

**UNIVERSITY OF OSLO
Faculty of Mathematics
and Natural Sciences**

**Regularized
parameter
estimation and the
choice of tuning
parameters**

Master thesis for the
programme
Modelling and Data
Analysis

Hilde Galleberg
Johnsen

May 2011



Contents

Preface	3
1 Introduction	7
2 The least squares estimator	11
2.1 Model assumptions	11
2.2 The least squares estimator and the maximum likelihood estimator	11
2.3 The least squares estimator from a Bayesian perspective	12
2.4 The least squares estimator and model selection	14
2.5 Discussion	15
3 Ridge regression	17
3.1 The ridge estimate $\hat{\beta}_{\text{ridge}}(\lambda)$	20
3.2 Estimating the penalty parameter τ	23
3.2.1 A simple approach based on the method of moments . . .	25
3.2.2 Cross validation	28
3.2.3 The maximum marginal likelihood estimator	33
3.3 A numerical example	36
3.3.1 Designing the experiment	36
3.3.2 Quantities of interest	38
3.3.3 Results	40
3.4 Discussion	44
4 The zero-mixed ridge estimator	47
4.1 The zero-mixed ridge estimate $\hat{\beta}_{\text{mix}}(\alpha, \tau)$	49
4.1.1 The connection to ordinary subset selection	53
4.2 Other possible ways of estimating β	55
4.2.1 The posterior distribution of β given α and τ	55
4.2.2 Alternative ways of estimating β	57
4.3 Estimating the parameters α and τ	59
4.3.1 A simple approach based on the method of moments . . .	59
4.3.2 Cross validation	61
4.3.3 The maximum marginal likelihood estimator	62
4.4 Discussion	63

5	The lasso	67
5.1	The lasso estimate $\hat{\beta}_{\text{lasso}}(\lambda)$	67
5.2	Estimating the penalty parameter λ	70
5.2.1	A simple approach based on the method of moments . . .	70
5.2.2	Cross validation	71
5.2.3	The maximum marginal likelihood estimator	72
5.3	A small numerical example	74
5.4	Discussion	75
6	Estimating the covariance matrix and its inverse in the Gaussian case	77
6.1	Estimating the covariance matrix	77
6.2	Estimating the precision matrix	79
6.2.1	Estimating the support of the precision matrix	80
6.2.2	Estimating the precision matrix using a lasso penalty . .	80
7	Discussion	85
	References	87

Preface

The process of finding a suitable topic for this work has been long, starting already early in the autumn of 2009. The first idea was connected to graphical models and covariance selection, starting with an article about high-dimensional graphs and variable selection using the lasso (Meinshausen and Bühlmann, 2006). At this time I had very little experience with estimation of covariance matrices, and I had only worked in the classical setting with a small number of parameters and a large number of observations. Also, I had not worked with graphical models, and never even heard of the lasso. Clearly, it took me a long time to get through this article, and in hope of making it easier, I tried to read some material on graphical models as well.

The spring of 2010 was spent reading a number of other articles. The one I had already studied was only concerned with finding the support of the inverse covariance matrix, but we now went on to a broader class of topics, including the estimation of the entire precision matrix and also the covariance matrix itself. Some articles restricted themselves to Gaussian data, while others were slightly more general. Some used the intuitive thresholding approach of simply removing all sufficiently small elements of the sample covariance matrix, whereas others used more complicated estimation methods such as an l_1 -penalty on each element of the precision matrix. It seemed like no two articles were estimating the same quantity, and that they all had their own slightly different set of assumptions, but even though I was still quite confused, I finally started to understand the motivation for these methods and get used to reading about matrices and the unfamiliar situation of high-dimensional data.

It soon became clear that although all the attempts of estimating the covariance matrix and its inverse were using some kind of regularization to obtain a sparse solution, the regularization itself were done in many different ways. My supervisors and I then felt that it was necessary to use some time studying the concept of regularization itself, which was new to me when I started reading all these articles. Probably it would have been easier for me to read the matrix articles if I had studied regularization in a simpler setting at an even earlier stage, but it is never too late, and in the late spring of 2010 I finally read the classical lasso article by Tibshirani (1996) and two other articles describing lasso-like penalty functions in the linear regression setting (Fan and Li (2001); Zou and Hastie (2005)).

Early in the autumn the same year, we had still not agreed on a specific topic for this thesis, and I used the time reading a few new articles related to the ones I had read in the spring. We agreed that we had to make a decision before Christmas, and finally it was decided that we would focus on the penalty parameter itself and how to choose it. All the regularization methods I had

studied were using one or more parameters to control the amount of regularization, but not all of them were using much time discussing how it should be chosen. Most of them used some kind of cross validation procedure, sometimes only considering a very small number of possible values, while others provided explicit rules for choosing the parameter in order to control the probability of certain errors.

It is well known that a lasso estimate in the linear regression setting may be seen as the mode of a posterior distribution obtained by using a certain prior distribution on the parameter vector. The idea was to exploit this fact when estimating the penalty parameter, by using the penalty parameter maximizing the marginal likelihood of the data. We also wanted to compare the result to other possible methods of estimation, primarily cross validation.

In the lasso case, the final parameter estimate has no simple closed form, and deriving the properties in the cross validation case is therefore very hard. Thus, we decided to start with ridge regression, another popular method of regularized parameter estimation in the linear regression case. Ridge regression provides closed form expressions for the parameter estimates, very much like the ones obtained using ordinary least squares, and we therefore expected to be able to derive a few results analytically.

We were able to derive a few results regarding the different estimators of the tuning parameter used in ridge regression, but it turned out to take quite a lot of time. The cross validation results were particularly hard to derive, as they involved very complicated formulas with a lot of similar-looking terms and therefore required very high precision.

Inspired by ridge regression and the concept of subset selection, we also invented a lasso-like method performing both shrinkage and model selection. This estimator is based on a simple idea, but the resulting expressions are not as simple as in the pure ridge regression case. We were not able to derive any results analytically, but we studied the behaviour of the resulting β -estimate, and we tried to use the same methods for estimating the tuning parameter as in the ridge setting.

We have given this new estimation method the name 'zero-mixed ridge estimation', and it was intended to be an alternative to the lasso. We have therefore also described the lasso estimate in an own chapter and tried to compare the results of the two estimation methods. Because of the more complicated expressions, the calculations were far more time-consuming than for ridge regression, and our main numerical example is therefore connected to the pure ridge case. We have performed a few computations related to zero-mixed ridge and lasso as well, and a small example is presented in Chapter 5.

Originally we were hoping to be able to apply our methods of penalty parameter estimation to the sparse precision matrix setting as well. However, the other parts were taking more time than we expected, and we decided to focus on the new idea of the zero-mixed ridge estimator instead. Since I have used a lot of time reading articles related to the problem of estimating (inverse) covariance matrices, I have included a short chapter on this topic at the end of the thesis. We have not had the time to do any work related to this, but I sum up some of the literature I have read and describe an idea for further work, based on the experience of the previous chapters.

We were also hoping to derive results for the high-dimensional setting of $p > n$, but we thought it was important to understand the low-dimensional

setting first. Even though the estimation methods proposed in this thesis are valid also in the high-dimensional setting, the results concerning their properties are generally derived in the classical small p , large n setting and can therefore not be expected to hold in general. We have not derived any results or performed any large simulation examples for the $p > n$ setting.

Even though the plans for this project have changed a couple of times during the process, I am happy with the resulting topics. I might have wished for a clear problem formulation at an earlier stage, but the long process of finding a suitable area of work was very interesting and instructive as well. I want to thank my supervisors Nils Lid Hjort and Arnolfo Frigessi for showing such great interest in my project during the last two years, even before my work had officially started, and in particular for using so much time trying to find a good project. The many months spent on reading various articles were tough, but I am grateful for their belief in me and their high ambitions on my behalf.

I particularly want to thank my main supervisor Nils Lid Hjort for taking the time to discuss my progress almost every week, and for always answering my emails very quickly and accurately. The feedback I have received from both my supervisors the last couple of weeks have been very useful, and I deeply appreciate all their help. Finally I want to thank Arnolfo Frigessi for remaining my co-supervisor during his stay in Italy, and for taking an active part in my project through his correspondence with my main supervisor Nils Lid Hjort.

Chapter 1

Introduction

Parameter estimation is one of the main topics of statistics. Often we have a set of observations $y_1, \dots, y_n$ which we assume to be stemming from some parametric distribution $f(y|\theta)$, and we would like to use these observations in estimating θ . Although both the observations and θ itself may be multidimensional, we will mainly consider one-dimensional data while allowing θ to be multidimensional.

Methods for estimating θ have been, and still are, a large research area. A large variety of estimators have been proposed, and even a first course in statistics will often present at least three. Popular choices include the method of moments, the maximum likelihood estimator and the family of Bayesian estimators.

The method of moments is based on the idea that if the population moments $E(X^k)$ of a parametric distribution may be expressed in terms of the elements of θ , the elements of θ may equally well be expressed in terms of the of population moments. Replacing the unobservable population moments with the observed sample moments $\frac{1}{n} \sum_{i=1}^n X_i^k$, we obtain an estimate $\hat{\theta}$. This method is both conceptually and computationally simple, but suffers from the drawback that it may sometimes produce estimates outside the proper range, for example estimated probabilities outside $[0, 1]$. In such cases, modifications need to be made in order to obtain meaningful estimates.

The Bayesian approach to parameter estimation is rather different from the method of moments. Here the parameter estimate is a result not only of the data, but also of a certain prior distribution $\pi(\theta)$ quantifying our a priori beliefs concerning the value of θ . The prior distribution and the likelihood function is combined in the following way, giving what is called the posterior distribution

$$\pi(\theta|\text{data}) \propto L(\theta)\pi(\theta).$$

Whereas the method of moments only presents us with a point estimate of θ , the Bayesian approach gives us a complete distribution (after properly normalizing the expression above to make sure it integrates to one). Using this posterior distribution, a point estimate $\hat{\theta}$ may be chosen in many different ways dependent on the choice of loss function. One example of a widely used Bayesian estimator is the posterior mean of β .

A third and very popular estimator is the maximum likelihood estimator. Here, the point estimate $\hat{\theta}$ is the value of θ which maximizes the likelihood function $L_n(\theta)$, or equivalently the log likelihood function $l_n(\theta)$. It has been shown

that when the maximum likelihood estimator exists, it is asymptotically normally distributed under some fairly general assumptions (see e.g. Lehmann (1999)). It is also consistent and asymptotically optimal in the sense that the variance of the limiting distribution is minimal. The asymptotic normality makes it easy to construct confidence intervals and joint confidence regions for the parameters, and together these properties make the estimator an attractive choice.

In the Bayesian framework, we have the possibility of specifying the expected range of θ by using a suitable prior distribution. A similar opportunity is provided by the family of regularized maximum likelihood estimators. Consider a non-negative function $\text{pen}(\theta)$ chosen such that it is small when θ is in the expected range and large otherwise. Instead of maximizing the ordinary log likelihood we may then maximize the penalized log likelihood function given by

$$H_n(\theta) = l_n(\theta) - \text{pen}(\theta). \quad (1.1)$$

The last term works as a penalty, penalizing us for choosing what is believed to be unsuitable values of θ . The resulting estimate will usually differ from the ordinary maximum likelihood estimate, in the sense that it is being pulled towards the θ region we have considered to be appropriate. Again we have a compromise between data and prior belief, and it is in fact closely related to the mode of the Bayesian posterior distribution, as will be seen later.

The penalty function $\text{pen}(\theta)$ may have many different forms, but we will focus on the popular version given by

$$\text{pen}(\theta) = \lambda \sum_{j=1}^p |\theta_j|^\gamma,$$

for $\theta = (\theta_1, \dots, \theta_p)$, with the tuning parameters λ and γ being non-negative.

Penalized maximum likelihood estimators may be applied to a wide range of estimation problems, but in this thesis we will mainly focus on linear regression. We assume that the observations $y = (y_1, \dots, y_n)$ are stemming from the model

$$y_i = x_i^T \beta + \varepsilon_i.$$

Here the x_i 's are fixed covariate vectors, and the ε_i 's are usually assumed to be i.i.d. with $\varepsilon_i \sim N(0, \sigma^2)$. We are mainly interested in the estimation of β , and for simplicity we will therefore assume σ^2 to be a known constant.

Using the penalty function above, the penalized log likelihood function is now given by

$$H_n(\beta) = l_n(\beta) - \lambda \sum_{j=1}^p |\beta_j|^\gamma.$$

When $d \geq 1$ this estimation method is called bridge regression (see e.g. Fu (1998)). In these cases we get

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \left\{ l_n(\beta) - \lambda \sum_{j=1}^p |\beta_j|^\gamma \right\}.$$

Our penalty function involves a non-negative number γ called the shrinkage parameter, and I will briefly discuss a few possible choices of its value. First of

all, if we let $\gamma \rightarrow 0$, we see that in the limit, $\text{pen}(\beta) = \lambda \sum (1 - \delta_0(|\beta_j|))$. Thus for $\gamma = 0$, the penalty is of the form λk , where k is the number of non-zero elements of β . Since the penalty term is independent of the actual value of any non-zero β_j , this corresponds to the best variable subset selection with a penalty proportional to the number of non-zero elements. This corresponds to the Akaike Information Criterion (AIC, see e.g. Claeskens and Hjort (2008)) if we let $\lambda = 1$.

Another popular choice is $\gamma = 2$, leading to ridge regression. Ridge regression was introduced by Hoerl and Kennard (1970), and has the advantage that the estimate of β has an explicit formula. In fact,

$$\hat{\beta}_{\text{ridge}}(\lambda) = (X^T X + 2\lambda\sigma^2 I_p)^{-1} X^T y,$$

if we use the parameterization above. It is easily seen that the ridge estimate is closely connected to the ordinary least squares estimate. The elements of the estimate are pulled towards zero, but none of them are set equal to zero. Hoerl and Kennard (1970) showed that a ridge estimate of β may have smaller MSE than the ordinary least squares estimate, especially when the columns of X are not orthogonal.

Ridge regression has another very attractive property. When the columns of X are (nearly) collinear, the least squares estimate is undefined (or at least very unstable) since $X^T X$ is (nearly) singular. Adding a small amount to each diagonal element makes the resulting matrix non-singular, and the ridge estimate is always defined. In the usual setting with small p and large n the columns are usually not collinear, but as soon as p gets bigger than n , $X^T X$ is always singular, and ridge regression is a good alternative to the undefined least squares estimate.

Even though the ridge estimate exists for $p > n$, it is very variable. An alternative in this situation is to let $\gamma = 1$. This gives us an estimate with some elements set equal to zero and the rest shrunk towards zero, named the 'least absolute shrinkage and selection operator' (LASSO) by Tibshirani (1996). As opposed to the $\gamma = 0$ case, this gives rise to a continuous penalized likelihood function. This allows us to do the selection and the shrinkage in one operation, without running through every possible sub-model, and gives us a continuous and thus more stable estimate. The choice of $\gamma = 1$ is the only one giving both subset selection and continuity, as we need $\gamma \leq 1$ to ensure sparsity (β_j 's being set equal to zero) and $\gamma > 1$ to get a continuous solution Fan and Li (2001).

The penalized likelihood function for LASSO is not smooth, and may be hard to maximize numerically, but several algorithms have been proposed, and numerous implementations of LASSO exists for R and other statistical software. The LASSO has become a popular way of treating high-dimensional data, and is also being used outside the linear regression setting, for example in estimating precision matrices from multinormal data.

In this thesis we will mainly be working with the linear regression model and various methods for estimating the parameter vector β . We begin with a short presentation of the least squares estimator and the best variable subset selection in Chapter 2. We then move on to ridge regression, focusing both on the β -estimator itself and on a number of different methods for estimating the tuning parameter λ . In this chapter we will also present our main numerical example. In Chapter 3 we introduce a new estimator based on ridge regression, having

the ability to perform variable selection as well as shrinkage. Here the main focus will be on the β -estimator, but we will also present three possible ways of estimating its two tuning parameters. As an alternative estimator performing both shrinkage and model selection, the lasso will be presented in Chapter 5. Chapter 6 is mainly a summary of the articles I have read concerning the estimation of (inverse) covariance matrices, with a few ideas for further work being described. Finally, in Chapter 7 we sum up the most important points from the previous chapters, and provide a couple of ideas for further research.

Chapter 2

The least squares estimator

The main part of this thesis concerns regularized parameter estimation in the multiple linear regression setting. To make the presentation complete, I will start with a brief presentation of the most widely used estimator in this setting, the least squares estimator. Most of the material in this chapter will probably be well-known to the reader, but reading it is nonetheless recommended, as the later chapters will follow the outline and notation of this one very closely.

2.1 Model assumptions

We assume to have observations $y_1, \dots, y_n$ stemming from the linear model

$$y_i = x_i^T \beta + \varepsilon_i.$$

Here, the vectors $x_1, \dots, x_n$ are known covariate vectors, while β_0 is some unknown p -dimensional coefficient vector. The ε_i 's are unobserved random error terms. Rather than using the individual equations, we often state the above relation as

$$y = X\beta + \varepsilon,$$

where $y = (y_1, \dots, y_n)^T$, $X = (x_1, \dots, x_n)^T$ is an $n \times p$ matrix and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$. We assume the error terms to be independent and normally distributed with mean zero and known variance σ^2 , that is,

$$\varepsilon \sim N_n(0, \sigma^2 I_n).$$

Finally, we assume the error terms to be independent of X .

2.2 The least squares estimator and the maximum likelihood estimator

Usually the goal of a linear regression analysis is to find a good point estimate $\hat{\beta}$ of the coefficient vector β_0 . A conceptually and computationally simple, and thus very common, estimator is the (ordinary) least squares estimator. The idea is to find the value of β that minimizes the sum of squared residuals, that is,

$$\hat{\beta}_{LS} = \underset{\beta}{\operatorname{argmin}} \{ (y - X\beta)^T (y - X\beta) \}.$$

It is easily shown that the least squares estimate $\hat{\beta}_{LS}$ is given by

$$\hat{\beta}_{LS} = \Sigma_n^{-1} \frac{X^T Y}{n}, \quad (2.1)$$

with

$$\Sigma_n = X^T X / n = \frac{1}{n} \sum_{i=1}^n x_i x_i^T.$$

Another popular estimator is the maximum likelihood estimator, which requires a parametric density for y . In our case of normally distributed error terms we know that

$$Y \sim N_n(X\beta, \sigma^2 I_n),$$

leading to the likelihood

$$L_n(\beta) = \frac{1}{\sqrt{2\pi}^n} \frac{1}{\sigma^n} \exp \left\{ -\frac{1}{2} \frac{(y - X\beta)^T (y - X\beta)}{\sigma^2} \right\}.$$

From this likelihood we immediately see that the ordinary least squares estimator and the maximum likelihood estimator coincide in this particular case, giving

$$\hat{\beta}_{MLE} = \hat{\beta}_{LS} = \Sigma_n^{-1} \frac{X^T Y}{n}.$$

The maximum likelihood/least squares estimate will be important throughout the thesis, and we will denote it by $\hat{\beta}_{LS}$ and refer to it as the (ordinary) least squares estimate. The matrix Σ_n and several modified versions of it will also be used extensively in later chapters.

The distributional properties of our least squares estimator are easily found due to its simple form, and we have

$$\hat{\beta}_{LS} \sim N_p \left(\beta, \frac{\sigma^2}{n} \Sigma_n^{-1} \right),$$

leading to marginal distributions of the following form:

$$\hat{\beta}_{LS,j} \sim N \left(\beta_j, \frac{\sigma^2}{n} [\Sigma_n^{-1}]_{(j,j)} \right). \quad (2.2)$$

From this we see that the least squares estimator is unbiased and consistent (as long as $\Sigma_n^{-1} = o(n)$, a fairly mild condition). We will discuss further properties and limitations of the estimator in the last section of this chapter, but first we want to connect the least squares estimator to a couple of other concepts.

2.3 The least squares estimator from a Bayesian perspective

A common mistake made by beginners is to reverse the distributional result above into

$$\beta \sim N_p \left(\hat{\beta}_{LS}, \frac{\sigma^2}{n} \Sigma_n^{-1} \right).$$

To be able to state such a result about the real unobserved β_0 would be very useful, but is unfortunately not possible following the ideas presented in the previous section. However, results like this concerning the real parameter vector may be obtained using Bayesian statistics, and in this and the next section we will look at a few aspects of this possibility.

In Bayesian statistics we do obtain distributional results about the real parameters, but this comes at a cost. The distribution of the parameter given the data, called the posterior distribution, can only be found by combining the likelihood function and a prior distribution $\pi(\beta)$ in a certain way:

$$\pi(\beta|\text{data}) = \frac{L_n(\beta)\pi(\beta)}{\int L_n(\beta)\pi(\beta)d\beta}. \quad (2.3)$$

We thus have to specify a prior distribution of β , quantifying our beliefs about β before actually seeing the data. It is not always clear how this distribution should be chosen, and without clear prior beliefs, one is often tempted to choose what is called a flat or uninformative prior. Such a prior allows all parameter values in a large set or interval to have approximately the same probability (or approximately the same value of the density function). The resulting posterior distribution will be almost entirely a result of the observed data alone.

The most uninformative prior distribution we can think of is a uniform distribution over the entire real line. Of course, this is not really a distribution, but as long as the integral in (2.3) is finite, we nonetheless obtain a meaningful posterior distribution. Such prior distributions are called improper priors and they are being used from time to time.

If we allow ourselves to use improper priors, we find that by using

$$\pi_1(\beta) = 1,$$

we get

$$\pi(\beta|\text{data}) = \frac{L_n(\beta)}{\int L_n(\beta)d\beta} \propto \exp \left\{ -\frac{n}{2} \frac{1}{\sigma^2} (\beta - \hat{\beta}_{LS})^T \Sigma_n (\beta - \hat{\beta}_{LS}) \right\}.$$

Now the posterior distribution of β is given by

$$\beta \sim N_p \left(\hat{\beta}_{LS}, \frac{\sigma^2}{n} \Sigma_n^{-1} \right),$$

with marginal distributions

$$\beta_j \sim N_p \left([\hat{\beta}_{LS}]_j, \frac{\sigma^2}{n} [\Sigma_n^{-1}]_{(j,j)} \right).$$

This is the same as the 'incorrect' result from the beginning of this section, this time derived by the use of Bayesian statistics and an improper prior.

In Bayesian statistics, we are often interested in the posterior distribution itself, but since we want to link these ideas to the least squares estimator, we are mainly interested in a sensible point estimate of β . Therefore, our next question is what this point estimate should be. Two possible Bayesian estimators are the mean and the median of the posterior distribution, and in this case they are both equal to $\hat{\beta}_{LS}$. Thus the least squares estimate is not only the maximum

likelihood estimate, but also a Bayesian estimate obtained by using the improper prior $\pi_1(\beta) = 1$.

Another possible choice is the posterior mode, closely linked to the maximum likelihood estimator. This is not a real Bayesian estimator, but it is called a maximum a posteriori probability (MAP) estimator and it is being used from time to time. In the case of our normal posterior distribution, the resulting point estimate is also equal to $\hat{\beta}_{LS}$.

2.4 The least squares estimator and model selection

In a regression setting we often have several possible covariate vectors to choose from, but we may not want to use all of them. There may be several reasons for this, the most important ones being model simplicity and stable parameter estimates, since estimating more parameters makes each estimate more variable. With p possible covariates, there are 2^p different subsets $S \subset \{1, \dots, p\}$ to consider, each corresponding to a different linear model with $|S|$ non-zero parameters.

Adding more (non-zero) parameters will never decrease the likelihood, so using the maximum likelihood alone as a selection criterion would always result in choosing the largest model, in our case the model with $|S| = p$ non-zero parameters. In order to be able to choose a model different from the full model, we therefore need a different model selection criterion. One simple and widely used criterion is Akaike's information criterion (AIC, see e.g. Claeskens and Hjort (2008)), where we choose the model maximizing

$$\text{AIC}(M) = 2 \log\text{-likelihood}_{\max}(M) - 2 \dim(M). \quad (2.4)$$

Here we introduce a penalty term proportional to the number of parameters, having the effect that an extra parameter is only added if it increases the log-likelihood by at least one. Another popular model selection criterion with a heavier penalty is the Bayesian information criterion (BIC), given by

$$\text{BIC}(M) = 2 \log\text{-likelihood}_{\max}(M) - (\log n) \dim(M). \quad (2.5)$$

Both the AIC and the BIC usually result in final models smaller than the full model, and for large n , the model chosen using BIC will often be much smaller than the one selected by AIC.

Comparing the AIC or BIC of all possible submodels is only feasible when p is small, since there are 2^p possible models to consider. When p is too large for this, one will typically perform a greedy forward or backward search. In a greedy forward search, we start with the empty set and add the parameter that increases the AIC the most. This way we consider first p possible models, then $p - 1$ and so on, thereby reducing the complexity of the process from $O(2^p)$ to $O(p^2)$, since we have to fit at most $p(p + 1)/2$ models. At some point none of the remaining parameters increase the AIC, and we stop the process and let the current model be the final one. Greedy backward search is a similar process, starting with the full set of parameters and then removing one parameter at the time until the AIC can not longer be increased. Clearly, at some point p

becomes too large even for this simpler process, and other methods have to be used, but such methods will not be discussed here.

An alternative to AIC/BIC in this situation involving several possibly unimportant parameters would be to quantify our beliefs in the form of a prior distribution on the full p -dimensional β . We have already more or less assumed that a certain fraction of the parameters are really zero or close to zero, while all the other parameters are free to take on any real non-zero value. A natural choice would be the improper prior distribution

$$\pi_\alpha(\beta) = \prod_{j=1}^p [\alpha \delta_0(\beta_j) + (1 - \alpha)(1 - \delta_0(\beta_j))\pi(\beta_j)]. \quad (2.6)$$

This prior distribution makes all the p components of β independent and identically distributed with probability α of being zero, and probability $1 - \alpha$ of having any other real value, and thus seems to be exactly what we want. Here $\pi(\beta_j) = c$ for some constant c is an improper prior like in the previous section. There we let $c = 1$ since the value was obviously not important. However, in this situation it turns out that the value of c will have a large effect on the resulting estimate of β , and we have no reasons for using any particular c , so we will not pursue this idea any further.

The problems concerning (2.6) are all due to the improper distribution $\pi(\beta_j) = c$. We could therefore consider using a sufficiently flat prior distribution instead, giving almost the same effect. One possibility is to let $\beta_j \sim N(0, \tau^2)$ with τ very large. The consequences of using such a prior will be described in chapter 4.

2.5 Discussion

The ordinary least squares estimator is computationally simple, unbiased and consistent. Also, under our assumption of i.i.d. normally distributed error terms, it equals the maximum likelihood estimator, and therefore has all the desirable properties of such estimators. However, there may be reasons for choosing other estimators instead of the least squares estimator. We have already seen that it may sometimes be desirable to set some of the coefficients to zero, thereby introducing a small bias, but at the same time reducing the total variance of the final estimate. In the next section we will see how shrinkage of the coefficients has the same effect, and in Chapter 4 and 5 we will introduce two different estimators, each of them performing both shrinkage and model selection. If these methods are able to reduce the variance so much that the mean square error of the estimator becomes smaller than in the least square case, then we should consider using them as long as the goal is not inference about the coefficients themselves. Clearly, in this case we do not want biased estimators, since we do not know how large the bias is.

The variability of the least squares estimate may be very large when the predictors are highly correlated. However, if they are perfectly correlated, or $p > n$ such that the rank of X is smaller than p , the least squares estimate does not even exist. In these cases we clearly need to use alternative estimators. Finally, we know that the least squares estimator may be influenced too heavily by one extreme observation, and various robust alternatives have been developed

to avoid this problem. All these problems make it reasonable to look for other estimators that may solve one or more of them, and the first method we will consider is the ridge regression.

Chapter 3

Ridge regression

The ordinary least squares estimate has been shown to perform well when Σ_n is close to the identity matrix. However, in other situations, it may be unsatisfactory. We have already seen a modified version with a number of parameter estimates set equal to zero in order to reduce the variability of the remaining ones, and there are several other examples of estimators introducing a small bias in order to reduce the mean square error. In this chapter we will discuss one of them, the ridge regression described by Hoerl and Kennard (1970).

We start by presenting ridge regression in its original form, following the lines of Hoerl and Kennard closely, but using our own notation. Again we let $y = X\beta + \varepsilon$, with X being an $n \times p$ -matrix of rank p . We assume $p < n$ and $\varepsilon \sim N_n(0, \sigma^2 I_n)$. Letting $\Sigma_n = X^T X/n$ we have already seen that this leads to $\hat{\beta}_{LS} = \Sigma_n^{-1} \frac{X^T y}{n}$ with $E(\hat{\beta}_{LS}) = \beta$ and $Var(\hat{\beta}_{LS}) = \frac{\sigma^2}{n} \Sigma_n^{-1}$. If we let L_1 be the distance from $\hat{\beta}_{LS}$ to β , defined as

$$L_1 = (\hat{\beta}_{LS} - \beta)^T (\hat{\beta}_{LS} - \beta),$$

we also find that

$$E(L_1^2) = \frac{\sigma^2}{n} Tr(\Sigma_n^{-1}), \quad (3.1)$$

and

$$Var(L_1^2) = \frac{2\sigma^4}{n^2} Tr(\Sigma_n^{-2}).$$

This demonstrates that the size of the square error of $\hat{\beta}_{LS}$ is closely connected to the form of the matrix Σ_n , or more precisely its eigenvalues. If we denote the eigenvalues of Σ_n by

$$\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p = \lambda_{\min} > 0,$$

then the expected distance from $\hat{\beta}_{LS}$ to β , the mean square error, is given by

$$E(L_1^2) = \frac{\sigma^2}{n} \sum_{j=1}^p (1/\lambda_j),$$

whereas the variance of the square error is given by

$$Var(L_1^2) = \frac{2\sigma^4}{n^2} \sum_{j=1}^p (1/\lambda_j)^2.$$

These equations clearly show that both the expected square error and the variance of the square error are large when $X^T X$ has one or more small eigenvalues. In particular, Σ_n nearly singular will often lead to large square errors. If we could somehow increase the eigenvalues, the mean and variance of the square error would clearly decrease, and this is done by adding a small amount λ to each diagonal element of Σ_n . The ridge estimate is given by

$$\hat{\beta}_{\text{ridge}}(\lambda) = (X^T X + \lambda I_p)^{-1} X^T y, \quad (3.2)$$

with $\lambda \geq 0$, or equivalently

$$\hat{\beta}_{\text{ridge}}(\lambda) = \Sigma_{n,\text{ridge}}^{-1}(\lambda) \frac{X^T y}{n},$$

with $\Sigma_{n,\text{ridge}}(\lambda) = \Sigma_n + \frac{\lambda}{n} I_p$. The last equation is very close to (2.1), and the ridge estimate is simply a scaled version of the least squares estimate:

$$\hat{\beta}_{\text{ridge}}(\lambda) = \left(I_p + \frac{\lambda}{n} \Sigma_n^{-1} \right)^{-1} \hat{\beta}_{LS},$$

where the matrix $(I_p + \frac{\lambda}{n} \Sigma_n^{-1})^{-1}$ is clearly symmetric and positive definite.

Dependent on the size of the penalty parameter $\lambda \geq 0$, the least squares estimate is shrunk towards zero by a certain amount, always being smaller than or equal to the least squares estimate measured in the l_2 norm. This is easily seen by noting that

$$\hat{\beta}_{\text{ridge}}^T(\lambda) \hat{\beta}_{\text{ridge}}(\lambda) = \hat{\beta}_{LS}^T \left(I_p + \frac{\lambda}{n} \Sigma_n^{-1} \right)^{-2} \hat{\beta}_{LS} \leq \xi_{\max}^2 \hat{\beta}_{LS}^T \hat{\beta}_{LS}, \quad (3.3)$$

where $\xi_{\max}$ is the largest eigenvalue of $(I_p + \frac{\lambda}{n} \Sigma_n^{-1})^{-1}$, given by $\xi_{\max} = \lambda_1 / (\lambda_1 + \lambda/n) \leq 1$. Thus we get equality if and only if $\lambda = 0$, and we see that the estimate approaches zero as $\lambda \rightarrow \infty$.

Of course the residual sum of squares is larger for any other estimate than it is for the least squares estimate. Actually, for a given estimate B of β , we have

$$\text{RSS}_B = \text{RSS}_{LS} + (B - \hat{\beta}_{LS})^T X^T X (B - \hat{\beta}_{LS}).$$

We see that the residual sum of squares is constant on the surface of hyperellipsoids centered at $\hat{\beta}_{LS}$. From (3.1) we get that

$$E[\hat{\beta}_{\text{ridge}}^T(\lambda) \hat{\beta}_{\text{ridge}}(\lambda)] = \beta^T \beta + \frac{\sigma^2}{n} \text{Tr}(\Sigma_n^{-1}),$$

and it seems that if we are willing to move away from the least squares estimate, we should do so in a direction that shortens the estimated vector. Hoerl and Kennard show that the ridge estimate is in fact the B with minimum length subject to

$$(B - \hat{\beta}_{LS})^T X^T X (B - \hat{\beta}_{LS}) = \phi_0$$

for a certain value of ϕ_0 . Thus, the ridge estimate is also the solution of a constrained optimization problem. The relation between λ and ϕ_0 is given by

$$\hat{\beta}_{\text{ridge}}(\lambda) = \underset{B}{\text{argmin}} \left\{ B^T B + (1/\lambda) (B - \hat{\beta}_{LS})^T X^T X (B - \hat{\beta}_{LS}) - \phi_0 \right\}, \quad (3.4)$$

and they point out that even though the derivation of the estimate is based on minimization with ϕ_0 fixed, and with $1/\lambda$ as a Lagrange multiplier, it is easier to first choose λ and then compute the corresponding ϕ_0 .

As the idea leading to the ridge estimate was to reduce the mean square error, we should use some time examining the square error of our new estimate. We let

$$L_1^2(\lambda) = (\hat{\beta}_{\text{ridge}}(\lambda) - \beta)^T (\hat{\beta}_{\text{ridge}}(\lambda) - \beta),$$

and find that the mean square error is given by

$$E[L_1^2(\lambda)] = \frac{\sigma^2}{n} \sum_{j=1}^p \frac{\lambda_j}{(\lambda_j + \lambda/n)^2} + \frac{\lambda^2}{n^2} \beta^T \Sigma_{n,\text{ridge}}^{-2}(\lambda) \beta. \quad (3.5)$$

The first term can be shown to equal the sum of the variances of the parameter estimates, while the second term can be considered the square of a bias introduced when $\hat{\beta}_{\text{ridge}}(\lambda)$ is used instead of the least squares estimate. The first term decreases when λ increases, while the second term increases with λ . The details may be found in Hoerl and Kennard (1970), where it is also shown that it is possible to choose a $\lambda > 0$, and thereby accept a small bias, such that the variance is reduced by an amount large enough to decrease the mean square error. In particular, they state and prove the following result (using a slightly different notation):

Theorem 3.0.1 (Existence Theorem). *There always exists a $\lambda > 0$ such that*

$$E[L_1^2(\lambda)] < E[L_1^2(0)] = \frac{\sigma^2}{n} \sum_{j=1}^p (1/\lambda_j).$$

One example of such a situation is demonstrated in 3.1. Here $p = 5$, $n = 15$ and $\sigma = 2$, and Σ_n has one large and four very small eigenvalues. We see that the mean square error of the ridge estimate may be much smaller than the MSE of the least squares estimate. We also note that a smaller MSE is obtained for quite a broad range of λ -values. This is important, since we will typically estimate the value of λ , and we do not want the λ value to fall outside the interval too often.

We have now presented the original motivation for the ridge estimate, and before we move on to studying the estimate itself and the choice of λ , we will note that the ridge estimate is connected to penalized maximum likelihood estimation and that it may also be regarded as a Bayesian estimate. First, from (3.4) it is easily seen that

$$\begin{aligned} \hat{\beta}_{\text{ridge}}(\lambda) &= \underset{\beta}{\operatorname{argmin}} \{ (y - X\beta)^T (y - X\beta) + \lambda \|\beta\|_2^2 \} \\ &= \underset{\beta}{\operatorname{argmax}} \left\{ l_n(\beta) - \frac{\lambda}{2\sigma^2} \|\beta\|_2^2 \right\}, \end{aligned} \quad (3.6)$$

where the last expression is the penalized maximum log-likelihood with penalty parameter $\lambda/(2\sigma^2)$. Here, the constant ϕ_0 is no longer present in the equation, but there is still an underlying unique ϕ_0 corresponding to each specified value of λ . Hoerl and Kennard also point out that the ridge estimate may be considered as the posterior mean based on giving the regression coefficients β a normal

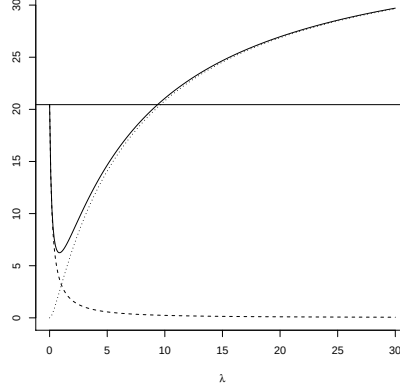


Figure 3.1: The mean square error of the ridge estimate as a function of λ . The dashed and dotted curves depict the variance and bias terms of (3.5), whereas the solid curve shows the total MSE. The horizontal line corresponds to the MSE of the least squares estimate.

prior distribution with mean zero and covariance matrix proportional to I_p/λ . These two alternative interpretations will be important in the remaining part of this chapter, and will be paralleled in later chapters where the estimates themselves have no simple closed form.

3.1 The ridge estimate $\hat{\beta}_{\text{ridge}}(\lambda)$

Of course the ridge estimate depends heavily on the value of λ , and we have not yet discussed how this parameter should be chosen. Estimating λ will be the main topic of this chapter, but since the target of the estimation is usually the β -estimate or some function of it, rather than the $\hat{\lambda}$ itself, we will first study the behaviour of $\hat{\beta}_{\text{ridge}}(\lambda)$ for a fixed value of λ .

In this section we will use the penalized maximum likelihood interpretation of the ridge estimate, and we recall that this corresponds to penalizing parameter estimates for being far from a certain anticipated region, in this case far from zero. First of all we will state the penalized log-likelihood function H_n , given by

$$H_n(\beta) = l_n(\beta) - \frac{1}{2\tau^2} \|\beta\|^2. \quad (3.7)$$

Here we have chosen a different parametrization of the penalty parameter in order to avoid complicated expressions later, letting $\lambda = \sigma^2/\tau^2$.

Maximizing (3.7) is equivalent to minimizing

$$\sum_{i=1}^n (y_i - x_i^T \beta)^2 + \frac{\sigma^2}{\tau^2} \|\beta\|^2,$$

and we find that

$$\hat{\beta}_{\text{ridge}}(\tau) = \left(X^T X + \frac{\sigma^2}{\tau^2} I_p \right)^{-1} X^T y. \quad (3.8)$$

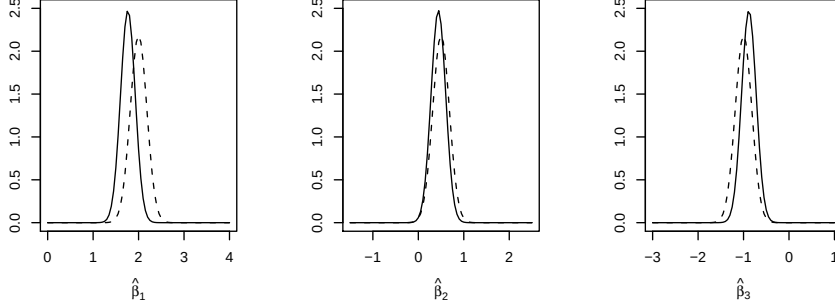


Figure 3.2: The distribution of the ridge estimate (solid curve) and the least squares estimate (dashed curve) when $\rho = 0$, corresponding to $\Sigma_n = I_3$.

Parallel to the expressions obtained for λ , we may also state this result as

$$\hat{\beta}_{\text{ridge}}(\tau) = \Sigma_{n,\text{ridge}}^{-1}(\tau) \frac{X^T y}{n},$$

with $\Sigma_{n,\text{ridge}}(\tau) = \Sigma_n + \frac{1}{n} \frac{\sigma^2}{\tau^2} I_p$.

Using the fact that $\hat{\beta}_{\text{ridge}}(\tau) = \Sigma_{n,\text{ridge}}^{-1}(\tau) \Sigma_n \hat{\beta}_{LS}$ we find that

$$\hat{\beta}_{\text{ridge}}(\tau) \sim N_p \left(\Sigma_{n,\text{ridge}}^{-1}(\tau) \Sigma_n \beta, \frac{\sigma^2}{n} \Sigma_{n,\text{ridge}}^{-1}(\tau) \Sigma_n \Sigma_{n,\text{ridge}}^{-1}(\tau) \right)$$

This expression shows us how a small bias is introduced, while the variance is reduced compared to the least squares estimate. In fact, the sum of the individual variances equals the first term of (3.5) with $\lambda = \sigma^2/\tau^2$, in which we clearly see how the variance is reduced when τ decreases (λ increases).

The marginal distribution of the individual parameter estimates is given by

$$\hat{\beta}_{\text{ridge},j}(\tau) \sim N_p \left(\left[\Sigma_{n,\text{ridge}}^{-1}(\tau) \Sigma_n \beta \right]_j, \frac{\sigma^2}{n} \left[\Sigma_{n,\text{ridge}}^{-1}(\tau) \Sigma_n \Sigma_{n,\text{ridge}}^{-1}(\tau) \right]_{(j,j)} \right).$$

This expression is not as simple as the one obtained for $\hat{\beta}_{LS,j}$ in (2.2). It is however easily computable, and two examples are given in Figure 3.2 and Figure 3.3. The figures show the marginal distribution of each parameter estimate when $\beta = (2, 0.5, -1)$, $\sigma^2 = 1$, $n = 30$ and Σ_n is the matrix given by $[\Sigma_n]_{(i,j)} = \rho^{|i-j|}$. In Figure 3.2 we use $\rho = 0$, giving the identity matrix, while in Figure 3.3 we use $\rho = 0.9$, giving a rather ill-conditioned matrix. The solid lines show the marginal distribution of the ridge estimate for $\tau = 1/2$, while the dashed lines show the marginal distribution of the ordinary least squares estimates.

Looking at the figures, we first observe how different the least squares estimate performs in the two cases. The estimate is much more variable in the second figure, as discussed in the introduction of this chapter. Also, we see that the ridge estimate is less variable than the least squares estimate, and that it is expected to be closer to zero and thus biased. Finally, these two figures clearly show that the bias and the decrease in variance depends on the form of Σ_n .

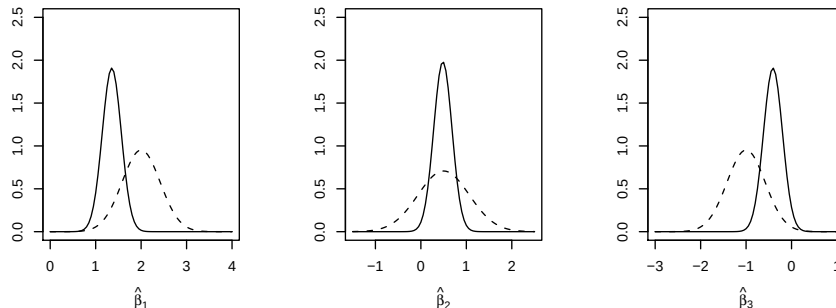


Figure 3.3: The distribution of the ridge estimate (solid curve) and the least squares estimate (dashed curve) when $\rho = 0.9$, corresponding to a rather ill-conditioned Σ_n .

One striking feature of 3.2 is the fact that the ridge estimate is so close to the least squares estimate. In the case of $\Sigma_n = I_p$ we may actually obtain very simple formulas describing the relation between the two estimates. Using (3.8) we find that

$$\hat{\beta}_{\text{ridge}}(\tau) = \frac{1}{1 + \frac{1}{n} \frac{\sigma^2}{\tau^2}} \hat{\beta}_{LS}.$$

Thus in this case each component of the ridge estimate is simply the corresponding component of the least squares estimate shrunk by a constant factor, and that the difference between the two estimates are in fact very small even for small n .

We have already discussed the interpretation of penalized maximum likelihood estimates. Penalizing parameters proportional to some function of their absolute value results in final estimates closer to zero than ordinary maximum likelihood estimates. The choice of penalty parameter is important, as it quantifies how much we expect the true parameters to be spread out. The penalty function gives even more accurate information about our expectations, allowing us to penalize differently inside the expected range as well. Another way of including our prior beliefs in the analysis is of course to use Bayesian techniques. The ridge estimate or the penalized maximum likelihood estimate is in fact also a Bayesian estimate found by using a certain normal prior, as pointed out by Hoerl and Kennard (1970) in their first ridge regression article. Before we move on to estimating τ , we will therefore take a closer look at the Bayesian interpretation of the ridge estimate.

Using (3.6), we observe that

$$\begin{aligned} \hat{\beta}_{\text{ridge}}(\lambda) &= \operatorname{argmax}_{\beta} \left\{ l_n(\beta) - \frac{\lambda}{2\sigma^2} \|\beta\|_2^2 \right\} \\ &= \operatorname{argmax}_{\beta} \left\{ l_n(\beta) - \frac{1}{2\tau^2} \|\beta\|_2^2 \right\} \\ &= \operatorname{argmax}_{\beta} \left\{ L_n(\beta) \frac{1}{\sqrt{2\pi^p}} \frac{1}{\tau^p} \exp \left\{ -\frac{1}{2} \frac{\beta^T \beta}{\tau^2} \right\} \right\}, \end{aligned}$$

which we recognize as the mode of the posterior distribution when the prior distribution is given by

$$\pi_\tau(\beta) = \frac{1}{\sqrt{2\pi^p}} \frac{1}{\tau^p} \exp \left\{ -\frac{1}{2} \frac{\beta^T \beta}{\tau^2} \right\},$$

that is,

$$\beta \sim N_p(0, \tau^2).$$

A normal prior distribution and a normal likelihood results in a normal posterior distribution, so clearly, the mode is also the mean, and we get $\hat{\beta}_{\text{ridge}}(\lambda) = E(\beta|\text{data})$. Thus, the ridge estimate is a Bayesian estimate as well.

Being Bayesian, we should not only consider the point estimate of β , but also the posterior distribution of β . We find that

$$\begin{aligned} \pi_\tau(\beta|\text{data}) &\propto L_n(\beta)\pi_\tau(\beta) \\ &\propto \exp \left\{ -\frac{n}{2\sigma^2} \left[\beta^T \left(\Sigma_n + \frac{1}{n} \frac{\sigma^2}{\tau^2} I_p \right) \beta - 2 \frac{y^T X}{n} \beta \right] \right\} \\ &\propto \exp \left\{ -\frac{n}{2\sigma^2} (\beta - \hat{\beta}_{\text{ridge}}(\tau))^T \Sigma_{n,\text{ridge}}(\tau) (\beta - \hat{\beta}_{\text{ridge}}(\tau)) \right\}, \end{aligned}$$

where $\hat{\beta}_{\text{ridge}}(\tau)$ and $\Sigma_{n,\text{ridge}}(\tau)$ are the same as before. Thus

$$\beta|\text{data} \sim N_p \left(\hat{\beta}_{\text{ridge}}(\tau), \frac{\sigma^2}{n} \Sigma_{n,\text{ridge}}^{-1}(\tau) \right),$$

with our now familiar estimate $\hat{\beta}_{\text{ridge}}(\tau)$ as posterior mean, mode and median. This distribution could be used to derive other estimates of β , but since this chapter is about the ridge estimate, we will not pursue this idea here, but simply note that both the posterior mean and the posterior mode equals the ridge estimate. Of course, the distribution of the estimate itself is exactly the same as before, this Bayesian approach is just a different idea leading to the same estimate.

3.2 Estimating the penalty parameter τ

All the results regarding the ridge estimate $\hat{\beta}_{\text{ridge}}(\tau)$ depend heavily on the choice of τ . It is not immediately clear how this value should be chosen, and in this chapter we will study a few possible ways of obtaining a suitable $\hat{\tau}$. In order to say that one value of $\hat{\tau}$ is better than another, we need some sort of criterion, and one possible and well-known alternative is the minimum sum of squared residuals, giving

$$\hat{\tau} = \underset{\beta}{\operatorname{argmin}} (y - X\hat{\beta}_{\text{ridge}}(\tau))^T (y - X\hat{\beta}_{\text{ridge}}(\tau)).$$

Of course, in this case the ordinary least squares estimate will minimize the right hand side, and we get $\hat{\tau} = \infty$ or $\hat{\lambda} = 0$ independent of the actual data. Therefore this is clearly not a suitable criterion for finding a good penalty parameter.

A natural starting point is to return to the original ridge regression article (Hoerl and Kennard, 1970) once again and find out what it says about choosing

the parameter $\lambda = \sigma^2/\tau^2$. The article gives a couple of qualitative guidelines for choosing a suitable value of the parameter k (our parameter λ). We are mainly interested in specific automatic procedures for choosing λ , but their qualitative criteria are interesting nonetheless, and we therefore quote them below.

- At a certain value of k the system will stabilize and have the general characteristics of an orthogonal system.
- Coefficients will not have unreasonable absolute values with respect to the factors for which they represent rates of change.
- Coefficients with apparently incorrect signs at $k = 0$ will have changed to have the proper sign.
- The residual sum of squares will not have been inflated to an unreasonable value. It will not be large relative to the minimum residual sum of squares or large relative to what would be a reasonable variance for the process generating the data.

These criteria seem sensible, but it may be too optimistic to expect that there exists a value of λ which simultaneously fulfill all of them. Also, terms like 'unreasonable absolute values' or 'apparently incorrect signs' seem to require a great deal of subjective judgement. Finally, using vague criteria like this, it is difficult to derive the properties of the resulting β -estimate or functions thereof. We would therefore like to use an automatic procedure, and thereby avoid these subjective judgement in every analysis. With the power of modern computers we also expect automatic procedures to be faster, even if they are in fact comparing a high number of possible values.

In their article Hoerl and Kennard do propose one automatic procedure as well, but it estimates a vector of $\hat{k}_i$'s instead of a single scalar $\hat{k}$. This leads to a diagonal matrix K with possibly different diagonal elements, inserted in the place of kI_p in the ridge estimate. The remaining part of this chapter is based on the original ridge estimate with a scalar k (λ), and we will therefore not discuss their procedure for estimating the vector.

In the years following the publication of the original ridge regression article many methods for choosing a good value $\hat{\lambda}$ for λ were proposed. Golub et al. (1979) provide references to more than 15 such attempts and propose one method themselves, using generalized cross-validation (GCV). Using our own notation, their estimate $\hat{\tau}$ is the minimizer of the function $V(\tau)$ given by

$$V(\tau) = \frac{1}{n} \|(I_n - A(\tau))y\|^2 \bigg/ \left[\frac{1}{n} \text{Tr}(I_n - A(\tau)) \right]^2,$$

where

$$A(\tau) = X(X^T X + \sigma^2/\tau^2 I_p)^{-1} X^T.$$

In the article, the estimated quantity is actually $\sigma^2/n\tau^2$, thereby avoiding the estimation of σ . But since we have assumed σ to be known, we will use the above version estimating the same quantity τ as in the following sections. However the fact that σ need not be estimated separately has the consequence that the GCV can be used also on problems where $n - p$ is small. The authors note that at the time of writing, the only other known methods with this property are cross validation and maximum likelihood estimation, both of which will be discussed and analyzed in later sections.

We will not discuss the generalized cross validation approach any further, but simply refer to Golub et al. (1979) for theoretical results and a numerical example. We will however include GCV in our numerical examples at the end of this chapter, together with the method of moments, ordinary cross validation and maximum likelihood estimation. In the next three sections we will derive a few theoretical results regarding the $\hat{\tau}$ obtained by each of these methods.

3.2.1 A simple approach based on the method of moments

Our first approach to choosing the value of τ will be one that is both intuitive and computationally simple. It is based on the method of moments, expressing the parameter in terms of population moments. As the population moments are clearly not available, we replace them by the corresponding sample moments. We recall that β has the prior distribution

$$\beta \sim N_p(0, \tau^2 I_p). \quad (3.9)$$

In order to estimate τ^2 we will assume for a moment that β_0 is really stemming from this distribution. Of course β_0 is unknown and we can therefore not use its value in estimating τ . However, we assume that the vector y is related to β_0 in the following way

$$y = X\beta_0 + \varepsilon \sim N_p(X\beta_0, \sigma^2 I_n), \quad (3.10)$$

that is, $y_1, \dots, y_n$ are independent, but not identically distributed. Combining (3.9) and (3.10) we are able to show that

$$y \sim N_n\left(0, \tau^2 \left(XX^T + \frac{\sigma^2}{\tau^2} I_n\right)\right).$$

Now, the first moment of y is clearly not helpful, but using the second moment we arrive at the following estimate of τ^2 :

$$\hat{\tau}^2 = \frac{y^T y - n\sigma^2}{Tr(XX^T)} = \frac{\frac{1}{n}y^T y - \sigma^2}{Tr(\Sigma_n)}. \quad (3.11)$$

We should note that there is a small probability of the right hand side being negative, but this situation is rarely encountered in practice. Clearly, we do not want a negative estimate, and in such cases we therefore let $\hat{\tau}^2 = 0$ (giving $\hat{\lambda} = \infty$). The exact form of our estimate is therefore given by

$$\hat{\tau}^2 = \max\left(0, \frac{\frac{1}{n}y^T y - \sigma^2}{Tr(\Sigma_n)}\right). \quad (3.12)$$

Later in this section we will derive the mean and variance of (3.11), and we will see that such zero estimates are increasingly unlikely as n increases. We will therefore not use the exact expression stated above, but rather use (3.11) in the following calculations. The probability of zero estimates will be discussed later in this section.

We are now ready to derive the properties of (3.11), but before we do so, we note that this particular estimate requires a known value of σ , or at least a good estimate of it. As noted earlier, this is not the case for the other methods

in this chapter when the goal is to find an estimate of the ridge parameter λ . However, we are not estimating λ itself, but rather $\tau^2 = \sigma^2\lambda$, and therefore σ is needed in all the other estimation methods as well. We will assume σ to be a known fixed constant in the following.

The first result of this section gives exact expressions for the first and second moment of our estimate. This is the only estimate in this chapter for which such simple closed form expressions are available.

Theorem 3.2.1. *For the estimate $\hat{\tau}^2$ defined by (3.11) we have*

$$E(\hat{\tau}^2|X, \beta_0) = \tau_{n,X}^2 \stackrel{\text{def}}{=} \frac{\beta_0^T \Sigma_n \beta_0}{\text{Tr}(\Sigma_n)}$$

and

$$\text{Var}(\hat{\tau}^2|X, \beta_0) = \frac{1}{n} v_{n,X}^2 \stackrel{\text{def}}{=} \frac{1}{n} \frac{4\sigma^2 \beta_0^T \Sigma_n \beta_0 + 2\sigma^4}{\text{Tr}(\Sigma_n)^2}.$$

Proof. Both results are direct consequences of (3.10) and the basic properties of the multivariate normal distribution. $\square$

We recall that Σ_n is assumed to be positive definite. Thus, the expected value and the variance are always positive. We do still consider our x_i 's to be fixed known vectors. However, sometimes it seems reasonable to assume that the x_i 's are such that $\Sigma_n \rightarrow^P \Sigma$ for some positive definite matrix Σ . One example for which this holds is when the x_i 's are i.i.d. with common mean 0 and covariance matrix Σ , since

$$\Sigma_n = \frac{1}{n} \sum_{i=1}^n x_i x_i^T,$$

and clearly each element of Σ_n converges in probability to the corresponding element of Σ . If we are willing to make this assumption, we see that both $\tau_{n,X}^2$ and $v_{n,X}^2$ converge in probability towards positive constants, and the $1/n$ factor ensures that the probability of a negative value of $\hat{\tau}^2$ tends to zero as n increases.

We will state a more specific theorem describing the limiting behaviour of (3.11), but before we do this we need to present a theorem and a condition. The theorem is a generalized version of the central limit theorem, and will be used in this and later sections. This particular version is by Lehmann (1999):

Theorem 3.2.2 (Liapounov). *Let X_i ($i = 1, \dots, n$) be independently distributed with means $E(X_i) = \xi_i$ and variances σ_i^2 , and with finite third moments. If*

$$Y_n = \frac{\bar{X} - E(\bar{X})}{\sqrt{\text{Var}(\bar{X})}} = \frac{\sqrt{n}(\bar{X} - \bar{\xi})}{\sqrt{(\sigma_1^2 + \dots + \sigma_n^2)/n}},$$

then

$$Y_n \xrightarrow{L} N(0, 1),$$

provided

$$\left[E \left(\sum |X_i - \xi_i|^3 \right) \right]^2 = o \left[\left(\sum \sigma_i^2 \right)^3 \right].$$

In our situation, the above condition has the following form:

Condition 3.2.1. For $Z_i = (y_i^2 - \sigma^2)/\text{Tr}(\Sigma_n)$ we have

$$\left[E \left(\sum |Z_i - \xi_i|^3 \right) \right]^2 = o \left[\left(\sum \sigma_i^2 \right)^3 \right],$$

with $\xi_i = E(Z_i) = \beta_0^T x_i x_i^T \beta_0 / \text{Tr}(\Sigma_n)$ and $\sigma_i^2 = \text{Var}(Z_i) = (4\sigma^2 \beta_0^T x_i x_i^T \beta_0 + 2\sigma^4) / \text{Tr}(\Sigma_n)^2$.

It is not clear exactly what is needed to satisfy this condition, and we will not try to derive any necessary or sufficient conditions here. However, in my simulations with x_i^T stemming from different multivariate normal distributions and linear transformations of uniformly distributed variables, the results below have always been valid. This may also be seen in the numerical example at the end of this chapter.

We are now ready to state a limiting result about our estimate $\hat{\tau}^2$:

Theorem 3.2.3. For the estimate $\hat{\tau}^2$ defined by (3.11) we have

$$\frac{\sqrt{n}(\hat{\tau}^2 - \tau_{n,X}^2)}{v_{n,X}} \xrightarrow{L} N(0, 1), \quad (3.13)$$

with $\tau_{n,X}^2$ and $v_{n,X}^2$ as defined in Theorem 3.2.1, whenever Condition 3.2.1 is satisfied.

Proof. If we let

$$Z_i = \frac{y_i^2 - \sigma^2}{\text{Tr}(\Sigma_n)},$$

we get

$$\hat{\tau}^2 = \bar{Z}.$$

The Z_i 's are clearly independently distributed, and since the y_i 's are normally distributed with finite mean and variance, the Z_i 's all have finite third moments. We find that $E(Z_i) = \xi_i = \beta_0^T x_i x_i^T \beta_0 / \text{Tr}(\Sigma_n)$ and $\text{Var}(Z_i) = \sigma_i^2 = (4\sigma^2 \beta_0^T x_i x_i^T \beta_0 + 2\sigma^4) / \text{Tr}(\Sigma_n)^2$. Now the result

$$\frac{\sqrt{n}(\hat{\tau}^2 - \tau_{n,X}^2)}{v_{n,X}} = \frac{\sqrt{n}(\bar{Z} - \bar{\xi})}{\sqrt{(\sigma_1^2 + \dots + \sigma_n^2)/n}} \xrightarrow{L} N(0, 1)$$

follows directly from Theorem 3.2.2 and Condition 3.2.1. $\square$

One immediate consequence of the previous theorem is the following:

Corollary 3.2.4. If Condition 3.2.1 is satisfied and $\Sigma_n \rightarrow^P \Sigma$, then

$$\sqrt{n}(\hat{\tau}^2 - \tau_{n,X}^2) \xrightarrow{L} N(0, v_{0X}^2) \quad (3.14)$$

with $\tau_{n,X}^2$ as defined in Theorem 3.2.1 and

$$v_{0X}^2 \stackrel{\text{def}}{=} \frac{4\sigma^2 \beta_0^T \Sigma \beta_0 + 2\sigma^4}{\text{Tr}(\Sigma)^2}.$$

Proof. The function

$$f(\Sigma_n) = \frac{4\sigma^2\beta_0^T\Sigma_n\beta_0 + 2\sigma^4}{\text{Tr}(\Sigma_n)^2}$$

is clearly continuous, and therefore $\Sigma_n \rightarrow^P \Sigma$ implies $f(\Sigma_n) \rightarrow^P f(\Sigma) = v_{0X}^2$. Now Slutsky's theorem completes the proof. $\square$

Finally, we obtain the limit in probability for our estimate $\hat{\tau}^2$:

Corollary 3.2.5. *When Condition 3.2.1 is satisfied and $\Sigma_n \rightarrow^P \Sigma$, then*

$$\hat{\tau}^2|\beta_0 \xrightarrow{P} \tau_0^2 \stackrel{\text{def}}{=} \frac{\beta_0^T\Sigma\beta_0}{\text{Tr}(\Sigma)}. \quad (3.15)$$

Proof. By Corollary 3.2.4

$$\hat{\tau}^2 - \frac{\beta_0^T\Sigma_n\beta_0}{\text{Tr}(\Sigma_n)} \xrightarrow{P} 0.$$

Also, by the result in the proof of the same corollary,

$$\frac{\beta_0^T\Sigma_n\beta_0}{\text{Tr}(\Sigma_n)} - \frac{\beta_0^T\Sigma\beta_0}{\text{Tr}(\Sigma)} \xrightarrow{P} 0.$$

Combining these two results we get

$$\hat{\tau}^2 - \frac{\beta_0^T\Sigma\beta_0}{\text{Tr}(\Sigma)} \xrightarrow{P} 0,$$

or equivalently

$$\hat{\tau}^2 \xrightarrow{P} \frac{\beta_0^T\Sigma\beta_0}{\text{Tr}(\Sigma)}.$$

$\square$

We have now derived a few properties of this estimate obtained using the method of moments. The interpretation of the estimate itself will be discussed when the other estimates have been presented. Also, the different estimates will be compared analytically and numerically in the last sections of this chapter.

3.2.2 Cross validation

Another popular method of estimation is the cross validation. We have already seen a generalized version of it, but we have not discussed cross validation itself. The method was introduced by Allen (1974) and was originally named PRESS, an acronym for 'Prediction Sum of Squares'. In his article, Allen suggests that a specific criterion of good prediction could be used in choosing the ridge parameter, and he proposes one such criterion given by

$$\text{PRESS}(D) = \sum_{i=1}^n (Y - \hat{Y}_{(i)})^2,$$

where $\hat{Y}_{(i)}$ is the estimator of $E(Y_i)$ *excluding* the i -th observation. Here D is a diagonal matrix generalizing our matrix $\sigma^2 I_p / \tau^2$. The estimate $\hat{\tau}^2$ obtained by using cross validation is given by

$$\hat{\tau} = \underset{\tau}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n \left[y_i - x_i^T \hat{\beta}_{\text{ridge},(i)}(\tau) \right]^2 \right\} \quad (3.16)$$

where $\hat{\beta}_{\text{ridge},(i)}(\tau)$ is the ridge estimate of β when the i th observation is left out, that is

$$\hat{\beta}_{\text{ridge},(i)}(\tau) = \left(\sum_{j \neq i} x_j x_j^T + \frac{\sigma^2}{\tau^2} I_p \right)^{-1} \sum_{j \neq i} x_j y_j.$$

The resulting $\hat{\tau}$ may now be found numerically by running through all values of τ in some suitable interval. In the numerical comparisons at the end of this chapter, this expression is used directly, and a numerical optimization procedure is used for finding the optimal value $\hat{\tau}$. minimization of a complicated function involving

Finding the estimate $\hat{\tau}$ analytically seems infeasible, as it requires minimization of a complicated function with several inverse matrixes involving the parameter τ . However, we are able to obtain approximate results if we assume n to be large. The remaining part of this section will present a few such results. Before we start, we will state a few assumptions and a lemma that will be used extensively throughout the thesis.

In the following we assume to have n large compared to p , and we assume the matrix Σ_n to be positive definite for all n . Also we assume $\Sigma_n = O_p(1)$, and in particular that the largest eigenvalue of Σ_n is $O_p(1)$. The following lemma (Lyche, 2010) will be used several times in this chapter:

Lemma 3.2.6 (Neumann Series). *Suppose $B \in \mathbb{C}^{n,n}$. If $\rho(B) < 1$ then $(I - B)$ is nonsingular and*

$$(I - B)^{-1} = \sum_{k=0}^{\infty} B^k.$$

Here $\rho(B)$ is the spectral radius defined as the maximum absolute value of the eigenvalues of B .

First we state a theorem describing the form of the estimate itself:

Theorem 3.2.7. *If Σ_n is positive definite for all n and $\Sigma_n = O_p(1)$, the estimate $\hat{\tau}^2$ given by (3.16) satisfies*

$$\hat{\tau}^2 = -\frac{2\sigma^2 U_n}{V_n} + O_p(n^{-1}),$$

with $U_n = \hat{\beta}_{LS}^T \Sigma_n^{-1} \hat{\beta}_{LS}$ and

$$V_n = \frac{1}{n} \sum_{i=1}^n \left[-2y_i x_i^T \Sigma_n^{-2} x_i y_i + 4y_i x_i^T \Sigma_n^{-1} (\Sigma_n^{-1} x_i x_i^T + x_i x_i^T \Sigma_n^{-1}) \hat{\beta}_{LS} - \hat{\beta}_{LS}^T (2x_i x_i^T \Sigma_n^{-1} + 4\Sigma_n^{-1} x_i x_i^T) \Sigma_n^{-1} x_i x_i^T \hat{\beta}_{LS} \right]$$

Proof. We want to transform the right hand side of (3.16) into a more understandable function of τ . To keep everything as simple as possible, we remove

all constant terms and focus on terms of order at least n^{-2} . Now

$$\begin{aligned}\hat{\tau} &= \operatorname{argmin}_{\tau} \left\{ \frac{1}{n} \sum_{i=1}^n \left[y_i - x_i^T \hat{\beta}_{\text{ridge},(i)}(\tau) \right]^2 \right\} \\ &= \operatorname{argmin}_{\tau} \left\{ \frac{1}{n} \sum_{i=1}^n \left[-2y_i x_i^T \hat{\beta}_{\text{ridge},(i)}(\tau) + \left(x_i^T \hat{\beta}_{\text{ridge},(i)}(\tau) \right)^2 \right] \right\}\end{aligned}$$

and, using the assumption of positive definiteness of Σ_n we find that

$$\begin{aligned}\hat{\beta}_{(i)}(\tau) &= \left(I_p + \frac{1}{n} \Sigma_n^{-1} \left(\frac{\sigma^2}{\tau^2} I_p - x_i x_i^T \right) \right)^{-1} \Sigma_n^{-1} \left(\frac{X^T y}{n} - \frac{x_i y_i}{n} \right) \\ &= \left(I_p + \frac{1}{n} Z_i(\tau) \right)^{-1} W_i.\end{aligned}$$

Here we have introduced two new quantities to make the following calculations easier to follow, and we note that $Z_i(\tau) = O_p(1)$ and $W_i = O_p(1)$. Now the eigenvalues of $Z_i(\tau)$ are also $O_p(1)$, and for large n we may use Lemma 3.2.6 and get

$$\begin{aligned}\hat{\tau} &= \operatorname{argmin}_{\tau} \left\{ \frac{1}{n} \sum_{i=1}^n -2y_i x_i^T \sum_{k=0}^{\infty} \left(-\frac{1}{n} Z_i(\tau) \right)^k W_i \right. \\ &\quad \left. + W_i^T \sum_{k=0}^{\infty} \left(-\frac{1}{n} Z_i^T(\tau) \right)^k x_i x_i^T \sum_{k=0}^{\infty} \left(-\frac{1}{n} Z_i(\tau) \right)^k W_i \right. \\ &= \operatorname{argmin}_{\tau} \left\{ \frac{1}{n} \sum_{i=1}^n -2y_i x_i^T \left(-\frac{1}{n} Z_i(\tau) + \frac{1}{n^2} Z_i^2(\tau) \right) W_i \right. \\ &\quad \left. + W_i^T \left(-\frac{2}{n} x_i x_i^T Z_i(\tau) + \frac{1}{n^2} Z_i^T(\tau) x_i x_i^T Z_i(\tau) + \frac{2}{n^2} x_i x_i^T Z_i^2(\tau) \right) W_i \right. \\ &\quad \left. + O_p(n^{-3}) \right\}.\end{aligned}$$

Substituting back for $Z_i(\tau)$ and W_i is a tedious task requiring high precision. We will not give the steps here, but simply state that after quite a lot of algebra we finally arrive at

$$\begin{aligned}\hat{\tau} &= \operatorname{argmin}_{\tau} \left\{ \frac{\sigma^4}{\tau^4} \hat{\beta}_{LS}^T \Sigma_n^{-1} \hat{\beta}_{LS} \right. \\ &\quad \left. + \frac{\sigma^2}{\tau^2} \frac{1}{n} \sum_{i=1}^n \left[-2y_i x_i^T \Sigma_n^{-2} x_i y_i + 4y_i x_i^T \Sigma_n^{-1} (\Sigma_n^{-1} x_i x_i^T + x_i x_i^T \Sigma_n^{-1}) \hat{\beta}_{LS} \right. \right. \\ &\quad \left. \left. - \hat{\beta}_{LS}^T (2x_i x_i^T \Sigma_n^{-1} + 4\Sigma_n^{-1} x_i x_i^T) \Sigma_n^{-1} x_i x_i^T \hat{\beta}_{LS} \right] + O_p(n^{-1}) \right\}.\end{aligned}$$

Now the right hand side is a polynomial in $1/\tau^2$, which may be written more compactly as

$$\frac{\sigma^4}{\tau^4} U_n + \frac{\sigma^2}{\tau^2} V_n + O_p(n^{-1})$$

with U_n and V_n as defined in the theorem. All the terms included in the $O_p(n^{-1})$ term are of the form

$$\frac{1}{n^k} \frac{1}{\tau^{2m}} \text{const}(x, y, \sigma)$$

with $k \geq 1$, $m \geq 0$ and different $O_p(1)$ constants dependent on the data but independent of τ . Therefore, differentiating the $O_p(n^{-1})$ term with respect to $1/\tau^2$ leads to another $O_p(n^{-1})$ term, and we find that

$$\hat{\tau}^2 = -\frac{2\sigma^2 U_n}{V_n} + O_p(n^{-1}).$$

□

Like in the previous chapter, this expression may be negative, but the probability decreases when n increases. Using numerical optimization we avoid this problem by optimizing over non-negative values only. If we would for some reason use this expression to obtain an approximation of the exact estimate, the natural choice would be to let $\hat{\tau}^2 = 0$ when the result is negative. However, since the approximation is only useful for large n , this problem is rarely encountered in practice.

The formula in Theorem 3.2.7 can of course not be used in finding the exact value of $\hat{\tau}$. Still, we are able to derive the limiting distribution of $\hat{\tau}$ as $n \rightarrow \infty$, and this result may be used in obtaining approximate expressions for the mean and variance of $\hat{\tau}$ when β_0 is known. This will be demonstrated in the numerical example at the end of this chapter.

Before we move on to deriving the limiting distribution of $\hat{\tau}^2$, we note that U_n and V_n may be expressed in terms of ε in the following way:

$$\begin{aligned} U_n &= \beta_0^T \Sigma_n^{-1} \beta_0 + 2\beta_0^T \Sigma_n^{-2} \frac{X^T \varepsilon}{n} + \frac{\varepsilon^T X}{n} \Sigma_n^{-3} \frac{X^T \varepsilon}{n} \\ V_n &= \frac{1}{n} \sum_{i=1}^n \left[(-2x_i^T \Sigma_n^{-2} x_i) \varepsilon_i^2 \right. \\ &\quad \left. + (4\beta_0^T \Sigma_n^{-1} x_i x_i^T \Sigma_n^{-1} x_i) \varepsilon_i + (-4\beta_0^T \Sigma_n^{-1} C_n \Sigma_n^{-1} x_i) \varepsilon_i \right] + O_p(n^{-1}) \end{aligned}$$

with

$$C_n = \frac{1}{n} \sum_{i=1}^n x_i x_i^T \Sigma_n^{-1} x_i x_i^T. \quad (3.17)$$

The derivation is straightforward and will not be given here. These alternative expressions will be used in the proof of our next theorem. In particular, we observe that V_n is close to a sum of independent but not identically distributed variables, or that

$$V_n = \frac{1}{n} \sum_{i=1}^n Z_i + O_p(n^{-1}),$$

with

$$Z_i = (-2x_i^T \Sigma_n^{-2} x_i) \varepsilon_i^2 + (4\beta_0^T \Sigma_n^{-1} x_i x_i^T \Sigma_n^{-1} x_i) \varepsilon_i + (-4\beta_0^T \Sigma_n^{-1} C_n \Sigma_n^{-1} x_i) \varepsilon_i. \quad (3.18)$$

Like in the previous section, we will need Theorem 3.2.2 in the proof of our limiting theorem, and we therefore state the following condition:

Condition 3.2.1. For Z_i as defined by (3.18) we have

$$\left[E \left(\sum |Z_i - \xi_i|^3 \right) \right]^2 = o \left[\left(\sum \sigma_i^2 \right)^3 \right],$$

with $\xi_i = E(Z_i) = -2\sigma^2 x_i^T \Sigma_n^{-2} x_i$ and

$$\begin{aligned} \sigma_i^2 = \text{Var}(Z_i) &= 8\sigma^2 (x_i^T \Sigma_n^{-2} x_i)^2 + 16\sigma^2 (\beta_0^T \Sigma_n^{-1} x_i x_i^T \Sigma_n^{-1} x_i)^2 \\ &+ 16\sigma^2 (\beta_0^T \Sigma_n^{-1} C_n \Sigma_n^{-1} x_i)^2 - 32\sigma^2 (\beta_0^T \Sigma_n^{-1} x_i x_i^T \Sigma_n^{-1} x_i) (\beta_0^T \Sigma_n^{-1} C_n \Sigma_n^{-1} x_i). \end{aligned}$$

Again it is not clear exactly what is needed to satisfy this condition, but again for my simulations with x_i^T stemming from different multivariate normal distributions and linear transformations of uniformly distributed variables, the results below have always been valid. This may once again be seen in the numerical example at the end of this chapter.

We are now ready to present our theorem regarding the limiting distribution of $\hat{\tau}^2$.

Theorem 3.2.8. Under the same conditions as in Theorem 3.2.7 we have

$$\frac{\sqrt{n}(\hat{\tau}^2 - \tau_{n,X}^2)}{v_{n,X}} \xrightarrow{L} N(0, 1),$$

with

$$\tau_{n,X}^2 = \frac{\beta_0^T \Sigma_n^{-1} \beta_0}{\text{Tr}(\Sigma_n^{-1})}.$$

and $v_{n,X}^2$ as defined by (3.22), provided that Condition 3.2.1 is satisfied.

Proof. Using the delta method and the fact that

$$\sqrt{n}(\Sigma_n^{1/2}(\hat{\beta}_{LS} - \beta_0)) \xrightarrow{L} N_p(0, \sigma^2 I_p),$$

we find that

$$\frac{\sqrt{n}(U_n - \beta_0^T \Sigma_n^{-1} \beta_0)}{\sigma_{U_n}} \xrightarrow{L} N(0, 1),$$

with

$$\sigma_{U_n}^2 = 4\sigma^2 \beta_0^T \Sigma_n^{-3} \beta_0. \quad (3.19)$$

Since we require Condition 3.2.1 to be satisfied, we may use Theorem 3.2.2 and find that

$$\frac{\sqrt{n}(V_n + 2\sigma^2 \text{Tr}(\Sigma_n^{-1}))}{\sigma_{V_n}} \xrightarrow{L} N(0, 1).$$

Here we have used the fact that $\bar{Z}_i - V_n + O_p(n^{-1})$, and that the two quantities are therefore asymptotically equal. Also, $\bar{\xi}_i = -2\sigma^2 \text{Tr}(\Sigma_n^{-1})$ and

$$\begin{aligned} \sigma_{V_n}^2 &= \frac{1}{n} \sum_{i=1}^n \sigma_i^2 = 8\sigma^4 \frac{1}{n} \sum_{i=1}^n (x_i^T \Sigma_n^{-2} x_i)^2 \\ &+ 16\sigma^2 \frac{1}{n} \sum_{i=1}^n (\beta_0^T \Sigma_n^{-1} x_i x_i^T \Sigma_n^{-1} x_i)^2 - 16\sigma^2 \beta_0^T \Sigma_n^{-1} C_n \Sigma_n^{-1} C_n \Sigma_n^{-1} \beta_0. \end{aligned}$$

We now have the limiting distribution of U_n and V_n , but in order to be able to derive the limiting distribution of the quotient, we need to calculate the covariance as well. It turns out that

$$\text{Cov}(U_n, V_n|X) = 0 + O(n^{-3/2}),$$

and the covariance is thus negligible compared to the variances. Using the delta method we find that

$$\frac{\sqrt{n}(\hat{\tau}^2 - \tau_{n,X}^2)}{v_{n,X}} \xrightarrow{L} N(0, 1), \quad (3.20)$$

with

$$\tau_{n,X}^2 = \frac{\beta_0^T \Sigma_n^{-1} \beta_0}{\text{Tr}(\Sigma_n^{-1})}, \quad (3.21)$$

and

$$v_{n,X}^2 = \frac{\sigma_{U_n}^2}{\text{Tr}(\Sigma_n^{-1})^2} + \frac{(\beta_0^T \Sigma_n^{-1} \beta_0)^2 \sigma_{V_n}^2}{4\sigma^4 \text{Tr}(\Sigma_n^{-1})^4}. \quad (3.22)$$

□

In the previous chapter we were able to obtain a similar result with the limiting variance on the right hand side only by assuming $\Sigma_n \rightarrow^P \Sigma$ for some positive definite matrix Σ . The variance term is far more complicated in this case, and we will therefore leave it on the left hand side.

Before we move on to our last estimate, we state one more result regarding the limiting behaviour of $\hat{\tau}^2$.

Corollary 3.2.9. *If the conditions of Theorem 3.2.7 are satisfied and $\Sigma_n \rightarrow^P \Sigma$ for some positive definite matrix Σ , then*

$$\hat{\tau}^2 \rightarrow^P \frac{\beta_0^T \Sigma^{-1} \beta_0}{\text{Tr}(\Sigma^{-1})}.$$

Proof. By Theorem 3.2.8

$$\hat{\tau}^2 - \frac{\beta_0^T \Sigma_n^{-1} \beta_0}{\text{Tr}(\Sigma_n^{-1})} \rightarrow^P 0.$$

Also, $\Sigma_n \rightarrow^P \Sigma$ implies

$$\frac{\beta_0^T \Sigma_n^{-1} \beta_0}{\text{Tr}(\Sigma_n^{-1})} - \frac{\beta_0^T \Sigma^{-1} \beta_0}{\text{Tr}(\Sigma^{-1})} \rightarrow^P 0.$$

Combining these two results completes the proof. □

3.2.3 The maximum marginal likelihood estimator

We have already described two possible methods for choosing τ . In this section we will consider a third possibility inspired by Bayesian statistics and the maximum likelihood estimator. In the introduction of this chapter we discussed the fact that the ridge estimate $\hat{\beta}_{\text{ridge}}(\tau)$ may be interpreted as the mean (and also

the mode) of the posterior distribution of β when the prior distribution is given by

$$\pi_\tau(\beta) = \frac{1}{\sqrt{2\pi}} \frac{1}{\tau^p} \exp \left\{ -\frac{1}{2} \frac{\beta^T \beta}{\tau^2} \right\},$$

or in other words

$$\beta \sim N_p(0, \tau^2 I_p).$$

We want to estimate τ , and the idea is to let $\hat{\tau}$ be the value of τ that maximizes the marginal likelihood of y . The marginal likelihood is given by

$$\begin{aligned} f_{\text{marg}}(y, \tau) &= \int \exp \{l_n(\beta)\} \pi_\tau(\beta) d\beta \\ &\propto \frac{1}{\tau^p} \int \exp \left\{ -\frac{1}{2} \frac{(y - X\beta)^T (y - X\beta)}{\sigma^2} - \frac{1}{2} \frac{1}{\tau^2} \beta^T \beta \right\} d\beta \\ &\propto \frac{1}{n^{p/2}} \frac{\sigma^p}{\tau^p} |\Sigma_{n,\text{ridge}}(\tau)|^{-1/2} \\ &\quad \exp \left\{ \frac{n}{2\sigma^2} \hat{\beta}_{\text{ridge}}^T(\tau) \Sigma_{n,\text{ridge}}(\tau) \hat{\beta}_{\text{ridge}}(\tau) \right\}. \end{aligned} \quad (3.23)$$

The maximum (marginal) likelihood estimate is therefore given by

$$\begin{aligned} \hat{\tau} &= \arg\max_{\tau} \{f_{\text{marg}}(y, \tau)\} \\ &= \arg\max_{\tau} \left\{ -p \log(\tau) - \frac{1}{2} \log |\Sigma_{n,\text{ridge}}(\tau)| + \frac{n}{2\sigma^2} \hat{\beta}_{\text{ridge}}^T(\tau) \Sigma_{n,\text{ridge}}(\tau) \hat{\beta}_{\text{ridge}}(\tau) \right\} \end{aligned} \quad (3.24)$$

The expression above is a complicated function of τ , and we are not able to find the maximizer analytically. However, it is easily maximized by any numerical optimization procedure, and in the numerical examples at the end of this chapter, a suitably normalized version of this expression is being maximized.

We want to compare the maximum likelihood estimate with the two other estimates, and we therefore present a theorem providing an approximate value of $\hat{\tau}^2$:

Theorem 3.2.10. *If Σ_n is positive definite for all n and $\Sigma_n = O_p(1)$, then for $\hat{\tau}$ given by (3.24) we have*

$$\hat{\tau}^2 = \frac{\hat{\beta}_{LS}^T \hat{\beta}_{LS}}{p} + O_p(n^{-1}).$$

Proof. To find the maximizer of (3.24), we differentiate it with respect to τ and let

$$0 = -p + \frac{1}{n} \frac{\sigma^2}{\hat{\tau}^2} \text{Tr}(\Sigma_{n,\text{ridge}}^{-1}(\hat{\tau})) + \frac{1}{\hat{\tau}^2} \frac{y^T X}{n} \Sigma_{n,\text{ridge}}^{-1}(\hat{\tau}) \Sigma_{n,\text{ridge}}^{-1}(\hat{\tau}) \frac{X^T y}{n}.$$

We now use Lemma 3.2.6 to expand $\Sigma_{n,\text{ridge}}^{-1}(\hat{\tau})$, and find that

$$0 = -p + \frac{1}{\hat{\tau}^2} \hat{\beta}_{LS}^T \hat{\beta}_{LS} + O_p(n^{-1}).$$

giving

$$\hat{\tau}^2 = \frac{\hat{\beta}_{LS}^T \hat{\beta}_{LS}}{p} + O_p(n^{-1}).$$

□

For very small n we may again get negative estimates, but since the dominant term is guaranteed to be non-negative, we see that this problem vanishes when n increases. Using this approximate expression we are able to derive the approximate mean and variance of the estimate as well.

Corollary 3.2.11. *For $\hat{\tau}$ given by (3.24) we have*

$$E(\hat{\tau}^2|X, \beta_0) = \frac{\beta_0^T \beta_0}{p} + O(n^{-1})$$

and

$$\text{Var}(\hat{\tau}^2|X, \beta_0) = \frac{4\sigma^2}{n} \frac{\beta_0^T \Sigma_n^{-1} \beta_0}{p^2} + O(n^{-3/2}).$$

under the conditions of Theorem 3.2.10.

These results are immediate consequences of the distribution of $\hat{\beta}_{LS}$.

Finally, we derive the limiting distribution of $\hat{\tau}^2$.

Theorem 3.2.12. *For $\hat{\tau}^2$ given by (3.24) and under the conditions of Theorem 3.2.10 we have*

$$\frac{\sqrt{n}(\hat{\tau}^2 - \tau_{n,X}^2)}{v_{n,X}} \xrightarrow{L} N_p(0, 1),$$

with

$$\tau_{n,X}^2 = \beta_0^T \beta_0 / p$$

and

$$v_{n,X}^2 = \frac{4\sigma^2 \beta_0^T \Sigma_n^{-1} \beta_0}{p^2}.$$

Proof. We know that

$$\sqrt{n}(\Sigma_n^{1/2}(\hat{\beta}_{LS} - \beta_0)) \xrightarrow{L} N_p(0, \sigma^2 I_p). \quad (3.25)$$

Then, by using the delta method with the function $g(\beta) = \beta^T \beta / p$ we find that,

$$\frac{\sqrt{n}(\hat{\tau}^2 - \tau_X^2)}{v_X} \xrightarrow{L} N_p(0, 1),$$

with $\tau_{n,X}^2$ and $v_{n,X}$ as given in the theorem. $\square$

Corollary 3.2.13. *If we assume that $\Sigma_n \rightarrow^P \Sigma$ for some positive definite matrix Σ , then*

$$\sqrt{n}(\hat{\tau}^2 - \tau_X^2) \xrightarrow{L} N(0, v_{0X}^2),$$

with

$$v_{0X}^2 = \frac{4\sigma^2 \beta_0^T \Sigma^{-1} \beta_0}{p^2}.$$

This result is a direct consequence of Theorem 3.2.12 and Slutsky's Theorem. Another consequence of Theorem 3.2.12 is given in the following corollary:

Corollary 3.2.14. *Under the conditions of Theorem 3.2.12*

$$\hat{\tau}^2 \xrightarrow{P} \frac{\beta_0^T \beta_0}{p}.$$

3.3 A numerical example

In addition to the ordinary least squares estimate we have in this chapter introduced the ridge estimate with a total of four different methods for estimating the ridge parameter τ . The generalized cross validation (GCV) was only briefly mentioned in the introduction of the previous section, whereas the method of moments (MoM), cross validation (CV) and maximum (marginal) likelihood estimation (MML) were studied in more detail. We derived results concerning the limiting distribution of each estimate $\hat{\tau}^2$, but the resulting quantities were dependent of σ^2 , the true β_0 and the sample covariance matrix Σ_n , sometimes even the matrix X . Thus, analytical comparison is a difficult task, as the performance of each estimate depends on the relationship between all these quantities. For example, the relation between the mean of the three different versions of $\hat{\tau}^2$ depends on how β is related to the eigenvectors of Σ_n . Since β is generally unknown, such results would not be of much help when deciding which method to choose. Therefore, instead of deriving more results analytically, we will use this section discussing a numerical example comparing all four methods. Not only will we discuss the estimation of τ itself, but also the performance of each estimate in terms of $\hat{\beta}_{\text{ridge}}(\hat{\tau})$ and predicted values $\hat{y} = X\hat{\beta}_{\text{ridge}}(\hat{\tau})$. When appropriate, we will include the ordinary least squares estimate (LS) as well.

3.3.1 Designing the experiment

This chapter is based on the assumption that

$$y \sim N_n(X\beta_0, \sigma^2 I_n),$$

with β_0 being one realization stemming from a certain multivariate distribution of β 's. We have also introduced a prior normal distribution for β , of the form

$$\beta \sim N_p(0, \tau^2 I_p).$$

In our numerical experiment we therefore have many parameters which need to be set. We have to choose values of p , n and σ , and we need to decide how X and β_0 should be defined. We want to analyze the setting of a fixed β_0 and a fixed X , but β_0 should be related to τ as well. Before we present the results, we will therefore explain how the different parameters have been chosen.

Golub et al. (1979) presents a numerical example with $p = 10$, $n = 21$ in their article about generalized cross validation. A main point in the article is the ability of obtaining reasonable estimates in cases where $n - p$ is small, and we will use a similar combination of $p = 10$ and $n = 20$ as an example of p large relative to n . We will also provide examples with small p and large n , and for this purpose we have chosen to use the same p , but increase n to 100. This makes it possible to use the same β_0 's in both settings, allowing us to study how the performance is connected to the value of n . In particular, we will use the limiting distributions derived in the previous chapter to obtain approximate values of the mean and standard deviation of each estimate, and we will compare them to the sample mean and standard deviation of the estimates. By using the same β_0 and two different values of n , we hope to see that the approximations are getting better for larger n .

We want β_0 to be a fixed vector, but we also want it to depend on τ . We have chosen to solve this problem by using two different vectors, which will be called

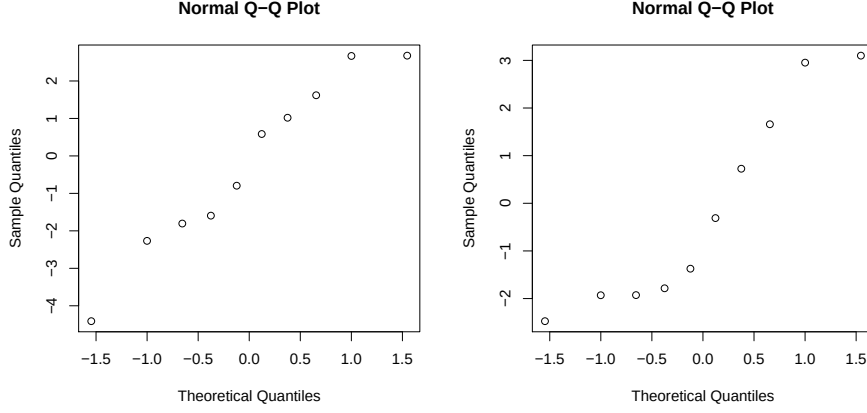


Figure 3.4: Q-Q plots of β_1 generated from a normal distribution with mean 0 and variance 4, and β_2 generated from a uniform distribution with the same mean and variance.

β_1 and β_2 in the remaining part of this section. The first vector (β_1) is generated from a normal distribution with mean 0 and covariance matrix $\tau^2 I_p$ with $\tau = 2$. The second vector (β_2) is generated from a uniform distribution with mean 0 and the same variance, with all components being independent. Golub et al. point out that the maximum likelihood estimate is expected to perform well when β_0 do really come from the prior distribution, but that its performance will be worse when this is not the case. We therefore include both possibilities in our experiment. Figure 3.4 shows normal Q-Q plots for the vectors β_1 and β_2 .

Another problem is how to choose the matrix X . In the introduction of this chapter we focused on the matrix Σ_n and its eigenvalues following the lines of Hoerl and Kennard (1970). We have chosen to generate X in a way that ensures a certain Σ_n , using the following approach:

1. Choose the matrix Ω you want Σ_n to equal.
2. Generate a random $n \times p$ matrix X with each element following a uniform distribution on $(-1/2, 1/2)$.
3. Use the Gram-Schmidt process to orthonormalize the matrix X , and multiply the result by $\sqrt{n}$. Now $\Sigma_n = I_p$.
4. Calculate the Cholesky decomposition $LL^T = \Omega$, and let XL^T be the final matrix X . Now $\Sigma_n = \Omega$.

The problem of choosing a matrix X is now reduced to choosing a matrix Ω . We will use three different matrices, which we name Ω_1 , Ω_2 and Ω_3 . The first matrix is the identity matrix, that is $\Omega_1 = I_p$. The second matrix Ω_2 is such that the ij -th element equals $0.9^{|i-j|}$ and is chosen in order to give us an ill-conditioned matrix as well. The third matrix is designed to be random, with no special structure except the fact that the sum of eigenvalues equals p like for the other two matrices. The matrix Ω_3 is only generated once, and is

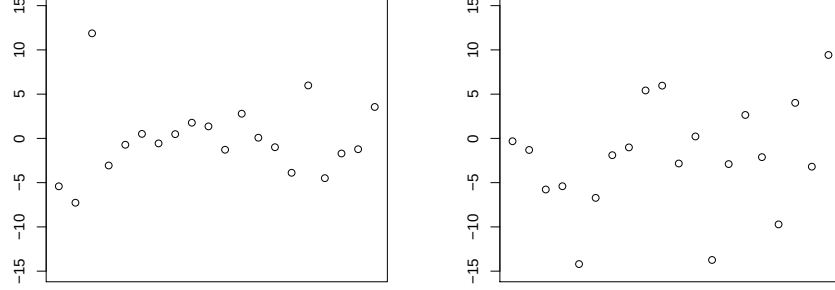


Figure 3.5: Plots of $X\beta$ with $n = 20$. In the left plot $\beta = \beta_1$ and X is generated using Ω_2 . In the right plot $\beta = \beta_2$ and X is generated using Ω_3 .

therefore the same matrix in all the experiments. The matrix is generated using the following algorithm:

1. Generate a random $p \times p$ matrix Q with each element following a uniform distribution on $(-1/2, 1/2)$.
2. Use the Gram-Schmidt process to orthonormalize the matrix Q .
3. Generate a set of p eigenvalues $(\lambda_1, \dots, \lambda_p)$ from the uniform distribution on $(0, 1)$ and normalize them to have sum p . Let $L = \text{diag}(\lambda_1, \dots, \lambda_p)$.
4. Finally, let $\Omega_3 = QLQ^T$. Now Ω_3 is a positive definite matrix with eigenvalues $(\lambda_1, \dots, \lambda_p)$ and eigenvectors equal to the columns of Q .

Finally, we need to choose a value of σ . Golub et al. (1979) use several different values in their numerical example, but to keep the number of possible combinations at a manageable level, we will use $\sigma = 1$ in all the experiments. To demonstrate how this compares to the values of $X\beta$, Figure 3.5 shows two different plots of $X\beta$. Inspecting these plots, a variance of 1 seems to be a reasonable choice when we restrict ourselves to one value of σ .

3.3.2 Quantities of interest

Now that we have set all the parameters, we need to decide which quantities should be computed and compared. I have chosen to look at three different kinds of estimates, with the main focus being on the first of them. The first estimate is $\hat{\tau}^2$ itself, the second is the resulting β -estimate $\hat{\beta}_{\text{ridge}}(\hat{\tau})$ and the third is the new predicted y -values $\hat{y} = X\hat{\beta}_{\text{ridge}}(\hat{\tau})$. In this section I will describe all the measured quantities and explain why they are selected.

For each combination of n , Σ_n and β described in the previous section we generate a matrix X using the described algorithm. Keeping the matrix X fixed, we perform 100 simulations, each with a new observation vector y . For each y we compute $\hat{\tau}^2$ using each of the three methods described in the previous

section, in addition to generalized cross validation. This gives four different estimates $\hat{\tau}^2$ leading to four different $\hat{\beta}_{\text{ridge}}(\hat{\tau})$. All these 8 estimates are stored, as well as the least squares estimate computed using the same X and y . For a given X we now have 100 estimates $\hat{\tau}^2$ for each of the four methods. Also, we have 100 estimates of β for each of the methods, as well as 100 least squares estimates.

First of all, since the main topic of this thesis is the estimation of parameters like τ^2 , we want to compare the different estimates $\hat{\tau}^2$. For each method we therefore calculate the sample mean and standard deviation of the estimates. We also calculate the sample mean and standard deviation of $1/\hat{\tau}^2$ for each method, remembering that $\lambda = \sigma^2/\tau^2$ is the parameter of interest in many articles, including the original ridge regression article.

We have seen in the previous section that the different methods may result in different expected values of $\hat{\tau}^2$ when $\Sigma_n \neq I_p$. Therefore, a small variance is not necessarily a sign that the estimate is good, since the bias may be large. To account for this problem, we also include an estimate of the mean square error of each estimate and its inverse. Since we have no hope of getting a better estimate of τ^2 than the maximum likelihood estimate of τ^2 when β is known, we use $\hat{\tau}_{MLE}^2$ rather than the true $\tau^2 = 4$ when calculating these quantities. Thus

$$\widehat{\text{MSE}}(\hat{\tau}^2) = \frac{1}{100} \sum_{j=1}^{100} (\hat{\tau}_{MLE}^2 - \hat{\tau}_j^2)^2,$$

with a similar expression for the $1/\hat{\tau}^2$ case.

Finally, we want to approximate the mean and variance of each estimate using $\tau_{n,X}^2$ and $v_{n,X}^2/n$ as defined in the theorems describing the limiting behaviour of each estimate. These quantities are therefore also calculated and included in the tables in the next section.

In the introduction of this chapter we focused on the mean square error of $\hat{\beta}_{\text{ridge}}(\tau)$. Since we want to use $\hat{\tau}$ instead of τ , this quantity is harder to compute. We therefore use

$$\widehat{\text{MSE}}(\hat{\beta}_{\text{ridge}}(\hat{\tau})) = \frac{1}{100} \sum_{j=1}^{100} (\hat{\beta}_{\text{ridge}}(\hat{\tau}_j) - \beta)^T (\hat{\beta}_{\text{ridge}}(\hat{\tau}_j) - \beta)$$

as an approximation.

Finally we want to say something about the ability to predict m new y -values using a new matrix X_m . We could choose to compare the quantity

$$\frac{1}{100} \sum_{j=1}^{100} \frac{1}{m} (X_m \hat{\beta}_{\text{ridge}}(\hat{\tau}_j) - X_m \beta)^T (X_m \hat{\beta}_{\text{ridge}}(\hat{\tau}_j) - X_m \beta)$$

or

$$\frac{1}{100} \sum_{j=1}^{100} (\hat{\beta}_{\text{ridge}}(\hat{\tau}_j) - \beta)^T \Sigma_m (\hat{\beta}_{\text{ridge}}(\hat{\tau}_j) - \beta),$$

where $\Sigma_m = X_m^T X_m / m$. Now, we see that the quantity depends on X_m and m only through Σ_m , and we have chosen to use $m = n$ and $\Sigma_m = \Sigma_n$, giving

$$\widehat{\text{MSE}}(X \hat{\beta}_{\text{ridge}}(\hat{\tau})) = \frac{1}{100} \sum_{j=1}^{100} (\hat{\beta}_{\text{ridge}}(\hat{\tau}_j) - \beta)^T \Sigma_n (\hat{\beta}_{\text{ridge}}(\hat{\tau}_j) - \beta),$$

This may seem artificial, but as demonstrated by our algorithm for finding an X giving a certain Σ_n , a lot of different X -matrices will give this result. Also, if we assume the x_i 's to be i.i.d. from some distribution with mean 0 and covariance matrix Σ , both Σ_n and Σ_m is close to Σ with high probability when n and m are large. In this situation we also have $\Sigma_n \approx \Sigma_m$ with high probability, so using $\Sigma_m = \Sigma_n$ should at least give us a good picture of the large sample behaviour.

3.3.3 Results

We have now explained the setup of our experiment and defined all the quantities to be compared. In this section we will present the results of the 12 experiments covering all possible combinations of n (2), β (2) and Σ_n (3). We recall that for each setup all the estimates are computed using the same matrix X , and that the same 100 y -vectors are used for all the different estimates. Thus for any given table presented in these section, all the estimates are computed using the exact same data.

Table 3.1: $\Sigma_n = \Omega_1$, $\beta = \beta_1$, $n = 20$

	$\hat{\tau}^2$	$\widehat{\text{MSE}}(\hat{\tau}^2)$	$\widehat{\text{MSE}}(1/\hat{\tau}^2)$	Appr. mean (sd)
LS	- (-)	-	0.041	- (-)
MoM	4.983 (0.321)	0.105	2E-04	4.930 (0.316)
CV	9E+04 (2E+05)	5E+10	0.067	4.930 (4.622)
GCV	6.934 (6.737)	48.942	0.010	- (-)
MML	4.983 (0.318)	0.103	2E-04	4.930 (0.314)

The first two experiments involves $\Sigma_n = \Omega_3 = I_p$ and $n = 20$, using β_1 and β_2 respectively. The results are similar, and we will use the results for β_1 as a basis for our qualitative assessments. The results concerning $\hat{\tau}^2$ are displayed in Table 3.1. First of all we want to compare the mean and variance (the numbers in parantheses are the standard deviations) of each estimate to each other and to the large sample approximations. For large n , the means should be approximately equal, but the variance of the cross validation estimate should be much larger than for the method of moments and the maximum marginal likelihood estimate. Here we only have $n = 20$, considered a small n compared to our $p = 10$, and the cross validation results are clearly not even close to their limiting distribution in terms of $\hat{\tau}^2$. The two other estimates however, have almost exactly the same mean and variance, and the sample results are very close to the large sample approximations (recall that for the method of moments the large sample results are exact). We do not know the large sample properties of the generalized cross validation estimate, but we see that it has a large variance and is thus quite unstable, but still far more stable than the ordinary cross validation estimate.

The high mean and variance of the CV estimate is a result of its tendency to choose very large values $\hat{\tau}^2$. Remembering that $\lambda = \sigma^2/\tau^2$ and that $\lambda = 0$ or $\tau^2 = \infty$ (or any sufficiently large value of $\hat{\tau}^2$) gives the ordinary least squares estimate, this simply means that for small n , cross validation often gives a β -estimate equal to the least squares estimate. Considering the function (3.16) to

be minimized by the CV estimate, this should not come as a surprise, as this is very close to the corresponding LS expression. Since $\hat{\tau}^2$ only appears in the form $1/\hat{\tau}^2$ in the formula for β , we have also included the mean square error of this quantity. Here we have also included the least squares estimate, as using least squares as a criterion for selecting $\hat{\tau}^2$ would always result in $1/\hat{\tau}^2 = 0$. This normalizes the behaviour of the CV estimate, but it is still the least successful estimate measured against the maximum likelihood estimate $\beta^T \beta / p = \beta_1^T \beta_1 / 10$. In fact, it performs even worse than the least squares estimate which is always $1/\hat{\tau}^2 = 0$. For n this small and $\Sigma_n = I_p$ it therefore seems like CV and GCV performs poorly, whereas our other two estimates are quite successful.

Table 3.2: $\Sigma_n = \Omega_2$, $\beta = \beta_1$, $n = 20$

	$\hat{\tau}^2$	$\widehat{\text{MSE}}(\hat{\tau}^2)$	$\widehat{\text{MSE}}(1/\hat{\tau}^2)$	Appr. mean (sd)
LS	- (-)	-	0.041	- (-)
Mom	1.700 (0.173)	10.463	0.157	1.686 (0.186)
CV	7E+03 (6E+04)	3E+09	0.017	7.475 (3.628)
GCV	8.838 (4.943)	39.459	0.012	- (-)
MML	4.535 (1.152)	1.469	0.004	4.930 (1.139)

To see the effect of using an ill-conditioned Σ_n we include the results of our third experiment. Now we still have $n = 20$, and we use Ω_2 and β_1 . Qualitative speaking, the four different combinations of $\Omega_2, \Omega_3, \beta_1$ and β_2 give approximately the same results, and the conclusions drawn from Table 3.2 are therefore valid for these experiments as well. We note that CV and GCV are still unstable compared to the other estimates, and that the large sample variance of CV is once again expected to be much larger than for MoM and MML. The most interesting feature about this table is the fact that the large sample means are so different. The three methods we have discussed are trying to estimate completely different quantities, and only MML is still close to the maximum likelihood estimate. To see the effect of changing the matrix Σ_n , you should compare Table 3.1 and Table 3.2. However, even though the MML and MoM estimates are now completely different, we observe that the large sample approximations of the mean and variance are still very good.

Table 3.3: $\Sigma_n = \Omega_2$, $\beta = \beta_2$, $n = 100$

	$\hat{\tau}^2$	$\widehat{\text{MSE}}(\hat{\tau}^2)$	$\widehat{\text{MSE}}(1/\hat{\tau}^2)$	Appr. mean (sd)
LS	- (-)	-	0.061	- (-)
MoM	1.070 (0.070)	8.786	0.481	1.071 (0.067)
CV	4.078 (0.921)	0.842	0.003	4.015 (0.904)
GCV	4.102 (0.738)	0.543	0.002	- (-)
MML	4.050 (0.357)	0.126	5E-04	4.033 (0.373)

We have performed all these six experiments for $n = 100$ as well, and there are a few points worth mentioning. Again there are several similarities between the results, and we will only provide the results for $\beta = \beta_2$ with $\Sigma_n = \Omega_2$ and

Table 3.4: $\Sigma_n = \Omega_3$, $\beta = \beta_2$, $n = 100$

	$\hat{\tau}^2$	$\widehat{\text{MSE}}(\hat{\tau}^2)$	$\widehat{\text{MSE}}(1/\hat{\tau}^2)$	Appr. mean (sd)
LS	- (-)	-	0.061	- (-)
MoM	4.037 (0.119)	0.014	5E-05	4.034 (0.128)
CV	2.500 (2.743)	9.796	0.137	0.462 (1.173)
GCV	2.556 (2.864)	10.299	0.130	- (-)
MML	4.236 (0.237)	0.097	3E-04	4.033 (0.204)

$\Sigma_n = \Omega_3$. The $\Sigma_n = \Omega_1 = I_p$ case is not very interesting, as the results are very similar to those of Table 3.1. The most striking difference between the $n = 20$ and the $n = 100$ situations are the fact that the CV estimates have become stable. The variance of the CV and GCV estimates are still relatively large compared to their means, but the improvement from $n = 20$ is significant. Also, for $\Sigma_n = \Omega_2$ it seems like $n = 100$ is large enough for the large sample mean and variance to be reasonable approximations, whereas for $\Sigma_n = \Omega_3$ n clearly needs to be even larger. Also, comparing the two tables we once again see how sensitive the estimates are to changes in the matrix Σ_n . The maximum marginal likelihood estimate is the only estimate with approximately the same mean and variance in the two cases. We also note that each of the three other estimates are close to MML in one of the tables, but not in the other one.

We have now looked at the results related to $\hat{\tau}^2$ itself, and it does have interesting interpretations on its own, but our main goal is often the estimation of β or some function thereof, like predicted y values for a new matrix X . We have already described the quantities to be measured, and we will here present some of the results.

We have seen already that when $\Sigma_n = \Omega_1 = I_p$, the MSE of $1/\hat{\tau}^2$ is small for all the estimation methods. This results in almost identical $\widehat{\text{MSE}}(\hat{\beta}_{\text{ridge}}(\hat{\tau}))$ and $\widehat{\text{MSE}}(X\hat{\beta}_{\text{ridge}}(\hat{\tau}))$ for all the estimation methods, and the results are therefore not presented here. Also, the results for β_1 and β_2 are qualitatively the same, and we will therefore restrict ourselves to the $\beta = \beta_2$ situation, with $\Sigma_n = \Omega_2$ and $\Sigma_n = \Omega_3$.

Table 3.5: $\Sigma_n = \Omega_2$, $\beta = \beta_2$, $n = 20$

	$\hat{\tau}^2$	$\widehat{\text{MSE}}(\hat{\beta}_{\text{ridge}}(\hat{\tau}))$	$\widehat{\text{MSE}}(X\hat{\beta}_{\text{ridge}}(\hat{\tau}))$
LS	- (-)	4.412	0.507
MoM	1.086 (0.157)	5.232	0.650
CV	7E+04 (3E+05)	4.216	0.497
GCV	5.539 (2.926)	4.211	0.493
MML	4.074 (0.862)	4.194	0.493

We first present the result for n small relative to p , with $n = 20$, displayed in Table 3.5 and Table 3.6. We find that the MSE of both the β -estimate and the predicted y -values are in most cases at least slightly smaller for the ridge parameter estimation methods than for the least squares estimate. However,

Table 3.6: $\Sigma_n = \Omega_3$, $\beta = \beta_2$, $n = 20$

	$\hat{\tau}^2$	$\widehat{\text{MSE}}(\hat{\beta}_{\text{ridge}}(\hat{\tau}))$	$\widehat{\text{MSE}}(X\hat{\beta}_{\text{ridge}}(\hat{\tau}))$
LS	- (-)	11.297	0.503
MoM	4.006 (0.304)	4.432	0.471
CV	4E+04 (3E+05)	7.572	0.493
GCV	8.891 (17.651)	7.779	0.489
MML	4.359 (0.421)	5.175	0.475

the performance of each estimate depends heavily on the form of Σ_n , and none of the methods seem to be uniformly best. The MML estimate performs well in both situations, whereas the MoM estimate is best for Ω_3 , but not even as good as the LS estimate for Ω_2 . Comparing the mean squared error for the β -estimate in both tables, we also see how the degree of improvement upon the least squares estimate is clearly connected to the form of Σ_n . This should not be surprising, given the results of the introductory section of this chapter.

We recall that the value of $\hat{\tau}^2$ stabilizes as n grows. At the other hand, the impact of the penalization is reduced when n grows, since it only appears inside the matrix $\Sigma_{n,\text{ridge}}$, and is there multiplied by a factor $1/n$. Therefore, the ridge estimate will become closer and closer to the ordinary least squares estimate as n increases. This may be seen in Table 3.7 and Table 3.8, where the difference between the MSE of the least squares method and the other methods have become smaller. It therefore seems that the improvement due to ridge regression is largest for small n , which is consistent with the motivation presented in the introduction.

Table 3.7: $\Sigma_n = \Omega_2$, $\beta = \beta_2$, $n = 100$

	$\hat{\tau}^2$	$\widehat{\text{MSE}}(\hat{\beta}_{\text{ridge}}(\hat{\tau}))$	$\widehat{\text{MSE}}(X\hat{\beta}_{\text{ridge}}(\hat{\tau}))$
LS	- (-)	0.781	0.094
MoM	1.070 (0.070)	0.873	0.105
CV	4.078 (0.921)	0.773	0.093
GCV	4.102 (0.738)	0.773	0.093
MML	4.050 (0.357)	0.772	0.093

Table 3.8: $\Sigma_n = \Omega_3$, $\beta = \beta_2$, $n = 100$

	$\hat{\tau}^2$	$\widehat{\text{MSE}}(\hat{\beta}_{\text{ridge}}(\hat{\tau}))$	$\widehat{\text{MSE}}(X\hat{\beta}_{\text{ridge}}(\hat{\tau}))$
LS	- (-)	2.169	0.100
MoM	4.037 (0.119)	1.727	0.098
CV	2.500 (2.743)	1.564	0.098
GCV	2.556 (2.864)	1.621	0.098
MML	4.236 (0.237)	1.796	0.098

The results of this experiment suggest that the generalized cross-validation gives more stable estimates than the ordinary cross validation. Also, the method of moments and the maximum marginal likelihood estimator are very stable in terms of the sample variance of $\hat{\tau}^2$, and the sample mean and variance are very close to their large sample approximations for both estimators. For the cross validation procedure, the mean and variance are being heavily influenced by the very large values of $\hat{\tau}^2$ corresponding to $\hat{\lambda} \approx 0$ and a least squares fit. Thus the large sample mean and variance are nowhere near their sample equivalents for $n = 20$, but for $n = 100$ we see some similarity. We note that the large sample approximations can in general not be used constructing for example confidence intervals, as they are based on the true but unknown β_0 .

We have seen that the performance of each estimator is closely connected to the form of Σ_n and β_0 , and that it is therefore in general not possible to say which of the estimators that will perform the best in one particular situation. In our experiment, it seems like the maximum marginal likelihood estimator is doing quite well in all the different categories, but we should be careful not to put too much weight on these few examples, as the other estimators have demonstrated that a change in Σ_n or β_0 may completely change the result. Also, both β_1 and β_2 are close to what we would expect from our prior, giving the method of moments and the MML an advantage, as they both explicitly use the prior assumption in their estimation procedures.

3.4 Discussion

We have now presented the ridge estimate and studied three different estimators for τ^2 . It seems clear that the performance of each estimator depends on the form of Σ_n and the true parameter vector β_0 , and that no one is uniformly better. However, if the true underlying distribution is normal, the MML estimator will perform best. Also, if the goal of the analysis is to predict new y -values, we expect the cross validation estimate to be working well, due to its form closely connected to prediction. The method of moments estimator on the other hand, is performing well in many situations and is very fast to compute, but we need to decide what to do when the formula gives a negative answer. The MML estimate is also computed relatively fast, while the computations of the CV estimator are far more time-consuming.

The methods mentioned in this chapter are not the only ones possible. As an example, we could consider a Bayesian approach with a specified prior distribution for τ . Also, we could pursue the idea of minimum mean square error. It may easily be shown that the τ^2 leading to the minimal mean square error is close to

$$\frac{\beta_0 \Sigma_n^{-2} \beta_0}{\text{Tr}(\Sigma_n^{-2})},$$

and we could try to find a suitable estimator based on this.

We should also note that all the methods in this chapter are valid also when $p > n$. This possibility was not included in the original ridge article, but the ridge estimate is defined for any $\lambda > 0$. We originally intended to study this situation, or alternatively a non-standard situation with $p < n$, for example $p/n \rightarrow c$ with c a constant such that $c \in (0, 1)$, or $n - p \rightarrow c$ for some $c \in \mathbb{N}$.

However, we did not have time to pursue these ideas, but as long as $p < n$ it may be possible to reuse some parts of our derivations.

Chapter 4

The zero-mixed ridge estimator

We have seen that the ridge regression may lead to better parameter estimates than the ordinary least squares method. The ridge estimate is biased, but the total variance of the parameters is so much smaller than in the least squares case that it often results in a smaller squared error. The ridge estimate has all the same non-zero components as the least squares estimate, but it is shorter in terms of the Euclidian norm.

In the chapter describing ordinary least squares, we introduced the concept of subset selection. When many covariates are available, the inclusion of all of them will often lead to too variable parameter estimates. Therefore, by choosing only a subset $S \subset \{1, \dots, p\}$ and forcing the remaining estimates to equal zero, we are able to reduce the total variance of the estimates. However, when the true parameter values are not zero this introduces a bias. Hopefully this bias is not too large, and the estimate will have smaller MSE than the full estimate. Since we of course do not know the true parameter vector, the mean square error can not be used as a criterion for choosing the best subset. Therefore, one will typically use a model selection criterion like the AIC (2.4) or BIC (2.5), possibly combined with greedy forward or backward selection.

In this chapter we will propose a new method combining the shrinkage of ridge regression with subset selection. The method is based on the relation between the ridge estimate and the posterior mode when β is given a normal prior distribution. We recall that

$$\begin{aligned}\hat{\beta}_{\text{ridge}}(\tau) &= \operatorname{argmax}_{\beta} \pi_{\tau}(\beta|\text{data}) \\ &= \operatorname{argmax}_{\beta} \{l_n(\beta) + \log(\pi_{\tau}(\beta))\},\end{aligned}$$

with

$$\pi_{\tau}(\beta) \propto \exp \left\{ -\frac{1}{2} \frac{\beta^T \beta}{\tau^2} \right\}.$$

Our idea is to use the same relation with a different prior in such a way that the final estimate may have one or more elements equal to zero. This is never the case for the ordinary least squares estimate or the ridge estimate, but it resembles the lasso estimate which will be discussed in the next chapter.

Since the goal is to obtain one or more zeroes in the final estimate, it seems natural to include this possibility in the prior distribution as well. A natural compromise between ridge regression and subset selection is therefore to let the prior distribution be such that each β_j has a certain probability α of being exactly zero, and probability $(1 - \alpha)$ of being distributed in the same way as for the ridge estimate, that is $\beta_j \sim N(0, \tau^2)$. Thus, for $\beta = (\beta_1, \dots, \beta_p)$ we let the components be i.i.d. with

$$\beta_j \sim \begin{cases} 0 & , \text{ with probability } \alpha \\ N(0, \tau^2) & , \text{ with probability } (1 - \alpha). \end{cases}$$

With this prior distribution, each of the 2^p subsets S of non-zero parameters has a certain probability $p_\alpha(S) = \alpha^{p-|S|}(1 - \alpha)^{|S|}$, and the $|S|$ non-zero parameters follow the normal distribution $N_{|S|}(0, \tau^2 I_{|S|})$. If we denote our prior distribution by $\pi_{\alpha, \tau}(\beta)$ we find that

$$\pi_{\alpha, \tau}(\beta) = \sum_S I_S(\beta) p_\alpha(S) \pi_{S, \tau}(\beta) \quad (4.1)$$

with $I_S(\beta) = \prod_{j \in S} (1 - \delta_0(\beta_j)) \prod_{j \notin S} \delta_0(\beta_j)$. This indicator variable ensures that each possible β does only result in one term on the right hand side. The distribution may be seen as a complicated distribution defined differently on 2^p subspaces of $\mathbb{R}^p$, or it may be seen as a set of 2^p different prior models, each with a certain prior probability. Each of the 2^p prior models will result in a corresponding posterior model with a different normal distribution and a different model probability. Both interpretations will be used in this chapter.

In (4.1) we have introduced the prior distributions $\pi_{S, \tau}(\beta)$ with respect to one particular subset S . These distributions are such that

$$\beta_j \sim \begin{cases} 0 & , \text{ for } j \notin S \\ N(0, \tau^2) & , \text{ for } j \in S, \end{cases}$$

and are given by

$$\pi_{S, \tau}(\beta) = \frac{1}{\sqrt{2\pi}^{|S|}} \frac{1}{\tau^{|S|}} \exp \left\{ -\frac{1}{2} \frac{\beta^T \beta}{\tau^2} \right\}.$$

Note that in this expression we have used $\beta^T \beta$ instead of the sum of the squared non-zero elements of β , but the use of the indicator variable $I_S(\beta)$ guarantees that these two quantities are the same whenever the corresponding term of the sum in (4.1) is non-zero.

Like in the ridge situation, we need to choose or estimate the parameters α and τ , but before we move on to this task, we will study the behaviour of the resulting β -estimate. We first present the estimate derived by following the idea of using the posterior mode, parallelling the ridge regression and also the lasso regression presented in the next chapter. This estimate will be denoted $\hat{\beta}_{\text{mix}}(\alpha, \tau)$ and will be discussed in the next section. This estimate will also be linked to the ordinary subset selection method presented in chapter 2.

In addition to the estimate $\hat{\beta}_{\text{mix}}(\alpha, \tau)$, we will also present a few other possible methods for estimating β based on the same prior distribution, but using other criterions than the posterior mode. These estimates are not connected to penalized maximum likelihood estimation and ridge regression in the same way, and will be treated in a separate section to avoid confusion.

4.1 The zero-mixed ridge estimate $\hat{\beta}_{\text{mix}}(\alpha, \tau)$

In order to follow the ridge regression and lasso regression as closely as possible, we have already suggested to let our estimate be given by

$$\hat{\beta}_{\text{mix}}(\alpha, \tau) = \underset{\beta}{\operatorname{argmax}} \{l_n(\beta) + \log(\pi_{\alpha, \tau}(\beta))\}.$$

In this section we will discuss the interpretation of this estimate and try to understand how it depends on the value of the parameters α and τ .

First of all we observe that for any value of β , exactly one of the 2^p terms of (4.1) will be non-zero, meaning that the logarithm of the sum is actually the sum of the logarithm of the individual terms if we keep the indicator variables $I_S(\beta)$. Thus

$$\hat{\beta}_{\text{mix}}(\alpha, \tau) = \underset{\beta}{\operatorname{argmax}} \left\{ l_n(\beta) + \sum_S I_S(\beta) \left(\log \left(\frac{p_S(\alpha)}{\sqrt{2\pi}^{|S|} \tau^{|S|}} \right) - \frac{1}{2} \frac{\beta^T \beta}{\tau^2} \right) \right\}.$$

If we let β_S^* be a β -vector with $\beta_j \neq 0$ for $j \in S$ and 0 otherwise, we may divide this optimization into a two-stage procedure. First, we find the optimal β_S^* and denote it by $\hat{\beta}_{S, \text{mix}}^*(\alpha, \tau)$. Then we find the subset S with the 'best' $\hat{\beta}_{S, \text{mix}}^*(\alpha, \tau)$ and denote it by $\hat{S}$. Now

$$\hat{\beta}_{\text{mix}}(\alpha, \tau) = \hat{\beta}_{\hat{S}, \text{mix}}^*(\alpha, \tau)$$

is our final estimate. Following this approach we find that

$$\begin{aligned} \hat{\beta}_{S, \text{mix}}^*(\alpha, \tau) &= \underset{\beta_S^*}{\operatorname{argmax}} \left\{ l_n(\beta_S^*) - \frac{1}{2} \frac{\beta_S^{*T} \beta_S^*}{\tau^2} + \log \left(\frac{p_S(\alpha)}{\sqrt{2\pi}^{|S|} \tau^{|S|}} \right) \right\} \\ &= \underset{\beta_S^*}{\operatorname{argmax}} \left\{ l_n(\beta_S^*) - \frac{1}{2} \frac{\beta_S^{*T} \beta_S^*}{\tau^2} \right\} \\ &= \hat{\beta}_{S, \text{ridge}}^*(\tau). \end{aligned}$$

Here $\hat{\beta}_{S, \text{ridge}}^*(\tau)$ is an expanded version of the ridge estimate obtained considering only the variables in S , with zeroes inserted for all $j \notin S$. If we let X_S be a matrix consisting of the columns X_j of X for which $j \in S$ and let $\Sigma_{n, S} = X_S^T X_S / n$, the non-zero components of the estimate $\hat{\beta}_{S, \text{ridge}}^*(\tau)$ are given by

$$\hat{\beta}_{S, \text{ridge}}^*(\tau) = \Sigma_{n, S, \text{ridge}}^{-1}(\tau) \frac{X_S^T Y}{n}, \quad (4.2)$$

with

$$\Sigma_{n, S, \text{ridge}}(\tau) = \Sigma_{n, S} + \frac{1}{n} \frac{\sigma^2}{\tau^2} I_{|S|}. \quad (4.3)$$

For a given subset S we therefore see a clear connection to the ridge estimate, and α is not involved at all.

We recall that in the ridge regression chapter we defined the penalized log-likelihood function:

$$H_n(\beta) = l_n(\beta) - \frac{1}{2} \frac{\beta^T \beta}{\tau^2}. \quad (4.4)$$

Now, we find that

$$\hat{S} = \operatorname{argmax}_S \left\{ H_n(\hat{\beta}_{S,\text{ridge}}^*(\tau)) + \log \left(\frac{p_S(\alpha)}{\sqrt{2\pi}^{|S|} \tau^{|S|}} \right) \right\}. \quad (4.5)$$

This looks a lot like the expression for AIC or BIC, except that the log-likelihood is replaced with a penalized version, and that the penalty term looks different. If we require $\alpha \in (0, 1)$ such that both α and $1 - \alpha$ is different from zero, we find that:

$$\begin{aligned} \hat{S} &= \operatorname{argmax}_S \left\{ 2H_n(\hat{\beta}_{S,\text{ridge}}^*(\tau)) - 2|S| \log \left(\frac{\alpha\tau\sqrt{2\pi}}{1-\alpha} \right) \right\} \\ &= \operatorname{argmax}_S \left\{ 2H_n(\hat{\beta}_{S,\text{ridge}}^*(\tau)) - c_{\text{pen}}(\alpha, \tau)|S| \right\}. \end{aligned}$$

with

$$c_{\text{pen}}(\alpha, \tau) = 2 \log \left(\frac{\alpha\tau\sqrt{2\pi}}{1-\alpha} \right).$$

This is very similar to AIC, but with H_n instead of l_n , and with a penalty parameter $c_{\text{pen}}(\alpha, \tau)$ different from the factor 2 of AIC. Our two-stage procedure therefore includes two different penalization processes, the first one performing shrinkage, the second model selection between all possible ridged submodels. Since the shrinkage part is the same as in the previous chapter, we will not discuss it here, but focus on the second step and particularly the rather unusual penalty term.

The new penalty parameter $c_{\text{pen}}(\alpha, \tau)$ may have any value in $\mathbb{R}$ dependent on α and τ , in particular it may be positive, zero or negative. The 'penalty term' may therefore either penalize, ignore or even reward a high value of $|S|$, but we will continue to call it a penalty term as it still penalizes 'unreasonable' choices of $|S|$. We note that when $c_{\text{pen}}(\alpha, \tau) \leq 0$, or equivalently

$$\frac{\alpha\tau\sqrt{2\pi}}{1-\alpha} \leq 1,$$

the full ridged model will always be chosen. This is consistent with viewing ridge regression as a special case of our new mixture with $\alpha = 0$. The above expression is not defined for $\alpha = 0$, but we observe that for any given τ we may choose α small enough to make the left hand side arbitrarily small, and in particular less than one. We also see from the original expression (4.5) that $\alpha = 0$ will result in $-\infty$ for all S except the full model, forcing the full model to be chosen.

We also observe that when $c_{\text{pen}}(\alpha, \tau) = 2$, or equivalently

$$\frac{\alpha\tau\sqrt{2\pi}}{1-\alpha} = e, \quad (4.6)$$

we get the AIC penalty, but since we are still considering the penalized log-likelihood, this is not equivalent to using AIC in the ordinary least squares setting. If we decide to choose α and τ subjectively instead of estimating them, we should consider the resulting value of the logged expression to ensure that we

obtain a suitably sparse model in the end. In this case, we should first choose τ according to our wish for shrinkage, and then choose α such that $c_{\text{pen}}(\alpha, \tau)$, is given a suitable value.

We have already mentioned the situation with α close to zero or even exactly zero, forcing us to choose the full model. The opposite situation arises when we let $\alpha \rightarrow 1$. A large value of α is used when we expect that very few (or even zero) of the β_j 's are non-zero. In this case we expect to have $c_{\text{pen}}(\alpha, \tau)$ to be positive and large such that adding an extra non-zero β_j is penalized heavily. A quick look at (4.6) confirms this, as an increasing α will make the logged term tend to infinity for any fixed value τ . At some point, the increasing α will make $c_{\text{pen}}(\alpha, \tau)$ so large that we get $|\hat{S}| = 0$.

The penalty parameter $c_{\text{pen}}(\alpha, \tau)$ is also dependent on the value of τ , so we will now assume $\alpha > 0$ (since the situation $\alpha = 0$ have already been covered) to be fixed, and see what happens when we vary τ . Recall that the penalized likelihood function is a function of τ , so the picture is a bit more complicated than in the previous case. However, we are able to see what happens in the extreme cases of a very small or very large τ . The simplest case is that of a very large τ . Then the penalized likelihood is approximately equal to the likelihood itself, and for τ above a certain threshold nearly unchanged by changes in τ . The penalty parameter $c_{\text{pen}}(\alpha, \tau)$ on the other hand increases towards infinity as τ increases, and it is possible to choose τ so large that the penalty term again ensures $|\hat{S}| = 0$.

This effect may seem strange, but we remember that the prior and thus also the posterior distribution is a mixture of two possibilities for each β_j : it is either zero or it follows a certain normal distribution. This posterior normal distribution will be influenced by the prior distribution with mean zero and variance τ^2 . If we choose τ large enough, the variance of the posterior distribution will still be very large due to this influence. Now, with very large τ , the ridge estimates, which are now approximately equal to the ordinary least squares estimates for the corresponding sets S , will not be spread out enough to defend the choice of any of the normal components, and the point mass in zero will be chosen for every component, giving $|\hat{S}| = 0$. This may easily be seen from the posterior distribution itself, presented later in the next section.

The last possibility we will consider is that of a very small τ . We first note that a very small τ will result in $\hat{\beta}_{S, \text{ridge}}^*(\tau)$ being very close to zero even in its non-zero components. Similar to the large τ situation above we find that τ may be chosen small enough to ensure a very small variance in the posterior distribution. Again, by forcing this variance to be sufficiently small, the normal possibility with small variance will be chosen over the zero possibility for all β_j , and the full set of parameters is chosen. We should note however, that even though the components are not exactly zero, they will all be very close to zero in this case.

We have now considered the extreme cases, and seen that α and τ control the complexity and value of our final estimate. For more sensible values of the parameters, the resulting effect is more difficult to quantify, and we have therefore made a few plots showing examples of how the resulting estimate is dependent on α and τ . In Figure 4.1 we see how the components of the β -estimate change when α moves from 0 to 1 and τ is held fixed. We see that the estimates are constant on an interval before they suddenly all jump to another value at the same time. Note in particular that when one $\hat{\beta}_j$ becomes zero, the

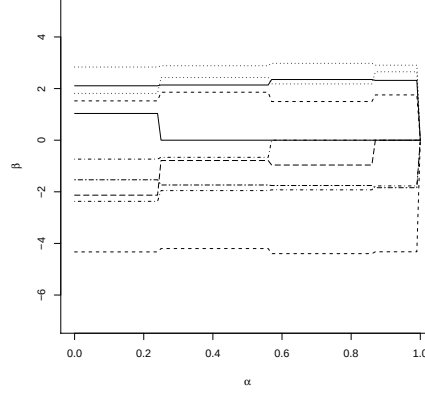


Figure 4.1: The paths of each component of $\hat{\beta}_{\text{mix}}(\alpha, 2.2)$. The sudden level changes should be interpreted as discontinuities.

others may move away from zero. One reason for this behaviour maybe that for highly correlated predictor variables, the decrease of one parameter estimate may be accounted for by the increase of another. What we see in the plot is consistent with the fact that α is not involved in estimating β_S^* at all for any fixed subset S , it does only affect the choice of $\hat{S}$.

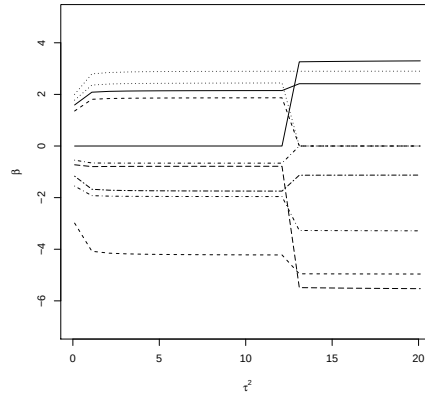


Figure 4.2: The paths of each component of $\hat{\beta}_{\text{mix}}(0.4, \tau)$. The sudden level changes at about $\tau^2 = 12$ should be interpreted as discontinuities.

The situation is more complicated when we fix α and vary τ . This situation is shown in Figure 4.2. Here we see that we have both jumps and continuous changes. This is because the parameter τ is involved in both stages of the optimization, first the ridge stage, corresponding to continuous changes, and then the subset selection stage, giving jumps. A combined picture is given in Figure 4.3, where the first component of $\hat{\beta}_{\text{mix}}(\alpha, \tau)$ is shown for a range of different α - and τ -values. Once again we observe that the function is piecewise

constant in α and piecewise continuous but not constant in τ . We also note that for a fixed τ^2 the function may be non-zero for small α , then zero for moderately-sized α and non-zero again for large α . This may seem strange, but we recall that this is only one component of the β -estimate, and that the change of this component is closely connected to the changes of the remaining components which is not displayed in this plot.

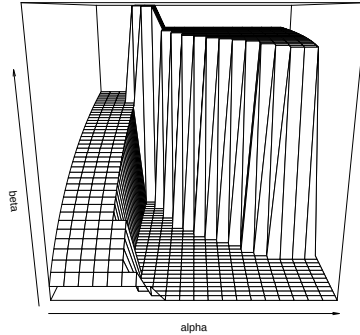


Figure 4.3: The estimate of the first component of $\hat{\beta}_{\text{mix}}(\alpha, \tau)$ displayed for a grid of possible pairs (α, τ^2) . The sudden level changes should be interpreted as discontinuities.

We have now discussed the behaviour of $\hat{\beta}_{\text{mix}}(\alpha, \tau)$, the maximizer of (4.5). The new and complicated part of this expression is a result of our desire to be able to compare models of different size. Parallel to the fact that the log-likelihood function itself can not be used to select a model of a suitable size, the penalized log likelihood in the ridge framework will always choose the full model. Therefore, we need some sort of selection criterion, and our proposed prior distribution has led to one such criterion. We have also seen that this criterion resembles AIC with the penalized log likelihood of the ridge model instead of the ordinary log likelihood, and that for any τ used in ridge regression, we may find a corresponding value α such that the penalization is the same as for the AIC. Using AIC on ridged models may therefore be seen as an approximation to our approach, ignoring the interpretation of the parameter α . The connection to ordinary subset selection using the AIC will be described more thoroughly in the next subsection.

4.1.1 The connection to ordinary subset selection

We have seen that the form of $\hat{\beta}_{\text{mix}}(\alpha, \tau)$ as defined by (4.5) resembles the parameter estimate obtained by ordinary subset selection using the AIC as the model selection criterion. In the chapter describing the ordinary least squares estimate, we also suggested that instead of using this standard procedure, we could impose a prior distribution on β and use the posterior mode as our estimate. This prior distribution should be such that $\beta_j = 0$ with a certain probability α , but such that a non-zero β_j would be almost equally likely to

have any value in $\mathbb{R}$. We saw that using an improper uniform prior was not an alternative, and suggested using an uninformative proper prior instead, more specifically a normal distribution with very large variance. In this section we will see how this idea is related to the estimate of this chapter, and what the resulting estimate looks like.

A normal prior does not make all values of β_j equally likely, but in real world applications we are often satisfied as long as it is reasonably flat inside some suitably large interval. We may therefore choose a very large value of τ , which we will denote by τ_0 , and let the prior be such that the components β_j are i.i.d. with

$$\beta_j \sim \begin{cases} 0 & , \text{ with probability } \alpha \\ N(0, \tau_0^2) & , \text{ with probability } (1 - \alpha). \end{cases}$$

This is exactly the same as the prior introduced in this chapter, with $\tau = \tau_0$. Using the results obtained earlier in this section we therefore get

$$\hat{\beta}_{\text{mix}}(\alpha, \tau_0) = \underset{\beta}{\operatorname{argmax}} \left\{ l_n(\beta) + \sum_S I_S(\beta) \left(\log \left(\frac{p_S(\alpha)}{\sqrt{2\pi}^{|S|} \tau_0^{|S|}} \right) - \frac{1}{2} \frac{\beta^T \beta}{\tau_0^2} \right) \right\}$$

as before, with

$$\hat{\beta}_{\text{mix}}(\alpha, \tau_0) = \hat{\beta}_{S, \text{mix}}^*(\alpha, \tau_0).$$

Here

$$\begin{aligned} \hat{\beta}_{S, \text{mix}}^*(\alpha, \tau_0) &= \underset{\beta_S^*}{\operatorname{argmax}} \left\{ l_n(\beta_S^*) - \frac{1}{2} \frac{\beta_S^{*T} \beta_S^*}{\tau_0^2} + \log \left(\frac{p_S(\alpha)}{\sqrt{2\pi}^{|S|} \tau_0^{|S|}} \right) \right\} \\ &= \underset{\beta_S^*}{\operatorname{argmax}} \left\{ l_n(\beta_S^*) - \frac{1}{2} \frac{\beta_S^{*T} \beta_S^*}{\tau_0^2} \right\} \\ &\approx \hat{\beta}_{S, LS}^*, \end{aligned}$$

with $\hat{\beta}_{S, LS}^*$ being an expanded version of the least squarese estimate obtained considering only the variables in S , with zeroes inserted for all $j \notin S$. The approximation in the last step is possible since τ_0 is chosen to be very large, and by (3.3) we must have

$$\hat{\beta}_{\text{mix}}(\alpha, \tau_0)^T \hat{\beta}_{\text{mix}}(\alpha, \tau_0) \leq \max_S \hat{\beta}_{S, LS}^{*T} \hat{\beta}_{S, LS}^* < \infty.$$

Thus, for τ_0 large enough, the effect of the penalty term will no longer be significant, and our estimate will be very close to the least squares estimate for the same subset S . Now if we assume $\alpha \in (0, 1)$, we get

$$\begin{aligned} \hat{S} &\approx \underset{S}{\operatorname{argmax}} \left\{ l_n(\hat{\beta}_{S, LS}^*) + \log \left(\frac{p_S(\alpha)}{\sqrt{2\pi}^{|S|} \tau_0^{|S|}} \right) \right\} \\ &= \underset{S}{\operatorname{argmax}} \left\{ 2l_n(\hat{\beta}_{S, LS}^*) - 2|S| \log \left(\frac{\alpha \tau_0 \sqrt{2\pi}}{1 - \alpha} \right) \right\} \\ &= \underset{S}{\operatorname{argmax}} \left\{ 2l_n(\hat{\beta}_{S, LS}^*) - c_{\text{pen}}(\alpha, \tau_0) |S| \right\}. \end{aligned}$$

with

$$c_{\text{pen}}(\alpha, \tau_0) = 2 \log \left(\frac{\alpha \tau_0 \sqrt{2\pi}}{1 - \alpha} \right).$$

Now, if we let

$$2 = c_{\text{pen}}(\alpha, \tau_0) = 2 \log \left(\frac{\alpha \tau_0 \sqrt{2\pi}}{1 - \alpha} \right).$$

or

$$\alpha_{\text{AIC}} = \frac{1}{1 + \frac{\tau_0 \sqrt{2\pi}}{e}},$$

this is clearly the same as subset selection with *AIC*. If we let α be defined like this, subset selection using AIC may therefore be seen as the limit of our new approach as $n \rightarrow \infty$. The last calculation may of course be done for any other constant than 2 as well, in order to penalize less or more than the AIC. Also, we remember that the BIC has penalty $|S| \log(n)$ instead of 2, and subset selection using BIC may therefore be seen as the limit of our approach as $n \rightarrow \infty$ with

$$\alpha_{\text{BIC}} = \frac{1}{1 + \frac{\tau_0 \sqrt{2\pi}}{\sqrt{n}}}.$$

We should note that both probabilities tend to 0 as $\tau_0 \rightarrow \infty$, but that $\alpha_{\text{AIC}} < \alpha_{\text{BIC}}$ for $n \geq 8$.

4.2 Other possible ways of estimating β

We have already introduced the estimate $\hat{\beta}_{\text{mix}}(\alpha, \tau)$ corresponding to the lasso and ridge estimate, but following the same idea with this new prior, we will propose a number of other possible estimates as well. They are all based on the posterior distribution of β , and we will therefore first derive the properties of this distribution.

Mye her

4.2.1 The posterior distribution of β given α and τ

With the help of a prior distribution and a likelihood, we are able to find not only a point estimate for β , but the entire posterior distribution of β given data. Particularly interesting in this case is the fact that the prior distribution contains a point mass in zero for each component β_j , leading to point masses in zero also in the posterior distribution. However, the probability of zero components will change according to the observed data.

The posterior distribution of β is given by

$$\begin{aligned} \pi_{\alpha, \tau}(\beta | \text{data}) &\propto L_n(\beta) \pi_{\alpha, \tau}(\beta) \\ &= \sum_S \left[I_S(\beta) p_\alpha(S) \exp \left\{ -\frac{1}{2} (y - X\beta)^T (y - X\beta) \right\} \pi_{S, \tau}(\beta), \right] \end{aligned}$$

or equivalently

$$\pi_{\alpha,\tau}(\beta|\text{data}) \propto \sum_S [I_S(\beta)p_\alpha(S)f_{\text{marg},S}(y,\tau)\pi_{S,\tau}(\beta|\text{data})]. \quad (4.7)$$

Here $f_{\text{marg},S}(y,\tau)$ is the marginal likelihood of y under the model corresponding to the subset S , that is, the model with $\beta_j = 0$ for $j \notin S$. We saw in the previous section that for a fixed subset S , we have a ridge problem, and the marginal likelihood is therefore given by an expression similar to that of (3.23):

$$f_{\text{marg},S}(y,\tau) = \frac{1}{\sqrt{2\pi}^n} \frac{1}{\sigma^n} \exp \left\{ -\frac{1}{2} \frac{y^T y}{\sigma^2} \right\} \frac{1}{n^{|S|/2}} \frac{\sigma^{|S|}}{\tau^{|S|}} |\Sigma_{n,S,\text{ridge}}(\tau)|^{-1/2} \\ \exp \left\{ \frac{n}{2\sigma^2} \hat{\beta}_{S,\text{ridge}}^T(\tau) \Sigma_{n,S,\text{ridge}}(\tau) \hat{\beta}_{S,\text{ridge}}(\tau) \right\},$$

with $\hat{\beta}_{S,\text{ridge}}(\tau)$ and $\Sigma_{n,S,\text{ridge}}(\tau)$ as defined by (4.2) and (4.3) respectively. In the special case of $|S| = 0$, all the terms dependent of S disappear.

The posterior distribution $\pi_{S,\tau}(\beta|\text{data})$ corresponding to the subset S is such that $\beta_j \sim 0$ for $j \notin S$ while

$$\beta_S \sim N_{|S|} \left(\hat{\beta}_{S,\text{ridge}}(\tau), \frac{\sigma^2}{n} \Sigma_{n,S,\text{ridge}}^{-1}(\tau) \right).$$

Thus the zero components in each of the 2^p distributions are the same as before, while the non-zero part still follows a normal distribution, but with mean and variance adjusted according to the data. Remembering that the indicator variable $I_S(\beta)$ guarantees that $\beta_j = 0$ for all $j \notin S$ whenever the term corresponding to S is positive, we may write

$$\pi_{S,\tau}(\beta|\text{data}) = \frac{1}{\sqrt{2\pi}^{|S|}} \frac{1}{n^{|S|/2}} \frac{\sigma^{|S|}}{\tau^{|S|}} |\Sigma_{n,S,\text{ridge}}(\tau)|^{-1/2} \\ \exp \left\{ -\frac{n}{2\sigma^2} (\beta_S - \hat{\beta}_{S,\text{ridge}}(\tau))^T \Sigma_{n,S,\text{ridge}}(\tau) (\beta_S - \hat{\beta}_{S,\text{ridge}}(\tau)) \right\},$$

ignoring the β_j with $j \notin S$.

Since each $\pi_{S,\tau}(\beta|\text{data})$ integrates to one, we find that the missing integration constant of (4.7) is

$$c(\alpha, \tau, y) = f_{\text{marg}}(y, \alpha, \tau) = \sum_S p_\alpha(S) f_{\text{marg},S}(y, \tau).$$

We now see that not only have the mean and variance of each normal distribution changed, but the probability corresponding to each distribution is changed as well, from $p_\alpha(S)$ to

$$\frac{p_\alpha(S) f_{\text{marg},S}(y, \tau)}{\sum_S p_\alpha(S) f_{\text{marg},S}(y, \tau)}.$$

Using the posterior distribution above, we may easily compute the marginal posterior distributions as well. We started with a point mass defined by $Pr(\beta_j = 0) = \alpha$ and a normal distribution $N(0, \tau^2)$. The posterior distribution consists of a point mass with a different probability and a sum of 2^{p-1} normal distributions with different means and variances. The derivation of the marginal posterior distribution will be given below:

We first observe that the set of all possible subsets S may be divided into two disjoint subsets, based on whether $j \in S$ or not. For all the sets not including j , the marginal distribution is a point mass in zero, while all sets including j result in a normal marginal distribution. Thus the posterior probability that $\beta_j = 0$ is the probability of being inside the first group of distributions, more precisely

$$p_{\beta_j=0,\alpha,\tau}(\beta_j|\text{data}) = \frac{\sum_{S:j \notin S} p_\alpha(S) f_{\text{marg},S}(y, \tau)}{\sum_S p_\alpha(S) f_{\text{marg},S}(y, \tau)}. \quad (4.8)$$

The other component of the posterior marginal distribution is a sum of normal distributions, and is given by

$$\pi_{\beta_j \neq 0,\alpha,\tau}(\beta_j|\text{data}) = \int \frac{\sum_{S:j \in S} p_\alpha(S) f_{\text{marg},S}(y, \tau) \pi_{S,\tau}(\beta|\text{data})}{\sum_{S:j \in S} p_\alpha(S) f_{\text{marg},S}(y, \tau)} d\beta_{S,-j}$$

where $\beta_{S,-j}$ is β_S without β_j . The marginal distribution of this mixture equals the sum of marginal distributions for all the 2^{p-1} components. Each of these distributions are easy to find since the marginal distribution of a multivariate normal distribution $N(\mu, \Sigma)$ is easily expressed in terms of μ_j and $\Sigma_{j,j}$. Thus the marginal distribution of nonzero β_j 's is

$$\pi_{\beta_j \neq 0,\alpha,\tau}(\beta_j|\text{data}) = \frac{\sum_{S:j \in S} p_\alpha(S) f_{\text{marg},S}(y, \tau) \pi_{\beta_j \neq 0,S,\tau}(\beta_j|\text{data})}{\sum_{S:j \in S} p_\alpha(S) f_{\text{marg},S}(y, \tau)} \quad (4.9)$$

with

$$\begin{aligned} \pi_{\beta_j \neq 0,S,\tau}(\beta_j|\text{data}) &= \frac{1}{\sqrt{2\pi}} \frac{\sqrt{n}}{\sigma} \frac{1}{\sqrt{[\Sigma_{n,S,\text{ridge}}^{-1}(\tau)]_{(j,j)}}} \\ &\quad \exp \left\{ -\frac{n}{2\sigma^2} \frac{(\beta_j - [\hat{\beta}_{S,\text{ridge}}(\tau)]_j)^2}{[\Sigma_{n,S,\text{ridge}}^{-1}(\tau)]_{(j,j)}} \right\}. \end{aligned}$$

Note that $[\Sigma_{n,S,\text{ridge}}^{-1}(\tau)]_{(j,j)}$ and $[\hat{\beta}_{S,\text{ridge}}(\tau)]_j$ are the elements corresponding to β_j , not necessarily the element in the j -th or (j,j) -th position of the vector or matrix.

The marginal posterior distribution of the parameter β_j is now given by

$$\beta_j = \begin{cases} 0 & , \text{ with probability } p_{\beta_j=0,\alpha,\tau}(\beta_j|\text{data}) \\ Z & , \text{ with probability } (1 - p_{\beta_j=0,\alpha,\tau}(\beta_j|\text{data})), \end{cases}$$

where Z is distributed according to $f(Z) = \pi_{\beta_j \neq 0,\alpha,\tau}(\beta_j|\text{data})$ as defined by (4.9). Two examples of such marginal distributions are given in Figure 4.4 and Figure 4.5. We note in particular that the continuous part of the distribution may be multimodal.

4.2.2 Alternative ways of estimating β

The posterior distribution of β and its marginals give rise to a couple of other possible ways of estimating β . For example we could try a proper Bayesian estimate with $\hat{\beta} = E(\beta|\text{data})$, given by

$$\hat{\beta}_{\text{mix,mean}}(\alpha, \tau) = E(\beta|\alpha, \tau \text{data}) = \frac{\sum_S p_\alpha(S) f_{\text{marg},S}(y, \tau) \hat{\beta}_{S,\text{ridge}}^*(\tau)}{\sum_S p_\alpha(S) f_{\text{marg},S}(y, \tau)}.$$

This estimate is easily calculated and is different from $\hat{\beta}_{\text{mix}}(\alpha, \tau)$, with the most striking difference being that it has only non-zero components. The estimate may therefore be seen as a two-parameter alternative to ridge regression, resulting in a different kind of shrinkage. The difference will be small for α close to 1, and will increase as α decreases. However, since the possibility of zero estimates was an important part of the motivation for our new prior distribution, we will not discuss the use of the posterior mean any further.

Another possible Bayesian approach is to choose the model with highest posterior probability. We recall that our prior distribution may be interpreted as a set of 2^p possible models M_S , each having a corresponding model probability $p_\alpha(S)$. We have seen that the posterior distribution has the same form, but with new probabilities and with adjusted mean and variance for the normal distribution of each model. From the previous section we know that

$$Pr(M_S|\text{data}) = \frac{p_\alpha(S)f_{\text{marg},S}(y, \tau)}{\sum_S p_\alpha(S)f_{\text{marg},S}(y, \tau)},$$

and selecting the model with the highest posterior probability is the same as setting

$$\widehat{M}_S = M_{\hat{S}_{\text{mod}}}$$

with

$$\hat{S}_{\text{mod}} = \underset{S}{\operatorname{argmax}} \{p_\alpha(S)f_{\text{marg},S}(y, \tau)\}.$$

When the model has been chosen, we need to obtain a point estimate of β under the chosen model $M_{\hat{S}_{\text{mod}}}$. The corresponding posterior distribution is such that $\beta_j \sim 0$ for $j \notin \hat{S}_{\text{mod}}$ while

$$\beta_{\hat{S}_{\text{mod}}} \sim N_{|\hat{S}_{\text{mod}}|} \left(\hat{\beta}_{\hat{S}_{\text{mod}}, \text{ridge}}(\tau), \frac{\sigma^2}{n} \Sigma_{n, \hat{S}_{\text{mod}}, \text{ridge}}^{-1}(\tau) \right).$$

For this distribution the mean and the mode are equal, and the natural estimate is therefore

$$\hat{\beta}_{\text{mix}, \text{mod}}(\alpha, \tau) = \hat{\beta}_{\hat{S}_{\text{mod}}, \text{ridge}}^*(\tau).$$

This estimate resembles our original estimate, but the final model $\hat{S}_{\text{mod}}$ is chosen based on a different criterion.

A third possible estimation method is based on the marginal posterior distributions as shown in Figure 4.4 and Figure 4.5. We may hope that a high marginal posterior probability of $\beta_j = 0$ is often the result of a true parameter value (close to) zero. Thus we might select a threshold c_{thresh} and let

$$\hat{S}_{\text{thresh}} = \{j : p_{\beta_j=0, \alpha, \tau}(\beta_j|\text{data}) < c_{\text{thresh}}\}.$$

The β_j 's with $j \in \hat{S}_{\text{thresh}}$ could be estimated as above, or we could even use the posterior mean or mode of the marginal distribution of each component. However, this approach will probably perform worse than the other methods when the variables are highly correlated, and the introduction of a third parameter c_{thresh} makes it even more complicated to derive the properties of this estimate.

We have now introduced our main parameter estimate $\hat{\beta}_{\text{mix}}(\alpha, \tau)$ and a few other possible estimates, but we have not yet discussed in detail how the parameters α and τ should be chosen. In the next section we will present the same

three approaches as in the ridge chapter, each of them estimating both parameters at the same time. Estimating one parameter at the time would probably lead to simpler expressions and less time-consuming computation, but also less accurate estimates, and we will not discuss such methods in this chapter.

4.3 Estimating the parameters α and τ

In this section we will look at three possible methods for estimating the parameters α and τ . The three methods will be the same as in the ridge section, but with our new prior it is more difficult to derive the properties of the estimates. We will therefore not derive limiting theorems corresponding to the theorems derived in the ridge chapter, but we will try to gain some insight about the estimates in other ways by studying various expressions and using the experience from a few numerical experiments.

4.3.1 A simple approach based on the method of moments

Once again we start by deriving a conceptually and computationally simple estimate based on the method of moments. If we were able to show that it had nice properties, it would perhaps be an interesting estimate itself, but even if the two other estimates were shown to be significantly better, this estimate could possibly be used as a reasonable start value in their complicated optimization procedures.

This method is based on the prior itself, and we estimate the parameters of the prior distribution by assuming for a moment that the elements of β_0 's do really come from this distribution. Following the lines of chapter 3, we will once again compute moments of y_i , trying to find simple expressions involving α and τ . We recall that

$$y_i = x_i^T \beta + \varepsilon_i,$$

with all ε_i 's independent and with

$$\varepsilon_i \sim N(0, \sigma^2),$$

giving

$$y_i | x_i, \beta_0 \sim N(x_i^T \beta_0, \sigma^2).$$

The first moment of y_i is $E(y_i | x_i, \beta_0) = x_i^T \beta_0$, giving $E(y_i | x_i) = 0$. Clearly, this will not help us finding reasonable estimates for α and τ , and we move on to the second order moment. We recall that the second order moment was the one used in the pure ridge case, and we expect it to work here as well. We find that

$$E(y_i^2 | x_i, \beta_0) = (x_i^T \beta_0)^2 + \sigma^2,$$

giving

$$E(y_i^2 | x_i) = x_i^T E(\beta \beta^T) x_i + \sigma^2 = (1 - \alpha) \tau^2 x_i^T x_i + \sigma^2.$$

The second order moment turns out to be useful, just as we expected it to. We now have one expression involving the two parameters, but to find estimates for α and τ we need one more equation. One idea might be to use the same structure as above, but with different weights $v_1, \dots, v_n$ on the observations, but unfortunately this will only give us one more equation in $(1 - \alpha) \tau^2$. Two

equations with only one unknown parameter give either zero or infinitely many solutions, and the weighted version is therefore of no interest, except perhaps as an alternative to our unweighted version. We continue computing higher order moments and see that

$$E(y_i^3|x_i, \beta_0) = (x_i^T \beta)^3 + 3x_i^T \beta_0 \sigma^2,$$

giving $E(y_i^3|x_i) = 0$. This is of no help to us, but we finally find that

$$E(y_i^4|x_i, \beta_0) = (x_i^T \beta_0)^4 + 6(x_i^T \beta_0)^2 \sigma^2 + 3\sigma^4,$$

giving

$$E(y_i^4|x_i) = 3(1-\alpha)\tau^4\|x_i\|_4^4 + (1-\alpha)^2\tau^4[\|x_i\|_2^4 - \|x_i\|_4^4] + 6\sigma^2(1-\alpha)\tau^2\|x_i\|_2^2 + 3\sigma^4.$$

Replacing the population mean by the sample mean in the first equation leads to the following:

$$1 - \hat{\alpha} = \frac{\frac{1}{n}\|y\|_2^2 - \sigma^2}{\hat{\tau}^2 \frac{1}{n} \sum_{i=1}^n \|x_i\|^2} = \frac{\hat{\tau}_{\text{ridge}}^2}{\hat{\tau}^2} \quad (4.10)$$

with $\hat{\tau}_{\text{ridge}}^2$ as defined by (3.11). Thus our two new estimates will be closely linked to our familiar pure ridge estimate (this holds of course for the parameters themselves as well). Inserting the above expression for $(1 - \hat{\alpha})$ into the second equation and using the sample mean here as well gives

$$\begin{aligned} \frac{1}{n}\|y\|_4^4 &= 3\hat{\tau}^2\hat{\tau}_{\text{ridge}}^2 \frac{1}{n} \sum_{i=1}^n \|x_i\|_4^4 + \hat{\tau}_{\text{ridge}}^4 \left(\frac{1}{n} \sum_{i=1}^n \|x_i\|_2^4 - \frac{1}{n} \sum_{i=1}^n \|x_i\|_4^4 \right) \\ &\quad + 6\sigma^2\hat{\tau}_{\text{ridge}}^2 \frac{1}{n} \sum_{i=1}^n \|x_i\|_2^2 + 3\sigma^4, \end{aligned}$$

leading to

$$\begin{aligned} \hat{\tau}^2 &= \frac{\frac{1}{n}\|y\|_4^4 - \hat{\tau}_{\text{ridge}}^4 \left(\frac{1}{n} \sum_{i=1}^n \|x_i\|_2^4 - \frac{1}{n} \sum_{i=1}^n \|x_i\|_4^4 \right)}{3\hat{\tau}_{\text{ridge}}^2 \frac{1}{n} \sum_{i=1}^n \|x_i\|_4^4} \\ &\quad - \frac{6\sigma^2\hat{\tau}_{\text{ridge}}^2 \frac{1}{n} \sum_{i=1}^n \|x_i\|_2^2 + 3\sigma^4}{3\hat{\tau}_{\text{ridge}}^2 \frac{1}{n} \sum_{i=1}^n \|x_i\|_4^4}. \end{aligned}$$

Now, $\hat{\alpha}$ may be computed using (4.10).

With this approach we should be careful and always make sure $\hat{\alpha} \in [0, 1]$ and $\hat{\tau}^2 \geq 0$. This is not always the case using the above expressions, and such 'illegal' estimates have occurred quite frequently during my numerical experiments. In these cases we might for example let $\hat{\alpha} = 0$ whenever the result of the expression is negative. However, since we are estimating two quantities simultaneously, we should consider changing the other estimate as well to satisfy (4.10).

The expressions we have derived are not as simple as in the ridge case, and we will not derive the mean, variance or limiting distribution of our estimates. It may not be too hard in this case, but we are not able to derive corresponding results for the two other estimators, and are therefore unable to compare them in terms of these quantities. We will therefore not make any attempt in this case either, and simply conclude that this estimate is at least easily computable, requiring only simple arithmetic, as opposed to the next to approaches which require numerical optimization of rather complicated functions.

4.3.2 Cross validation

Another possible approach for estimating α and τ is the cross validation. In this case our estimate will be given by

$$(\hat{\alpha}, \hat{\tau}) = \underset{\alpha \in [0,1], \tau \geq 0}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n \left[y_i - x_i^T \hat{\beta}_{\text{mix},(i)}(\alpha, \tau) \right]^2 \right\}.$$

Here $\hat{\beta}_{\text{mix},(i)}(\alpha, \tau)$ is again the β -estimate obtained by leaving out the i -th observation. Remembering the form of the estimate $\hat{\beta}_{\text{mix}}(\alpha, \tau)$ we see that this function has discontinuities in both directions, and we are not able to approximate it easily like in the ridge case.

The cross validation estimator may however easily be implemented using a numerical optimizer. Here we should be careful to make sure that the numerous discontinuities are not causing any problems, and the safest solution may be to minimize over a suitably chosen grid. We should note that for every pair (α, τ) we have to compute n corresponding estimates $\hat{\beta}_{\text{mix},(i)}(\alpha, \tau)$, each of them requiring the fit and comparison of 2^p ridge models. We therefore expect the computing time to be approximately 2^p times as long for each point (α, τ) here as for one point τ for ordinary ridge regression,. Also, since we are optimizing in two dimensions, we expect the number of different points (α, τ) we need to consider to be larger than the number of different τ -values considered in the ridge case. Even if we use the fact that the 2^p ridge models are independent of α and therefore only need to be computed once for each τ , the cross validation is so time-consuming that it is impossible to use in practise even for quite small n and p . Various approximations may be implemented, for example replacing the comparison of all 2^p ridge models by a greedy forward or backward search. Also, we could consider only a few different points chosen in a suitable way instead of a complete grid. I will provide one example of such a set of points below.

Like for the two other estimation methods we assume for a moment that the β_j 's do really come from our prior. If the β_j 's were known, we could have used the maximum likelihood estimates for α and τ , which may be shown to have the form

$$\hat{\alpha}_{\text{MLE}} = \frac{\sum_{j=1}^p \delta_0(\beta_j)}{p}$$

and

$$\hat{\tau}_{\text{MLE}}^2 = \sum_{j: \beta_j \neq 0} \frac{\beta_j^2}{p(1 - \hat{\alpha})}.$$

We do of course not know the β_j 's, but we may try to use the least squares estimate and a set of thresholds $t_1, \dots, t_k$ to obtain a set of approximations

$$\alpha_i = \frac{\sum_{j=1}^p I(\beta_{LS,j} < t_i)}{p}$$

and

$$\tau_i^2 = \sum_{j: \beta_{LS,j} \geq t_i} \frac{\beta_{LS,j}^2}{p(1 - \alpha_i)}.$$

The idea is that if $\beta_j = 0$ this will hopefully be reflected in a small value of $\beta_{LS,j}$, at least for large n . Letting the t_i 's equal the $\hat{\beta}_{LS,i}$'s we get a set of p possible points (α_i, τ_i^2) we could try to compare using cross validation.

We have not studied the effect of these possible simplifications and approximations of the cross validation procedure, except for the grid approach, which is also too time-consuming for any reasonably spaced grid. We therefore conclude that exact cross-validation is not useful in this setting, except perhaps for very small p and n .

4.3.3 The maximum marginal likelihood estimator

The last estimator we will consider is the one based on the maximum marginal likelihood of y , given by

$$\begin{aligned} (\hat{\alpha}, \hat{\tau}) &= \operatorname{argmax}_{\alpha \in [0,1], \tau \geq 0} f_{\text{marg}}(y, \alpha, \tau) \\ &= \operatorname{argmax}_{\alpha \in [0,1], \tau \geq 0} \int L_n(\beta) \pi_{\alpha, \tau}(\beta) d\beta. \end{aligned}$$

We recall that the posterior distribution of β is given by

$$\begin{aligned} \pi_{\alpha, \tau}(\beta | \text{data}) &\propto L_n(\beta) \pi_{\alpha, \tau}(\beta) \\ &\propto \sum_S [I_S(\beta) p_\alpha(S) f_{\text{marg}, S}(y, \tau) \pi_{S, \tau}(\beta | \text{data})], \end{aligned}$$

and we thus have

$$\begin{aligned} (\hat{\alpha}, \hat{\tau}) &= \operatorname{argmax}_{\alpha \in [0,1], \tau \geq 0} f_{\text{marg}}(y, \alpha, \tau) \\ &= \operatorname{argmax}_{\alpha \in [0,1], \tau \geq 0} \sum_S [p_\alpha(S) f_{\text{marg}, S}(y, \tau)], \end{aligned} \quad (4.11)$$

since all the distributions $\pi_{S, \tau}(\beta | \text{data})$ integrate to one. Here

$$\begin{aligned} f_{\text{marg}, S}(y, \tau) &= \frac{1}{\sqrt{2\pi}^n} \frac{1}{\sigma^n} \exp \left\{ -\frac{1}{2} \frac{y^T y}{\sigma^2} \right\} \frac{1}{n^{|S|/2}} \frac{\sigma^{|S|}}{\tau^{|S|}} |\Sigma_{n, S, \text{ridge}}(\tau)|^{-1/2} \\ &\quad \exp \left\{ \frac{n}{2\sigma^2} \hat{\beta}_{S, \text{ridge}}^T(\tau) \Sigma_{n, S, \text{ridge}}(\tau) \hat{\beta}_{S, \text{ridge}}(\tau) \right\} \end{aligned}$$

as before, with $\hat{\beta}_{S, \text{ridge}}(\tau)$ and $\Sigma_{n, S, \text{ridge}}(\tau)$ as defined by (4.2) and (4.3) respectively.

Recalling that $p_S(\alpha) = \alpha^{p-|S|}(1-\alpha)^{|S|}$, we see that the function in (4.11) is a polynomial of degree p in α , and an even more complicated function of τ , so we have no hope of finding its maximizer analytically. However, it is easily solved using numerical optimization (for example the `nlm` procedure in R). All the terms of the form $f_{\text{marg}, S}(y, \tau)$ are growing rapidly with n , so numerical optimization requires a simple normalization of these terms. The obvious choice of using the same expression (except the $\tau^{|S|}$ term) with S being the full set and $\tau = \infty$, corresponding to the ordinary least squares solution, seems to be working well. As the function grows with p as well it may also be useful to divide by the absolute value of $f_{\text{marg}}(y, \alpha_0, \tau_0)$ for some (α_0, τ_0) .

Although we are not able to solve this problem analytically, we should use a few moments to inspect the function to be maximized. We note that the part of (4.11) not including α is exactly the expression to be maximized if we want to find the best ridge parameter τ for estimating one particular β_S using MML. Thus, for a fixed α , our problem corresponds to finding a ridge parameter τ that is optimal over the entire set of $2^p - 1$ ridge submodels with at least one parameter, with each submodel given a certain weight dependent on α .

For each possible value of α , there exists a corresponding optimal $\hat{\tau}(\alpha)$. Thus the maximization may be implemented as a two-stage procedure, first computing $\hat{\tau}(\alpha)$ for each α by maximizing the sum of weighted ridge marginal likelihoods, and then maximizing the entire function over all values of α with $\hat{\tau}(\alpha)$ inserted for τ . We note that all models with the same number of non-zero parameters are given the same weight, and that $\hat{\alpha}$ is therefore chosen based on the average performance of all possible models of the same size $|S|$.

Again the optimization requires the fit of 2^p ridge models for each τ , but we avoid the factor n from the cross validation procedure. Also, in this case we do not have to find a $\hat{S}$, we simply use a sum of 2^p continuous function in α and τ , making this estimation method much more suitable for numerical optimization procedures.

4.4 Discussion

We will not give any numerical examples in this chapter, but a small experiment involving the zero-mixed ridge estimator will be provided at the end of the next chapter. In this section we will only provide a very brief summary of this chapter and discuss a few positive and negative features of the estimator.

In this chapter we have introduced a new regularized regression estimator which we have called the zero-mixed ridge estimator. The estimator performs both variable selection and shrinkage of β , and may therefore be seen as an alternative to the lasso estimator presented in the next chapter. Since this estimator is using two tuning parameters, we expect it to do better than the lasso in some situations, as the lasso has only one parameter dictating both selection and shrinkage. Our new estimator is thus more flexible than the lasso, and it is somewhat more intuitive in the sense that it provides actual point masses in zero, whereas the lasso only has singular values leading to estimated zeroes. In particular, in order to obtain zeroes in the final estimate, we include actual zeroes in the prior distribution. The lasso on the other hand, uses a continuous prior distribution and therefore has $Pr(\beta_j = 0) = 0$ both in the prior and in the posterior distribution.

Our point masses may look appealing, but they come at a cost. Our β -estimates are not continuous in α and τ , and therefore a small change in the data may result in a completely different model being chosen, since the data are being used to estimate α and τ . A similar effect is present for ordinary subset selection, but is avoided by the lasso due to its continuous prior distribution.

The estimation of β for a given pair (α, τ) is relatively fast for moderately sized p and n . The method of moments estimator and the MML estimator are also fast, whereas the cross validation has turned out to be too slow to be of much interest. Since the cross validation is so computationally heavy, we have not had the chance to perform any large experiments studying its

performance. The method of moments provide 'illegal' estimates quite often, typically the estimated $\hat{\alpha}$ is outside $[0, 1]$, and our recommendation is therefore to use the MML estimator, and to implement it using a constrained optimization procedure, for example `nlminb` in R.

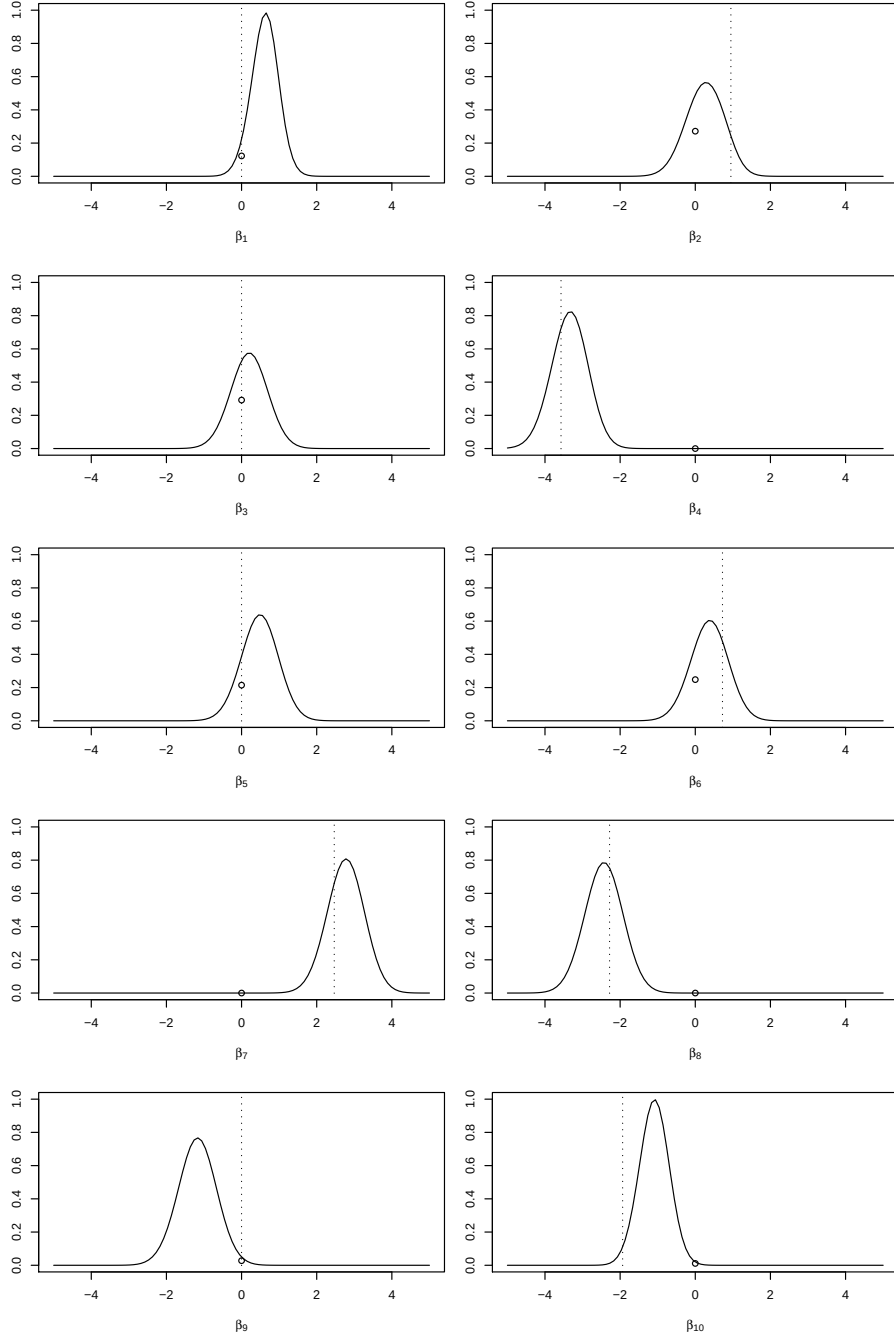


Figure 4.4: An example of the posterior distribution of each component of β with $p = 10$. In each plot the vertical dotted line shows the true value of β_j , while the point represents the point mass in 0. The area under the curve and the second coordinate of the point depict the probability of the continuous distribution and the point mass respectively.

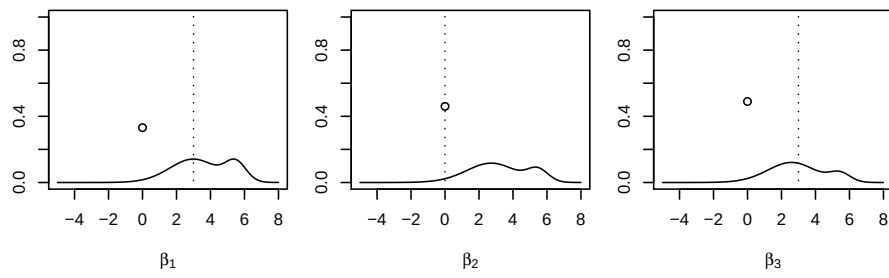


Figure 4.5: An example of the posterior distribution of each component of β with $p = 3$. In each plot the vertical dotted line shows the true value of β_j , while the point represents the point mass in 0. The area under the curve and the second coordinate of the point depict the probability of the continuous distribution and the point mass respectively.

Chapter 5

The lasso

We have now seen two different methods for doing model selection for linear regression models. The first one fits a linear model for each subset $S \subset \{1, \dots, p\}$ and uses a model selection criterion like the AIC (2.4) or BIC (2.5) to select the final model. The other method, which we have called the zero-ridge estimator, was presented in the previous chapter and is based on a penalized likelihood with the penalty term constructed from a certain prior distribution for β involving a point mass in zero for each β_j . In this chapter we will present a third alternative, known as the 'lasso', introduced by Tibshirani (1996). The word lasso is an acronym for 'least absolute shrinkage and selection operator', and as the name suggests, it is performing both variable selection and shrinkage of the remaining parameters. This was also the case in the previous chapter, whereas the ordinary subset selection with least squares estimates did not involve shrinkage. In this chapter we will present a few properties of the lasso and discuss the similarities and differences between the lasso and the zero-ridge estimator.

5.1 The lasso estimate $\hat{\beta}_{\text{lasso}}(\lambda)$

In this section we will discuss the lasso estimate of β_0 . We will use the same assumptions as in the previous chapters, letting

$$y|X, \beta_0 \sim N_n(X\beta_0, \sigma^2 I_n),$$

with σ^2 being a known constant.

The lasso estimate of β_0 is given in several different forms corresponding to different assumptions and parameterizations, but for mathematical convenience we will use the maximizer of the penalized likelihood function

$$H_n(\beta) = l_n(\beta) - \lambda \|\beta\|_1.$$

This gives

$$\begin{aligned} \hat{\beta}_{\text{lasso}}(\lambda) &= \underset{\beta}{\operatorname{argmax}} \left\{ -\frac{1}{2} \frac{(y - X\beta)^T (y - X\beta)}{\sigma^2} - \lambda \sum_{j=1}^p |\beta_j| \right\} \\ &= \underset{\beta}{\operatorname{argmax}} \left\{ -\frac{n}{2} \frac{(\beta - \hat{\beta}_{LS})^T \Sigma_n (\beta - \hat{\beta}_{LS})}{\sigma^2} - \lambda \sum_{j=1}^p |\beta_j| \right\}. \end{aligned}$$

The form of this penalty is different from that of the ridge penalty, and unlike in the ridge case, we may here get zeroes in our parameter estimate. The function is complicated and has many singular points, and it is usually not possible to solve analytically. Because of these points it may also be unsuitable for standard numerical optimization, but a number of algorithms have been developed for optimizing this particular function. In my numerical experiments I have used the package `lars` in R, based on the algorithm LARS proposed by Efron et al. (2004). They proved that the lasso solution paths grow piecewise linearly in a predictable way from 0 to $\hat{\beta}_{LS}$. Thus, they were able to solve the entire lasso solution path using the same order of computations as a standard least squares fit. An example is given in Figure 5.1, showing the solution paths as functions of $\|\hat{\beta}_{\text{lasso}}(\lambda)\|/\|\hat{\beta}_{LS}\|$ for the same ten-dimensional vector we used in the previous chapter. The solution paths in terms of our parameter λ are shown in Figure 5.2 and should be compared to those of Figure 4.1 and Figure 4.2 (note that p , n , β_0 and Σ_n are the same, but X and y are different).

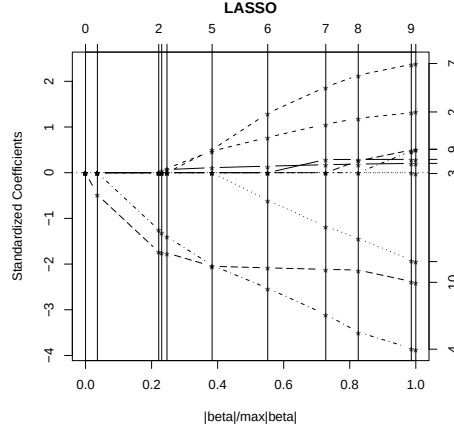


Figure 5.1: The solution path of $\hat{\beta}_{\text{lasso}}(\lambda)$ as a function of $\|\hat{\beta}_{\text{lasso}}(\lambda)\|/\|\hat{\beta}_{LS}\|$.

Unlike for ridge regression, the lasso problem has no simple closed form solution in general. However, when $\Sigma_n = \text{diag}(s_{n,1}^2, \dots, s_{n,p}^2)$ is diagonal we are able to obtain explicit solutions. In this case the problem is reduced to p one-dimensional lasso problems with solutions of the form

$$\hat{\beta}_{\text{lasso},j}(\lambda) = \underset{\beta_j}{\text{argmin}} \frac{1}{2}(\beta_j - \hat{\beta}_{LS,j})^2 + \frac{\lambda\sigma^2}{ns_{n,j}^2}|\beta_j|.$$

This gives

$$\hat{\beta}_{\text{lasso},j}(\lambda) = \text{sgn}(\hat{\beta}_{LS,j}) \left(|\hat{\beta}_{LS,j}| - \frac{\lambda\sigma^2}{ns_{n,j}^2} \right)_+,$$

see for example Fan and Li (2001). A plot showing the behaviour of $\hat{\beta}_{\text{lasso},j}(\lambda)$ and $\hat{\beta}_{\text{ridge},j}(\lambda)$ in the situation of $\Sigma_n = I_p$ is provided in Figure 5.3. We see that the ridge estimate is closer to $\hat{\beta}_{LS,j}$ than the lasso estimate when $\hat{\beta}_{LS,j}$ is small, but not when $\hat{\beta}_{LS,j}$ is large.

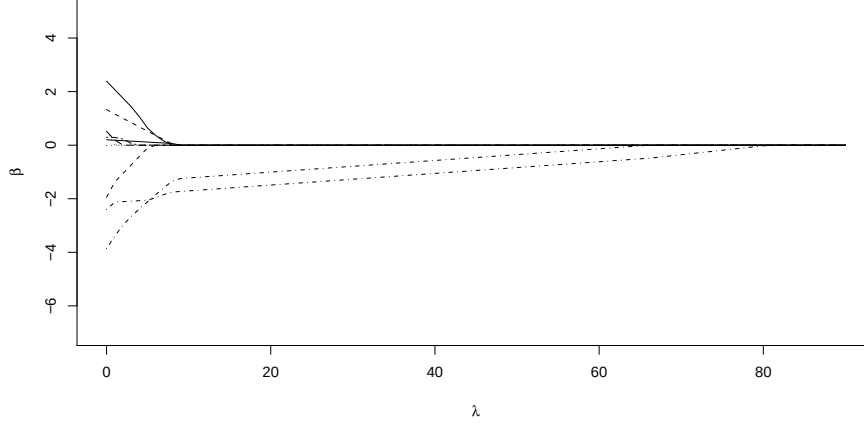


Figure 5.2: The solution path of $\hat{\beta}_{\text{lasso}}(\lambda)$ as a function of λ .

For non-diagonal matrices Σ_n , we are not able to say much more about the estimate $\hat{\beta}_{\text{lasso}}(\lambda)$. However, before we move on to estimating the tuning parameter λ , we will discuss interpretation of $\hat{\beta}_{\text{lasso},j}(\lambda)$. Like in the ridge case the lasso estimate is also the posterior mode of β when we give the elements of β a certain prior distribution. This prior distribution is double exponential with parameter $1/\lambda$, that is

$$\beta_j \sim \text{DE}(0, 1/\lambda),$$

with all β_j 's independent, in other words

$$\pi_\lambda(\beta) = \prod_{j=1}^p \frac{\lambda}{2} \exp\{-\lambda|\beta_j|\}.$$

Now

$$\begin{aligned} \pi_\lambda(\beta|\text{data}) &\propto L_n(\beta)\pi_\lambda(\beta) \\ &\propto \exp\{l_n(\beta) - \lambda \sum_{j=1}^p |\beta_j|\}, \end{aligned}$$

and we therefore have

$$\hat{\beta}_{\text{lasso}}(\lambda) = \underset{\beta}{\operatorname{argmax}} \pi_\lambda(\beta|\text{data})$$

as well. We could of course derive the entire posterior distribution of β and try to invent other estimators as well. The most natural idea is to use the posterior mean, but this will leave us with no zero entries. The zero entries were an important part of the motivation for the lasso, and we will therefore not pursue this idea. However, we note that it could be used as an alternative to the ridge estimate, performing shrinkage of a different kind.

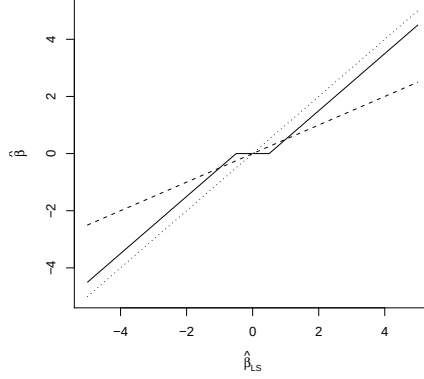


Figure 5.3: The estimates $\hat{\beta}_{\text{lasso},j}(\lambda)$ (solid line) and $\hat{\beta}_{\text{ridge},j}(\lambda)$ (dashed line) as a function of $\hat{\beta}_{LS,j}$ (dotted line added for easy comparison) in the situation of $\Sigma_n = I_p$. Here $\lambda\sigma^2/n = 0.5$ in both cases.

In the ridge section we used a reparameterization of λ in terms of τ . We might want to do the same here for easy comparison with the ridge case, but it is not clear what τ should be. We have two obvious choices, one being $\lambda = 1/(2\tau^2)$ in order to follow the ridge regression case as closely as possible w.r.t. the penalized log likelihood expression. However, we recall that the reason for replacing λ by $1/(2\tau^2)$ was to get nice-looking expressions for the corresponding prior distribution of β . In the lasso case, the penalty corresponds to a double exponential prior with location parameter 0 and scale parameter $1/\lambda$ for each β_j with all β_j i.i.d.. This prior distribution has mean 0 and variance $2/\lambda^2$. Another way of following the ridge case 'as closely as possible' would therefore be to make sure the prior distribution had the same mean and variance, that is $\tau^2 = 2/\lambda^2$, giving $\lambda = \sqrt{2}/\tau$. There may be even more appropriate ways of expressing λ in terms of τ , so to avoid confusion, the results in this chapter will mainly concern λ itself. In the next section, we will discuss possible ways of estimating this tuning parameter.

5.2 Estimating the penalty parameter λ

The lasso estimate $\hat{\beta}_{\text{lasso}}(\lambda)$ could range from 0 to $\hat{\beta}_{LS}$ dependent on the value of λ . The choice of this parameter is therefore a very important part of the estimation, and in this section we will present the same three estimation methods as in the previous chapters.

5.2.1 A simple approach based on the method of moments

We have already mentioned that using the lasso may be seen as finding the mode of the posterior distribution when the prior is double exponential with mean 0 and variance $2/\lambda^2$ for each component β_j . We will use this to derive a simple

estimate of λ , following the same idea as in the ridge chapter. We know that

$$y|X, \beta_0 \sim N(X\beta_0, \sigma^2 I_n),$$

so

$$E(y^T y|X, \beta_0) = (X\beta_0)^2 + n\sigma^2.$$

Also

$$E(y^T y|X) = E(\beta^T X^T X \beta) + \sigma^2 = \frac{2}{\lambda^2} \text{Tr}(X^T X) + n\sigma^2.$$

Inserting the sample moment for the population moment leads to

$$\hat{\lambda} = \sqrt{\frac{2\text{Tr}(\Sigma_n)}{\frac{1}{n}y^T y - \sigma^2}}. \quad (5.1)$$

We observe that $\hat{\lambda} = \sqrt{2/\hat{\tau}_{\text{ridge}}^2}$ where $\hat{\tau}_{\text{ridge}}^2$ is the corresponding estimate for ridge regression. We may therefore use the results from the ridge section and obtain the following result:

Theorem 5.2.1. *For $\hat{\lambda}$ as defined in (5.1) we have*

$$\frac{\sqrt{n}(\hat{\lambda} - \lambda_{n,X})}{v_{n,X} \lambda_{n,X}^3 / 4} \rightarrow N(0, 1) \quad (5.2)$$

whenever Condition 3.2.1 is satisfied. Here $v_{n,X}$ is as defined in Theorem 3.2.1, and $\lambda_{n,X} = \sqrt{2/\tau_{n,X}^2}$.

Proof. This is a direct result of Theorem 3.2.3 and the delta method. $\square$

5.2.2 Cross validation

Our second approach for estimating λ is the cross validation. In the lasso case our estimate is given by

$$(\hat{\lambda}) = \underset{\lambda \geq 0}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n \left[y_i - x_i^T \hat{\beta}_{\text{lasso},(i)}(\lambda) \right]^2 \right\}.$$

Here $\hat{\beta}_{\text{lasso},(i)}(\lambda)$ is once again the β -estimate obtained by leaving out the i -th observation. Since the lasso estimate is a continuous function of λ , this is a continuous function, as opposed to the function in the zero-mixed ridge case. However, it is still not smooth, and we will not try to approximate it by a simpler function like we did in the ridge chapter.

This function is easier to optimize numerically than the zero-mixed ridge function. First of all, we only need to estimate one parameter, so the number of grid points to consider will be smaller. Also, we may use the fast algorithm LARS to find the lasso solutions in each calculation. However, since we always need to compute n different lasso estimates for each value of λ , this method will be very time-consuming for large n .

5.2.3 The maximum marginal likelihood estimator

Our last estimator will once again try to use the fact that the lasso corresponds to finding the mode of the posterior distribution for β . Thinking Bayesian, we will again try to use the λ with the highest marginal likelihood:

$$\begin{aligned}\hat{\lambda} &= \operatorname{argmax}_{\lambda \geq 0} f_{\text{marg}}(y, \lambda) \\ &= \operatorname{argmax}_{\lambda \geq 0} \int l_n(\beta) \pi_\lambda(\beta) d\beta.\end{aligned}$$

Inserting the normal likelihood and our double exponential prior we find that

$$\begin{aligned}\hat{\lambda} &= \operatorname{argmax}_{\lambda \geq 0} \int l_n(\beta) \pi_\lambda(\beta) d\beta \\ &= \operatorname{argmax}_{\lambda \geq 0} \int \exp \left\{ -\frac{1}{2} \frac{(y - X\beta)^T (y - X\beta)}{\sigma^2} \right\} \lambda^p \exp \left\{ -\lambda \sum_{j=1}^p |\beta_j| \right\} d\beta \\ &= \operatorname{argmax}_{\lambda \geq 0} \int \exp \left\{ -\frac{n}{2} \frac{(\beta - \hat{\beta}_{LS})^T \Sigma_n (\beta - \hat{\beta}_{LS})}{\sigma^2} \right\} \lambda^p \exp \left\{ -\lambda \sum_{j=1}^p |\beta_j| \right\} d\beta\end{aligned}\tag{5.3}$$

This may always be solved numerically, but in the simple case of a diagonal Σ_n we are able to go a few steps further. In this case we assume $\Sigma_n = \text{diag}(s_{n,1}^2, \dots, s_{n,p}^2)$, obtain the following result

Theorem 5.2.2. *For $\hat{\lambda}$ as defined by (5.3) and $\Sigma_n = \text{diag}(s_{n,1}^2, \dots, s_{n,p}^2)$ we have*

$$\hat{\lambda} \approx \frac{p}{\sum_{j=1}^p \hat{\beta}_{LS,j}}.$$

Proof. For such Σ_n we find that

$$\begin{aligned}\hat{\lambda} &= \operatorname{argmax}_{\lambda \geq 0} \prod_{j=1}^p \int \exp \left\{ -\frac{n}{2} \frac{s_{n,j}^2}{\sigma^2} (\beta_j - \hat{\beta}_{LS,j})^2 \right\} \lambda \exp \{ -\lambda |\beta_j| \} d\beta_j \\ &= \operatorname{argmax}_{\lambda \geq 0} \prod_{j=1}^p \int \sqrt{2\pi} \phi(u_j) \lambda \exp \left\{ -\lambda \left| \frac{\sigma}{\sqrt{n} s_{n,j}} u_j + \hat{\beta}_{LS,j} \right| \right\} \frac{\sqrt{n} s_{n,j}}{\sigma} du_j \\ &= \operatorname{argmax}_{\lambda \geq 0} \prod_{j=1}^p \lambda \int \phi(u_j) \exp \left\{ -\lambda \frac{\sigma}{\sqrt{n} s_{n,j}} \left| u_j + \frac{\sqrt{n} s_{n,j}}{\sigma} \hat{\beta}_{LS,j} \right| \right\} du_j\end{aligned}$$

with $u_j = \frac{\sqrt{n} s_{n,j}}{\sigma} (\beta_j - \hat{\beta}_{LS,j})$ and $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp \{ -\frac{1}{2} x^2 \}$.

In order to continue, we need the following formula:

$$\begin{aligned}g(a, c) &= \int \phi(x) \exp \{ -c|x + a| \} dx \\ &= \{1 - \Phi(a + c)\} \exp \left\{ ac + \frac{1}{2} c^2 \right\} + \Phi(a - c) \exp \left\{ -ac + \frac{1}{2} c^2 \right\}.\end{aligned}$$

Letting $a_j = \frac{\sqrt{ns_{n,j}}}{\sigma} \hat{\beta}_{LS,j}$ and $c_j = \frac{\lambda\sigma}{\sqrt{ns_{n,j}}}$ we now have

$$\hat{\lambda} = \operatorname{argmax}_{\lambda \geq 0} \prod_{j=1}^p \left[\lambda \left[\left\{ 1 - \Phi \left(\frac{\sqrt{ns_{n,j}}}{\sigma} \hat{\beta}_{LS,j} + \frac{\lambda\sigma}{\sqrt{ns_{n,j}}} \right) \right\} \exp \left\{ \lambda \hat{\beta}_{LS,j} + \frac{1}{2} \frac{\lambda^2 \sigma^2}{ns_{n,j}^2} \right\} \right. \right. \\ \left. \left. + \Phi \left(\frac{\sqrt{ns_{n,j}}}{\sigma} \hat{\beta}_{LS,j} - \frac{\lambda\sigma}{\sqrt{ns_{n,j}}} \right) \exp \left\{ -\lambda \hat{\beta}_{LS,j} + \frac{1}{2} \frac{\lambda^2 \sigma^2}{ns_{n,j}^2} \right\} \right] \right].$$

If we consider the three possibilities $\beta_j < 0$, $\beta_j = 0$ and $\beta_j > 0$ separately, and use the fact that $\Phi(-x) = 1 - \Phi(x)$, we find that

$$\hat{\lambda} = \operatorname{argmax}_{\lambda \geq 0} \prod_{j=1}^p \left[\lambda \left[\left\{ 1 - \Phi \left(\frac{\sqrt{ns_{n,j}}}{\sigma} |\hat{\beta}_{LS,j}| + \frac{\lambda\sigma}{\sqrt{ns_{n,j}}} \right) \right\} \exp \left\{ \lambda |\hat{\beta}_{LS,j}| + \frac{1}{2} \frac{\lambda^2 \sigma^2}{ns_{n,j}^2} \right\} \right. \right. \\ \left. \left. + \Phi \left(\frac{\sqrt{ns_{n,j}}}{\sigma} |\hat{\beta}_{LS,j}| - \frac{\lambda\sigma}{\sqrt{ns_{n,j}}} \right) \exp \left\{ -\lambda |\hat{\beta}_{LS,j}| + \frac{1}{2} \frac{\lambda^2 \sigma^2}{ns_{n,j}^2} \right\} \right] \right]$$

now, for large n and $\hat{\beta}_{LS,j} \neq 0$ for all j , this function is very close to

$$\prod_{j=1}^p \left[\lambda \exp \left\{ -\lambda |\hat{\beta}_{LS,j}| \right\} \right]$$

which is maximized by

$$\tilde{\lambda} = \frac{p}{\sum_{j=1}^p |\hat{\beta}_{LS,j}|}.$$

□

Our $\hat{\lambda}$ is not exactly equal to $\tilde{\lambda}$, but it is very close. We note that the diagonal values of Σ_n do not appear in the expression. We also note that from the prior distribution of β we have

$$\lambda = \frac{p}{E(\sum_{j=1}^p |\beta_{0,j}|)},$$

leading directly to the estimate $\tilde{\lambda}$ if we substitute the observed $\hat{\beta}_{LS,j}$ for the unobservable β_j . In fact, if we still assume that the β_j 's are really stemming from our double exponential prior, the estimate also resembles the maximum likelihood estimate of λ when β_0 is known, given by

$$\hat{\lambda}_{MLE} = \frac{p}{\sum_{j=1}^p |\hat{\beta}_{0,j}|}. \quad (5.4)$$

Even though $\tilde{\lambda}$ is only close to the maximum marginal likelihood estimate when $\Sigma_n = I_p$, the observation above leads us to consider using $\tilde{\lambda}$ as our estimate also in other situations. This alternative estimate also has the nice property of always being positive and finite. Since $\tilde{\lambda}$ is a simple function of the ordinary least squares estimate, we also easily find that

$$\frac{\sqrt{n} \left(\tilde{\lambda} - \frac{p}{\|\hat{\beta}\|_1} \right)}{\frac{p\sigma}{\|\hat{\beta}_0\|_1^2} \sqrt{(\operatorname{sgn}(\beta_0))^T \Sigma_n^{-1} \operatorname{sgn}(\beta_0)}} \rightarrow N(0, 1), \quad (5.5)$$

where $\text{sgn}(\beta)$ is a vector with elements 1 and -1 corresponding to the signs of β_0 . For a given β_0 we may use this result to compare the limiting performance of $\hat{\lambda}$ and the method of moments estimate with limiting distribution given in Theorem 5.2.1.

We will conclude this section by showing that for large n our maximum marginal likelihood estimate is actually very close to the maximum likelihood estimate given in (5.4), independent of the form of Σ_n .

Theorem 5.2.3. *For $\hat{\lambda}$ as defined by (5.3) and large n we have*

$$\hat{\lambda} \approx \frac{p}{\|\beta_0\|_1}.$$

Proof. We have

$$\begin{aligned} \hat{\lambda} &= \arg\max_{\lambda \geq 0} \int \exp \left\{ -\frac{n}{2} \frac{(\beta - \hat{\beta}_{LS})^T \Sigma_n (\beta - \hat{\beta}_{LS})}{\sigma^2} \right\} \lambda^p \exp \{-\lambda \|\beta\|_1\} d\beta \\ &\approx \arg\max_{\lambda \geq 0} \int \exp \left\{ -\frac{n}{2} \frac{(\beta - \beta_0)^T \Sigma_n (\beta - \beta_0)}{\sigma^2} \right\} \lambda^p \exp \{-\lambda \|\beta\|_1\} d\beta \\ &\approx \arg\max_{\lambda \geq 0} \int \exp \left\{ -\frac{n}{2} \frac{(\beta - \beta_0)^T \Sigma_n (\beta - \beta_0)}{\sigma^2} \right\} \lambda^p \exp \{-\lambda \|\beta_0\|_1\} d\beta \\ &= \arg\max_{\lambda \geq 0} \lambda^p \exp \{-\lambda \|\beta_0\|_1\} \\ &= \frac{p}{\|\beta_0\|_1}. \end{aligned}$$

The first approximation is due to the fact that $\hat{\beta}_{LS}$ is a consistent estimate of β_0 , and is thus close to β_0 for large n . The second approximation is based on the observation that the first term of the integral is very small unless β is very close to β_0 . Therefore, replacing β_j by β_{0j} have almost no effect on the integral. $\square$

5.3 A small numerical example

In this section we will provide a small numerical example comparing three different β -estimators. The first one is the lasso estimator presented in this chapter, the second one is the zero-mixed ridge estimator from the previous chapter, and the third one is the ordinary variable selection approach comparing different least squares fits using the AIC. The main purpose of this experiment is to show that these three estimators perform different in some way. We have once again used a randomly generated β_0 , not stemming from the assumed prior of any of the estimators. The vector β_0 has four zero components and six non-zero components, and we have used $\sigma = 1$, $n = 50$ and $[\Sigma_n]_{(i,j)} = \rho^{|i-j|}$ for $\rho = 0$, $\rho = 0.5$ and $\rho = 0.9$. The matrix X has been generated the same way as in Chapter 3.

For each ρ we have generated one X and 100 y -vectors, and for each y -vector we have first estimated the necessary tuning parameters using the method of moments. This estimation method was chosen because we wanted to use the 'same' method for lasso and zero-mixed ridge, and the two other methods were too time-consuming in the lasso case. Using the estimated tuning parameters,

the three different β -estimates were computed. The tables below display the performance of each estimator measured in three different ways:

- Error A: The average number of $\beta_j = 0$ not estimated as zero.
- Error B: The average number of $\beta_j \neq 0$ estimated as zero.
- $\widehat{\text{MSE}}(\hat{\beta})$: The average value of $(\beta - \hat{\beta})^T(\beta - \hat{\beta})$.

Table 5.1: $\rho = 0$

	Error A	Error B	$\widehat{\text{MSE}}(\hat{\beta})$
LASSO	3.67	0.00	0.20
MIX	0.09	0.00	0.13
LS, AIC	0.69	0.00	0.17

Table 5.2: $\rho = 0.5$

	Error A	Error B	$\widehat{\text{MSE}}(\hat{\beta})$
LASSO	3.52	0.00	0.28
MIX	0.23	0.00	0.20
LS, AIC	0.62	0.00	0.23

Table 5.3: $\rho = 0.9$

	Error A	Error B	$\widehat{\text{MSE}}(\hat{\beta})$
LASSO	3.21	0.03	1.72
MIX	0.73	0.67	2.10
LS, AIC	0.72	0.41	1.89

The results of the experiment are given in Tables 5.1, 5.2 and 5.3. We see a clear difference between the lasso and the two other methods. The lasso does not give as sparse results as the other two methods, but the mean square error is approximately the same. This means that the lasso estimate in this case produces small non-zero estimated values, whereas the two other methods result in actual zeroes. In this particular case with four real zeroes, the zero-mixed ridge and the ordinary subset selection method clearly performs better than the lasso in terms of choosing the correct model, even though the mean square error is approximately the same for all three methods.

5.4 Discussion

We have now presented the lasso, the last of our estimators for use in the linear regression setting. The lasso estimate performs both shrinkage and variable

selection using one single tuning parameter. This is therefore a less flexible estimator than the zero-mixed ridge estimator, but at the other hand we only need to estimate one tuning parameter, and we expect this to be an easier task than estimating two. However, the functions to be optimized are different, and finding the CV or MML estimate of $\hat{\lambda}$ has turned out to be very time-consuming. It is doable if you are interested in analyzing only one set of data, but these estimators are not suitable for large numerical experiments. Having neither analytical nor numerical results, we can not recommend using these estimators at this time. The method of moments estimator or the estimator $\tilde{\lambda}$ introduced in the section describing the MML estimate may be used as fast alternatives with known asymptotic properties.

Even though I can not provide large numerical simulation studies involving the CV and MML estimators at this time, I will mention a few observations from my simulations over the last few weeks. First, as n increases the MML estimate of λ gets closer to the approximate MML estimate provided by Theorem 5.2.3, no matter how ill-conditioned the matrix Σ_n is. Second, the estimate obtained using cross validation seems to become 0 as n increases, being non-zero only for very small n . However, these statements are only based on a few observations, and can only be seen as possible hypotheses for further research.

Chapter 6

Estimating the covariance matrix and its inverse in the Gaussian case

So far we have only considered the linear regression model, trying to estimate the coefficient vector β by using different methods of regularization. However, the methodology may be applied in many other statistical problems as well. An area of application very close to linear regression is the more general regression setting with for example robust linear models or likelihood-based generalized linear models (see e.g. Fan and Li (2001)). Other applications are not connected to regression at all, and in this chapter I will briefly discuss one of them, the estimation of covariance matrices and their inverses.

Covariance matrices are important in many areas of statistical analysis, such as principal component analysis, factor analysis and constructing confidence intervals for linear functions of the means of the components of a p -variate random vector. Since the covariance matrix is often unknown, it has to be estimated, and in this chapter we will discuss a few topics linked to the estimation of covariance matrices, especially in the high-dimensional setting with $p > n$.

6.1 Estimating the covariance matrix

We assume X to be a p -dimensional random vector with covariance matrix Σ , that is,

$$E[(X - E(X))(X - E(X))^T] = \Sigma,$$

and we let $X_1, \dots, X_n$ be a set of observations stemming from the distribution of X .

One natural way of estimating Σ is to replace the population mean by the sample mean given by

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n [(X_i - \bar{X})(X_i - \bar{X})^T],$$

with $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. This is also the maximum likelihood estimate when X is normally distributed, but it has a bias and is therefore often replaced by the

unbiased sample covariance matrix

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^n [(X_i - \bar{X})(X_i - \bar{X})^T].$$

The bias is a consequence of replacing the population mean $E(X)$ by its sample version $\bar{X}$, so in the cases where the sample mean is known, we may instead use the unbiased version

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n [(X_i - E(X))(X_i - E(X))^T].$$

These three estimates are almost equal in the case of a small p and a large n , and they are all performing well in this situation. However, as p approaches n they become unstable, and for $p > n$ they clearly become singular and thus no longer positive definite. When this happens, they not only become highly unstable, but they can no longer be used to estimate the precision matrix Σ^{-1} . In the increasingly important situation with $p > n$ we therefore need to find other and more stable estimation methods. In this section, I will present a few such methods proposed in recent years. Most of the methods are assuming the data to be Gaussian or 'nearly Gaussian' in a certain sense, so for the rest of this chapter we will assume the data to be multinormal with mean zero and covariance matrix Σ .

Several methods of consistent estimation have been proposed for the multinormal $p > n$ situation, most of them imposing some special structure on $\hat{\Sigma}$ in order to reduce the number of estimated parameters. One possible solution is to introduce a sparsity assumption on Σ , requiring a certain amount of the elements to be zero. One of the simplest estimation methods based on this idea is called thresholding, and was developed simultaneously and independently by Bickel and Levina (2008a) and El Karoui (2008).

Bickel and Levina (2008a) propose an estimation method where all elements of the estimated covariance matrix smaller than a certain threshold value s are set equal to zero. They show how this leads to a symmetric, permutationally invariant and under a certain condition positive definite estimate of Σ . Also they prove consistency (in the operator norm) of the estimate for a certain class of sparse matrices as long as $\log(p)/n \rightarrow 0$ and obtain explicit rates of convergence. A cross validation procedure is proposed for estimating the threshold s .

The estimated covariance matrix obtained by thresholding is permutationally invariant, and therefore a suitable choice in the situations where the data has no natural ordering. This is often the case with for example genetics or financial data. In other situations, like time series or spatial data, there is a notion of distance between variables, and variables 'far apart' are often assumed to be only weakly correlated. These two settings correspond to two large categories of covariance estimation methods, and we will give an example from the latter category as well.

Bickel and Levina (2008b) consider estimating the covariance matrix by banding the sample covariance matrix. They focus on situations where $|i - j|$ large is expected to imply near independence or conditional (given the intervening indexes) independence of X_i and X_j . In the paper they show how, by either banding the sample covariance matrix or estimating a banded version of

the inverse population covariance matrix, it is possible to obtain consistent estimates (in the operator norm) as long as $\log(p)/n \rightarrow 0$ and the real Σ belongs to a specified group of matrices. They also obtain explicit convergence rates for the estimates.

Banding the sample covariance matrix corresponds to setting $\sigma_{ij} = 0$ if $|i-j| > k$ for a certain $0 \leq k < p$. This is both conceptually and computationally simple, as only the parameter k needs to be estimated. A cross-validation procedure is once again proposed for estimating k . However, banding the sample covariance matrix does not generally guarantee positive definiteness. Another technique shrinking all the off-diagonal elements towards zero, called 'tapering' (Furrer and Bengtsson, 2007) is therefore also considered in the article, as it avoids the problem.

Both thresholding and banding replace a number of elements in the sample covariance matrix by zero, but the remaining elements are left unchanged. In the regression setting we studied methods involving both model selection and shrinkage, and such methods have also been proposed for estimating covariance matrices. One such approach is the generalized thresholding, as described by Rothman et al. (2009), having the thresholding described above as a special case. They give several examples of popular shrinkage and thresholding functions covered by their definition of generalized thresholding, and establish convergence rates for a class of 'approximately sparse' matrices. They also show that the estimates are consistent as long as $\log(p)/n \rightarrow 0$, and that true zeroes are estimated as zeroes with probability tending to 1. Under an additional condition, non-zero elements are also estimated as non-zero and with the correct sign with probability tending to 1.

6.2 Estimating the precision matrix

Sometimes, we are not interested in the covariance matrix itself, but rather its inverse, known as the precision matrix. The precision matrix has no natural pivotal estimator, and is often estimated by inverting the sample covariance matrix. In the case of $p > n$, the sample covariance matrix is singular, and other estimation methods are needed. Some of the estimates mentioned in the previous section are positive definite, and thus non-singular, and clearly one possibility is to use one of them, and then invert the resulting estimated covariance matrix. Unfortunately, a sparse matrix does not necessarily have a sparse inverse. One example of such a situation is when $\sigma_{ij} = \rho^{|i-j|}$ with ρ close to 1. Then Σ itself has no zeroes, while Σ^{-1} is tridiagonal and thus very sparse. Introducing a sparse $\hat{\Sigma}$ in such situations may therefore introduce unnecessary error. Also, when Σ^{-1} is estimated in order to be used in further calculations, we may not need it to be positive definite, making this an unnecessarily restrictive method of estimation. These and other problems have led to several attempts of estimating the precision matrix directly, without first estimating the covariance matrix.

Like in the covariance matrix setting, sparsity is often imposed in order to make the resulting estimate more stable in the case of $p > n$. Bickel and Levina (2008b) propose a method based on banding the Cholesky factor of the precision matrix, giving a banded estimated precision matrix. Further examples of sparse precision matrix estimators will be given later in this chapter, but first we will

discuss the sparsity pattern itself and its interpretation.

6.2.1 Estimating the support of the precision matrix

A sparse estimate of the inverse covariance matrix is particularly interesting when the data is Gaussian, as in this case, zero elements in the precision matrix correspond to conditional independence between variables (see e.g. Lauritzen (1996)). Thus $[\Sigma^{-1}]_{(j,k)} = 0$ if and only if X_{ij} and X_{ik} are independent given all the other components. The inverse covariance matrix therefore has a corresponding undirected graphical model with p nodes, where j and k are connected if and only if $\sigma_{ij} \neq 0$. Undirected graphical models like this offer a way to describe and explain the relationships among a set of variables, a central element of multivariate data analysis. Using the principle of parsimony, we should select the simplest graphical model that adequately explains the data, and for large p the graphs will become too complicated unless Σ^{-1} is very sparse. In the graphical model situation we are only interested in identifying the zeroes of the precision matrix, and do not care about the value of the remaining elements. The task may therefore be seen as conceptually simpler than estimating the entire precision matrix.

The problem is to find the best of $2^{p(p+1)/2}$ possible models due to some specified model selection criterion, but comparing all the models is clearly possible only for very small p . The problem has therefore traditionally been solved by performing a greedy forward or backward search based on the MLE and a suitable stopping criterion. Even though this reduces the problem from $O(2^{p(p+1)/2})$ to $O(p(p+1)/2)$ models to be considered, for moderately large p this method becomes computationally infeasible as well. Another problem is that the MLE may not exist when $p > n$. Therefore, new and faster procedures have been proposed, and I will here present one such method closely linked to the lasso regression case presented in the previous chapter.

Meinshausen and Bühlmann (2006) propose a computationally efficient method where the neighbours of each of the p nodes are found consecutively, and then combined to a final graph structure. The neighbours of a node a are found by fitting a linear regression model with a lasso penalty, letting a be the response variable and regarding all the other variables as predictor variables. If the coefficient estimate corresponding to node b is non-zero, this corresponds to an edge between a and b in the final graph. We should note here that even though we have $\sigma_{ab} = 0 \Leftrightarrow \sigma_{ba} = 0$, the estimation method above may result in b being a neighbour of a without the opposite being true. Therefore, in the final graph an edge is included due to an 'and' or 'or' rule, but it is shown that the choice of rule is not important. The penalty parameter used in the lasso regression may be chosen in a way that controls the probability of falsely connecting two nodes, and this is shown to lead to a consistent estimator under some additional assumptions. The result holds for p growing as any power of n , and the number of neighbours of any variable growing at most slightly slower than the number of observations.

6.2.2 Estimating the precision matrix using a lasso penalty

We have seen that an l_1 penalty may be used to estimate the support of the precision matrix. In this section we will discuss how such a penalty may help us

estimate the entire precision matrix, including the value of non-zero elements. We assume our data to come from a multinormal distribution with mean μ and covariance matrix Σ . The maximum likelihood estimate of Σ^{-1} is given by

$$\hat{\Sigma}_{\text{MLE}}^{-1} = \operatorname{argmax}_{Z \succ 0} \{\log |Z| - \operatorname{Tr}(SZ)\},$$

where we let

$$S := \frac{1}{n} \sum_{i=1}^n (X_i - \mu)(X_i - \mu)^T.$$

In the regular situation with S non-singular, this leads to the familiar MLE, $X = S^{-1}$. However, for $p > n$ the matrix S is singular and the MLE may not exist. We therefore impose an l_1 -penalty on each element of the matrix, and obtain a new regularized estimate given by

$$\hat{\Sigma}^{-1} = \operatorname{argmax}_{Z \succ 0} \{\log |Z| - \operatorname{Tr}(SZ) - \lambda \|Z\|_1\}, \quad (6.1)$$

with $\|Z\|_1 = \sum_{j,k} |z_{jk}|$. This regularization ensures that $\hat{\Sigma}^{-1}$ (and thus the corresponding $\hat{\Sigma}$) is always invertible, no matter how small the ratio of samples to variables is, as long as $\lambda > 0$. Note that even when S is invertible, S^{-1} may not be sparse even if the real Σ^{-1} is, and the method may be used to find a sparse solution in these cases as well. Banerjee et al. (2008) shows that the optimal solution of (6.1) is unique and has bounded eigenvalues.

Algorithms

Although the expression for $\hat{\Sigma}^{-1}$ looks simple, the computation itself is no simple task. The objective function in (6.1) is convex and continuous, but non-smooth due to the last term. Maximizing the function in $p(p+1)/2$ parameters over all $p \times p$ matrices is therefore computationally difficult and very time-consuming. Finding $\hat{\Sigma}^{-1}$ is even harder, since our estimate should be the maximum over the class of positive definite matrices only, giving an unbounded constraint set. Solving this particular non-smooth, high-dimensional interior point optimization problem has become an area of research itself, and I will give two examples from the literature.

Banerjee et al. (2008) present an algorithm for solving the problem with at least a thousand nodes. The algorithm is iterative, optimizing over one row and column at the time until convergence is reached. The initial estimate is $W^{(0)} := S + \lambda I_p$, which is clearly positive definite, and they prove that the resulting solution, as well as all the other iterates produced by the algorithm, are strictly positive definite. Thus even though they are actually estimating $\hat{\Sigma}$ rather than its inverse, the result is guaranteed to be invertible, giving a sparse positive definite $\hat{\Sigma}^{-1}$. They also prove convergence of the algorithm, and provide a simple rule for choosing λ such that all elements of a certain column (and the corresponding row) of $\hat{\Sigma}$ are zero, except for the diagonal element.

A related approach called the graphical lasso is given by Friedman et al. (2008). They use the blockwise coordinate descent approach of Banerjee et al. (2008) as a starting point, and the algorithm is the same except for the updating step. While Banerjee et al. are using the solution of a box-constrained quadratic program, Friedman et al. is solving a problem close to ordinary lasso regression.

As fast coordinate descent algorithms have already been developed for the lasso problem, this new algorithm is able to reach convergence much faster than the original one. Again, $\hat{\Sigma}$ is computed and is not inverted before convergence is reached, using a simple method for inverting the matrix. It is also described how the simpler approach of Meinshausen and Bühlmann (2006) is connected to this solution.

Choosing the penalty parameter

Using the lasso penalty requires a suitable choice of λ , as this parameter controls the degree of regularization. We have already mentioned that Banerjee et al. (2008) provide a rule for choosing λ such that one component becomes conditionally independent of all the others. They also provide a rule for choosing λ in such a way that the probability of falsely connecting two different connectivity components is controlled. Other articles are being less concerned with this choice, for example simply stating that the penalty parameter may be estimated using fivefold crossvalidation, not discussing this stage of the estimation any further.

One of the first ideas for my project was to work with the estimation of sparse high-dimensional precision matrices, and when we decided to focus on the estimation of the tuning parameter we still hoped to be able to say something about estimating λ in this case. We have not had the time to actually do this, but I will present the main idea as a possible topic for further work. We have already seen that the maximum likelihood estimate of Σ^{-1} is given by

$$\hat{\Sigma}_{\text{MLE}}^{-1} = \operatorname{argmax}_{Z \succ 0} \{\log |Z| - \operatorname{Tr}(SZ)\},$$

and that adding an l_1 -penalty on each element of the matrix Z leads to the maximum penalized likelihood estimate

$$\hat{\Sigma}^{-1} = \operatorname{argmax}_{Z \succ 0} \{\log |Z| - \operatorname{Tr}(SZ) - \lambda \|Z\|_1\}.$$

This expression may be written on the equivalent form

$$\begin{aligned} \hat{\Sigma}^{-1} &= \operatorname{argmax}_{Z \succ 0} \left\{ |Z|^{n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (X_i - \mu)^T Z (X_i - \mu) \right\} \exp \left\{ -\frac{n\lambda}{2} \|Z\|_1 \right\} \right\} \\ &\propto \operatorname{argmax}_{Z \succ 0} L_n(Z) \pi_\lambda(Z). \end{aligned}$$

Thus once again we find that the penalized maximum likelihood estimate is also the posterior mode corresponding to a certain prior distribution. This particular prior is a product of independent priors for each $z_{ij:i \leq j}$ given by

$$\pi_\lambda(z_{ij:i < j}) \propto \exp \{-n\lambda |z_{ij}|\}$$

and

$$\pi_\lambda(z_{ii}) \propto \exp \left\{ -\frac{n\lambda}{2} |z_{ii}| \right\}.$$

Following the lines of the derivation of the maximum marginal likelihood estimator in the previous chapters we may now let

$$\hat{\lambda} = \operatorname{argmax}_{\lambda \geq 0} \int_{Z^*: Z \succ 0} L_n(Z^*) \pi_\lambda(Z^*) dZ^*,$$

where Z^* is the lower triangular matrix of Z . We are not able to evaluate this integral analytically, but we may be able to evaluate it numerically, at least for a suitably large set of possible λ -values, by using for example importance sampling.

We have now presented our basic idea for deriving a maximum marginal likelihood estimator for the precision matrix situation. We have not had the time to implement it, but considering the good performance of the MML estimator in the other situations, we think that this might lead to a good estimator.

Chapter 7

Discussion

In this thesis we have presented three different alternatives to ordinary least squares estimation, the ridge regression, our new zero-mixed ridge and the lasso. The idea of these estimation methods was to allow a small bias in order to reduce the total variance of the estimate, and hopefully also the mean square error. The numerical example of Chapter 3 showed how the ridge estimator with a set of different ridge parameters did sometimes have significantly smaller MSE than the ordinary least squares estimator. We also saw that our three estimators for τ had different strengths and weaknesses. The maximum marginal likelihood estimator had the smallest variance, while the cross validation procedure obtained its estimates by minimizing the predicted sum of squares, and this estimate should therefore be expected to perform well with respect to prediction. The method of moments estimator did not have a corresponding natural criterion for which it seemed to be optimal, but its great advantage was its simplicity and extremely fast computation. These properties should also be valid for the corresponding estimators for the zero-mixed ridge and the lasso, but unfortunately we were not able to derive the same amount of analytical and numerical results for these estimators.

We also saw how our proposed zero-mixed ridge estimator was connected to ordinary subset selection using the information criterions AIC and BIC. In the small numerical example of Chapter 5 we saw that this new estimator resulted in point estimates quite close to those of the least squares estimator combined with subset selection using the AIC. However, this was only a very small experiment, and we should not draw any conclusions based on such a small material. Still, the fact that it did not perform completely different from our well-known alternative is a good sign, and we should consider this new estimator a worthy competitor to the lasso, at least in situations with moderately small p .

The zero-mixed ridge estimator and the lasso have many of the same characteristics, but also important differences. They both perform model selection, and they both shrink the remaining components, as opposed to the ordinary subset selection procedure. However, the amount of shrinkage is different in the two cases, as the zero-mixed ridge estimator performs shrinkage in the same way as the ordinary ridge estimator. Another important difference between the two estimators is that the lasso do only have one parameter controlling both shrinkage and model selection. In the zero-ridged case we have two parameters, and we have seen that the parameter α only influence the model selection, whereas

the other parameter τ is involved in both shrinkage and model selection. This makes the zero-mixed ridge a more flexible approach, and we could probably construct examples in which it would perform much better than the lasso. At the other hand, the estimates provided by the lasso are more stable, as they are maximizers of continuous functions, as opposed to the zero-mixed ridge and the ordinary subset selection procedure. These two estimators are discontinuous in the data, and a small change in the data could result in a completely different model. Considering all these differences, the choice between the lasso and the zero-mixed ridge is not an obvious one. There are also a number of other similar estimators, one example being the elastic net of Zou and Hastie (2005), combining the lasso penalty and the ridge penalty.

The work in this thesis has been carried out using a number of implicit and explicit assumptions. Some of them may be too restrictive, for example requiring the error terms to be normally distributed. Also, we have assumed the variance of the error term to be known, a rather unusual situation in practise. It would therefore be interesting to go through the arguments again, checking if these and other assumptions are really necessary. If they turn out to be necessary, we should try to derive the results again without using them. This would be a far more difficult task, but the results would be more useful.

Another point I wish we had had time for, is the derivation of confidence intervals for the various estimates presented. We did derive limiting distributions, but these involved the true but unknown β_0 and can therefore not be used directly. Of course we could try to estimate them using a consistent estimate of β_0 , but we should be careful, since the initial problem is the estimation of β_0 , and some of the estimates are expected to be poor.

With more time I would also have wanted to derive the properties of our β_0 -estimates as functions of the estimates of the penalization parameters. We are not the only ones avoiding this part of the work, but if we are interested in the final estimates, we should take the time to derive this kind of results. After all, estimates of the kind $\hat{\beta}(\hat{\lambda})$ have a much more complicated form than the corresponding $\hat{\beta}(\lambda)$, especially when the two estimates are computed using the same data. Similarly, we should try to derive the properties of the predicted y -values given by $\hat{y} = X\hat{\beta}(\hat{\lambda})$ or similar expressions.

The reuse of data is another point I would have liked to study in more detail. In the case of a large number of observations, we should consider using one part of them for the penalty parameter estimation and the other part for the β_0 -estimation. This way it would also be easier to derive the limiting properties of our final β_0 -estimate.

There are many things I would have done if I had had more time, the most important one being to perform a larger numerical experiment like the one in the ridge chapter for the lasso and the zero-mixed ridge estimator and the different methods for estimating their tuning parameters. Unfortunately some of the functions are so slow that we would probably have needed weeks to complete this. For example I tried to make a single contour plot for the cross validation function in the zero-mixed ridge case, but after waiting three hours for a plot with what I considered properly spaced grid lines, I gave up, finding that it had completed the computations for about 5-10% of the points. The method of moments estimates on the other hand were very fast. The time consumption becomes increasingly important as n and p grow, and it is not surprising that the

articles describing methods for estimation of very high-dimensional structures are using much space on time comparisons for different methods. An interesting idea for the future might be to develop fast approximations for the zero-mixed ridge and lasso estimators. Here I mean the estimators for the tuning parameters, as the estimation of β_0 itself were fast enough for my choices of n and p .

With faster procedures we could also try to attack the problem of p larger than n , or at least almost as large as n . This is a problem we have not considered in this thesis, but all the estimators exist in these settings as well. Deriving the properties of such estimators analytically would probably require a lot of work, but at least we could try to gain some insight from numerical experiments.

Even though I wish I had had time to derive more results and perform more numerical experiments, I am satisfied with what we were able to do. In my opinion the zero-mixed ridge estimator is the most interesting result of our work, and I am very curious about how it will perform in a larger experiment compared to subset selection with AIC/BIC and the lasso. I will perform such an experiment as soon as possible, even if my project is now officially over.

Bibliography

- Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127.
- Banerjee, O., El Ghaoui, L., and d’Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516.
- Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.*, 36(6):2577–2604.
- Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227.
- Claeskens, G. and Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32(2):407–499. With discussion, and a rejoinder by the authors.
- El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.*, 36(6):2717–2756.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360.
- Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.
- Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *J. Comput. Graph. Statist.*, 7(3):397–416.
- Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivariate Anal.*, 98(2):227–255.
- 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.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

- Lauritzen, S. L. (1996). *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York. Oxford Science Publications.
- Lehmann, E. L. (1999). *Elements of large-sample theory*. Springer Texts in Statistics. Springer-Verlag, New York.
- Lyche, T. (2010). Lecture notes for inf-mat 4350, 2010. University of Oslo.
- Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462.
- Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.*, 104(485):177–186.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320.