regression estimator $\hat{\beta}(\alpha)$.

Question 3.12

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$. Goldstein & Smith (1974) proposed a novel generalized ridge estimator of its p -dimensional regression parameter:

$$\hat{\beta}_m(\lambda) = [(\mathbf{X}^\top \mathbf{X})^m + \lambda \mathbf{I}_{pp}]^{-1} (\mathbf{X}^\top \mathbf{X})^{m-1} \mathbf{X}^\top \mathbf{Y},$$

with penalty parameter $\lambda > 0$ and ‘shape’ or ‘rate’ parameter m .

- a) Assume, only for part a), that $n = p$ and the design matrix is orthonormal. Show that, irrespectively of the choice of m , this generalized ridge regression estimator coincides with the ‘regular’ ridge regression estimator.
- b) Consider the generalized ridge loss function $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \beta^\top \mathbf{A}\beta$ with \mathbf{A} a $p \times p$ -dimensional symmetric matrix. For what \mathbf{A} , does $\hat{\beta}_m(\lambda)$ minimize this loss function?
- c) Let d_j be the j -th singular value of \mathbf{X} . Show that in $\hat{\beta}_m(\lambda)$ the singular values are shrunken as $(d_j^{2m} + \lambda)^{-1} d_j^{2m-2}$. Hint: use the singular value decomposition of \mathbf{X} .
- d) Do, for positive singular values, larger m lead to more shrinkage? Hint: Involve particulars of the singular value in your answer.
- e) Express $\mathbb{E}[\hat{\beta}_m(\lambda)]$ in terms of the design matrix, model and shrinkage parameters λ and m .
- f) Express $\text{Var}[\hat{\beta}_m(\lambda)]$ in terms of the design matrix, model and shrinkage parameters λ and m .

4 Mixed model

Here the mixed model introduced by Henderson (1953), which generalizes the linear regression model, is studied and estimated in unpenalized (!) fashion. Nonetheless, it will turn out to have an interesting connection to ridge regression. This connection may be exploited to arrive at an informed choice of the ridge penalty parameter.

The linear regression model, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, assumes the effect of each covariate to be fixed. In certain situations it may be desirable to relax this assumption. For instance, a study may be replicated. Conditions need not be exactly constant across replications. Among others this may be due to batch effects. These may be accounted for and are then incorporated in the linear regression model. But it is not the effects of these particular batches included in the study that are of interest. Would the study have been carried out at a later date, other batches may have been involved. Hence, the included batches are thus a random draw from the population of all batches. With each batch possibly having a different effect, these effects may also be viewed as random draws from some hypothetical ‘effect’-distribution. From this point of view the effects estimated by the linear regression model are realizations from the ‘effect’-distribution. But interest is not in the particular but the general. Hence, a model that enables a generalization to the distribution of batch effects would be more suited here.

Like the linear regression model the *mixed model*, also called *mixed effects model* or *random effects model*, explains the variation in the response by a linear combination of the covariates. The key difference lies in the fact that the latter model distinguishes two sets of covariates, one with fixed effects and the other with random effects. In matrix notation mirroring that of the linear regression model, the mixed model can be written as:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon,$$

where \mathbf{Y} is the response vector of length n , \mathbf{X} the $(n \times p)$ -dimensional design matrix with the fixed vector β with p fixed effects, \mathbf{Z} the $(n \times q)$ -dimensional design matrix with an associated $q \times 1$ dimensional vector γ of random effects, and distributional assumptions $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma_\varepsilon^2 \mathbf{I}_{nn})$, $\gamma \sim \mathcal{N}(\mathbf{0}_q, \mathbf{R}_\theta)$ and ε and γ independent. In this \mathbf{R}_θ is symmetric, positive definite and parametrized by a low-dimensional parameter θ .

The distribution of \mathbf{Y} is fully defined by the mixed model and its accompanying assumptions. As \mathbf{Y} is a linear combination of normally distributed random variables, it is itself normally distributed. Its mean is:

$$\mathbb{E}(\mathbf{Y}) = \mathbb{E}(\mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon) = \mathbb{E}(\mathbf{X}\beta) + \mathbb{E}(\mathbf{Z}\gamma) + \mathbb{E}(\varepsilon) = \mathbf{X}\beta + \mathbf{Z}\mathbb{E}(\gamma) = \mathbf{X}\beta,$$

while its variance is:

$$\text{Var}(\mathbf{Y}) = \text{Var}(\mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon) = \mathbf{Z}\text{Var}(\gamma)\mathbf{Z}^\top + \sigma_\varepsilon^2 \mathbf{I}_{nn} = \mathbf{Z}\mathbf{R}_\theta\mathbf{Z}^\top + \sigma_\varepsilon^2 \mathbf{I}_{nn}$$

in which the independence between ε and γ and the standard algebra rules for the $\text{Var}(\cdot)$ and $\text{Cov}(\cdot)$ operators have been used. Put together, this yields: $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta, \mathbf{Z}\mathbf{R}_\theta\mathbf{Z}^\top + \sigma_\varepsilon^2 \mathbf{I}_{nn})$. Hence, the random effects term $\mathbf{Z}\gamma$ of the mixed model does not contribute to the explanation of the mean of \mathbf{Y} , but aids in the decomposition of its variance around the mean. From this formulation of the model it is obvious that the random part of two distinct observations of the response are – in general – not independent: their covariance is given by the corresponding element of $\mathbf{Z}\mathbf{R}_\theta\mathbf{Z}^\top$. Put differently, due to the independence assumption on the error two observations can only be (marginally) dependent through the random effect which is attenuated by the associated design matrix \mathbf{Z} . To illustrate this, temporarily set $\mathbf{R}_\theta = \sigma_\gamma^2 \mathbf{I}_{qq}$. Then, $\text{Var}(\mathbf{Y}) = \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}^\top + \sigma_\varepsilon^2 \mathbf{I}_{nn}$. From this it is obvious that two variates of \mathbf{Y} are now independent if and only if the corresponding rows of \mathbf{Z} are orthogonal. Moreover, two pairs of variates have the same covariance if they have the same covariate information in \mathbf{Z} . Two distinct observations of the same individual have the same covariance as one of these observations with that of another individual with identical covariate information as the left-out observation on the former individual. In particular, their ‘between-covariance’ equals their individual ‘within-covariance’.

The mixed model and the linear regression model are clearly closely related: they share a common mean, a normally distributed error, and both explain the response by a linear combination of the explanatory variables.

Moreover, when γ is known, the mixed model reduces to a linear regression model. This is seen from the conditional distribution of \mathbf{Y} : $\mathbf{Y} | \gamma \sim \mathcal{N}(\mathbf{X}\beta + \mathbf{Z}\gamma, \sigma_e^2 \mathbf{I}_{nn})$. Conditioning on the random effect γ thus pulls in the term $\mathbf{Z}\gamma$ to the systematic, non-random explanatory part of the model. In principle, the conditioned mixed model could now be rewritten as a linear regression model by forming a new design matrix and parameter from \mathbf{X} and \mathbf{Z} and β and γ , respectively.

Example 4.1 (*Mixed model for a longitudinal study*)

A longitudinal study looks into the growth rate of cells. At the beginning of the study cells are placed in n petri dishes, with the same growth medium but at different concentrations. The initial number of cells in each petri dish is counted as is done at several subsequent time points. The change in cell count is believed to be – at the log-scale – a linear function of the concentration of the growth medium. The linear regression model may suffice. However, variation is omnipresent in biology. That is, apart from variation in the initial cell count, each cell – even if they are from common decent – will react (slightly?) differently to the stimulus of the growth medium. This intrinsic cell-to-cell variation in growth response may be accommodated for in the linear mixed model by the introduction of a random cell effect, both in off-set and slope. The (log) cell count of petri dish i at time point t , denoted Y_{it} , is thus described by:

$$Y_{it} = \beta_0 + X_i\beta_1 + \mathbf{Z}_i\gamma + \varepsilon_{it},$$

with intercept β_0 , growth medium concentration X_i in petri dish i , and fixed growth medium effect β_1 , and $\mathbf{Z}_i = (1, X_i)$, γ the 2 dimensional random effect parameter bivariate normally distributed with zero mean and diagonal covariance matrix, and finally $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_e^2)$ the error in the cell count of petri dish i at time t . In matrix notation the matrix \mathbf{Z} would comprise of $2n$ columns: two columns for each cell, \mathbf{e}_i and $X_i\mathbf{e}_i$ (with \mathbf{e}_i the n -dimensional unit vector with a one at the i -th location and zeros elsewhere), corresponding to the random intercept and slope effect, respectively. The fact that the number columns of \mathbf{Z} , i.e. the explanatory random effects,

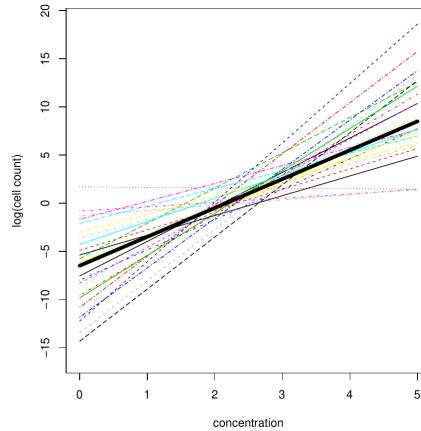


Figure 4.1: Linear regression (thick black solid line) vs. mixed model fit (thin colored and patterned lines).

equals $2n$ does not pose identifiability problems as per column only a single parameter is estimated. Finally, to illustrate the difference between the linear regression and the linear mixed model their fits on artificial data are plotted (top left panel, Figure 4.1). Where the linear regression fit shows the ‘grand mean relationship’ between cell count and growth medium, the linear mixed model fit depicts the petri dish specific fits. \square

The mixed model was motivated by its ability to generalize to instances not included in the study. From the examples above another advantage can be deduced. E.g., the cells’ effects are modelled by a single parameter (rather than one per cell). More degrees of freedom are thus left to estimate the noise level. In particular, a test for the presence of a cell effect will have more power.

The parameters of the mixed model are estimated either by means of likelihood maximization or a related procedure known as restricted maximum likelihood. Both are presented, with the exposé loosely based on Bates and DebRoy (2004). First the maximum likelihood procedure is introduced, which requires the derivation of the likelihood. Hereto the assumption on the random effects is usually transformed. Let $\tilde{\mathbf{R}}_\theta = \sigma_e^{-2} \mathbf{R}_\theta$, which is the covariance of the random effects parameter relative to the error variance, and $\tilde{\mathbf{L}}_\theta = \mathbf{L}_\theta \mathbf{L}_\theta^\top$ its Cholesky decomposition. Next define the change-of-variables $\gamma = \mathbf{L}_\theta \tilde{\gamma}$. This transforms the model to: $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{L}_\theta \tilde{\gamma} + \varepsilon$

but now with the assumption $\tilde{\gamma} \sim \mathcal{N}(\mathbf{0}_q, \sigma_\varepsilon^2 \mathbf{I}_{qq})$. Under this assumption the conditional likelihood, conditional on the random effects, is:

$$L(\mathbf{Y} | \tilde{\gamma} = \mathbf{g}) = (2\pi\sigma_\varepsilon^2)^{-n/2} \exp(-\frac{1}{2}\sigma_\varepsilon^{-2} \|\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{L}_\theta\mathbf{g}\|_2^2).$$

From this the unconditional likelihood is obtained through:

$$\begin{aligned} L(\mathbf{Y}) &= \int_{\mathbb{R}^q} L(\mathbf{Y} | \tilde{\gamma} = \mathbf{g}) f_{\tilde{\gamma}}(\mathbf{g}) d\mathbf{g} \\ &= \int_{\mathbb{R}^q} (2\pi\sigma_\varepsilon^2)^{-(n+q)/2} \exp[-\frac{1}{2}\sigma_\varepsilon^{-2}(\|\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{L}_\theta\mathbf{g}\|_2^2 + \|\mathbf{g}\|_2^2)] d\mathbf{g}. \end{aligned}$$

To evaluate the integral, the exponent needs rewriting. Hereto first note that:

$$\|\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{L}_\theta\mathbf{g}\|_2^2 + \|\mathbf{g}\|_2^2 = (\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{L}_\theta\mathbf{g})^\top (\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{L}_\theta\mathbf{g}) + \mathbf{g}^\top \mathbf{g}.$$

Now expand the right-hand side as follows:

$$\begin{aligned} &(\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{L}_\theta\mathbf{g})^\top (\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{L}_\theta\mathbf{g}) + \mathbf{g}^\top \mathbf{g} \\ &= \mathbf{Y}^\top \mathbf{Y} + \mathbf{g}^\top (\mathbf{L}_\theta^\top \mathbf{Z}^\top \mathbf{Z}\mathbf{L}_\theta + \mathbf{I}_{qq})\mathbf{g} + \beta^\top \mathbf{X}^\top \mathbf{X}\beta - \mathbf{Y}^\top \mathbf{X}\beta - \beta^\top \mathbf{X}^\top \mathbf{Y} \\ &\quad - (\mathbf{Y}^\top \mathbf{Z}\mathbf{L}_\theta - \beta^\top \mathbf{X}^\top \mathbf{Z}\mathbf{L}_\theta)\mathbf{g} - \mathbf{g}^\top (\mathbf{L}_\theta^\top \mathbf{Z}^\top \mathbf{Y} - \mathbf{L}_\theta^\top \mathbf{Z}^\top \mathbf{X}\beta) \\ &= (\mathbf{g} - \boldsymbol{\mu}_{\tilde{\gamma} | \mathbf{Y}})^\top (\mathbf{L}_\theta^\top \mathbf{Z}^\top \mathbf{Z}\mathbf{L}_\theta + \mathbf{I}_{qq})(\mathbf{g} - \boldsymbol{\mu}_{\tilde{\gamma} | \mathbf{Y}}) \\ &\quad + (\mathbf{Y} - \mathbf{X}\beta)^\top [\mathbf{I}_{nn} - \mathbf{Z}\mathbf{L}_\theta(\mathbf{L}_\theta^\top \mathbf{Z}^\top \mathbf{Z}\mathbf{L}_\theta + \mathbf{I}_{qq})^{-1}\mathbf{L}_\theta^\top \mathbf{Z}^\top](\mathbf{Y} - \mathbf{X}\beta) \\ &= (\mathbf{g} - \boldsymbol{\mu}_{\tilde{\gamma} | \mathbf{Y}})^\top (\mathbf{L}_\theta^\top \mathbf{Z}^\top \mathbf{Z}\mathbf{L}_\theta + \mathbf{I}_{qq})(\mathbf{g} - \boldsymbol{\mu}_{\tilde{\gamma} | \mathbf{Y}}) \\ &\quad + (\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{I}_{nn} + \mathbf{Z}\tilde{\mathbf{R}}_\theta\mathbf{Z}^\top)^{-1}(\mathbf{Y} - \mathbf{X}\beta), \end{aligned}$$

where $\boldsymbol{\mu}_{\tilde{\gamma} | \mathbf{Y}} = (\mathbf{L}_\theta^\top \mathbf{Z}^\top \mathbf{Z}\mathbf{L}_\theta + \mathbf{I}_{qq})^{-1}\mathbf{L}_\theta^\top \mathbf{Z}^\top(\mathbf{Y} - \mathbf{X}\beta)$ and the Woodbury identity has been used in the last step. As the notation suggests $\boldsymbol{\mu}_{\tilde{\gamma} | \mathbf{Y}}$ is the conditional expectation of the random effect conditional on the data: $\mathbb{E}(\tilde{\gamma} | \mathbf{Y})$. This may be verified from the conditional distribution $\tilde{\gamma} | \mathbf{Y}$ when exploiting the equality derived in the preceding display. Substitute the latter in the integral of the likelihood and use the change-of-variables: $\mathbf{h} = (\mathbf{L}_\theta^\top \mathbf{Z}^\top \mathbf{Z}\mathbf{L}_\theta + \mathbf{I}_{qq})^{1/2}(\mathbf{g} - \boldsymbol{\mu}_{\tilde{\gamma} | \mathbf{Y}})$ with Jacobian $|(\mathbf{L}_\theta^\top \mathbf{Z}^\top \mathbf{Z}\mathbf{L}_\theta + \mathbf{I}_{qq})^{1/2}|$:

$$\begin{aligned} L(\mathbf{Y}) &= \int_{\mathbb{R}^q} (2\pi\sigma_\varepsilon^2)^{-(n+q)/2} |\mathbf{L}_\theta^\top \mathbf{Z}^\top \mathbf{Z}\mathbf{L}_\theta + \mathbf{I}_{qq}|^{-1/2} \exp(-\frac{1}{2}\sigma_\varepsilon^{-2}\mathbf{h}^\top \mathbf{h}) \\ &\quad \exp[-\frac{1}{2}\sigma_\varepsilon^{-2}(\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{I}_{nn} + \mathbf{Z}\tilde{\mathbf{R}}_\theta\mathbf{Z}^\top)^{-1}(\mathbf{Y} - \mathbf{X}\beta)] d\mathbf{g} \\ &= (2\pi\sigma_\varepsilon^2)^{-n/2} |\mathbf{I}_{nn} + \mathbf{Z}\tilde{\mathbf{R}}_\theta\mathbf{Z}^\top|^{-1/2} \\ &\quad \exp[-\frac{1}{2}\sigma_\varepsilon^{-2}(\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{I}_{nn} + \mathbf{Z}\tilde{\mathbf{R}}_\theta\mathbf{Z}^\top)^{-1}(\mathbf{Y} - \mathbf{X}\beta)], \end{aligned} \tag{4.1}$$

where in the last step Sylvester's determinant identity has been used.

The maximum likelihood estimators of the mixed model parameters β , σ_ε^2 and $\tilde{\mathbf{R}}_\theta$ are found through the maximization of the logarithm of the likelihood (4.1). Find the roots of the partial derivatives of this log-likelihood with respect to the mixed model parameters. For β and σ_ε^2 this yields:

$$\begin{aligned} \hat{\beta} &= [\mathbf{X}^\top (\mathbf{I}_{nn} + \mathbf{Z}\tilde{\mathbf{R}}_\theta\mathbf{Z}^\top)^{-1} \mathbf{X}]^{-1} \mathbf{X}^\top (\mathbf{I}_{nn} + \mathbf{Z}\tilde{\mathbf{R}}_\theta\mathbf{Z}^\top)^{-1} \mathbf{Y}, \\ \hat{\sigma}_\varepsilon^2 &= \frac{1}{n} (\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{I}_{nn} + \mathbf{Z}\tilde{\mathbf{R}}_\theta\mathbf{Z}^\top)^{-1} (\mathbf{Y} - \mathbf{X}\beta). \end{aligned}$$

The former estimate can be substituted into the latter to remove its dependency on β . However, both estimators still depend on θ . An estimator of θ may be found by substitution of $\hat{\beta}$ and $\hat{\sigma}_\varepsilon^2$ into the log-likelihood followed by its maximization. For general parametrizations of $\tilde{\mathbf{R}}_\theta$ by θ there are no explicit solutions. Then, resort to standard nonlinear solvers such as the Newton-Raphson algorithm and the like. With a maximum likelihood estimate of θ at hand, those of the other two mixed model parameters are readily obtained from the formula's above. As θ is unknown at the onset, it needs to be initiated followed by sequential updating of the parameter estimates until convergence.

Restricted maximum likelihood (REML) considers the fixed effect parameter β as a ‘nuisance’ parameter and concentrates on the estimation of the variance components. The nuisance parameter is integrated out of the likelihood, $\int_{\mathbb{R}^p} L(\mathbf{Y}) d\beta$, which is referred to as the restricted likelihood. Those values of θ (and thereby $\tilde{\mathbf{R}}_\theta$) and

σ_ε^2 that maximize the restricted likelihood are the REML estimators. The restricted likelihood, by an argument similar to that used in the derivation of the likelihood, simplifies to:

$$\begin{aligned} \int_{\mathbb{R}^p} L(\mathbf{Y}) d\beta &= (2\pi\sigma_\varepsilon^2)^{-n/2} |\tilde{\mathbf{Q}}|^{-1/2} \exp\left\{-\frac{1}{2}\sigma_\varepsilon^{-2} \mathbf{Y}^\top [\tilde{\mathbf{Q}}_\theta^{-1} - \tilde{\mathbf{Q}}_\theta^{-1} \mathbf{X} (\mathbf{X}^\top \tilde{\mathbf{Q}}_\theta^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \tilde{\mathbf{Q}}_\theta^{-1}] \mathbf{Y}\right\} \\ &\quad \int_{\mathbb{R}^p} \exp\left\{-\frac{1}{2}\sigma_\varepsilon^{-2} [\beta - (\mathbf{X}^\top \tilde{\mathbf{Q}}_\theta^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \tilde{\mathbf{Q}}_\theta^{-1} \mathbf{Y}]^\top \mathbf{X}^\top \tilde{\mathbf{Q}}_\theta^{-1} \mathbf{X} \right. \\ &\quad \left. [\beta - (\mathbf{X}^\top \tilde{\mathbf{Q}}_\theta^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \tilde{\mathbf{Q}}_\theta^{-1} \mathbf{Y}]\right\} d\beta \\ &= (2\pi\sigma_\varepsilon^2)^{-(n-p)/2} |\tilde{\mathbf{Q}}_\theta|^{-1/2} |\mathbf{X}^\top \tilde{\mathbf{Q}}_\theta^{-1} \mathbf{X}|^{-1/2} \\ &\quad \exp\left\{-\frac{1}{2}\sigma_\varepsilon^{-2} \mathbf{Y}^\top [\tilde{\mathbf{Q}}_\theta^{-1} - \tilde{\mathbf{Q}}_\theta^{-1} \mathbf{X} (\mathbf{X}^\top \tilde{\mathbf{Q}}_\theta^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \tilde{\mathbf{Q}}_\theta^{-1}] \mathbf{Y}\right\}, \end{aligned}$$

where $\tilde{\mathbf{Q}}_\theta = \mathbf{I}_{nn} + \mathbf{Z}\tilde{\mathbf{R}}_\theta\mathbf{Z}^\top$ is the relative covariance (relative to the error variance) of \mathbf{Y} . The REML estimators are now found by equating the partial derivatives of this restricted loglikelihood to zero and solving for σ_ε^2 and θ . The former, given the latter, is:

$$\begin{aligned} \hat{\sigma}_\varepsilon^2 &= \frac{1}{n-p} \mathbf{Y}^\top [\tilde{\mathbf{Q}}_\theta^{-1} - \tilde{\mathbf{Q}}_\theta^{-1} \mathbf{X} (\mathbf{X}^\top \tilde{\mathbf{Q}}_\theta^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \tilde{\mathbf{Q}}_\theta^{-1}] \mathbf{Y} \\ &= \frac{1}{n-p} \mathbf{Y}^\top \tilde{\mathbf{Q}}_\theta^{-1/2} [\mathbf{I}_{nn} - \tilde{\mathbf{Q}}_\theta^{-1/2} \mathbf{X} (\mathbf{X}^\top \tilde{\mathbf{Q}}_\theta^{-1/2} \tilde{\mathbf{Q}}_\theta^{-1/2} \mathbf{X})^{-1} \mathbf{X}^\top \tilde{\mathbf{Q}}_\theta^{-1/2}] \tilde{\mathbf{Q}}_\theta^{-1/2} \mathbf{Y}, \end{aligned}$$

where the rewritten form reveals a projection matrix and, consequently, a residual sum of squares. Like the maximum likelihood estimator of θ , its REML counterpart is generally unknown analytically and to be found numerically. Iterating between the estimation of both parameters until convergence yields the REML estimators. Obviously, REML estimation of the mixed model parameters does not produce an estimate of the fixed parameter β (as it has been integrated out). Should however a point estimate be desired, then in practice the ML estimate of β with the REML estimates of the other parameters is used.

An alternative way to proceed (and insightful for the present purpose) follows the original approach of Henderson, who aimed to construct a linear predictor for \mathbf{Y} .

Definition 4.1

A predictand is the function of the parameters that is to be predicted. A predictor is a function of the data that predicts the predictand. When this latter function is linear in the observation it is said to be a linear predictor.

In case of the mixed model the predictand is $\mathbf{X}_{\text{new}}\beta + \mathbf{Z}_{\text{new}}\gamma$ for $(n_{\text{new}} \times p)$ - and $(n_{\text{new}} \times q)$ -dimensional design matrices \mathbf{X}_{new} and \mathbf{Z}_{new} , respectively. Similarly, the predictor is some function of the data \mathbf{Y} . When it can be expressed as \mathbf{AY} for some matrix \mathbf{A} it is a linear predictor.

The construction of the aforementioned linear predictor requires estimates of β and γ . To obtain these estimates first derive the joint density of (γ, \mathbf{Y}) :

$$\begin{pmatrix} \gamma \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0}_q \\ \mathbf{X}\beta \end{pmatrix}, \begin{pmatrix} \mathbf{R}_\theta & \mathbf{R}_\theta \mathbf{Z}^\top \\ \mathbf{Z}\mathbf{R}_\theta & \sigma_\varepsilon^2 \mathbf{I}_{nn} + \mathbf{Z}\mathbf{R}_\theta\mathbf{Z}^\top \end{pmatrix}\right)$$

From this the likelihood is obtained and after some manipulations the loglikelihood can be shown to be proportional to:

$$\sigma_\varepsilon^{-2} \|\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\gamma\|_2^2 + \gamma^\top \mathbf{R}_\theta^{-1} \gamma, \quad (4.2)$$

in which – following Henderson – \mathbf{R}_θ and σ_ε^2 are assumed known (for instance by virtue of maximum likelihood or REML estimation). The estimators of β and γ are now the minimizers of loss criterion (4.2). Effectively, the random effect parameter γ is now temporarily assumed to be ‘fixed’. That is, it is temporarily treated as fixed in the derivations below that lead to the construction of the linear predictor. However, γ is a random variable and one therefore speaks of a linear predictor rather than linear estimator.

To find the estimators of β and γ , defined as the minimizer of loss function, equate the partial derivatives of mixed model loss function (4.2) with respect to β and γ to zero. This yields the estimating equations (also referred to as Henderson’s mixed model equations):

$$\begin{aligned} \mathbf{X}^\top \mathbf{Y} - \mathbf{X}^\top \mathbf{X}\beta - \mathbf{X}^\top \mathbf{Z}\gamma &= \mathbf{0}_p, \\ \sigma_\varepsilon^{-2} \mathbf{Z}^\top \mathbf{Y} - \sigma_\varepsilon^{-2} \mathbf{Z}^\top \mathbf{Z}\gamma - \sigma_\varepsilon^{-2} \mathbf{Z}^\top \mathbf{X}\beta - \mathbf{R}_\theta^{-1} \gamma &= \mathbf{0}_q. \end{aligned}$$

Solve each estimating equation for the parameters individually and find:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{Y} - \mathbf{Z}\gamma), \quad (4.3)$$

$$\hat{\gamma} = (\mathbf{Z}^\top \mathbf{Z} + \sigma_\varepsilon^2 \mathbf{R}_\theta^{-1})^{-1} \mathbf{Z}^\top (\mathbf{Y} - \mathbf{X}\beta). \quad (4.4)$$

Note, using the Cholesky decomposition of $\tilde{\mathbf{R}}_\theta$ and applying the Woodbury identity twice (in both directions), that:

$$\begin{aligned} \hat{\gamma} &= (\mathbf{Z}^\top \mathbf{Z} + \tilde{\mathbf{R}}_\theta^{-1})^{-1} \mathbf{Z}^\top (\mathbf{Y} - \mathbf{X}\beta) \\ &= [\tilde{\mathbf{R}}_\theta - \tilde{\mathbf{R}}_\theta \mathbf{Z}^\top (\mathbf{I}_{nn} + \mathbf{Z}\tilde{\mathbf{R}}_\theta \mathbf{Z}^\top)^{-1} \mathbf{Z}\tilde{\mathbf{R}}_\theta] \mathbf{Z}^\top (\mathbf{Y} - \mathbf{X}\beta) \\ &= \mathbf{L}_\theta [\mathbf{I}_{qq} - \mathbf{L}_\theta^\top \mathbf{Z}^\top (\mathbf{I}_{nn} + \mathbf{Z}\mathbf{L}_\theta \mathbf{L}_\theta^\top \mathbf{Z}^\top)^{-1} \mathbf{Z}\mathbf{L}_\theta] \mathbf{L}_\theta^\top \mathbf{Z}^\top (\mathbf{Y} - \mathbf{X}\beta) \\ &= \mathbf{L}_\theta (\mathbf{L}_\theta^\top \mathbf{Z}^\top \mathbf{Z}\mathbf{L}_\theta + \mathbf{I}_{qq})^{-1} \mathbf{L}_\theta^\top \mathbf{Z}^\top (\mathbf{Y} - \mathbf{X}\beta) \\ &= \mathbf{L}_\theta \boldsymbol{\mu}_{\tilde{\gamma}|\mathbf{Y}}. \end{aligned}$$

It thus coincides with the conditional estimate of the γ found in the derivation of the maximum likelihood estimator of the mixed model. This expression could also have been found by conditioning with the multivariate normal above which would have given $\mathbb{E}(\gamma | \mathbf{Y})$.

The estimator of both β and γ can be expressed fully and explicitly in terms of \mathbf{X} , \mathbf{Y} , \mathbf{Z} and \mathbf{R}_θ . To obtain that of β substitute the estimator of γ of equation (4.4) into that of β given by equation 4.3):

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top [\mathbf{Y} - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z} + \sigma_\varepsilon^2 \mathbf{R}_\theta^{-1})^{-1} \mathbf{Z}^\top (\mathbf{Y} - \mathbf{X}\beta)] \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \{ \mathbf{Y} - [\mathbf{I}_{nn} - (\sigma_\varepsilon^{-2} \mathbf{Z}\mathbf{R}_\theta \mathbf{Z}^\top + \mathbf{I}_{nn})^{-1}] (\mathbf{Y} - \mathbf{X}\beta) \}, \end{aligned}$$

in which the Woodbury identity has been used. Now group terms and solve for β :

$$\hat{\beta} = [\mathbf{X}^\top (\mathbf{Z}\mathbf{R}_\theta \mathbf{Z}^\top + \sigma_\varepsilon^2 \mathbf{I}_{nn})^{-1} \mathbf{X}]^{-1} \mathbf{X}^\top (\mathbf{Z}\mathbf{R}_\theta \mathbf{Z}^\top + \sigma_\varepsilon^2 \mathbf{I}_{nn})^{-1} \mathbf{Y}. \quad (4.5)$$

This coincides with the maximum likelihood estimator of β presented above (for known \mathbf{R}_θ and σ_ε^2). Moreover, in the preceding display one recognizes a generalized least squares (GLS) estimator. The GLS regression estimator is BLUE (Best Linear Unbiased Estimator) when \mathbf{R}_θ and σ_ε are known. To find an explicit expression for γ use \mathbf{Q}_θ as previously defined and substitute the explicit expression (4.5) for the estimator of β in the estimator of γ , shown in display (4.4) above. This gives:

$$\hat{\gamma} = (\mathbf{Z}^\top \mathbf{Z} + \sigma_\varepsilon^2 \mathbf{R}_\theta^{-1})^{-1} \mathbf{Z}^\top [\mathbf{I}_{nn} - \mathbf{X}(\mathbf{X}^\top \mathbf{Q}_\theta^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Q}_\theta^{-1}] \mathbf{Y},$$

an explicit expression for the estimator of γ .

The linear predictor constructed from these estimator can be shown (cf. Theorem 4.1) to be optimal, in the BLUP sense.

Definition 4.2

A Best Linear Unbiased Predictor (BLUP) *i*) is linear in the observations, *ii*) is unbiased, and *iii*) has a minimum (variance of its) predictor error, i.e. the difference among the predictor and predictand, among all unbiased linear predictors.

Theorem 4.1

The predictor $\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\gamma}$ is the BLUP of $\tilde{\mathbf{Y}} = \mathbf{X}\beta + \mathbf{Z}\gamma$.

Proof. The predictor of $\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\gamma}$ is:

$$\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\gamma} = [\mathbf{I}_{nn} - \sigma_\varepsilon^2 \mathbf{Q}_\theta^{-1} + \mathbf{Q}_\theta^{-1} \mathbf{X}(\mathbf{X}^\top \mathbf{Q}_\theta^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Q}_\theta^{-1}] \mathbf{Y} := \mathbf{B}\mathbf{Y}.$$

Clearly, this is a linear function in \mathbf{Y} .

The expectation of the linear predictor is

$$\begin{aligned} \mathbb{E}(\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\gamma}) &= \mathbb{E}[\mathbf{X}\hat{\beta} + \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z} + \sigma_\varepsilon^2 \mathbf{R}_\theta)^{-1} \mathbf{Z}^\top (\mathbf{Y} - \mathbf{X}\hat{\beta})] \\ &= \mathbf{X}\mathbb{E}(\hat{\beta}) + (\mathbf{I}_{nn} - \sigma_\varepsilon^2 \mathbf{Q}_\theta^{-1})[\mathbb{E}(\mathbf{Y}) - \mathbb{E}(\mathbf{X}\hat{\beta})] = \mathbf{X}\beta. \end{aligned}$$

This is also the expectation of the predictand $\mathbf{X}\beta + \mathbf{Z}\gamma$. Hence, the predictor is unbiased.

To show the predictor \mathbf{BY} has minimum prediction error variance within the class of unbiased linear predictors, assume the existence of another unbiased linear predictor \mathbf{AY} of $\mathbf{X}\beta + \mathbf{Z}\gamma$. The predictor error variance of the latter predictor is:

$$\begin{aligned}\text{Var}(\mathbf{X}\beta + \mathbf{Z}\gamma - \mathbf{AY}) &= \text{Var}(\mathbf{X}\beta + \mathbf{Z}\gamma - \mathbf{BY} - \mathbf{AY} + \mathbf{BY}) \\ &= \text{Var}[(\mathbf{A} - \mathbf{B})\mathbf{Y}] + \text{Var}(\mathbf{X}\beta + \mathbf{Z}\gamma - \mathbf{BY}) \\ &\quad - 2 \text{Cov}[\mathbf{X}\beta + \mathbf{Z}\gamma - \mathbf{BY}, (\mathbf{A} - \mathbf{B})\mathbf{Y}].\end{aligned}$$

The last term vanishes as:

$$\begin{aligned}\text{Cov}[\mathbf{X}\beta + \mathbf{Z}\gamma - \mathbf{BY}, (\mathbf{A} - \mathbf{B})\mathbf{Y}] &= [\mathbf{Z}\text{Cov}(\gamma, \mathbf{Y}) - \mathbf{B}\text{Var}(\mathbf{Y})](\mathbf{A} - \mathbf{B})^\top \\ &= \{\mathbf{Z}\mathbf{R}_\theta\mathbf{Z}^\top - [\mathbf{I}_{nn} - \sigma_\varepsilon^2 \mathbf{Q}_\theta^{-1} + \mathbf{Q}_\theta^{-1} \mathbf{X}(\mathbf{X}\mathbf{Q}_\theta^{-1}\mathbf{X}^\top)^{-1} \mathbf{X}^\top \mathbf{Q}_\theta^{-1}]\mathbf{Q}_\theta\}(\mathbf{A} - \mathbf{B})^\top \\ &= \mathbf{Q}_\theta^{-1} \mathbf{X}(\mathbf{X}\mathbf{Q}_\theta^{-1}\mathbf{X}^\top)^{-1} \mathbf{X}^\top (\mathbf{A} - \mathbf{B})^\top \\ &= \mathbf{Q}_\theta^{-1} \mathbf{X}(\mathbf{X}\mathbf{Q}_\theta^{-1}\mathbf{X}^\top)^{-1} [(\mathbf{A} - \mathbf{B})\mathbf{X}]^\top = \mathbf{0}_{nn},\end{aligned}$$

where the last step uses $\mathbf{AX} = \mathbf{BX}$, which follows from the fact that

$$\mathbf{AX}\beta = \mathbb{E}(\mathbf{AY}) = \mathbb{E}(\mathbf{BY}) = \mathbf{BX}\beta,$$

for all $\beta \in \mathbb{R}^p$. Hence,

$$\text{Var}(\mathbf{X}\beta + \mathbf{Z}\gamma - \mathbf{AY}) = \text{Var}[(\mathbf{A} - \mathbf{B})\mathbf{Y}] + \text{Var}(\mathbf{X}\beta + \mathbf{Z}\gamma - \mathbf{BY}),$$

from which the minimum variance follows as the first summand on the right-hand side is nonnegative and zero if and only if $\mathbf{A} = \mathbf{B}$. ■

4.1 Link to ridge regression

The link with ridge regression, implicit in the exposé on the mixed model, is now explicated. Recall that ridge regression fits the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ by means of a penalized maximum likelihood procedure, which defines – for given penalty parameter λ – the estimator as:

$$\hat{\beta}(\lambda) = \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \beta^\top \beta.$$

Contrast this to a mixed model void of covariates with fixed effects and comprising only covariates with a random effects: $\mathbf{Y} = \mathbf{Z}\gamma + \varepsilon$ with distributional assumptions $\gamma \sim \mathcal{N}(\mathbf{0}_q, \sigma_\gamma^2 \mathbf{I}_{qq})$ and $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma_\varepsilon^2 \mathbf{I}_{nn})$. This model, when temporarily considering γ as fixed, is fitted by the minimization of loss function (4.2). The corresponding estimator of γ is then defined, with the current mixed model assumptions in place, as:

$$\hat{\gamma} = \arg \min_{\gamma \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{Z}\gamma\|_2^2 + \sigma_\gamma^{-2} \gamma^\top \gamma.$$

The estimators are – up to a reparametrization of the penalty parameter – defined identically. This should not come as a surprise after the discussion of Bayesian regression (cf. Chapter 2) and the alert reader would already have recognized a generalized ridge loss function in Equation (4.2). The fact that we discarded the fixed effect part of the mixed model is irrelevant for the analogy as those would correspond to unpenalized covariates in the ridge regression problem.

The link with ridge regression is also imminent from the linear predictor of the random effect. Recall: $\hat{\gamma} = (\mathbf{Z}^\top \mathbf{Z} + \mathbf{R}_\theta^{-1})^{-1} \mathbf{Z}^\top (\mathbf{Y} - \mathbf{X}\beta)$. When we ignore \mathbf{R}_θ^{-1} , the predictor reduces to a least squares estimator. But with a symmetric and positive definite matrix \mathbf{R}_θ^{-1} , the predictor is actually of the shrinkage type as is the ridge regression estimator. This shrinkage estimator also reveals, through the term $(\mathbf{Z}^\top \mathbf{Z} + \mathbf{R}_\theta^{-1})^{-1}$, that a q larger than n does not cause identifiability problems as long as \mathbf{R}_θ is parametrized low-dimensionally enough.

The following mixed model result provide an alternative approach to choice of the penalty parameter in ridge regression. It assumes a mixed model comprising of the random effects part only. Or, put differently, it assumes the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\beta \sim \mathcal{N}(\mathbf{0}_p, \sigma_\beta^2 \mathbf{I}_{pp})$ and $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma_\varepsilon^2 \mathbf{I}_{nn})$.

Theorem 4.2 (Theorem 2, Golub et al)

The expected generalized cross-validation error $\mathbb{E}_\beta \{ \mathbb{E}_\varepsilon [GCV(\lambda)] \}$ is minimized for $\lambda = \sigma_\varepsilon^2 / \sigma_\beta^2$.

Proof. The proof first finds an analytic expression of the expected $GCV(\lambda)$, then its minimum. Its expectation can be re-expressed as follows:

$$\begin{aligned}\mathbb{E}_{\beta}\{\mathbb{E}_{\varepsilon}[GCV(\lambda)]\} &= \mathbb{E}_{\beta}\left[\mathbb{E}_{\varepsilon}\left(\frac{1}{n}\{\text{tr}[\mathbf{I}_{nn} - \mathbf{H}(\lambda)]/n\}^{-2}\|\mathbf{I}_{nn} - \mathbf{H}(\lambda)\mathbf{Y}\|_2^2\right)\right] \\ &= n\{\text{tr}[\mathbf{I}_{nn} - \mathbf{H}(\lambda)]\}^{-2}\mathbb{E}_{\beta}\left(\mathbb{E}_{\varepsilon}\{\mathbf{Y}^\top [\mathbf{I}_{nn} - \mathbf{H}(\lambda)]^2 \mathbf{Y}\}\right) \\ &= n\{\text{tr}[\mathbf{I}_{nn} - \mathbf{H}(\lambda)]\}^{-2}\mathbb{E}_{\beta}\left[\mathbb{E}_{\varepsilon}\left(\text{tr}\{(\mathbf{X}\beta + \varepsilon)^\top [\mathbf{I}_{nn} - \mathbf{H}(\lambda)]^2 (\mathbf{X}\beta + \varepsilon)\}\right)\right] \\ &= n\{\text{tr}[\mathbf{I}_{nn} - \mathbf{H}(\lambda)]\}^{-2}\left(\text{tr}\{[\mathbf{I}_{nn} - \mathbf{H}(\lambda)]^2 \mathbf{X} \mathbf{X}^\top \mathbb{E}_{\beta}[(\beta\beta^\top)]\}\right. \\ &\quad \left.+ \text{tr}\{[\mathbf{I}_{nn} - \mathbf{H}(\lambda)]^2 \mathbb{E}_{\beta}[\varepsilon\varepsilon^\top]\}\right) \\ &= n(\sigma_{\beta}^2 \text{tr}\{[\mathbf{I}_{nn} - \mathbf{H}(\lambda)]^2 \mathbf{X} \mathbf{X}^\top\} + \sigma_{\varepsilon}^2 \text{tr}\{[\mathbf{I}_{nn} - \mathbf{H}(\lambda)]^2\})\{\text{tr}[\mathbf{I}_{nn} - \mathbf{H}(\lambda)]\}^{-2}.\end{aligned}$$

To get a handle on this expression, use $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{X} = \mathbf{I}_{pp} - \lambda(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} = \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}$, the cyclic property of the trace, and define $A(\lambda) = \sum_{j=1}^p (d_{x,j}^2 + \lambda)^{-1}$, $B(\lambda) = \sum_{j=1}^p (d_{x,j}^2 + \lambda)^{-2}$, and $C(\lambda) = \sum_{j=1}^p (d_{x,j}^2 + \lambda)^{-3}$. The traces in the expectation of $GCV(\lambda)$ can now be written as:

$$\begin{aligned}\text{tr}[\mathbf{I}_{nn} - \mathbf{H}(\lambda)] &= \lambda \text{tr}[(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}] = \lambda A(\lambda), \\ \text{tr}\{[\mathbf{I}_{nn} - \mathbf{H}(\lambda)]^2\} &= \lambda^2 \text{tr}[(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-2}] = \lambda^2 B(\lambda), \\ \text{tr}\{[\mathbf{I}_{nn} - \mathbf{H}(\lambda)]^2 \mathbf{X} \mathbf{X}^\top\} &= \lambda^2 \text{tr}[(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}] - \lambda^3 \text{tr}[(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{pp})^{-2}] \\ &= \lambda^2 A(\lambda) - \lambda^3 B(\lambda).\end{aligned}$$

The expectation of $GCV(\lambda)$ can then be reformulated as:

$$\mathbb{E}_{\beta}\{\mathbb{E}_{\varepsilon}[GCV(\lambda)]\} = n\{\sigma_{\beta}^2[A(\lambda) - \lambda B(\lambda)] + \sigma_{\varepsilon}^2 B(\lambda)\}[A(\lambda)]^{-2}.$$

Equate the derivative of this expectation w.r.t. λ to zero, which can be seen to be proportional to:

$$2(\lambda\sigma_{\beta}^2 - \sigma_{\varepsilon}^2)[B(\lambda)]^2 + 2(\lambda\sigma_{\beta}^2 - \sigma_{\varepsilon}^2)A(\lambda)C(\lambda) = 0.$$

Indeed, $\lambda = \sigma_{\varepsilon}^2/\sigma_{\beta}^2$ is the root of this equation. ■

Theorem 4.2 can be extended to include unpenalized covariates. This leaves the result unaltered: the optimal (in the expected GCV sense) ridge penalty is the same signal-to-noise ratio.

We have encountered the result of Theorem 4.2 before. Revisit Example 1.6 which derived the mean squared error (MSE) of the ridge regression estimator when \mathbf{X} is orthonormal. It was pointed out that this MSE is minimized for $\lambda = p\sigma_{\varepsilon}/\beta^\top \beta$. As $\beta^\top \beta/p$ is an estimator for σ_{β}^2 , this implies the same optimal choice of the penalty parameter.

To point out the relevance of Theorem 4.2 for the choice of the ridge penalty parameter still assume the regression parameter random. The theorem then says that the optimal penalty parameter (in the GCV sense) equals the ratio of the error variance and that of the regression parameter. Both variances can be estimated by means of the mixed model machinery (provided for instance by the `lme4` package in R). These estimates may be plugged in the ratio to arrive at a choice of ridge penalty parameter (see Section 4.3 for an illustration of this usage).

4.2 REML consistency, high-dimensionally

Here a result on the asymptotic quality of the REML estimators of the random effect and error variance parameters is presented and discussed. It is the ratio of these parameters that forms the optimal choice (in the expected GCV sense) of the penalty parameter of the ridge regression estimator. As in practice the parameters are replaced by estimates to arrive at a choice for the penalty parameter, the quality of these estimators propagates to the chosen penalty parameter.

Consider the standard linear mixed model $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon$, now with equivariant and uncorrelated random effects: $\gamma \sim \mathcal{N}(\mathbf{0}_q, \sigma_{\gamma}^2 \mathbf{I}_{qq})$. Write $\theta = \sigma_{\gamma}^2/\sigma_{\varepsilon}^2$. The REML estimators of θ and σ_{ε}^2 are to be found from the estimating equations:

$$\begin{aligned}\text{tr}(\mathbf{P}_{\theta} \mathbf{Z} \mathbf{Z}^\top) &= \sigma_{\varepsilon}^{-2} \text{tr}(\mathbf{Y}^\top \mathbf{P}_{\theta} \mathbf{Z} \mathbf{Z}^\top \mathbf{P}_{\theta} \mathbf{Y}), \\ \sigma_{\varepsilon}^2 &= (n-p)^{-1} \mathbf{Y}^\top \mathbf{P}_{\theta} \mathbf{Y},\end{aligned}$$

where $\mathbf{P}_\theta = \tilde{\mathbf{Q}}_\theta^{-1} - \tilde{\mathbf{Q}}_\theta^{-1}\mathbf{X}(\mathbf{X}^\top\tilde{\mathbf{Q}}_\theta^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\tilde{\mathbf{Q}}_\theta^{-1}$ and $\tilde{\mathbf{Q}}_\theta = \mathbf{I}_{nn} + \theta\mathbf{Z}\mathbf{Z}^\top$. To arrive at the REML estimators choose initial values for the parameters. Choose one of the estimating equations substitute the initial value of the one of the parameters and solve for the other. The found root is then substituted into the other estimating equation, which is subsequently solved for the remaining parameter. Iterate between these two steps until convergence. The discussion of the practical evaluation of a root for θ from these estimating equations in a high-dimensional context is postponed to the next section.

The employed linear mixed model assumes that each of the q covariates included as a column in \mathbf{Z} contributes to the variation of the response. However, it may be that only a fraction of these covariates exerts any influence on the response. That is, the random effect parameter γ is sparse, which could be operationalized as γ having q_0 zero elements while the remaining $q_c = q - q_0$ elements are non-zero. Only for the latter q_c elements of γ the normal assumption makes sense, but is invalid for the q_0 zeros in γ . The posed mixed model is then misspecified.

The next theorem states that the REML estimators of $\theta = \sigma_\gamma^2/\sigma_\varepsilon^2$ and σ_ε^2 are consistent (possibly after adjustment, see the theorem), even under the above mentioned misspecification.

Theorem 4.3 (Theorem 3.1, Jiang *et al.* (2016))

Let \mathbf{Z} be standardized column-wise and with its unstandardized entries i.i.d. from a sub-Gaussian distribution. Furthermore, assume that $n, q, q_c \rightarrow \infty$ such that

$$\frac{n}{q} \rightarrow \tau \quad \text{and} \quad \frac{q_c}{q} \rightarrow \omega,$$

where $\tau, \omega \in (0, 1]$. Finally, suppose that σ_ε^2 and σ_γ^2 are positive. Then:

i) The ‘adjusted’ REML estimator of the variance ratio $\sigma_\gamma^2/\sigma_\varepsilon^2$ is consistent:

$$\frac{q}{q_c} (\widehat{\sigma_\gamma^2/\sigma_\varepsilon^2}) \xrightarrow{P} \sigma_\gamma^2/\sigma_\varepsilon^2.$$

ii) The REML estimator of the error variance is consistent: $\hat{\sigma}_\varepsilon^2 \xrightarrow{P} \sigma_\varepsilon^2$.

Proof. Confer Jiang *et al.* (2016). ■

Before the interpretation and implication of Theorem 4.3 are discussed, its conditions for the consistency result are reviewed:

- The standardization and distribution assumption on the design matrix of the random effects has no direct practical interpretation. These conditions warrant the applicability of certain results from random matrix theory upon which the proof of the theorem hinges.
- The positive variance assumption $\sigma_\varepsilon^2, \sigma_\gamma^2 > 0$, in particular that of the random effect parameter, effectively states that some – possibly misspecified – form of the mixed model applies.
- Practically most relevant are the conditions on the sample size, random effect dimension, and sparsity. The τ and ω in Theorem 4.3 are the limiting ratio’s of the sample size n and non-zero random effects q_c , respectively, to the total number of random effects q . The number of random effects thus exceeds the sample size, as long as the latter grows (in the limit) at some fixed rate with the former. Independently, the model may be misspecified. The sparsity condition only requires that (in the limit) a fraction of the random effects is nonzero.

Now discuss the interpretation and relevance of the theorem:

- Theorem 4.3 complements the classical low-dimensional consistency results on the REML estimator.
- Theorem 4.3 shows that not all (i.e. consistency) is lost when the model is misspecified.
- The practical relevance of the part *i*) of Theorem 4.3 is limited as the number of nonzero random effects q_c , or ω for that matter, is usually unknown. Consequently, the REML estimator of the variance ratio $\sigma_\gamma^2/\sigma_\varepsilon^2$ cannot be adjusted correctly to achieve asymptotically unbiasedness and – thereby – consistency
- Part *ii*) in its own right may not seem very useful. But it is surprising that high-dimensionally (i.e. when the dimension of the random effect parameter exceeds the sample size) the standard (that is, derived for low-dimensional data) REML estimator of σ_ε^2 is consistent. Beyond this surprise, a good estimator of σ_ε^2 indicates how much of the variation in the response cannot be attributed to the covariates represented by the columns \mathbf{Z} . A good indication of the noise level in the data finds use at many place. In particular, it is helpful in deciding on the order of the penalty parameter.
- Theorem 4.2 suggests to choose the ridge penalty parameter equal to the ratio of the error variance and that of the random effects. Confronted with data the reciprocal of the REML estimator of $\theta = \sigma_\gamma^2/\sigma_\varepsilon^2$ may be

used as value for the penalty parameter. Without the adjustment for the fraction of nonzero random effects, this value is off. But in the worst case this value is an over-estimation of the optimal (in the GCV sense) ridge penalty parameter. Consequently, too much penalization is applied and the ridge regression estimate of the regression parameter is conservative as it shrinks the elements too much to zero.

4.3 Illustration: P-splines

An organism's internal circadian clock enables it to synchronize its activities to the earth's day-and-night cycle. The circadian clock maintains, due to environmental feedback, oscillations of approximately 24 hours. Molecularly, these oscillations reveal themselves in the fluctuation of the transcription levels of genes. The molecular core of the circadian clock is made up of ± 10 genes. Their behaviour (in terms of their expression patterns) is described by a dynamical system with feedback mechanisms. Linked to this core are genes that tap into the clock's rhythm and use it to regulate the molecular processes. As such many genes are expected to exhibit circadian rhythms. This is investigated in a mouse experiment in which the expression levels of several transcripts have been measured during two days with a resolution of one hour, resulting in a time-series of 48 time points publicly available from the R-package `MetaCycle`. Circadian rhythms may be identified simply by eye-balling the data. But to facilitate this identification the data are smoothed to emphasize the pattern present in these data.

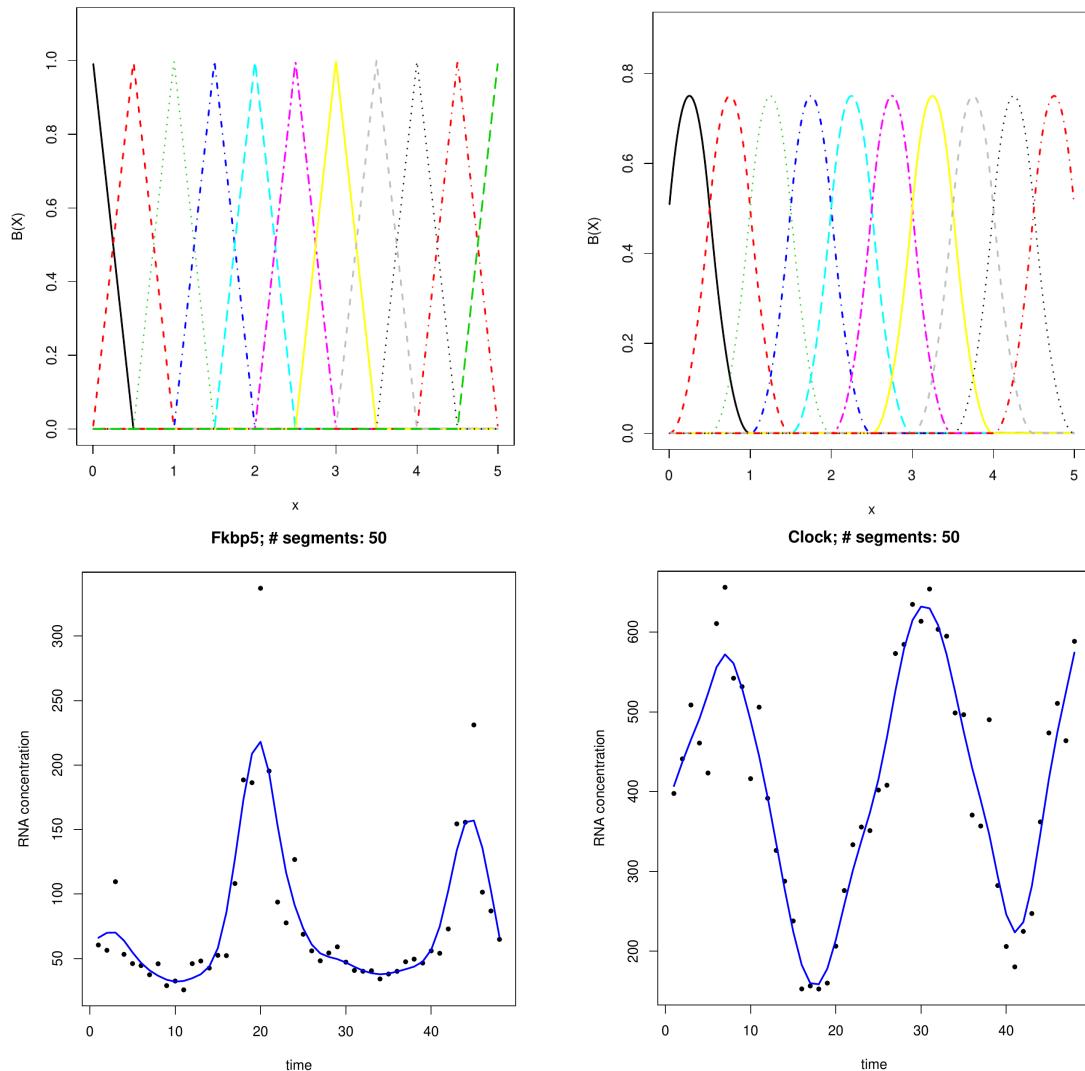


Figure 4.2: Top left and right panels: B-spline basis functions of degree 1 and 2, respectively. Bottom left and right panel: P-spline fit to transcript levels of circadian clock experiment in mice.

Smoothing refers to nonparametric – in the sense that parameters have no tangible interpretation – description

of a curve. For instance, one may wish to learn some general functional relationship between two variable, X and Y , from data. Statistically, the model $Y = f(X) + \varepsilon$, for unknown and general function $f(\cdot)$, is to be fitted to paired observations $\{(y_i, x_i)\}_{i=1}^n$. Here we use P-splines, penalized B-splines with B for Basis (Eilers and Marx, 1996).

A B-spline is formed through a linear combination of (pieces of) polynomial basis functions of degree r . For their construction specify the interval $[x_{\text{start}}, x_{\text{end}}]$ on which the function is to be learned/approximated. Let $\{t_j\}_{j=0}^{m+2r}$ be a grid, overlapping the interval, of equidistantly placed points called knots given by $t_j = x_{\text{start}} + (j - r)h$ for all $j = 0, \dots, m + 2r$ with $h = \frac{1}{m}(x_{\text{end}} - x_{\text{start}})$. The B-spline base functions are then defined as:

$$B_j(x; r) = (-1)^{r+1} (h^r r!)^{-1} \Delta^{r+1}[(x - t_j)^r \mathbb{1}_{\{x \geq t_j\}}]$$

where $\Delta^r[f_j(\cdot)]$ is the r -th difference operator applied to $f_j(\cdot)$. For $r = 1$: $\Delta[f_j(\cdot)] = f_j(\cdot) - f_{j-1}(\cdot)$, while $r = 2$: $\Delta^2[f_j(\cdot)] = \Delta\{\Delta[f_j(\cdot)]\} = \Delta[f_j(\cdot) - f_{j-1}(\cdot)] = f_j(\cdot) - 2f_{j-1}(\cdot) + f_{j-2}(\cdot)$, et cetera. The top right and bottom left panels of Figure 4.2 show a 1st and 2nd degree B-spline basis functions. A P-spline is a curve of the form $\sum_{j=0}^{m+2r} \alpha_j B_j(x; r)$ fitted to the data by means of penalized least squares minimization. The least squares are $\|\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha}\|_2^2$ where \mathbf{B} is a $n \times (m + 2r)$ -dimensional matrix with the j -th column equalling $(B_j(x_1; r), B_j(x_2; r), \dots, B_j(x_n; r))^T$. The employed penalty is of the ridge type: the sum of the squared difference among contiguous α_j . Let \mathbf{D} be the first order differencing matrix. The penalty can then be written as $\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 = \sum_{j=2}^{m+2r} (\alpha_j - \alpha_{j-1})^2$. A second order difference matrix would amount to $\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 = \sum_{j=3}^{m+2r} (\alpha_j - 2\alpha_{j-1} + \alpha_{j-2})^2$. Eilers (1999) points out how P-splines may be interpret as a mixed model. Hereto choose $\tilde{\mathbf{X}}$ such that its columns span the null space of $\mathbf{D}^\top \mathbf{D}$, which comprises a single column representing the intercept when \mathbf{D} is a first order differencing matrix, and $\tilde{\mathbf{Z}} = \mathbf{D}^\top (\mathbf{D}\mathbf{D}^\top)^{-1}$. Then, for any $\boldsymbol{\alpha}$:

$$\mathbf{B}\boldsymbol{\alpha} = \mathbf{B}(\tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{Z}}\boldsymbol{\gamma}) := \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}.$$

This parametrization simplifies the employed penalty to:

$$\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 = \|\mathbf{D}(\tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{Z}}\boldsymbol{\gamma})\|_2^2 = \|\mathbf{D}\mathbf{D}^\top(\mathbf{D}\mathbf{D}^\top)^{-1}\boldsymbol{\gamma}\|_2^2 = \|\boldsymbol{\gamma}\|_2^2,$$

where $\mathbf{D}\tilde{\mathbf{X}}\boldsymbol{\beta}$ has vanished by the construction of $\tilde{\mathbf{X}}$. Hence, the penalty only affects the random effect parameter, leaving the fixed effect parameter unshrunken. The resulting loss function, $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}\|_2^2 + \lambda\|\boldsymbol{\gamma}\|_2^2$, coincides for suitably chosen λ to that of the mixed model (as will become apparent later). The bottom panels of Figure 4.2 shows the flexibility of this approach.

The following R-script fits a P-spline to a gene's transcript levels of the circadian clock study in mice. It uses a basis of $m = 50$ truncated polynomial functions of degree $r = 3$ (cubic), which is generated first alongside several auxillary matrices. This basis forms, after post-multiplication with a projection matrix onto the space spanned by the columns of the difference matrix \mathbf{D} , the design matrix for the random coefficient of the mixed model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \varepsilon$ with $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}_q, \sigma_\gamma^2 \mathbf{I}_{qq})$ and $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma_\varepsilon^2 \mathbf{I}_{nn})$. The variance parameters of this model are then estimated by means of restricted maximum likelihood (REML). The final P-spline fit is obtained from the linear predictor using, in line with Theorem 4.2, $\lambda = \sigma_\varepsilon^2 / \sigma_\gamma^2$ in which the REML estimates of these variance parameters are substituted. The resulting P-spline fit of two transcripts is shown in the bottom panels of Figure 4.2.

Listing 4.1 R code

```
# load libraries
library(gridExtra)
library(MetaCycle)
library(MASS)

#-----
# intermezzo: declaration of functions used analysis
#-----

tpower <- function(x, knots, p) {
  # function for evaluation of truncated p-th
  # power functions positions x, given knots
  return((x - knots)^p * (x > knots))
}
```

```

bbase <- function(x, m, r) {
  # function for B-spline basis generation
  # evaluated at the extremes of x
  # with m segments and spline degree r
  h      <- (max(x) - min(x))/m
  knots <- min(x) + (c(0:(m+2*r)) - r) * h
  P      <- outer(x, knots, tpower, r)
  D      <- diff(diag(m+2*r+1), diff=r+1) / (gamma(r+1) * h^r)
  return((-1)^(r+1) * P %*% t(D))
}

thetaEstEqREML <- function(theta, Z, Y, X, sigma2e) {
  # function for REML estimation:
  # estimating equation of theta
  QthetaInv <- solve(diag(length(Y)) + theta * Z %*% t(Z))
  Ptheta    <- QthetaInv -
              QthetaInv %*% X %*%
              solve(t(X) %*% QthetaInv %*% X) %*% t(X) %*% QthetaInv
  return(sum(diag(Ptheta %*% Z %*% t(Z))) -
         as.numeric(t(Y) %*% Ptheta %*% Z %*% t(Z) %*% Ptheta %*% Y) / sigma2e)
}

#-----

# load data
data(cycMouseLiverRNA)
id <- 14
Y <- as.numeric(cycMouseLiverRNA[id,-1])
X <- 1:length(Y)

# set P-spline parameters
m <- 50
r <- 3
B <- bbase(X, m=m, r=r)

# prepare some matrices
D <- diff(diag(m+r), diff = 2)
Z <- B %*% t(D) %*% solve(D %*% t(D))
X <- B %*% Null(t(D) %*% D)

# initiate
theta <- 1
for (k in 1:100) {
  # for-loop, alternating between theta and error variance estimation
  thetaPrev <- theta
  QthetaInv <- solve(diag(length(Y)) + theta * Z %*% t(Z))
  Ptheta    <- QthetaInv -
              QthetaInv %*% X %*%
              solve(t(X) %*% QthetaInv %*% X) %*% t(X) %*% QthetaInv
  sigma2e   <- t(Y) %*% Ptheta %*% Y / (length(Y)-2)
  theta     <- uniroot(thetaEstEqREML, c(0, 100000),
                        Z=Z, Y=Y, X=X, sigma2e=sigma2e)$root
  if (abs(theta - thetaPrev) < 10^(-5)) { break }
}

# P-spline fit
bgHat <- solve(t(cbind(X, Z)) %*% cbind(X, Z) +
                diag(c(rep(0, ncol(X)), rep(1/theta, ncol(Z))))) %*%
                t(cbind(X, Z)) %*% Y

```

```
# plot fit
plot(Y, pch=20, xlab="time", ylab="RNA_concentration",
  main=paste(strsplit(cycMouseLiverRNA[id,1], "_")[[1]][1],
  "#segments:", m, sep=""))
lines(cbind(X, Z) %*% bgHat, col="blue", lwd=2)
```

The fitted splines displayed Figure 4.2 nicely match the data. From the circadian clock perspective it is especially the fit in the right-hand side bottom panel that displays the archetypical sinusoidal behaviour associated by the layman with the sought-for rythm. Close inspection of the fits reveals some minor discontinuities in the derivative of the spline fit. These minor discontinuities are indicative of a little overfitting, due to too large an estimate of σ_γ^2 . This appears to be due to numerical instability of the solution of the estimating equations of the REML estimators of the mixed model's variance parameter estimators when m is large compared to the sample size n .

5 Ridge logistic regression

Ridge penalized estimation is not limited to the standard linear regression model, but may be used to estimate (virtually) any model. Here we illustrate how it may be used to fit the logistic regression model. To this end we first recap this model and the (unpenalized) maximum likelihood estimation of its parameters. After which the model is estimated by means of ridge penalized maximum likelihood, which will turn out to be a relatively straightforward modification of unpenalized estimation.

5.1 Logistic regression

The logistic regression model explains a binary response variable (through some transformation) by a linear combination of a set of covariates (as in the linear regression model). Denote this response of the i -th sample by Y_i with $Y_i \in \{0, 1\}$ for $i = 1, \dots, n$. The n -dimensional column vector \mathbf{Y} stacks these n responses. For each sample information on the p explanatory variables $X_{i,1}, \dots, X_{i,p}$ is available. In row vector form this information is denoted $\mathbf{X}_{i,*} = (X_{i,1}, \dots, X_{i,p})$. The $(n \times p)$ -dimensional matrix \mathbf{X} aggregates these vectors, such that $\mathbf{X}_{i,*}$ is the i -th row vector.

The binary response cannot be modelled as in the linear model like $Y_i = \mathbf{X}_{i,*}\boldsymbol{\beta} + \varepsilon_i$. With each element of $\mathbf{X}_{i,*}$ and $\boldsymbol{\beta}$ assuming a value in \mathbb{R}^p , the linear predictor is not restricted to the domain of the response. This is resolved by modeling $p_i = P(Y_i = 1)$ instead. Still the linear predictor may exceed the domain of the response ($p_i \in [0, 1]$). Hence, a transformation is applied to map p_i to \mathbb{R} , the range of the linear predictor. The transformation associated with the logistic regression model is the logarithm of the odds, with the odds defined as: $odds = P(\text{succes})/P(\text{failure}) = p_i/(1 - p_i)$. The logistic model is then written as $\log[p_i/(1 - p_i)] = \mathbf{X}_{i,*}\boldsymbol{\beta}$ for all i . Or, expressed in terms of the response:

$$p_i = P(Y_i = 1) = g^{-1}(\mathbf{X}_{i,*}; \boldsymbol{\beta}) = \exp(\mathbf{X}_{i,*}\boldsymbol{\beta})[1 + \exp(\mathbf{X}_{i,*}\boldsymbol{\beta})]^{-1}.$$

The function $g(\cdot; \cdot)$ is called the *link function*. It links the response to the explanatory variables. The one above is called the logistic link function. Or short, logit. The regression parameters have tangible interpretations. When the first covariate represents the intercept, i.e. $X_{i,j} = 1$ for all i , then β_1 determines where the link function equals a half when all other covariates fail to contribute to the linear predictor (i.e. where $P(Y_i = 1 | \mathbf{X}_{i,*}) = 0.5$ when $\mathbf{X}_{i,*}\boldsymbol{\beta} = \beta_1$). This is illustrated in the top-left panel of Figure 5.1 for various choices of the intercept. On the other hand, the regression parameters are directly related to the odds ratio: $odds\ ratio = odds(X_{i,j} = 1)/odds(X_{i,j} = 0) = \exp(\beta_j)$. Hence, the effect of a unit change in the j -th covariate on the odds ratio is $\exp(\beta_j)$ (see Figure 5.1, top-right panel). Other link functions (depicted in Figure 5.1, bottom-left panel) are common, e.g. the *probit*: $p_i = \Phi_{0,1}(\mathbf{X}_{i,*}\boldsymbol{\beta})$; the *cloglog*: $p_i = \frac{1}{\pi} \arctan(\mathbf{X}_{i,*}\boldsymbol{\beta}) + \frac{1}{2}$; the *Cauchit*: $p_i = \exp[-\exp(\mathbf{X}_{i,*}\boldsymbol{\beta})]$. All these link functions are invertible. Irrespective of the choice of the link function, the binary data are thus modelled as $Y_i \sim \mathcal{B}[g^{-1}(\mathbf{X}_{i,*}; \boldsymbol{\beta}), 1]$. That is, as a single draw from the Binomial distribution with success probability $g^{-1}(\mathbf{X}_{i,*}; \boldsymbol{\beta})$.

Let us now estimate the parameter of the logistic regression model by means of the maximum likelihood method. The likelihood of the experiment is then:

$$L(\mathbf{Y} | \mathbf{X}; \boldsymbol{\beta}) = \prod_{i=1}^n [P(Y_i = 1 | \mathbf{X}_{i,*})]^{Y_i} [P(Y_i = 0 | \mathbf{X}_{i,*})]^{1-Y_i}.$$

After taking the logarithm and some ready algebra, the log-likelihood is found to be:

$$\mathcal{L}(\mathbf{Y} | \mathbf{X}; \boldsymbol{\beta}) = \sum_{i=1}^n \{Y_i \mathbf{X}_{i,*}\boldsymbol{\beta} - \log[1 + \exp(\mathbf{X}_{i,*}\boldsymbol{\beta})]\}.$$

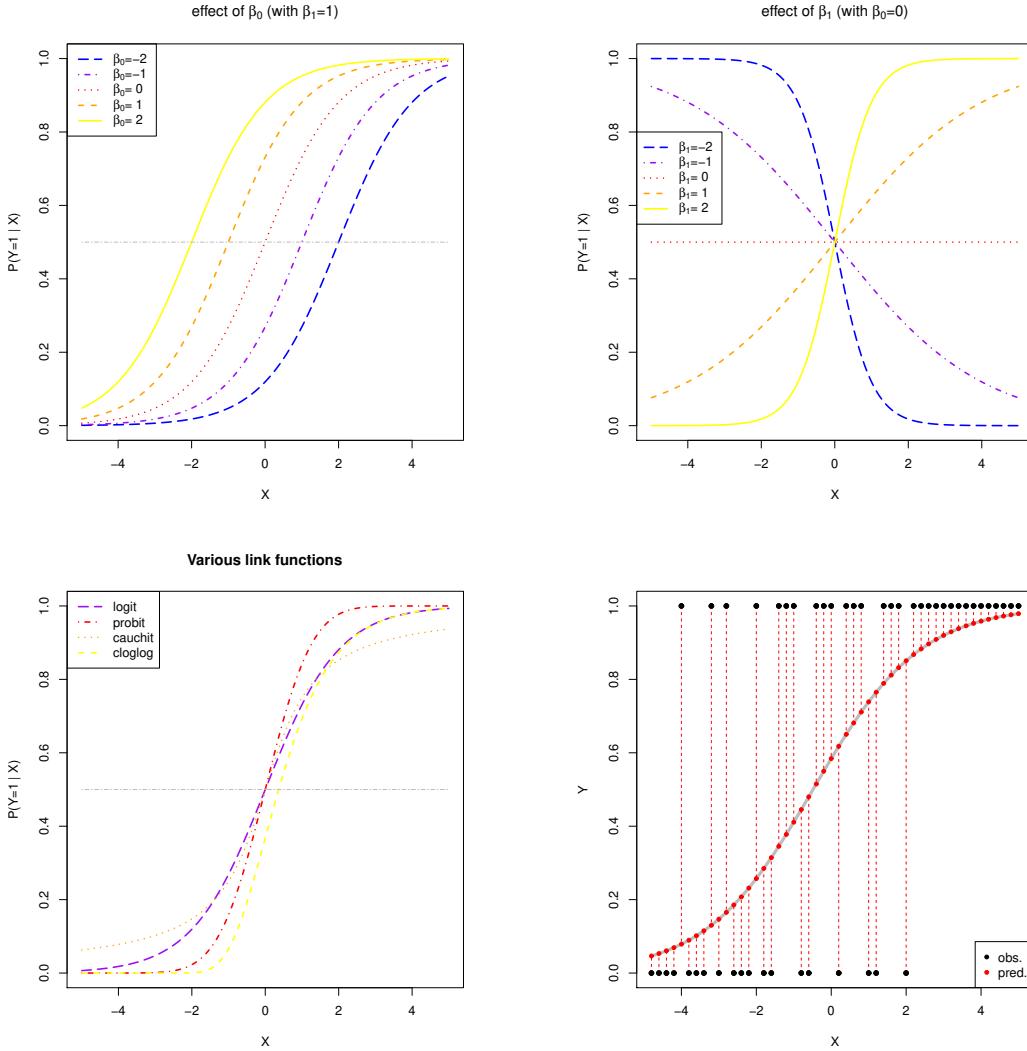


Figure 5.1: Top row, left panel: the response curve for various choices of the intercept β_0 . Top row, right panel: the response curve for various choices of the regression coefficient β_1 . Bottom row, left panel: the response curve for various choices of the link function. Bottom panel, right panel: observations, fits and their deviations.

Differentiate the log-likelihood with respect to β , equate it zero, and obtain the estimating equation for β :

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum_{i=1}^n \left[Y_i - \frac{\exp(\mathbf{X}_{i,*}\beta)}{1 + \exp(\mathbf{X}_{i,*}\beta)} \right] \mathbf{X}_{i,*}^\top = \mathbf{0}_p. \quad (5.1)$$

The ML estimate of β strikes a (weighted by the $\mathbf{X}_{i,*}$) balance between observation and model. Put differently (and illustrated in the bottom-right panel of Figure 5.1), a curve is fit through data by minimizing the distance between them: at the ML estimate of β a weighted average of their deviations is zero.

The maximum likelihood estimate of β is evaluated by solving Equation (5.1) with respect to β by means of the Newton-Raphson algorithm. The Newton-Raphson algorithm iteratively finds the zeros of a smooth enough function $f(\cdot)$. Let x_0 denote an initial guess of the zero. Then, approximate $f(\cdot)$ around x_0 by means of a first order Taylor series: $f(x) \approx f(x_0) + (x - x_0)(df/dx)|_{x=x_0}$. Solve this for x and obtain: $x = x_0 - [(df/dx)|_{x=x_0}]^{-1}f(x_0)$. Let x_1 be the solution for x , use this as the new guess and repeat the above until convergence. When the function $f(\cdot)$ has multiple arguments, is vector-valued and denoted by \vec{f} , and the Taylor

approximation becomes: $\vec{f}(\mathbf{x}) \approx f(\mathbf{x}_0) + J\vec{f}|_{\mathbf{x}=\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0)$ with

$$J\vec{f} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_p} \\ \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_q}{\partial x_1} & \frac{\partial f_q}{\partial x_2} & \cdots & \frac{\partial f_q}{\partial x_p} \end{pmatrix},$$

the Jacobi matrix. An update of x_0 is now readily constructed by solving (the approximation for) $\vec{f}(\mathbf{x}) = \mathbf{0}$ for \mathbf{x} .

When applied here to the maximum likelihood estimation of the regression parameter β of the logistic regression model, the Newton-Raphson update is:

$$\hat{\beta}^{\text{new}} = \hat{\beta}^{\text{old}} - \left(\frac{\partial^2 \mathcal{L}}{\partial \beta \partial \beta^\top} \right)^{-1} \Big|_{\beta=\hat{\beta}^{\text{old}}} \frac{\partial \mathcal{L}}{\partial \beta} \Big|_{\beta=\hat{\beta}^{\text{old}}}$$

where the Hessian of the log-likelihood equals:

$$\frac{\partial^2 \mathcal{L}}{\partial \beta \partial \beta^\top} = - \sum_{i=1}^n \frac{\exp(\mathbf{X}_{i,*}\beta)}{[1 + \exp(\mathbf{X}_{i,*}\beta)]^2} \mathbf{X}_{i,*}^\top \mathbf{X}_{i,*}.$$

Iterative application of this updating formula converges to the ML estimate of β .

The Newton-Raphson algorithm is often reformulated as an iteratively re-weighted least squares (IRLS) algorithm. Hereto, first write the gradient and Hessian in matrix notation:

$$\frac{\partial \mathcal{L}}{\partial \beta} = \mathbf{X}^\top [\mathbf{Y} - \vec{g}^{-1}(\mathbf{X}; \beta)] \quad \text{and} \quad \frac{\partial^2 \mathcal{L}}{\partial \beta \partial \beta^\top} = -\mathbf{X}^\top \mathbf{W} \mathbf{X},$$

where $\vec{g}^{-1}(\mathbf{X}; \beta) = [g^{-1}(\mathbf{X}_{1,*}; \beta), \dots, g^{-1}(\mathbf{X}_{n,*}; \beta)]^\top$ with $g^{-1}(\cdot, \cdot) = \exp(\cdot, \cdot)/[1 + \exp(\cdot, \cdot)]$ and \mathbf{W} diagonal with $(\mathbf{W})_{ii} = \exp(\mathbf{X}_{i,*}\hat{\beta}^{\text{old}})[1 + \exp(\mathbf{X}_{i,*}\hat{\beta}^{\text{old}})]^{-2}$. The notation \mathbf{W} was already used in Chapter 3 and, generally, refers to a (diagonal) weight matrix with the choice of the weights depending on the context. The updating formula of the estimate then becomes:

$$\begin{aligned} \hat{\beta}^{\text{new}} &= \hat{\beta}^{\text{old}} + (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top [\mathbf{Y} - \vec{g}^{-1}(\mathbf{X}; \beta^{\text{old}})] \\ &= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \{\mathbf{X} \hat{\beta}^{\text{old}} + \mathbf{W}^{-1} [\mathbf{Y} - \vec{g}^{-1}(\mathbf{X}; \beta^{\text{old}})]\} \\ &= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{Z}, \end{aligned}$$

where $\mathbf{Z} = \mathbf{X} \hat{\beta}^{\text{old}} + \mathbf{W}^{-1} [\mathbf{Y} - \vec{g}^{-1}(\mathbf{X}; \beta^{\text{old}})]$. The Newton-Raphson update is thus the solution to the following weighted least squares problem:

$$\hat{\beta}^{\text{new}} = \arg \min_{\beta \in \mathbb{R}^p} (\mathbf{Z} - \mathbf{X} \beta)^\top \mathbf{W} (\mathbf{Z} - \mathbf{X} \beta).$$

Effectively, at each iteration the *adjusted response* \mathbf{Z} is regressed on the covariates that comprise \mathbf{X} . For more on logistic regression confer the monograph of Hosmer Jr *et al.* (2013).

5.2 Separable and high-dimensional data

The binary nature of the response may bring about another problem, called *separable data*, that frustrates the estimation of the logistic regression estimator. Separable data refer to the situation where the covariate space can be separated by a hyperplane such that the samples with indices in the set $\{i : Y_i = 0\}$ fall on one side of this hyperplane while those with an index in the set $\{i : Y_i = 1\}$ on the other. More formally, the data are separable if there exist $\mathbf{w} \in \mathbb{R}^p$ and $b \in \mathbb{R}$ such that

$$\begin{cases} Y_i = 0 & \text{if } \mathbf{X}_{i,*} \mathbf{w} + b > 0, \\ Y_i = 1 & \text{if } \mathbf{X}_{i,*} \mathbf{w} + b < 0, \end{cases}$$

for $i = 1, \dots, n$. Separable data mostly occur of either of the two response outcomes has a low prevalence.

The logistic regression parameter cannot be learned from separable data by means of the maximum likelihood method. For the existence of a separating hyperplane implies that the optimal fit is perfect and all samples have –

according to the fitted logistic regression model – a probability of one of being assigned to the correct outcome, i.e. $P(Y_i = 1 | \mathbf{X}_i)$ equals either zero or one. Consequently, the loglikelihood vanishes, cf.:

$$\mathcal{L}(\mathbf{Y} | \mathbf{X}; \boldsymbol{\beta}) = \sum_{i=1}^n Y_i \log[P(Y_i = 1 | \mathbf{X}_{i,*})] + (1 - Y_i) \log[P(Y_i = 0 | \mathbf{X}_{i,*})] = 0,$$

and does no longer involve the logistic regression parameter. The logistic regression parameter is then to be chosen such that $P(Y_i = 1 | \mathbf{X}_{i,*}) = \exp(\mathbf{X}_{i,*}\boldsymbol{\beta})[1 + \exp(\mathbf{X}_{i,*}\boldsymbol{\beta})]^{-1} \in \{0, 1\}$ (depending on whether Y_i is indeed of the ‘1’ class). This only occurs when (some) elements of $\boldsymbol{\beta}$ equal (minus) infinity. Hence, the maximum likelihood estimator is not well-defined.

The common workaround to learn the logistic regression parameter from separable data is a technique called Firth penalized estimation (Firth, 1993; Heinze and Schemper, 2002). It amounts to the maximization of the loglikelihood augmented with the so-called Firth penalty:

$$\mathcal{L}(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta}) + \frac{1}{2} \log\{\det[\mathcal{I}(\boldsymbol{\beta})]\},$$

where $\mathcal{I}(\boldsymbol{\beta}) = \mathbf{X}^\top \mathbf{W}(\boldsymbol{\beta}) \mathbf{X}$ is the Fisher information matrix. The Firth penalty corrects for the first order bias of the maximum likelihood estimator due to the unbalancedness in the class prevalences. The larger this unbalance, the more likely the data are separable. Alternatively, the Firth penalty can be motivated as a (very) weakly informative Jeffrey’s prior.

High-dimensionally, a separable hyperplane can always be found, unless there is at least one pair of samples with a common variate vector, i.e. $\mathbf{X}_{i,*} = \mathbf{X}_{i',*}$ with $i \neq i'$, and different responses $Y_i \neq Y_{i'}$. Moreover, the maximum likelihood estimator of the logistic regression parameter is not well-defined high-dimensionally. To see this, assume $p > n$ and an estimate $\hat{\boldsymbol{\beta}}$ to be available. Due to the high-dimensionality, the null space of \mathbf{X} is non-trivial. Hence, let $\boldsymbol{\gamma} \in \text{null}(\text{span}(\mathbf{X}))$. Then: $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\boldsymbol{\gamma} = \mathbf{X}(\hat{\boldsymbol{\beta}} + \boldsymbol{\gamma})$. As the null space is a $p - n$ -dimensional subspace, $\boldsymbol{\gamma}$ need not equal zero. Hence, an infinite number of estimators of the logistic regression parameter exists that yield the same loglikelihood.

Firth penalization may resolve the separable data issue low-dimensionally. It does not ensure the existence of the estimator high-dimensionally. This is illustrated by the next example.

Example 5.1

Let the covariates be mutually exclusive indicators. Then, for all $i = 1, \dots, n$, there is a $j \in \{1, \dots, p\}$ such that $\mathbf{X}_{i,*} = \mathbf{e}_j^\top$, where \mathbf{e}_j is the unit vector comprising all zero’s except for a one at the j -th position. The Firth penalty then is

$$\sum_{j=1}^p \sum_{i=1}^n \mathbb{1}_{\{\mathbf{X}_{i,*}=\mathbf{e}_j^\top\}} \left\{ \frac{1}{2} \beta_j - \log[1 + \exp(\beta_j)] \right\}.$$

High-dimensionally, this Firth penalty is not strictly concave as its Hessian matrix is diagonal with diagonal elements $\mathbf{H}_{jj} = -\exp(\beta_j)[1 + \exp(\beta_j)]^{-2} \sum_{i=1}^n \mathbb{1}_{\{\mathbf{X}_{i,*}=\mathbf{e}_j^\top\}}$ for $j = 1, \dots, p$. Hence, the maximizer of the Firth penalized loglikelihood is not well-defined high-dimensionally. \square

5.3 Ridge estimation

Ridge maximum likelihood estimates of the logistic model parameters are found by the maximization of the ridge penalized loglikelihood (cf. Schaefer *et al.* 1984; Le Cessie and Van Houwelingen 1992):

$$\mathcal{L}^{\text{pen}}(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta}, \lambda) = \mathcal{L}(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta}) - \frac{1}{2} \lambda \|\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^n \{Y_i \mathbf{X}_{i,*}\boldsymbol{\beta} - \log[1 + \exp(\mathbf{X}_{i,*}\boldsymbol{\beta})]\} - \frac{1}{2} \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta},$$

where the second summand is the ridge penalty (the sum of the square of the elements of $\boldsymbol{\beta}$) with λ the penalty parameter. Penalization of the loglikelihood now amounts to the subtraction of the ridge penalty. This is due to the fact that the estimator is now defined as the maximizer (instead of a minimizer) of the loss function. Moreover, the $\frac{1}{2}$ in front of the penalty is only there to simplify derivations later, and could in principle be absorbed in the penalty parameter. Finally, the augmentation of the loglikelihood with the ridge penalty ensures the existence of a unique estimator. The loglikelihood need not be strictly concave, but the ridge penalty, $-\frac{1}{2} \|\boldsymbol{\beta}\|_2^2$, is. Together they form the ridge penalized loglikelihood above, which is thus also strictly concave. This warrants the existence of a unique maximizer, and a well-defined ridge logistic regression estimator.

The optimization of the ridge penalized loglikelihood associated with the logistic regression model proceeds, due to the differentiability of the penalty, fully analogous to the unpenalized case and uses the Newton-Raphson

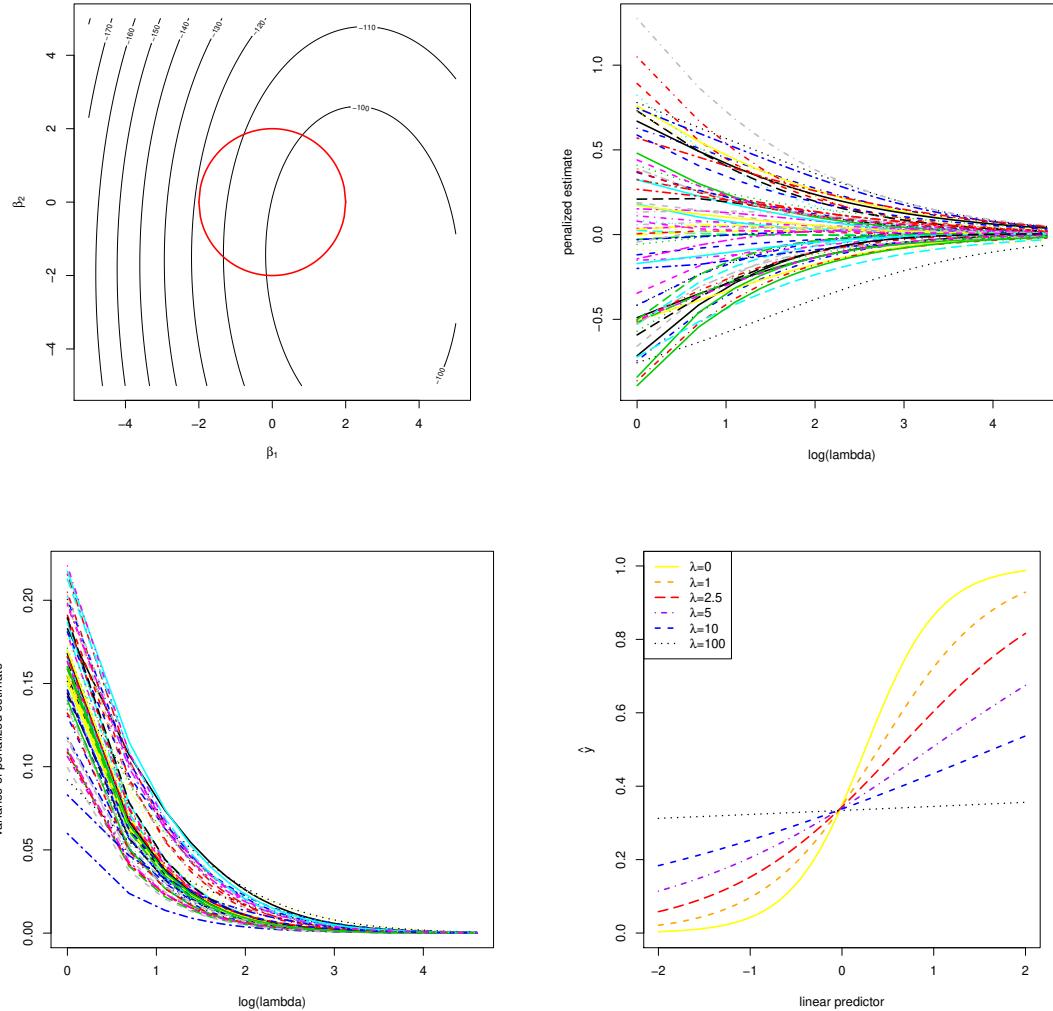


Figure 5.2: Top row, left panel: contour plot of the penalized log-likelihood of a logistic regression model with the ridge constraint (red line). Top row, right panel: the regularization paths of the ridge estimator of the logistic regression parameter. Bottom row, left panel: variance of the ridge estimator of the logistic regression parameter against the logarithm of the penalty parameter. Bottom panel, right panel: the predicted success probability versus the linear predictor for various choices of the penalty parameter.

algorithm for solving the (penalized) estimating equation. Hence, the unpenalized ML estimation procedure is modified straightforwardly by replacing gradient and Hessian by their ‘penalized’ counterparts:

$$\frac{\partial \mathcal{L}^{\text{pen}}}{\partial \beta} = \frac{\partial \mathcal{L}}{\partial \beta} - \lambda \beta \quad \text{and} \quad \frac{\partial^2 \mathcal{L}^{\text{pen}}}{\partial \beta \partial \beta^\top} = \frac{\partial^2 \mathcal{L}}{\partial \beta \partial \beta^\top} - \lambda \mathbf{I}_{pp}.$$

With these at hand, the Newton-Raphson algorithm is (again) reformulated as an iteratively re-weighted least squares algorithm with the updating step changing accordingly to:

$$\begin{aligned} \hat{\beta}^{\text{new}}(\lambda) &= \hat{\beta}^{\text{old}}(\lambda) + \mathbf{V}^{-1} (\mathbf{X}^\top \{ \mathbf{Y} - \bar{\mathbf{g}}^{-1}[\mathbf{X}; \beta^{\text{old}}(\lambda)] \} - \lambda \beta^{\text{old}}(\lambda)) \\ &= \mathbf{V}^{-1} \mathbf{V} \hat{\beta}^{\text{old}} - \lambda \mathbf{V}^{-1} \hat{\beta}^{\text{old}}(\lambda) + \mathbf{V}^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{W}^{-1} \{ \mathbf{Y} - \bar{\mathbf{g}}^{-1}[\mathbf{X}; \beta^{\text{old}}(\lambda)] \} \\ &= \mathbf{V}^{-1} \mathbf{X}^\top \mathbf{W} (\mathbf{X} \hat{\beta}^{\text{old}}(\lambda) + \mathbf{W}^{-1} \{ \mathbf{Y} - \bar{\mathbf{g}}^{-1}[\mathbf{X}; \beta^{\text{old}}(\lambda)] \}) \\ &= (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{Z}, \end{aligned}$$

where $\mathbf{V} = \mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_{pp}$ and \mathbf{W} and \mathbf{Z} as before. Hence, use this to update the estimate of β until convergence, which yields the desired ridge ML estimate.

Obviously, the ridge estimate of the logistic regression parameter tends to zero as $\lambda \rightarrow \infty$. Now consider a linear predictor with an intercept that is left unpenalized. When λ tends to infinity, all regression coefficients but the intercept vanish. The intercept is left to model the success probability. Hence, in this case $\lim_{\lambda \rightarrow \infty} \hat{\beta}_0(\lambda) = \log[\frac{1}{n} \sum_{i=1}^n Y_i / \frac{1}{n} \sum_{i=1}^n (1 - Y_i)]$.

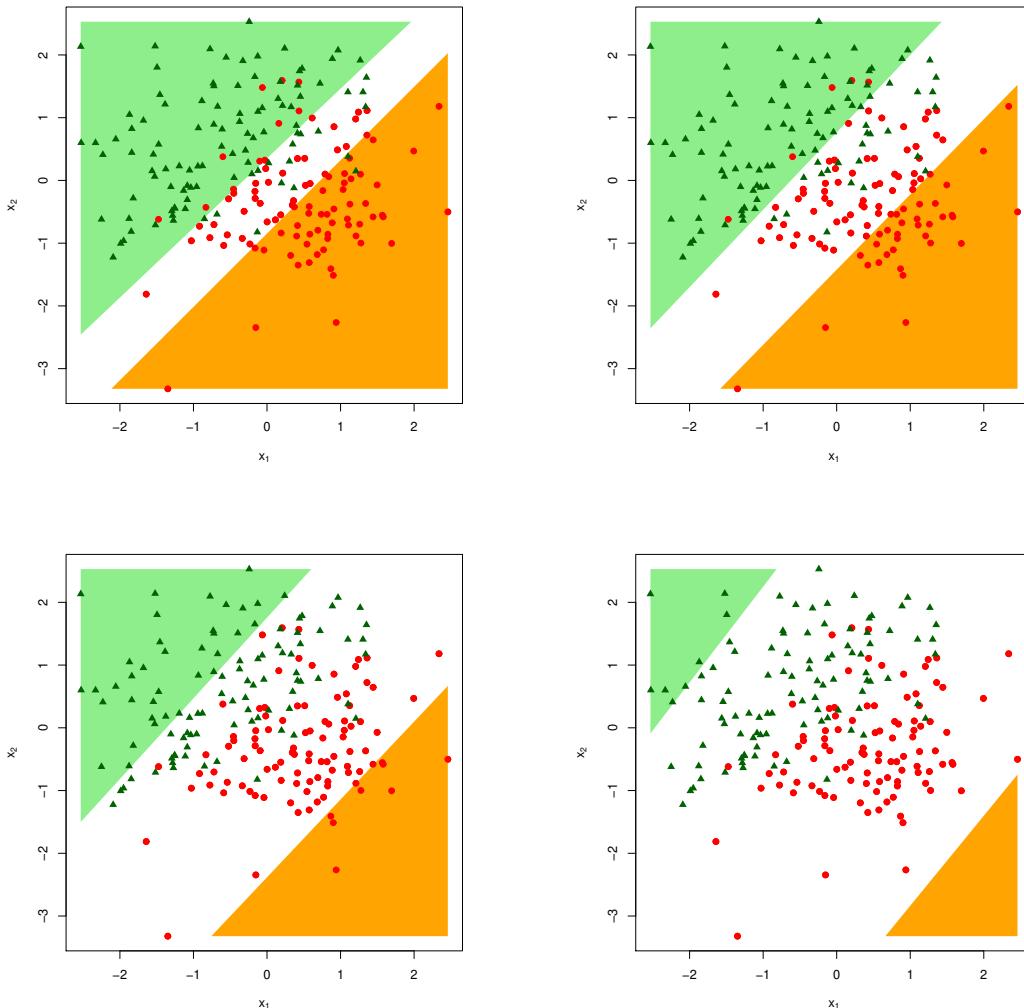


Figure 5.3: The realized design as scatter plot (X_1 vs X_2) overlayed by the success (red) and failure regions (green) for various choices of the penalty parameter: $\lambda = 0$ (top row, left panel), $\lambda = 10$ (top row, right panel), $\lambda = 40$ (bottom row, left panel), $\lambda = 100$ (bottom row, right panel).

The effect of the ridge penalty on parameter estimates propagates to the predictor \hat{p}_i . The linear predictor of the logistic regression model involving the ridge estimator $\mathbf{X}_i \hat{\beta}(\lambda)$ shrinks towards a common value for each i , leading to a scale difference between observation and predictor (as seen before in Section 1.10). This behaviour transfers to the ridge logistic regression predictor, as is illustrated on simulated data. The dimension and sample size of these data are $p = 2$ and $n = 200$, respectively. The covariate data are drawn from the standard normal, while that of the response is sampled from a Bernoulli distribution with success probability $P(Y_i = 1) = \exp(2X_{i,1} - 2X_{i,2})/[1 + \exp(2X_{i,1} - 2X_{i,2})]$. The logistic regression model is estimated from these data by means of ridge penalized likelihood maximization with various choices of the penalty parameter. The bottom right plot in Figure 5.2 shows the predicted success probability versus the linear predictor for various choices of the penalty parameter. Larger values of the penalty parameter λ flatten the slope of this curve. Consequently, for larger λ more excessive values of the covariates are needed to achieve the same predicted success probability as those obtained with smaller λ at more moderate covariate values. The implications for the resulting classification may become clearer when studying the effect of the penalty parameter on the ‘failure’ and ‘success regions’ respectively defined by:

$$\begin{aligned} & \{(x_1, x_2) : P(\textcolor{blue}{Y} = 0 | X_1 = x_1, X_2 = x_2, \hat{\beta}(\lambda)) > 0.75\}, \\ & \{(x_1, x_2) : P(\textcolor{red}{Y} = 1 | X_1 = x_1, X_2 = x_2, \hat{\beta}(\lambda)) > 0.75\}. \end{aligned}$$

This separates the design space in a light red ('failure') and light green ('success') domain. The white bar between them is the domain where samples cannot be classified with high enough certainty. As λ grows, so does the white area that separates the failure and success regions. Hence, as stronger penalization shrinks the logistic regression parameter estimate towards zero, it produces a predictor that is less outspoken in its class assignments.

Remark 5.1

Historically, confer Schaefer *et al.* (1984), the ridge logistic regression estimator is defined, analogously to the ridge regression estimator, as an ad-hoc fix to the collinearity among the covariates. The collinearity now causes $\mathbf{X}^\top \mathbf{W} \mathbf{X}$ to be ill-conditionedness, which is circumvented by the addition of the term $\lambda \mathbf{I}_{pp}$. This results in the following alternative definition of the ridge logistic regression estimator:

$$\hat{\beta}_a(\lambda) = (\mathbf{X}^\top \mathbf{W}_{ml} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{W}_{ml} \mathbf{X} \hat{\beta}^{ml},$$

where \mathbf{W}_{ml} is defined as in the Iteratively Weighted Least Squares algorithm but with $\hat{\beta}^{ml}$ substituted for the previous update $\hat{\beta}^{old}(\lambda)$. This alternative definition assumes the availability of the maximum likelihood estimator and is thus not applicable to separable, and in particular high dimensional, data.

An alternative motivation of this approximate ridge logistic regression estimator follows from the following asymptotic argument. Hereto we assume that for large enough n the maximum likelihood estimator of the logistic regression parameter $\hat{\beta}^{ml}$ exists. We can then develop a second order Taylor approximation of the penalized loglikelihood's summand $\log[1 + \exp(\mathbf{X}\beta)]$ in β around the point $\beta = \hat{\beta}^{ml}$:

$$\log[1 + \exp(\mathbf{X}\beta)] \approx \log[1 + \exp(\mathbf{X}\hat{\beta}^{ml})] + [\vec{g}(\mathbf{X}; \hat{\beta}^{ml})]^\top \mathbf{X}(\beta - \hat{\beta}^{ml}) + \frac{1}{2}(\beta - \hat{\beta}^{ml})^\top \mathbf{X}^\top \mathbf{W}_{ml} \mathbf{X}(\beta - \hat{\beta}^{ml}),$$

where \mathbf{W}_{ml} defined as above. We substitute this Taylor approximation in the penalized loglikelihood to obtain an approximation of the latter:

$$\mathcal{L}^{\text{pen}}(\mathbf{Y}, \mathbf{X}; \beta, \lambda) \approx [\mathbf{Y} - \vec{g}(\mathbf{X}; \hat{\beta}^{ml})]^\top \mathbf{X} \beta - \frac{1}{2} \beta^\top (\mathbf{X}^\top \mathbf{W}_{ml} \mathbf{X} + \lambda \mathbf{I}_{pp}) \beta + \beta^\top \mathbf{X}^\top \mathbf{W}_{ml} \mathbf{X} \hat{\beta}^{ml} + a_n,$$

where a_n contains all remaining terms not involving β . The right-hand side is maximized at

$$(\mathbf{X}^\top \mathbf{W}_{ml} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{W}_{ml} \{ \mathbf{X} \hat{\beta}^{ml} + \mathbf{W}_{ml}^{-1} [\mathbf{Y} - \vec{g}(\mathbf{X}; \hat{\beta}^{ml})] \},$$

which equals $\hat{\beta}_a(\lambda)$.

5.4 Moments

The 1st and 2nd order moments of the ridge maximum likelihood estimator of the logistic regression parameter is unknown analytically but typically approximated by that of the final update of the Newton-Raphson algorithm. This approximation assumes the one-to-last update $\hat{\beta}^{old}$ to be non-random and then proceeds as before with the regular ridge regression estimator of the linear regression model parameter to arrive at:

$$\begin{aligned} \mathbb{E}[\hat{\beta}^{\text{new}}(\lambda)] &= (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{W} \mathbb{E}(\mathbf{Z}), \\ \text{Var}[\hat{\beta}^{\text{new}}(\lambda)] &= (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{W} [\text{Var}(\mathbf{Z})] \mathbf{W} \mathbf{X} (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1}, \end{aligned}$$

with

$$\begin{aligned} \mathbb{E}(\mathbf{Z}) &= \{ \mathbf{X} \hat{\beta}^{old} + \mathbf{W}^{-1} [\mathbb{E}(\mathbf{Y}) - \vec{g}^{-1}(\mathbf{X}; \beta^{old})] \}, \\ \text{Var}(\mathbf{Z}) &= \mathbf{W}^{-1} \text{Var}(\mathbf{Y}) \mathbf{W}^{-1} = \mathbf{W}^{-1}, \end{aligned}$$

where the identity $\text{Var}(\mathbf{Y}) = \mathbf{W}$ follows from the variance of a Binomial distributed random variable. From these expressions similar properties as for the ridge maximum likelihood estimator of the regression parameter of the linear model may be deduced. For instance, the ridge maximum likelihood estimator of the logistic regression parameter converges to zero as the penalty parameter tends to infinity (confer the top right panel of Figure 5.2). Similarly, their variances vanish as $\lambda \rightarrow \infty$ (illustrated in the bottom left panel of Figure 5.2).

The form of these moment approximation corroborates asymptotically with that of the approximated ridge logistic regression estimator $\hat{\beta}(\lambda)$ introduced in Remark 5.1.

5.5 Constrained estimation

The maximization of the penalized loglikelihood of the logistic regression model can be reformulated as a constrained estimation problem, as was done in Section 1.5 for the linear regression model. The ridge logistic regression estimator is thus defined equivalently as:

$$\hat{\beta}(\lambda) = \arg \max_{\{\beta \in \mathbb{R}^p : \|\beta\|_2^2 \leq c(\lambda)\}} \mathcal{L}(\mathbf{Y}, \mathbf{X}; \beta).$$

This is illustrated by the top left panel of Figure 5.2 for the ‘ $p = 2$ ’-case. It depicts the contours (black lines) of the loglikelihood and the spherical domain of the parameter constraint $\{\beta \in \mathbb{R}^2 : \beta_1^2 + \beta_2^2 \leq c(\lambda)\}$ (red line). The parameter constraint can again be interpreted as a means to harness against overfitting, as it prevents the ridge logistic regression estimator from assuming very large values, thereby avoiding a perfect description of the data.

The parameter constraint of the logistic regression estimator exhibits behavior in λ as that of the regular ridge estimator. Hereto we note that the squared radius of the parameter constraint is – by identical argumentation as provided in Section 1.5 – equal to $c(\lambda) = \|\hat{\beta}(\lambda)\|_2^2$. However, no explicit expression for this squared radius exists, as none exists for the logistic ridge regression estimator, to verify its the constraint shrinkage behavior. Proposition 5.1 characterizes the essential properties of this behavior.

Proposition 5.1

The squared norm of the ridge logistic regression estimator satisfies:

- i) $d\|\hat{\beta}(\lambda)\|_2^2/d\lambda < 0$ for $\lambda > 0$,
- ii) $\lim_{\lambda \rightarrow \infty} \|\hat{\beta}(\lambda)\|_2^2 = 0$.

Proof. For part i) take the derivative of the estimating equation with respect to λ , solve for $\frac{d}{d\lambda}\hat{\beta}(\lambda)$, and find that

$$\frac{d}{d\lambda}\hat{\beta}(\lambda) = -(\mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \hat{\beta}(\lambda).$$

Using this derivative and the chain rule, we obtain the derivative of the estimator’s (squared) Euclidean length:

$$\frac{d}{d\lambda}\|\hat{\beta}(\lambda)\|_2^2 = \frac{d}{d\lambda}\{[\hat{\beta}(\lambda)]^\top \hat{\beta}(\lambda)\} = 2[\hat{\beta}(\lambda)]^\top \frac{d}{d\lambda}\hat{\beta}(\lambda) = -2[\hat{\beta}(\lambda)]^\top (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \hat{\beta}(\lambda).$$

As the right-hand side is a quadratic form multiplied by a negative scalar, we conclude that $\frac{d}{d\lambda}\|\hat{\beta}(\lambda)\|_2^2 < 0$ for all $\lambda > 0$.

To verify the claimed λ -limit of the squared length of the ridge logistic regression estimator, note that it satisfies: $\hat{\beta}(\lambda) = \lambda^{-1} \mathbf{X}^\top \{\mathbf{Y} - \vec{\mathbf{g}}^{-1}[\mathbf{X}; \hat{\beta}(\lambda)]\}$, which can be derived from the estimating equation. But as all elements of the vector $\vec{\mathbf{g}}^{-1}[\mathbf{X}; \hat{\beta}(\lambda)]$ are in the interval $(0, 1)$ and $Y_i \in \{0, 1\}$, we obtain the inequality:

$$\|\hat{\beta}(\lambda)\|_2^2 = \lambda^{-2} \{\mathbf{Y} - \vec{\mathbf{g}}^{-1}[\mathbf{X}; \hat{\beta}(\lambda)]\}^\top \mathbf{X} \mathbf{X}^\top \{\mathbf{Y} - \vec{\mathbf{g}}^{-1}[\mathbf{X}; \hat{\beta}(\lambda)]\} \leq \lambda^{-2} \mathbf{1}_n^\top \mathbf{X} \mathbf{X}^\top \mathbf{1}_n.$$

This bound can be sharpened by using the worst possible fit, where either $\lim_{\lambda \rightarrow \infty} \vec{\mathbf{g}}^{-1}[\mathbf{X}; \hat{\beta}(\lambda)] = \frac{1}{2}$ or $\lim_{\lambda \rightarrow \infty} \vec{\mathbf{g}}^{-1}[\mathbf{X}; \hat{\beta}(\lambda)] = \frac{1}{n} \sum_{i=1}^n Y_i$, depending on the presence of an unpenalized intercept in the model. Irrespectively, the derived bound vanishes as $\lambda \rightarrow \infty$. ■

Part i) of Proposition 5.1 tells us that the shrinkage behaviour of the ridge logistic regression estimator is monotone in λ . Monotone in the sense that the squared Euclidean length of the estimator is, as $\|\hat{\beta}(\lambda)\|_2^2 \geq 0$ for all $\lambda > 0$, is strictly decreasing in λ . Furthermore, part ii) of Proposition 5.1 then ensures that ultimately the constraint collapses onto zero.

5.6 Degrees of freedom

The degrees of freedom consumed by the ridge logistic regression estimator can be derived from the definition (1.11) previously introduced in Section 1.6. It uses $\hat{\beta}(\lambda_1)$ as obtained from the last iteration of the IRLS algorithm, and assumes the weight matrix employed in this last iteration to be nonrandom. Then, but see also Park and Hastie

(2008),

$$\begin{aligned}
df(\lambda) &= \sum_{i=1}^n [\text{Var}(Y_i)]^{-1} \text{Cov}(\hat{Y}_i, Y_i) \\
&= \sum_{i=1}^n \{\exp(\mathbf{X}_{i,*}\beta)[1 + \exp(\mathbf{X}_{i,*}\beta)]^{-2}\}^{-1} \text{Cov}[\mathbf{X}_{i,*}\hat{\beta}(\lambda_1), Y_i] \\
&= \sum_{i=1}^n [(\mathbf{W})_{ii}]^{-1} \text{Cov}(\mathbf{X}_{i,*}(\mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{Z}, Y_i) \\
&= \sum_{i=1}^n [(\mathbf{W})_{ii}]^{-1} \mathbf{X}_{i,*} (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{W} \\
&\quad \times \text{Cov}\{\mathbf{X}\hat{\beta}^{\text{old}} + \mathbf{W}^{-1}[\mathbf{Y} - \vec{g}^{-1}(\mathbf{X}; \beta^{\text{old}})], Y_i\} \\
&= \sum_{i=1}^n [(\mathbf{W})_{ii}]^{-1} \mathbf{X}_{i,*} (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{W} \text{Cov}[\mathbf{W}^{-1}\mathbf{Y}, Y_i] \\
&= \text{tr}[\mathbf{X}(\mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^\top],
\end{aligned}$$

where we have used the independence among the individual observations. The degrees of freedom of the ridge logistic regression estimator too decrease monotonically to zero as λ increases.

5.7 The Bayesian connection

All penalized estimators can be formulated as Bayesian estimators, including the ridge logistic estimator. In particular, ridge estimators correspond to Bayesian estimators with a multivariate normal prior on the regression coefficients. Thus, assume $\beta \sim \mathcal{N}(\mathbf{0}_p, \Delta^{-1})$. The posterior distribution of β then is:

$$f_\beta(\beta | \mathbf{Y}, \mathbf{X}) \propto \left\{ \prod_{i=1}^n [P(Y_i = 1 | \mathbf{X}_{i,*})]^{Y_i} [P(Y_i = 0 | \mathbf{X}_{i,*})]^{1-Y_i} \right\} \exp(-\frac{1}{2}\beta^\top \Delta \beta).$$

This does not coincide with any standard distribution. But, under appropriate conditions, the posterior distribution is asymptotically normal. This invites a (multivariate) normal approximation to the posterior distribution above. The Laplace's method provides (cf. Bishop, 2006).

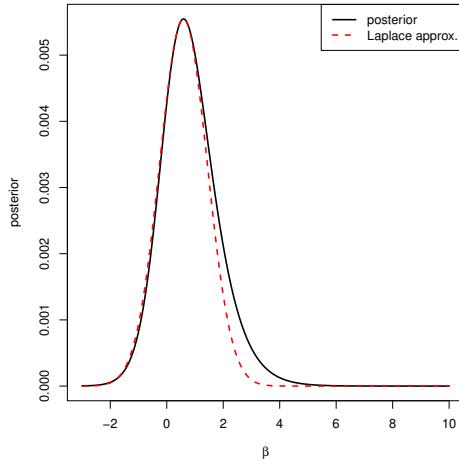


Figure 5.4: Laplace approximation to the posterior density of the Bayesian logistic regression parameter.

Laplace's method *i*) centers the normal approximation at the mode of the posterior, and *ii*) chooses the covariance to match the curvature of the posterior at the mode. The posterior mode is the location of the maximum of the posterior distribution. The location of this maximum coincides with that of the logarithm of the posterior. The latter is the log-likelihood augmented with a ridge penalty. Hence, the posterior mode, which is taken as the mean of the approximating Gaussian, coincides with the ridge logistic regression estimator. For the covariance of the approximating Gaussian, the logarithm of the posterior is approximated by a second order Taylor series around

the posterior mode and limited to second order terms:

$$\begin{aligned} \log[f_{\beta}(\beta | \mathbf{Y}, \mathbf{X})] &\propto \log[f_{\beta}(\beta | \mathbf{Y}, \mathbf{X})] \Big|_{\beta=\hat{\beta}_{\text{MAP}}} \\ &+ \frac{1}{2}(\beta - \hat{\beta}_{\text{MAP}})^{\top} \frac{\partial^2}{\partial \beta \partial \beta^{\top}} \log[f_{\beta}(\beta | \mathbf{Y}, \mathbf{X})] \Big|_{\beta=\hat{\beta}_{\text{MAP}}} (\beta - \hat{\beta}_{\text{MAP}})^{\top}, \end{aligned}$$

in which the first order term cancels as the derivative of $f_{\beta}(\beta | \mathbf{Y}, \mathbf{X})$ with respect to β vanishes at the posterior mode – its maximum. Take the exponential of this approximation and match its arguments to that of a multivariate Gaussian $\exp[-\frac{1}{2}(\beta - \mu_{\beta})^{\top} \Sigma_{\beta}^{-1}(\beta - \mu_{\beta})]$. The covariance of the sought Gaussian approximation is thus the inverse of the Hessian of the negative penalized log-likelihood. Put together the posterior is approximated by:

$$\beta | \mathbf{Y}, \mathbf{X} \sim \mathcal{N}[\hat{\beta}_{\text{MAP}}, (\Delta + \mathbf{X}^{\top} \mathbf{W} \mathbf{X})^{-1}].$$

The Gaussian approximation is convenient but need not be good. Fortunately, the Bernstein-Von Mises Theorem (van der Vaart, 2000) tells us that it is very accurate when the model is regular, the prior smooth, and the sample size sufficiently large. The quality of the approximation for an artificial example data set is shown in Figure 5.4.

5.8 Computationally efficient evaluation

High-dimensionally, the IRLS algorithm of the ridge logistic regression estimator can be evaluated computationally efficiently as follows. Hereto, effectively, the problem of penalized likelihood maximization over all $\beta \in \mathbb{R}^p$ is converted into maximization of the same criterion but now over the weights $\text{diag}(\mathbf{W}) \in \mathbb{R}^n$, as those weights are all that are needed for the evaluation of the estimator. The key observations, that enable this conversion, are the following reformulations of the updated linear predictor and the penalty (see Question 5.5):

$$\mathbf{X}\hat{\beta}(\lambda) = \lambda^{-1}\mathbf{X}\mathbf{X}^{\top}(\lambda^{-1}\mathbf{X}\mathbf{X}^{\top} + \mathbf{W}^{-1})^{-1}\mathbf{Z}, \quad (5.2)$$

$$\lambda\|\hat{\beta}(\lambda)\|_2^2 = \mathbf{Z}^{\top}(\lambda^{-1}\mathbf{X}\mathbf{X}^{\top} + \mathbf{W}^{-1})^{-1}(\lambda^{-1}\mathbf{X}\mathbf{X}^{\top})(\lambda^{-1}\mathbf{X}\mathbf{X}^{\top} + \mathbf{W}^{-1})^{-1}\mathbf{Z} \quad (5.3)$$

Both reformulations hinge upon the Woodbury matrix identity. On one hand, these reformulations avoid the inversion of a $p \times p$ dimensional matrix (as been shown for the regular ridge regression estimator, see Section 1.7). On the other, only the term $\lambda^{-1}\mathbf{X}\mathbf{X}^{\top}$ in the above expressions involves a matrix multiplication over the p dimension. Exactly this term is not updated at each iteration of the IRLS algorithm. It can thus be evaluated and stored prior to the first iteration step, and at each iteration be called upon. Furthermore, the remaining quantities updated at each iteration are the weights and the penalized loglikelihood, but those are obtained straightforwardly from the linear predictor and the penalty. The pseudo-code of an efficient version of the IRLS algorithm of the ridge logistic estimator, as has been implemented in the `ridgeGLM`-function of the `porridge`-package (van Wieringen and Aflakparast, 2021), thus comprises the following steps:

- 1) Evaluate and store $\lambda^{-1}\mathbf{X}\mathbf{X}^{\top}$.
- 2) Initiate the linear predictor.
- 3) Update the weights, the adjusted response, the penalized loglikelihood, and the linear predictor.
- 4) Repeat the previous step until convergence.
- 5) Evaluate the estimator using the weights from the last iteration.

In the above, convergence refers to no further or a negligible improvement of the penalized loglikelihood. In the final step of this version of the IRLS algorithm, the ridge logistic regression estimator is obtained by:

$$\hat{\beta}(\lambda) = \lambda^{-1}\mathbf{X}^{\top}(\lambda^{-1}\mathbf{X}\mathbf{X}^{\top} + \mathbf{W}^{-1})^{-1}\mathbf{Z}.$$

In the above display the weights \mathbf{W} and adjusted response \mathbf{Z} are obtained from the last iteration of the IRLS algorithm's third step. Alternatively, computational efficiency is obtained by the ‘SVD-trick’ (see Question 5.6).

5.9 Penalty parameter selection

As before the penalty parameter may be chosen through K -fold cross-validation. For the $K = n$ case, Meijer and Goeman (2013) describe a computationally efficient approximation of the leave-one-out cross-validated loglikelihood. It is based on the exact evaluation of the LOOCV loss, discussed in Section 1.8.2, that avoided resampling.

The approach of Meijer and Goeman (2013) hinges upon the first-order Taylor expansion of the left-out penalized loglikelihood of the left-out estimate $\hat{\beta}_{-i}(\lambda)$ around $\hat{\beta}(\lambda)$, which yields an approximation of the former:

$$\begin{aligned}\hat{\beta}_{-i}(\lambda) &\approx \hat{\beta}(\lambda) - \left(\frac{\partial^2 \mathcal{L}_{-i}^{\text{pen}}}{\partial \beta \partial \beta^\top} \Big|_{\beta=\hat{\beta}(\lambda)} \right)^{-1} \frac{\partial \mathcal{L}_{-i}^{\text{pen}}}{\partial \beta} \Big|_{\beta=\hat{\beta}(\lambda)} \\ &= \hat{\beta}(\lambda) + (\mathbf{X}_{-i,*}^\top \mathbf{W}_{-i,-i} \mathbf{X}_{-i,*} + \lambda \mathbf{I}_{pp})^{-1} \{ \mathbf{X}_{-i,*}^\top [\mathbf{Y}_{-i} - \vec{\mathbf{g}}^{-1}(\mathbf{X}_{-i,*}; \hat{\beta}(\lambda))] - \lambda \hat{\beta}(\lambda) \}.\end{aligned}$$

This approximation involves the inverse of a $p \times p$ dimensional matrix, which amounts to the evaluation of n such inverses for the LOOCV loss. As in Section 1.8.2 this may be avoided. Rewrite both the gradient and the Hessian of the left-out loglikelihood in the approximation of the preceding display:

$$\begin{aligned}& \mathbf{X}_{-i,*}^\top \{ \mathbf{Y}_{-i} - \vec{\mathbf{g}}^{-1}(\mathbf{X}_{-i,*}; \hat{\beta}(\lambda)) \} - \lambda \hat{\beta}(\lambda) \\ &= \mathbf{X}^\top \{ \mathbf{Y} - \vec{\mathbf{g}}^{-1}(\mathbf{X}; \hat{\beta}(\lambda)) \} - \lambda \hat{\beta}(\lambda) - \mathbf{X}_{i,*}^\top \{ Y_i - g^{-1}(\mathbf{X}_{i,*}; \hat{\beta}(\lambda)) \} \\ &= -\mathbf{X}_{i,*}^\top \{ Y_i - g^{-1}(\mathbf{X}_{i,*}; \hat{\beta}(\lambda)) \}\end{aligned}$$

and

$$(\mathbf{X}_{-i,*}^\top \mathbf{W}_{-i,-i} \mathbf{X}_{-i,*} + \lambda \mathbf{I}_{pp})^{-1} = (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} + \mathbf{W}_{ii} (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}_{i,*}^\top [1 - \mathbf{H}_{ii}(\lambda)]^{-1} \mathbf{X}_{i,*} (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1},$$

where the Woodbury identity has been used and now $\mathbf{H}_{ii}(\lambda) = \mathbf{W}_{ii} \mathbf{X}_{i,*} (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}_{i,*}^\top$. Substitute both in the approximation of the left-out ridge logistic regression estimator and manipulate as in Section 1.8.2 to obtain:

$$\hat{\beta}_{-i}(\lambda) \approx \hat{\beta}(\lambda) - (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}_{i,*}^\top [1 - \mathbf{H}_{ii}(\lambda)]^{-1} [Y_i - g^{-1}(\mathbf{X}_{i,*}; \hat{\beta}(\lambda))].$$

Hence, the leave-one-out cross-validated loglikelihood $\sum_{i=1}^n \mathcal{L}[Y_i | \mathbf{X}_{i,*}, \hat{\beta}_{-i}(\lambda)]$ can now be evaluated by means of a single inverse of a $p \times p$ dimensional matrix and some matrix multiplications. And even this single inversion can be reduced to one of an $n \times n$ -dimensional matrix when rewritten by means of the Woodbury matrix identity. For the performance of this approximation in terms of accuracy and speed, see Meijer and Goeman (2013).

For general K -fold cross-validation, computational efficiency is achieved through the same trick as used in Section 1.8.2 and reported by (van de Wiel *et al.*, 2021). It exploits the fact that the logistic regression parameter only appears in combination with the design matrix, forming the linear predictor, in the loglikelihood. For cross-validation there is thus no need to evaluate the estimator itself. To elaborate on the particulars of the logistic regression case, let $\mathcal{G}_1, \dots, \mathcal{G}_K \subset \{1, \dots, n\}$ be the mutually exclusive and exhaustive K -fold sample index sets. The linear predictor can then be written as:

$$\mathbf{X}_{\mathcal{G}_k,*} \hat{\beta}_{-\mathcal{G}_k}(\lambda) = \lambda^{-1} \mathbf{X}_{\mathcal{G}_k,*} \mathbf{X}_{-\mathcal{G}_k,*}^\top (\lambda^{-1} \mathbf{X}_{-\mathcal{G}_k,*} \mathbf{X}_{-\mathcal{G}_k,*}^\top + \mathbf{W}^{-1})^{-1} \mathbf{Z},$$

which can be verified by means of the Woodbury matrix identity. Furthermore, all matrix products of submatrices of \mathbf{X} are themselves submatrices of \mathbf{XX}^\top . Further efficiency is gained if the latter is evaluated before the start of the cross-validation loop. It then only remains to subset \mathbf{XX}^\top inside this loop.

5.10 Generalizing ridge logistic regression

The ridge logistic regression estimator may be generalized as the ridge regression counterpart. That is, the loglikelihood may be augmented by a nonzero centered quadratic form penalty $(\beta - \beta_0)^\top \Delta (\beta - \beta_0)$, where $\beta_0 \in \mathbb{R}^p$ is the shrinkage target and Δ is the nonnegative definite penalty matrix. The latter should be positive definite in high-dimensional settings. Moreover, we may include covariates that we wish to leave unpenalized in order to estimate them as unbiasedly as possible. To accommodate that, we introduce \mathbf{U} , the $n \times q$ -dimensional design matrix of the unpenalized covariates, and $\gamma \in \mathbb{R}^q$, the associated regression parameter. The inclusion of the additional unpenalized covariates alters the distribution of the response variable. We now have:

$$P(Y_i = 1 | \mathbf{U}_{i,*}, \mathbf{X}_{i,*}) = \exp(\mathbf{U}_{i,*} \gamma + \mathbf{X}_{i,*} \beta) [1 + \exp(\mathbf{U}_{i,*} \gamma + \mathbf{X}_{i,*} \beta)]^{-1}.$$

The derivation of the generalized ridge penalized loglikelihood of the logistic regression model is then analogous to that of the regular ridge penalized case. Its maximizer is the sought estimator:

$$\hat{\gamma}(\cdot), \hat{\beta}(\cdot) = \arg \max_{\gamma \in \mathbb{R}^q, \beta \in \mathbb{R}^p} \sum_{i=1}^n Y_i (\mathbf{U}_{i,*} \gamma + \mathbf{X}_{i,*} \beta) - \log[1 + \exp(\mathbf{U}_{i,*} \gamma + \mathbf{X}_{i,*} \beta)] \\ - (\beta - \beta_0)^\top \Delta (\beta - \beta_0).$$

No analytic expression for this estimator is available. It is again evaluated numerically by means of the Iteratively Reweighted Least Squares (IRLS) algorithm that generates a sequence $\{\hat{\gamma}^{(t)}(\cdot), \hat{\beta}^{(t)}(\cdot)\}_{t=1}^\infty$ that converges to the generalized ridge logistic regression estimator. Given the t -th values of this sequence, the next are found by:

$$\hat{\gamma}^{(t+1)}(\cdot) = \{\mathbf{U}^\top (\mathbf{W}^{(t)})^{-1} [\mathbf{X} \Delta^{-1} \mathbf{X}^\top + (\mathbf{W}^{(t)})^{-1}]^{-1} (\mathbf{W}^{(t)})^{-1} \mathbf{U}\}^{-1} \\ \times \mathbf{U}^\top [\mathbf{X} \Delta^{-1} \mathbf{X}^\top + (\mathbf{W}^{(t)})^{-1}]^{-1} (\mathbf{Y}^{(\text{adj}, t)} - \mathbf{X} \beta_0), \\ \hat{\beta}^{(t+1)}(\cdot) = \beta_0 + (\mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{X} + \Delta)^{-1} \mathbf{X}^\top \mathbf{W}^{(t)} [\mathbf{Y}^{(\text{adj}, t)} - \mathbf{U} \hat{\gamma}^{(t+1)}(\cdot) - \mathbf{X} \beta_0].$$

These expression have been obtained by application of the analytic expression of the inverse of a 2×2 block matrix, the Woodbury matrix identity, and some linear algebraic manipulations (see Lettink *et al.*, 2023 for details). The IRLS algorithm is terminated after a finite number of iterations, either when the loss does no longer substantially improve or subsequent iterations show little difference in their evaluation of the parameter estimates.

5.11 Application

The ridge logistic regression is used here to explain the status (dead or alive) of ovarian cancer samples at the close of the study from gene expression data at baseline. Data stem from the TCGA study (Cancer Genome Atlas Network, 2011), which measured gene expression by means of sequencing technology. Available are 295 samples with both status and transcriptomic profiles. These profiles are composed of 19990 transcript reads. The sequencing data, being representative of the mRNA transcript count, is heavily skewed. Zwiener *et al.* (2014) show that a simple transformation of the data prior to model building generally yields a better model than tailor-made approaches. Motivated by this observation the data were – to accommodate the zero counts – asinh-transformed. The logistic regression model is then fitted in ridge penalized fashion, leaving the intercept unpenalized. The ridge penalty parameter is chosen through 10-fold cross-validation minimizing the cross-validated error. R-code, and that for the sequel of this example, is to be found below.

Listing 5.1 R code

```
# load libraries
library(glmnet)
library(TCGA2STAT)

# load data
OVdata <- getTCGA(disease="OV", data.type="RNASeq", type="RPKM", clinical=TRUE)
Y      <- as.numeric(OVdata[[3]][,2])
X      <- asinh(data.matrix(OVdata[[3]][,-c(1:3)]))

# start fit
# optimize penalty parameter
cv.fit   <- cv.glmnet(X, Y, alpha=0, family=c("binomial"),
                      nfolds=10, standardize=FALSE)
optL2    <- cv.fit$lambda.min

# estimate model
glmFit   <- glmnet(X, Y, alpha=0, family=c("binomial"),
                      lambda=optL2, standardize=FALSE)

# construct linear predictor and predicted probabilities
linPred <- as.numeric(glmFit$a0 + X %*% glmFit$beta)
predProb <- exp(linPred) / (1+exp(linPred))

# visualize fit
boxplot(linPred ~ Y, pch=20, border="lightblue", col="blue",
```

```

    ylab="linear_predictor", xlab="response",
    main="fit")

# evaluate predictive performance
# generate k-folds balanced w.r.t. status
fold      <- 10
folds1   <- rep(1:fold, ceiling(sum(Y)/fold)) [1:sum(Y)]
folds0   <- rep(1:fold, ceiling((length(Y)-length(folds1))
                                /fold)) [1:(length(Y)-length(folds1))]
shuffle1 <- sample(1:length(folds1), length(folds1))
shuffle0 <- sample(1:length(folds0), length(folds0))
folds1   <- split(shuffle1, as.factor(folds1))
folds0   <- split(shuffle0, as.factor(folds0))
folds    <- list()
for (f in 1:fold) {
  folds[[f]] <- c(which(Y==1) [folds1[[f]]], which(Y==0) [folds0[[f]]])
}
for (f in 1:fold) {
  print(sum(Y[folds[[f]]]))
}

# build model
pred2obsL2 <- matrix(nrow=0, ncol=4)
colnames(pred2obsL2) <- c("optLambda", "linPred", "predProb", "obs")
for (f in 1:length(folds)) {
  print(f)
  cv.fit      <- cv.glmnet(X[-folds[[f]],], Y[-folds[[f]]], alpha=0,
                           family=c("binomial"), nfolds=10, standardize=FALSE)
  optL2       <- cv.fit$lambda.min
  glmFit     <- glmnet(X[-folds[[f]],], Y[-folds[[f]]], alpha=0,
                        family=c("binomial"), lambda=optL2, standardize=FALSE)
  linPred    <- glmFit$a0 + X[folds[[f]],,drop=FALSE] %*% glmFit$beta
  predProb   <- exp(linPred) / (1+exp(linPred))
  pred2obsL2 <- rbind(pred2obsL2, cbind(optL2, linPred, predProb, Y[folds[[f]]]))
}

# visualize fit
boxplot(pred2obsL2[,3] ~ pred2obsL2[,4], pch=20, border="lightblue", col="blue",
        ylab="linear_predictor", xlab="response",
        main="prediction")

```

The fit of the resulting model is studied. Hereto the fitted linear predictor $\mathbf{X}\hat{\beta}(\lambda_{\text{opt}})$ is plotted against the status (Figure 5.5, left panel). The plot shows some overlap between the boxes, but also a clear separation. The latter suggests gene expression at baseline thus enables us to distinguish surviving from the to-be-diseased ovarian cancer patients. Ideally, a decision rule based on the linear predictor can be formulated to predict an individual's outcome.

The fit, however, is evaluated on the samples that have been used to build the model. This gives no insight on the model's predictive performance on novel samples. A replication of the study is generally costly and comparable data sets need not be at hand. A common workaround is to evaluate the predictive performance on the same data (Subramanian and Simon, 2010). This requires to put several samples aside for performance evaluation while the remainder is used for model building. The left-out sample may accidentally be chosen to yield an exaggerated (either dramatically poor or overly optimistic) performance. This is avoided through the repetition of this exercise, leaving (groups of) samples out one at the time. The left-out performance evaluations are then averaged and believed to be representative of the predictive performance of the model on novel samples. Note that, effectively, as the model building involves cross-validation and so does the performance evaluation, a double cross-validation loop is applied. This procedure is applied with a ten-fold split in both loops. Denote the outer folds by $f = 1, \dots, 10$. Then, \mathbf{X}_f and \mathbf{X}_{-f} represent the design matrix of the samples comprising fold f and that of the remaining samples, respectively. Define \mathbf{Y}_f and \mathbf{Y}_{-f} similarly. The linear prediction for the left-out fold f is then $\mathbf{X}_f\hat{\beta}_{-f}(\lambda_{\text{opt},f})$. For reference to the fit, this is compared to \mathbf{Y}_f visually by means of a boxplot as used above (see Figure 5.5, right panel). The boxes overlap almost perfectly. Hence, little to nothing remains of the predictive

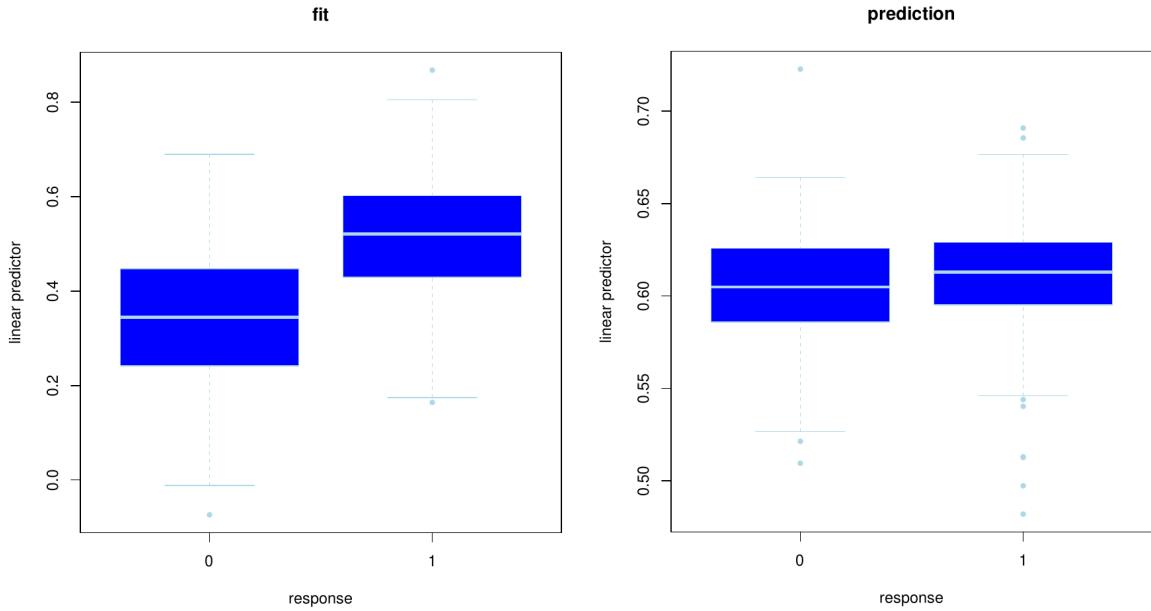


Figure 5.5: Left panel: Box plot of the status vs. the fitted linear predictor using the full data set. Right panel: Box plot of the status vs. the linear prediction in the left-out samples of the 10 folds.

power suggested by the boxplot of the fit. The fit may thus give a reasonable description of the data at hand, but it extrapolates poorly to new samples.

5.12 Beyond logistic regression

Logistic regression separates – by virtue of the linear predictor – the response classes by a hyperplane in the design space (see Figure 5.3). This may appear a substantial limitation of logistic regression as indeed there is no ground why the phenomenon under study would exhibit such neat linear behavior. This limitation can be circumvented by the expansion of our design matrix by functions of the covariates, e.g. $X_{i,1}, X_{i,1}X_{i,2}, X_{i,1}^2, \sin(X_{i,1}), \dots$. Such an expanded set of covariates facilitates nonlinear boundaries of the classification domains, as illustrated in Figure 5.6. This greatly expands the class of classifiers, with the potential of a substantial higher performance but with the risk of overfitting.

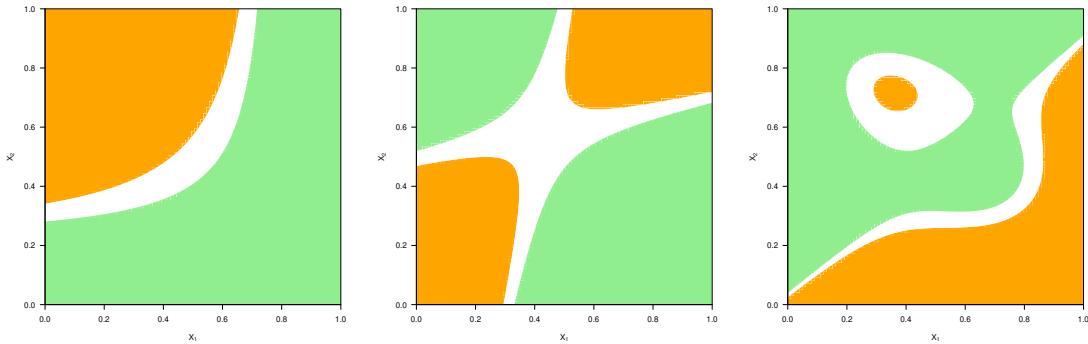


Figure 5.6: Classification domains of logistic regression with, from left to right, a design matrix expanded by the interaction between the covariates, and the covariates squared and cubed, respectively. The red and green domain represent regions where the classification is reasonably certain, while in the white domain it is not.

The key question in order to facilitate nonlinear classification boundaries is what functions to choose, or, put differently, what basis to form, from the original covariates that yield the best classifier. The easy (?) answer would be to let the data decide, e.g. by means of a feedforward neural network. A *feedforward neural network* is

a function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ of the form:

$$f : \mathbb{R}^p = \mathbb{R}^{L_0} \xrightarrow{(\alpha_1, \mathbf{W}_1)} \mathbb{R}^{L_1} \xrightarrow{(\alpha_2, \mathbf{W}_2)} \mathbb{R}^{L_2} \xrightarrow{(\alpha_3, \mathbf{W}_3)} \mathbb{R}^{L_3} \longrightarrow \dots \\ \dots \longrightarrow \mathbb{R}^{L_{m-1}} \xrightarrow{(\alpha_m, \mathbf{W}_m)} \mathbb{R}^{L_m} = \mathbb{R},$$

where:

- the $\{\mathbf{W}_u\}_{u=1}^m$ are parameter matrices that define linear combinations of the input which are propagated,
- the $\{\alpha\}_{u=1}^m$ are parameter vectors – called *bias terms* – that offset the linear combination of the input,
- the $f_u : \mathbb{R}^{L_{u-1}} \rightarrow \mathbb{R}^{L_u}$ are nonlinear *activation* functions that operate on the offsetted linear combination.

This compositional map of the feedforward neural network can be written as:

$$f(\mathbf{x}) = f_m(\dots f_3(\alpha_3 + \mathbf{W}_3 f_2(\alpha_2 + \mathbf{W}_2 f_1(\alpha_1 + \mathbf{W}_1 \mathbf{x})))),$$

with input vector $\mathbf{x} \in \mathbb{R}^p$. The feedforward neural network thus alternates between an affine linear combination and a nonlinear transformation. Logistic regression is a single layered forward neural network with $\alpha = \beta_0$, $\mathbf{W} = \boldsymbol{\beta}$, and $f(\alpha + \mathbf{W}\mathbf{x}) = [1 + \exp(-\alpha - \mathbf{W}\mathbf{x})]^{-1}$. A common alternative and nonlinear choice, due to its continuity and piece-wise linearity, of the $f_m(\cdot)$ is the so-called the ReLu (Rectified Linear Unit) function defined as $f(x) = \max\{0, x\}$. Refer to the monograph of Goodfellow *et al.* (2016) for more on forward neural networks.

The forward neural network effectively creates itself the basis functions from the covariates. It can approximate virtually any function if m and the $\{L_u\}_{u=1}^m$ are large (Schmidt-Hieber, 2019). There is of course no free lunch. The feedforward neural network's parameters may be non-identifiable. Irrespectively, the estimation of the network's many parameter requires strong regularization. Additionally, it requires massive computational power to learn a deep neural network. Finally, the utilization of a learned forward neural network may be hampered by its opaqueness: the input-output relationship is hard to interpret, a desirable property for their acceptance (Rudin, 2019).

5.13 Conclusion

To deal with response variables other than continuous ones, ridge logistic regression was discussed. High-dimensionally, the empirical identifiability problem then persists. Again, penalization came to the rescue: the ridge penalty may be combined with other link functions than the identity. Properties of ridge regression were shown to carry over to its logistic equivalent.

5.14 Exercises

Question 5.1

The variation in a binary response $Y_i \in \{0, 1\}$ due to two covariates $\mathbf{X}_{i,*} = (X_{i,1}, X_{i,2})$ is described by the logistic regression model: $P(Y_i = 1) = \exp(\mathbf{X}_{i,*}\boldsymbol{\beta})[1 + \exp(\mathbf{X}_{i,*}\boldsymbol{\beta})]$. The study design and the observed response are given by:

$$\mathbf{X} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

- Write down the loglikelihood and show that $\hat{\boldsymbol{\beta}} \in \{(\beta_1, \beta_2)^\top : \beta_1 - \beta_2 = \infty\}$.
- Augment the loglikelihood with the ridge penalty $\frac{1}{2}\lambda\|\boldsymbol{\beta}\|_2^2$ and show that $|\hat{\beta}_j(\lambda)| < \lambda$ for $j = 1, 2$.

Question 5.2

Consider an experiment involving n cancer samples. For each sample i the transcriptome of its tumor has been profiled and is denoted $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^\top$ where X_{ij} represents the gene $j = 1, \dots, p$ in sample i . Additionally, the overall survival data, (Y_i, c_i) for $i = 1, \dots, n$ of these samples is available. In this Y_i denotes the survival time of sample i and c_i the event indicator with $c_i = 0$ and $c_i = 1$ representing non- and censoredness, respectively. You may ignore the possibility of ties in the remainder.

- Write down the Cox proportional regression model that links overall survival times (as the response variable) to the expression levels.
- Specify its loss function for penalized maximum partial (!) likelihood estimation of the parameters. Penalization is via the regular ridge penalty.

- c) From this loss function, derive the estimating equation for the Cox regression coefficients.
d) Describe (in words) how you would evaluate the ridge ML estimator numerically.

Question 5.3

Load the leukemia data available via the `multtest`-package (downloadable from BioConductor) through the following R-code:

Listing 5.2 R code

```
# activate library and load the data
library(multtest)
data(golub)
```

The objects `golub` and `golub.cl` are now available. The matrix-object `golub` contains the expression profiles of 38 leukemia patients. Each profile comprises expression levels of 3051 genes. The numeric-object `golub.cl` is an indicator variable for the leukemia type (AML or ALL) of the patient.

- a) Relate the leukemia subtype and the gene expression levels by a logistic regression model. Fit this model by means of penalized maximum likelihood, employing the ridge penalty with penalty parameter $\lambda = 1$. This is implemented in the `penalized`-package available from CRAN. *Note:* center (gene-wise) the expression levels around zero.
- b) Obtain the fits from the regression model. The fit is almost perfect. Could this be due to overfitting the data? Alternatively, could it be that the biological information in the gene expression levels indeed determines the leukemia subtype almost perfectly?
- c) To discern between the two explanations for the almost perfect fit, randomly shuffle the subtypes. Refit the logistic regression model and obtain the fits. On the basis of this and the previous fit, which explanation is more plausible?
- d) Compare the fit of the logistic model with different penalty parameters, say $\lambda = 1$ and $\lambda = 1000$. How does λ influence the possibility of overfitting the data?
- e) Describe what you would do to prevent overfitting.

Question 5.4

Load the breast cancer data available via the `breastCancerNKI`-package (downloadable from BioConductor) through the following R-code:

Listing 5.3 R code

```
# activate library and load the data
library(breastCancerNKI)
data(nki)

# subset and center the data
X <- exprs(nki)
X <- X[-which(rowSums(is.na(X)) > 0),]
X <- apply(X[1:1000,], 1, function(X) { X - mean(X) })

# define ER as the response
Y <- pData(nki)[,8]
```

The `eSet`-object `nki` is now available. It contains the expression profiles of 337 breast cancer patients. Each profile comprises expression levels of 24481 genes. The R-code above extracts the expression data from the object, removes all genes with missing values, centers the gene expression gene-wise around zero, and subsets the data set to the first thousand genes. The reduction of the gene dimensionality is only for computational speed. Furthermore, it extracts the estrogen receptor status (short: ER status), an important prognostic indicator for breast cancer, that is to be used as the response variable in the remainder of the exercise.

- a) Relate the ER status and the gene expression levels by a logistic regression model, which is fitted by means of ridge penalized maximum likelihood. First, find the optimal value of the penalty parameter of λ by means of cross-validation. This is implemented in `optL2`-function of the `penalized`-package available from CRAN.
- b) Evaluate whether the cross-validated likelihood indeed attains a maximum at the optimal value of λ . This can be done with the `profL2`-function of the `penalized`-package available from CRAN.
- c) Investigate the sensitivity of the penalty parameter selection with respect to the choice of the cross-validation fold.

d) Does the optimal lambda produce a reasonable fit?

Question 5.5

Derive the equivalent of Equations (5.2) and (5.3) for the targeted ridge logistic regression estimator:

$$\hat{\beta}(\lambda; \beta_0) = \arg \min_{\beta \in \mathbb{R}^p} \mathbf{Y}^\top \mathbf{X} \hat{\beta}(\lambda; \beta_0) - \sum_{i=1}^n \log\{1 + \exp[\mathbf{X}_{i,*} \hat{\beta}(\lambda; \beta_0)]\} - \frac{1}{2}\lambda \|\hat{\beta}(\lambda; \beta_0) - \beta_0\|_2^2,$$

with nonrandom $\beta_0 \in \mathbb{R}^p$.

Question 5.6

The iteratively reweighted least squares (IRLS) algorithm for the numerical evaluation of the ridge logistic regression estimator requires the inversion of a $p \times p$ -dimensional matrix at each iteration. In Section 1.7 the singular value decomposition (SVD) of the design matrix is exploited to avoid the inversion of such a matrix in the numerical evaluation of the ridge regression estimator. Use this trick to show that the computational burden of the IRLS algorithm may be reduced to one SVD prior to the iterations and the inversion of an $n \times n$ dimensional matrix at each iteration (as is done in Eilers *et al.*, 2001).

Question 5.7

The ridge estimator of parameter β of the logistic regression model is the maximizer of the ridge penalized loglikelihood:

$$\mathcal{L}^{\text{pen}}(\mathbf{Y}, \mathbf{X}; \beta, \lambda) = \sum_{i=1}^n \{Y_i \mathbf{X}_i \beta - \log[1 + \exp(\mathbf{X}_i \beta)]\} - \frac{1}{2}\lambda \|\beta\|_2^2.$$

The maximizer is found numerically by the iteratively reweighted least squares (IRLS) algorithm which is outlined in Section 5.3. Modify the algorithm, as is done in van Wieringen and Binder (2022), to find the generalized ridge logistic regression estimator of β defined as:

$$\mathcal{L}^{\text{pen}}(\mathbf{Y}, \mathbf{X}; \beta, \lambda) = \sum_{i=1}^n \{Y_i \mathbf{X}_i \beta - \log[1 + \exp(\mathbf{X}_i \beta)]\} - \frac{1}{2}(\beta - \beta_0)^\top \Delta(\beta - \beta_0),$$

where β_0 and Δ are as in Chapter 3.

Question 5.8

Consider the logistic regression model to explain the variation of a binary random variable by a covariate. Let $Y_i \in \{0, 1\}$, $i = 1, \dots, n$, represent n binary random variables that follow a Bernoulli distribution with parameter $P(Y_i = 1 | X_i) = \exp(X_i \beta)[1 + \exp(X_i \beta)]$ and corresponding covariate $X_i \in \{-1, 1\}$. Data $\{(y_i, X_i)\}_{i=1}^n$ are summarized as a contingency table (Table 5.1):

	$y_i = 0$	$y_i = 1$
$X_i = -1$	301	196
$X_i = 1$	206	297

Table 5.1: Data for Exercise 5.8.

The data of Table 5.1 is only to be used in parts c) and g).

a) Show that the estimating equation of the ridge logistic regression estimator of β is of the form:

$$c - n \exp(\beta)[1 + \exp(\beta)]^{-1} - \lambda \beta = 0,$$

and specify c .

- b) Show that $\hat{\beta}(\lambda) \in (\lambda^{-1}(c - n), \lambda^{-1}c)$ for all $\lambda > 0$. Ensure that this is a meaningful interval, i.e. that it is non-empty, and verify that $c > 0$. Moreover, conclude from the interval that $\lim_{\lambda \rightarrow \infty} \hat{\beta}(\lambda) = 0$.
- c) The Taylor expansion of $\exp(\beta)[1 + \exp(\beta)]^{-1}$ around $\beta = 0$ is:

$$\exp(\beta)[1 + \exp(\beta)]^{-1} = \frac{1}{2} + \frac{1}{4}x - \frac{1}{48}x^3 + \frac{1}{480}x^5 + \mathcal{O}(x^6).$$

Substitute the 3rd order Taylor approximation into the estimating equations and use the data of Table 5.1 to evaluate its roots using the `polyroot`-function (of the `base`-package). Do the same for the 1st and 2nd order Taylor approximations of $\exp(\beta)[1 + \exp(\beta)]^{-1}$. Compare these approximate estimates to the one provided by the `ridgeGLM`-function (of the `porridge`-package, van Wieringen and Aflakparast, 2021).

In the remainder consider the 1st order Taylor approximated ridge logistic regression estimator: $\hat{\beta}_{1^{\text{st}}}(\lambda) = (\lambda + \frac{1}{4}n)^{-1}(c - \frac{1}{2}n)$.

- d) Find an expression for $\mathbb{E}[\hat{\beta}_{1^{\text{st}}}(\lambda)]$.
- e) Find an expression for $\text{Var}[\hat{\beta}_{1^{\text{st}}}(\lambda)]$.
- f) Combine the answers of parts d) and e) to find an expression for $\text{MSE}[\hat{\beta}_{1^{\text{st}}}(\lambda)]$.
- g) Plot this MSE against λ for the data provided in Table 5.1 and reflect on an analogue of Theorem 1.2 for the ridge logistic regression estimator.

Question 5.9

Consider the logistic regression model $Y_i \sim \text{Bernoulli}\{\exp(\mathbf{X}_{i,*}\boldsymbol{\beta})[\exp(\mathbf{X}_{i,*}\boldsymbol{\beta})]^{-1}\}$ for $i = 1, \dots, n$ and regression parameter $\boldsymbol{\beta} \in \mathbb{R}^p$, response variable Y_i and the p -dimensional row vector $\mathbf{X}_{i,*}$ with the covariate information. Suppose the samples can be divided into two groups, one of size n_0 as the covariate information and the other of size n_1 such that $n = n_0 + n_1$, with data $Y_i = 0$ and $\mathbf{X}_{i,*} = \mathbf{X}_{1,*}$ for $i = 1, \dots, n_0$ and $Y_i = 1$ and $\mathbf{X}_{i,*} = -\mathbf{X}_{1,*}$ for $i = n_0 + 1, \dots, n_0 + n_1$. From these data, the regression parameter is estimated with the ridge logistic regression parameter which maximizes the loglikelihood augmented with the ridge penalty $\frac{1}{2}\lambda\|\boldsymbol{\beta}\|_2^2$ with penalty parameter $\lambda_2 > 0$. Show that the ridge logistic regression estimator $\hat{\boldsymbol{\beta}}(\lambda_2)$ is of the form $c(\lambda_2)\mathbf{X}_{1,*}^\top$ with $c(\lambda_2) \in \mathbb{R}$.

6 Lasso regression

In this chapter we return to the linear regression model, which is still fitted in penalized fashion but this time with a so-called lasso penalty. Yet another penalty? Yes, but the resulting estimator will turn out to have interesting properties. The outline of this chapter loosely follows that of its counterpart on ridge regression (Chapter 1). The chapter can – at least partially – be seen as an elaborated version of the original work on lasso regression, i.e. Tibshirani (1996), with most topics covered and visualized more extensively and incorporating results and examples published since.

Recall that ridge regression finds an estimator of the parameter of the linear regression model through the minimization of:

$$\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + f_{\text{pen}}(\beta, \lambda), \quad (6.1)$$

with $f_{\text{pen}}(\beta, \lambda) = \lambda\|\beta\|_2^2$. The particular choice of the penalty function originated in a post-hoc motivation of the ad-hoc fix to the singularity of the matrix $\mathbf{X}^\top \mathbf{X}$, stemming from the design matrix \mathbf{X} not being of full rank (i.e. $\text{rank}(\mathbf{X}) < p$). The ad-hoc nature of the fix suggests that the choice for the squared Euclidean norm of β as a penalty is arbitrary and other choices may be considered, some of which were already encountered in Chapter 3.

One such choice is the so-called lasso penalty, giving rise to lasso regression, as introduced by Tibshirani (1996). Like ridge regression, lasso regression fits the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with the standard assumption on the error ε . Like ridge regression, it does so by minimization of the sum of squares augmented with a penalty. Hence, lasso regression too minimizes loss function (6.1). The difference with ridge regression is in the penalty function. Instead of the squared Euclidean norm, lasso regression uses the ℓ_1 -norm: $f_{\text{pen}}(\beta, \lambda_1) = \lambda_1\|\beta\|_1$, the sum of the absolute values of the regression parameters multiplied by the lasso penalty parameter λ_1 . To distinguish the ridge and lasso penalty parameters, they are henceforth denoted λ_2 and λ_1 , respectively, with the subscript referring to the norm used in the penalty. The lasso regression loss function is thus:

$$\mathcal{L}_{\text{lasso}}(\beta; \lambda) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1 = \sum_{i=1}^n (Y_i - \mathbf{X}_{i*}\beta)^2 + \lambda_1 \sum_{j=1}^p |\beta_j|. \quad (6.2)$$

The lasso regression estimator is then defined as the minimizer of this loss function. As with the ridge regression loss function, the maximum likelihood estimator (should it exists) of β minimizes the first part, while the second part is minimized by setting β equal to the p dimensional zero vector. For λ_1 close to zero, the lasso estimate is close to the maximum likelihood estimate. Whereas for large λ_1 , the penalty term overshadows the sum-of-squares, and the lasso estimate is small (in some sense). Intermediate choices of λ_1 mold a compromise between those two extremes, with the penalty parameter determining the contribution of each part to this compromise. The lasso regression estimator thus is not one but a whole sequence of estimators of β , one for every $\lambda_1 \in \mathbb{R}_{>0}$. This sequence is the lasso regularization path, defined as $\{\hat{\beta}(\lambda_1) : \lambda_1 \in \mathbb{R}_{>0}\}$. To arrive at a final lasso estimator of β , like its ridge counterpart, the lasso penalty parameter λ_1 needs to be chosen (see Section 6.8).

The ℓ_1 penalty of lasso regression is equally arbitrary as the ℓ_2 -penalty of ridge regression. The latter ensured the existence of a well-defined estimator of the regression parameter β in the presence of super-collinearity in the design matrix \mathbf{X} , in particular when the dimension p exceeds the sample size n . The augmentation of the sum-of-squares with the lasso penalty achieves the same. This is illustrated in Figure 6.1. For the high-dimensional setting with $p = 2$ and $n = 1$ and arbitrary data the level sets of the sum-of-squares $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$ and the lasso regression loss $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1$ are plotted (left and right panel, respectively). In both panels the minimum is indicated in red. For the sum-of-squares the minimum is a line. As pointed out before in Section 1.2 of Chapter 1 on ridge regression, this minimum is determined up to an element of the null set of the design matrix \mathbf{X} , which in this case is non-trivial. In contrast, the lasso regression loss exhibits a unique well-defined minimum. Hence, the augmentation of the sum-of-squares with the lasso penalty yields a well-defined estimator of the regression parameter. (This needs some attenuation: in general the minimum of the lasso regression loss need not be unique, confer Section 6.1).

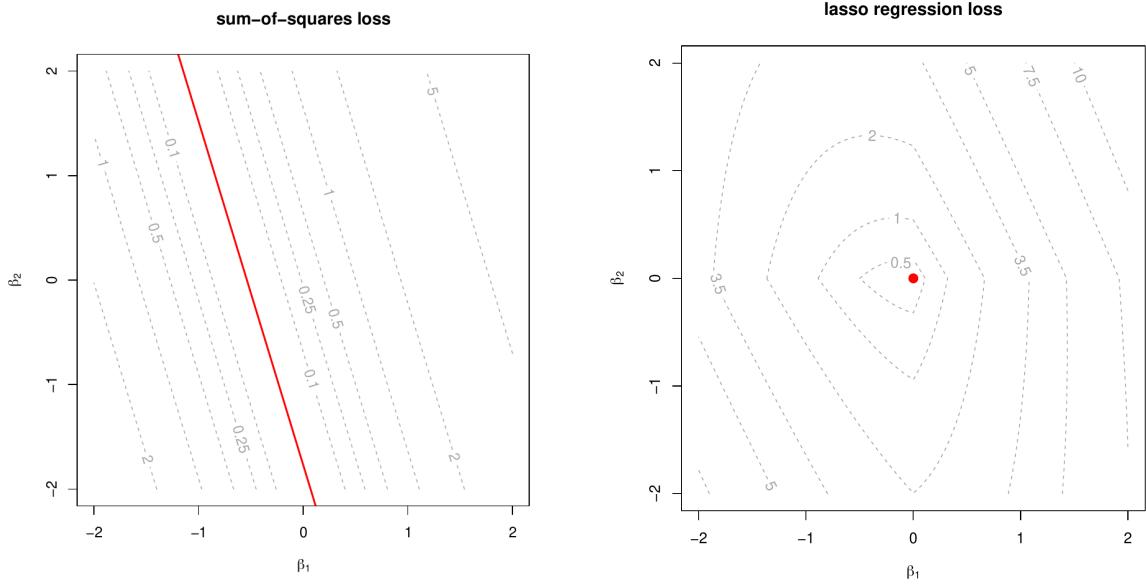


Figure 6.1: Contour plots of the sum-of-squares and the lasso regression loss (left and right panel, respectively). The dotted grey line represent level sets. The red line and dot represent the location of minimum in both panels.

The mathematics involved in the derivation in this chapter tends to be more intricate than for ridge regression. This is due to the non-differentiability of the lasso penalty at zero. This has consequences on all aspects of the lasso regression estimator as is already obvious in the right-hand panel of Figure 6.1: confer the non-differentiable points of the lasso regression loss level sets.

6.1 Uniqueness

The lasso regression estimator need not be unique. This can loosely be concluded from the lasso regression loss function, which is the sum of the sum-of-squares criterion and a sum of absolute value functions. Both are convex in β : the former is not strict convex due to the high-dimensionality and the absolute value function is convex due to its piece-wise linearity. Thereby the lasso loss function too is convex but not strict. Consequently, its minimum need not be uniquely defined. But, the set of solutions of a convex minimization problem is convex (Theorem 9.4.1, Fletcher, 1987). Hence, would there exist multiple minimizers of the lasso loss function, they can be used to construct a convex set of minimizers. Thus, if $\hat{\beta}_a(\lambda_1)$ and $\hat{\beta}_b(\lambda_1)$ are lasso estimators, then so are $(1 - \theta)\hat{\beta}_a(\lambda_1) + \theta\hat{\beta}_b(\lambda_1)$ for $\theta \in (0, 1)$. This is illustrated in Example 6.1.

Example 6.1 (Super-collinear covariates)

Consider the standard linear regression model $Y_i = \mathbf{X}_{i,*}\beta + \varepsilon_i$ for $i = 1, \dots, n$ and with the ε_i i.i.d. normally distributed with zero mean and a common variance. The rows of the design matrix \mathbf{X} are of length two, neither column represents the intercept, but $\mathbf{X}_{*,1} = \mathbf{X}_{*,2}$. Suppose an estimate of the regression parameter β of this model is obtained through the minimization of the sum-of-squares augmented with a lasso penalty, $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1$ with penalty parameter $\lambda_1 > 0$. To find the minimizer define $u = \beta_1 + \beta_2$ and $v = \beta_1 - \beta_2$ and rewrite the lasso loss criterion to:

$$\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1 = \|\mathbf{Y} - \mathbf{X}_{*,1}u\|_2^2 + \frac{1}{2}\lambda_1(|u+v| + |u-v|).$$

The function $|u+v| + |u-v|$ is minimized with respect to v for any v such that $|v| < |u|$ and the corresponding minimum equals $2|u|$. The estimator of u thus minimizes:

$$\|\mathbf{Y} - \mathbf{X}_{*,1}u\|_2^2 + \lambda_1|u|.$$

For sufficiently small values of λ_1 the estimate of u will be unequal to zero. Then, any v such that $|v| < |u|$ will yield the same minimum of the lasso loss function. Consequently, $\hat{\beta}(\lambda_1)$ is not uniquely defined as $\hat{\beta}_1(\lambda_1) = \frac{1}{2}[\hat{u}(\lambda_1) + \hat{v}(\lambda_1)]$ need not equal $\hat{\beta}_2(\lambda_1) = \frac{1}{2}[\hat{u}(\lambda_1) - \hat{v}(\lambda_1)]$ for any $\hat{v}(\lambda_1)$ such that $0 < |\hat{v}(\lambda_1)| < |\hat{u}(\lambda_1)|$. \square

The non-uniqueness of the lasso regression estimator may not be its most appealing property, it is unproblematic in many practical settings. The following lemma states that the estimator is, with high probability and under a mild condition on the design matrix, unique.

Lemma 6.1 (Lemma 4, Tibshirani, 2013)

If the elements of $\mathbf{X} \in \mathcal{M}^{n,p}$ are drawn from a continuous probability distribution on $\mathbb{R}^{n \times p}$, then for any \mathbf{Y} and $\lambda > 0$, the lasso solution is unique with probability one. ■

Proof. See Tibshirani (2013).

Moreover, while the lasso regression estimator $\hat{\beta}(\lambda_1)$ need not be unique, its linear predictor $\mathbf{X}\hat{\beta}(\lambda_1)$ is. This can be proven by contradiction (Tibshirani, 2013). Suppose there exists two lasso regression estimators of β , denoted $\hat{\beta}_a(\lambda_1)$ and $\hat{\beta}_b(\lambda_1)$, such that $\mathbf{X}\hat{\beta}_a(\lambda_1) \neq \mathbf{X}\hat{\beta}_b(\lambda_1)$. Define c to be the minimum of the lasso loss function. Then, by definition of the lasso estimators $\hat{\beta}_a(\lambda_1)$ and $\hat{\beta}_b(\lambda_1)$ satisfy:

$$\|\mathbf{Y} - \mathbf{X}\hat{\beta}_a(\lambda_1)\|_2^2 + \lambda_1\|\hat{\beta}_a(\lambda_1)\|_1 = c = \|\mathbf{Y} - \mathbf{X}\hat{\beta}_b(\lambda_1)\|_2^2 + \lambda_1\|\hat{\beta}_b(\lambda_1)\|_1.$$

For $\theta \in (0, 1)$ we then have:

$$\begin{aligned} & \|\mathbf{Y} - \mathbf{X}[(1-\theta)\hat{\beta}_a(\lambda_1) + \theta\hat{\beta}_b(\lambda_1)]\|_2^2 + \lambda_1\|(1-\theta)\hat{\beta}_a(\lambda_1) + \theta\hat{\beta}_b(\lambda_1)\|_1 \\ &= \|(1-\theta)[\mathbf{Y} - \mathbf{X}\hat{\beta}_a(\lambda_1)] + \theta[\mathbf{Y} - \mathbf{X}\hat{\beta}_b(\lambda_1)]\|_2^2 + \lambda_1\|(1-\theta)\hat{\beta}_a(\lambda_1) + \theta\hat{\beta}_b(\lambda_1)\|_1 \\ &< (1-\theta)\|\mathbf{Y} - \mathbf{X}\hat{\beta}_a(\lambda_1)\|_2^2 + \theta\|\mathbf{Y} - \mathbf{X}\hat{\beta}_b(\lambda_1)\|_2^2 + (1-\theta)\lambda_1\|\hat{\beta}_a(\lambda_1)\|_1 + \theta\lambda_1\|\hat{\beta}_b(\lambda_1)\|_1 \\ &= (1-\theta)c + \theta c = c, \end{aligned}$$

by the strict convexity of $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$ in $\mathbf{X}\beta$ and the convexity of $\|\beta\|_1$ on $\theta \in (0, 1)$. This implies that $(1-\theta)\hat{\beta}_a(\lambda_1) + \theta\hat{\beta}_b(\lambda_1)$ yields a lower minimum of the lasso loss function and contradicts our assumption that $\hat{\beta}_a(\lambda_1)$ and $\hat{\beta}_b(\lambda_1)$ are lasso regression estimators. Hence, the lasso linear predictor is unique.

Example 6.2 (*Perfectly super-collinear covariates, revisited*)

Revisit the setting of Example 6.1, where a linear regression model without intercept and only two but perfectly correlated covariates is fitted to data. The example revealed that the lasso estimator need not be unique. The lasso predictor, however, is

$$\hat{\mathbf{Y}}(\lambda_1) = \mathbf{X}\hat{\beta}(\lambda_1) = \mathbf{X}_{*,1}\hat{\beta}_1(\lambda_1) + \mathbf{X}_{*,2}\hat{\beta}_2(\lambda_1) = \mathbf{X}_{*,1}[\hat{\beta}_1(\lambda_1) + \hat{\beta}_2(\lambda_1)] = \mathbf{X}_{*,1}\hat{u}(\lambda_1),$$

with u defined and (uniquely) estimated as in Example 6.1 and v dropping from the predictor. □

Example 6.3

The issues, non- and uniqueness of the lasso-estimator and predictor, respectively, raised above are illustrated in a numerical setting. Hereto data are generated in accordance with the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ where the $n = 5$ rows of \mathbf{X} are sampled from $\mathcal{N}[\mathbf{0}_p, (1-\rho)\mathbf{I}_{pp} + \rho\mathbf{1}_{pp}\mathbf{1}_{pp}^\top]$ with $p = 10$, $\rho = 0.99$, $\beta = (\mathbf{1}_3^\top, \mathbf{0}_{p-3}^\top)^\top$ and $\varepsilon \sim \mathcal{N}(\mathbf{0}_p, \frac{1}{10}\mathbf{I}_{nn})$. With these data the lasso estimator of the regression parameter β for $\lambda_1 = 1$ is evaluated using two different algorithms (see Section 6.4). Employed implementations of the algorithms are those available through the R-packages `penalized` and `glmnet`. Both estimates, denoted $\hat{\beta}_p(\lambda_1)$ and $\hat{\beta}_g(\lambda_1)$ (the subscript refers to the first letter of the package), are given in Table 6.3. The table reveals that the estimates differ, in par-

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	
penalized	$\hat{\beta}_p(\lambda_1)$	0.267	0.000	1.649	0.093	0.000	0.000	0.000	0.571	0.000	0.269
glmnet	$\hat{\beta}_g(\lambda_1)$	0.269	0.000	1.776	0.282	0.195	0.000	0.000	0.325	0.000	0.000

Table 6.1: Lasso estimates of the linear regression β for both algorithms.

ticular in their support (i.e. the set of nonzero values of the estimate of β). This is troublesome when it comes to communication of the optimal model. From a different perspective the realized loss $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1$ for each estimate is approximately equal to 2.99, with the difference possibly due to convergence criteria of the algorithms. On another note, their corresponding predictors, $\mathbf{X}\hat{\beta}_p(\lambda_1)$ and $\mathbf{X}\hat{\beta}_g(\lambda_1)$, correlate almost perfectly:

$\text{cor}[\hat{\mathbf{X}}\hat{\boldsymbol{\beta}}_p(\lambda_1), \hat{\mathbf{X}}\hat{\boldsymbol{\beta}}_g(\lambda_1)] = 0.999$. These results thus corroborate the non-uniqueness of the estimator and the uniqueness of the predictor.

The R-script provides the code to reproduce the analysis.

Listing 6.1 R code

```
# set the random seed
set.seed(4)

# load libraries
library(penalized)
library(glmnet)
library(mvtnorm)

# set sample size
p <- 10

# create covariance matrix
Sigma <- matrix(0.99, p, p)
diag(Sigma) <- 1

# sample the design matrix
n <- 5
X <- rmvnorm(10, sigma=Sigma)

# create a sparse beta vector
betas <- c(rep(1, 3), rep(0, p-3))

# sample response
Y <- X %*% betas + rnorm(n, sd=0.1)

# evaluate lasso estimator with two methods
Bhat1 <- matrix(as.numeric(coef(penalized(Y, X, lambda=1, unpenalized=~0),
                                "all")), ncol=1)
Bhat2 <- matrix(as.numeric(coef(glmnet(X, Y, lambda=1/(2*n), standardize=FALSE,
                                         intercept=FALSE)))[-1], ncol=1)

# compare estimates
cbind(Bhat1, Bhat2)

# compare the loss
sum((Y - X %*% Bhat1)^2) + sum(abs(Bhat1))
sum((Y - X %*% Bhat2)^2) + sum(abs(Bhat2))

# compare predictor
cor(X %*% Bhat1, X %*% Bhat2)
```

Note that in the code above the evaluation of the lasso estimator appears to employ a different lasso penalty parameter λ_1 for each package. This is due to the fact that internally (after removal of standardization of \mathbf{X} and \mathbf{Y}) the loss functions optimized are $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1$ vs. $(2n)^{-1}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1$. Rescaling of λ_1 resolves this issue. \square

6.2 Analytic solutions

In general, no explicit expression for the lasso regression estimator exists. There are exceptions, as illustrated in Examples 6.4 and 6.6. Nonetheless, it is possible to show properties of the lasso estimator, amongst others of the smoothness of its regularization path (Theorem 6.1) and the limiting behaviour as $\lambda_1 \rightarrow \infty$ (see the end of this section).

Theorem 6.1 (Theorem 2, Rosset and Zhu, 2007)

The lasso regression loss function (6.2) yields a piecewise linear, in λ_1 , regularization path $\{\hat{\boldsymbol{\beta}}(\lambda_1) : \mathbb{R}_{>0}\}$.

Proof. Confer Rosset and Zhu (2007). ■

This piecewise linear nature of the lasso solution path is illustrated in the left-hand panel of Figure 6.2 of an arbitrary data set. At each vertical dotted line a discontinuity in the derivative with respect to λ_1 of the regularization path of a lasso estimate of an element of β may occur. The plot also foreshadows the $\lambda_1 \rightarrow \infty$ limiting behaviour of the lasso regression estimator: the estimator tends to zero. This is no surprise knowing that the ridge regression estimator exhibits the same behaviour and the lasso regression loss function is of similar form as that of ridge regression: a sum-of-squares plus a penalty term (which is linear in the penalty parameter).

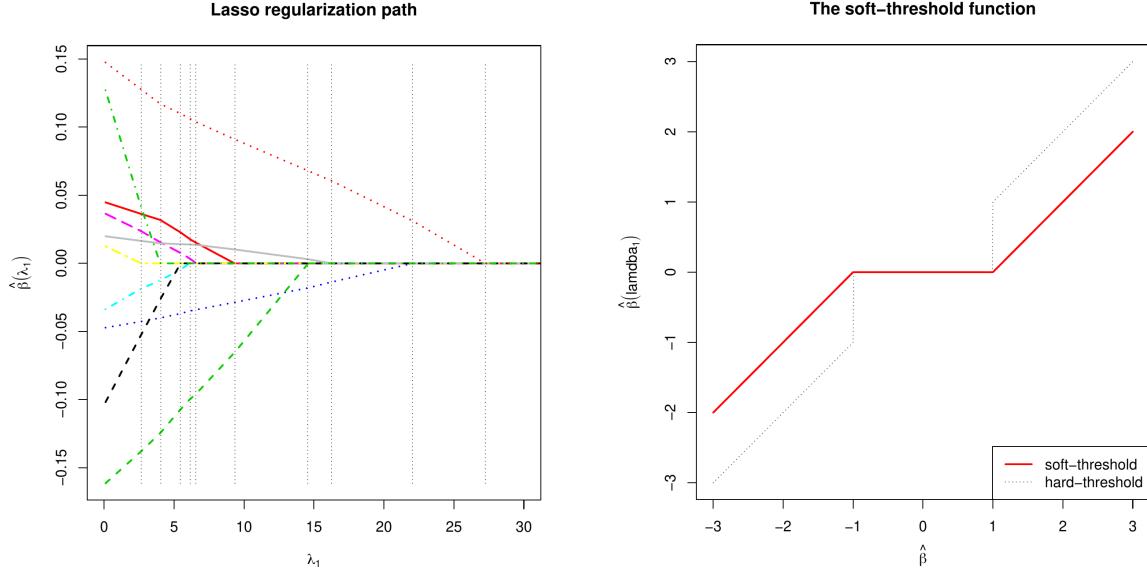


Figure 6.2: The left panel shows the regularization path of the lasso regression estimator for simulated data. The vertical grey dotted lines indicate the values of λ_1 at which there is a discontinuity in the derivative (with respect to λ_1) of the lasso regularization path of one the regression estimates. The right panel displays the soft (solid, red) and hard (grey, dotted) threshold functions.

For particular cases, an orthonormal design (Example 6.4) and $p = 2$ (Example 6.6), an analytic expression for the lasso regression estimator exists. While the latter is of limited use, the former is exemplary and will come of use later in the numerical evaluation of the lasso regression estimator in the general case (see Section 6.4).

Example 6.4 (Orthonormal design matrix)

Consider an orthonormal design matrix \mathbf{X} , i.e. $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_{pp} = (\mathbf{X}^\top \mathbf{X})^{-1}$. The lasso estimator then is:

$$\hat{\beta}_j(\lambda_1) = \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \frac{1}{2}\lambda_1)_+,$$

where $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \mathbf{X}^\top \mathbf{Y}$ is the maximum likelihood estimator of β and $\hat{\beta}_j$ its j -th element and $f(x) = (x)_+ = \max\{x, 0\}$. This expression for the lasso regression estimator can be obtained as follows. Rewrite the lasso regression loss criterion:

$$\begin{aligned} \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1 &= \min_{\beta} \mathbf{Y}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X}\beta - \beta^\top \mathbf{X}^\top \mathbf{Y} + \beta^\top \mathbf{X}^\top \mathbf{X}\beta + \lambda_1 \sum_{j=1}^p |\beta_j| \\ &\propto \min_{\beta} -\hat{\beta}^\top \beta - \beta^\top \hat{\beta} + \beta^\top \beta + \lambda_1 \sum_{j=1}^p |\beta_j| \\ &= \min_{\beta_1, \dots, \beta_p} \sum_{j=1}^p (-2\hat{\beta}_j \beta_j + \beta_j^2 + \lambda_1 |\beta_j|) \\ &= \sum_{j=1}^p \left(\min_{\beta_j} -2\hat{\beta}_j \beta_j + \beta_j^2 + \lambda_1 |\beta_j| \right). \end{aligned}$$

The minimization problem can thus be solved per regression coefficient. This gives:

$$\min_{\beta_j} -2\hat{\beta}_j \beta_j + \beta_j^2 + \lambda_1 |\beta_j| = \begin{cases} \min_{\beta_j} -2\hat{\beta}_j \beta_j + \beta_j^2 + \lambda_1 \beta_j & \text{if } \beta_j > 0, \\ \min_{\beta_j} -2\hat{\beta}_j \beta_j + \beta_j^2 - \lambda_1 \beta_j & \text{if } \beta_j < 0. \end{cases}$$

The minimization within the sum over the covariates is with respect to each element of the regression parameter separately. Optimization with respect to the j -th one gives:

$$\hat{\beta}_j(\lambda_1) = \begin{cases} \hat{\beta}_j - \frac{1}{2}\lambda_1 & \text{if } \hat{\beta}_j(\lambda_1) > 0 \\ \hat{\beta}_j + \frac{1}{2}\lambda_1 & \text{if } \hat{\beta}_j(\lambda_1) < 0 \\ 0 & \text{otherwise} \end{cases}$$

This case-wse solution can be compressed to the form of the lasso regression estimator above.

The analytic expression for the lasso regression estimator above provides insight in how it relates to the maximum likelihood estimator of β . The right-hand side panel of Figure 6.2 depicts this relationship. Effectively, the lasso regression estimator thresholds (after a translation) its maximum likelihood counterpart. The function is also referred to as the *soft-threshold function* (for contrast the hard-threshold function is also plotted – dotted line – in Figure 6.2). \square

Example 6.5 (Orthogonal design matrix)

The analytic solution of the lasso regression estimator for experiments with an orthonormal design matrix applies to those with an orthogonal design matrix. This is illustrated by a numerical example. Use the lasso estimator with $\lambda_1 = 10$ to fit the linear regression model to the response data and the design matrix:

$$\begin{aligned} \mathbf{Y}^\top &= (-4.9 \quad -0.8 \quad -8.9 \quad 4.9 \quad 1.1 \quad -2.0), \\ \mathbf{X}^\top &= \begin{pmatrix} 1 & -1 & 3 & -3 & 1 & 1 \\ -3 & -3 & -1 & 0 & 3 & 0 \end{pmatrix}. \end{aligned}$$

Note that the design matrix is orthogonal, i.e. its columns are orthogonal (but not normalized to one). The orthogonality of \mathbf{X} yields a diagonal $\mathbf{X}^\top \mathbf{X}$, and so is its inverse $(\mathbf{X}^\top \mathbf{X})^{-1}$. Here $\text{diag}(\mathbf{X}^\top \mathbf{X}) = (22, 28)$. Rescale \mathbf{X} to an orthonormal design matrix, denoted $\tilde{\mathbf{X}}$, and rewrite the lasso regression loss function to:

$$\begin{aligned} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1 &= \left\| \mathbf{Y} - \mathbf{X} \begin{pmatrix} \sqrt{22} & 0 \\ 0 & \sqrt{28} \end{pmatrix}^{-1} \begin{pmatrix} \sqrt{22} & 0 \\ 0 & \sqrt{28} \end{pmatrix} \beta \right\|_2^2 + \lambda_1\|\beta\|_1 \\ &= \|\mathbf{Y} - \tilde{\mathbf{X}}\gamma\|_2^2 + (\lambda_1/\sqrt{22})|\gamma_1| + (\lambda_1/\sqrt{28})|\gamma_2|, \end{aligned}$$

where $\gamma = (\sqrt{22}\beta_1, \sqrt{28}\beta_2)^\top$. By the same argument this loss can be minimized with respect to each element of γ separately. In particular, the soft-threshold function provides an analytic expression for the estimates of γ :

$$\begin{aligned} \hat{\gamma}_1(\lambda_1/\sqrt{22}) &= \text{sign}(\hat{\gamma}_1)[|\hat{\gamma}_1| - \frac{1}{2}(\lambda_1/\sqrt{22})]_+ = -[9.892513 - \frac{1}{2}(10/\sqrt{22})]_+ = -8.826509, \\ \hat{\gamma}_2(\lambda_1/\sqrt{28}) &= \text{sign}(\hat{\gamma}_2)[|\hat{\gamma}_2| - \frac{1}{2}(\lambda_1/\sqrt{28})]_+ = [5.537180 - \frac{1}{2}(10/\sqrt{28})]_+ = 4.592269, \end{aligned}$$

where $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are the ordinary least square estimates of γ_1 and γ_2 obtained from regressing \mathbf{Y} on the corresponding column of $\tilde{\mathbf{X}}$. Rescale back and obtain the lasso regression estimate: $\hat{\beta}(10) = (-1.881818, 0.8678572)^\top$. \square

Example 6.6 ($p = 2$ with equivariant covariates, Leng et al., 2006)

Let $p = 2$ and suppose the design matrix \mathbf{X} has equivariant covariates. Without loss of generality they are assumed to have unit variance. We may thus write

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

for some $\rho \in (-1, 1)$. The lasso regression estimator is then of similar form as in the orthonormal case but the soft-threshold function now depends on λ_1 , ρ and the maximum likelihood estimate $\hat{\beta}$ (see Exercise 6.8). \square

Apart from the specific cases outlined in the two examples above no other explicit solution for the minimizer of the lasso regression loss function appears to be known. Locally though, for large enough values of λ_1 , an analytic expression for solution can also be derived. Hereto we point out that (details to be included later) the lasso regression estimator satisfies the following estimating equation:

$$\mathbf{X}^\top \mathbf{X} \hat{\beta}(\lambda_1) = \mathbf{X}^\top \mathbf{Y} - \frac{1}{2}\lambda_1 \hat{\mathbf{z}}$$

for some $\hat{\mathbf{z}} \in \mathbb{R}^p$ with $(\hat{\mathbf{z}})_j = \text{sign}\{[\hat{\beta}(\lambda_1)]_j\}$ whenever $[\hat{\beta}_j(\lambda_1)]_j \neq 0$ and $(\hat{\mathbf{z}})_j \in [-1, 1]$ if $[\hat{\beta}_j(\lambda_1)]_j = 0$. Then:

$$0 \leq [\hat{\beta}(\lambda_1)]^\top \mathbf{X}^\top \mathbf{X} \hat{\beta}(\lambda_1) = [\hat{\beta}(\lambda_1)]^\top (\mathbf{X}^\top \mathbf{Y} - \frac{1}{2} \lambda_1 \hat{\mathbf{z}}) = \sum_{j=1}^p [\hat{\beta}(\lambda_1)]_j (\mathbf{X}^\top \mathbf{Y} - \frac{1}{2} \lambda_1 \hat{\mathbf{z}})_j.$$

For $\lambda_1 > 2\|\mathbf{X}^\top \mathbf{Y}\|_\infty$ the summands on the right-hand side satisfy:

$$\begin{aligned} [\hat{\beta}(\lambda_1)]_j (\mathbf{X}^\top \mathbf{Y} - \frac{1}{2} \lambda_1 \hat{\mathbf{z}})_j &< 0 \quad \text{if } [\hat{\beta}(\lambda_1)]_j > 0, \\ [\hat{\beta}(\lambda_1)]_j (\mathbf{X}^\top \mathbf{Y} - \frac{1}{2} \lambda_1 \hat{\mathbf{z}})_j &= 0 \quad \text{if } [\hat{\beta}(\lambda_1)]_j = 0, \\ [\hat{\beta}(\lambda_1)]_j (\mathbf{X}^\top \mathbf{Y} - \frac{1}{2} \lambda_1 \hat{\mathbf{z}})_j &< 0 \quad \text{if } [\hat{\beta}(\lambda_1)]_j < 0. \end{aligned}$$

This implies that $\hat{\beta}(\lambda_1) = \mathbf{0}_p$ if $\lambda_1 > 2\|\mathbf{X}^\top \mathbf{Y}\|_\infty$, where $\|\mathbf{a}\|_\infty$ is the supremum norm of vector \mathbf{a} defined as $\|\mathbf{a}\|_\infty = \max\{|a_1|, |a_2|, \dots, |a_p|\}$.

6.3 Sparsity

The change from the ℓ_2^2 -norm to the ℓ_1 -norm in the penalty may seem only a detail. Indeed, both ridge and lasso regression fit the same linear regression model. But the attractiveness of the lasso lies not in *what* it fits, but in a *consequence of how* it fits the linear regression model. The lasso estimator of the vector of regression parameters may contain some or many zero's. In contrast, ridge regression yields an estimator of β with elements (possibly) close to zero, but unlikely to be equal to zero. Hence, lasso penalization results in $\hat{\beta}_j(\lambda_1) = 0$ for some j (in particular for large values of λ_1 , see Section 6.1), while ridge penalization yields an estimate of the j -th element of the regression parameter $\hat{\beta}_j(\lambda_2) \neq 0$. A zero estimate of a regression coefficient means that the corresponding covariate has no effect on the response and can be excluded from the model. Effectively, this amounts to variable selection. Where traditionally the linear regression model is fitted by means of maximum likelihood followed by a testing step to weed out the covariates with effects indistinguishable from zero, lasso regression is a one-step-go procedure that simultaneously estimates and selects.

The in-built variable selection of the lasso regression estimator is a geometric accident. To understand how it comes about the lasso regression loss optimization problem (6.2) is reformulated as a constrained estimation problem (using the same argumentation as previously employed for ridge regression, see Section 1.5):

$$\hat{\beta}(\lambda_1) = \arg \min_{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq c(\lambda_1)} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2.$$

where $c(\lambda_1) = \|\hat{\beta}(\lambda_1)\|_1$. Again, this is the standard least squares problem, with the only difference that the sum of the (absolute value of the) regression parameters $\beta_1, \beta_2, \dots, \beta_p$ is required to be smaller than $c(\lambda_1)$. The effect of this requirement is that the lasso estimator of the regression parameter coefficients can no longer assume any value (from $-\infty$ to ∞ , as is the case in standard linear regression), but are limited to a certain range of values. With the lasso and ridge regression estimators minimizing the same sum-of-squares, the key difference with the constrained estimation formulation of ridge regression is not in the explicit form of $c(\lambda_1)$ (and is set to some arbitrary convenient value in the remainder of this section) but in what is bounded by $c(\lambda_1)$ and the domain of acceptable values for β that it implies. For the lasso regression estimator the domain is specified by a bound on the ℓ_1 -norm of the regression parameter while for its ridge counterpart the bound is applied to the squared ℓ_2 -norm of β . The parameter constraints implied by the lasso and ridge norms result in balls in different norms:

$$\begin{aligned} \{\beta \in \mathbb{R}^p : |\beta_1| + |\beta_2| + \dots + |\beta_p| \leq c_1(\lambda_1)\}, \\ \{\beta \in \mathbb{R}^p : \beta_1^2 + \beta_2^2 + \dots + \beta_p^2 \leq c_2(\lambda_2)\}, \end{aligned}$$

respectively, and where $c(\cdot)$ is now equipped with a subscript referring to the norm to stress that it is different for lasso and ridge. The left-hand panel of Figure 6.3 visualizes these parameter constraints for $p = 2$ and $c_1(\lambda_1) = 2 = c_2(\lambda_2)$. In the Euclidean space ridge yields a spherical constraint for β , while a diamond-like shape for the lasso. The lasso regression estimator is then that β inside this diamond domain that yields the smallest sum-of-squares (as is visualized by right-hand panel of Figure 6.3).

The right-hand panel of Figure 6.3) illustrates Lemma 6.1. The lasso solution is non-unique if the levels sets of the sum-of-squares are parallel to one of the sides of the diamond parameter constraint. This occurs with probability zero.

The selection property of the lasso is due to the fact that the diamond-shaped parameter constraint has its corners falling on the axes. For a point to lie on an axis, one coordinate needs to equal zero. The lasso regression

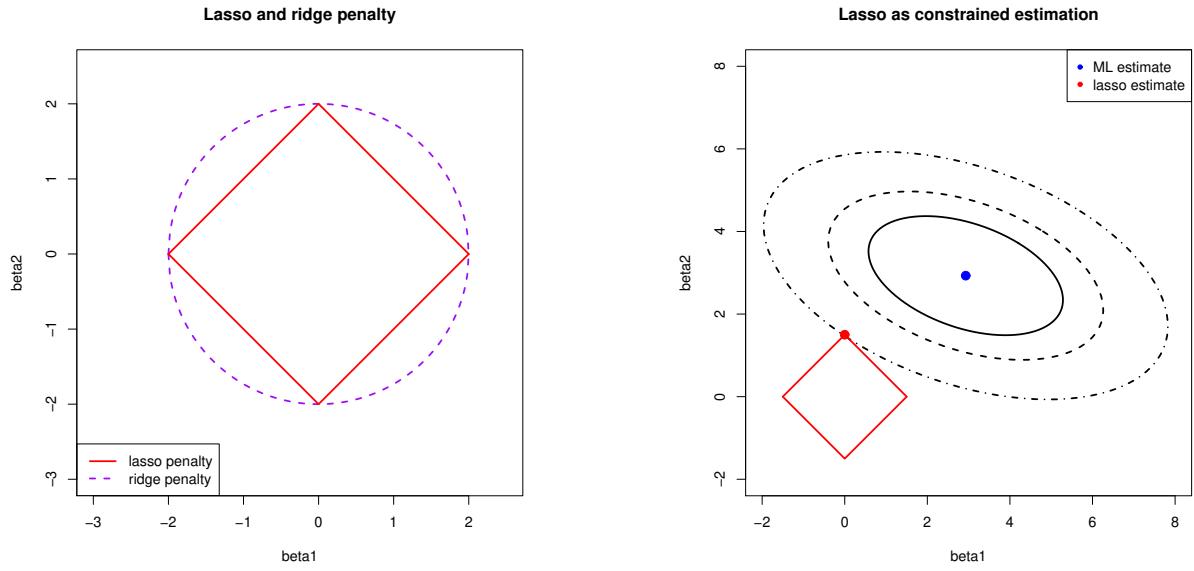


Figure 6.3: Left panel: The lasso parameter constraint ($|\beta_1| + |\beta_2| \leq 2$) and its ridge counterpart ($\beta_1^2 + \beta_2^2 \leq 2$). Solution path of the ridge estimator and its variance. Right panel: the lasso regression estimator as a constrained least squares estimator.

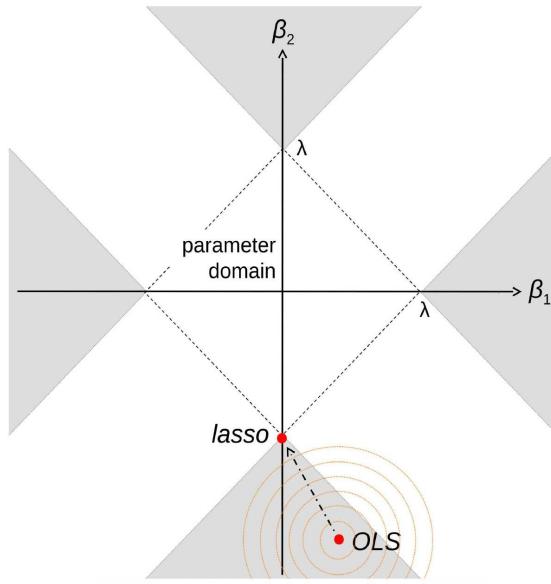


Figure 6.4: Shrinkage with the lasso. The range of possible lasso estimates is demarcated by the diamond around the origin. The grey areas contain all points that are closest to one of the diamond's corners than to any other point inside the diamond. If the maximum likelihood estimate falls inside any of these grey areas, the lasso shrinks it to the closest diamond tip (which corresponds to a sparse solution). For example, let the red dot in the fourth quadrant be a maximum likelihood estimate. It is in a grey area. Hence, its lasso estimate is the red dot at the lowest tip of the diamond.

estimator coincides with the point inside the diamond closest to the maximum likelihood estimator. This point may correspond to a corner of the diamond, in which case one of the coordinates (regression parameters) equals zero and, consequently, the lasso regression estimator does not select this element of β . Figure 6.4 illustrates the selection property for the case with $p = 2$ and an orthonormal design matrix. An orthonormal design matrix yields level sets (orange dotted circles in Figure 6.4) of the sum-of-squares that are spherical and centered around

the maximum likelihood estimate (red dot in Figure 6.4). For maximum likelihood estimates inside the grey areas the closest point in the diamond-shaped parameter domain will be on one of its corners. Hence, for these maximum likelihood estimates the corresponding lasso regression estimate will include on a single covariate in the model. The geometrical explanation of the selection property of the lasso regression estimator also applies to non-orthonormal design matrices and in dimensions larger than two. In particular, high-dimensionally, the sum-of-squares may be a degenerated ellipsoid, that can and will still hit a corner of the diamond-shaped parameter domain. Finally, note that a zero value of the lasso regression estimate does imply neither that the parameter is indeed zero nor that it will be statistically indistinguishable from zero.

Larger values of the lasso penalty parameter λ_1 induce tighter parameter constraints. Consequently, the number of zero elements in the lasso regression estimator of β increases as λ_1 increases. However, where $\|\hat{\beta}(\lambda_1)\|_1$ decreases monotonically as λ_1 increases (left panel of Figure 6.5 for an example and Exercise 6.13), the number of non-zero coefficients does not. Locally, at some finite λ_1 , the number of non-zero elements in $\hat{\beta}(\lambda_1)$ may temporarily increase with λ_1 , to only go down again as λ_1 is sufficiently increased (as in the $\lambda_1 \rightarrow \infty$ limit the number of non-zero elements is zero, see the argumentation at the end of Section 6.2). The right panel of Figure 6.5 illustrates this behavior for an arbitrary data set.

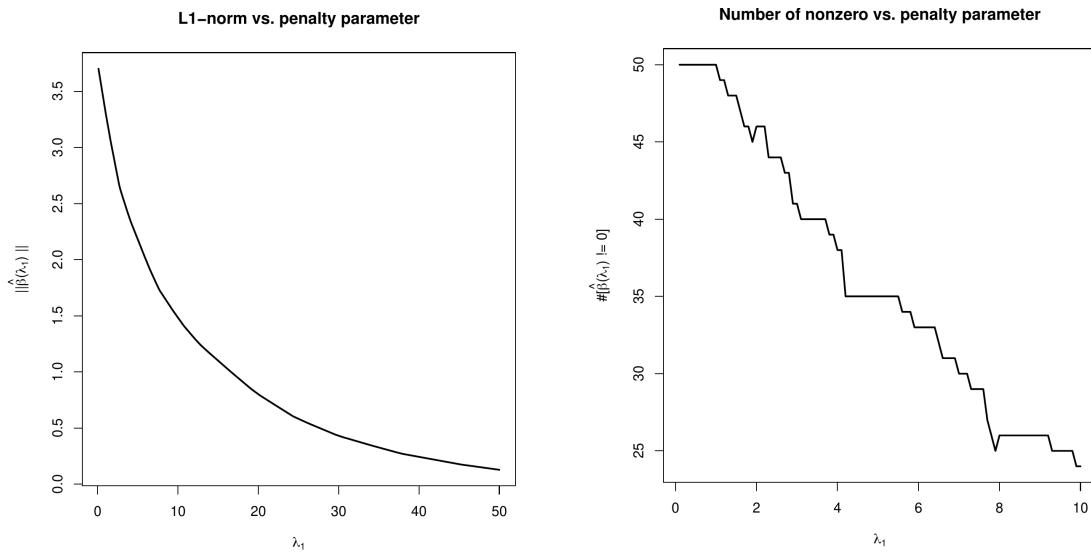


Figure 6.5: Contour plots of the sum-of-squares and the lasso regression loss (left and right panel, respectively). The dotted grey line represent level sets. The red line and dot represent the the location of minimum in both panels.

The attractiveness of the lasso regression estimator is in its simultaneous estimation and selection of parameters. For large enough values of the penalty parameter λ_1 the estimated regression model comprises only a subset of the supplied covariates. In high-dimensions (demanding a large penalty parameter) the number of selected parameters by the lasso regression estimator is usually small (relative to the total number of parameters), thus producing a so-called sparse model. Would one adhere to the parsimony principle, such a sparse and thus simpler model is preferable over a full model. Simpler may be better, but too simple is worse. The phenomenon or system that is to be described by the model need not be sparse. For instance, in molecular biology the regulatory network of the cell is no longer believed to be sparse (Boyle *et al.*, 2017). Similarly, when analyzing brain image data, the connectivity of the brain is not believed to be sparse.

6.3.1 Maximum number of selected covariates

The number of parameter/covariates selected by the lasso regression estimator is bounded non-trivially. The cardinality (i.e. the number of included covariates) of every lasso estimated linear regression model is smaller than or equal to $\min\{n, p\}$ (Bühlmann and Van De Geer, 2011). As Bühlmann and Van De Geer (2011) point out this is obvious from the analysis of the LARS algorithm of Efron *et al.* (2004) (which is to be discussed in Section ??). For now we just provide an R-script that generates the regularization paths using the `lars`-package for the `diabetes` data included in the package for a random number of samples n not exceeding the number of covariates p .

Listing 6.2 R code

```
# activate library
library(lars)

# load data
data(diabetes)
X <- diabetes$x
Y <- diabetes$y

# set sample size
n <- sample(1:ncol(X), 1)
id <- sample(1:length(Y), n)

# plot regularization paths
plot(lars(X[id,], Y[id], intercept=FALSE))
```

Irrespective of the drawn sample size n the plotted regularization paths all terminate before the $n + 1$ -th variate enters the model. This could of course be circumstantial evidence at best, or even be labelled a bug in the software.

But even without the LARS algorithm the nontrivial part of the inequality, that the number of selected variates p does not exceed the sample size n , can be proven (Osborne *et al.*, 2000).

Theorem 6.2 (Theorem 6, Osborne *et al.*, 2000)

If $p > n$ and $\hat{\beta}(\lambda_1)$ is a minimizer of the lasso regression loss function (6.2), then $\hat{\beta}(\lambda_1)$ has at most n non-zero entries.

Proof. Confer Osborne *et al.* (2000). ■

In the high-dimensional setting, when p is large compared to n small, this implies a considerable dimension reduction. It is, however, somewhat unsatisfactory that it is the study design, i.e. the inclusion of the number of samples, that determines the upperbound of the model size.

6.4 Estimation

In the absence of an analytic expression for the optimum of the lasso loss function (6.2), much attention is devoted to numerical procedures to find it.

6.4.1 Quadratic programming

In the original lasso paper Tibshirani (1996) reformulates the lasso optimization problem to a quadratic program. A quadratic problem optimizes a quadratic form subject to linear constraints. This is a well-studied optimization problem for which many readily available implementations exist (e.g., the quadprog-package in R). The quadratic program that is equivalent to the lasso regression problem, which minimizes the least squares criterion, $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$ subject to the constraint $\beta \in \mathbb{R}^p$ such that $\|\beta\|_1 < c(\lambda_1)$, is:

$$\min_{\{\beta \in \mathbb{R}^p : \mathbf{R}\beta \geq \mathbf{0}_q\}} \frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta), \quad (6.3)$$

where \mathbf{R} is a suitably chosen $q \times p$ dimensional linear constraint matrix that specifies the linear constraints on the parameter β . For $p = 2$ the domain implied by lasso parameter constraint $\{\beta \in \mathbb{R}^2 : \|\beta\|_1 < c(\lambda_1)\mathbf{1}_4\}$ is equal to:

$$\begin{aligned} \{\beta \in \mathbb{R}^2 : \beta_1 + \beta_2 \leq c(\lambda_1)\} \cap \{\beta \in \mathbb{R}^2 : \beta_1 - \beta_2 \geq -c(\lambda_1)\} \cap \{\beta \in \mathbb{R}^2 : \beta_1 - \beta_2 \leq c(\lambda_1)\} \\ \cap \{\beta \in \mathbb{R}^2 : \beta_1 + \beta_2 \geq -c(\lambda_1)\}. \end{aligned}$$

This collection of linear parameter constraints can be reformulated, using:

$$\mathbf{R} = \begin{pmatrix} 1 & 1 \\ -1 & -1 \\ 1 & -1 \\ -1 & 1 \end{pmatrix},$$

into $\{\beta \in \mathbb{R}^2 : \mathbf{R}\beta \geq -c(\lambda_1)\mathbf{1}_4\}$.

To solve the quadratic program (6.3) it is usually reformulated in terms of its dual. Hereto we introduce the Lagrangian:

$$L(\beta, \nu) = \frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) - \nu^T[\mathbf{R}\beta + c(\lambda_1)\mathbf{1}_q], \quad (6.4)$$

where $\nu = (\nu_1, \dots, \nu_q)^T$ is the vector of non-negative multipliers. The dual function is now defined as $\inf_{\beta} L(\beta, \nu)$. This infimum is attained at:

$$\beta^* = (\mathbf{X}^T \mathbf{X})^{-1}(\mathbf{X}^T \mathbf{Y} + \mathbf{R}^T \nu), \quad (6.5)$$

which can be verified by equating the first order partial derivative with respect to β of the Lagrangian to zero and solving for β . Substitution of $\beta = \beta^*$ into the dual function gives, after changing the minus sign:

$$\frac{1}{2}\nu^T \mathbf{R}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{R}^T \nu + \nu^T [\mathbf{R}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} + c(\lambda_1)\mathbf{1}_q] - \frac{1}{2}\mathbf{Y}^T [\mathbf{I}_{nn} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] \mathbf{Y}.$$

The dual problem minimizes this expression (from which the last term is dropped as it does not involve ν) with respect to ν , subject to $\nu \geq 0$. Although also a quadratic programming problem, the dual problem *i*) has a simpler formulation of the linear constraints and *ii*) is defined on a lower dimensional space than the primal problem (should the number of columns of \mathbf{R} exceeds its number of rows). If $\hat{\nu}$ is the solution of the dual problem, the solution of the primal problem is obtained from Equation (6.5). Refer to, e.g., Bertsekas (2014) for more on quadratic programming.

Example 6.5 (Orthogonal design matrix, continued)

The evaluation of the lasso regression estimator by means of quadratic programming is illustrated using the data from the numerical Example 6.5. The R-script below solves, the implementation of the quadprog-package, the quadratic program associated with the lasso regression problem of the aforementioned example.

Listing 6.3 R code

```
# load library
library(quadprog)

# data
Y <- matrix(c(-4.9, -0.8, -8.9, 4.9, 1.1, -2.0), ncol=1)
X <- t(matrix(c(1, -1, 3, -3, 1, 1, -3, -3, -1, 0, 3, 0), nrow=2, byrow=TRUE))

# constraint radius
L1norm <- 1.881818 + 0.8678572

# solve the quadratic program
solve.QP(t(X) %*% X, t(X) %*% Y,
         t(matrix(c(1, 1, -1, -1, 1, -1, -1, 1), ncol=2, byrow=TRUE)),
         L1norm*c(-1, -1, -1, -1))$solution
```

The resulting estimates coincide with those found earlier. □

For relatively small p quadratic programming is a viable option to find the lasso regression estimator. For large p it is practically not feasible. Above the linear constraint matrix \mathbf{R} is 4×2 dimensional for $p = 2$. When $p = 3$, it requires a linear constraint matrix \mathbf{R} with six rows corresponding to the six side of a 3-dimensional cube. In general, $2p$ linear constraints are required to fully specify the parameter constraint of the lasso regression estimator. If p large, the specification of only the linear constraint matrix \mathbf{R} will take endlessly, as it is a $2p \times p$ matrix, leave alone solving the corresponding quadratic program.

6.4.2 Iterative ridge

Why develop something new, when one can also make do with existing tools? The loss function of the lasso regression estimator can be optimized by iterative application of ridge regression (as pointed out in Fan and Li, 2001). It requires an approximation of the lasso penalty, or the absolute value function. Set $p = 1$ and let β_0 be an initial parameter value for β around which the absolute value function $|\beta|$ is to be approximated. Its quadratic approximation then is:

$$|\beta| \approx |\beta_0| + \frac{1}{2}|\beta_0|^{-1}(\beta^2 - \beta_0^2).$$

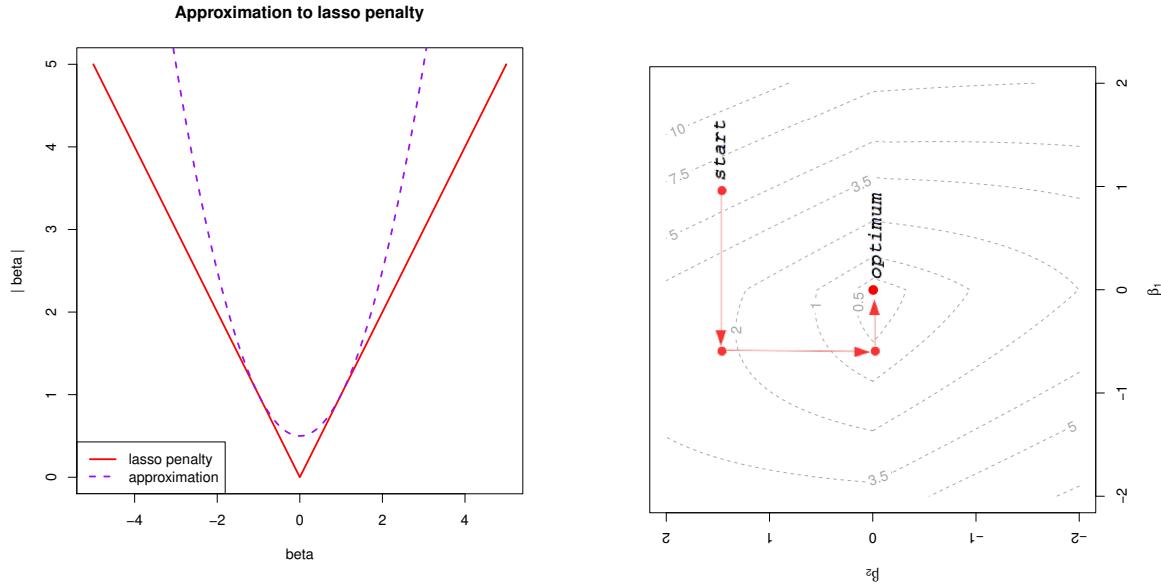


Figure 6.6: Left panel: quadratic approximation (i.e. the ridge penalty) to the absolute value function (i.e. the lasso penalty). Right panel: Illustration of the coordinate descent algorithm. The dashed grey lines are the level sets of the lasso regression loss function. The red arrows depict the parameter updates. These arrows are parallel to either the β_1 or the β_2 parameter axis, thus indicating that the regression parameter β is updated coordinate-wise.

An illustration of this approximation is provided in the left panel of Figure 6.6.

The lasso regression estimator is evaluated through iterative application of the ridge regression estimator. This iterative procedure needs initiation by some guess $\beta^{(0)}$ for β . For example, the ridge estimator itself may serve as such. Then, at the $k + 1$ -th iteration an update $\beta^{(k+1)}$ of the lasso regression estimator of β is to be found. Application of the quadratic approximation to the absolute value functions of the elements of β (around the k -th update $\beta^{(k)}$) in the lasso penalty yields an approximation to the lasso regression loss function:

$$\begin{aligned} \|\mathbf{Y} - \mathbf{X}\beta^{(k+1)}\|_2^2 + \lambda_1\|\beta^{(k+1)}\|_1 &\approx \|\mathbf{Y} - \mathbf{X}\beta^{(k+1)}\|_2^2 + \lambda_1\|\beta^{(k)}\|_1 \\ &\quad + \frac{1}{2}\lambda_1 \sum_{j=1}^p |\beta_j^{(k)}|^{-1} [\beta_j^{(k+1)}]^2 - \frac{1}{2}\lambda_1 \sum_{j=1}^p |\beta_j^{(k)}|^{-1} [\beta_j^{(k)}]^2 \\ &\propto \|\mathbf{Y} - \mathbf{X}\beta^{(k+1)}\|_2^2 + \frac{1}{2}\lambda_1 \sum_{j=1}^p |\beta_j^{(k)}|^{-1} [\beta_j^{(k+1)}]^2. \end{aligned}$$

The loss function now contains a weighted ridge penalty. In this one recognizes a generalized ridge regression loss function (see Chapter 3). As its minimizer is known, the approximated lasso regression loss function is optimized by:

$$\beta^{(k+1)}(\lambda_1) = \{\mathbf{X}^\top \mathbf{X} + \lambda_1 \Psi[\beta^{(k)}(\lambda_1)]\}^{-1} \mathbf{X}^\top \mathbf{Y}$$

where

$$\text{diag}\{\Psi[\beta^{(k)}(\lambda_1)]\} = (1/|\beta_1^{(k)}|, 1/|\beta_2^{(k)}|, \dots, 1/|\beta_p^{(k)}|).$$

The thus generated sequence of updates $\{\beta^{(k)}(\lambda_1)\}_{k=0}^\infty$ converges (under ‘nice’ conditions) to the lasso regression estimator $\hat{\beta}(\lambda_1)$.

A note of caution. The in-built variable selection property of the lasso regression estimator may – for large enough choices of the penalty parameter λ_1 – cause elements of $\beta^{(k)}(\lambda_1)$ to become arbitrary close to zero (or, in R exceed machine precision and thereby being effectively zero) after enough updates. Consequently, the ridge penalty parameter for the j -th element of regression parameter may approach infinity, as the j -th element of $\Psi[\beta^{(k)}(\lambda_1)]$ equals $|\beta_j^{(k)}|^{-1}$. To accommodate this, the iterative ridge regression algorithm for the evaluation of the lasso regression estimator requires a modification. Effectively, that amounts to the removal of j -th covariate from the model all together (for its estimated regression coefficient is indistinguishable from zero). After its removal, it does not return to the set of covariates. This may be problematic if two covariates are (close to) super-collinear.

6.4.3 Gradient ascent

Another method of finding the lasso regression estimator and implemented in the `penalized`-package (Goeman, 2010) makes use of gradient ascent. Gradient ascent/descent is an maximization/minimization method that finds the optimum of a smooth function by iteratively updating a first-order local approximation to this function. Gradient ascents runs through the following sequence of steps repetitively until convergence:

- Choose a starting value.
- Calculate the derivative of the function, and determine the direction in which the function increases most. This direction is the path of steepest ascent.
- Proceed in this direction, until the function no longer increases.
- Recalculate at this point the gradient to determine a new path of steepest ascent.
- Repeat the above until the (region around the) optimum is found.

The procedure above is illustrated in Figure 6.7. The top panel shows the choice of the initial value. From this point the path of the steepest ascent is followed until the function no longer increases (right panel of Figure 6.7). Here the path of steepest ascent is updated along which the search for the optimum is proceeded (bottom panel of Figure 6.7).

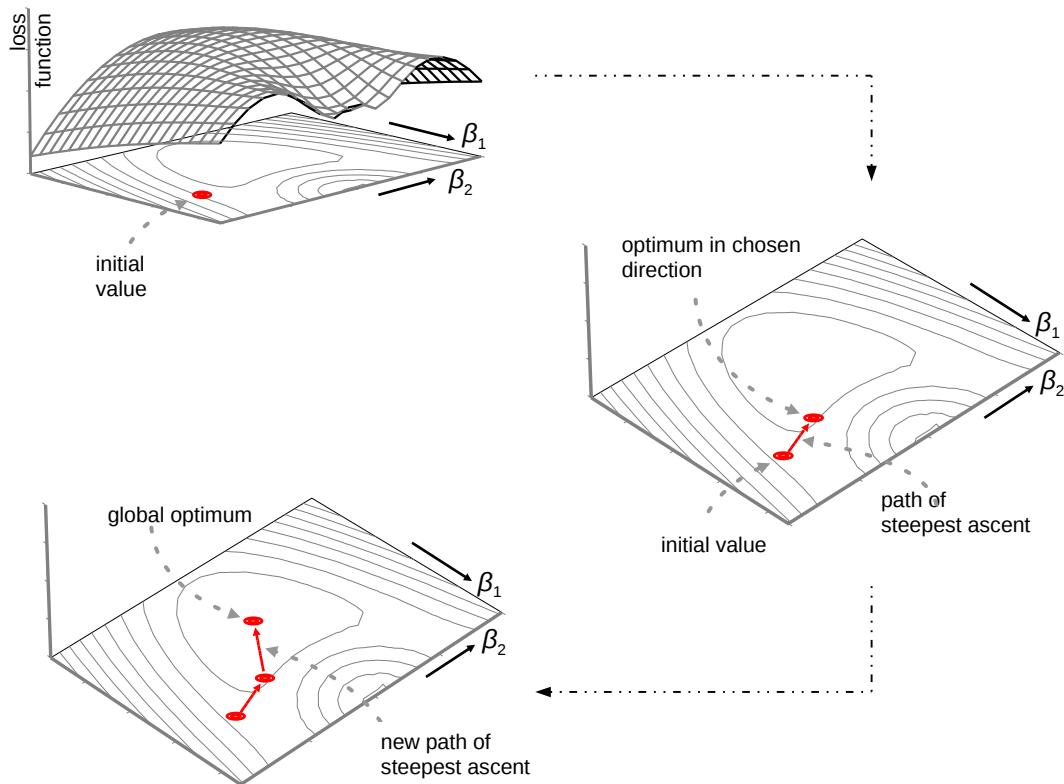


Figure 6.7: Illustration of the gradient ascent procedure.

The application of gradient ascent to find the lasso regression estimator is frustrated by the non-differentiability (with respect to any of the regression parameters) of the lasso penalty function at zero. In Goeman (2010) this is overcome by the use of a generalized derivative. Define the *directional* or *Gâteaux* derivative of the function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ at $\mathbf{x} \in \mathbb{R}^p$ in the direction of $\mathbf{v} \in \mathbb{R}^p$ as:

$$f'(\mathbf{x}) = \lim_{\tau \downarrow 0} \tau^{-1} [f(\mathbf{x} + \tau \mathbf{v}) - f(\mathbf{x})],$$

assuming this limit exists. The Gâteaux derivative thus gives the infinitesimal change in f at \mathbf{x} in the direction of \mathbf{v} . As such $f'(\mathbf{x})$ is a scalar (as is immediate from the definition after noting that $f(\cdot) \in \mathbb{R}$) and should not be confused with the gradient (the vector of partial derivatives). Furthermore, at each point \mathbf{x} there are infinitely many Gâteaux differentials (as there are infinitely many choices for $\mathbf{v} \in \mathbb{R}^p$). In the particular case when $\mathbf{v} = \mathbf{e}_j$, \mathbf{e}_j the unit vector along the axis of the j -th coordinate, the directional derivative coincides with the partial derivative

of f in the direction of x_j . Relevant for the case at hand is the absolute value function $f(x) = |x|$ with $x \in \mathbb{R}$. Evaluation of the limits in its Gâteaux derivative yields:

$$f'(x) = \begin{cases} v \frac{x}{|x|} & \text{if } x \neq 0, \\ v & \text{if } x = 0, \end{cases}$$

for any $v \in \mathbb{R} \setminus \{0\}$. Hence, the Gâteaux derivative of $|x|$ does exist at $x = 0$. In general, the Gâteaux differential may be uniquely defined by limiting the directional vectors v to *i*) those with unit length (i.e. $\|v\| = 1$) and *ii*) the direction of steepest ascent. Using the Gâteaux derivative a gradient of $f(\cdot)$ at $\mathbf{x} \in \mathbb{R}^p$ may then be defined as:

$$\nabla f(\mathbf{x}) = \begin{cases} f'(\mathbf{x}) \circ \mathbf{v}_{\text{opt}} & \text{if } f'(\mathbf{x}) \geq 0, \\ \mathbf{0}_p & \text{if } f'(\mathbf{x}) < 0, \end{cases} \quad (6.6)$$

in which \circ is the Hadamard (i.e., element-wise) product and $\mathbf{v}_{\text{opt}} = \arg \max_{\{v : \|v\|=1\}} f'(v)$. This is the direction of steepest ascent, \mathbf{v}_{opt} , scaled by Gâteaux derivative, $f'(\mathbf{x})$, in the direction of \mathbf{v}_{opt} .

Goeman (2010) applies the definition of the Gâteaux gradient to the lasso penalized likelihood (6.2) using the direction of steepest ascent as \mathbf{v}_{opt} . The resulting partial Gâteaux derivative with respect to the j -th element of the regression parameter β is:

$$\frac{\partial}{\partial \beta_j} \mathcal{L}_{\text{lasso}}(\mathbf{Y}, \mathbf{X}; \beta) = \begin{cases} \frac{\partial}{\partial \beta_j} \mathcal{L}(\mathbf{Y}, \mathbf{X}; \beta) - \lambda_1 \text{sign}(\beta_j) & \text{if } \beta_j \neq 0 \\ \frac{\partial}{\partial \beta_j} \mathcal{L}(\mathbf{Y}, \mathbf{X}; \beta) - \lambda_1 \text{sign}\left[\frac{\partial}{\partial \beta_j} \mathcal{L}(\mathbf{Y}, \mathbf{X}; \beta)\right] & \text{if } \beta_j = 0 \text{ and } |\partial \mathcal{L} / \partial \beta_j| > \lambda_1 \\ 0 & \text{otherwise} \end{cases},$$

where $\partial \mathcal{L} / \partial \beta_j = \sum_{j'=1}^p (\mathbf{X}^\top \mathbf{X})_{j',j} \beta_j - (\mathbf{X}^\top \mathbf{Y})_j$. This can be understood through a case-by-case study. The partial derivative above is assumed to be clear for the $\beta_j \neq 0$ and the ‘otherwise’ cases. That leaves the clarification of the middle case. When $\beta_j = 0$, the direction of steepest ascent of the penalized loglikelihood points either into $\{\beta \in \mathbb{R}^p : \beta_j > 0\}$, or $\{\beta \in \mathbb{R}^p : \beta_j < 0\}$, or stays in $\{\beta \in \mathbb{R}^p : \beta_j = 0\}$. When the direction of steepest ascent points into the positive or negative half-hyperplanes, the contribution of $\lambda_1 |\beta_j|$ to the partial Gâteaux derivative is simply λ_1 or $-\lambda_1$, respectively. Then, only when the partial derivative of the log-likelihood together with this contribution is larger than zero, the penalized loglikelihood improves and the direction is of steepest ascent. Similarly, the direction of steepest ascent may be restricted to $\{\beta \in \mathbb{R}^p | \beta_j = 0\}$ and the contribution of $\lambda_1 |\beta_j|$ to the partial Gâteaux derivative vanishes. Then, only if the partial derivative of the loglikelihood is positive, this direction is to be pursued for the improvement of the penalized loglikelihood.

Convergence of gradient ascent can be slow close to the optimum. This is due to its linear approximation of the function. Close to the optimum the linear term of the Taylor expansion vanishes and is dominated by the second-order quadratic term. To speed-up convergence close to the optimum the gradient ascent implementation offered by the penalized-package switches to a Newton-Raphson procedure.

6.4.4 Coordinate descent

Coordinate descent is another optimization algorithm that may be used to evaluate the lasso regression estimator numerically, as is done by the implementation offered via the `glmnet`-package. Coordinate descent, instead of following the gradient of steepest descent (as in Section 6.4.3), minimizes the loss function along the coordinates one-at-the-time. For the j -th regression parameter this amounts to finding:

$$\begin{aligned} \arg \min_{\beta_j} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 &= \arg \min_{\beta_j} \|\mathbf{Y} - \mathbf{X}_{*, \setminus j} \beta_{\setminus j} - \mathbf{X}_{*,j} \beta_j\|_2^2 + \lambda_1 |\beta_j|_1 \\ &= \arg \min_{\beta_j} \|\tilde{\mathbf{Y}} - \mathbf{X}_{*,j} \beta_j\|_2^2 + \lambda_1 |\beta_j|_1, \end{aligned}$$

where $\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}_{*, \setminus j} \beta_{\setminus j}$. After a simple rescaling of both $\mathbf{X}_{*,j}$ and β_j , the minimization of the lasso regression loss function with respect to β_j is equivalent to one with an orthonormal design matrix. From Example 6.4 it is known that the minimizer is obtained by application of the soft-threshold function to the corresponding maximum likelihood estimator (now derived from $\tilde{\mathbf{Y}}$ and $\mathbf{X}_{*,j}$). The coordinate descent algorithm iteratively runs over the p elements until convergence. The right panel of Figure 6.6 provides an illustration of the coordinate descent algorithm.

Convergence of the coordinate descent algorithm to the minimum of the lasso regression loss function (6.2) is warranted by the convexity of this function. At each minimization step the coordinate descent algorithm yields an

update of the parameter estimate that corresponds to an equal or smaller value of the loss function. It, together with the compactness of diamond-shaped parameter domain and the boundedness (from below) of the lasso regression loss function, implies that the coordinate descent algorithm converges to the minimum of this lasso regression loss function. Although convergence is assured, it may take many steps for it to be reached. In particular, when *i*) two covariates are strongly collinear, *ii*) one of the two covariates contributes only slightly more to the response, and *iii*) the algorithm is initiated with the weaker explanatory covariate. The coordinate descent algorithm will then take many iterations to replace the latter covariate by the preferred one. In such cases simultaneous updating, as is done by the gradient ascent algorithm (Section 6.4.3), may be preferable.

6.5 Moments

In general the moments of the lasso regression estimator appear to be unknown. In certain cases an approximation can be given. This is pointed out here. Use the quadratic approximation to the absolute value function of Section 6.4.2 and approximate the lasso regression loss function around the lasso regression estimate:

$$\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 \approx \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \frac{1}{2} \lambda_1 \sum_{j=1}^p |\hat{\beta}(\lambda_1)|^{-1} \beta_j^2.$$

Optimization of the right-hand side of the preceding display with respect to β gives a ‘ridge approximation’ to the lasso estimator:

$$\hat{\beta}(\lambda_1) \approx \{\mathbf{X}^\top \mathbf{X} + \lambda_1 \Psi[\hat{\beta}(\lambda_1)]\}^{-1} \mathbf{X}^\top \mathbf{Y},$$

with $(\Psi[\hat{\beta}(\lambda_1)])_{jj} = |\hat{\beta}_j(\lambda_1)|^{-1}$ if $\hat{\beta}_j(\lambda_1) \neq 0$. Now use this ‘ridge approximation’ to obtain the approximation to the moments of the lasso regression estimator:

$$\mathbb{E}[\hat{\beta}(\lambda_1)] \approx \mathbb{E}(\{\mathbf{X}^\top \mathbf{X} + \lambda_1 \Psi[\hat{\beta}(\lambda_1)]\}^{-1} \mathbf{X}^\top \mathbf{Y}) = \{\mathbf{X}^\top \mathbf{X} + \lambda_1 \Psi[\hat{\beta}(\lambda_1)]\}^{-1} \mathbf{X}^\top \mathbf{X} \beta$$

and

$$\begin{aligned} \text{Var}[\hat{\beta}(\lambda_1)] &\approx \text{Var}(\{\mathbf{X}^\top \mathbf{X} + \lambda_1 \Psi[\hat{\beta}(\lambda_1)]\}^{-1} \mathbf{X}^\top \mathbf{Y}) \\ &= \sigma^2 \{\mathbf{X}^\top \mathbf{X} + \lambda_1 \Psi[\hat{\beta}(\lambda_1)]\}^{-1} \mathbf{X}^\top \mathbf{X} \{\mathbf{X}^\top \mathbf{X} + \lambda_1 \Psi[\hat{\beta}(\lambda_1)]\}^{-1}. \end{aligned}$$

These approximations can only be used if the lasso regression estimate is not sparse, which is at odds with its attractiveness. A better approximation of the variance of the lasso regression estimator can be found in Osborne *et al.* (2000), but even this becomes poor when many elements of β are estimated as zero.

Although the above approximations are only crude, they indicate that the moments of the lasso regression estimator exhibit similar behaviour as those of its ridge counterpart. The (approximation of the) mean $\mathbb{E}[\hat{\beta}(\lambda_1)]$ tends to zero as $\lambda_1 \rightarrow \infty$. This was intuitively already expected from the form of the lasso regression loss function (6.2), in which the penalty term dominates for large λ_1 and is minimized for $\hat{\beta}(\lambda_1) = \mathbf{0}_p$. This may also be understood geometrically when appealing to the equivalent constrained estimation formulation of the lasso regression estimator. The parameter constraint shrinks to zero with increasing λ_1 . Hence, so must the estimator. Similarly, the (approximation of the) variance of the lasso regression estimator vanishes as the penalty parameter λ_1 grows. Again, its loss function (6.2) provides the intuition: for large λ_1 the penalty term, which does not depend on data, dominates. Or, from the perspective of the constrained estimation formulation, the parameter constraint shrinks to zero as $\lambda_1 \rightarrow \infty$. Hence, so must the variance of the estimator as less and less room is left for it to fluctuate.

The behaviour of the mean squared error, bias squared plus variance, of the lasso regression estimator in terms of λ_1 is hard to characterize exactly without knowledge of the quality of the approximations. In particular, does a λ_1 exist such that the MSE of the lasso regression estimator outperforms that of its maximum likelihood counterpart? Nonetheless, a first observation may be obtained from reasoning in extremis. Suppose $\beta = \mathbf{0}_p$, which corresponds to an empty or maximally sparse model. A large value of λ_1 then yields a zero estimate of the regression parameter: $\hat{\beta}(\lambda_1) = \mathbf{0}_p$. The bias squared is thus minimized as $\|\hat{\beta}(\lambda_1) - \beta\|_2^2 = 0$. With the bias vanished and the (approximation of the) variance decreasing in λ_1 , so must the MSE decrease for λ_1 larger than some value. So, for an empty model the lasso regression estimator with a sufficiently large penalty parameter yields a better MSE than the maximum likelihood estimator. For very sparse models this property may be expected to uphold, but for non-sparse models the bias squared will have a substantial contribution to the MSE, and it is thus not obvious whether a λ_1 exists that yields a favourable MSE for the lasso regression estimator. This is investigated *in silico* in Hansen (2015). The simulations presented there indicate that the MSE of the lasso regression estimator is particularly sensitive to the actual β . Moreover, for a large part of the parameter space $\beta \in \mathbb{R}^p$ the MSE of $\hat{\beta}(\lambda_1)$ is behind that of the maximum likelihood estimator.

6.6 Degrees of freedom

The degrees of freedom consumed by the lasso regression estimator can be estimated simply by its number of nonzero elements.

Theorem 6.3 (Zou *et al.*, 2007, Tibshirani and Taylor, 2012)

Let $\mathbf{X} \in \mathcal{M}^{n,p}$ any design matrix and $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$. Then,

$$\text{df}(\lambda_1) = \mathbb{E}(|\{j : [\hat{\boldsymbol{\beta}}(\lambda_1)]_j \neq 0\}|),$$

for $\lambda_1 > 0$. If $\text{rank}(\mathbf{X}) = p$, the lasso regression estimator's number of nonzero's is even a consistent estimator of the degrees of freedom.

Proof. The full proof is beyond the scope of these notes, and here we limit ourselves to show the unbiasedness of the degrees of freedom estimator for orthonormal \mathbf{X} . The proof makes use of Lemma 2 of Stein (1981), which states that if $Y \sim \mathcal{N}(\mu, \sigma^2)$ and $g(\cdot)$ is an absolute continuous function with $\mathbb{E}[g'(Y)] < \infty$, then $\mathbb{E}[g(Y)(Y - \mu)] = \sigma^2 \mathbb{E}[g'(Y)]$. Starting from the degrees of freedom definition introduced in Section 1.6, we then manipulate (with minor abuse of notation) as follows:

$$\begin{aligned} \text{df}(\lambda_1) &= \sum_{i=1}^n [\text{Var}(Y_i)]^{-1} \text{Cov}(\hat{Y}_i, Y_i) \\ &= \sum_{i=1}^n \sigma^{-2} \mathbb{E}\{[\hat{Y}_i - \mathbb{E}(\hat{Y}_i)][Y_i - \mathbb{E}(Y_i)]\} \\ &= \sum_{i=1}^n \sigma^{-2} \mathbb{E}\{\mathbf{X}_{i,*} \hat{\boldsymbol{\beta}}(\lambda_1)[Y_i - \mathbb{E}(Y_i)]\} \\ &= \sum_{i=1}^n \sigma^{-2} \sum_{j=1}^p X_{i,j} \mathbb{E}\{[\hat{\boldsymbol{\beta}}(\lambda_1)]_j [Y_i - \mathbb{E}(Y_i)]\} \\ &= \sum_{i=1}^n \sigma^{-2} \sum_{j=1}^p X_{i,j} \mathbb{E}\{\text{sign}(\hat{\beta}_j^{\text{ml}})(|\hat{\beta}_j^{\text{ml}}| - \frac{1}{2}\lambda_1)_+ [Y_i - \mathbb{E}(Y_i)]\} \\ &= \sum_{i=1}^n \sum_{j=1}^p X_{i,j} \mathbb{E}\left\{ \frac{d}{dY_i} \text{sign}\left(\sum_{i'=1}^n X_{i',j} Y_{i'} \right) \left(\left| \sum_{i'=1}^n X_{i',j} Y_{i'} \right| - \frac{1}{2}\lambda_1 \right)_+ \right\} \\ &= \sum_{i=1}^n \sum_{j=1}^p X_{i,j}^2 \mathbb{E}\left[\mathbb{1}_{\{|\sum_{i'=1}^n X_{i',j} Y_{i'}| > \frac{1}{2}\lambda_1\}} \right] \\ &= \mathbb{E}\left[\sum_{j=1}^p \mathbb{1}_{\{|\sum_{i=1}^n X_{i,j} Y_i| > \frac{1}{2}\lambda_1\}} \right], \end{aligned}$$

where we have used the definition of the covariance, the analytic expression of the lasso regression estimator in the orthonormal case, and the independence among the samples. ■

Would one select the lasso regression parameter's penalty parameter on the basis of an information criterion, this simple unbiased estimator of its degrees of freedom is rather convenient.

6.7 The Bayesian connection

The lasso regression estimator, being a penalized estimator, knows a Bayesian formulation, much like the (generalized) ridge regression estimator could be viewed as a Bayesian estimator if a Gaussian prior is imposed on the regression parameter (cf. Chapter 2 and Section 3.2). Instead of normal prior, the lasso regression estimator requires (as suggested by the form of the lasso penalty) a zero-centered Laplacian (or double exponential) prior for it to be viewed as a Bayesian estimator. A zero-centered Laplace distributed random variable X has density $f_X(x) = \frac{1}{2}b^{-1} \exp(-|x|/b)$ with scale parameter $b > 0$. The top panel of Figure 6.8 shows the Laplace prior, and for contrast the normal prior of the ridge regression estimator. This figure reveals that the 'lasso prior' puts more mass close to zero and in the tails than the Gaussian 'ridge prior'. This corroborates with the tendency of the lasso regression estimator to produce either zero or large (compared to ridge) estimates.

The lasso regression estimator corresponds to the maximum a posteriori (MAP) estimator of $\boldsymbol{\beta}$, when the prior is a Laplace distribution. The posterior distribution is then proportional to:

$$\prod_{i=1}^n (2\pi\sigma^2)^{-1/2} \exp[-(2\sigma^2)^{-1}(Y_i - \mathbf{X}_{i,*}\boldsymbol{\beta})^2] \times \prod_{j=1}^p (2b)^{-1} \exp(-|\beta_j|/b).$$

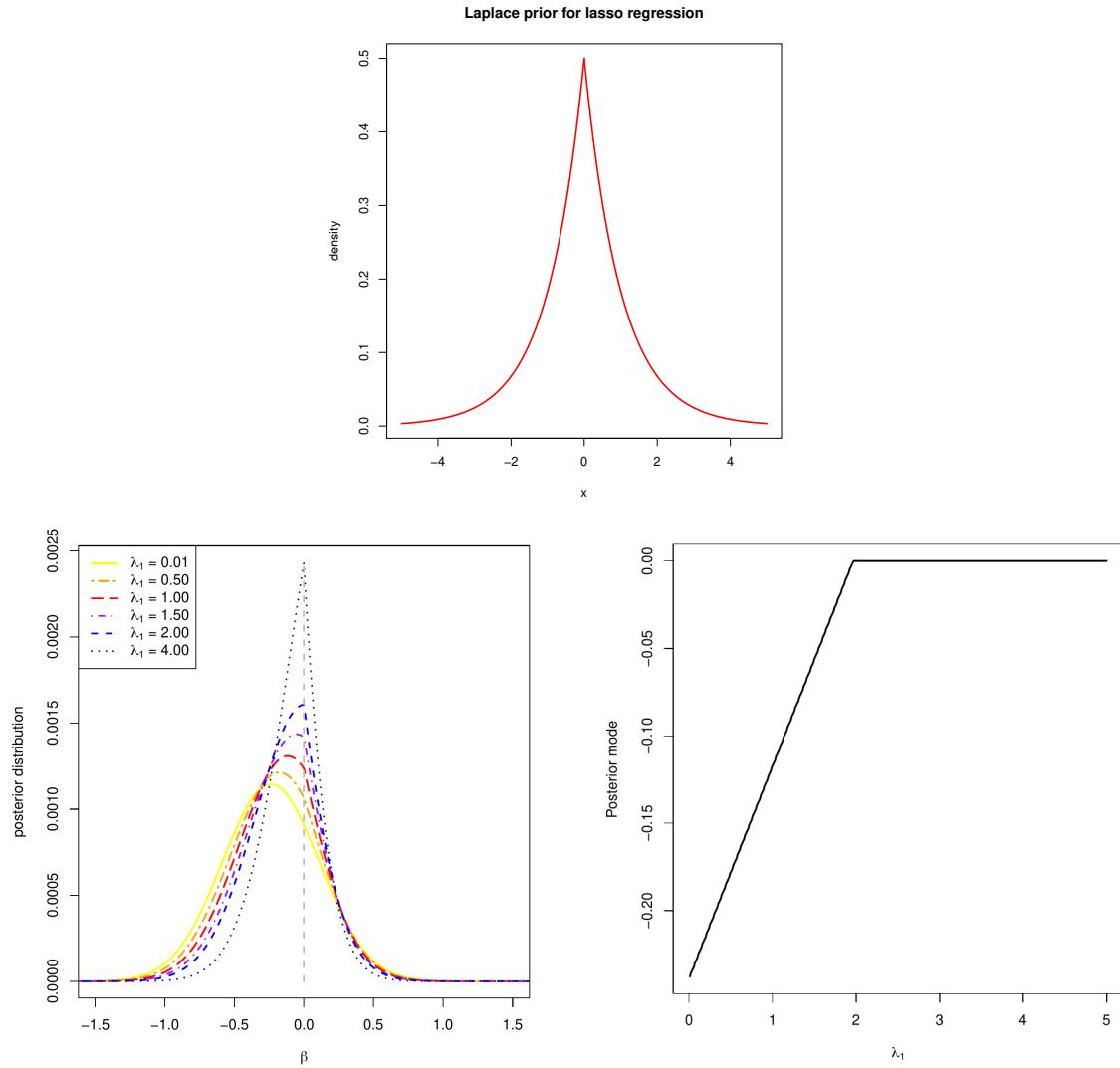


Figure 6.8: Top panel: the Laplace prior associated with the Bayesian counterpart of the lasso regression estimator. Bottom left panel: the posterior distribution of the regression parameter for various Laplace priors. Bottom right panel: posterior mode vs. the penalty parameter λ_1 .

The posterior is not a well-known and characterized distribution. This is not necessary as interest concentrates here on its maximum. The location of the posterior mode coincides with the location of the maximum of logarithm of the posterior. The log-posterior is proportional to: $-(2\sigma^2)^{-1}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 - b^{-1}\|\beta\|_1$, with its maximizer minimizing $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + (2\sigma^2/b)\|\beta\|_1$. In this one recognizes the form of the lasso regression loss function (6.2). It is thus clear that the scale parameter of the Laplace distribution reciprocally relates to lasso penalty parameter λ_1 , similar to the relation of the ridge penalty parameter λ_2 and the variance of the Gaussian prior of the ridge regression estimator.

The posterior may not be a standard distribution, in the univariate case ($p = 1$) it can be visualized. Specifically, the behaviour of the MAP can then be illustrated, which – as the MAP estimator corresponds to the lasso regression estimator – should also exhibit the selection property (see Exercise 6.15). The bottom left panel of Figure 6.8 shows the posterior distribution for various choices of the Laplace scale parameter (i.e. lasso penalty parameter). Clearly, the mode shifts towards zero as the scale parameter decreases / lasso penalty parameter increases. In particular, the posterior obtained from the Laplace prior with the smallest scale parameter (i.e. largest penalty parameter λ_1), although skewed to the left, has a mode placed exactly at zero. The Laplace prior may thus produce MAP estimators that select. However, for smaller values of the lasso penalty parameter the Laplace prior is not concentrated enough around zero and the contribution of the likelihood in the posterior outweighs that of the prior. The mode is then not located at zero and the parameter is ‘selected’ by the MAP estimator. The bottom

right panel of Figure 6.8 plots the mode of the normal-Laplace posterior vs. the Laplace scale parameter. In line with Theorem 6.1 it is piece-wise linear.

Park and Casella (2008) go beyond the elementary correspondence of the frequentist lasso estimator and the Bayesian posterior mode and formulate the Bayesian lasso regression model. To this end they exploit the fact that the Laplace distribution can be written as a scale mixture of normal distributions with an exponential mixing density. This allows the construction of a Gibbs sampler for the Bayesian lasso estimator. Finally, they suggest to impose a gamma-type hyperprior on the (square of the) lasso penalty parameter. Such a full Bayesian formulation of the lasso problem enables the construction of credible sets (i.e. the Bayesian counterpart of confidence intervals) to express the uncertainty of the maximum a posterior estimator. However, the lasso regression estimator may be seen as a Bayesian estimator, in the sense that it coincides with the posterior mode, the ‘lasso’ posterior distribution cannot be blindly used for uncertainty quantification. In high-dimensional sparse settings the ‘lasso’ posterior distribution of β need not concentrate around the true parameter, even though its mode is a good estimator of the regression parameter (cf. Section 3 and Theorem 7 of Castillo *et al.*, 2015).

6.8 Penalty parameter selection

The informed choice of the lasso penalty parameter can be made by similar means as for the ridge penalty parameter, e.g. cross-validation, information criteria (Section 1.8). Here we describe an alternative procedure for the selection of the lasso penalty parameter based on another heuristic, which aims to ensure a reliable covariate selection by the estimator.

6.8.1 Stability selection

Stability selection is an alternative procedure to choose the penalty parameter of the lasso regression estimator (Meinshausen and Bühlmann, 2010). Central to the procedure is a map of λ_1 from its unitless scale of the positive reals to one with a tangible interpretation. This requires the generation of many perturbed versions of the data set. The lasso regression estimator is fitted on each of these version. The number of times a covariate is selected by the estimator over all versions is called its *selection frequency*. The selection frequency is a quantity that is directly related to λ_1 as λ_1 controls which covariates enter the model. Moreover, by their stability covariates with a high selection frequency are preferred over those with a low one. The stability selection procedure then chooses λ_1 such that the resulting model includes only ‘stable’ covariates. In particular, the procedure bounds the number of falsely selected covariates based a cut-off on the selection frequency, in other words, what is considered a stable covariate. This guides an informed choice on the amount of penalization.

Let us detail the stability selection procedure. Define the sets $\mathcal{S} = \{j : (\beta)_j \neq 0\}$ and $\mathcal{N} = \{j : (\beta)_j = 0\}$ that indicate at which elements the regression parameter β has support and where it does not, respectively. An estimate of the set \mathcal{S} is the set of nonzero coefficient of the lasso regression estimate $\hat{\mathcal{S}}_{\lambda_1} = \{j : [\beta(\lambda_1)]_j \neq 0\}$, and its complement is an estimate of \mathcal{N} . Ideally, we choose λ_1 such that $\hat{\mathcal{S}}_{\lambda_1} \cap \mathcal{S} = \mathcal{S}$ and $\hat{\mathcal{S}}_{\lambda_1} \cap \mathcal{N} = \emptyset$.

The stability selection procedure chooses λ_1 on the basis of the covariates’ stability. Meinshausen and Bühlmann (2010) assess this stability over perturbed versions of the data created by subsampling. Each subsample yields an evaluation of the lasso regression estimator and, thus, an estimate of \mathcal{S} . Selection behavior over the subsamples is described by the concept of selection frequency/probability.

Definition 6.1

Let \mathcal{I} be random drawn without replacement from $\{1, \dots, n\}$ such that $|\mathcal{I}| = \lfloor n/2 \rfloor$. For the j -th covariate, the probability of being in the selected set $\hat{\mathcal{S}}_{\lambda_1}(\mathcal{I})$ is $\hat{\Pi}_{\lambda_1}(j) = P[j \in \hat{\mathcal{S}}_{\lambda_1}(\mathcal{I})]$, where the dependence on the subsample is explicated in the notation.

The probability definition above is with respect to all possible subsamples of the same size and for the particular value of λ_1 . This probability is studied over a set of penalty parameters denoted Λ . Hereto we introduce the *stability path* of the j -th covariate, which is the set of the selection probabilities $\hat{\Pi}_{\lambda_1}(j)$ obtained by running the regularization parameter λ_1 over the elements of Λ for subsample \mathcal{I} . Furthermore, we define $\hat{\mathcal{S}}_{\Lambda}(\mathcal{I}) = \bigcup_{\lambda_1 \in \Lambda} \hat{\mathcal{S}}_{\lambda_1}(\mathcal{I})$, the set of selected covariates at any point of their stability path. For a covariate to be in $\hat{\mathcal{S}}_{\Lambda}(\mathcal{I})$ for a particular subsample \mathcal{I} is not of interest. But if it is for many subsamples, it is considered stable.

Definition 6.2

The set of stable covariates is:

$$\hat{\mathcal{S}}^{\text{stable}} = \{j : \max_{\lambda_1 \in \Lambda} \hat{\Pi}_{\lambda_1}(j) \geq \pi_{\text{thr}}\},$$

for threshold $\pi_{\text{thr}} \in (0, 1)$ and the set of regularization parameters Λ .

A covariate is thus considered stable if at some point on its stability path it is selected in more than $100\pi_{\text{thr}}\%$ of the subsamples. Meinshausen and Bühlmann (2010) claim that selection on the basis of stability is relatively insensitive to either the choice of π_{thr} or that of λ_1 and Λ . Meinshausen and Bühlmann (2010) then prove that error control of selection based on stability is possible.

Theorem 6.4 (Theorem 1, Meinshausen and Bühlmann, 2010)

Assume the distribution of $\mathbb{1}_{\{j \in \hat{\mathcal{S}}_{\lambda_1}\}}$ to be exchangeable over the random subsamples for all $\lambda_1 \in \Lambda$ and $j \in \mathcal{N}$. Also, assume the original procedure is not worse than random guessing, i.e.

$$\frac{\mathbb{E}(|\mathcal{S} \cap \hat{\mathcal{S}}_{\Lambda}|)}{\mathbb{E}(|\mathcal{N} \cap \hat{\mathcal{N}}_{\Lambda}|)} = \frac{|\mathcal{S}|}{|\mathcal{N}|}.$$

The expected number $V = |\mathcal{N} \cap \hat{\mathcal{S}}^{\text{stable}}|$ of falsely selected but stable variables is then bounded for $\pi_{\text{thr}} \in (\frac{1}{2}, 1)$ by $V \leq (2\pi_{\text{thr}} - 1)^{-1}p^{-1}q_{\Lambda}^2$, where $q_{\Lambda} = \mathbb{E}_{\mathcal{I}}[|\hat{\mathcal{S}}_{\Lambda}(\mathcal{I})|]$ is the expected number of selected covariates over all subsamples of equal size.

Exchangeability thus requires that for the covariates with a zero coefficient the probability of being selected is invariant under subsampling over the full range of penalty parameters. The validity of the exchangeability may be hard to assess in practice. The random guessing assumption, however, seems unproblematic for any minimally sophisticated method.

Theorem 6.4 can be put to practical use and guide the decision on the optimal value(s) of λ_1 as follows. Specify the stability threshold, i.e. the bound on the covariates' selection frequency beyond which they are considered stable, π_{thr} . Then, to ensure (say) $\mathbb{E}(V) \leq 1$ choose the penalty parameter λ_1 such that $q_{\Lambda} \leq \sqrt{(2\pi_{\text{thr}} - 1)p}$. While q_{Λ} is in principle unknown, it can be estimated by means of the resampling.

6.9 Comparison to ridge

Here an inventory of the similarities and differences between the lasso and ridge regression estimators is presented. To recap what we have seen so far: both estimators optimize a loss function of the form (6.1) and can be viewed as Bayesian estimators. But in various respects the lasso regression estimator exhibited differences from its ridge counterpart: *i*) the former need not be uniquely defined (for a given value of the penalty parameter) whereas the latter is, *ii*) an analytic form of the lasso regression estimator does in general not exists, but *iii*) it is sparse (for large enough values of the lasso penalty parameter). The remainder of this section expands this inventory.

6.9.1 Linearity

The ridge regression estimator is a linear (in the observations) estimator, while the lasso regression estimator is not. This is immediate from the analytic expression of the ridge regression estimator, $\hat{\beta}(\lambda_2) = (\mathbf{X}^\top \mathbf{X} + \lambda_2 \mathbf{I}_{pp})^{-1} \mathbf{X}^\top \mathbf{Y}$, which is a linear combination of the observations \mathbf{Y} . To show the non-linearity of the lasso regression estimator available, it suffices to study the analytic expression of j -th element of $\hat{\beta}(\lambda_1)$ in the orthonormal case: $\hat{\beta}_j(\lambda_1) = \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \frac{1}{2}\lambda_1)_+ = \text{sign}(\mathbf{X}_{*,j}^\top \mathbf{Y})(|\mathbf{X}_{*,j}^\top \mathbf{Y}| - \frac{1}{2}\lambda_1)_+$. This clearly is not linear in \mathbf{Y} . Consequently, the response \mathbf{Y} may be scaled by some constant c , denoted $\tilde{\mathbf{Y}} = c\mathbf{Y}$, and the corresponding ridge regression estimators are one-to-one related by this same factor $\hat{\beta}(\lambda_2) = c\hat{\beta}(\lambda_2)$. The lasso regression estimator based on the unscaled data is not so easily recovered from its counterpart obtained from the scaled data.

6.9.2 Shrinkage

Both lasso and ridge regression estimation minimize the sum-of-squares plus a penalty. The latter encourages the estimator to be small, in particular closer to zero. This behavior is called shrinkage. The particular form of the penalty yields different types of this shrinkage behavior. This is best grasped in the case of an orthonormal design matrix. The j -the element of the ridge regression estimator then is: $\hat{\beta}_j(\lambda_2) = (1 + \lambda_2)^{-1} \hat{\beta}_j$, while that of the lasso regression estimator is: $\hat{\beta}_j(\lambda_1) = \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \frac{1}{2}\lambda_1)_+$. In Figure 6.9 these two estimators $\hat{\beta}_j(\lambda_2)$ and $\hat{\beta}_j(\lambda_1)$ are plotted as a function of the maximum likelihood estimator $\hat{\beta}_j$. Figure 6.9 shows that lasso and ridge regression estimator translate and scale, respectively, the maximum likelihood estimator, which could also have

been concluded from the analytic expression of both estimators. The scaling of the ridge regression estimator amounts to substantial and little shrinkage (in an absolute sense) for elements of the regression parameter β with a large and small maximum likelihood estimate, respectively. In contrast, the lasso regression estimator applies an equal amount of shrinkage to each element of β , irrespective of the coefficients' sizes.

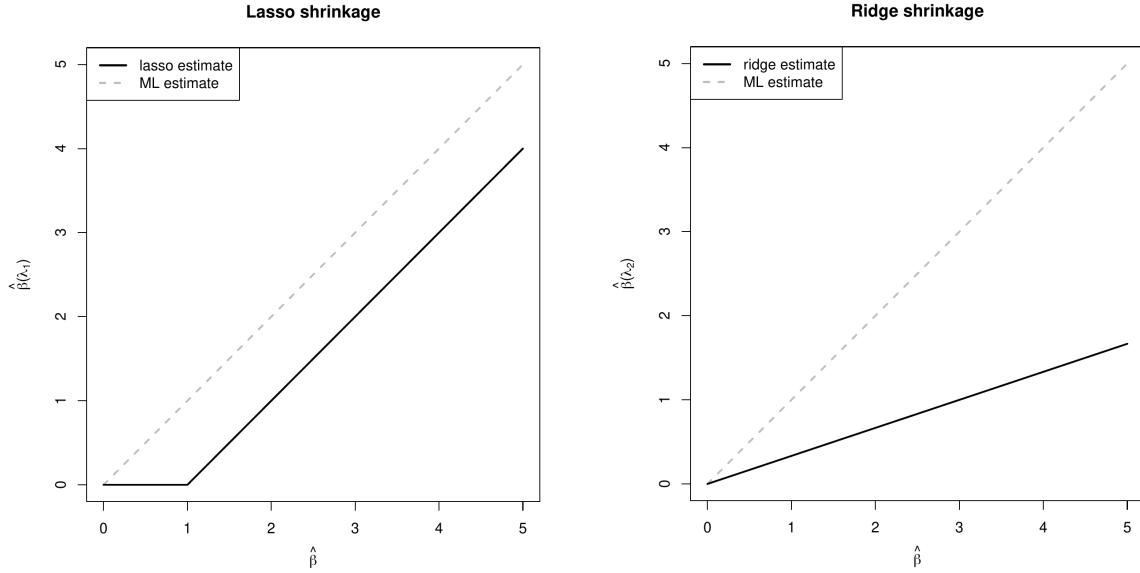


Figure 6.9: Solution path of the lasso and ridge regression estimators, left and right panel, respectively, for data with an orthonormal design matrix.

6.9.3 Simulation I: covariate selection

Here it is investigated whether lasso regression exhibits the same behaviour as ridge regression in the presence of covariates with differing variances. Recall: the simulation of Section 1.9.1 showed that ridge regression shrinks the estimates of covariates with a large spread less than those with a small spread. That simulation has been repeated, with the exact same parameter choices and sample size, but now with the ridge regression estimator replaced by the lasso regression estimator. To refresh the memory: in the simulation of Section 1.9.1 the linear regression model is fitted, now with the lasso regression estimator. The $(n = 1000) \times (p = 50)$ dimensional design matrix \mathbf{X} is sampled from a multivariate normal distribution: $\mathbf{X}_{i,*}^\top \sim \mathcal{N}(\mathbf{0}_{50}, \Sigma)$ with Σ diagonal and $(\Sigma)_{jj} = j/10$ for $j = 1, \dots, p$. The response \mathbf{Y} is generated through $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with β a vector of all ones and ε sampled from the multivariate standard normal distribution. Hence, all covariates contribute equally to the response.

The results of the simulation are displayed in Figure 6.10, which shows the regularization paths of the $p = 50$ covariates. The regularization paths are demarcated by color and style to indicate the size of the spread of the corresponding covariate. These regularization paths show that the lasso regression estimator shrinks – like the ridge regression estimator – the covariates with the smallest spread most. For the lasso regression this translates (for sufficiently large values of the penalty parameter) into a preference for the selection of covariates with largest variance.

Intuition for this behavior of the lasso regression estimator may be obtained through geometrical arguments analogous to that provided for the similar behaviour of the ridge regression estimator in Section 1.9.1. Algebraically it is easily seen when assuming an orthogonal design with $\text{Var}(X_1) \gg \text{Var}(X_2)$. The lasso regression loss function can then be rewritten, as in Example 6.5, to:

$$\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1 = \|\mathbf{Y} - \tilde{\mathbf{X}}\gamma\|_2^2 + \lambda_1[\text{Var}(X_1)]^{-1/2}|\gamma_1| + \lambda_1[\text{Var}(X_2)]^{-1/2}|\gamma_2|,$$

where $\gamma_1 = [\text{Var}(X_1)]^{1/2}\beta_1$ and $\gamma_2 = [\text{Var}(X_2)]^{1/2}\beta_2$. The rescaled design matrix $\tilde{\mathbf{X}}$ is now orthonormal and analytic expressions of estimators of γ_1 and γ_2 are available. The former parameter is penalized substantially less than the latter as $\lambda_1[\text{Var}(X_1)]^{-1/2} \ll \lambda_1[\text{Var}(X_2)]^{-1/2}$. As a result, if for large enough values of λ_1 one variable is selected, it is more likely to be γ_1 .

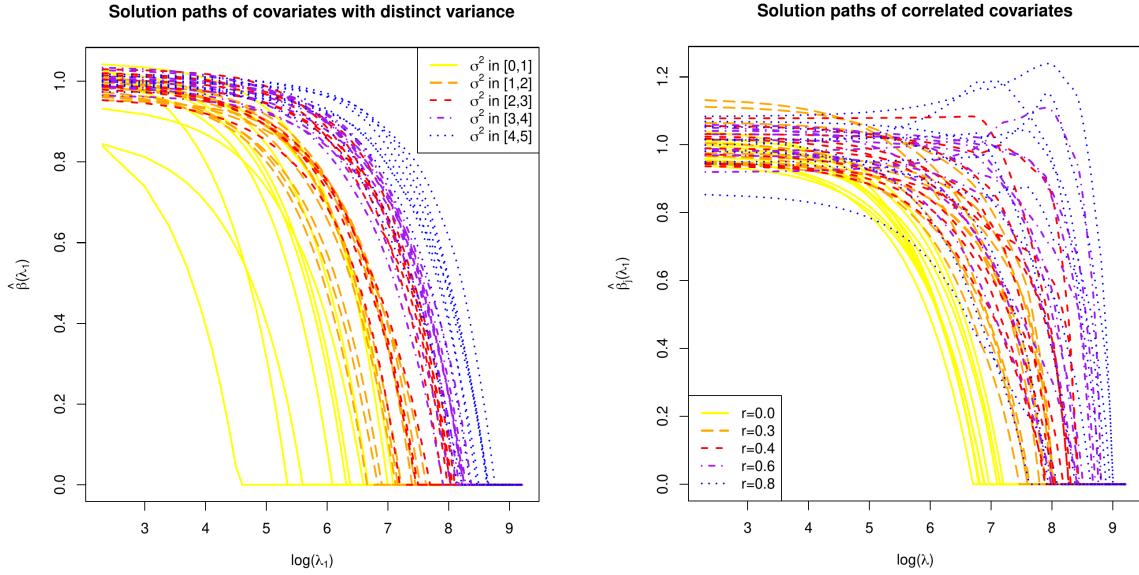


Figure 6.10: Regularization paths of the lasso regression estimator related to the simulations presented in Sections 6.9.3 and 6.9.4 in the left and right panel, respectively.

6.9.4 Simulation II: correlated covariates

The behaviour of the lasso regression estimator is now studied in the presence of collinearity among the covariates. Previously, in simulation, Section 1.9.2, the ridge regression estimator was shown to exhibit the joint shrinkage of strongly collinear covariates. This simulation is repeated for the lasso regression estimator. The details of the simulation are recapped. The linear regression model is fitted by means of the lasso regression estimator. The $(n = 1000) \times (p = 50)$ dimensional design matrix \mathbf{X} is samples from a multivariate normal distribution: $\mathbf{X}_{i,*}^\top \sim \mathcal{N}(\mathbf{0}_{50}, \boldsymbol{\Sigma})$ with a block-diagonal $\boldsymbol{\Sigma}$. The k -the, $k = 1, \dots, 5$, diagonal block, denoted $\boldsymbol{\Sigma}_{kk}$ comprises ten covariates and equals $\frac{k-1}{5} \mathbf{1}_{10 \times 10} + \frac{6-k}{5} \mathbf{I}_{10 \times 10}$ for $k = 1, \dots, 5$. The response vector \mathbf{Y} is then generated by $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon}$ sampled from the multivariate standard normal distribution and $\boldsymbol{\beta}$ containing only ones. Again, all covariates contribute equally to the response.

The results of the above simulation results are captured in Figure 6.10. It shows the lasso regularization paths for all elements of the regression parameter $\boldsymbol{\beta}$. The regularization paths of covariates corresponding to the same block of $\boldsymbol{\Sigma}$ (indicative of the degree of collinearity) are now marcate by different colors and styles. Whereas the ridge regularization paths nicely grouped per block, the lasso counterparts do not. The selection property spoils the party. Instead of shrinking the regression parameter estimates of collinear covariates together, the lasso regression estimator (for sufficiently large values of its penalty parameter λ_1) tends to pick one covariates to enters the model while forcing the others out (by setting their estimates to zero).

6.9.5 Simulation III: sparsity

Our next simulation investigates the effect of sparsity on the predictive performance of the lasso and ridge regression estimators. Hereto we draw again from the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$. Throughout we set the sample size to $n = 100$ and adopt a unit error variance $\sigma^2 = 1$. Furthermore, to see the effect of the dimension, we vary it such that $p \in \{50, 500, 5000\}$. An $n \times p$ -dimensional design matrix \mathbf{X} is then formed by subsetting the `breastCancerVDX`-package's gene expression data (REF), with covariates and samples randomly drawn. Subsequently, we center the covariates at zero. This data-driven choice of the design matrix harbors strongly collinear covariates that have different variances. Furthermore, with respect to the regression parameter we assume either a sparse or dense structure:

- *sparse* : i.e. $\beta_j = j$ for $j = 1, \dots, 5$ and $\beta_j = 0$ otherwise, or
- *dense* : i.e. $\beta_j = 10p^{-1}$ for all $j = 1, \dots, p$.

In the dense case the size of the elements of $\boldsymbol{\beta}$ depends on p . This aims to keep the variance of the response constant over the various employed dimensions. With the design matrix and regression parameter at hand, we draw the response in accordance with the distributional assumption of the linear regression model. Next we

choose the penalty parameter of both estimators by means of leave-one-out cross-validation. The estimators are then evaluated with the optimal penalty parameter, and subsequently, the prediction of the left-out samples are obtained. These predictions are then compared to the corresponding observations by means of Spearman's rank correlation. This performance measure is hardly affected by the shrunken scale of the predictions due the regularization. For the lasso regression estimate we also register which elements of the parameter are nonzero. The aforementioned has been repeated a thousand times for each setting.

Listing 6.4 R code

```
# activate libraries
library(penalized)
library(breastCancerVDX)
library(Bioconductor)

# set parameters
structure <- c("dense", "sparse") [2]
p           <- c(50, 500, 5000) [1]
n           <- 100

# load data
data(vdx)

# define storage variables
selFreqsL1    <- matrix(0, nrow=p, ncol=1)
cors          <- matrix(nrow=0, ncol=4)
colnames(cors) <- c("n", "p", "corL1", "corL2")
for (u in 1:1000) {
  # create design matrix, regression parameter, and response
  X           <- exprs(vdx)
  idCovariates <- sample(1:nrow(X), p)
  idSamples    <- sample(1:ncol(X), n)
  X           <- t(X[idCovariates, idSamples])
  X           <- sweep(X, 2, apply(X, 2, mean), "-")
  if (structure == "dense"){
    betas <- matrix(rep(10/p, p), ncol=1)
  }
  if (structure == "sparse"){
    betas   <- matrix(rep(0, p), ncol=1)
    betas[1:5] <- c(1:5)
  }
  Y <- X %*% betas + rnorm(n)

  # lasso: optimal lambda, estimate model, extract prediction and selection
  optLasso    <- optL1(Y, X, trace=FALSE)
  YpredL1    <- optLasso$prediction[,1]
  selFreqsL1 <- selFreqsL1 + 1*(coef(optLasso$fullfit, "all") [-1] != 0)

  # ridge: optimal lambda, estimate model and extract prediction
  optRidge   <- optL2(Y, X, trace=FALSE)
  YpredL2    <- optRidge$prediction[,1]

  # results
  cors <- rbind(cors, c(n, p, cor(Y, YpredL1, m="s"),
                        cor(Y, YpredL2, m="s")))
}

}
```

The predictive performances as measured by Spearman's rank correlation over all iterations of the simulation is depicted by violinplots in Figure 6.11. In the setting with a dense regression parameter, the ridge regression estimates show a better performance than its lasso counterpart. This becomes more pronounced for larger dimensions: the former's performance is reasonably constant over the dimensions, while the latter's performance dilutes slightly and, additionally, becomes more unstable, i.e. the spread among the thousand correlations increases. In the sparse case, the roles are reversed and the lasso regression estimates exhibit a better performance. But with

larger dimensions the spread among both estimates' performance measures increases, although most for the ridge regression estimator. Overall, the ridge and lasso regression estimators are preferred if the true regression parameter is dense and sparse, respectively.

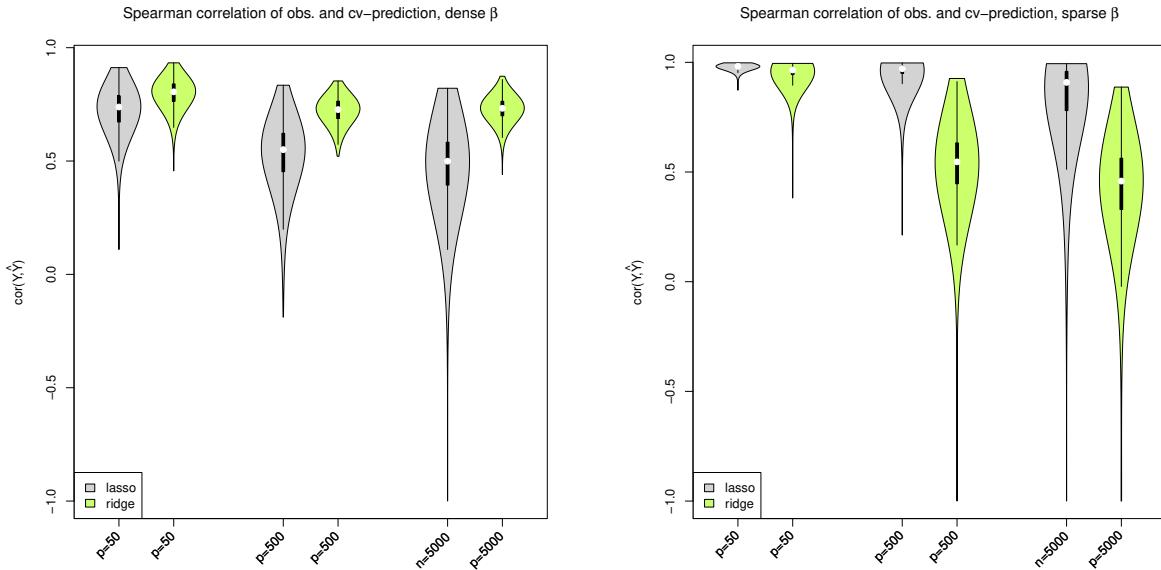


Figure 6.11: Violinplots of the predictive performance, operationalized as Spearman's rank correlation among observation and prediction, of the lasso and ridge regression estimates with fixed absolute sparsity at 10% for various increasing p and a dense (left) and sparse (right) regression parameter.

The selection frequency of the lasso regression estimator (not shown) reveals that larger (in absolute sense) elements of the parameter are selected more often than smaller ones. The selection frequency of either waters down as the dimension increases. In principle, we could do a similar exercise for the ridge estimator by studying the ranks of the estimate's absolute value. This indicates similar behavior, i.e. the ranks are concordant with the true parameter values, but a formal selection procedure is absent, although we could formulate some thresholding-type procedure.

We have repeated the simulation but fix the sparsity in a relative sense instead of an absolute one. That is, the relative sparsity is fixed at 10%, i.e. $\lceil p/10 \rceil$ elements of the regression parameter are non-zero, instead of the five non-zero elements irrespective of the dimension. The elements of the regression parameter are set to $\beta_j = 500 j p^{-7/4}$ for $j = 1, \dots, \lceil p/10 \rceil$ and $\beta_j = 0$ for $j = \lceil p/10 \rceil + 1, \dots, p$. The particular dependence of the nonzero elements of the regression parameter on the dimension is a clumsy attempt to fix the variance of the response over the employed range of dimensions. The latter now ranges over $p \in \{50, 250, 500, 750, \dots, 2500\}$. Figure 6.12 shows the violinplots of the resulting thousand Spearman's rank correlations between the predictions and observations of the lasso and ridge estimates for the various dimensions. For small dimensions, $p = 50$ and $p = 250$, the predictive performance of the ridge regression estimates falls behind that of its lasso counterpart. But as the dimension grows beyond $p = 500$, the roles reverse and the ridge regression estimate performs better than its lasso counterpart.

Let us close with some intuition why we see those results. The lasso regression estimator selects, and thereby performs dimension reduction. This may be unstable and thereby less reproducible in high-dimensional settings. The ridge regression estimator can be conceived as doing some form of averaging, which is relatively robust and reproducible.

In all, these simulations may suggest that overall the 'ridge predictor' tends to have a better predictive performance than its lasso counterpart. Should we prefer thus the ridge over the lasso regression estimator? No, that would be shortsighted. The simulations are far too limited in nature, they are only meant to illustrate the dependence of the behavior of the estimators under various choices of the regression parameter. One may interpret the

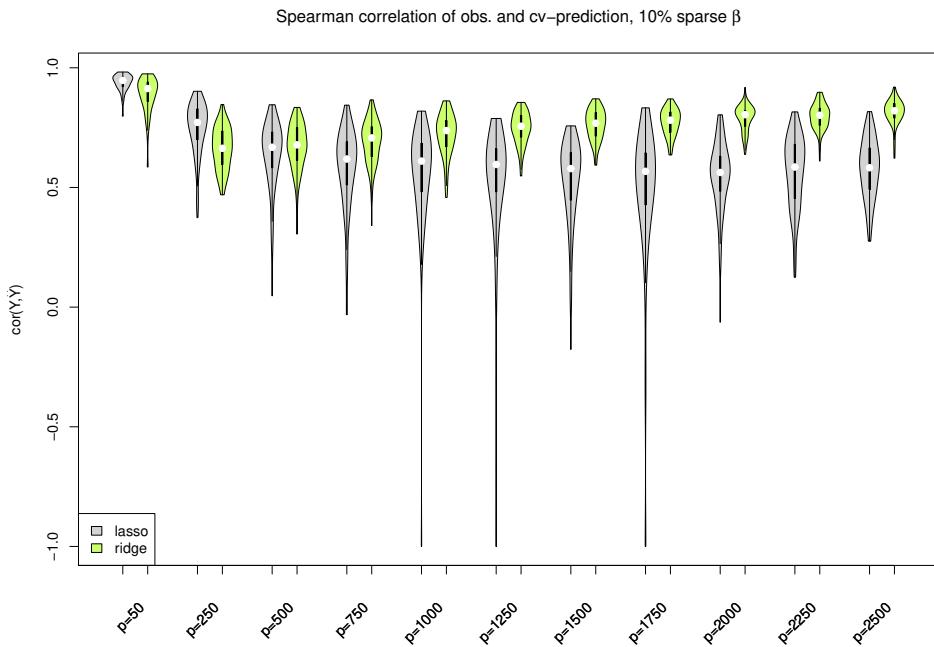


Figure 6.12: Violinplots of the predictive performance, operationalized as Spearman's rank correlation among observation and prediction, of the lasso and ridge regression estimates with fixed relative sparsity at 10% and increasing p .

simulations as suggesting that a choice for either estimator comes with a belief on the structure of the regression parameter. E.g., the lasso regression estimator implicitly assumes the system under study is sparse (and the nonzero elements are clearly distinguishable – in some sense – from zero). The validity of this assumption is not directly obvious is, e.g., biological phenomena. Similarly, the ridge regression estimator implicitly assumes all (or at least many) covariates contribute to the explanation of the variation of the response. Such an assumption too is difficult to verify.

6.10 Application

A seminal high-dimensional data set, that has often been re-analyzed in many articles to illustrate novel (regularization) methods, is presented in Van't Veer *et al.* (2002); Van De Vijver *et al.* (2002). It is part of a study into breast cancer and comprises the overall survival information and expression levels of 24158 genes of 291 breast cancer samples. Alongside the high-throughput omics data, clinical information like age and sex is available but discarded here. The data are provided via the `breastCancerNKI`-package (Schroeder *et al.*, 2022).

In the original work of Van't Veer *et al.* (2002); Van De Vijver *et al.* (2002) the data are used to build a linear predictor of overall survival. This resulted in a so-called ‘signature’ of 70 genes – quite a dimension reduction! – that together are predictive of overall survival. Upon this work, a commercial enterprise has been founded that introduced a diagnostic test called the MammaPrint (<https://en.wikipedia.org/wiki/MammaPrint>). In a nutshell, the workings of the test (essentially) amount to the evaluation of the linear predictor, which is subsequently compared to a reference value to decide on the expected survival outcome. Within the context of our data set, the reference value will be the median of all linear predictions:

$$\begin{cases} \text{poor prognosis} & \text{if } \mathbf{X}_{i,*}\hat{\beta} > \text{median}\{\mathbf{X}_{1,*}\hat{\beta}, \dots, \mathbf{X}_{n,*}\hat{\beta}\}, \\ \text{good prognosis} & \text{if } \mathbf{X}_{i,*}\hat{\beta} < \text{median}\{\mathbf{X}_{1,*}\hat{\beta}, \dots, \mathbf{X}_{n,*}\hat{\beta}\}. \end{cases}$$

The prognosis determines the individual’s follow-up treatment. For instance, should the test indicate a ‘good prognosis’, the individual may be spared chemo-therapy without a substantial overall survival reduction but a considerable gain in quality of life.

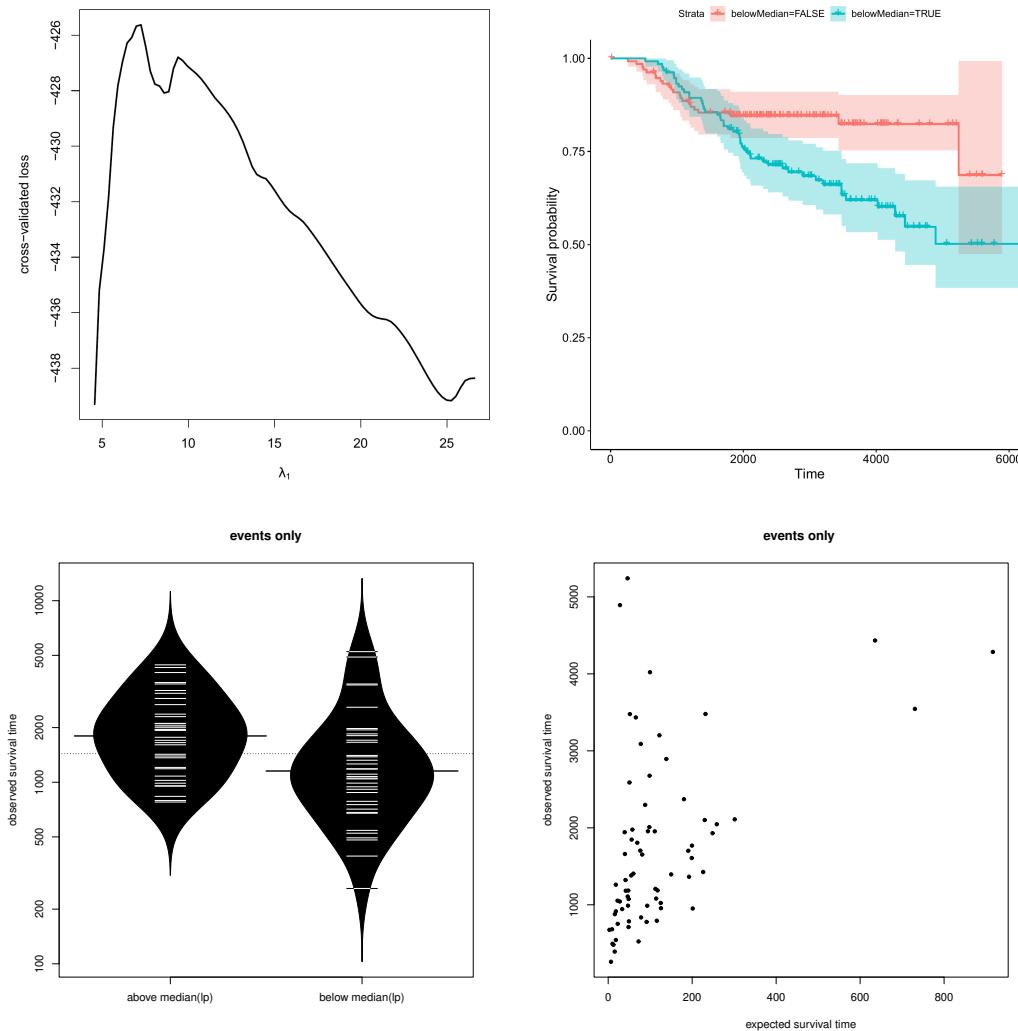


Figure 6.13: State diagrams of two Markov chains with corresponding transition matrices. The '*' in the transition matrices indicate unknown values of the transition probabilities.

Here we re-analyse the data set of Van't Veer *et al.* (2002); Van De Vijver *et al.* (2002) and predict the breast cancer overall survival by gene expression levels. To this end we adopt the Cox proportional hazards model that describes the instantaneous rate of death at time t conditional on survival until then, or later. Mathematically, the hazard of individual i is then $h_0(t) \exp(\mathbf{X}_{i,*}\boldsymbol{\beta})$, where $h_0(t)$ is the baseline hazard function common to all. The covariates then affect the hazard in a multiplicative manner, that is independent of time. The regression parameter is estimated by maximization of the (partial) likelihood. Due to the high-dimensionality, the latter is augmented with a penalty. To mimick the parameter selection used in the construction of the gene signature, we employ the lasso penalty. The penalty parameter λ_1 is chosen through leave-one-out cross-validation. The following R-code loads the data, trains the model, and performs some simple diagnostics of the model's fit.

Listing 6.5 R code

```
# activate libraries
library(breastCancerNKI)
library(BioBase)
library(penalized)
library(impute)
library(beanplot)
library(survival)
library(survminer)
```

```

# load data to memory
data(nki)
X      <- t(exprs(nki))
time   <- pData(nki) [,20]
status <- pData(nki) [,21]

# remove samples without outcome
X      <- X[ -which(is.na(time)), ]
status <- status[-which(is.na(time))]
time   <- time[ -which(is.na(time))]

# remove genes with too many missings
X <- X[, -which(colSums(is.na(X)) > 25)]

# remove samples with too many missings
time   <- time[ -which(rowSums(is.na(X)) > 100)]
status <- status[-which(rowSums(is.na(X)) > 100)]
X      <- X[ -which(rowSums(is.na(X)) > 100),]

# impute missing values
X <- t(impute.knn(t(X)) $data)

# make Surv object
Y <- Surv(time, status)

# cross-validated plot
profL1list <- profL1(Y ~ X, model="cox")
plot(profL1list[[3]] ~ profL1list[[1]], type="l", lwd=2,
     xlab=expression(lambda[1]), ylab="cross-validated loss", main="")

# find optimal lambda + estimate model for this lambda
optLasso <- optL1(Y, X, model="cox")

# generate Kaplan-Meier plot
linPredL1  <- X %*% coef(optLasso$fullfit, "all")
belowMedian <- (linePredL1 > median(linPredL1))
ggsurvplot(survfit(Y ~ belowMedian,
                    data=as.data.frame(belowMedian)), conf.int=TRUE)

# obtain fit
coxFit <- coxph(Y ~ linPredL1)

# compare expectation to observation
expTime <- predict(coxFit, type="expected") [status==1]*365
plot(expTime, time=status==1, pch=20, xlab="expected survival time",
     ylab="observed survival time", main="events only")

# compare grouping to observation
belowMedian      <- 1*(expTime > median(expTime))
slh              <- rep("below_median(lp)", length(belowMedian))
slh[belowMedian==1] <- "above_median(lp)"
beanplot(time$status==1 ~ as.factor(slh),
         main="events only", ylab="observed survival time")

```

The upper left panel of Figure 6.13 shows the cross-validated loglikelihood versus the penalty parameter. The plot reveals that has it has multiple local maxima. This is frequently seen for the lasso-type estimators when trained through cross-validation. Hence, care should be exercised when adopting the outcome of a search algorithm for the optimal penalty parameter: it may correspond to a local maximum.

The estimated linear predictor with the optimal cross-validated penalty parameter comprises 20 genes, an even stronger dimension reduction than in the original analysis of) (but they used different statistical machinery to do so). The resulting fit is depicted in various ways by three panels of Figure 6.13. The upper right panel shows

the Kaplan-Meier curves of the groups of individuals falling below and above the linear predictor's median. The bottom left panels compares the observed non-censored survival times of these groups, while the bottom right panel plots the these survival times against their expected values. The diagnostic plots reveal some value on the formed linear predictor.

One may now ask oneself whether the 20 genes selected by our lasso estimator are well-established (or novel) contributors to breast cancer. Let us first answer this anecdotally. A few years after work of Van't Veer *et al.* (2002); Van De Vijver *et al.* (2002), the results of a similar study were published. It too presented a gene signature for overall survival of breast cancer patients based on expression levels. This signature comprises a similar amount of genes, but the two gene signatures, that became known as the 'Amsterdam' and 'Rotterdam' signatures, showed little overlap. Clinicians did not know which signature to prefer (football fans may see a potential rivalry along familiar lines here), as neither set of signature genes was clearly implicated in breast cancer. The signatures' minimal overlap and their lack of a clear relation to breast cancer was a cause for concern. To investigate this, Ein-Dor *et al.* (2005) conducted a simple *in silico* experiment. They took the original Amsterdam signature of 70 genes. Subsequently, they removed these 70 genes from the set of covariates, and built another predictor comprising 70 covariates/genes. In turn the latter 70 covariates were removed too, and again a novel predictor of 70 covariates was built. And so on, until ten predictors of equal size were obtained. These predictors differed little in terms of their performance. Hence, predictors with non-overlapping gene sets may perform equally well. The line of thought behind this experiment was pushed *in extremis*. The authors skipped the training of a model and simply formed a signature from a randomly selected gene set of size $p \approx 100$. The title of their article captures the essence of their conclusion: "Most random gene expression signatures are significantly associated with breast cancer outcome". To conclude, Ein-Dor *et al.* (2006) showed by simulation that a training set of thousands of samples is needed to produce a predictor with a stable gene set. The lasso estimator selects, but it does so to explain the response best, not to facilitate a convenient contextual interpretation. Moreover, as supercollinearity abounds in high-dimensional settings, there typically is an alternative parameter selection with equal performance. Hence, show restrained when assigning (too?) much interpretational weight to the selected set of covariates.

6.11 Exercises

Question 6.1

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma_\varepsilon^2 \mathbf{I}_{nn})$. This model (without intercept) is fitted to data using the ridge regression estimator $\hat{\beta}(\lambda_1) = \arg \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1$ with $\lambda_1 > 0$. The data are:

$$\mathbf{X}^\top = \begin{pmatrix} 1/2 & -1/2 & 1/2 & -1/2 \end{pmatrix} \text{ and } \mathbf{Y}^\top = \begin{pmatrix} -1.5 & 2.9 & -3.5 & 0.7 \end{pmatrix}.$$

- a) Verify that the design matrix is orthonormal.
- b) Evaluate the maximum likelihood/ordinary least squares estimator of the regression parameter, i.e. $\hat{\beta}(\lambda)$ for $\lambda = 0$.
- c) Evaluate the lasso regression estimator for $\lambda = 1$.
- d) Verify that the lasso regression estimator $\hat{\beta}(\lambda_1)$ shrinks as λ_1 increases. Hereto combine your answer to parts b) and c) and the evaluation of the ridge regression estimator for $\lambda_1 = 4$ and $\lambda_1 = 8$. Is the order of the employed choices of λ_1 and the absolute value of the corresponding estimates concordant?

Question 6.2

Find the lasso regression solution for the data below for a general value of λ and for the straight line model $Y = \beta_0 + \beta_1 X + \varepsilon$ (only apply the lasso penalty to the slope parameter, not to the intercept). Show that when λ_1 is chosen as 14, the lasso solution fit is $\hat{Y} = 40 + 1.75X$. Data: $\mathbf{X}^\top = (X_1, X_2, \dots, X_8)^\top = (-2, -1, -1, -1, 0, 1, 2, 2)^\top$, and $\mathbf{Y}^\top = (Y_1, Y_2, \dots, Y_8)^\top = (35, 40, 36, 38, 40, 43, 45, 43)^\top$.

Question 6.3

Consider the standard linear regression model $Y_i = \mathbf{X}_{i,*}\beta + \varepsilon_i$ for $i = 1, \dots, n$ and with $\varepsilon_i \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$. The model comprises a single covariate and, depending on the subquestion, an intercept. Data on the response and the covariate are: $\{(y_i, x_{i,1})\}_{i=1}^4 = \{(1.4, 0.0), (1.4, -2.0), (0.8, 0.0), (0.4, 2.0)\}$.

- a) Evaluate the lasso regression estimator of the model without intercept for the data at hand with $\lambda_1 = 0.2$.
- b) Evaluate the lasso regression estimator of the model with intercept for the data at hand with $\lambda_1 = 0.2$ that does not apply to the intercept (which is left unpenalized).

Question 6.4

Plot the regularization path of the lasso regression estimator over the range $\lambda_1 \in (0, 160]$ using the data of Example 1.2.

Question 6.5

Consider the standard linear regression model $Y_i = X_{i,1}\beta_1 + X_{i,2}\beta_2 + \varepsilon_i$ for $i = 1, \dots, n$ and with the ε_i i.i.d. normally distributed with zero mean and some known common variance. In the estimation of the regression parameter $(\beta_1, \beta_2)^\top$ a lasso penalty is used: $\lambda_{1,1}|\beta_1| + \lambda_{1,2}|\beta_2|$ with penalty parameters $\lambda_{1,1}, \lambda_{1,2} > 0$.

- Let $\lambda_{1,1} = \lambda_{1,2}$ and assume the covariates are orthogonal with the spread of the first covariate being much larger than that of the second. Draw a plot with β_1 and β_2 on the x - and y -axis, respectively. Sketch the parameter constraint as implied by the lasso penalty. Add the levels sets of the sum-of-squares, $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$, loss criterion. Use the plot to explain why the lasso tends to select covariates with larger spread.
- Assume the covariates to be orthonormal. Let $\lambda_{1,2} \gg \lambda_{1,1}$. Redraw the plot of part a of this exercise. Use the plot to explain the effect of differencing $\lambda_{1,1}$ and $\lambda_{1,2}$ on the resulting lasso estimate.
- Show that the two cases (i.e. the assumptions on the covariates and penalty parameters) of part a and b of this exercise are equivalent, in the sense that their loss functions can be rewritten in terms of the other.

Question 6.6

Investigate the effect of the variance of the covariates on variable selection by the lasso. Hereto consider the toy model: $Y_i = X_{i,1} + X_{i,2} + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, 1)$, $X_{i,1} \sim \mathcal{N}(0, 1)$, and $X_{i,2} = a X_{i,1}$ with $a \in [0, 2]$. Draw a hundred samples for both $X_{i,1}$ and ε_i and construct both $X_{i,2}$ and Y_i for a grid of a 's. Fit the model by means of the lasso regression estimator with $\lambda_1 = 1$ for each choice of a . Plot e.g. in one figure a) the variance of $X_{i,1}$, b) the variance of $X_{i,2}$, and c) the indicator of the selection of $X_{i,2}$. Which covariate is selected for which values of scale parameter a ?

Question 6.7

Show the non-uniqueness of the lasso regression estimator for $p > 2$ when the design matrix \mathbf{X} contains linearly dependent columns.

Question 6.8

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and an $n \times 2$ -dimensional design matrix with zero-centered and standardized but collinear columns, i.e.:

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

with $\rho \in (-1, 1)$. Then, an analytic expression for the lasso regression estimator exists. Show that:

$$\hat{\beta}_j(\lambda_1) = \begin{cases} \text{sgn}(\hat{\beta}_j^{(ml)})[|\hat{\beta}_j^{(ml)}| - \frac{1}{2}\lambda_1(1+\rho)^{-1}]_+ & \text{if } \text{sgn}[\hat{\beta}_1(\lambda_1)] = \text{sgn}[\hat{\beta}_2(\lambda_1)], \\ & \hat{\beta}_j(\lambda_1) \neq 0 \neq \hat{\beta}_2(\lambda_1), \\ \text{sgn}(\hat{\beta}_j^{(ml)})[|\hat{\beta}_j^{(ml)}| - \frac{1}{2}\lambda_1(1-\rho)^{-1}]_+ & \text{if } \text{sgn}[\hat{\beta}_1(\lambda_1)] \neq \text{sgn}[\hat{\beta}_2(\lambda_1)], \\ & \hat{\beta}_1(\lambda_1) \neq 0 \neq \hat{\beta}_2(\lambda_1), \\ \begin{cases} 0 & \text{if } j \neq \arg \max_{j'} \{|\hat{\beta}_{j'}^{(ml)}|\} \\ \text{sgn}(\tilde{\beta}_j^{(ml)})(|\tilde{\beta}_j^{(ml)}| - \frac{1}{2}\lambda_1)_+ & \text{if } j = \arg \max_{j'} \{|\hat{\beta}_{j'}^{(ml)}|\} \end{cases} & \text{otherwise} \end{cases} .$$

where $\tilde{\beta}_j^{(ml)} = (\mathbf{X}_{*,j}^\top \mathbf{X}_{*,j})^{-1} \mathbf{X}_{*,j}^\top \mathbf{Y}$.

Question 6.9

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma_\varepsilon^2 \mathbf{I}_{nn})$. This model (without intercept) is fitted to data using the lasso regression estimator $\hat{\boldsymbol{\beta}}(\lambda_1) = \arg \min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1$. The relevant summary statistics of the data are:

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 1 & -\frac{1}{5} \\ -\frac{1}{5} & 1 \end{pmatrix}, \text{ and } \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} -7 \\ 5 \end{pmatrix}.$$

- Evaluate for $\lambda_1 = 6$ the lasso regression estimator.
- For which λ_1 does the lasso regression estimator have a single non-zero element?

Question 6.10

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\beta \in \mathbb{R}^p$ and $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma_\varepsilon^2 \mathbf{I}_{nn})$. This model is fitted to data using the lasso regression estimator $\hat{\beta}(\lambda_1) = \arg \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1$.

- Suppose $n = 2$ and $p = 3$. Could it be that $\hat{\beta}(\lambda_1) = (-1.3, 2.7, 0.9)^\top$ for some $\lambda_1 > 0$? Motivate.
- Suppose $n = 3$ and $p = 2$. Could it be that $\hat{\beta}(\lambda_1) = (-1.3, 0.9)^\top$ for $\lambda_1 = 1$ and $\hat{\beta}(\lambda_1) = (-1.5, 0.8)^\top$ for $\lambda_1 = 4$? Motivate.
- Suppose $n = 3$ and $p = 2$. Could it be that $\hat{\beta}(\lambda_1) = (-4, 2)^\top$ for $\lambda_1 = 1$, $\hat{\beta}(\lambda_1) = (-2, 1.5)^\top$ for $\lambda_1 = 2$, and $\hat{\beta}(\lambda_1) = (-1, 1)^\top$ for $\lambda_1 = 3$? Motivate.

Question 6.11

Consider the standard linear regression model $Y_i = \mathbf{X}_{i,*}\beta + \varepsilon_i$ for $i = 1, \dots, n$ and with the ε_i i.i.d. normally distributed with zero mean and a common variance. Moreover, $\mathbf{X}_{*,j} = \mathbf{X}_{*,j'}$ for all $j, j' = 1, \dots, p$ and $\sum_{i=1}^n X_{i,j}^2 = 1$. Question 1.19 revealed that in this case all elements of the ridge regression estimator are equal, irrespective of the choice of the penalty parameter λ_2 . Does this hold for the lasso regression estimator? Motivate your answer.

Question 6.12

Consider the linear regression model $Y_i = \mathbf{X}_{i,*}\beta + \varepsilon_i$ for $i = 1, \dots, n$ and with the ε_i i.i.d. normally distributed with zero mean and a common variance. Relevant information on the response and design matrix are summarized as:

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 3 & -2 \\ -2 & 2 \end{pmatrix}, \quad \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} 3 \\ -1 \end{pmatrix}.$$

The lasso regression estimator is used to learn parameter β .

- Show that the lasso regression estimator is given by:

$$\hat{\beta}(\lambda_1) = \arg \min_{\beta \in \mathbb{R}^2} 3\beta_1^2 + 2\beta_2^2 - 4\beta_1\beta_2 - 6\beta_1 + 2\beta_2 + \lambda_1 |\beta_1| + \lambda_1 |\beta_2|.$$

- For $\lambda_1 = 0.2$ the lasso estimate of the second element of β is $\hat{\beta}_2(\lambda_1) = 1.25$. Determine the corresponding value of $\hat{\beta}_1(\lambda_1)$.
- Determine the smallest λ_1 for which it is guaranteed that $\hat{\beta}(\lambda_1) = \mathbf{0}_2$.

Question 6.13

Show $\|\hat{\beta}(\lambda_1)\|_1$ is monotone decreasing in λ_1 . In this assume orthonormality of the design matrix \mathbf{X} .

Question 6.14

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$. This model is fitted to data, $\mathbf{X}_{1,*} = (4, -2)$ and $Y_1 = 10$, using the lasso regression estimator $\hat{\beta}(\lambda_1) = \arg \min_{\beta} \|Y_1 - \mathbf{X}_{1,*}\beta\|_2^2 + \lambda_1 \|\beta\|_1$.

- How many nonzero elements does the lasso regression estimator with an arbitrary $\lambda_1 > 0$ have for these data?
- Ignore the second covariate and evaluate the lasso regression estimator for $\lambda_1 = 8$.
- Suppose that, when regressing the response on each covariate separately, the corresponding lasso regression estimates with $\lambda_1 = 8$ are $\hat{\beta}_1(\lambda_1) = 2\frac{1}{4}$ and $\hat{\beta}_2(\lambda_1) = -4$. Now consider the regression problem with both covariates in the model. Does the lasso regression estimate with $\lambda_1 = 8$ then equal $\hat{\beta}(\lambda_1) = (2\frac{1}{4}, 0)^\top$, $\hat{\beta}(\lambda_1) = (0, -4)^\top$, or some other value? Motivate!

Question 6.15

Consider a single draw, denoted \mathbf{Y} , from the normal means model $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_{pp})$ with $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\sigma^2 = 1$. Assume the μ_i to be i.i.d. distributed following the double exponential distribution with density $f(x) = (2b)^{-1} \exp(-b^{-1}|x|)$. Show that, if $b = \lambda_1^{-1}$, the MAP (Mode A Posteriori) estimator $\hat{\mu}_i^{\text{map}}$ equals $\text{sign}(Y_j)(|Y_j| - \lambda_1)_+$ for $j = 1, \dots, p$.

Question 6.16

A researcher has measured gene expression measurements for 1000 genes in 40 subjects, half of them cases and the other half controls.

- Describe and explain what would happen if the researcher would fit an ordinary logistic regression to these data, using case/control status as the response variable.

- b) Instead, the researcher chooses to fit a lasso regression, choosing the tuning parameter lambda by cross-validation. Out of 1000 genes, 37 get a non-zero regression coefficient in the lasso fit. In the ensuing publication, the researcher writes that the 963 genes with zero regression coefficients were found to be ‘irrelevant’. What is your opinion about this statement?

Question 6.17

Consider the standard linear regression model $Y_i = \mathbf{X}_{i,*}\boldsymbol{\beta} + \varepsilon_i$ for $i = 1, \dots, n$ and with the ε_i i.i.d. normally distributed with zero mean and a common variance. Let the first covariate correspond to the intercept. The model is fitted to data by means of the minimization of the sum-of-squares augmented with a lasso penalty in which the intercept is left unpenalized: $\lambda_1 \sum_{j=2}^p |\beta_j|$ with penalty parameter $\lambda_1 > 0$. The penalty parameter is chosen through leave-one-out cross-validation (LOOCV). The predictive performance of the model is evaluated, again by means of LOOCV. Thus, creating a double cross-validation loop. At each inner loop the optimal λ_1 yields an empty intercept-only model, from which a prediction for the left-out sample is obtained. The vector of these prediction is compared to the corresponding observation vector through their Spearman correlation (which measures the monotonicity of a relationship and – as a correlation measure – assumed values on the $[-1, 1]$ interval with an analogous interpretation to the ‘ordinary’ correlation). The latter equals -1 . Why?

Question 6.18

Load the breast cancer data available via the `breastCancerNKI`-package (downloadable from BioConductor) through the following R-code:

Listing 6.6 R code

```
# activate library and load the data
library(breastCancerNKI)
data(nki)

# subset and center the data
X <- exprs(nki)
X <- X[-which(rowSums(is.na(X)) > 0),]
X <- apply(X[1:1000,], 1, function(X) { X - mean(X) })

# define ER as the response
Y <- pData(nki)[,8]
```

The eSet-object `nki` is now available. It contains the expression profiles of 337 breast cancer patients. Each profile comprises expression levels of 24481 genes. The R-code above extracts the expression data from the object, removes all genes with missing values, centers the gene expression gene-wise around zero, and subsets the data set to the first thousand genes. The reduction of the gene dimensionality is only for computational speed. Furthermore, it extracts the estrogen receptor status (short: ER status), an important prognostic indicator for breast cancer, that is to be used as the response variable in the remainder of the exercise.

- Relate the ER status and the gene expression levels by a logistic regression model, which is fitted by means of the lasso penalized maximum likelihood method. First, find the optimal value of the penalty parameter of λ_1 by means of cross-validation. This is implemented in `optL1`-function of the `penalized`-package available from CRAN.
- Evaluate whether the cross-validated likelihood indeed attains a maximum at the optimal value of λ_1 . This can be done with the `profL1`-function of the `penalized`-package available from CRAN.
- Investigate the sensitivity of the penalty parameter selection with respect to the choice of the cross-validation fold.
- Does the optimal lambda produce a reasonable fit? And how does it compare to the ‘ridge fit’?

7 Generalizing lasso regression

Many variants of penalized regression, in particular of lasso regression, have been presented in the literature. Here we give an overview of some of the more current ones. Not a full account is given, but rather a brief introduction with emphasis on their motivation and use.

7.1 Elastic net

The elastic net regression estimator is a modification of the lasso regression estimator that preserves its strength and harnesses its weaknesses. The biggest appeal of the lasso regression estimator is clearly its ability to perform selection. Less pleasing are *i*) the non-uniqueness of the lasso regression estimator due to the non-strict convexity of its loss function, *ii*) the bound on the number of selected variables, i.e. maximally $\min\{n, p\}$ can be selected, and *iii*) the observation that strongly (positively) collinear covariates are not shrunken together: the lasso regression estimator selects among them while it is hard to distinguish their contributions to the variation of the response. While it does not select, the ridge regression estimator does not exhibit these less pleasing features. These considerations led Zou and Hastie (2005) to combine the strengths of the lasso and ridge regression estimators and form a ‘best-of-both-worlds’ estimator, called the *elastic net* regression estimator, defined as:

$$\hat{\beta}(\lambda_1, \lambda_2) = \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \frac{1}{2} \lambda_2 \|\beta\|_2^2.$$

The elastic net penalty – defined implicitly in the preceding display – is thus simply a linear combination of the lasso and ridge penalties. Consequently, the elastic net regression estimator encompasses its lasso and ridge counterparts. Hereto just set $\lambda_2 = 0$ or $\lambda_1 = 0$, respectively. A novel estimator is defined if both penalties act simultaneously, i.e. if their corresponding penalty parameters are both nonzero.

Does this novel elastic net estimator indeed inherit the strengths of the lasso and ridge regression estimators? Let us turn to the aforementioned motivation behind the elastic net estimator. Starting with the uniqueness, the strict convexity of the ridge penalty renders the elastic net loss function strictly convex, as it is a combination of the ridge penalty and the lasso function – notably non-strict convex when the dimension p exceeds the sample size n . This warrants the existence of a unique minimizer of the elastic net loss function. To assess the preservation of the selection property, now without the bound on the maximum number of selectable variables, exploit the equivalent constraint estimation formulation of the elastic net estimator. Figure 7.1 shows the parameter constraint of the elastic net estimator for the ‘ $p = 2$ ’-case, which is defined by the set:

$$\{(\beta_1, \beta_2) \in \mathbb{R}^2 : \lambda_1(|\beta_1| + |\beta_2|) + \frac{1}{2} \lambda_2(\beta_1^2 + \beta_2^2) \leq c(\lambda_1, \lambda_2)\}.$$

Visually, the ‘elastic net parameter constraint’ is a compromise between the circle and the diamond shaped constraints of the ridge and lasso regression estimators. This compromise inherits exactly the right geometrical features: the strict convexity of the ‘ridge circle’ and the ‘corners’ (referring to points at which the constraint’s boundary is non-smoothness/non-differentiability) falling at the axes of the ‘lasso constraint’. The latter feature, by the same argumentation as presented in Section 6.3, endows the elastic net estimator with the selection property. Moreover, it can – in principle – select p features as the point in the parameter space where the smallest level set of the unpenalized loss hits the elastic net parameter constraint need not fall on any axis. For example, in the ‘ $p = 2, n = 1$ ’-case the level sets of the sum-of-squares loss are straight lines that, if running almost parallel to the edges of the ‘lasso diamond’, are unlikely to first hit the elastic net parameter constraint at one of its corners. Finally, the largest penalty parameter relates (reciprocally) to the volume of the elastic net parameter constraint, while the ratio between λ_1 and λ_2 determines whether it is closer to the ‘ridge circle’ or to the ‘lasso diamond’.

Whether the elastic net regression estimator also delivers on the joint shrinkage property is assessed by simulation (not shown). The impression given by these simulations is that the elastic net has joint shrinkage potential. This, however, usually requires a large ridge penalty, which then dominates the elastic net penalty.

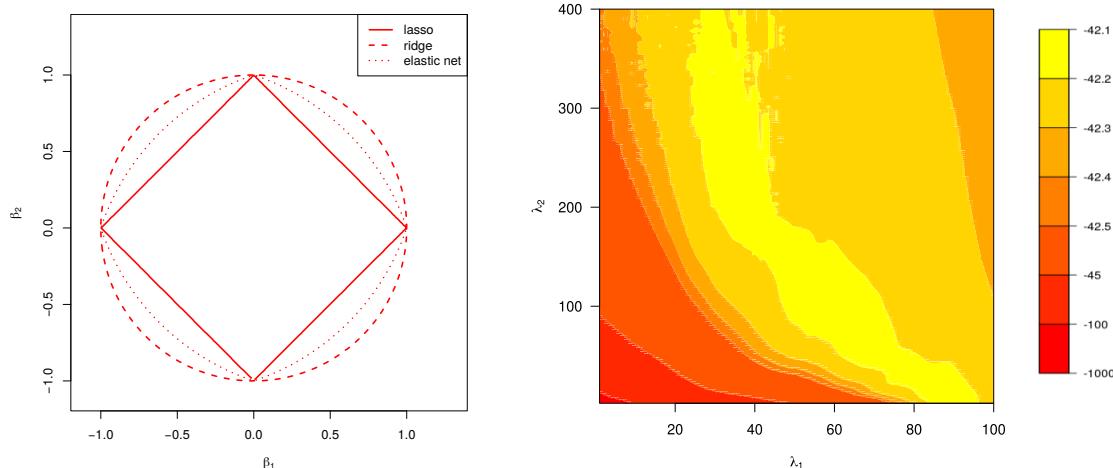


Figure 7.1: The left panel depicts the parameter constraint induced by the elastic net penalty and, for reference, those of the lasso and ridge are added. The right panel shows the contour plot of the cross-validated loglikelihood vs. the two penalty parameters of the elastic net estimator.

The elastic net regression estimator can be found with procedures similar to those that evaluate the lasso regression estimator (see Section 6.4) as the elastic net loss can be reformulated as a lasso loss. Hereto the ridge part of the elastic net penalty is absorbed into the sum of squares using the data augmentation trick of Exercise 1.6 which showed that the ridge regression estimator is the ML regression estimator of the related regression model with p zeros and rows added to the response and design matrix, respectively. That is, write $\tilde{\mathbf{Y}} = (\mathbf{Y}^\top, \mathbf{0}_p^\top)^\top$ and $\tilde{\mathbf{X}} = (\mathbf{X}^\top, \sqrt{\lambda_2} \mathbf{I}_{pp})^\top$. Then:

$$\|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2.$$

Hence, the elastic net loss function can be rewritten to $\|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1$. This is familiar territory and the lasso algorithms of Section 6.4 can be used. Zou and Hastie (2005) present a different algorithm for the evaluation of the elastic net estimator that is faster and numerically more stable (see also Exercise 7.5). Irrespectively, the reformulation of the elastic net loss in terms of augmented data also reveals that the elastic net regression estimator can select p variables. Even for $p > n$, which is immediate from the observation that $\text{rank}(\tilde{\mathbf{X}}) = p$.

The penalty parameters need tuning, e.g. by cross-validation. They are subject to empirically indeterminacy. That is, often a large range of (λ_1, λ_2) -combinations will yield a similar cross-validated performance as can be witnessed from Figure 7.1. It shows the contourplot of the penalty parameters vs. this performance. There is a yellow ‘banana-shaped’ area that corresponds to the same and optimal performance. Hence, no single (λ_1, λ_2) -combination can be distinguished as all yield the best performance. This behaviour may be understood intuitively. Reasoning loosely, while the lasso penalty $\lambda_1 \|\boldsymbol{\beta}\|_1$ ensures the selection property and the ridge penalty $\frac{1}{2} \lambda_2 \|\boldsymbol{\beta}\|_2^2$ warrants the uniqueness and joint shrinkage of coefficients of collinear covariates, they have a similar effect on the size of the estimator. Both shrink, although in different norms. But a reduction in the size of $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ in one norm implies a reduction in another. An increase in either the lasso and ridge penalty parameter will have a similar effect on the elastic net estimator: it shrinks. The selection and ‘joint shrinkage’ properties are only consequences of the employed penalty and are not criteria in the optimization of the elastic net loss function. There, only size matters. The size of $\boldsymbol{\beta}$ refers to $\lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{1}{2} \lambda_2 \|\boldsymbol{\beta}\|_2^2$. As in the size the lasso and ridge penalties appear as a linear combination in the elastic net loss function and have a similar effect on the elastic net estimator: there are many positive (λ_1, λ_2) -combinations that constrain the size of the elastic net estimator equally. In contrast, for both the lasso and ridge regression estimators, different penalty parameters yield estimators of different sizes (defined accordingly). Moreover, it is mainly the size that determines the cross-validated performance as the size determines the shrinkage of the estimator and, consequently, the size of the errors. But only a fixed size leaves enough freedom to distribute this size over the p elements of the regression parameter estimator $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ and, due to the collinearity, among them many that yield a comparable performance. Hence, if a particularly sized elastic net estimator $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ optimizes the cross-validated performance, then high-dimensionally there are likely many others with a different (λ_1, λ_2) -combination but of equal size and similar performance.

The empirical indeterminacy of penalty parameters touches upon another issues. In principle, the elastic net

regression estimator can decide whether a sparse or non-sparse solution is most appropriate. The indeterminacy indicates that for any sparse elastic net regression estimator a less sparse one can be found with comparable performance, and vice versa. Care should be exercised when concluding on the sparsity of the linear relation under study from the chosen elastic net regression estimator.

A solution to the indeterminacy of the optimal penalty parameter combination is to fix their ratio. For interpretation purposes this is done through the introduction of a ‘mixing parameter’ $\alpha \in [0, 1]$. The elastic net penalty is then written as $\lambda[\alpha\|\beta\|_1 + \frac{1}{2}(1 - \alpha)\|\beta\|_2^2]$. The mixing parameter is set by the user while $\lambda > 0$ is typically found through cross-validation (cf. the implementation in the `glmnet`-package) (Friedman *et al.*, 2009). Generally, no guidance on the choice of mixing parameter α can be given. In fact, it is a tuning parameter and as such needs tuning rather than setting out of the blue.

7.2 Fused lasso

The fused lasso regression estimator proposed by Tibshirani *et al.* (2005) is the counterpart of the fused ridge regression estimator encountered in Example 3.1. It is a generalization of the lasso regression estimator for situations where the order of index j , $j = 1, \dots, p$, of the covariates has a certain meaning such as a spatial or temporal one. The fused lasso regression estimator minimizes the sum-of-squares augmented with the lasso penalty, the sum of the absolute values of the elements of the regression parameter, and the ℓ_1 -*fusion* (or simply *fusion* if clear from the context) penalty, the sum of the first order differences of the regression parameter. Formally, the fused lasso estimator is defined as:

$$\hat{\beta}(\lambda_1, \lambda_{1,f}) = \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_{1,f} \sum_{j=2}^p |\beta_j - \beta_{j-1}|,$$

which involves two penalty parameters λ_1 and $\lambda_{1,f}$ for the lasso and fusion penalties, respectively. As a result of adding the fusion penalty the fused lasso regression estimator not only shrinks elements of β towards zero but also the difference of neighboring elements of β . In particular, for large enough values of the penalty parameters the estimator selects elements of and differences of neighboring elements of β . This corresponds to a sparse estimate of β while the vector of its first order differences too is dominated by zeros. That is many elements of $\hat{\beta}(\lambda_1, \lambda_{1,f})$ equal zero with few changes in $\hat{\beta}(\lambda_1, \lambda_{1,f})$ when running over j . Hence, the fused lasso regression penalty encourages large sets of neighboring elements of j to have a common (or at least a comparable) regression parameter estimate. This is visualized – using simulated data from a simple toy model with details considered irrelevant for the illustration – in Figure 7.2 where the elements of the fused lasso regression estimate $\hat{\beta}(\lambda_1, \lambda_{1,f})$ are plotted against the index of the covariates. For reference the true β and the lasso regression estimate with the same λ_1 are added to the plot. Ideally, for a large enough fusion penalty parameter, the elements of $\hat{\beta}(\lambda_1, \lambda_{1,f})$ would form a step-wise function in the index j , with many equalling zero and exhibiting few changes, as the elements of β do. While this is not the case, it is close, especially in comparison to the elements of its lasso cousin $\hat{\beta}(\lambda_1)$, thus showing the effect of the inclusion of the fusion penalty.

It is insightful to view the fused lasso regression problem as a constrained estimation problem. The fused lasso penalty induces a parameter constraint: $\{\beta \in \mathbb{R}^p : \lambda_1\|\beta\|_1 + \lambda_{1,f} \sum_{j=2}^p |\beta_j - \beta_{j-1}| \leq c(\lambda_1, \lambda_{1,f})\}$. This constraint is plotted for $p = 2$ in the right panel of Figure 7.2 (clearly, it is not the intersection of the constraints induced by the lasso and fusion penalty separately as one might accidentally conclude from Figure 2 in Tibshirani *et al.*, 2005). The constraint is convex, although not strict, which is convenient for optimization purposes. Moreover, the geometry of this fused lasso constraint reveals why the fused lasso regression estimator selects the elements of β as well as its first order difference. Its boundary, while continuous and generally smooth, has six points at which it is non-differentiable. These all fall on the grey dotted lines in the right panel of Figure 7.2 that correspond to the axes and the diagonal, put differently, on either the ‘ $\beta_1 = 0$ ’, ‘ $\beta_2 = 0$ ’, or ‘ $\beta_1 = \beta_2$ ’-lines. The fused lasso regression estimate is the point where the smallest level set of the sum-of-squares criterion $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$, be it an ellipsoid or hyper-plane, hits the fused lasso constraint. For an element or the first order difference to be zero it must fall on one of the dotted grey lines of the right panel of Figure 7.2. Exactly this happens on one of the aforementioned six point of the constraint. Finally, the fused lasso regression estimator has, when – for reasonably comparable penalty parameters λ_1 and $\lambda_{1,f}$ – it shrinks the first order difference to zero, a tendency to also estimate the corresponding individual elements as zero. In part, this is due to the fact that $|\beta_1| = 0 = |\beta_2|$ implies that $|\beta_1 - \beta_2| = 0$, while the reverse does not necessarily hold. Moreover, if $|\beta_1| = 0$, then $|\beta_1 - \beta_2| = |\beta_2|$. The fusion penalty thus converts to a lasso penalty of the remaining nonzero element of this first order difference, i.e. $(\lambda_1 + \lambda_{1,f})|\beta_2|$, thus furthering the shrinkage of this element to zero.

The evaluation of the fused lasso regression estimator is more complicated than that of the ‘ordinary’ lasso regression estimator. For ‘moderately sized’ problems Tibshirani *et al.* (2005) suggest to use a variant of the

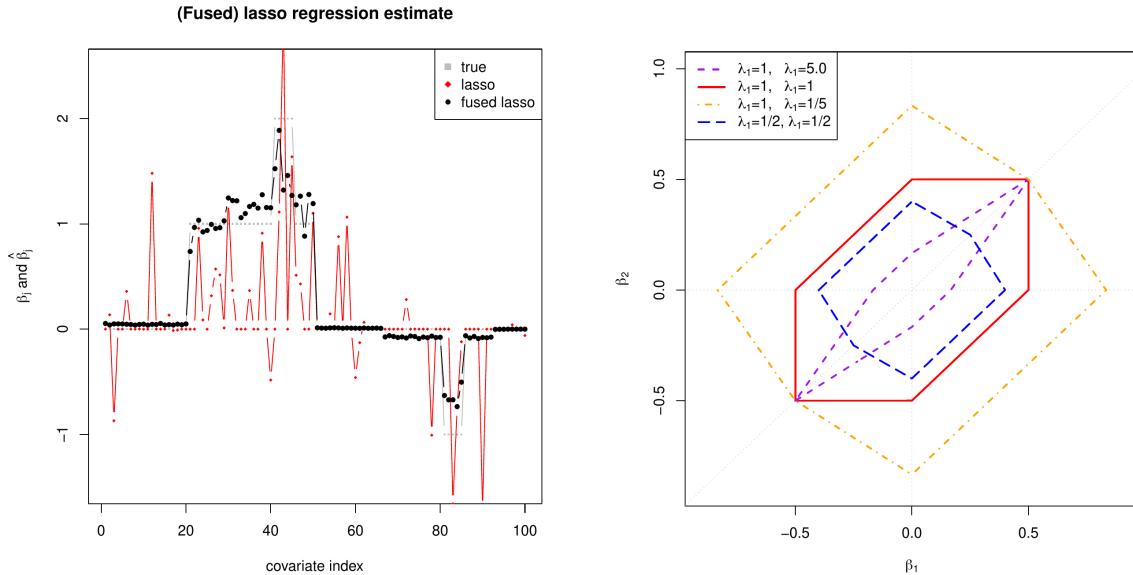


Figure 7.2: The left panel shows the lasso (red, diamonds) and fused lasso (black circles) regression parameter estimates and the true parameter values (grey circles) plotted against their index j . The (β_1, β_2) -parameter constraint induced by the fused lasso penalty for various combinations of the lasso and fusion penalty parameters λ_1 and $\lambda_{1,f}$, respectively. The grey dotted lines corresponds to the ‘ $\beta_1 = 0$ ’-, ‘ $\beta_2 = 0$ ’- and ‘ $\beta_1 = \beta_2$ ’-lines where selection takes place.

quadratic programming method (see also Section 6.4.1) that is computationally efficient when many linear constraints are active, i.e. if many elements of and first order difference of β are zero. Chaturvedi *et al.* (2014) extend the gradient ascent approach discussed in Section 6.4.3 to solve the minimization of the fused lasso loss function. For the limiting ‘ $\lambda_1 = 0$ ’-case the fused lasso loss function can be reformulated as a lasso loss function (see Exercise 7.1). Then, the algorithms of Section 6.4 may be applied to find the estimate $\hat{\beta}(0, \lambda_{1,f})$.

7.3 The (sparse) group lasso

The lasso regression estimator selects covariates, irrespectively of the relation among them. However, groups of covariates may be discerned. For instance, a group of covariates may be dummy variables representing levels of a categorical factor. Or, within the context of gene expression studies such groups may be formed by so-called pathways, i.e. sets of genes that work in concert to fulfill a certain function in the cell. In such cases a group-structure can be overlayed on the covariates and it may be desirable to select the whole group, i.e. all covariates together, rather than an individual covariate of the group. To achieve this Yuan and Lin (2006) proposed the group lasso regression estimator. It minimizes the sum-of-squares now augmented with the *group lasso* penalty, i.e.:

$$\lambda_{1,G} \sum_{g=1}^G \sqrt{|\mathcal{J}_g|} \|\beta_g\|_2 = \lambda_{1,G} \sum_{g=1}^G \sqrt{|\mathcal{J}_g|} \sqrt{\sum_{j \in \mathcal{J}_g} \beta_j^2},$$

where $\lambda_{1,G}$ is the group lasso penalty parameter (with subscript G for Group), G is the total number of groups, $\mathcal{J}_g \subset \{1, \dots, p\}$ is covariate index set of the g -th group such that the \mathcal{J}_g are mutually exclusive and exhaustive, i.e. $\mathcal{J}_{g_1} \cap \mathcal{J}_{g_2} = \emptyset$ for all $g_1 \neq g_2$ and $\cup_{g=1}^G \mathcal{J}_g = \{1, \dots, p\}$, and $|\mathcal{J}_g|$ denotes the cardinality (the number of elements) of \mathcal{J}_g .

The group lasso estimator performs covariate selection at the group level but does not result in a sparse within-group estimate. This may be achieved through employment of the *sparse group lasso* regression estimator (Simon *et al.*, 2013):

$$\hat{\beta}(\lambda_1, \lambda_{1,G}) = \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_{1,G} \sum_{g=1}^G \sqrt{|\mathcal{J}_g|} \|\beta_g\|_2,$$

which combines the lasso with the group lasso penalty. The inclusion of the former encourages within-group sparsity, while the latter performs selection at the group level. The sparse group lasso penalty resembles the elastic net penalty with the $\|\beta\|_2$ -term replacing the $\|\beta\|_2^2$ -term of the latter.

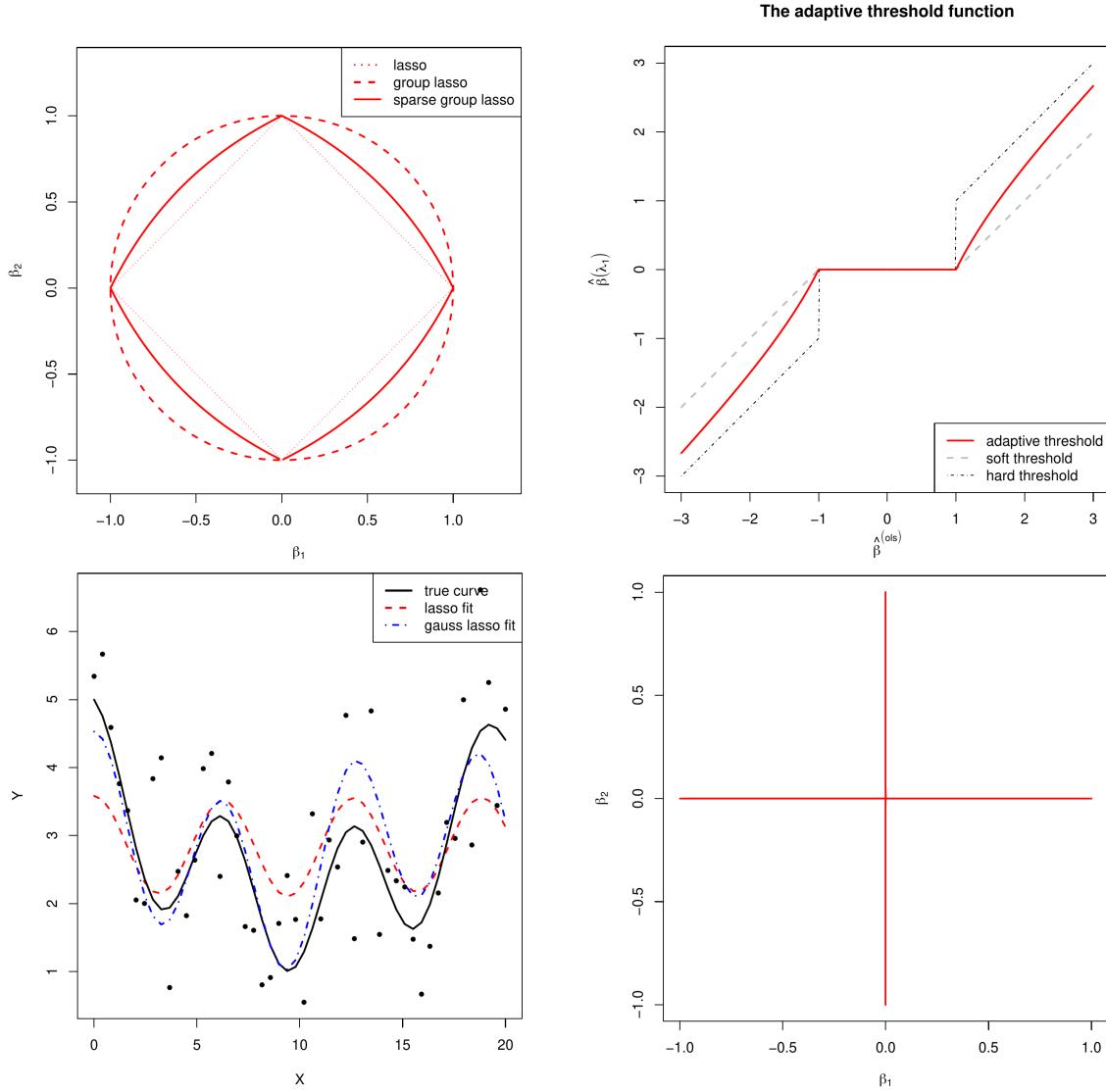


Figure 7.3: Top left panel: The (β_1, β_2) -parameter constraint induced by the lasso, group lasso and sparse group lasso penalty. Top right panel: the adaptive threshold function associated with the adaptive lasso regression estimator for orthonormal design matrices. Bottom left panel: illustration of the lasso and adaptive lasso fit and the true curve. Bottom right panel: the parameter constraint induced by the ℓ_0 -penalty.

The (sparse) group lasso regression estimation problems can be reformulated as constrained estimation problems. Their parameter constraints are depicted in Figure 7.3. By now the reader will be familiar with the characteristic feature, i.e. the non-differentiability of the boundary at the axes, of the constraint that endows the estimator with the potential to select. This feature is clearly present for the sparse group lasso regression estimator, covariate-wise. Although both the group lasso and the sparse group lasso regression estimator select group-wise, illustration of the associated geometrical feature requires plotting in dimensions larger than two and is not attempted. However, if all groups are singletons, the (sparse) group lasso penalties are equivalent to the regular lasso penalty.

The sparse group lasso regression estimator is found through exploitation of the convexity of the loss function (Simon *et al.*, 2013). It alternates between group-wise and within-group optimization. The resemblance of the sparse group lasso and elastic net penalties propagates to the optimality conditions of both estimators. In fact, the within-group optimization amounts to the evaluation of an elastic net regression estimator (Simon *et al.*, 2013). When within each group the design matrix is orthonormal and $\lambda_{1,G} = 0$, the group lasso regression estimator can be found by a group-wise coordinate descent procedure for the evaluation of the estimator (cf. Exercise ??).

Both the sparse group lasso and the elastic net regression estimators have two penalty parameters that need tuning. In both cases the corresponding penalties have a similar effect: shrinkage towards zero. If the g -th group's

contribution to the group lasso penalty has vanished, then so has the contribution of all covariates to the regular lasso penalty. And vice versa. This complicates the tuning of the penalty parameters as it is hard to distinguish which shrinkage effect is most beneficial for the estimator. Simon *et al.* (2013) resolve this by setting the ratio of the two penalty parameters λ_1 and $\lambda_{1,G}$ to some arbitrary but fixed value, thereby simplifying the tuning.

7.4 Adaptive lasso

The lasso regression estimator does not exhibit some key and desirable asymptotic properties. Zou (2006) proposed the adaptive lasso regression estimator to achieve these properties. The adaptive lasso regression estimator is a two-step estimation procedure. First, an initial estimator of the regression parameter β , denoted $\hat{\beta}^{\text{init}}$, is to be obtained. The adaptive lasso regression estimator is then defined as:

$$\hat{\beta}^{\text{adapt}}(\lambda_1) = \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \sum_{j=1}^p |\hat{\beta}_j^{\text{init}}|^{-1} |\beta_j|.$$

Hence, it is a generalization of the lasso penalty with covariate-specific weighing. The weight of the j -th covariate is reciprocal to the j -th element of the initial regression parameter estimate $\hat{\beta}^{\text{init}}$. If the initial estimate of β_j is large or small, the corresponding element in the adaptive lasso estimator will be penalized less or more and thereby determine the amount of shrinkage, which may now vary between the estimates of the elements of β . In particular, if $\hat{\beta}_j^{\text{init}} = 0$, the adaptive lasso penalty parameter corresponding to the j -th element is infinite and yields $\hat{\beta}^{\text{adapt}} = 0$.

The adaptive lasso regression estimator, given an initial regression estimate $\hat{\beta}^{\text{init}}$, can be found numerically by minor changes of the algorithms presented in Section 6.4. In case of an orthonormal design an analytic expression of the adaptive lasso estimator exists (see Exercise 7.7):

$$\hat{\beta}_j^{\text{adapt}}(\lambda_1) = \text{sign}(\hat{\beta}_j^{\text{ols}})(|\hat{\beta}_j^{\text{ols}}| - \frac{1}{2}\lambda_1/|\hat{\beta}_j^{\text{ols}}|)_+.$$

This adaptive lasso estimator can be viewed as a compromise between the soft thresholding function, associated with the lasso regression estimator for orthonormal design matrices (Section 6.2), and the hard thresholding function, associated with truncation of the ML regression estimator (see the top right panel of Figure 7.3 for an illustration of these thresholding functions).

How is the initial regression parameter estimate $\hat{\beta}^{\text{init}}$ to be chosen? Low-dimensionally the maximum likelihood regression estimator one may used. The resulting adaptive lasso regression estimator is sometimes referred to as the *Gauss-Lasso* regression estimator. High-dimensionally, the lasso or ridge regression estimators will do. Any other estimator may in principle be used. But not all yield the desirable asymptotic properties.

A different motivation for the adaptive lasso is found in its ability to undo some or all of the shrinkage of the lasso regression estimator due to penalization. This is illustrated in the left bottom panel of Figure 7.3. It shows the lasso and adaptive lasso regression fits. The latter clearly undoes some of the bias of the former.

7.5 The ℓ_0 penalty

An alternative estimators, considered to be superior to both the lasso and ridge regression estimators, is the ℓ_0 -penalized regression estimator. It too minimizes the sum-of-squares now with the ℓ_0 -penalty. The ℓ_0 -penalty, denoted $\lambda_0 \|\beta\|_0$, is defined as $\lambda_0 \|\beta\|_0 = \lambda_0 \sum_{j=1}^p \mathbb{1}_{\{\beta_j \neq 0\}}$ with penalty parameter λ_0 and $\mathbb{1}_{\{\cdot\}}$ the indicator function. The parameter constraint associated with this penalty is shown in the right panel of Figure 7.3. From the form of the penalty it is clear that the ℓ_0 -penalty penalizes only for the presence of a covariate in the model. As such it is concerned with the number of covariates in the model, and not the size of their regression coefficients (or a derived quantity thereof). The latter is considered only a surrogate of the number of covariates in the model. As such the lasso and ridge estimators are proxies to the ℓ_0 -penalized regression estimator.

The ℓ_0 -penalized regression estimator is not used for large dimensional problems as its evaluation is computationally too demanding. It requires a search over all possible subsets of the p covariates to find the optimal model. As each covariate can either be in or out of the model, in total 2^p models need to be considered. This is not feasible with present-day computers.

The adaptive lasso regression estimator may be viewed as an approximation of the ℓ_0 -penalized regression estimator. When the covariate-wise weighing employed in the penalization of the adaptive lasso regression estimation is equal to $\lambda_{1,j} = \lambda/|\beta_j|$ with β_j known true value of the j -th regression coefficient, the j -th covariate's estimated regression coefficient contributes only to the penalty if it is nonzero. In practice, this weighing involves an initial estimate of β and is therefore an approximation at best, which the quality of the approximation hinging upon that of the weighing.

7.6 Conclusion

Finally, a note of caution in similar spirit as that which concludes Chapter 3. It is a joy to play with the penalty and see how it encourages the regression parameter estimate to exhibit hypothesized behaviour. Nonetheless, for a deeper and profound understanding of the data generating mechanism, such knowledge is ideally incorporated explicitly in the model itself.

7.7 Exercises

Question 7.1

Augment the lasso penalty with the sum of the absolute differences all pairs of successive regression coefficients:

$$\lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_{1,f} \sum_{j=2}^p |\beta_j - \beta_{j-1}|.$$

This augmented lasso penalty is referred to as the *fused lasso penalty*.

- a) Consider the standard multiple linear regression model: $Y_i = \sum_{j=1}^p X_{ij} \beta_j + \varepsilon_i$. Estimation of the regression parameters takes place via minimization of penalized sum of squares, in which the fused lasso penalty is used with $\lambda_1 = 0$. Rewrite the corresponding loss function to the standard lasso problem by application of the following change-of-variables: $\gamma_1 = \beta_1$ and $\gamma_j = \beta_j - \beta_{j-1}$.
- b) Investigate on simulated data the effect of the second summand of the fused lasso penalty on the parameter estimates. In this, temporarily set $\lambda_1 = 0$.
- c) Let λ_1 equal zero still. Compare the regression estimates of part b) to the ridge estimates with a first-order autoregressive prior. What is qualitatively the difference in the behavior of the two estimates? *Hint:* plot the full solution path for the penalized estimates of both estimation procedures.
- d) How do the estimates of part b) of this question change if we allow $\lambda_1 > 0$?

Question 7.2

Consider the standard linear regression model $Y_i = \mathbf{X}_{i,*}\boldsymbol{\beta} + \varepsilon_i$ for $i = 1, \dots, n$ and with $\varepsilon_i \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$. The rows of the design matrix \mathbf{X} are of length 2, neither column represents the intercept. Relevant summary statistics from the data on the response \mathbf{Y} and the covariates are:

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 40 & -20 \\ -20 & 10 \end{pmatrix} \quad \text{and} \quad \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} 26 \\ -13 \end{pmatrix}.$$

- a) Use lasso regression to regress without intercept the response on the first covariate. Draw (i.e. not sketch!) the regularization path of lasso regression estimator.
- b) The two covariates are perfectly collinear. However, their regularization paths do not coincide. Why?
- c) Fit the linear regression model with both covariates (and still without intercept) by means of the fused lasso, i.e. $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_f) = \arg \min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_f |\beta_1 - \beta_2|$ with $\lambda_1 = 10$ and $\lambda_F = \infty$. *Hint:* at some point in your answer you may wish to write $\mathbf{X} = (\mathbf{X}_{*,1} \ c\mathbf{X}_{*,1})$ and deduce c from $\mathbf{X}^\top \mathbf{X}$.

Question 7.3

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$. This model (without intercept) is fitted to data using the lasso regression estimator $\hat{\boldsymbol{\beta}}(\lambda_1) = \arg \min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1$. The relevant summary statistics of the data are:

$$\mathbf{X} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} -5 \\ 4 \end{pmatrix}, \mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}, \text{ and } \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} -9 \\ 9 \end{pmatrix}.$$

- a) Specify the full set of lasso regression estimates with $\lambda_1 = 2$ that minimize the lasso loss function for these data.
- b) Now consider fitting the linear regression model with the fused lasso estimator $\hat{\boldsymbol{\beta}}(\lambda_f) = \arg \min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_f |\beta_1 - \beta_2|$. Determine $\lambda_f^{(0)} > 0$ such that $\hat{\boldsymbol{\beta}}(\lambda_f) = (0, 0)^\top$ for all $\lambda_f > \lambda_f^{(0)}$.

Question 7.4

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_{nn})$. This model (without intercept) is fitted to data using the lasso regression estimator $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_f) = \arg \min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_f \sum_{j=2}^p |\beta_j - \beta_{j-1}|$. The data are:

$$\mathbf{X} = \begin{pmatrix} 3 & -1 \\ 1 & 1 \end{pmatrix}, \text{ and } \mathbf{Y} = \begin{pmatrix} -2 \\ 3 \end{pmatrix}.$$

- a) Evaluate the fused lasso regression estimator $\hat{\beta}(\lambda_1, \lambda_f)$ for $\lambda_1 = 1$ and $\lambda_f = \infty$.
b) How many covariates can the fused lasso regression estimator select? Motivate.

Question 7.5 (The elastic net regression estimator)

Consider fitting the linear regression model by means of the elastic net regression estimator.

- a) Recall the data augmentation trick of Question 1.6 of the ridge regression exercises. Use the same trick to show that the elastic net least squares loss function can be reformulated to the form of the traditional lasso function. Hint: absorb the ridge part of the elastic net penalty into the sum of squares.
b) The elastic net regression estimator can be evaluated by a coordinate descent procedure outlined in Section 6.4.4. Show that in such a procedure at each step the j -th element of the elastic net regression estimate is updated according to:

$$\hat{\beta}_j(\lambda_1, \lambda_2) = (\|\mathbf{X}_{*,j}\|_2^2 + \lambda_2)^{-1} \text{sign}(\mathbf{X}_{*,j}^\top \tilde{\mathbf{Y}})[|\mathbf{X}_{*,j}^\top \tilde{\mathbf{Y}}| - \frac{1}{2}\lambda_1]_+.$$

with $\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}_{*,\setminus j} \beta_{\setminus j}$.

Question 7.6 (The elastic net regression estimator)

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma_\varepsilon^2 \mathbf{I}_{nn})$. This model (without intercept) is fitted to data using the elastic net estimator $\hat{\beta}(\lambda_1, \lambda_2) = \arg \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1 + \frac{1}{2}\lambda_2\|\beta\|_2^2$. The relevant summary statistics of the data are:

$$\mathbf{X} = \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} -5 \\ 4 \\ 1 \end{pmatrix}, \mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 3 \end{pmatrix}, \text{ and } \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} -10 \end{pmatrix}.$$

- a) Evaluate for $(\lambda_1, \lambda_2) = (3, 2)$ the elastic net regression estimator of the linear regression model.
b) Now consider the evaluation the elastic net regression estimator of the linear regression model for the same penalty parameters, $(\lambda_1, \lambda_2) = (3, 2)$, but this time involving two covariates. The first covariate is as in part a), the second is orthogonal to that one. Do you expect the resulting elastic net estimate of the first regression coefficient $\hat{\beta}_1(\lambda_1, \lambda_2)$ to be larger, equal or smaller (in an absolute sense) than your answer to part a)? Motivate.
c) Now take in part b) the second covariate equal to the first one. Show that the first coefficient of elastic net estimate $\hat{\beta}_1(\lambda_1, 2\lambda_2)$ is half that of part a). Note: there is no need to know the exact answer to part a).

Question 7.7 *

Consider the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$. It is fitted to data from a study with an orthonormal design matrix by means of the adaptive lasso regression estimator initiated by the OLS/ML regression estimator. Show that the j -th element of the resulting adaptive lasso regression estimator equals:

$$\hat{\beta}_j^{\text{adapt}}(\lambda_1) = \text{sign}(\hat{\beta}_j^{\text{ols}})(|\hat{\beta}_j^{\text{ols}}| - \frac{1}{2}\lambda_1/|\hat{\beta}_j^{\text{ols}}|)_+.$$

*This question is freely copied from Bühlmann and Van De Geer (2011): Problem 2.5a, page 43.

Bibliography

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.
- Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, **16**(1), 125–127.
- Ambs, S., Prueitt, R. L., Yi, M., Hudson, R. S., Howe, T. M., Petrocca, F., Wallace, T. A., Liu, C.-G., Volinia, S., Calin, G. A., Yfantis, H. G., Stephens, R. M., and Croce, C. M. (2008). Genomic profiling of microrna and messenger RNA reveals deregulated microrna expression in prostate cancer. *Cancer Research*, **68**(15), 6162–6170.
- Anatolyev, S. (2020). A ridge to homogeneity for linear models. *Journal of Statistical Computation and Simulation*, **90**(13), 2455–2472.
- Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**(2), 281–297.
- Bates, D. and DebRoy, S. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, **91**(1), 1–17.
- Berger, J. . (2013). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.
- Bertsekas, D. P. (2014). *Constrained Optimization and Lagrange Multiplier Methods*. Academic press.
- Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, **103**(4), 985–991.
- Bijma, F., Jonker, M. A., and van der Vaart, A. W. (2017). *An Introduction to Mathematical Statistics*. Amsterdam University Press.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell*, **169**(7), 1177–1186.
- Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, **19**(4), 1212–1242.
- Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.
- Cancer Genome Atlas Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**(7353), 609–615.
- Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**(7407), 330–337.
- Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. W. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, **43**(5), 1986–2018.
- Chaturvedi, N., de Menezes, R. X., and Goeman, J. J. (2014). Fused lasso algorithm for Cox' proportional hazards and binomial logit models with application to copy number profiles. *Biometrical Journal*, **56**(3), 477–492.
- Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, **49**(4), 327–335.
- de Vlaming, R. and Groenen, P. J. F. (2015). The current and future use of ridge regression for prediction in quantitative genetics. *BioMed Research International*, page Article ID 143712.
- Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis (3rd edition)*. John Wiley & Sons.
- Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American statistical Association*, **81**(394), 461–470.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, **32**(2), 407–499.

- Eilers, P. (1999). Discussion on: The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **48**(3), 307–308.
- Eilers, P. and Marx, B. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, **11**(2), 89–102.
- Eilers, P. H. C., Boer, J. M., van Ommen, G.-J., and van Houwelingen, H. C. (2001). Classification of microarray data with penalized logistic regression. In *Microarrays: Optical technologies and informatics*, volume 4266, pages 187–198.
- Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**(2), 171–178.
- Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences*, **103**(15), 5923–5928.
- Esquela-Kerscher, A. and Slack, F. J. (2006). Oncomirs: microRNAs with a role in cancer. *Nature Reviews Cancer*, **6**(4), 259–269.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96**(456), 1348–1360.
- Farebrother, R. W. (1976). Further results on the mean square error of ridge regression. *Journal of the Royal Statistical Society, Series B (Methodological)*, pages 248–250.
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**(1), 27–38.
- Fletcher, R. (2008). *Practical Methods of Optimization, 2nd Edition*. John Wiley, New York.
- Flury, B. D. (1990). Acceptance-rejection sampling made easy. *SIAM Review*, **32**(3), 474–476.
- Friedman, J., Hastie, T., and Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, **1**(4).
- García, C. B., García, J., López Martín, M., and Salmerón, R. (2015). Collinearity: Revisiting the variance inflation factor in ridge regression. *Journal of Applied Statistics*, **42**(3), 648–661.
- Geyer, C. J. (2011). Introduction to Markov chain Monte Carlo. *Handbook of Markov chain Monte Carlo*, **20116022**, 45.
- Goeman, J. J. (2008). Autocorrelated logistic ridge regression for prediction based on proteomics spectra. *Statistical Applications in Genetics and Molecular Biology*, **7**(2).
- Goeman, J. J. (2010). L_1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, **52**, 70–84.
- Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, **21**(2), 215–223.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.
- Guilkey, D. K. and Murphy, J. L. (1975). Directed ridge regression techniques in cases of multicollinearity. *Journal of the American Statistical Association*, **70**(352), 769–775.
- Hansen, B. E. (2015). The risk of James–Stein and lasso shrinkage. *Econometric Reviews*, **35**(8-10), 1456–1470.
- Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.
- Harville, D. A. (2008). *Matrix Algebra From a Statistician's Perspective*. Springer, New York.
- Hastie, T. and Tibshirani, R. (2004). Efficient quadratic regularization for expression arrays. *Biostatistics*, **5**(3), 329–340.
- Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The Elements of Statistical Learning*. Springer.
- Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine*, **21**(16), 2409–2419.
- Hemmerle, W. J. (1975). An explicit solution for generalized ridge regression. *Technometrics*, **17**(3), 309–314.
- Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics*, **9**(2), 226–252.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**(1), 55–67.
- Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge University Press.
- Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*, volume 398. John Wiley & Sons.
- Hua, T. A. and Gunst, R. F. (1983). Generalized ridge regression: a note on negative ridge parameters. *Communi-*

- nlications in Statistics-Theory and Methods*, **12**(1), 37–45.
- Ishwaran, H. and Rao, J. S. (2014). Geometry and properties of generalized ridge regression in high dimensions. *Contemporary Mathematics*, **622**, 81–93.
- Jiang, J., Li, C., Paul, D., Yang, C., and Zhao, H. (2016). On high-dimensional misspecified mixed model analysis in genome-wide association study. *The Annals of Statistics*, **44**(5), 2127–2160.
- Kim, V. N. and Nam, J.-W. (2006). Genomics of microRNA. *TRENDS in Genetics*, **22**(3), 165–173.
- Lawless, J. F. (1981). Mean squared error properties of generalized ridge estimators. *Journal of the American Statistical Association*, **76**(374), 462–466.
- Le Cessie, S. and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, **41**(1), 191–201.
- Leng, C., Lin, Y., and Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica*, pages 1273–1284.
- Lettink, A., Chinapaw, M., and van Wieringen, W. N. (2023). Two-dimensional fused targeted ridge regression for health indicator prediction from accelerometer data. *Journal of the Royal Statistical Society, Series C*.
- Marquardt, D. W. (1970). Generalized inverses, ridge regression and biased linear estimation. *Technometrics*, **12**, Technometrics, 591–612.
- Mathai, A. M. and Provost, S. B. (1992). *Quadratic Forms in Random Variables: Theory and Applications*. Dekker.
- Meijer, R. J. and Goeman, J. J. (2013). Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal*, **55**(2), 141–155.
- Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(4), 417–473.
- Mersmann, O. (2014). *microbenchmark: Accurate Timing Functions*. R package version 1.4-2.
- Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics*, **9**(2), 319–337.
- Padmanabhan, V., Callas, P., Philips, G., Trainer, T., and Beatty, B. (2004). DNA replication regulation protein MCM7 as a marker of proliferation in prostate cancer. *Journal of Clinical Pathology*, **57**(10), 1057–1062.
- Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, **9**(1), 30–50.
- Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, **103**(482), 681–686.
- Peeters, C. F. W., van de Wiel, M. A., and van Wieringen, W. N. (2019). The spectral condition number plot for regularization parameter evaluation. *Computational Statistics*, **35**, 629–646.
- Pust, S., Klokk, T., Musa, N., Jenstad, M., Risberg, B., Erikstein, B., Tcachoff, L., Liestøl, K., Danielsen, H., Van Deurs, B., and K, S. (2013). Flotillin as regulators of ErbB2 levels in breast cancer. *Oncogene*, **32**(29), 3443–3451.
- Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. John Wiley & Sons.
- Ross-Adams, H., Lamb, A., Dunning, M., Halim, S., Lindberg, J., Massie, C., Egevad, L., Russell, R., Ramos-Montoya, A., Vowler, S., et al. (2015). Integration of copy number and transcriptomics provides risk stratification in prostate cancer: a discovery and validation cohort study. *EBioMedicine*, **2**(9), 1133–1144.
- Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics*, pages 1012–1030.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, **1**(5), 206–215.
- Saleh, A. K. M. E., Arashi, M., and Kibria, B. M. G. (2019). *Theory of ridge regression estimation with applications*, volume 285. John Wiley & Sons.
- Sardy, S. (2008). On the practice of rescaling covariates. *International Statistical Review*, **76**(2), 285–297.
- Schaefer, R. L., Roi, L. D., and Wolfe, R. A. (1984). A ridge logistic estimator. *Communications in Statistics: Theory and Methods*, **13**(1), 99–113.
- Schmidt-Hieber, J. (2019). Deep relu network approximation of functions on a manifold. *arXiv preprint arXiv:1908.00695*.
- Schroeder, M., Haibe-Kains, B., Culhane, A., Sotiriou, C., Bontempi, G., and Quackenbush, J. (2022).

- breastCancerNKI: Gene expression dataset published by Schmidt et al. [2008] (NKI).* R package version 1.34.0.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**(2), 461–464.
- Shao, J. and Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics*, **40**(2), 812–831.
- Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, **22**(2), 231–245.
- Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 1135–1151.
- Sternier, J. M., Dew-Knight, S., Musahl, C., Kornbluth, S., and Horowitz, J. M. (1998). Negative regulation of DNA replication by the retinoblastoma protein is mediated by its association with MCM7. *Molecular and Cellular Biology*, **18**(5), 2748–2757.
- Subramanian, J. and Simon, R. (2010). Gene expression-based prognostic signatures in lung cancer: ready for clinical use? *Journal of the National Cancer Institute*, **102**(7), 464–474.
- Theobald, C. M. (1974). Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, **36**(1), 103–106.
- Tibshirani, R. (1996). Regularized shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58**(1), 267–288.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(1), 91–108.
- Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, **7**, 1456–1490.
- Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics*, **40**(2), 1198–1232.
- Tye, B. K. (1999). MCM proteins in DNA replication. *Annual Review of Biochemistry*, **68**(1), 649–686.
- Van De Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, **347**(25), 1999–2009.
- Van De Wiel, M. A., Lien, T. G., Verlaat, W., van Wieringen, W. N., and Wilting, S. M. (2016). Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statistics in Medicine*, **35**(3), 368–381.
- van de Wiel, M. A., van Nee, M. M., and Rauschenberger, A. (2021). Fast cross-validation for multi-penalty ridge regression. *Journal of Computational and Graphical Statistics*, **30**(4), 835–847.
- van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge University Press.
- van Wieringen, W. and Aflakparast, M. (2021). *porridge: Ridge-Type Estimation of a Potpourri of Models*. R package version 0.2.1, <https://CRAN.R-project.org/package=porridge>.
- van Wieringen, W. N. and Binder, H. (2022). Sequential learning of regression models by penalized estimation. *Journal of Computational and Graphical Statistics*, **31**(3), 877–886.
- Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., Van Der Kooy, K., Marton, M. J., Witteveen, A. T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**(6871), 530–536.
- Wang, L., Tang, H., Thayanthi, V., Subramanian, S., Oberg, L., Cunningham, J. M., Cerhan, J. R., Steer, C. J., and Thibodeau, S. N. (2009). Gene networks and microRNAs implicated in aggressive prostate cancer. *Cancer research*, **69**(24), 9490–9497.
- Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. John Wiley & Sons, Chichester, England.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(1), 49–67.
- Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: essays in honor of Bruno De Finetti*, **6**, 233–243.
- Zhang, Y. and Politis, D. N. (2022). Ridge regression revisited: Debiasing, thresholding and bootstrap. *Annals of Statistics*, **50**(3), 1401–1422.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**(476), 1418–1429.

- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320.
- Zou, H., Hastie, T., and Tibshirani, R. (2007). On the “degrees of freedom” of the lasso. *The Annals of Statistics*, **35**(5), 2173–2192.
- Zwiener, I., Frisch, B., and Binder, H. (2014). Transforming RNA-seq data to improve the performance of prognostic gene signatures. *PloS one*, **9**(1), e85150.