UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

LASSO - Simultaneous shrinkage and selection via the ℓ_1 norm

by

Lisa-Ann Kirkland

Submitted in partial fulfillment of the requirements for the degree

MSc Mathematical Statistics

In the Faculty of Natural & Agricultural Sciences

University of Pretoria

Pretoria

October 2014

Declaration

I, Lisa-Ann Kirkland, declare that this dissertation, which I hereby submit for the degree MSc Mathematical Statistics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

SIGNATURE: 

DATE: October 2014

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Acknowledgments

I would like to dedicate this paper to my mother who has always provided the opportunity for me to further my studies, I am sincerely grateful for all her support which allowed me the time to complete it. The gratitude extends to all my family and friends who were always willing to help, even if they didn't understand a word I was saying!

A big thanks to my supervisors, Dr Frans Kanfer and Mr Sollie Millard, who always provided useful insights and different perspectives. To the Department of Statistics, thank you for your patience and allowing me an extension. Special thanks for the financial support I was granted with the STATOMET bursary. Funding from the DST/NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS) is also acknowledged.

Lastly, thanks to all the authors in my bibliography for your valuable contributions. I can only express an overwhelming appreciation to Trevor Hastie, Robert Tibshirani and Jerome Friedman. The Elements of Statistical Learning has been a faithful companion to me since 2010, offering equal doses of inspiration and frustration. It constantly challenges the way I think and I will forever endeavour to soak up the wealth of knowledge between its covers. *Thank you!*

Abstract

LASSO - Simultaneous shrinkage and selection via the ℓ_1 norm

by Lisa-Ann Kirkland

Two major purposes of regression models are explanation and prediction of scientific phenomena. Explanation is obtained by producing interpretable models through variable selection, while prediction accuracy is optimised by balancing the bias and variance of predictions. This dissertation explores the LASSO, a shrinkage method that simultaneously performs selection and estimation, yielding interpretable models with high prediction accuracy. By penalizing the regression model, the variance is substantially reduced and sparsity is promoted by using the ℓ_1 norm. It often outperforms traditional methods like subset selection and ridge regression, each focusing either on variable selection or prediction, respectively. The LASSO has favourable statistical properties and can also be applied to high dimensional data. Applied in two-stage procedures, the bias is controlled to achieve consistency for both prediction and selection. Concave penalties reduce the bias more effectively by applying different penalty functions over fixed ranges of each coefficient's size. Adaptations of the LASSO penalty allow incorporating different structures between predictors, such as ordering predictors in a meaningful way or including known groups of predictors like dummy variables or polynomials. Penalties combining the ℓ_1 norm with other norms allow the identification of unknown groups of correlated variables. Overall the LASSO provides an elegant foundation for a class of methods which improves the way that sparse regression problems are solved.

Keywords: lasso, lars, shrinkage, regularization, variable selection, high-dimensional data, sparsity, oracle property, prediction accuracy, model selection, linear regression

Contents

List of Tables

List of Figures

Abbreviations

1 SE	one standard error
AIC	Akaike information criterion
BIC	Bayesian information criterion
BLUE	best linear unbiased estimator
CAP	composite absolute penalties
CV	cross-validation
DF	degrees of freedom
EN	elastic net
GCV	generalized cross-validation
GLM	generalized linear model
KKT	Karush-Kuhn-Tucker
LAR	least angle regression
LASSO	least angle selection and shrinkage operator
LAR	least angle regression
LAR-EN	LAR algorithm for elastic net
LAR-LASSO	LAR algorithm for LASSO
LAR-OLS	LAR algorithm used with least squares
LDP	least distance programming
LQA	local quadratic approximations

LSE	least squares estimate
LSI	least squares with linear inequality constraints
LOOCV	leave one out cross-validation
MCP	minimax concave penalty
MDL	minimum description length
MLE	maximum likelihood estimate
MSE	mean squared error
NNLS	nonnegative least squares
OSCAR	octagonal shrinkage and clustering algorithm for regression
PACS	pairwise absolute clustering and sparsity
PASS	prediction and stability selection
PE	prediction error
PLUS	penalized linear unbiased selection
PRESS	prediction sum of squares statistic
RSS	residual sum of squares
SCAD	smoothly clipped absolute deviation
SEA-LASSO	standard error adjusted LASSO
SIS	sure independence screening
SNR	signal to noise ratio
SVD	singular value decomposition
UMVUE	uniformly minimum variance unbiased estimator
VIF	variance inflation factor

Notation

Information

n	number of training observations
m	number of test observations
p	number of variables
d	number of relevant variables
g	number of groups
r	rank of $\mathbf{X}$

Random Variables

In general:

A	a random variable
a	real number, scalar, realization of A
A_n	a sequence of random variables
a_n	a sequence of real numbers
$\mathcal{A}$	a set of random variables
$\mathcal{A}^c$	the complement of set $\mathcal{A}$
$\mathbf{A}$	matrix
$\mathbf{A}_{\mathcal{A}}$	matrix with columns indexed by $\mathcal{A}$
$\mathbf{A}_{\mathcal{AB}}$	matrix with rows indexed by $\mathcal{A}$ and columns indexed by $\mathcal{B}$

$\mathbf{a}$	vector in $\mathbb{R}^n$
$\underline{a}$	vector in $\mathbb{R}^p, \mathbb{R}^{p+1}, \mathbb{R}^{p-d}$, etc.
$\mathbf{a}_{\mathcal{A}}, \underline{a}_A$	vector with elements indexed by $\mathcal{A}$

Some random vectors used:

$\mathbf{y}$	$n \times 1$ vector of response observations
$\mathbf{v}$	$n \times 1$ vector of centered response observations
ε	$n \times 1$ vector of error terms

Pertaining to each random vector, example for $\mathbf{y}$*:*

Y	random response variable
y_i	i-th observation of response variable Y
$\bar{y}$	sample mean of response variable Y
y_0	new response observation

Some random matrices used:

$\mathbf{X}$	$n \times (p+1)$ matrix of predictor observations (intercept column has index 0)
$\mathbf{C}$	$n \times p$ matrix of centered predictor observations (excl intercept)
$\mathbf{Z}$	$n \times p$ matrix of centered and scaled predictor observations (excl intercept)

Pertaining to each random matrix, example for $\mathbf{X}$*:*

X_j	j-th random predictor variable
$\mathbf{X}_k$	$n \times (k+1)$ matrix of predictor observations incl k predictors
$\mathbf{X}_{\mathcal{A}}$	$n \times \lvert\mathcal{A}\rvert$ matrix of predictor observations incl the columns indexed by $\mathcal{A}$
x_{ij}	i-th observation of j-th predictor variable
$\underline{x}_i$	$(p+1) \times 1$ vector, i-th row vector of $\mathbf{X}$, i-th observation of all predictors
$\underline{x}_0$	$(p+1) \times 1$ vector of new observations

$\mathbf{x}_j$	$n \times 1$ vector, j-th column vector of $\mathbf{X}$, observations of the j-th predictor X_j
$\bar{x}_j$	j-th column mean of $\mathbf{X}$, sample mean of the j-th predictor X_j
$\underline{\bar{x}}$	$p \times 1$ vector of column means of $\mathbf{X}$

Covariance and Correlation:

$\mathbf{S}$	$p \times p$ matrix, $(n-1) \times$ sample variance-covariance matrix of the predictors
s_{jj}	$(n-1) \times$ sample variance of the j-th predictor X_j
s_{jk}	$(n-1) \times$ sample covariance between X_j and X_k
$\mathbf{R}$	$p \times p$ matrix, sample correlation matrix of the predictors
$r_{jk}, \hat{\rho}_{jk}$	sample correlation between X_j and X_k
e_j	j-th eigenvalue of $\mathbf{R}$, unless specified otherwise
d_j	j-th singular value of $\mathbf{Z}$, unless specified otherwise

Parameters

In general:

θ	some parameter θ
$\hat{\theta}$	estimate of some parameter θ
$\hat{\theta}_n$	a sequence of estimators

Some parameters used:

σ^2	variance of the error term
β	unknown regression parameter
β_0	intercept parameter
β_j	parameter corresponding to the j-th predictor
$\underline{\beta}$	$(p+1) \times 1$ vector of regression parameters (intercept indexed by 0)
$\underline{\beta}_k$	$(k+1) \times 1$ vector of regression parameters for k predictors

$\underline{\beta}_{\mathcal{A}}$	$\lvert\mathcal{A}\rvert \times 1$ vector of regression parameters indexed by $\mathcal{A}$
$\mathcal{D}$	the correct subset of relevant variables of size $\lvert\mathcal{D}\rvert = d$
α_j	standardized parameter corresponding to the j-th predictor
$\underline{\alpha}$	$p \times 1$ vector of standardized parameters (excl intercept)
τ	constraint parameter, constrain the size of estimates
t	ℓ_1 norm of the estimates
λ	shrinkage parameter, controls shrinkage and selection
γ	bridge parameter, controls rotation of solution
ϕ	relaxation parameter in relaxed LASSO
w_j	weight factors in adaptive penalties
ζ	adaptive weight parameter in adaptive penalties
ψ	combined penalties parameter, controls the grouping effect
ξ	convex penalties parameter, controls the size of unpenalized estimates

Vectors and Matrices

$\mathbf{1}_n, \mathbf{1}$	$n \times 1$ vector of 1s	
$\mathbf{I}_n, \mathbf{I}$	$n \times n$ identity matrix	
$\text{diag}(\mathbf{a})$	diagonal matrix with diagonal elements $\mathbf{a}$	
$\text{rank}(\mathbf{A})$	rank of matrix $\mathbf{A}$	
$\text{tr}(\mathbf{A})$	trace of matrix $\mathbf{A}$	
$\mathbf{A}^T$	transpose of matrix $\mathbf{A}$	
$\mathbf{A}^{-1}$	inverse of matrix $\mathbf{A}$	
$\ell_q(\mathbf{a}), \lVert\mathbf{a}\rVert_q$	ℓ_q-norm of a vector $\mathbf{a}$, $0 < q < \infty$	A.1.1
$\mathbf{A}^-$	generalized inverse of $\mathbf{A}$	A.1.2

$\mathbf{A}^+$	pseudoinverse, Moore-Penrose inverse of $\mathbf{A}$	A.1.3
$\mathcal{C}(\mathbf{A})$	column space of matrix $\mathbf{A}$	A.1.4 (1)
$\mathcal{N}(\mathbf{A})$	null space of matrix $\mathbf{A}$	A.1.4 (3)
$\mathcal{C}^\perp(\mathbf{A})$	orthogonal complement of $\mathcal{C}(\mathbf{A})$	
$\text{var}(\mathbf{A})$	variance-covariance matrix if $\mathbf{A}$ is a vector/matrix	
$\text{corr}(\mathbf{A})$	correlation matrix if $\mathbf{A}$ is a vector/matrix	
$e_j(\mathbf{A})$	j-th eigenvalue of some matrix $\mathbf{A}$	
$d_j(\mathbf{A})$	j-th singular value of some matrix $\mathbf{A}$	
$\kappa_2(\mathbf{A})$	spectral condition number of matrix $\mathbf{A}$	2.1.29

Functions

$a'(b)$, $\frac{d}{db}a(b)$	first order derivative of function a with respect to b
$a''(b)$, $\frac{d^2}{db^2}a(b)$	second order derivative of function a with respect to b
$\frac{\partial}{\partial b_1}a(\underline{b})$	first order partial derivative of function a with respect to b_1
$\frac{\partial^2}{\partial b_1^2}a(\underline{b})$	second order partial derivative of function a with respect to b_1
$\nabla a(\underline{b})$	gradient of function a, vector of all first order partial derivatives
$\nabla^2 a(\underline{b})$	Hessian of function a, matrix of all second order partial derivatives
$P(a \in \mathcal{A})$	probability that a is in set $\mathcal{A}$
$\delta(a \in \mathcal{A})$	indicator that a is in set $\mathcal{A}$
$E(A)$	expectation of random variable A
$\text{var}(A)$	variance of random variable A
$\text{cov}(A_1, A_2)$	covariance between A_1 and A_2
$\text{corr}(A_1, A_2)$	correlation between A_1 and A_2

$P_\lambda(\lvert\underline{\alpha}\rvert)$	penalty function depending on λ	2
$B(\hat{\theta})$	bias of an estimate $\hat{\theta}$	A.2.4 (1)
$MSE(\hat{\theta})$	mean squared error of an estimate $\hat{\theta}$	A.2.4 (3)
$A_n \xrightarrow{p} A$	A_n converges to A in probability	A.2.5 (1)
$A_n \xrightarrow{d} A$	A_n converges to A in distribution	A.2.5 (4)
$O(n)$	order of at most n	A.2.6 (1)
$o(n)$	of order smaller than n	A.2.6 (2)
$O_p(n)$	order in probability of at most n	A.2.6 (3)
$o_p(n)$	of order in probability smaller than n	A.2.6 (4)
$f(\underline{x}_i)$	regression function	2.1.1
$RSS(\underline{\beta})$	residual sum or squares	2.1.3
$L(\underline{\beta}, \sigma^2)$	likelihood function	2.1.11
$D(\underline{\beta})$	deviance	2.1.12
$TE(f(\underline{x}))$	training error of regression predictor	3.1.1
$PE(\hat{f}(\underline{x}_0))$	expected prediction error of regression predictor	3.1.4
$\omega(\hat{f}(\underline{x}))$	expected optimism of the training error	3.1.7
$\varphi_n(k)$	penalty function of information criteria	3.2.6
$\text{sign}(a)$	sign of the scalar a	4.1.4
$(a)_+$	positive part of the scalar a	4.1.12
$h_\lambda(\underline{\alpha})$	objective function for penalized regression in orthogonal design	4.1.5
$l(\underline{\theta}, \underline{\eta}, \underline{\mu})$	Lagrangian function	A.3.6

Essentially,
all models are wrong,
but some are useful.

George E. P. Box

Chapter 1

Introduction

Statistical models are formulated to solve specific problems. The regression model expresses the relationship between the response variable, Y, and the predictor variables, X, with a systematic component and an additive random component,

$$Y = f(X;\beta) + \varepsilon.$$

It is assumed that Y is subject to the error ε and that X and ε are independent. That is, the predictor variables may be fixed variables or they may be random variables measured without error. In the latter case, the regression is conditional on the observed values of X. The error, or noise, is assumed to contain any deviations from the deterministic relationship $Y = f(X;\beta)$, including measurement errors in Y and any unmeasured variables that have an effect on Y. Therefore, it is considered as a random variable and we assume that $E(\varepsilon) = 0$ so that the regression function is

$$E(Y) = f(X;\beta).$$

The regression function thus attempts to estimate the mean of the response variable using the information contained by the predictor variables. When the relationship between the predictor variables and the response variable is approximately linear, the regression function is

$$f(X;\beta) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p,$$

where β_0 is an intercept term and p is the number of predictor variables. The estimate,

$$\hat{f}(X;\beta) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p,$$

is calculated using a sample of n observations and can be used to predict the response variable at future values of the predictor variables.

The quality of the estimate often depends on two critical aspects: prediction accuracy, which is a trade-off between the bias and the variance of the estimate, and model interpretability. Least squares estimation provides a simple approach to solve the linear regression model by minimizing the squared error loss, or residual sum of squares (RSS),

$$RSS(\beta) = (y - f(X;\beta))^2.$$

The least squares estimate (LSE) is unbiased and has the lowest variance, and consequently the best prediction accuracy, among all linear unbiased estimates. When $n \gg p$ then this variance will be small and the prediction accuracy will be satisfactory. However, when p is near n, the LSE can be highly variable and result in poor prediction accuracy. Furthermore, least squares cannot be used when $p > n$ since the estimate is not unique and the variability is infinite. Collinearity and overfitting can also inflate the variance of the LSE. Removing irrelevant predictors that have little effect on the response produces a more interpretable model and can significantly improve the estimate. However, it is highly unlikely that least squares will set any of the parameter estimates to zero. In any of these situations, a biased estimate may perform better than the LSE provided that the bias is small and the reduction in variance is substantial.

Traditional methods to overcome these drawbacks of least squares include subset selection methods and ridge regression. They produce biased estimates and can be used when p is near n or $p > n$. However, they address only one of the aspects while falling short on the other. Subset selection methods focus on model interpretability. A subset of size $d < p$ relevant variables are identified and the model is estimated using least squares, essentially setting the parameter estimates of the remaining $p - d$ variables to zero. However, since the process is discrete, the estimate can be very unstable and sensitive to small perturbations in the data, often yielding low prediction accuracy. In contrast, ridge regression (proposed by Hoerl & Kennard (1970)) focuses on stabilizing the variance to give greater prediction accuracy. A constraint is placed on the size of the parameters so that the estimates are shrunk towards zero. This is equivalent to minimizing the penalized RSS,

$$RSS(\beta) + P_\lambda(|\beta|),$$

where $\lambda > 0$ is called the tuning parameter (also called the decay, shrinkage, regularization or penalty

parameter) and $P_\lambda(\beta)$ is the penalty function which depends on λ. When $\lambda = 0$ the least squares estimate is obtained and the amount of shrinkage increases as λ increases. The ridge penalty is the squared ℓ_2 norm of the parameters and is differentiable at zero. Thus, ridge regression does not set any parameter estimates exactly to zero and produces a less interpretable model.

Tibshirani (1996) proposed the least angle selection and shrinkage operator (LASSO) as a method which could provide interpretable models with high prediction accuracy. It is thought to contain the best of both the traditional approaches, shrinking estimates of the relevant variables to control the variance and setting estimates of the irrelevant variables to zero to yield interpretability. The LASSO is similar to ridge regression, it minimizes the penalized RSS, but the penalty is the ℓ_1 norm of the parameters. The LASSO has the ability to set parameter estimates to zero since its penalty function is non-differentiable at zero. Although the non-differentiable nature of the problem prevents us from establishing an explicit expression for the estimate and its standard error, the LASSO solution path is piecewise-linear and efficient algorithms have been developed to compute the entire path, from the null model up to the least squares fit (if $p < n$), and standard errors can be calculated using either the bootstrap or approximations.

Subset selection methods, ridge regression and the LASSO are part of a general class of estimates, called bridge estimates, with penalty function

$$P_\lambda(\beta) = \lambda\left(\ell_\gamma(\beta)\right)^\gamma,$$

where the ℓ_γ-norm is given by

$$\ell_\gamma\left(\underline{\beta}\right) = \left\|\underline{\beta}\right\|_\gamma = \left(\sum_{j=1}^{p}|\beta_j|^\gamma\right)^{\frac{1}{\gamma}}.$$

The idea was suggested by Frank & Friedman (1993) as a paradigm for understanding subset selection and ridge regression. The ℓ_0-norm can be interpreted as the number of nonzero parameters and corresponds to the subset selection methods. They noted that it would be beneficial to estimate the parameters λ and γ simultaneously to widen the choice of possible models but did not develop the method any further. The parameter λ controls the size of the parameters and the γ parameter determines the directions in which the parameters are aligned with respect to the coordinate axes. Estimates are only likely to occur on the axes when $\gamma \in [0,1]$ and in this case parameters are set to zero. The problem is discrete when $\gamma = 0$ and the penalty function is concave when $\gamma \in (0,1)$, making $\gamma = 1$ (LASSO) an attractive choice for the ℓ_γ-norm. Knight & Fu (2000) showed that bridge estimates are consistent and have asymptotic normal

distributions.

The LASSO estimate relies strongly on the choice of the tuning parameter, which can be estimated using cross-validation (CV). For lower computational expense, information criteria and generalized cross-validation (GCV) can also be utilized by using an approximation of the effective degrees of freedom and an estimate of the error variance. Both approaches select the λ which minimizes the prediction error (PE). Greenshtein & Ritov (2004) show that the LASSO is consistent for prediction, a property which they call persistence. While the LASSO performs shrinkage and selection, it can fail to be optimal in both aspects with the use of only one tuning parameter. If we could choose λ so that the correct model is selected, its value would have to be large in order to shrink parameter estimates exactly to zero. But large values of λ tend to overshrink large parameters so the estimate can suffer large bias and thus poor prediction accuracy. On the other hand, if λ is chosen for optimal prediction (which is the usual case), its value will be smaller and it tends to overfit the model by including irrelevant variables. Although, it has been shown that all the relevant variables are included with high probability. This suggests using the LASSO in a two-stage procedure, performing selection in one stage and estimation in another.

Meinshausen (2007) proposed the relaxed LASSO to control the bias of the LASSO. Firstly, the entire path of the LASSO is computed. The LASSO is then applied to each model in the path with a smaller tuning parameter $\phi\lambda$, where $\phi \in (0,1]$, to obtain the entire path of the relaxed LASSO. The tuning parameter λ in the first step performs variable selection. In the second step, the tuning parameter ϕ relaxes the penalty so that the parameters are estimated with less bias. The tuning parameters ϕ and λ are chosen simultaneously using CV. He shows that the choice of tuning parameters for optimal prediction also yields consistent selection. Another two-stage procedure is the adaptive LASSO proposed by Zou (2006). He controls the bias by scaling each parameter in the penalty function with different weight factors. The weights depend on an initial estimate and are chosen adaptively. The amount of shrinkage applied to each parameter is inversely proportional to the size of its initial estimate so that the parameter estimates of relevant variables remain large and those of irrelevant variables are shrunk to zero. As with the relaxed LASSO, good selection properties are achieved using a prediction-optimal tuning parameter. In addition, he showed that provided the initial estimate is consistent for estimation, the adaptive LASSO has the oracle property. An oracle procedure estimates the parameters with efficience that is asymptotic to using least squares with the correct subset of variables, as if the correct model were known in advance. For the initial estimate, he proposed using the LSE when $p < n$ and no collinearity is present, and the ridge estimate

otherwise. Around the same time that Tibshirani (1996) proposed the LASSO, the nonnegative garrote was proposed by Breiman (1995). The goal of the nonnegative garrote is the same as for the LASSO, but the LSE is scaled directly by nonnegative constants. Although the method is often criticized for its dependence on the LSE, Zou (2006) showed that the nonnegative garrote is equivalent to the adaptive LASSO (when the LSE is used as the initial estimate) and thus also enjoys oracle properties. Yuan & Lin (2007) also generalized the nonnegative garrote to use estimates from ridge regression, the LASSO or the elastic net (EN) instead of the LSE and showed promising results.

The LASSO has also been modified to incorporate different structures among the predictor variables. Tibshirani *et al.* (2005) proposed the fused LASSO to handle predictor variables that can be ordered in some meaningful way, where parameters are similar for predictors that are near to each other. An additional LASSO penalty is imposed on the difference between adjacent parameters to encourage similar estimates for nearby variables. For handling known groups of predictors, such as a categorical variable coded as a group of dummy variables or a set of basis functions for polynomial or nonparametric components, the group LASSO was proposed, first by Bakin (1999) and developed further by Yuan & Lin (2006). Each group of predictors, also called factors, corresponds to one observed variable so that variable selection should consider the importance of factors rather than the derived variables within them. The penalty function is the sum of the ℓ_2 norms of each group, weighted by the size of the group (the number of variables within it). An ungrouped variable is just a group of size 1 and the penalty reduces to the LASSO penalty but for groups including two or more variables it is similar to the ridge penalty. Thus the group LASSO promotes sparsity between groups but not within groups. Zhao *et al.* (2009) proposed composite absolute penalties (CAP), a generalization of the group LASSO to use any ℓ_{q_k}-norm for the k-th group with $q_k > 1$ instead of the ℓ_2 norm. In particular, they focus on using the ℓ_∞-norm,

$$\ell_\infty\left(\underline{\beta}\right) = \lim_{q\to\infty}\left\|\underline{\beta}\right\|_q = \max\left\{|\beta_1|, |\beta_2|, \ldots, |\beta_p|\right\},$$

which encourages estimates within the group to be equally sized. They further generalize to combine the group norms with the ℓ_γ-norm instead of the sum. The ℓ_γ-norm is called the overall norm and controls the relationship between groups, while the group ℓ_{q_k}-norm controls how variables within each group are related. Using $\gamma = 1$ ensures sparsity between groups. Besides performing group selection, they allow overlapping groups to be defined and describe how to use them to enforce hierarchical information. If higher order effects are included in the model, it is usually desirable to include any main effects involved

so that the model is shift invariant. Huang *et al.* (2009) proposed a similar idea, called the group bridge, which uses the ℓ_1-norm as the group norm and the ℓ_γ-norm with $0 < \gamma < 1$ as the overall norm. The penalty performs variable selection at the group level and within groups, so that if a group is included, estimates for variables within that group may be set to zero. They show that the method has oracle properties for group selection.

Predictor variables can also occur in unknown groups with high pairwise correlations between variables in the group and it may be desirable to select all the variable in the group. When variables are highly correlated, the LASSO tends to randomly pick one of the variables and discard the rest. Ridge regression often outperforms the LASSO in terms of prediction when collinearity is present. While the LASSO penalty is convex, the ridge penalty is strictly convex (in fact, all bridge penalties with $\gamma > 1$ are). The strict convexity encourages a grouping effect but these penalties do not promote sparsity. Zou & Hastie (2005) proposed the EN as a combination of ridge regression and the LASSO. The ridge penalty performs decorrelation while the LASSO penalty performs both shrinkage and selection. The EN also performs very well with high-dimensional data, when $p \gg n$. The LASSO selects at most $\min(n, p)$ variables so that no more than n variables can be selected for high dimensional data. In contrast, the EN can potentially select all p variables. Zou & Zhang (2009) extended the EN as a two-stage procedure, the adaptive EN. They combine the ridge penalty with the adaptive LASSO penalty and suggest using the EN estimate to calculate the weights. They showed that the adaptive EN has the oracle property. Bondell & Reich (2008) proposed a similar method called octagonal shrinkage and clustering algorithm for regression (OSCAR). The penalty function is a combination of the LASSO penalty and an ℓ_∞-norm on pairs of parameters. The motivation is similar to the EN but the ℓ_∞-norm sets estimates exactly equal for highly correlated variables that have a similar effect on the response. The method thus performs supervised variable clustering automatically as part of the estimation procedure. Sharma *et al.* (2013) recently proposed pairwise absolute clustering and sparsity (PACS) as a generalization of both OSCAR and the EN. The penalty consists of an adaptive LASSO penalty and weighted penalties on the sums and differences of pairs of parameters. They pose four alternatives for calculating the weights. In particular, they show that the method has the oracle property when using data adaptive weights.

Fan & Li (2001) were the first to propose a shrinkage method with the oracle properties, smoothly clipped absolute deviation (SCAD). They realize that a good selection procedure requires a penalty function that is continuous to ensure stability, discontinuous at zero to enable sparsity and bounded by a

constant to control bias. They achieve this with a penalty function that is concave and applies a different penalty over three regions based on the size of the parameters. Another concave penalty is minimax concave penalty (MCP) proposed by Zhang (2010) who shows that the method has superior selection accuracy over SCAD.

Overall, shrinkage methods provide a convenient way to simultaneously select variables and estimate parameters. Modifications of the LASSO method and other shrinkage methods have been developed to overcome any shortcomings with the LASSO. Note that the LASSO is a special case in every method so that the LASSO estimate can be calculated using any of the methods. Unfortunately, no method works well in every situation and prior information should be used to match the problem at hand to the relevant procedure - see the "no free lunch" theorems by Wolpert & Macready (1997).

The organization of this paper is as follows. Chapter 2 examines the relevant aspects of the traditional methods available for linear regression prior to the development of modern shrinkage methods, including least squares, subset selection and ridge regression. Methods like subset selection and ridge regression produce a set of models which vary with complexity and we need some way to choose between them. Methods for model selection are discussed in Chapter 3, where most of them are focussed on selecting the best model for prediction. Chapter 4 explores the LASSO in detail, including an examination of the penalty function; methods for computing the parameter estimates and their standard errors; the statistical properties of the LASSO; and model selection methods. The chapter further explores two-stage LASSO methods for controlling the bias and modified LASSO methods which allow for different structures between variables. Other shrinkage methods utilizing combined penalties and concave penalties are briefly discussed in Chapter 5. Appendices containing some background theory and calculations are provided in Appendix A and Appendix B.

Comprehensive simulation studies are presented in Chapter 6 to support the theory and identify scenarios in which the LASSO performs well. An application of the LASSO and other shrinkage methods is presented in Chapter 7, in which the data collected for a study of possible factors influencing the progression of diabetes is analysed. Some R packages that are available for fitting these models are given in Appendix C, and the R code used to produce the figures, simulations and application can be downloaded at `http://www.filedropper.com/rcode`. Concluding remarks and recommendations for further research are given in Chapter 8.

Chapter 2

Least Squares and Traditional Methods

This chapter provides an overview of least squares estimation and the traditional methods of subset selection and ridge regression. Least squares provides the framework on which every other method is based and is the topic of Section 2.1. The basics of estimation for both the full rank and rank deficient cases are covered in Section 2.1.1. Properties of the LSE are given in Section 2.1.2, while properties of the least squares predictor are given in Section 2.1.3. Section 2.1.4 deals with centering and scaling the data, which will be necessary for the shrinkage methods, and provides formulae for converting back to the original location and scale. Section 2.1.5 is a brief overview of the drawbacks faced when using least squares with a large number of variables. Overfitting and collinearity are identified as the two main causes of instability and their effect on the variance is shown in Section 2.1.6 and Section 2.1.7, respectively. The remaining two sections of Chapter 2 look at the traditional methods available for overcoming these two problems in least squares. Subset selection methods are useful for preventing overfitting and some well known methods are discussed briefly in Section 2.2, including all possible subsets (Section 2.2.2), forward selection (Section 2.2.3) and backward elimination (Section 2.2.4), while some less known methods and modifications are mentioned in Section 2.2.5. Ridge regression was designed to combat collinearity and is discussed in Section 2.3. Since ridge regression is part of the class of shrinkage methods, it is afforded a closer examination. A formulation of the problem and its estimation is discussed in Section 2.3.1, while Section 2.3.2 shows how collinearity is eliminated during estimation. The ridge estimates are then compared to least squares estimates, showing how they are shrunk in Section 2.3.3 and the effect of shrinkage on the their properties in Section 2.3.4. Lastly, Section 2.3.5 provides some recommendations for selecting the ridge shrinkage parameter.

2.1 Least Squares

2.1.1 Estimation

The linear regression model assumes that the regression function is linear in the parameters, that is

$$\begin{aligned} f(X) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \\ &= \beta_0 + \sum_{j=1}^{p} X_j \beta_j, \end{aligned}$$

where β_0 is a constant term representing the intercept. The predictor variables can consist of quantitative variables, nonlinear transformations, polynomials or interactions between them, or qualitative variables that are coded as dummy variables.

In order to estimate the parameters of the linear regression model, a set of training data is collected. A sample of n observations $(\underline{x}_1, y_1), (\underline{x}_2, y_2), \ldots, (\underline{x}_n, y_n)$ should be drawn randomly and independently from the population, where $\underline{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$ are realizations of the predictor variables (including 1 for the intercept) and y_i is a realization of the response variable for the i-th case. The model can then be written as

$$y_i = \beta_0 + \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij} \beta_j + \varepsilon_i \text{ for } i = 1, 2, \ldots, n.$$

Let

$$\underset{n\times 1}{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \underset{n\times(p+1)}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}, \quad \underset{(p+1)\times 1}{\underline{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \text{ and } \underset{n\times 1}{\underline{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where the first column of $\mathbf{X}$ will have subscript $j = 0$ so that $\mathbf{x}_0 = \mathbf{1}_n$ and each column $\mathbf{x}_j$ corresponds to the observations of the j-th predictor variable X_j. Then the regression function is

$$f(\underline{x}_i) = \underline{x}_i^T \underline{\beta} \tag{2.1.1}$$

and the model can be written in the matrix form

$$\mathbf{y} = \mathbf{X}\underline{\beta} + \varepsilon. \tag{2.1.2}$$

There are a number of methods available to estimate the parameters in equation (2.1.2). The parameter

estimates $\hat{\underline{\beta}}$, are usually found by optimizing some objective function. Least squares estimation is the most common estimation method used for linear regression because of its simplicity and the good properties it has without any distributional assumptions. The only assumptions made are that the random errors have zero expectation, constant variance and are uncorrelated with each other. However, a further assumption of Gaussian errors improves its properties and is necessary for making any inferences on the model. Under the assumption of normality, the random errors are independent.

Least Squares Assumptions:

1. $E(\varepsilon) = \mathbf{0}$ and $\text{var}(\varepsilon) = \sigma^2\mathbf{I}$
2. optionally, $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$

The objective function in least squares is the residual sum of squares (RSS),

$$RSS(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2,$$

or in matrix form

$$\begin{aligned} RSS(\underline{\beta}) &= \varepsilon^T\varepsilon \\ &= \left(\mathbf{y} - \mathbf{X}\underline{\beta}\right)^T \left(\mathbf{y} - \mathbf{X}\underline{\beta}\right) \\ &= \left(\ell_2\left(\mathbf{y} - \mathbf{X}\underline{\beta}\right)\right)^2 \\ &= \left\|\mathbf{y} - \mathbf{X}\underline{\beta}\right\|^2, \end{aligned} \tag{2.1.3}$$

where $\ell_2(\cdot)$ is the ℓ_2 norm (see Definition A.1.1). The least squares estimate (LSE) is the minimizer of RSS, that is

$$\hat{\underline{\beta}} = \underset{\underline{\beta}}{\arg\min} \left\|\mathbf{y} - \mathbf{X}\underline{\beta}\right\|^2. \tag{2.1.4}$$

Differentiating with respect to $\underline{\beta}$, we have the gradient

$$\nabla RSS(\underline{\beta}) = -2\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\underline{\beta}\right)$$

and the Hessian

$$\nabla^2 RSS(\underline{\beta}) = 2\mathbf{X}^T\mathbf{X}.$$

The matrix $\mathbf{X}^T\mathbf{X}$, known as the Gramian matrix of $\mathbf{X}$, is always nonnegative definite. Thus, the Hessian is nonnegative definite for all $\underline{\beta}$ and $RSS\left(\underline{\beta}\right)$ is a convex function (Definition A.3.2). By Definition A.3.4, $\hat{\underline{\beta}}$ is optimal if $\nabla RSS\left(\hat{\underline{\beta}}\right) = 0$,

$$\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\hat{\underline{\beta}}\right) = \mathbf{0}. \tag{2.1.5}$$

This leads to a system of linear equations known as the normal equations,

$$\mathbf{X}^T\mathbf{X}\hat{\underline{\beta}} = \mathbf{X}^T\mathbf{y}.$$

If $\mathbf{X}$ has full column rank then $\mathbf{X}^T\mathbf{X}$ has full rank since $\text{rank}\left(\mathbf{X}^T\mathbf{X}\right) = \text{rank}\left(\mathbf{X}\right) = p + 1$. Thus, the Hessian is positive definite for all $\underline{\beta}$. This means that $RSS\left(\underline{\beta}\right)$ is a strictly convex function and $\hat{\underline{\beta}}$ is a unique global minimum. Consequently, $\mathbf{X}^T\mathbf{X}$ is nonsingular so that the LSE of $\underline{\beta}$ is given by

$$\hat{\underline{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{2.1.6}$$

If $\text{rank}\left(\mathbf{X}\right) < p + 1$ then $\mathbf{X}^T\mathbf{X}$ is rank deficient and the Hessian is positive semidefinite for all $\underline{\beta}$. This occurs when the predictors are not linearly independent or when $p > n$ since $\text{rank}\left(\mathbf{X}\right) \leqslant \min\left(n, p+1\right)$. In this case $RSS\left(\underline{\beta}\right)$ is a convex function and all stationary points are global minimums. Since $\mathbf{X}^T\mathbf{X}$ is singular, a solution to the normal equations is given by

$$\hat{\underline{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-}\mathbf{X}^T\mathbf{y},$$

where $\left(\mathbf{X}^T\mathbf{X}\right)^{-}$ is any generalized inverse of $\mathbf{X}^T\mathbf{X}$. However, the solution is not unique since the generalized inverse of a matrix is not unique. In particular, the solution with minimum ℓ_2 norm is obtained when using the unique pseudoinverse or Moore-Penrose inverse. See Gentle (2007:227-228) for a proof. This solution is given by

$$\hat{\underline{\beta}} = \mathbf{X}^{+}\mathbf{y}.$$

Definitions of the generalized inverse and Moore-Penrose inverse are given in Definition A.1.2 and Definition A.1.3. Seber & Lee (2003:470) show that $\hat{\underline{\beta}} = \mathbf{X}^{-}\mathbf{y}$ is a solution to the normal equations for any generalized inverse of $\mathbf{X}$ that satisfies conditions (1) and (3) in Definition A.1.3. Since $\hat{\underline{\beta}}$ is not unique, these solutions are not estimators for $\underline{\beta}$. However, if $\mathbf{a}$ is in the row space of $\mathbf{X}$ then $\mathbf{a}^T\underline{\beta}$ is an estimable function (see Definition A.2.1) and $\mathbf{a}^T\hat{\underline{\beta}}$ is a unique estimator of $\mathbf{a}^T\underline{\beta}$, it is invariant to the choice of $\hat{\underline{\beta}}$.

The fitted model is a linear combination of the response vector $\mathbf{y}$,

$$\hat{f}(\mathbf{X}) = \mathbf{X}\underline{\hat{\beta}} = \mathbf{H}\mathbf{y}, \tag{2.1.7}$$

where

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-}\mathbf{X}^T. \tag{2.1.8}$$

Note that $\left(\mathbf{X}^T\mathbf{X}\right)^{-} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ when $\mathbf{X}$ has full column rank. Let $\text{rank}(\mathbf{X}) = r$. It is easy to show that $\mathbf{H}$ is a symmetric idempotent matrix and $\text{rank}(\mathbf{H}) = r$. The properties in Theorem A.1.1 can be used in the rank deficient case, see Searle (1971:20-21) for details. It can also be shown that $\mathbf{H}$ is invariant to the choice of $\left(\mathbf{X}^T\mathbf{X}\right)^{-}$.

Geometrically, the regression function $f(\mathbf{X}) = \mathbf{X}\underline{\beta}$ is a vector in $\mathbb{R}^n$ and lies in the column space of $\mathbf{X}$ denoted by $\mathcal{C}(\mathbf{X})$. See Definition A.1.4 for a definition of the four fundamental subspaces of a matrix. Least squares attempts to find $\underline{\hat{\beta}}$ so that the vector $\hat{f}(\mathbf{X}) = \mathbf{X}\underline{\hat{\beta}} \in \mathcal{C}(\mathbf{X})$ is the closest to the vector $\mathbf{y}$. The distance between $\mathbf{y}$ and $\mathbf{X}\underline{\hat{\beta}}$ will be a minimum when $\left(\mathbf{y} - \mathbf{X}\underline{\hat{\beta}}\right) \perp \mathcal{C}(\mathbf{X})$. So $\mathbf{H}$ is an orthogonal projection matrix that projects the response vector $\mathbf{y}$ onto $\mathcal{C}(\mathbf{X})$ to produce $\mathbf{X}\underline{\hat{\beta}}$. The complementary projection matrix of $\mathbf{H}$ is given by $\mathbf{I} - \mathbf{H}$ and projects $\mathbf{y}$ onto the null space of $\mathbf{X}^T$, or $\mathcal{C}^{\perp}(\mathbf{X})$, to produce the residual vector $\left(\mathbf{y} - \mathbf{X}\underline{\hat{\beta}}\right)$, which can clearly be seen by equation (2.1.5). $RSS\left(\underline{\hat{\beta}}\right) = \left\|\mathbf{y} - \mathbf{X}\underline{\hat{\beta}}\right\|^2$ measures the length of this vector, or the orthogonal distance between $\mathbf{y}$ and the subspace spanned by the columns of $\mathbf{X}$. In $\mathbb{R}^{r+1}$, the regression function $\hat{f}(\mathbf{X}) = \mathbf{X}\underline{\hat{\beta}}$ is an r-dimensional hyperplane and each residual vector, $y_i - \underline{x}_i\underline{\hat{\beta}}$, is a vector going from the hyperplane to the point y_i. Least squares finds $\underline{\hat{\beta}}$ so that the sum of the squared residuals is minimized. This $\underline{\hat{\beta}}$ is the point in $\mathbb{R}^{r+1}$ where $RSS\left(\underline{\beta}\right)$, an r-dimensional hypersurface, attains its minimum.

The minimum RSS is given by

$$RSS\left(\underline{\hat{\beta}}\right) = \left\|\mathbf{y} - \mathbf{X}\underline{\hat{\beta}}\right\|^2 = \|(\mathbf{I} - \mathbf{H})\mathbf{y}\|^2 = \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y}, \tag{2.1.9}$$

since $\mathbf{I} - \mathbf{H}$ is symmetric and idempotent. Its expected value is

$$\begin{aligned}
&E\left[RSS\left(\underline{\hat{\beta}}\right)\right] \\
&= E\left[\mathbf{y}^T\left(\mathbf{I}-\mathbf{H}\right)\mathbf{y}\right] \\
&= \operatorname{tr}\left[\left(\mathbf{I}-\mathbf{H}\right)\operatorname{var}\left(\mathbf{y}\right)\right] + E\left(\mathbf{y}\right)^T\left(\mathbf{I}-\mathbf{H}\right)E\left(\mathbf{y}\right) \text{ by Theorem A.1.2} \\
&= \sigma^2 \operatorname{tr}\left(\mathbf{I}-\mathbf{H}\right) + \left(\mathbf{X}\underline{\beta}\right)^T\left(\mathbf{I}-\mathbf{H}\right)\mathbf{X}\underline{\beta} \text{ since } E\left(\varepsilon\right) = \mathbf{0} \text{ and } \operatorname{var}\left(\varepsilon\right) = \sigma^2\mathbf{I} \\
&= \sigma^2 \operatorname{tr}\left(\mathbf{I}-\mathbf{H}\right) \text{ since } \left(\mathbf{I}-\mathbf{H}\right)\mathbf{X} = \mathbf{0} \\
&= \sigma^2 \operatorname{rank}\left(\mathbf{I}-\mathbf{H}\right) \text{ since } \mathbf{I}-\mathbf{H} \text{ is idempotent} \\
&= \sigma^2\left(n-r\right).
\end{aligned}$$

This leads to the LSE of the error variance

$$\hat{\sigma}^2 = \frac{RSS\left(\underline{\hat{\beta}}\right)}{n-r}. \tag{2.1.10}$$

2.1.2 Properties of Least Squares Estimates

When $\mathbf{X}$ has full column rank, the LSE $\underline{\hat{\beta}}$ has all of the following properties:

1. $\underline{\hat{\beta}}$ is unique and linear in the response vector $\mathbf{y}$.
2. $\underline{\hat{\beta}}$ is unbiased, $E\left(\underline{\hat{\beta}}\right) = \underline{\beta}$.
3. $\operatorname{var}\left(\underline{\hat{\beta}}\right) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$.

The mean squared error (MSE) measures the accuracy of an estimate and is the sum of its variance and squared bias. Here, $MSE\left(\underline{\hat{\beta}}\right) = E\left\|\underline{\hat{\beta}} - \underline{\beta}\right\|^2 = \operatorname{tr}\left[\operatorname{var}\left(\underline{\hat{\beta}}\right)\right] + \left\|E\left(\underline{\hat{\beta}}\right) - \underline{\beta}\right\|^2 = \sum_{j=0}^{p}\operatorname{var}\left(\hat{\beta}_j\right)$ (see Section B.1.1). So MSE is just the total variance when the estimate is unbiased. By comparing the MSE of an estimate to that of another estimate, we can get an idea of its efficiency relative to the other estimate. If $MSE\left(\hat{\theta}_1\right) < MSE\left(\hat{\theta}_2\right)$ then $\hat{\theta}_1$ is a relatively more efficient estimate of θ than $\hat{\theta}_2$ (see Spanos (1989:234-237) for a discussion).

For any estimable function $\underline{a}^T\underline{\beta}$, the LSE $\underline{a}^T\underline{\hat{\beta}}$ has all of the following properties:

1. $\underline{a}^T\underline{\hat{\beta}}$ is a linear combination of $\mathbf{y}$.

2. $\underline{a}^T\hat{\underline{\beta}}$ is unbiased, $E\left(\underline{a}^T\hat{\underline{\beta}}\right) = \underline{a}^T\underline{\beta}$.

3. $\text{var}\left(\underline{a}^T\hat{\underline{\beta}}\right) = \sigma^2\underline{a}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-}\underline{a}$.

4. $\underline{a}^T\hat{\underline{\beta}}$ is the best linear unbiased estimator (BLUE) of $\underline{a}^T\underline{\beta}$ (see Definition A.2.2 and Theorem A.2.1). Since $\underline{a}^T\hat{\underline{\beta}}$ has the lowest variance among linear unbiased estimates, it also has the lowest MSE. It follows that $\underline{a}^T\hat{\underline{\beta}}$ is relatively more efficient than any other linear unbiased estimate.

5. $\underline{a}^T\hat{\underline{\beta}}$ is the uniformly minimum variance unbiased estimator (UMVUE) $\underline{a}^T\underline{\beta}$ if $\varepsilon \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$ (see Definition A.2.3 and Theorem A.2.2 (1)). Under normality $\underline{a}^T\hat{\underline{\beta}}$ has the lowest variance (and thus lowest MSE) among all unbiased estimates, not just linear estimates. In this case, $\text{var}\left(\underline{a}^T\hat{\underline{\beta}}\right)$ equals the Cramér-Rao lower bound $CR\left(\underline{a}^T\underline{\beta}\right) = I_n\left(\underline{a}^T\underline{\beta}\right)^{-1}$ (Theorem A.2.3) and is fully efficient. The Fisher information $I_n\left(\underline{a}^T\underline{\beta}\right)$ involves differentiating the likelihood function and can be related to maximum likelihood estimation.

Note that when $\mathbf{X}$ is rank deficient, both $\underline{a}^T\hat{\underline{\beta}}$ and $\text{var}\left(\underline{a}^T\hat{\underline{\beta}}\right)$ are invariant to the choices of $\hat{\underline{\beta}}$ or $\left(\mathbf{X}^T\mathbf{X}\right)^{-}$. Shao (1999:159-161) shows that, under certain conditions and without assuming normality, the LSE $\underline{a}^T\hat{\underline{\beta}}$ also has the following asymptotic properties (see Definition A.2.7):

1. $\underline{a}^T\hat{\underline{\beta}}$ is consistent in MSE, $\underline{a}^T\hat{\underline{\beta}}_n \xrightarrow{\ell_2} \underline{\beta}$.

2. asymptotic normality, $\underline{a}^T\left(\hat{\underline{\beta}}_n - \underline{\beta}\right)\Big/\sqrt{\text{var}\left(\underline{a}^T\hat{\underline{\beta}}\right)} \xrightarrow{d} N(0,1)$.

The LSE $\hat{\sigma}^2$ has all the following properties:

1. $\hat{\sigma}^2$ is unbiased, $E\left(\hat{\sigma}^2\right) = \sigma^2$.

2. $\hat{\sigma}^2$ is the UMVUE of σ^2 if $\varepsilon \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$ (see Theorem A.2.2 (2)).

Under the assumption of normality the distribution of $\hat{\underline{\beta}}$, and hence $\underline{a}^T\hat{\underline{\beta}}$, can easily be obtained. Similarly, the distribution and variance of $\hat{\sigma}^2$ can easily be derived. Inferences about the LSEs can then be drawn by making use of the their distributional properties. This is beyond the scope of this paper and the interested reader is referred to Seber & Lee (2003:47-49) or Searle (1971:99-130,174-180) for a more thorough explanation.

Maximum likelihood estimation is another method that can be used to estimate the linear regression model. The objective function is the likelihood function which, under the assumption of normality, is given by

$$\begin{aligned} L\left(\underline{\beta}, \sigma^2|\underline{\varepsilon}\right) &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\underline{\varepsilon}^T\underline{\varepsilon}}{2\sigma^2}\right) \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{RSS\left(\underline{\beta}\right)}{2\sigma^2}\right) \end{aligned} \tag{2.1.11}$$

Unlike with least squares, the maximum likelihood estimate (MLE) of $\underline{\beta}$ and σ^2 are found simultaneously by maximizing the likelihood,

$$\left(\tilde{\underline{\beta}}, \tilde{\sigma}^2\right) = \underset{\left(\underline{\beta}, \sigma^2\right)}{\arg\max}\, L\left(\underline{\beta}, \sigma^2|\underline{\varepsilon}\right),$$

or equivalently, minimizing the negative log-likelihood,

$$-\ln L\left(\underline{\beta}, \sigma^2|\underline{\varepsilon}\right) = \frac{1}{2}\left[n\ln\left(2\pi\right) + n\ln\left(\sigma^2\right) + \frac{RSS\left(\underline{\beta}\right)}{\sigma^2}\right].$$

It is easily shown that the MLE of $\underline{\beta}$ is equal to the LSE,

$$\tilde{\underline{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \hat{\underline{\beta}}.$$

Thus, when the errors are Gaussian, $\hat{\underline{\beta}}$ also enjoys the desirable properties of MLEs. Under certain regularity conditions, MLEs have the following asymptotic properties:

1. Consistency, $\hat{\underline{\beta}}_n \xrightarrow{p} \underline{\beta}$.
2. Asymptotic normality, $\sqrt{n}\left(\hat{\underline{\beta}}_n - \underline{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, V\left(\underline{\beta}\right)\right)$.
3. Asymptotic efficiency, $V\left(\theta\right) = \lim_{n\to\infty}\left(\frac{1}{n}I_n\left(\theta\right)\right)$.

See Definition A.2.7 for an explanation of these properties. The MLE of the error variance is given by

$$\tilde{\sigma}^2 = \frac{RSS\left(\hat{\underline{\beta}}\right)}{n}.$$

Note that the MLE of σ^2 is biased and its variance only attains the Cramér-Rao lower bound asymptotically, see Seber & Lee (2003:49-50). When using maximum likelihood estimation, the deviance is a statistic that

is similar to RSS for least squares estimation and is given by

$$D\left(\underline{\beta}\right) = -2\ln L\left(\underline{\beta}, \sigma^2|\varepsilon\right) \tag{2.1.12}$$

and the minimum deviance is

$$D\left(\hat{\underline{\beta}}\right) = n\ln\left(2\pi\tilde{\sigma}^2\right) + \frac{RSS\left(\hat{\underline{\beta}}\right)}{\tilde{\sigma}^2}. \tag{2.1.13}$$

For more information about MLEs and the regularity conditions necessary, Spanos (1989:267-281) or Casella & Berger (2002:470,472,516) can be consulted.

2.1.3 Prediction

Once we've estimated the regression parameters, we can formulate the predictor used to predict a new response at a new observation $\underline{x}_0$,

$$y_0 = f\left(\underline{x}_0\right) + \varepsilon_0.$$

Assume that y_0 has the same probability structure as the elements of $\mathbf{y}$. That is, $E\left(y_0\right) = f\left(\underline{x}_0\right)$, $\text{var}\left(y_0\right) = \sigma^2$ and $\text{cov}\left(y_0, \mathbf{y}\right) = 0$. Then the predictor is given by

$$\hat{f}\left(\underline{x}_0\right) = \underline{x}_0^T\hat{\underline{\beta}}$$

and

$$\text{var}\left(\hat{f}\left(\underline{x}_0\right)\right) = \text{var}\left(\underline{x}_0^T\hat{\underline{\beta}}\right) = \sigma^2\underline{x}_0^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\underline{x}_0.$$

The expected prediction error (PE) is a measure of how well the estimated model predicts the new response and is given by

$$PE\left(\hat{f}\left(\underline{x}_0\right)\right) = E\left(y_0 - \hat{f}\left(\underline{x}_0\right)\right)^2 = \sigma^2 + MSE\left(\hat{f}\left(\underline{x}_0\right)\right).$$

The predictor $\hat{f}\left(\underline{x}_0\right)$ can be seen as an estimator of $E\left(y_0\right)$. The expected PE consists of the MSE of this estimator and includes $\text{var}\left(y_0\right) = \sigma^2$ to account for the variation in the new data. Notice that $\hat{f}\left(\underline{x}_0\right) = \underline{x}_0^T\hat{\underline{\beta}}$ is a linear function of $\mathbf{y}$ and is estimable if $\underline{x}_0$ is in the row space of $\mathbf{X}$ (see Definition A.2.1). Thus, the least squares predictor has all the properties of linear unbiased estimates as discussed in Section 2.1.2. In particular, it is the BLUE.

2.1.4 Centering and Invariance to Scale

Centering the data shifts the location of the data and only affects the estimation of the intercept, which gives the location at which the regression hyperplane crosses the y-axis. The intercept can be interpreted as the expected value of the response when all the predictors are set to zero. When the predictors are centered at their means, the interpretation changes to having the predictors set at their average values. While this may be meaningful for interpretation in some situations, it is generally not necessary to center the data since least squares performs a natural centering of the data during estimation.

The parameter estimates corresponding to the predictor variables are not affected by centering. Since they provide the gradients or slopes of the regression hyperplane, a shift in location has no influence on them. However, they are affected when the scale of the data is changed. If a predictor is scaled by a constant, the LSEs will be scaled by the inverse of that constant. This can be helpful for interpretation, changing the scale of a predictor with a large order of magnitude can make the estimate more readable. Scaling the data can also improve the accuracy of calculations when the scale of different predictors vary by a large order of magnitude. In some situations it is desirable to scale the data so that all of the predictors are on the same scale. In particular, this is necessary when using shrinkage methods. Therefore, this section examines the effects that centered and scaled data have on the estimates. In general, least squares is scale invariant and does not benefit from a change in location or scale.

Centered Data

Suppose we fit the model using the original variables. Assume $\mathbf{X}$ has full column rank so that $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ exists. Let $\mathbf{X} = (\mathbf{1}_n, \mathbf{X}_{\mathcal{A}})$, where $\mathcal{A} = \{j : 1, 2, \ldots, p\}$ so that $\mathbf{X}_{\mathcal{A}}$ is the last p columns of $\mathbf{X}$, to separate the intercept vector from the predictors. Then

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{X}_{\mathcal{A}} \\ \mathbf{X}_{\mathcal{A}}^T\mathbf{1} & \mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}} \end{bmatrix} = \begin{bmatrix} n & n\bar{\underline{x}}^T \\ n\bar{\underline{x}} & \mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}} \end{bmatrix},$$

where $\bar{\underline{x}} = \left(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p\right)^T$ is a vector of the column means $\bar{x}_j = n^{-1}\sum_{i=1}^{n} x_{ij}$. The partitioned matrix can be inverted using Theorem A.1.3. It turns out that

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{bmatrix} \frac{1}{n} + \bar{\underline{x}}^T\left(\mathbf{C}^T\mathbf{C}\right)^{-1}\bar{\underline{x}} & -\bar{\underline{x}}^T\left(\mathbf{C}^T\mathbf{C}\right)^{-1} \\ -\left(\mathbf{C}^T\mathbf{C}\right)^{-1}\bar{\underline{x}} & \left(\mathbf{C}^T\mathbf{C}\right)^{-1} \end{bmatrix},$$

where $\mathbf{C}$ is the centered predictor matrix with typical element $c_{ij} = x_{ij} - \bar{x}_j$. The LSEs are given by

$$\begin{aligned}\underline{\hat{\beta}} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\ &= \begin{bmatrix} \frac{1}{n} + \underline{\bar{x}}^T\left(\mathbf{C}^T\mathbf{C}\right)^{-1}\underline{\bar{x}} & -\underline{\bar{x}}^T\left(\mathbf{C}^T\mathbf{C}\right)^{-1} \\ -\left(\mathbf{C}^T\mathbf{C}\right)^{-1}\underline{\bar{x}} & \left(\mathbf{C}^T\mathbf{C}\right)^{-1} \end{bmatrix}\begin{bmatrix} n\bar{y} \\ \mathbf{X}_{\mathcal{A}}^T\mathbf{y} \end{bmatrix},\end{aligned}$$

which leads to

$$\hat{\beta}_0 = \bar{y} - \underline{\bar{x}}^T\underline{\hat{\beta}}_p \tag{2.1.14}$$

and

$$\underline{\hat{\beta}}_{\mathcal{A}} = \left(\mathbf{C}^T\mathbf{C}\right)^{-1}\left(\mathbf{X}_p^T\mathbf{y} - n\bar{y}\underline{\bar{x}}\right), \tag{2.1.15}$$

where $\underline{\hat{\beta}}_{\mathcal{A}} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p\right)^T$ with

$$\operatorname{var}\left(\hat{\beta}_0\right) = \frac{\sigma^2}{n} + \sigma^2\underline{\bar{x}}^T\left(\mathbf{C}^T\mathbf{C}\right)^{-1}\underline{\bar{x}}, \tag{2.1.16}$$

$$\operatorname{cov}\left(\hat{\beta}_0, \underline{\hat{\beta}}_{\mathcal{A}}\right) = -\sigma^2\left(\mathbf{C}^T\mathbf{C}\right)^{-1}\underline{\bar{x}}, \tag{2.1.17}$$

$$\operatorname{var}\left(\underline{\hat{\beta}}_{\mathcal{A}}\right) = \sigma^2\left(\mathbf{C}^T\mathbf{C}\right)^{-1}.$$

Since $\mathbf{C}^T\mathbf{y}$ has typical element

$$\begin{aligned}\textstyle\sum_{i=1}^n\left(x_{ij} - \bar{x}_j\right)y_i &= \textstyle\sum_i\left(x_{ij}y_i - \bar{x}_jy_i\right) \\ &= \textstyle\sum_i x_{ij}y_i - \bar{x}_j\sum_i y_i \\ &= \textstyle\sum_i x_{ij}y_i - n\bar{y}\bar{x}_j,\end{aligned}$$

we have that $\mathbf{X}_{\mathcal{A}}^T\mathbf{y} - n\bar{y}\underline{\bar{x}} = \mathbf{C}^T\mathbf{y}$ in Equation (2.1.15). Thus

$$\underline{\hat{\beta}}_{\mathcal{A}} = \left(\mathbf{C}^T\mathbf{C}\right)^{-1}\mathbf{C}^T\mathbf{y}, \tag{2.1.18}$$

which is the LSE when estimating the model using the centered predictor matrix. This shows how the data is naturally centered as a result of estimation. See Searle (1971:83-86) for a derivation of this result.

Suppose now that we estimate the model using the centered matrix $\mathbf{C}$. The above shows that $\underline{\hat{\beta}}_{\mathcal{A}} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p\right)$ remained unchanged, but what happens to the intercept? We are fitting the same model

that has just been reparameterized,

$$\begin{aligned}\hat{y}_i &= \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + \cdots + x_{ip}\hat{\beta}_p \\ &= \left(\hat{\beta}_0 + \bar{x}_1\hat{\beta}_1 + \cdots + \bar{x}_p\hat{\beta}_p\right) + \left(x_{i1} - \bar{x}_1\right)\hat{\beta}_1 + \cdots + \left(x_{ip} - \bar{x}_p\right)\hat{\beta}_p \\ &= \hat{\beta}_0' + c_{i1}\hat{\beta}_1 + \cdots + c_{ip}\hat{\beta}_p.\end{aligned}$$

Here it is clear that $\hat{\underline{\beta}}_{\mathcal{A}}$ are unchanged but the intercept is now estimated by

$$\begin{aligned}\hat{\beta}_0' &= \hat{\beta}_0 + \bar{x}_1\hat{\beta}_1 + \cdots + \bar{x}_p\hat{\beta}_p \\ &= \hat{\beta}_0 + \underline{\bar{x}}^T\hat{\underline{\beta}}_{\mathcal{A}} = \bar{y} \text{ from (2.1.14)}.\end{aligned}$$

More formally, we have that $\sum_{i=1}^n c_{ij} = \sum_{i=1}^n \left(x_{ij} - \bar{x}_j\right) = 0$ so that $\mathbf{1}^T\mathbf{C} = \underline{0}$. Then the LSE, is given by

$$\begin{aligned}\begin{bmatrix} \hat{\beta}_0' \\ \hat{\underline{\beta}}_{\mathcal{A}} \end{bmatrix} &= \left[\begin{bmatrix} \mathbf{1}^T \\ \mathbf{C}^T \end{bmatrix}\begin{bmatrix} \mathbf{1} & \mathbf{C} \end{bmatrix}\right]^{-1}\begin{bmatrix} \mathbf{1}^T \\ \mathbf{C}^T \end{bmatrix}\mathbf{y} \\ &= \begin{bmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{C} \\ \mathbf{C}^T\mathbf{1} & \mathbf{C}^T\mathbf{C} \end{bmatrix}^{-1}\begin{bmatrix} \mathbf{1}^T\mathbf{y} \\ \mathbf{C}^T\mathbf{y} \end{bmatrix} \\ &= \begin{bmatrix} n & \underline{0}^T \\ \underline{0} & \mathbf{C}^T\mathbf{C} \end{bmatrix}^{-1}\begin{bmatrix} n\bar{y} \\ \mathbf{C}^T\mathbf{y} \end{bmatrix} \\ &= \begin{bmatrix} \bar{y} \\ \left(\mathbf{C}^T\mathbf{C}\right)^{-1}\mathbf{C}^T\mathbf{y} \end{bmatrix}\end{aligned} \tag{2.1.19}$$

and

$$\operatorname{var}\left(\hat{\beta}_0'\right) = \frac{\sigma^2}{n}, \tag{2.1.20}$$

$$\operatorname{cov}\left(\hat{\beta}_0', \hat{\underline{\beta}}_{\mathcal{A}}\right) = \mathbf{0}, \tag{2.1.21}$$

$$\operatorname{var}\left(\hat{\underline{\beta}}_{\mathcal{A}}\right) = \sigma^2\left(\mathbf{C}^T\mathbf{C}\right)^{-1}. \tag{2.1.22}$$

Thus, the intercept term is always estimated by $\bar{y}$ with constant variance σ^2/n and is uncorrelated with the other estimates. This suggests fitting a reparameterized model,

$$\mathbf{v} = \mathbf{C}\underline{\beta}_{\mathcal{A}} + \varepsilon, \tag{2.1.23}$$

where $\mathbf{v}$ is the centered response vector with $v_i = y_i - \bar{y}$. In this form, the model does not explicitly

estimate the intercept term, which can be convenient. Thus,

$$\hat{v}_i = c_{i1}\hat{\beta}_1 + \cdots + c_{ip}\hat{\beta}_p.$$

Note that if $\mathbf{S} = \mathbf{C}^T\mathbf{C}$ so that $\mathbf{S}$ has typical element $s_{jk} = \mathbf{c}_j^T\mathbf{c}_k = \sum_{i=1}^n \left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)$ for $j, k = 1, 2, \ldots, p$, then $\frac{1}{n-1}\mathbf{S}$ is the sample variance-covariance matrix of the predictors.

Centered and Scaled Data

Suppose now that the data is standardized. Let $\mathbf{Z}$ be the centered and scaled predictor matrix with typical element $z_{ij} = c_{ij}\Big/\sqrt{s_{jj}} = \left(x_{ij} - \bar{x}_j\right)\Big/\sqrt{\sum_{i=1}^n \left(x_{ij} - \bar{x}_j\right)^2}$. Consequently, $\sum_{i=1}^n z_{ij}^2 = \mathbf{z}_j^T\mathbf{z}_j = 1$ so that the columns of $\mathbf{Z}$ have unit ℓ_2 norm and all the predictors are now on the same scale. Note also that $\sum_{i=1}^n z_{ij} = 0$ so that $\mathbf{1}^T\mathbf{Z} = \underline{0}$. Therefore, equations (2.1.19) to (2.1.22) hold, with $\mathbf{C}$ replaced by $\mathbf{Z}$. By the same argument as above, we can fit the reparameterized model,

$$\mathbf{v} = \mathbf{Z}\underline{\alpha} + \varepsilon. \tag{2.1.24}$$

We have,

$$\begin{aligned}\hat{v}_i &= c_{i1}\hat{\beta}_1 + \cdots + c_{ip}\hat{\beta}_p \\ &= \left(c_{i1}\Big/\sqrt{s_{11}}\right)\sqrt{s_{11}}\hat{\beta}_1 + \cdots + \left(c_{i1}\Big/\sqrt{s_{11}}\right)\sqrt{s_{pp}}\hat{\beta}_p \\ &= z_{i1}\hat{\alpha}_1 + \cdots + z_{ip}\hat{\alpha}_p.\end{aligned}$$

So the standardized estimates are given by

$$\underline{\hat{\alpha}} = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{v}, \tag{2.1.25}$$

with

$$\hat{\alpha}_j = \sqrt{s_{jj}}\hat{\beta}_j.$$

This form of the standardized estimates shows that least squares is scale invariant. If any predictor is scaled by some constant, say k, then its regression coefficient will be scaled by $1/k$ so that $\hat{\beta}_j X_j$ always

remains unchanged. Let $\mathbf{D} = \text{diag}\left(1/\sqrt{s_{jj}}\right)$, then $\mathbf{Z} = \mathbf{CD}$ so that

$$\begin{aligned}\text{var}\left(\underline{\hat{\alpha}}\right) &= \sigma^2\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} \\ &= \sigma^2\left(\mathbf{D}\mathbf{C}^T\mathbf{C}\mathbf{D}\right)^{-1} \\ &= \sigma^2\mathbf{D}^{-2}\left(\mathbf{C}^T\mathbf{C}\right)^{-1} \\ &= \text{diag}\left(s_{jj}\right)\text{var}\left(\underline{\hat{\beta}}_p\right).\end{aligned}$$

Thus,

$$\text{var}\left(\hat{\alpha}_j\right) = s_{jj}\,\text{var}\left(\hat{\beta}_j\right), \tag{2.1.26}$$

$$\text{cov}\left(\hat{\alpha}_j, \hat{\alpha}_k\right) = \sqrt{s_{jj}s_{kk}}\,\text{cov}\left(\hat{\beta}_j, \hat{\beta}_k\right). \tag{2.1.27}$$

Note that $\mathbf{R} = \mathbf{Z}^T\mathbf{Z}$ is the sample correlation matrix of the predictors since it has typical element

$$\begin{aligned}r_{jk} &= \mathbf{z}_j^T\mathbf{z}_k \\ &= \frac{\sum_{i=1}^n\left(x_{ij}-\bar{x}_j\right)\left(x_{ik}-\bar{x}_k\right)}{\sqrt{\sum_{i=1}^n\left(x_{ij}-\bar{x}_j\right)^2}\sqrt{\sum_{i=1}^n\left(x_{ik}-\bar{x}_k\right)^2}} \\ &= \frac{\frac{1}{n-1}\sum_{i=1}^n\left(x_{ij}-\bar{x}_j\right)\left(x_{ik}-\bar{x}_k\right)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n\left(x_{ij}-\bar{x}_j\right)^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^n\left(x_{ik}-\bar{x}_k\right)^2}} \\ &= \frac{\text{cov}\left(X_j, X_k\right)}{\sqrt{\text{var}\left(X_j\right)}\sqrt{\text{var}\left(X_k\right)}} = \hat{\rho}_{jk},\end{aligned}$$

where ρ_{jk} is the correlation between X_j and X_k.

Inference and Prediction

Centering and scaling the data does not have an effect on the regression function $\hat{f}(\mathbf{X})$ or on $RSS\left(\underline{\hat{\beta}}\right)$. Inspection of the projection matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ indicates that it is not affected by any scalar change made to $\mathbf{X}$. It can easily be verified using the partitioned matrices above that $\mathbf{H}$ remains unchanged when substituting $\mathbf{X}$ with either one of $\left(\mathbf{1}, \mathbf{X}_p\right)$, $(\mathbf{1}, \mathbf{C})$ or $(\mathbf{1}, \mathbf{Z})$. It follows that $\hat{f}(\mathbf{X}) = \mathbf{Hy}$ and $RSS\left(\underline{\hat{\beta}}\right) = \mathbf{y}^T(\mathbf{I} - \mathbf{H})\,\mathbf{y}$ are also unaffected by any of these changes. Thus, RSS always attains the same minimum and $\hat{\sigma}^2$ remains unchanged.

If any inferences are to be made on the model or the model is used for prediction, it is best to convert back to the original location and scale. Equations (2.1.14), (2.1.16) and (2.1.17) can be used to correct for the intercept, while equations (2.1.25), (2.1.26) and (2.1.27) can be used to convert back to the original scale. For more information on centering and scaling the data, (Draper & Smith, 1998:371-375) or Seber & Lee (2003:69-72) can be consulted

2.1.5 Large Number of Variables

When there is a large number of variables the model may be very difficult to interpret. If explanation is the primary goal of the analysis then the analyst is challenged to find the most parsimonious model which fits the data well. The time and cost of including predictor variables can also play an important role. If observations for a predictor variable are expensive to collect, or difficult to measure, then it may be preferable to use other predictors that could explain its effect on the response.

Least squares depends largely on the sample size in relation to the number of variables. When $n \gg p$ then it is likely to estimate the parameters accurately and efficiently. When $n < p$, as seen in Section 2.1.1, the $\mathbf{X}^T\mathbf{X}$ matrix is singular and generalized inverses must be utilized. If the purpose of the regression is explanation then least squares cannot be used since $\underline{\beta}$ cannot be estimated uniquely. However, Section 2.1.3 shows that least squares can still be used effectively for prediction purposes, provided that the new observations lie in the row space of $\mathbf{X}$. When $n > p$ but the sample size is inadequate, LSEs are likely to suffer from high variance. The variation of the estimates and the predictions are both proportionally dependent on $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ and σ^2. When p is near n, the $\mathbf{X}^T\mathbf{X}$ matrix can become ill-conditioned and its inversion will be very unstable with extremely large elements in $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$. Another difficulty is that σ^2 is often unknown and is estimated by $\hat{\sigma}^2 = RSS\left(\underline{\hat{\beta}}\right)/(n-p-1)$. As p approaches n, $(n-p-1)$ becomes smaller and $\hat{\sigma}^2$ can become very large. When $n = p+1$ exactly, then there are no residual degrees of freedom and $\hat{\sigma}^2$ is undefined.

There are at least two other causes of high variance in LSEs: overfitting and collinearity. Overfitting can invalidate the model by including too many irrelevant variables which are not related to the response. If the model starts to explain the noise in the training data then it will not generalize well to new observations. In contrast, collinearity can weaken the model by including too many relevant variables. If there are many variables that are highly correlated with the response, they could contain similar information and exhibit high pairwise correlations between themselves. Sections 2.1.6 and 2.1.7 look at why these sit-

uations inflate the variance. The risk of encountering them is higher when there are a large number of predictors available.

2.1.6 Overfitting

The regression function attempts to estimate the population average of the response variable via the deterministic relationship between the response and the predictors. When too many predictors are included in the model the regression function starts adapting to the specific training sample. It is possible to include enough predictors so that the model fits the training sample perfectly. However, such models are likely to be fitting the noise in the data and as a result, the model will perform poorly on any other data set because it does not represent the relationship inherent in the population. Including too many predictors in the model is called overfitting. The effect of overfitting is to increase the variances of the estimates and predictions.

Estimation

Suppose that the true model includes only d predictors. Without loss of generality, assume that these are the first $d+1$ columns of $\mathbf{X}$ (d predictors plus the intercept). Thus, $\mathbf{X} = (\mathbf{X}_{\mathcal{D}}, \mathbf{X}_{\mathcal{D}^c})$, where $\mathcal{D} = \{j : 0, 1, \ldots, d\}$ and $\underline{\beta}^T = \left(\underline{\beta}_{\mathcal{D}}^T, \underline{\beta}_{\mathcal{D}^c}^T\right) = \left(\underline{\beta}_{\mathcal{D}}^T, \underline{0}^T\right)$. Thus the true model is given by

$$E(\mathbf{y}) = f^{true}(\mathbf{X}) = \mathbf{X}_{\mathcal{D}}\underline{\beta}_{\mathcal{D}}.$$

Suppose we overfit the model by including all p predictors,

$$\hat{f}(\mathbf{X}) = \mathbf{X}\underline{\hat{\beta}} = \mathbf{X}_{\mathcal{D}}\underline{\hat{\beta}}_{\mathcal{D}} + \mathbf{X}_{\mathcal{D}^c}\underline{\hat{\beta}}_{\mathcal{D}^c}.$$

Expressions for the estimates, their expected values and variances are derived in Section B.2.1. It is shown that,

$$E\left(\underline{\hat{\beta}}_{\mathcal{D}}\right) = \underline{\beta}_{\mathcal{D}},$$
$$E\left(\underline{\hat{\beta}}_{\mathcal{D}^c}\right) = \underline{0},$$

and

$$\text{var}\left(\hat{\underline{\beta}}_{\mathcal{D}}\right) = \sigma^2\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1} + \mathbf{B}\mathbf{M}^{-1}\mathbf{B}^T,$$
$$\text{var}\left(\hat{\underline{\beta}}_{\mathcal{D}^c}\right) = \sigma^2\mathbf{M}^{-1},$$

where

$$\mathbf{B} = \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}^c}$$

is the estimate obtained when regressing $\mathbf{X}_{\mathcal{D}^c}$ on $\mathbf{X}_{\mathcal{D}}$ and

$$\mathbf{M} = \mathbf{X}_{\mathcal{D}^c}^T\left(\mathbf{I} - \mathbf{H}_{\mathcal{D}}\right)\mathbf{X}_{\mathcal{D}^c}$$

is the minimum RSS from that regression.

Thus, when overfitting the model, $\hat{\underline{\beta}}_{\mathcal{D}}$ is an unbiased estimate of $\underline{\beta}_{\mathcal{D}}$ but its variance is inflated by an amount of $\sigma^2\,\text{tr}\left(\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^T\right)$. Although $\hat{\underline{\beta}}_{\mathcal{D}^c}$ has zero expectation, its presence in the model further inflates the total variance of the model by $\sigma^2\,\text{tr}\left(\mathbf{M}^{-1}\right)$. Since there is no bias and the variance is increased, the MSE will always be larger when the model is overfitted.

Prediction

When using the overfitted model to predict a new response y_0 at $\underline{x}_0 = \left(\underline{x}_{0,\mathcal{D}}, \underline{x}_{0,\mathcal{D}^c}\right)$, it is shown in Section B.2.1 that

$$E\left(\hat{f}\left(\underline{x}_0\right)\right) = \underline{x}_{0,\mathcal{D}}^T\underline{\beta}_{\mathcal{D}}$$

so that the bias is

$$B\left(\hat{f}\left(\underline{x}_0\right)\right) = 0 = B\left(\hat{f}^{true}\left(\underline{x}_0\right)\right)$$

and the variance is

$$\begin{aligned}\text{var}\left(\hat{f}\left(\underline{x}_0\right)\right) &= \sigma^2\underline{x}_{0,\mathcal{D}}^T\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\underline{x}_{0,\mathcal{D}} + \sigma^2\underline{d}^T\mathbf{M}^{-1}\underline{d}\\ &= \text{var}\left(\hat{f}^{true}\left(\underline{x}_0\right)\right) + \sigma^2\underline{d}^T\mathbf{M}^{-1}\underline{d},\end{aligned}$$

where

$$\underline{d} = \mathbf{B}^T\underline{x}_{0,\mathcal{D}} - \underline{x}_{0,\mathcal{D}^c}.$$

Thus, the predictor of the true model and the predictor of the overfitted model are both unbiased. But the variance of the overfitted predictor is larger, so the expected PE when overfitting will always be larger.

2.1.7 Collinearity

When using least squares, problems arise when the $\mathbf{X}$ matrix does have full rank but is ill-conditioned so that the $\mathbf{X}^T\mathbf{X}$ matrix is nearly singular. The problem is usually due to some of the columns of $\mathbf{X}$ being highly correlated so that they are nearly linear dependent or collinear. Geometrically, collinearity occurs when a column vector $\mathbf{x}_j$ is nearly parallel to a subspace spanned by a set of other column vectors, or the angle between them is small (see Gentle (2007:202)). The effect of collinearity is to inflate the variances of the parameter estimates.

Estimation

Supposed that we fit the standardized model (2.1.24). Let $p = 2$, then

$$\mathbf{Z}^T\mathbf{Z} = \mathbf{R} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix},$$

where r_{12} is the sample correlation between X_1 and X_2. Then

$$\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} = \mathbf{R}^{-1} = \frac{1}{1-r_{12}^2}\begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}.$$

So the variances of the estimates are given by $\text{var}\left(\underline{\hat{\alpha}}\right) = \sigma^2\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}$, or

$$\text{var}\left(\hat{\alpha}_1\right) = \text{var}\left(\hat{\alpha}_2\right) = \frac{\sigma^2}{1-r_{12}^2}.$$

Thus, it is clear that the variances of the parameter estimates depend only on σ^2 and the correlation between the two variables. If the two variables are very highly correlated with r_{12}^2 close to 1 then the estimates will have very large variances.

Consider the straight line regression of X_1 on X_2 then, $r_{12}^2 = R_1^2$, where R_1^2 is the coefficient of determination (see Definition A.2.8) and measures how well the variation in X_1 is explained by X_2. The same

is true for the regression of X_2 on X_1, $r_{12} = R_2^2$. Therefore,

$$\text{var}\left(\hat{\alpha}_1\right) = \frac{\sigma^2}{1 - R_1^2} \text{ and } \text{var}\left(\hat{\alpha}_2\right) = \frac{\sigma^2}{1 - R_2^2}.$$

However, $r_{12}^2 = R_1^2 = R_2^2$ only holds for a straight line regression. Seber & Lee (2003:252-254) show that this result can be generalized for all p variables. By partitioning the correlation matrix $\mathbf{R}$ to isolate the correlations between the j-th variable and the other variables, the diagonal elements of $\mathbf{R}^{-1}$ can be related to R_j^2, the coefficient of determination when regressing $\mathbf{z}_j$ on the other columns of $\mathbf{Z}$. Thus, they show that

$$\text{var}\left(\hat{\alpha}_j\right) = \frac{\sigma^2}{1 - R_j^2},$$

where $1/(1 - R_j^2)$ are the diagonal elements of $\mathbf{R}^{-1}$,

$$1 - R_j^2 = \left\|\left(\mathbf{I} - \mathbf{H}_{-jj}\right)\mathbf{z}_j\right\|^2$$

and $\mathbf{H}_{-jj}$ is the projection matrix onto $\mathcal{C}\left(\mathbf{Z}_{-j}\right)$, the column space of the matrix $\mathbf{Z}$ with the j-th column removed.

Geometrically, $1 - R_j^2 = \left\|\left(\mathbf{I} - \mathbf{H}_{-jj}\right)\mathbf{z}_j\right\|^2$ measures the orthogonal distance between $\mathbf{z}_j$ and the subspace spanned by all the other columns of $\mathbf{Z}$. The smaller this orthogonal distance is, the smaller the angle between them and X_j is likely to be nearly collinear to the other variables. This idea can be used as a way to detect collinearity. For the j-th predictor, the variance inflation factor (VIF) is

$$VIF_j = \frac{1}{1 - R_j^2}.$$

Since R_j is a measure of the relationship between X_j and the remaining variables, VIF_j measures how much σ^2 / s_{jj} is inflated by that relationship. Now, $0 \leqslant R_j^2 \leqslant 1$ and $R_j^2 = 0$ when $\mathbf{z}_j$ is orthogonal to the other columns of $\mathbf{Z}$ because then $\underline{r}_j = \mathbf{z}_j^T \mathbf{Z}_{-j} = \underline{0}$. Thus, the minimum value of VIF_j is 1 when X_j is uncorrelated with all the other predictor variables. Very large values of VIF_j can indicate that X_j is nearly collinear with the other variables. Since the VIFs are the diagonal elements of $\mathbf{R}^{-1}$, we can detect collinearity by examining the eigenvalues and eigenvectors of $\mathbf{R}$. Consider the spectral decomposition of $\mathbf{R}$ (see Definition A.1.6),

$$\mathbf{R} = \mathbf{V}\mathbf{E}\mathbf{V}^T,$$

where $\mathbf{V}$ is an orthogonal matrix and $\mathbf{E}$ is a diagonal matrix. The columns of $\mathbf{V}$ are orthonormal eigenvectors of $\mathbf{R}$ and the diagonal elements of $\mathbf{E}$ are the eigenvalues of $\mathbf{R}$, denoted by e_j. Then

$$\begin{aligned}\mathbf{R}^{-1} &= \mathbf{V}\mathbf{E}^{-1}\mathbf{V}^T \\ &= \mathbf{V}\,\text{diag}\left(\frac{1}{e_j}\right)\mathbf{V}^T \\ &= \mathbf{V}\,\text{diag}\left(\frac{1}{d_j^2}\right)\mathbf{V}^T, \end{aligned} \tag{2.1.28}$$

since the eigenvalues of $\mathbf{R} = \mathbf{Z}^T\mathbf{Z}$ are the squared singular values of $\mathbf{Z}$, denoted by d_j^2. Thus,

$$VIF_j = \left(\mathbf{R}^{-1}\right)_{jj} = \sum_{k=1}^{p}\frac{v_{jk}^2}{e_j},$$

where $\left(\mathbf{R}^{-1}\right)_{jj}$ denotes the j-th diagonal element of $\mathbf{R}$. So any eigenvalue e_j of the correlation matrix that is close to zero could lead to large VIFs and indicates the presence of collinearity. See Seber & Lee (2003:255) or see Draper & Smith (1998:375-378) for more information.

More generally, any ill-conditioning of the $\mathbf{X}$ matrix with $\text{rank}(\mathbf{X}) = r$ can be determined by its spectral condition number,

$$\kappa_2(\mathbf{X}) = \frac{\max d_j(\mathbf{X})}{\min d_j(\mathbf{X})}, \tag{2.1.29}$$

where $d_j(\mathbf{X}) > 0$ are the singular values of $\mathbf{X}$ for $j = 1, 2, \ldots, r$. If $\mathbf{X}$ has full rank then the condition number can also be specified as

$$\kappa_2(\mathbf{X}) = \sqrt{\frac{\max e_j(\mathbf{X}^T\mathbf{X})}{\min e_j(\mathbf{X}^T\mathbf{X})}},$$

where $e_j\left(\mathbf{X}^T\mathbf{X}\right) > 0$ are the eigenvalues of $\mathbf{X}^T\mathbf{X}$. If $\mathbf{X}$ has orthonormal columns then its condition number is 1. It will be infinite when $\mathbf{X}$ is rank deficient and very large when $\mathbf{X}$ is ill-conditioned, or nearly rank deficient. Note that $\kappa_2\left(\mathbf{X}^T\mathbf{X}\right) = \left[\kappa_2(\mathbf{X})\right]^2$. See Gentle (2007:129-131,203-206) for more information about condition numbers and matrix norms. Seber & Lee (2003:256-260) show how changes in the data can affect the parameter estimates. Small changes in the observed data shouldn't cause a big change in the estimates, provided that the condition number of $\mathbf{X}$ is not too large, so that the regression is stable. The condition number of the correlation matrix can be examined as a first step in detecting collinearity.

Seber & Lee (2003:315-319) show that

$$\kappa_2(\mathbf{R}) \geqslant \frac{1}{\min e_j(\mathbf{R})} \geqslant VIF_j \text{ for all } j. \tag{2.1.30}$$

Thus, if the condition number of the correlation matrix is small then none of its eigenvalues would be too small and none of the VIFs will be large, so there shouldn't be any collinearity present. However, because small changes in the original data could cause large changes in the centered and scaled matrix, they suggest rather examining the matrix which is scaled but not centered. The condition number can also be used to determine the accuracy of the estimates, see Gentle (2007:218-219) for details.

Prediction

If we use the parameter estimates to predict a new response at $\underline{x}_0$,

$$\hat{f}(\underline{x}_0) = \underline{x}_0^T \hat{\underline{\beta}}$$

then the variance of the prediction is

$$\text{var}\left(\hat{f}(\underline{x}_0)\right) = \sigma^2 \underline{x}_0^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \underline{x}_0.$$

Thus the variance depends on $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ and it appears that they will suffer from collinearity, resulting in unstable predictions. However, Seber & Lee (2003:261) assert that the predictions are not affected by collinearity. They consider predicting the new response using the centered model (2.1.23) ,

$$\hat{f}(\underline{x}_0) = \bar{y} + (\underline{x}_0 - \bar{\underline{x}})^T \hat{\underline{\beta}}_{\mathcal{A}},$$

where $\mathcal{A} = \{j | j = 1, 2, \ldots, p\}$. Then the variance is

$$\begin{aligned}
\text{var}\left(\hat{f}(\underline{x}_0)\right) &= \text{var}(\bar{y}) + (\underline{x}_0 - \bar{\underline{x}})^T \text{var}\left(\hat{\underline{\beta}}_{\mathcal{A}}\right)(\underline{x}_0 - \bar{\underline{x}}) + 2\,\text{cov}\left(\bar{y}, \hat{\underline{\beta}}_{\mathcal{A}}\right) \\
&= \frac{1}{n}\sigma^2 + \sigma^2 (\underline{x}_0 - \bar{\underline{x}})^T \mathbf{S}^{-1} (\underline{x}_0 - \bar{\underline{x}}) \\
&= \sigma^2 \left[\frac{1}{n} + (\underline{x}_0 - \bar{\underline{x}})^T \mathbf{S}^{-1} (\underline{x}_0 - \bar{\underline{x}})\right],
\end{aligned}$$

since $\text{cov}\left(\bar{y}, \underline{\hat{\beta}}_{\mathcal{A}}\right) = 0$. Therefore predictions made at points that are close to the average observed values will have small variances regardless of any collinearity present. For outlying points far from $\underline{\bar{x}}$, the predictions will have large variances, although they argue that making predictions outside the range of observed data is generally discouraged.

2.2 Subset Selection Methods

2.2.1 Estimation

LSEs have the desirable properties of Section 2.1.2, provided that the model $f(\mathbf{X})$ has been specified correctly. Subset selection methods attempt to find the correct subset of the available predictors for the model specification. Once the subset is identified, the model is fitted using least squares. Assume that the true model is linear and has the form

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_p\beta_p + \varepsilon.$$

The problem of selecting a subset of $d < p$ predictors can be interpreted as finding the $p-d$ predictors that are not related to the response and setting their parameters to zero. Without loss of generality, suppose that the first d variables are selected. Gentle (2007:347) notes that the parameter estimates in the model

$$y_i = \sum_{i=1}^{d} x_i\beta_i + \varepsilon_i$$

are the same as the parameter estimates in the model

$$\hat{y}_i = \sum_{i=1}^{d} x_i\beta_i + \varepsilon_i,$$

where $\hat{y}_i$ are the fitted values from the model including all p variables. Thus, fitting a subset of variables is equivalent to approximating the least squares predictions. Subset selection could be helpful in addressing some of the problems with least squares, it eases interpretation and provides a direct way to prevent overfitting. Furthermore, the forward selection method discussed below can be used when $p > n$.

The LSEs obtained for a subset of variables will be biased. Suppose we estimate a subset $\widehat{\mathcal{D}}$ of size $\hat{d}$. Section B.2.2 derives the properties of the LSE when the model is underfitted. Suppose that the true model

is given by

$$E(\mathbf{y}) = f^{true}(\mathbf{X}) = \mathbf{X}_{\widehat{\mathcal{D}}}\underline{\beta}_{\widehat{\mathcal{D}}} + \mathbf{X}_{\widehat{\mathcal{D}}^c}\underline{\beta}_{\widehat{\mathcal{D}}^c},$$

where $\widehat{\mathcal{D}} = \{j : 0, 1, \ldots, \hat{d}\}$ and the model is underfitted using only $\hat{d}$ predictors,

$$\hat{f}(\mathbf{X}) = \mathbf{X}_{\widehat{\mathcal{D}}}\underline{\beta}_{\widehat{\mathcal{D}}}.$$

The LSE is $\underline{\hat{\beta}}^T = \left(\underline{\hat{\beta}}_{\widehat{\mathcal{D}}}^T, \underline{0}^T\right)$, with

$$E\left(\underline{\hat{\beta}}_{\widehat{\mathcal{D}}}\right) = \underline{\beta}_{\widehat{\mathcal{D}}} + \mathbf{B}\underline{\beta}_{\widehat{\mathcal{D}}^c}$$

and

$$\text{var}\left(\underline{\hat{\beta}}_{\widehat{\mathcal{D}}}\right) = \sigma^2\left(\mathbf{X}_{\widehat{\mathcal{D}}}^T\mathbf{X}_{\widehat{\mathcal{D}}}\right)^{-1}.$$

So, it is clear that the estimate is biased. Miller (2002:3-6) interprets $\mathbf{B}\underline{\beta}_{\widehat{\mathcal{D}}^c}$ as the bias in the first $\hat{d}+1$ estimates resulting from the omission of the last $p-\hat{d}$ variables, and calls it appropriately the omission bias. Although the estimate is biased, $\text{var}\left(\underline{\hat{\beta}}\right) \leqslant \text{var}\left(\underline{\hat{\beta}}^{true}\right)$ and the MSE is given by

$$MSE\left(\underline{\hat{\beta}}\right) = \sigma^2\,\text{tr}\left(\left(\mathbf{X}_{\widehat{\mathcal{D}}}^T\mathbf{X}_{\widehat{\mathcal{D}}}\right)^{-1}\right) + \underline{\beta}_{\widehat{\mathcal{D}}^c}^T\left(\mathbf{B}^T\mathbf{B} + \mathbf{I}\right)\underline{\beta}_{\widehat{\mathcal{D}}^c}.$$

Comparison with the true MSE,

$$MSE\left(\underline{\hat{\beta}}^{true}\right) = \sigma^2\left(\text{tr}\left(\mathbf{X}_{\widehat{\mathcal{D}}}^T\mathbf{X}_{\widehat{\mathcal{D}}}\right)^{-1}\right) + \text{tr}\left(\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^T\right) + \text{tr}\left(\mathbf{M}^{-1}\right)$$

shows that the biased estimator may be more efficient if

$$\underline{\beta}_{\widehat{\mathcal{D}}^c}^T\left(\mathbf{B}^T\mathbf{B} + \mathbf{I}\right)\underline{\beta}_{\widehat{\mathcal{D}}^c} < \text{tr}\left(\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^T\right) + \text{tr}\left(\mathbf{M}^{-1}\right).$$

Similarly, the predictions are biased but have reduced variance. The bias of the prediction at a new observation $\underline{x}_0 = \left(\underline{x}_{0,\widehat{\mathcal{D}}}, \underline{x}_{0,\widehat{\mathcal{D}}^c}\right)$ is

$$B\left(\hat{f}\left(\underline{x}_0\right)\right) = \underline{d}^T\underline{\beta}_{\widehat{\mathcal{D}}^c}$$

and the variance is

$$\text{var}\left(\hat{f}\left(\underline{x}_0\right)\right) = \sigma^2\underline{x}_{0,\widehat{\mathcal{D}}}^T\left(\mathbf{X}_{\widehat{\mathcal{D}}}^T\mathbf{X}_{\widehat{\mathcal{D}}}\right)^{-1}\underline{x}_{0,\widehat{\mathcal{D}}},$$

so that the expected PE is

$$PE\left(\hat{f}\left(\underline{x}_0\right)\right) = \sigma^2 + \sigma^2 \underline{x}_{0,\widehat{\mathcal{D}}}^T \left(\mathbf{X}_{\widehat{\mathcal{D}}}^T \mathbf{X}_{\widehat{\mathcal{D}}}\right)^{-1} \underline{x}_{0,\widehat{\mathcal{D}}} + \left(\underline{d}^T \underline{\beta}_{\widehat{\mathcal{D}}^c}\right)^2.$$

Comparison with the true PE,

$$PE\left(\hat{f}^{true}\left(\underline{x}_0\right)\right) = \sigma^2 + \sigma^2 \underline{x}_{0,\widehat{\mathcal{D}}}^T \left(\mathbf{X}_{\widehat{\mathcal{D}}}^T \mathbf{X}_{\widehat{\mathcal{D}}}\right)^{-1} \underline{x}_{0,\widehat{\mathcal{D}}} + \sigma^2 \underline{d}^T \mathbf{M}^{-1} \underline{d}$$

shows that the biased predictor has smaller expected PE when $\left(\underline{d}^T \underline{\beta}_{\widehat{\mathcal{D}}^c}\right)^2 < \sigma^2 \underline{d}^T \mathbf{M}^{-1} \underline{d}$. Seber & Lee (2003:398) show that this will be the case when $\underline{\beta}_{\widehat{\mathcal{D}}^c}^T \mathbf{M} \underline{\beta}_{\widehat{\mathcal{D}}^c} < \sigma^2$.

Any discussion of the LSEs above is under the premise that the model $f(\mathbf{X})$ is chosen *a priori*. Besides omission bias, subset selection will also suffer from selection bias unless independent data sets are used for model selection and estimation. Since the variables are selected adaptively (using the response) the selection bias will be high. All possible subsets suffers from very large bias since a large number of models are considered. Little is known about the properties of the parameters when subset selection methods are used, but it is clear that the effective degrees of freedom for the model is larger than the number of parameters in the model because of the adaptive selection (see page 3.1 for a discussion). The subset selection methods are not oracle estimators (see Definition A.3.8). They are not consistent because the greedy searches often select a local minimum. Further evidence of their inconsistency is due to instability, removing one observation from the training data could result in a completely different sequence of models. The estimation is inefficient due to their discrete nature, since parameters are either included or forced to zero, a particular parameter estimate can vary dramatically depending on which covariates are excluded.

2.2.2 All Possible Subsets

The all possible subsets procedure involves fitting every possible model including an intercept term and any number of predictor variables. If there are $p = 2$ predictor variables available then the following

models would be fitted:

$$Y = \beta_0 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

$$Y = \beta_0 + \beta_2 X_2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Each predictor variable is either included or excluded so that, if there are p available predictor variables, then there are a total of 2^p possible models. Thus, an exhaustive search is performed so that the method is capable of finding the globally optimal subset of variables. However, it can quickly become too expensive computationally. Table 2.2.1 below shows that this methods quickly becomes unfeasible as p is increased, with over a thousand possible models when $p = 10$, over 1 million when $p = 20$, and over 1 billion when $p = 30$!

Predictor Variables	All Possible Subsets	Forward / Backward
2	4	4
4	16	11
6	64	22
8	256	37
10	1,024	56
12	4,096	79
14	16,384	106
16	65,536	137
18	262,144	172
20	1,048,576	211
30	1,073,741,824	466

Table (2.2.1) Number of models considered in subset selection methods. All possible subsets quickly becomes infeasible, while forward selection and backward elimination require far less computation.

Algorithm 2.2.1, from James *et al.* (2013:205), describes the all possible subsets procedure. First the null model, including only the intercept, is fitted. For each subset size $k = 1, 2, \ldots, p$ all $\binom{p}{k}$ possible models are fitted and the best one is selected. This leads to a set of $p+1$ models, including the null model. The best of these models is selected using one of the information criteria in Section 3.2 or CV which is discussed in Section 3.3.1. See Draper & Smith (1998:329-334) for more information.

Algorithm 2.2.1 ***All possible subsets***

1. Fit the null model including only the intercept $\widehat{\mathcal{D}}_0$
2. For $k = 1, 2, \ldots, p$:
   (a) Fit all $\binom{p}{k}$ possible subsets of size k
   (b) Select the subset of size k which has the smallest RSS and call it $\widehat{\mathcal{D}}_k$
3. Using some selection criteria, select the best subset $\widehat{\mathcal{D}}_{\hat{d}}$ among $\widehat{\mathcal{D}}_0, \widehat{\mathcal{D}}_1, \ldots, \widehat{\mathcal{D}}_p$

Seber & Lee (2003:439-442) discuss algorithms for computing all possible subsets and show that the calculations can be reduced by 50% using sums of squares and cross-product matrices. Although, the use of these matrices can be inaccurate if there are high correlations between the predictors. Instead, orthogonal matrix reductions, such as the QR-decomposition (see Definition A.1.7), can be used for updating models to add or delete variables (see Seber & Lee (2003:446-447)). Miller (2002:11-36) and Lawson & Hanson (1974) also discuss efficient algorithms for performing least squares.

A way to further reduce computation is to only consider the c best models of each size, for some constant c. Clarke *et al.* (2009:572-573) discuss two variations of branch and bound optimization methods which eliminate subsets of variables in an efficient way. The computation for one of these methods, the leaps and bounds procedure, is described in Seber & Lee (2003:442-446). Miller (2002:48-54) also discusses algorithms and computational issues for these selection algorithms.

2.2.3 Forward Selection

The forward selection method begins with the null model, including only the intercept term so that $E(y) = \bar{y}$, and selects variables for inclusion one by one. Once a variable is included, it is retained in all further subsets. Miller (2002:39-40) describes the process with regards to the RSS. The variable that produces the smallest RSS is selected first. At each subsequent step, the variable selected is that which minimizes the RSS when added to the variables previously selected. The process continues until all the variables are included or until satisfying some stopping rule. He also relates the process to a comparison of correlations. If the variables have been centered, then the first variable selected has the largest correlation with the response. Each subsequent variable selected has the largest partial correlation with

the response, given the variables previously selected. James *et al.* (2013:207-208) describe the procedure, shown in Algorithm 2.2.2, to continue until all the variables are added and then to use some selection criteria to decide upon the best model. This could be more beneficial because there are $p+1$ models which can be examined instead of just the final model produced by a stopping rule. Their algorithm is shown below. It involves fitting $1+\ p\,(p+1)/\,2$ models, which is far more efficient than all possible subsets, see Table 2.2.1.

Algorithm 2.2.2 ***Forward selection***

1. Fit the null model including only the intercept $\widehat{\mathcal{D}}_0$
2. For $k = 0, 2, \ldots, p-1$
   (a) Fit all $p-k$ subsets of size $k+1$ which add one variable to subset $\widehat{\mathcal{D}}_k$
   (b) Select the subset of size $k+1$ which has the smallest RSS and call it $\widehat{\mathcal{D}}_{k+1}$
3. Using some selection criteria, select the best subset $\widehat{\mathcal{D}}_{\hat{d}}$ among $\widehat{\mathcal{D}}_0, \widehat{\mathcal{D}}_1, \ldots, \widehat{\mathcal{D}}_p$

Alternatively, the process has been described as performing a sequence of hypothesis tests. A version of the procedure where F-tests are used to decide which variable to include is described in Seber & Lee (2003:414). At each step, the variable which produces the largest value of the F-statistic is included if the statistic exceeds a specified value, say F_{in}. If no such variable is found then the procedure stops. F_{in} can be a specified value or it can be calculated as the critical value for the F-distribution corresponding to a significance level α. A drawback with this formulation is that the F-tests performed are not strictly valid since the F-statistics do not meet the necessary distributional requirements. Suppose we are at a single stage in the algorithm we want to test whether a variable should be added. The F-test assumes that, under the null hypothesis, the current model is the true model with normal residuals that are independently and identically distributed. If one of the remaining variables is chosen randomly then the F-statistic will follow the F-distribution. However, since the F-statistic selected for testing is the maximum among a set of statistics which are correlated, the statistic will not follow an F-distribution. Thus it seems arbitrary to choose values F_{in} and F_{out} with any meaning. Attempts have been made to correct this problem, see Draper & Smith (1998:343), Miller (2002:43-44) and Seber & Lee (2003:419), but there is no uncomplicated solution.

An advantage of using forward selection, as pointed out in Hastie *et al.* (2009:59), is that it can be performed even when $p \gg n$. Since we begin with the null model, the process can continue while $n > k$. However, there is no guarantee that the best subset of variables will be found with forward selection. The procedure follows a greedy algorithm, once a variable is chosen it cannot be reversed. Since variables are never discarded, all models of size $1, 2, \ldots, k-1$ are nested within the model of size k. It is possible that, when comparing the best fitting models of two different sizes, only some of the variables are present in both. Miller (2002:67-69) shows an example where the best fitting model of size 3 does not contain any of the variables that are in the best fitting model of size 2. Furthermore, forward selection is not likely to include any groups of variables. Miller (2002:41) provides an example where a linear combination of variables, $(X_1 - X_2)$, is an excellent predictor of the response but X_1 and X_2 are both poor predictors on their own. Since forward selection includes one variable at a time, it will often fail to include both variables. Essentially, the algorithm performs optimization locally. At each step, variables are considered that improve the current state of the model. Therefore, it is possible that the best fitting model could be completely overlooked by forward selection.

2.2.4 Backward Elimination

Backward elimination follows the same principal as forward selection but in reverse order. The method begins by including all the variables and at each step the variable is removed which produces the smallest residual sum of squares after its deletion. The process continues until all the variables are removed or until satisfying some stopping rule. James *et al.* (2013:208-209) provide the Algorithm 2.2.3 below for backward elimination, simply as the reverse of forward selection, again considering a total of $1 + p\,(p+1)/2$ models.

Algorithm 2.2.3 ***Backward selection***

1. Fit the full model including all the variables $\widehat{\mathcal{D}}_p$
2. For $k = p, p-1, \ldots, 1$
   (a) Fit all k subsets of size $k-1$ which removes one variable from subset $\widehat{\mathcal{D}}_k$
   (b) Select the subset of size $k-1$ which has the smallest RSS and call it $\widehat{\mathcal{D}}_{k-1}$
3. Using some selection criteria, select the best subset $\widehat{\mathcal{D}}_{\hat{d}}$ among $\widehat{\mathcal{D}}_0, \widehat{\mathcal{D}}_1, \ldots, \widehat{\mathcal{D}}_p$

Since backward elimination begins with the full model, it cannot be used when $p > n$ unless some unsupervised screening is utilized to obtain a smaller subset of variables to start with. However, backward elimination could be susceptible to similar drawbacks as least squares in the presence of collinearity. As with forward selection, the algorithm is greedy and searches are local so that the true model may be overlooked entirely. Although, Miller (2002:45) states that backward elimination would tend to keep groups of variables in the model. See Draper & Smith (1998:339-341), Miller (2002:44-45), Seber & Lee (2003:416,418) and Clarke *et al.* (2009:574) for more information.

2.2.5 Other Subset Selection Methods

Stepwise regression, or Efroymson's algorithm, is a combination of forward selection and backward elimination. At each step, variables can be entered or removed, thereby overcoming the greedy aspect of the forward and backward procedures. However, the algorithm proceeds by performing a series of hypothesis tests and its use is strongly discouraged for same the reasons mentioned on page 34. The method is discussed in Draper & Smith (1998:335-338), Miller (2002:42-43) and Seber & Lee (2003:418-419).

Hastie *et al.* (2009:60) describe a procedure called forward stagewise regression. The procedure is carried out on centered variables and starts with the null model. At each step, the variable which has the highest correlation with the residuals is found. The residuals are then regressed on this variable and its coefficient in the model is incremented by that regression coefficient. The process continues in that manner until none of the variables are correlated with the residuals. When $n > p$, the final model is the least squares model and it can take very many steps to reach it.

Miller (2002:46-48,54-47) discusses variations of the forward and backward procedures described above. He talks about sequential replacement algorithms which can be used in conjunction with one of the procedures. Once a number of variables are included in the model, a search is made to see if replacing any of the variables will lead to a reduction in RSS. The process continues until RSS cannot be reduced any further. He also discusses the possibility of considering pairs of variables for inclusion or exclusion instead of individual variables. Lastly, he mentions an untried method whereby variables are placed into smaller groups and all possible subsets is performed on each group. The groups would have to defined so that the RSS for variables in different groups are additive. This situation occurs when the variables are orthogonal. He does provide another case when this condition is met, but ultimately leaves the problem for further research.

2.3 Ridge Regression

2.3.1 Estimation

Ridge regression, proposed by Hoerl & Kennard (1970), is the oldest and most well known shrinkage method. A constraint is placed on the size of the parameters which has the effect of shrinking them towards zero. Since the constraint is focused on the size of the parameters, it is important that predictors are on the same scale to allow them equal consideration. Ridge regression is not scale invariant, a parameter estimate can change drastically if the scale of the predictor is changed and can be also be affected by the scale of other predictors. Therefore, the data is standardized before estimation. The ridge estimator is given by

$$\hat{\underline{\alpha}}^R = \underset{\underline{\alpha}}{\arg\min} \ \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 \ \text{subject to} \ \|\underline{\alpha}\|^2 \leqslant \tau, \tag{2.3.1}$$

where $\tau > 0$ and $RSS(\underline{\alpha}) = \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2$. The constrained problem is equivalent to the penalizing the RSS,

$$\hat{\underline{\alpha}}^R = \underset{\underline{\alpha}}{\arg\min} \ \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 + \lambda \|\underline{\alpha}\|^2, \tag{2.3.2}$$

where $\lambda \geqslant 0$ is chosen so that $\left\|\hat{\underline{\alpha}}^R\right\|^2 = \tau$. This can be shown by looking at the Karush-Kuhn-Tucker (KKT) optimality conditions (see Definition A.3.6). Suppose that $\hat{\underline{\alpha}}^R$ is the optimal solution for problem (2.3.2), then

$$\nabla \left\|\mathbf{v} - \mathbf{Z}\hat{\underline{\alpha}}^R\right\|^2 + \lambda \nabla \left\|\hat{\underline{\alpha}}^R\right\|^2 = 0. \tag{2.3.3}$$

For problem (2.3.1), the KKT conditions imply that $\underline{\alpha}^*$ and λ^* are optimal if:

1. $\|\underline{\alpha}^*\|^2 - \tau \leqslant 0$,
2. $\lambda^* \geqslant 0$,
3. $\lambda^* \left(\|\underline{\alpha}^*\|^2 - \tau\right) = 0$, and
4. $\nabla \|\mathbf{v} - \mathbf{Z}\underline{\alpha}^*\|^2 + \lambda^* \nabla \left(\|\underline{\alpha}^*\|^2 - \tau\right) = \nabla \|\mathbf{v} - \mathbf{Z}\underline{\alpha}^*\|^2 + \lambda^* \nabla \|\underline{\alpha}^*\|^2 = 0$.

Let $\left\|\hat{\underline{\alpha}}^R\right\|^2 = \tau$. If we set $\underline{\alpha}^* = \hat{\underline{\alpha}}^R$ then conditions (1) and (3) are met. If we also set $\lambda^* = \lambda$ then condition (4) is met because of equation (2.3.3). Therefore, problems (2.3.1) and (2.3.2) are equivalent when $\lambda \geqslant 0$ (2) and $\left\|\hat{\underline{\alpha}}^R\right\|^2 = \tau$ since they have the same solution.

The ridge penalty function is

$$P_R(\underline{\alpha}) = \lambda \|\underline{\alpha}\|^2 = \lambda \underline{\alpha}^T \underline{\alpha} = \lambda \sum_{j=1}^{p} \alpha_j^2. \tag{2.3.4}$$

We have that $P_R'(\underline{\alpha}) = 2\lambda\underline{\alpha}$ and $P_R''(\underline{\alpha}) = 2\lambda \geqslant 0$. Thus, the ridge penalty is differentiable and is strictly convex when $\lambda > 0$ for all $\underline{\alpha}$. Since both $RSS(\underline{\alpha})$ and $P_R(\underline{\alpha})$ are both positive quantities, each of them is minimized and it is clear that $P_R(\underline{\alpha})$ will be a minimum when $\alpha_1, \alpha_2, \ldots, \alpha_p$ are all close to zero. λ is called the shrinkage parameter and it controls the amount of shrinkage. A set of estimates can be produced, one for each value of λ, and we can follow the path of each parameter estimate as λ increases. If we set $\lambda = 0$, we would obtain the least squares estimates and the penalty would have no effect. As $\lambda \to \infty$, the effect of the penalty increases and $\hat{\underline{\alpha}}^R \to \underline{0}$. Note also that the gradient $P_R'(\underline{\alpha}) \propto \underline{\alpha}$ so that the shrinkage applied by ridge regression is proportional to the parameters.

Since the penalty function is differentiable, ridge regression can be solved explicitly. Setting the partial derivatives of (2.3.2) to zero, we have

$$\begin{aligned}
& \nabla\left(RSS(\underline{\alpha}) + P_\lambda^R(|\underline{\alpha}|)\right) = 0 \\
\Leftrightarrow\quad & \nabla\left\{(\mathbf{v} - \mathbf{Z}\underline{\alpha})^T(\mathbf{v} - \mathbf{Z}\underline{\alpha}) + \lambda\underline{\alpha}^T\underline{\alpha}\right\} = 0 \\
\Leftrightarrow\quad & -2\mathbf{Z}^T(\mathbf{v} - \mathbf{Z}\underline{\alpha}) + 2\lambda\underline{\alpha} = 0 \\
\Leftrightarrow\quad & \mathbf{Z}^T\mathbf{Z}\underline{\alpha} + \lambda\underline{\alpha} = \mathbf{Z}^T\mathbf{v} \\
\Leftrightarrow\quad & \left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)\underline{\alpha} = \mathbf{Z}^T\mathbf{v}.
\end{aligned}$$

Thus, the ridge regression estimator is

$$\hat{\underline{\alpha}}^R = \left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{v}. \tag{2.3.5}$$

with variance

$$\operatorname{var}\left(\hat{\underline{\alpha}}^R\right) = \sigma^2\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}\right)^{-1}. \tag{2.3.6}$$

2.3.2 Collinearity

Ridge regression corrects any problems with the correlation matrix, $\mathbf{R} = \mathbf{Z}^T\mathbf{Z}$. Adding the positive constant λ to each diagonal element of $\mathbf{R}$ has the following effect:

- If $\mathbf{R}$ is singular, $(\mathbf{R}+\lambda\mathbf{I})$ is nonsingular. Therefore, $\mathbf{Z}$ is not required to have full column rank and a unique estimator can be found when $p > n$.
- If $\mathbf{R}$ is nonsingular but is ill-conditioned, $(\mathbf{R}+\lambda\mathbf{I})$ is not ill-conditioned. Therefore, ridge regression overcomes any problems with collinearity.

Elaborating on the second point, Gentle (2007:206) states that, when $\mathbf{R}$ has full rank, the condition number of $(\mathbf{R}+\lambda\mathbf{I})$ is lower than the condition number of $\mathbf{R}$ since

$$\frac{\max\left(d_j+\lambda\right)}{\min\left(d_j+\lambda\right)} < \frac{\max\left(d_j\right)}{\min\left(d_j\right)},$$

where $\lambda > 0$ and here, $d_j = d_j(\mathbf{R})$, the singular values of $\mathbf{R}$. See equations (2.1.29) and (2.1.30). This can easily be confirmed. Since $\mathbf{R}$ has full rank, the singular values are all positive. Suppose that $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_p > 0$, then we can write

$$\begin{aligned}
&\frac{d_1+\lambda}{d_p+\lambda} < \frac{d_1}{d_p} \\
\Leftrightarrow\quad & (d_1+\lambda)\,d_p < d_1\left(d_p+\lambda\right) \\
\Leftrightarrow\quad & d_1d_p + \lambda d_p < d_1d_p + d_1\lambda \\
\Leftrightarrow\quad & \lambda d_p < \lambda d_1 \\
\Leftrightarrow\quad & d_p < d_1.
\end{aligned}$$

More directly, we have that

$$\begin{aligned}
(\mathbf{R}+\lambda\mathbf{I})^{-1} &= \begin{bmatrix} 1+\lambda & r_{12} & \cdots & r_{1p} \\ r_{12} & 1+\lambda & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1+\lambda \end{bmatrix}^{-1} \\
&= \left[(1+\lambda)\begin{bmatrix} 1 & \frac{r_{12}}{1+\lambda} & \cdots & \frac{r_{1p}}{1+\lambda} \\ \frac{r_{12}}{1+\lambda} & 1 & \cdots & \frac{r_{2p}}{1+\lambda} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{r_{1p}}{1+\lambda} & \frac{r_{2p}}{1+\lambda} & \cdots & 1 \end{bmatrix}\right]^{-1} \\
&= \frac{1}{1+\lambda}\begin{bmatrix} 1 & \frac{r_{12}}{1+\lambda} & \cdots & \frac{r_{1p}}{1+\lambda} \\ \frac{r_{12}}{1+\lambda} & 1 & \cdots & \frac{r_{2p}}{1+\lambda} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{r_{1p}}{1+\lambda} & \frac{r_{2p}}{1+\lambda} & \cdots & 1 \end{bmatrix}^{-1}.
\end{aligned} \tag{2.3.7}$$

Thus, ridge regression shrinks each of the correlations by a factor of $1/(1+\lambda)$, an operation called decorrelation by Zou & Hastie (2005). In addition, a direct shrinkage factor of $1/(1+\lambda)$ is applied to control

the variance.

Another way to deal with collinearity is to collect additional data. Seber & Lee (2003:322) point out that ridge regression can be interpreted as the least squares solution if the data is augmented with additional observations $\left(\sqrt{\lambda}\mathbf{I}_p, \mathbf{0}\right)$. Let

$$\underline{\hat{\alpha}}^R = \underset{\underline{\alpha}}{\arg\min} \ \|\mathbf{v}^* - \mathbf{Z}^* \underline{\alpha}\|^2,$$

where

$$\underset{(n+p)\times p}{\mathbf{Z}^*} = \begin{bmatrix} \mathbf{Z} \\ \sqrt{\lambda}\mathbf{I}_p \end{bmatrix} \text{ and } \underset{(n+p)\times 1}{\mathbf{v}^*} = \begin{bmatrix} \mathbf{v} \\ \mathbf{0} \end{bmatrix}. \tag{2.3.8}$$

Then the ridge estimate is obtained by applying least squares,

$$\begin{aligned} \underline{\hat{\alpha}}^R &= \left(\mathbf{Z}^{*T}\mathbf{Z}^*\right)^{-1}\mathbf{Z}^{*T}\mathbf{v}^* \\ &= \left[\begin{bmatrix} \mathbf{Z}^T & \sqrt{\lambda}\mathbf{I} \end{bmatrix}\begin{bmatrix} \mathbf{Z} \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix}\right]^{-1}\begin{bmatrix} \mathbf{Z}^T & \sqrt{\lambda}\mathbf{I} \end{bmatrix}\begin{bmatrix} \mathbf{v} \\ \mathbf{0} \end{bmatrix} \\ &= \left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{v}. \end{aligned}$$

2.3.3 Shrinkage

It is useful to compare ridge regression with least squares to examine how ridge regression shrinks the estimates and the effect that shrinkage has on the properties of the estimates. Assume that $\mathbf{Z}$ has full column rank so that the LSEs exists. We can write the ridge estimate as

$$\begin{aligned} \underline{\hat{\alpha}}^R &= \left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{v} \\ &= \left[\mathbf{Z}^T\mathbf{Z} + \lambda\left(\mathbf{Z}^T\mathbf{Z}\right)\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\right]^{-1}\mathbf{Z}^T\mathbf{v} \\ &= \left\{\left(\mathbf{Z}^T\mathbf{Z}\right)\left[\mathbf{I} + \lambda\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\right]\right\}^{-1}\mathbf{Z}^T\mathbf{v} \\ &= \left[\mathbf{I} + \lambda\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\right]^{-1}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{v} \\ &= \left[\mathbf{I} + \lambda\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\right]^{-1}\hat{\underline{\alpha}} = \mathbf{B}_\lambda\hat{\underline{\alpha}}, \end{aligned} \tag{2.3.9}$$

where $\mathbf{B}_\lambda = \left[\mathbf{I} + \lambda\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\right]^{-1}$ and $\hat{\underline{\alpha}}$ are the least squares estimates.

Consider the singular value decomposition (SVD) of $\mathbf{Z}$ (see Definition A.1.5),

$$\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}^T,$$

where $\mathbf{D}$ is a $p \times p$ diagonal matrix, $\mathbf{U}$ is an $n \times p$ matrix with orthogonal columns and $\mathbf{V}$ is a $p \times p$ orthogonal matrix. The diagonal elements of $\mathbf{D}$ are the singular values of $\mathbf{Z}$, with $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_p > 0$. The columns of $\mathbf{U}$ span $\mathcal{C}(\mathbf{Z})$ and the columns of $\mathbf{V}$ span $\mathcal{C}\left(\mathbf{Z}^T\right)$. Then,

$$\mathbf{R} = \mathbf{Z}^T\mathbf{Z} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T.$$

So,

$$\begin{aligned}
\mathbf{B}_\lambda &= \left[\mathbf{I} + \lambda\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\right]^{-1} \\
&= \left[\mathbf{V}\mathbf{V}^T + \lambda\left(\mathbf{V}\mathbf{D}^2\mathbf{V}^T\right)^{-1}\right]^{-1} \text{ since } \mathbf{V}\mathbf{V}^T = \mathbf{I} \\
&= \left[\mathbf{V}\mathbf{V}^T + \lambda\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^T\right]^{-1} \text{ since } \mathbf{V}^T = \mathbf{V}^{-1} \\
&= \left[\mathbf{V}\left(\mathbf{I} + \lambda\mathbf{D}^{-2}\right)\mathbf{V}^T\right]^{-1} \\
&= \mathbf{V}\left(\mathbf{I} + \lambda\mathbf{D}^{-2}\right)^{-1}\mathbf{V}^T \text{ since } \mathbf{V}^T = \mathbf{V}^{-1} \\
&= \mathbf{V}\,\mathrm{diag}\left(\frac{1}{1+\lambda/d_j^2}\right)\mathbf{V}^T \\
&= \mathbf{V}\,\mathrm{diag}\left(\frac{d_j^2}{d_j^2+\lambda}\right)\mathbf{V}^T
\end{aligned}$$

and $\underline{\hat{\alpha}}^R$ is given by

$$\underline{\hat{\alpha}}^R = \mathbf{B}_\lambda\underline{\hat{\alpha}} = \mathbf{V}\,\mathrm{diag}\left(\frac{d_j^2}{d_j^2+\lambda}\right)\mathbf{V}^T\underline{\hat{\alpha}}$$

It is clear that:

1. When $\lambda = 0$, $d_j^2 \Big/ \left(d_j^2 + \lambda\right) = 1$
2. When $\lambda > 0$, $d_j^2 < d_j^2 + \lambda$ so that $d_j^2 \Big/ \left(d_j^2 + \lambda\right) \in (0,1)$
   (a) $d_j^2 \Big/ \left(d_j^2 + \lambda\right) \to 0$ as $\lambda \to \infty$, and
   (b) $d_j^2 \Big/ \left(d_j^2 + \lambda\right) \to 0$ as $d_j^2 \to 0$.

With these results we can make some deductions about the form of shrinkage in ridge regression.

Since $\mathbf{V}$ is an orthonormal basis for $\mathcal{C}\left(\mathbf{Z}^T\right)$, $\mathbf{V}^T\underline{\hat{\alpha}}$ are the coordinates of $\underline{\hat{\alpha}}$ with respect to the basis $\mathbf{V}$ and are affected by a factor of $d_j^2 \Big/ \left(d_j^2 + \lambda\right)$. Points (1) and (2) above verify that the LSEs are not affected when $\lambda = 0$ and are shrunk towards zero when $\lambda > 0$, while (2a) clarifies that the amount of shrinkage increases as λ increases. Point (2b) states that the shrinkage affects the coordinates corresponding to the smaller singular values more. Now, $\mathbf{p}_j = \mathbf{Z}\underline{v}_j$ are the principal components of $\mathbf{X}$ and each $\mathbf{p}_j$ accounts for a proportion of the sample variance of $\mathbf{X}$. From the SVD we have that $\mathbf{Z}\underline{v}_j = \mathbf{u}_j d_j$ so that $\text{var}\left(\mathbf{p}_j\right) = d_j^2/n$. Since $d_1 \geqslant \cdots \geqslant d_p$, the first principal component $\mathbf{p}_1$ has the largest sample variance, $\mathbf{p}_2$ the second largest, and so forth until the last $\mathbf{p}_p$ which has the minimum variance. Thus, since the columns of $\mathbf{U}$ span the column space of $\mathbf{Z}$, the small singular values of $\mathbf{Z}$ correspond to directions of $\mathcal{C}\left(\mathbf{Z}\right)$ that have low variance. That is, those predictor variables whose observations are have no spread. It is these coordinates that experience the most shrinkage. See Hastie *et al.* (2009:66-67,79) and Seber & Lee (2003:325-326) for information about principal components.

2.3.4 Properties of Ridge Estimates

If we assume the linear model is correct, $E\left(\mathbf{v}\right) = \mathbf{Z}\underline{\alpha}$, then $\underline{\hat{\alpha}}^R$ is biased since

$$\begin{aligned} E\left(\underline{\hat{\alpha}}^R\right) &= E\left(\mathbf{B}_\lambda \underline{\hat{\alpha}}\right) \\ &= \mathbf{B}_\lambda E\left(\underline{\hat{\alpha}}\right) \\ &= \mathbf{B}_\lambda \underline{\alpha} \\ &= \mathbf{V}\,\text{diag}\left(\frac{d_j^2}{d_j^2 + \lambda}\right)\mathbf{V}^T\underline{\alpha}. \end{aligned}$$

Hastie *et al.* (2009:224-225) refer to this bias as estimation bias because the model is estimated in a restricted model space. The variance-covariance matrix of $\underline{\hat{\alpha}}$ in terms of the SVD is (also seen from (2.1.28)),

$$\begin{aligned} \text{var}\left(\underline{\hat{\alpha}}\right) &= \sigma^2\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} \\ &= \sigma^2\left(\mathbf{V}\mathbf{D}^2\mathbf{V}^T\right)^{-1} \\ &= \sigma^2\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^T \text{ since } \mathbf{V}^T = \mathbf{V}^{-1} \\ &= \sigma^2\mathbf{V}\,\text{diag}\left(\frac{1}{d_j^2}\right)\mathbf{V}^T, \end{aligned}$$

so that

$$\begin{aligned}
\text{var}\left(\underline{\hat{\alpha}}^R\right) &= \text{var}\left(\mathbf{B}_\lambda \underline{\hat{\alpha}}\right) \\
&= \mathbf{B}_\lambda \,\text{var}\left(\underline{\hat{\alpha}}\right) \mathbf{B}_\lambda^T \\
&= \mathbf{V}\left(\mathbf{I} + \lambda \mathbf{D}^{-2}\right)^{-1} \mathbf{V}^T \left(\sigma^2 \mathbf{V}\mathbf{D}^{-2}\mathbf{V}^T\right) \mathbf{V}\left(\mathbf{I} + \lambda \mathbf{D}^{-2}\right)^{-1} \mathbf{V}^T \\
&= \sigma^2 \mathbf{V}\left(\mathbf{I} + \lambda \mathbf{D}^{-2}\right)^{-2} \mathbf{D}^{-2} \mathbf{V}^T \text{ since } \mathbf{V}^T\mathbf{V} = \mathbf{I} \\
&= \sigma^2 \mathbf{V}\,\text{diag}\left(\frac{1}{d_j^2\left(1 + \lambda/d_j^2\right)^2}\right) \mathbf{V}^T \\
&= \sigma^2 \mathbf{V}\,\text{diag}\left(\frac{d_j^2}{\left(d_j^2 + \lambda\right)^2}\right) \mathbf{V}^T
\end{aligned}$$

Thus the ridge estimates have much lower variance than the least squares estimates for $\lambda > 0$. For small singular values or large values of λ, the decrease in the variances of the ridge estimates are extreme. Draper & Smith (1998:396-397) show that there is a value of λ for which $MSE\left(\underline{\hat{\alpha}}^R\right) < MSE\left(\underline{\hat{\alpha}}\right)$ and Seber & Lee (2003:322-323) show that if λ is sufficiently small then $MSE\left(\underline{\hat{\alpha}}^R\right)$ will decrease as λ increases. However, finding the value of λ that minimizes $MSE\left(\underline{\hat{\alpha}}^R\right)$ depends on the unknown parameters $\underline{\alpha}$ and σ^2.

As with least squares, the fitted model is a linear combination of the response,

$$\mathbf{Z}\underline{\hat{\alpha}}^R = \mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{v} = \mathbf{H}_\lambda \mathbf{v},$$

where $\mathbf{H}_\lambda = \mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}\right)^{-1}\mathbf{Z}^T$. The effective degrees of freedom for the model (see (3.1.10)) is therefore,

$$\begin{aligned}
df &= \text{tr}\left(\mathbf{H}_\lambda\right) = \text{tr}\left[\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}\right)^{-1}\mathbf{Z}^T\right] \\
&= \text{tr}\left[\mathbf{U}\mathbf{D}\mathbf{V}^T\left(\mathbf{V}\mathbf{D}^2\mathbf{V}^T + \lambda\mathbf{V}\mathbf{V}^T\right)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^T\right] \text{since } \mathbf{V}\mathbf{V}^T = \mathbf{I} \\
&= \text{tr}\left\{\mathbf{U}\mathbf{D}\mathbf{V}^T\left[\mathbf{V}\left(\mathbf{D}^2 + \lambda\mathbf{I}\right)\mathbf{V}^T\right]^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^T\right\} \\
&= \text{tr}\left[\mathbf{U}\mathbf{D}^2\left(\mathbf{D}^2 + \lambda\mathbf{I}\right)^{-1}\mathbf{U}^T\right] \text{ since } \mathbf{V}^T = \mathbf{V}^{-1} \text{ and } \mathbf{V}^T\mathbf{V} = \mathbf{I} \\
&= \text{tr}\left[\mathbf{U}^T\mathbf{U}\mathbf{D}^2\left(\mathbf{D}^2 + \lambda\mathbf{I}\right)^{-1}\right] \\
&= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} \text{ since } \mathbf{U}^T\mathbf{U} = \mathbf{I}.
\end{aligned}$$

Thus the degrees of freedom for the model decreases as λ increases. For a least squares fit the degrees of freedom is equal to the number of parameters, $p + 1$.

2.3.5 Model Selection

Ridge regression requires little computational effort. James *et al.* (2013:219) state that finding the ridge estimates simultaneously for all λ requires almost the same amount of computation as using least squares. One difficulty is the selection of the tuning parameter but many suggestions are available in the literature. Draper & Smith (1998:388-389) and Seber & Lee (2003:324) discuss the ridge trace, a plot of each $\hat{\alpha}_j^R$ against λ - the original authors suggested using the plot to determine at which value of λ the parameters are stabilized. Draper & Smith (1998:390-391) and Miller (2002:59) provide some formulae for calculating the optimal value of λ and also discuss using an iterative ridge regression algorithm for this purpose. Seber & Lee (2003:424) recommend using resampling methods to determine which value of λ to use.

Ridge regression provides an attractive method which can be used to fit the linear regression model when $p > n$ or when there is collinearity present. Although the parameter estimates are shrunk towards zero, they are unable to attain the value of zero exactly and no variable selection is performed. Draper & Smith (1998:391) explain how one could use ridge regression for variable selection purposes. The process involves performing ridge regression in two stages together with some kind of thresholding rule. For the first stage, the model is estimated and the optimal value for λ is selected. After inspection of the estimates, any estimates that are smaller than some pre-specified value, or threshold, are removed from the model. The model is then estimated using ridge regression again in a second stage where these parameters have been removed . For further details about ridge regression, see Draper & Smith (1998:387-400), Seber & Lee (2003:321-324,423-425) or Hastie *et al.* (2009:61-68).

Chapter 3

Model Selection

Subset selection methods and ridge regression both produce a set of models with varying complexity from which a best-fitting model must be selected. This chapter focuses on methods that are available for performing model selection, which are usually based on prediction error (PE). Section 3.1 examines the composition of PE and its relation to model complexity. Models with low complexity have high bias and models that are too complex have high variance. The model selected has the best balance of bias and variance, the one with minimum PE. However, we cannot calculate PE directly since we usually do not know the true model function. If enough data is available then model selection can be carried out using a validation data set. In the absence of extra data, we need to estimate PE using the training data. Although the training error is an over optimistic estimate of PE, the optimism of the training error can be estimated. Section 3.2 looks at some information criteria which adjust the training error to form a better estimate of PE. If Gaussian errors are assumed then methods that penalize the likelihood function can also be used. Alternatively, Section 3.3 explains how PE can be estimated by resampling data from the training set using CV(Section 3.3.1) or bootstrapping (Section 3.3.2). Choosing the best model for variable selection is slightly more difficult. The Bayesian information criterion (BIC) in Section 3.2 is consistent for variable selection and the one-standard-error rule mentioned in Section 3.3.1 can be also used to select more parsimonious models.

3.1 Prediction Error

Prediction error (PE) provides us with a measure of how well a predictive model performs so that we can assess the quality of the model. It also aids in model selection since it allows us to make comparisons among different models. Hastie *et al.* (2009:222) provide some guidelines on how to use the available data effectively when choosing and assessing a model. When there is sufficient data, they recommend splitting the data into three parts:

1. 50% training sample for estimating the models,

2. 25% validation sample for model selection,

3. 25% test sample for assessing the accuracy of the final model.

When the observations available are too few to allow for such a split of the data, they suggest using information criteria or resampling methods for the model selection step. PE was introduced in Section 2.1.3. This section takes a closer look at PE and how it relates to model complexity, particularly when used for model selection.

Suppose that our predictive model $\hat{f}(X)$ is estimated from a set of training data given by $\mathcal{T} = \{(\underline{x}_1, y_1), (\underline{x}_2, y_2), \ldots, (\underline{x}_n, y_n)\}$ that is drawn randomly from the population. The training error of the model is the average loss over the training sample. For squared error loss,

$$TE\left(\hat{f}(\underline{x})\right) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(\underline{x}_i)\right)^2 = \frac{RSS\left(\hat{\underline{\beta}}\right)}{n}. \tag{3.1.1}$$

The test error is the PE over an independent test sample. Hastie *et al.* (2009:220) provide these definitions for PE. Using a specific training sample $\mathcal{T}$, the test error, or extra-sample error, is

$$PE_{\mathcal{T}}\left(\hat{f}(\underline{x}_0)\right) = E\left[\left(y_0 - \hat{f}(\underline{x}_0)\right)^2 \middle| \mathcal{T}\right]. \tag{3.1.2}$$

The training set in this conditional expectation is fixed. (X_0, Y_0) is a new observation from the joint distribution of X and Y, and the expectation is over this distribution. The expected test error, or expected PE averages the randomness in the training data,

$$PE\left(\hat{f}(\underline{x}_0)\right) = E_{\mathcal{T}}E\left[\left(y_0 - \hat{f}(\underline{x}_0)\right)^2 \middle| \mathcal{T}\right] = E\left[PE_{\mathcal{T}}\left(\hat{f}(\underline{x}_0)\right)\right]. \tag{3.1.3}$$

Section B.1.2 shows that the expected PE can be divided into three components,

$$\begin{aligned} PE\left(\hat{f}(\underline{x}_0)\right) &= E\left[y_0 - \hat{f}(\underline{x}_0)\right]^2 \\ &= E\left[y_0 - E(y_0)\right]^2 + E\left[\hat{f}(\underline{x}_0) - f(\underline{x}_0)\right]^2 \\ &= \text{var}(y_0) + E\left[\hat{f}(\underline{x}_0) - E\left(\hat{f}(\underline{x}_0)\right)\right]^2 + \left[E\left(\hat{f}(\underline{x}_0)\right) - f(\underline{x}_0)\right]^2 \\ &= \sigma^2 + \text{var}\left(\hat{f}(\underline{x}_0)\right) + B\left(\hat{f}(\underline{x}_0)\right)^2. \end{aligned} \tag{3.1.4}$$

The first term $\text{var}(y_0) = \sigma^2$ is unavoidable, it is the variance of the new response around the true mean $f(\underline{x}_0)$. Hastie *et al.* (2009:223) calls it the irreducible error since our estimate of $f(\underline{x}_0)$ is unable to change it. The second term is the variance of $\hat{f}(\underline{x}_0)$ around its mean and the third term is the squared bias of $\hat{f}(\underline{x}_0)$, the squared difference between the mean of $\hat{f}(\underline{x}_0)$ and the true mean $f(\underline{x}_0)$. Hastie *et al.* (2009:224) split the average squared bias further,

$$\begin{aligned} E\left[B\left(\hat{f}(\underline{x}_0)\right)^2\right] &= E\left[E\left(\hat{f}(\underline{x}_0)\right) - f(\underline{x}_0)\right]^2 \\ &= E\left[f(\underline{x}_0) - \underline{x}_0^T\underline{\hat{\beta}}_*\right]^2 + E\left[\underline{x}_0^T\underline{\hat{\beta}}_* - E\left(\hat{f}(\underline{x}_0)\right)\right]^2 \\ &= \text{Ave}\left[\text{Model Bias}\right]^2 + \text{Ave}\left[\text{Estimation Bias}\right]^2, \end{aligned} \tag{3.1.5}$$

where they call $\underline{\hat{\beta}}_*$ the "best-fitting linear approximation to f", defined by

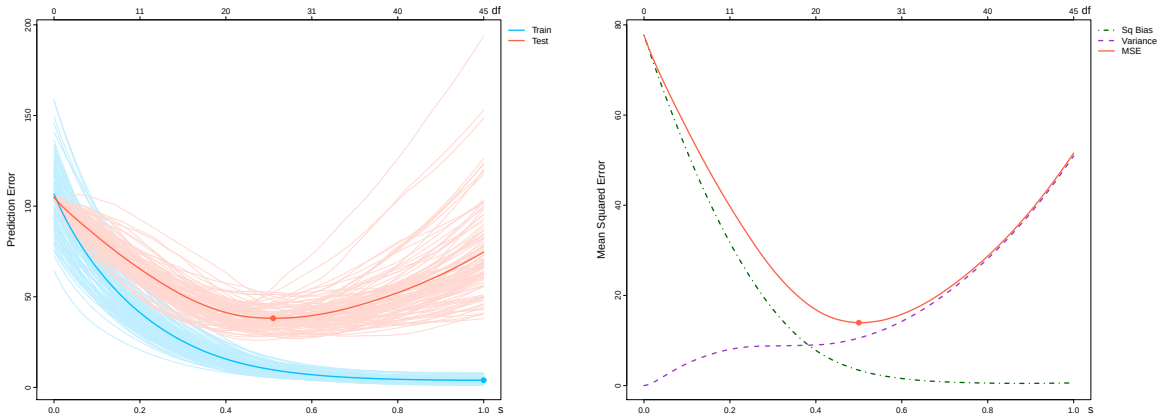
$$\underline{\hat{\beta}}_* = \underset{\underline{\beta}}{\arg\min} E\left(f(X) - X^T\underline{\beta}\right)^2.$$

The model bias is the difference between the true model and the best-fitting linear approximation. This bias is irreducible unless a larger class of linear models, including transformations and interactions, is considered (see Hastie *et al.* (2009:224)). The estimation bias is the difference between the best-fitting linear approximation and the average model estimate $E\left(\hat{f}(\underline{x}_0)\right)$.

For a least squares fit, model complexity is controlled by p, the number of variables in the model. We can denote the model by $\hat{f}_p(\underline{x}) = \underline{x}^T\underline{\hat{\beta}}_p$ with residual sum of squares RSS_p. If the model is too complex, the model starts to fit the noise in the training data - overfitting occurs and the model will be too variable. However, if the model is underfitted then it will be very biased. In both cases, the model will not generalize well to new data. The training error is given by RSS_p/n and is over optimistic since RSS_p always decreases as p is increased. Thus, training error is unable to detect overfitting. To asses the predictive power of the model, we need to look at the test error. The test error accounts for model complexity and starts to increase as we begin overfitting. The minimum test error enables us to choose the correct balance of bias and variance. The bias and variance are the last two terms in (3.1.4) and can be controlled by choosing different values of p. Generally, the model has high bias and low variance when p is small. As p increases, the bias decreases and the variance increases. Least squares does not have any estimation bias since $E\left(\hat{f}_p(\underline{x}_0)\right) = E\left(\underline{x}_0^T\underline{\hat{\beta}}_p\right) = \underline{x}_0^T\underline{\hat{\beta}}_*$.

For a ridge fit, model complexity is controlled by the parameter λ. We can denote the model by $\hat{f}_\lambda(\underline{x}) = \underline{x}^T\hat{\underline{\beta}}_\lambda$, where $\hat{\underline{\beta}}_\lambda$ is the vector of restricted estimates. The situation is similar and the last two terms in (3.1.4) and can be controlled by choosing different values of λ. The model has low bias and high variance when λ is small. As λ increases, the bias increases and the variance decreases. The only difference is that the ridge model has additional estimation bias since $E\left(\hat{f}_\lambda(\underline{x}_0)\right) = E\left(\underline{x}_0^T\hat{\underline{\beta}}_\lambda\right) \neq \underline{x}_0^T\hat{\underline{\beta}}_*$.

Figure 3.1.1 demonstrates these concepts for the LASSO using 100 training samples of size 60 each. There were 45 predictor variables, of which 15 were relevant. Model complexity is indexed by the fraction of the ℓ_1 norm, $s = \left\|\hat{\underline{\beta}}^L\right\|_1 \Big/ \left\|\hat{\underline{\beta}}\right\|_1$, where the null model occurs at 0 and the LSE at 1. Figure 3.1.1a shows how the training error is an optimistic estimate of the test error. The light curves represent the extra-sample error in equation (3.1.2) and the thick curves are the averages given by equation (3.1.3). The large test error near 0 is due to high bias and the variation in the test error increases as we approach 1. The trade-off between the bias and variance is seen clearly in Figure 3.1.1b. The squared bias steadily decreases and the variance increases as the complexity is increased, while the minimum MSE (marked with a point) provides the best balance between them.



(a) (b)

Figure (3.1.1) Prediction error and model complexity: (a) the training error and test error for 100 training samples of size 60, the average error is indicated by the thick solid lines; and (b) the composition of prediction error - the variance, squared bias and MSE for the simulated data.

Assume we have a sample of m test observations $\left(\underline{x}_{0,1}, y_{0,1}\right), \left(\underline{x}_{0,2}, y_{0,2}\right), \ldots, \left(\underline{x}_{0,m}, y_{0,m}\right)$ drawn randomly and independently from the same population as the training data. Let the test data be given by $\underset{1\times m}{\mathbf{y}_0^T} = (y_{0,1}, y_{0,2}, \ldots, y_{0,m})$ and $\underset{(p+1)\times m}{\mathbf{X}_0^T} = \left(\underline{x}_{0,1}, \underline{x}_{0,2}, \ldots, \underline{x}_{0,m}\right)$. .Assume that $\mathbf{y}_0$ has the same probability

structure as $\mathbf{y}$. That is, $E(\mathbf{y}_0) = f(\mathbf{X}_0)$, $\text{var}(\mathbf{y}_0) = \sigma^2\mathbf{I}$ and $\mathbf{y}_0$ is independent of $\mathbf{y}$. Then the expected PE is (see Section B.1.2),

$$PE\left(\hat{f}_p(\mathbf{X}_0)\right) = m\sigma^2 + MSE\left(\hat{f}_p(\mathbf{X}_0)\right).$$

Since the MSE is the major part of the expected PE, results are often reported for the MSE of the test data set $(\mathbf{y}_0, \mathbf{X}_0)$. For the linear model $f(\mathbf{X}_0) = \mathbf{X}_0\underline{\beta}$, we have

$$\begin{aligned}
MSE\left(\hat{f}(\mathbf{X}_0)\right) &= E_{X_0,Y_0}\left[\left(\mathbf{X}_0\underline{\hat{\beta}} - \mathbf{X}_0\underline{\beta}\right)^T\left(\mathbf{X}_0\underline{\hat{\beta}} - \mathbf{X}_0\underline{\beta}\right)\right] \\
&= E\left(\underline{\hat{\beta}}^T\mathbf{X}_0^T\mathbf{X}_0\underline{\hat{\beta}} - \underline{\beta}^T\mathbf{X}_0^T\mathbf{X}_0\underline{\hat{\beta}} + \underline{\beta}^T\mathbf{X}_0^T\mathbf{X}_0\underline{\beta} - \underline{\hat{\beta}}^T\mathbf{X}_0^T\mathbf{X}_0\underline{\beta}\right) \\
&= E\left[\left(\underline{\hat{\beta}} - \underline{\beta}\right)^T\mathbf{X}_0^T\mathbf{X}_0\underline{\hat{\beta}} - \left(\underline{\hat{\beta}} - \underline{\beta}\right)^T\mathbf{X}_0^T\mathbf{X}_0\underline{\beta}\right] \\
&= E\left[\left(\underline{\hat{\beta}} - \underline{\beta}\right)^T\mathbf{X}_0^T\mathbf{X}_0\left(\underline{\hat{\beta}} - \underline{\beta}\right)\right] \\
&= \left(\underline{\hat{\beta}} - \underline{\beta}\right)^T E\left(\mathbf{X}_0^T\mathbf{X}_0\right)\left(\underline{\hat{\beta}} - \underline{\beta}\right).
\end{aligned}$$

If we assume $E(X) = 0$ then $E\left(\mathbf{X}_0^T\mathbf{X}_0\right) = \mathbf{X}_0^T\mathbf{X}_0/n$, is the population variance-covariance matrix of $\mathbf{X}_0$.

In practice, the true form of the model is unknown. The expected test error cannot be evaluated directly and must be estimated. The training error TE is an underestimate of the test error $PE_{\mathcal{T}}$. Hastie *et al.* (2009:228) also call $PE_{\mathcal{T}}$ the extra-sample error since the observations of the predictor variables differ from the training observations. Conversely, they define the in-sample error as the error when new response values y_0 are observed at each of the training observations of the predictor variables. For squared error loss,

$$PE_{in}\left(\hat{f}(\underline{x})\right) = \frac{1}{n}\sum_{i=1}^{n} E\left[\left(y_{0,i} - \hat{f}\left(\underline{x}_i\right)\right)^2 \middle| \mathcal{T}\right]. \tag{3.1.6}$$

Hastie *et al.* (2009:228-229) define optimism as the difference between the training error and the in-sample error. Furthermore, they have the expected optimism, where the average is over the response values of the training sample (the parameters in the training set are fixed),

$$\begin{aligned}
\omega\left(\hat{f}(\underline{x})\right) &= E\left(PE_{in}\left(\hat{f}(\underline{x})\right)\right) - E\left(TE\left(\hat{f}(\underline{x})\right)\right) \\
&= \frac{2}{n}\sum_{i=1}^{n}\text{cov}\left(y_i, \hat{f}\left(\underline{x}_i\right)\right).
\end{aligned} \tag{3.1.7}$$

See Section B.1.3 for a proof. For a linear model where $\hat{f}(\mathbf{X}) = \mathbf{Hy}$,

$$\begin{aligned}
&\operatorname{cov}\left(y_i, \hat{f}\left(\underline{x}_i\right)\right) \\
&= \operatorname{cov}\left(y_i, \sum_{j=1}^{n} h_{ij} y_j\right) \\
&= h_{ii} \operatorname{var}\left(y_i\right) + \sum_{i \neq j} h_{ij} \operatorname{cov}\left(y_i, y_j\right) \\
&= h_{ii}\sigma^2 \text{ since } \operatorname{var}\left(y_i\right) = \sigma^2 \text{ and } \operatorname{cov}\left(y_i, y_j\right) = 0
\end{aligned}$$

and therefore,

$$\begin{aligned}
\omega\left(\hat{f}\left(\underline{x}\right)\right) &= \frac{2}{n} \sum_{i=1}^{n} \operatorname{cov}\left(y_i, \hat{f}\left(\underline{x}_i\right)\right) \\
&= \frac{2}{n}\sigma^2 \sum_{i=1}^{n} h_{ii} \\
&= \frac{2}{n}\sigma^2 \operatorname{tr}\left(\mathbf{H}\right). \qquad (3.1.8)
\end{aligned}$$

This leads to general definition for the effective degrees of freedom (DF) of a fitted model $\hat{f}(\underline{x})$. For any adaptively fitted model with additive error, $Y = f(X) + \varepsilon$, where $\operatorname{var}(\varepsilon) = \sigma^2$, the effective DF is given by

$$df = \frac{1}{\sigma^2} \sum_{i=1}^{n} \operatorname{cov}\left(y_i, \hat{f}\left(\underline{x}_i\right)\right) \qquad (3.1.9)$$

As we fit the model to the data using the response, the covariance between the response and the fitted model increases - the harder we fit, the larger the covariance. For a linear model $\hat{f}(\mathbf{X}) = \mathbf{Hy}$ we have

$$df = \operatorname{tr}\left(\mathbf{H}\right). \qquad (3.1.10)$$

For least squares estimation the DF is the number of estimated parameters in the model, $df = p + 1$, the number of variables plus the intercept. For maximum likelihood estimation, the DF is also defined as the number of estimated parameters in the model and in this case, $df = p + 2$ since σ^2 is also estimated. If a subset of d variables is fitted using least squares then $df = d + 1$ if the subset $\mathcal{D}$ is specified a priori. However, if subset selection is used to find the best subset $\widehat{\mathcal{D}}$ of size $\hat{d}$, the search for the optimal subset uses extra DF so that $df > \hat{d} + 1$. In this case, (3.1.9) can be estimated by simulation. See Hastie *et al.*

(2009:77-79,232-233) for a discussion. The expected in-sample error is therefore given by

$$E\left[PE_{in}\left(\hat{f}\left(\underline{x}\right)\right)\right] = E\left[TE\left(\hat{f}\left(\underline{x}\right)\right)\right] + \frac{2df}{n}\sigma^2. \tag{3.1.11}$$

For a least squares fit the in-sample error is (see Hastie *et al.* (2009:224) or Seber & Lee (2003:394-397)),

$$\begin{aligned} PE_{in}\left(\hat{f}\left(\underline{x}\right)\right) &= \frac{1}{n}\sum_{i=1}^{n} PE\left(\hat{f}\left(\underline{x}_i\right)\right) \\ &= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\text{var}\left(\hat{f}\left(\underline{x}_i\right)\right) + \frac{1}{n}\sum_{i=1}^{n}\left[E\left(\hat{f}\left(\underline{x}_i\right)\right) - f\left(\underline{x}_i\right)\right]^2 \text{ from (3.1.4)} \\ &= \sigma^2 + \frac{p+1}{n}\sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\left[E\left(\hat{f}\left(\underline{x}_i\right)\right) - f\left(\underline{x}_i\right)\right]^2. \end{aligned}$$

since

$$\begin{aligned} \sum_{i=1}^{n}\text{var}\left(\hat{f}\left(\underline{x}_i\right)\right) &= \text{tr}\left[\text{var}\left(\hat{f}\left(\mathbf{X}\right)\right)\right] \\ &= \text{tr}\left[\text{var}\left(\mathbf{Hy}\right)\right] \text{ from (2.1.7)} \\ &= \sigma^2\,\text{tr}\left(\mathbf{H}\right) = \left(p+1\right)\sigma^2 \end{aligned}$$

if $\mathbf{X}$ has full column rank. The in-sample error is therefore directly related to the number of variables p. However, $f\left(x_i\right)$ must be known to calculate it. The expected optimism is given by

$$\omega\left(\hat{f}\left(\underline{x}\right)\right) = \frac{2}{n}\sigma^2\,\text{tr}\left(\mathbf{H}\right) = \frac{2\left(p+1\right)}{n}\sigma^2.$$

So for least squares, the expected optimism increases as p increases, and decreases as the number of training observations n increases. Adding the expected optimism to the training error provides an estimate of the in-sample error,

$$\begin{aligned} \widehat{PE}_{in}\left(\hat{f}\left(\underline{x}\right)\right) &= TE\left(\hat{f}\left(\underline{x}\right)\right) + \omega\left(\hat{f}\left(\underline{x}\right)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}\left(\underline{x}_i\right)\right)^2 + \frac{2\left(p+1\right)}{n}\sigma^2 \\ &= \frac{RSS\left(\hat{\underline{\beta}}\right)}{n} + \frac{2\left(p+1\right)}{n}\sigma^2. \end{aligned}$$

In summary, prediction error provides a useful measure to assess the predictive performance of a model. Since PE is related to model complexity, it can be used for model selection to select the best balance of bias and variance. However, the expected PE cannot be calculated directly because it depends on the true form of the model $f(X)$ which is usually unknown. PE can be estimated by estimating the expected optimism and adding it to the training error. Although this leads to an estimate of the in-sample error, where the observations coincide with the training sample, Hastie *et al.* (2009:230) assert that it can be used effectively for model selection. Some of the information criteria in Section 3.2 are estimates of the expected in-sample error. The resampling methods in Section 3.3 estimate the PE, or expected extra-sample error, directly.

3.2 Information Criteria

Information criteria are used to make comparisons between models for the purpose of model selection. To select the best model of size $d \leqslant p$, the selected criterion is calculated for models of size $k = 1, 2, \ldots, p$ and the model corresponding to the best value is selected. The RSS and R^2 measures cannot be used to compare models of different sizes because they always improve when more variables are added to the model. The estimated error variance s^2 and the adjusted R^2 account for model complexity by adjusting these measures with the DF. Mallow's C_p works in a similar way and is also an estimate of the expected in-sample error (see 3.1.6). If the errors are Gaussian, nested models can be compared using F-statistics or the likelihood ratio test. These test will not be discussed here, the interested reader is referred to Searle (1971:124-125), Seber & Lee (2003:98-102) and Johnson & Wichern (2007:219-220). For non-nested models, measures that penalize the likelihood, like Akaike information criterion (AIC) and BIC can be used when the errors are Gaussian.

An advantage of information criteria is that they have considerably less computational expense than resampling methods - once the models are estimated, it is simply a matter of evaluating an expression for each model. However, there are some drawbacks. Although the criteria are defined by the number of variables in the model, this is only strictly correct for least squares or maximum likelihood estimation. Note that for s_k^2, $\bar{R}_k^2$ and C_k we use $k+1$ for the k predictors plus the intercept. When using the maximum likelihood, $\tilde{\sigma}^2$ is another estimable parameter and we use $k+2$ to account for it, except when σ^2 is assumed known in (3.2.12) and (3.2.13). The correct adjustment is thus for the effective DF . Now, DF is not always easy to specify. For a linear fit such as $\mathbf{y} = \mathbf{H}\mathbf{y}$, equation (3.1.10) can be used. However, for a nonlinear fit or

if the parameters are chosen adaptively, the equation does not hold and DF can be estimated by simulation using (3.1.9). Alternatively, Hastie *et al.* (2009:237-241) discuss the Vapnik–Chervonenkis Dimension, which is a generalization of (3.1.9). Another difficulty is that σ^2 is needed to calculate each criterion and a model that is roughly correct is necessary to estimate $\hat{\sigma}^2$. The criteria that penalize the likelihood function have the further disadvantage of relying strongly on distributional properties.

s^2 and Adjusted R^2

Let $RSS_k = RSS\left(\underline{\hat{\beta}}_k\right)$, the minimum RSS for a least squares model containing k predictors. Since least squares seeks to minimize RSS it can be used as a measure of how well the model fits. However, the RSS decreases as more variables are added to the model and should only be used to compare different models of the same size k. To compare models of different sizes, RSS_k can be adjusted by the residual DF to arrive at the estimated residual variance,

$$s_k^2 = \frac{RSS_k}{n-k-1}. \tag{3.2.1}$$

Let $\bar{s}^2(k)$ be the average of the s_k^2 for all models of size k. Draper & Smith (1998:331) suggest that s^2 starts to stabilize and approach the true error variance σ^2 as the model is being more and more overfitted. Thus, a plot of $\bar{s}^2(k)$ against k should reveal an approximate σ^2 and the ideal number of variables k. They assert that such a plot is most informative when there is a large number of variables and a large number of observations, specifically $p > 10$ and $5p \leqslant n \leqslant 10p$.

The coefficient of determination,

$$R_k^2 = 1 - \frac{RSS_k}{\sum (y_i - \bar{y})^2} \tag{3.2.2}$$

is also a well known measure for assessing least squares models. Similarly to RSS, the value of R^2 always increases as more variables are included in the model. The adjusted R^2 can be used to compare models of different sizes and and is given by

$$\bar{R}_k^2 = 1 - \frac{RSS_k/(n-k-1)}{\sum (y_i - \bar{y})^2/(n-1)}, \tag{3.2.3}$$

where both RSS and $\sum (y_i - \bar{y})^2$ are adjusted by their DF (see Draper & Smith (1998:139-140)). It is easily shown (see Seber & Lee (2003:400-401)) that the model with the maximum $\bar{R}^2$ is the same model with the minimum s^2.

Mallow's C_p

Mallows' C_p is defined for a model containing k predictors as

$$C_k = \frac{RSS_k}{\hat{\sigma}^2} + 2(k+1) - n, \tag{3.2.4}$$

where an estimate of σ^2 from a low bias model can be used. It is suggested to use the estimate $\hat{\sigma}^2 = s_p^2$ from the least squares fit including all the predictors available. See Clarke *et al.* (2009:572,579). If the largest model contains p variables, then $\hat{\sigma}^2 = RSS_p/(n-p-1)$ and for that model we have

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2(p+1) - n = \frac{RSS_p}{RSS_p/(n-p-1)} + 2(p+1) - n = p+1.$$

So by using $\hat{\sigma}^2$, we are comparing each model to the full model where $C_p = p+1$. In fact, for all k, the expected value of C_k is

$$\begin{aligned} E(C_k) &= \frac{E(RSS_k)}{E(\hat{\sigma}^2)} + 2(k+1) - n \\ &= \frac{(n-k-1)\sigma^2}{\sigma^2} + 2(k+1) - n \\ &= k+1, \end{aligned}$$

if the model is correct, $E(\mathbf{y}) = \mathbf{X}_k \underline{\beta}_k$, so that $E(RSS_k) = (n-k-1)\sigma^2$. Thus, a plot of C_k against k should reveal adequate models close to the $C_k = k+1$ line. According to Draper & Smith (1998:332), biased models will appear above the $C_k = k+1$ line. This is because RSS_k is larger when the model is biased so that $C_k > k+1$. Seber & Lee (2003:402) show that when the estimate $\hat{\sigma}^2$ from the largest model is used, and if n is much larger than k, then the lowest value of $C_k - k - 1$ coincides with the highest value of $\bar{R}_k^2$. Efron & Tibshirani (1993:242), Hastie *et al.* (2009:230) and James *et al.* (2013:211) define a version of the C_k statistic as an estimate of the expected in-sample PE,

$$\begin{aligned} C_k &= \widehat{PE}_{in}\left(\hat{f}\left(\underline{x}_k\right)\right) \\ &= TE\left(\hat{f}\left(\underline{x}_k\right)\right) + \frac{2(k+1)}{n}\hat{\sigma}^2 \text{ from (3.1.11)} \\ &= \frac{RSS_k}{n} + \frac{2(k+1)}{n}\hat{\sigma}^2. \end{aligned} \tag{3.2.5}$$

Seber & Lee (2003:402) show that C_k is an estimate of $E\left[MSE\left(\hat{f}\left(\mathbf{X}\right)\right)\right]/\sigma^2$. This result can be used to show that the two versions in equations (3.2.4) and (3.2.5) are proportional and differ by a factor of $\hat{\sigma}^2/n$. For more information, Miller (2002:116-127) discusses Mallow's C_p and modifications thereof in detail.

AIC and BIC

If the errors are Gaussian, we can use measures based on the likelihood function. If the true model contains $d \leqslant p$ predictors then the best model is selected having size

$$\hat{d} = \arg\min_k IC_k,$$

with

$$IC_k = D_k + \varphi_n\left(k\right) \tag{3.2.6}$$

where D_k is the minimum deviance (see (2.1.13)) for a model containing k predictors,

$$\begin{aligned} D_k &= -2\ln L_k\left(\underline{\tilde{\beta}}_k, \tilde{\sigma}^2|\varepsilon\right) \\ &= n\ln\left(2\pi\tilde{\sigma}^2\right) + \frac{RSS_k}{\tilde{\sigma}^2} && (3.2.7) \\ &= n\ln\left(2\pi\tilde{\sigma}^2\right) + n && (3.2.8) \\ &= n + n\ln\left(2\pi\right) + n\ln\left(\tilde{\sigma}^2\right), && (3.2.9) \end{aligned}$$

and $\varphi_n\left(k\right)$ is the penalty function which increases with n. Thus, these information criteria penalize the likelihood function L_k and the idea is quite similar to shrinkage methods which penalize RSS_k. However, the penalty function for shrinkage methods applies to the parameters $\underline{\beta}_k$ and depends on the shrinkage parameter λ, whereas the penalty function for these criteria applies to the number of predictors k and depends on the sample size n. These criteria are consistent for model selection if $\lim_{n\to\infty} P\left(\hat{d} = d\right) = 1$. Clarke *et al.* (2009:579) show that if $\lim_{n\to\infty} \varphi_n\left(k\right)/n = 0$ then $\lim_{n\to\infty} P\left(\hat{d} < d\right) = 0$ so that consistency for selection can be shown by proving that $\lim_{n\to\infty} P\left(\hat{d} > d\right) = 0$.

AIC has penalty function $\varphi_n\left(k\right) = 2\left(k+2\right)$. From (3.2.8), the AIC is given by

$$AIC_k = n\ln\left(2\pi\tilde{\sigma}^2\right) + n + 2\left(k+2\right). \tag{3.2.10}$$

It is derived from the Kullback-Leibler discrepancy and is an estimate of the discrepancy between the density of the true distribution of $\mathbf{y}$ and that specified by the model. Seber & Lee (2003:407-410) show that an unbiased estimate of the discrepancy between the densities is given by the modified criterion

$$\begin{aligned} CAIC_k &= n \ln\left(2\pi\tilde{\sigma}^2\right) + \frac{n\left(n+k+1\right)}{n-k-3} \\ &= AIC_k + \frac{2\left(k+2\right)\left(k+3\right)}{n-k-3} \\ &= D_k + \frac{2n\left(k+2\right)}{n-k-3}. \end{aligned} \tag{3.2.11}$$

The bias correction is necessary for small samples, Burnham & Anderson (2002:66) recommend using this version when $n/\left(k+2\right) < 40$. When n is large AIC_k and $CAIC_k$ are asymptotically equivalent since

$$\lim_{n\to\infty} \varphi_n\left(k\right) = \lim_{n\to\infty} \frac{2n\left(k+2\right)}{n-k-3} = 1.$$

When σ^2 is known, AIC is

$$AIC_k = \frac{RSS_k}{\sigma^2} + 2\left(k+1\right), \tag{3.2.12}$$

by using (3.2.7) and omitting the constant $n\log\left(2\pi\sigma^2\right)$ which does not dependent on the model. This form of AIC is proportional to C_p in (3.2.5) with σ^2 known, they differ by a factor of σ^2/n. In the literature, AIC has many similar definitions based on either equations (3.2.7) to (3.2.9), possibly removing constants and dividing by n or $\hat{\sigma}^2$. For example James *et al.* (2013:212) define

$$AIC_k = \frac{RSS_k}{n\hat{\sigma}^2} + \frac{2\left(k+2\right)}{n}. \tag{3.2.13}$$

Clarke *et al.* (2009:580-581) shows that $\lim_{n\to\infty} P\left(\hat{d} > d\right) > 0$ so that AIC_k is not consistent for selection. But $\lim_{n\to\infty} P\left(\hat{d} \geqslant d\right) = 1$, so asymptotically, AIC_k will select too many variables. In the finite sense, this can be seen by its small penalty function which does not depend on n. Because the deviance is not heavily penalized, more variables are allowed to enter the model. For this reason, Clarke *et al.* (2009:585-586) say that AIC is robust and should be used when selecting a model for prediction purposes. Furthermore, Clarke *et al.* (2009:580-581) show that the model chosen by AIC is minimax optimal. For small samples, or when p is large compared to n, $CAIC_k$ will select less variables than AIC_k because its penalty function is larger. Miller (2002:162-163) shows that AIC_k and C_k are equivalent, but if $\hat{\sigma}^2 = RSS_k/\left(n-k-1\right)$ is used for AIC_k then it will select more variables than C_k. For more information, Burnham & Anderson

(2002) provide extensive details about AIC.

The Bayesian information criterion (BIC), also known as the Schwarz criterion, penalizes the deviance with penalty function $\varphi_n(k) = \ln(n)(k+2)$. It is given by

$$BIC_k = n \ln\left(2\pi\tilde{\sigma}^2\right) + n + \ln(n)(k+2) \tag{3.2.14}$$

and when σ^2 is known then

$$BIC_k = \frac{RSS_k}{\sigma^2} + \ln(n)(k+1). \tag{3.2.15}$$

The BIC was developed from a Bayesian approach: choosing the model with minimum BIC is equivalent to choosing the model with the largest posterior probability. For a discussion about the Bayesian perspective of BIC, see Seber & Lee (2003:410-413), or Hastie *et al.* (2009:233-235). As with AIC, there appear to be many forms of BIC in the literature. In particular, James *et al.* (2013:212) define it as

$$BIC_k = \frac{RSS_k}{n} + \frac{(\ln n)(k+2)\hat{\sigma}^2}{n}. \tag{3.2.16}$$

The BIC penalty depends on n and is much stricter than AIC since $\ln n > 2$ when $n > e^2 = 7.389$. Thus, BIC allows less predictors into the model and is more appropriate for identifying the correct model. In fact, Clarke *et al.* (2009:584) show that BIC is consistent for selection because $\lim_{n\to\infty} P\left(\hat{d} > d\right) = 0$. However, it is not minimax optimal.

Other Information Criteria

The information criteria discussed above are among the most popular criteria used but there are many others. The minimum description length (MDL) minimizes the negative log-posterior distribution and is thus equivalent to the BIC which maximizes the posterior probability (see Miller (2002:158-160) and Hastie *et al.* (2009:235-237)). There are also a number of modified AIC and BIC criteria. Miller (2002:157) explains two modifications of the AIC, Risannen's criterion and the Hannan and Quinn (HQ) criterion. Clarke *et al.* (2009:586-587) also mentions the HQ criterion including some others, the deviance information criterion (DIC), the focused information criterion (FIC) and the covariance inflation criterion (CIC). Miller (2002:127-129) discusses the risk inflation criterion (RIC), which has a much smaller dependency on the number of variables. Shao (1997) examines the asymptotic properties of various information criteria for linear model selection.

3.3 Resampling Methods

Resampling methods consist of repeatedly drawing random samples from the training data and estimating the same model or measure on each sample. They can be immensely beneficial if there is not a wealth of data available since the same observations are reused in different subsets of data. A disadvantage of resampling is that it can be very computationally expensive. If the model is computationally expensive then resampling methods will be even more so, since they involve fitting the same model repeatedly on different subsets of data. These methods can be used to measure accuracy, such as finding standard errors of estimates; for model selection, like choosing the level of complexity for a method or deciding between different methods; and for model assessment, to ascertain how well the model performs. A major advantage is that they do not make any distributional assumptions.

3.3.1 Cross-Validation

Cross-validation is a resampling method that can be used for model selection and model assessment. In Section 3.1, it was recommended to split the data into three parts, a training set, a validation set and a test set, provided there is sufficient data to do so. If there is insufficient data to warrant having a designated validation set and test set, the training data can be split into parts to obtain a validation set. James *et al.* (2013:176-178) describe the validation set approach which randomly splits the observations into two equal parts, a training set and a validation set. The training observations are used to estimate the parameters of the model and those estimates are used to make predictions on the validation observations. The estimated test error for the validation set is used to assess the performance of the model. However, they show that estimating the test error in this way can be highly variable. They repeat the process ten times, each time randomly splitting the data into two equal parts, estimating the parameters on the training set and predicting the validation set. They point out that the variation of the test error among the ten models is large. The test error is very different each time, and highly depends on which observations are included in each data set. Furthermore, since the model is fit on only a subset of the data, they suggest that the estimated test error is an overestimate of the actual test error for the model fit on the full data set.

Repeatedly performing the process of splitting the data, estimating and predicting is the basis of CV. Each method splits the data in different ways and provides an estimate of PE. Hastie *et al.* (2009:241) assert that CV directly estimates the expected PE. Using CV avoids the problems incurred in the validation set approach.

K-Fold Cross-Validation

K-Fold CV repeatedly splits the data into K different parts, approximately equal in size, and uses one part for predictions and the remaining parts for estimation. For each k, the k-th part is used for validation and the other $k-1$ parts are used for training.

Algorithm 3.3.1 ***K-fold cross-validation***

1. *Randomly split the data into K parts*
2. *For* $k = 1, 2, \ldots, K$
   (a) *The training set includes all the observations except those in the k-th part. Fit the model to the training set and denote it by* $\hat{f}^{-k}(\underline{x})$.
   (b) *Use the model* $\hat{f}^{-k}(\underline{x})$ *to make predictions on the k-th part of the data which includes* n_k *observations.*
   (c) *Calculated the estimated test error for the k-th fold*

$$\widehat{PE}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\left(y_i - \hat{f}^{-k}\left(\underline{x}_i\right)\right).$$

3. *Calculate the CV estimate of PE is the average*

$$CV_K\left(\hat{f}\right) = \frac{1}{K}\sum_{k=1}^{K}\widehat{PE}_k. \tag{3.3.1}$$

Usually we choose $K = 5$ or $K = 10$. James *et al.* (2013:183-184) state that these values yield estimates which are not excessively biased or excessively variable. The size of each training sample is $(K-1)\,n/K$, so as K increases, the model is fit on a larger data set and the bias is reduced. However, the variance increases as K increases. The larger K is, the more the training sets overlap each other and the correlations increase between the fitted models, and therefore between the $\widehat{PE}_k$. Since $CV_K\left(\hat{f}\right)$ is the mean of the $\widehat{PE}_k$, it's variance is given by

$$\operatorname{var}\left[CV_K\left(\hat{f}\right)\right] = \frac{1}{K^2}\sum_{k=1}^{K}\operatorname{var}\left(\widehat{PE}_k\right) + \frac{2}{K^2}\sum_{j\neq k}\operatorname{cov}\left(\widehat{PE}_j, \widehat{PE}_k\right). \tag{3.3.2}$$

So the variance of the estimate increases as the correlations between the $\widehat{PE}_k$ increases. The choice of K thus plays an important role in estimating the PE. Similar discussions about the bias and variance of K-fold CV can be found in Clarke *et al.* (2009:593-594) and Hastie *et al.* (2009:242-243). James *et al.* (2013:182) notice that the variability of the K-fold CV estimate of PE is much lower than that of the validation set approach, there is still variation in how the data is split but not as much as when the data is split into two equal parts. They also note that the K-fold CV estimates are subject to moderate bias, an improvement over the validation set approach.

James *et al.* (2013:182-183) do simulation studies to examine how well K-fold CV estimates the true PE for different levels of complexity of a model. They find that, although the estimates can sometimes be biased and underestimate the true PE, they usually do a good job in identifying the minimum value of the PE. Thus, K-fold CV may be more suited for model selection, where the minimum is of importance, rather than for model assessment, where the actual value if of interest.

Breiman *et al.* (1984:Section 3.4.3) proposed the one standard error (1 SE) rule which they use for pruning classification and regression trees. Suppose we use CV for model selection and CV_K is plotted against model complexity. There is often a steep initial decrease in the curve, followed by a long flat tail and then possibly an increase. The minimum CV_K often lies somewhere on the long flat tail and can be unstable with slight up and down fluctuations. We can estimate the variance of each CV_K using equation (3.3.2), then the 1 SE rule selects the smallest model for which CV_K lies within one standard error of the minimum CV_K. The idea is to stabilize the selection and promote parsimony without losing accuracy. Hastie *et al.* (2009:244) recommend using the 1 SE rule for subset selection.

Leave One Out Cross-Validation

leave one out cross-validation (LOOCV) is a special case of K-fold CV with $K = n$. For $k = 1, 2, \ldots, n$, we remove the k-th observation $\left(\underline{x}_k, y_k\right)$ from the data set and fit the model $\hat{f}^{-k}\left(\underline{x}\right)$ on the remaining $n-1$ observations. The model is then used to predict the k-th observation and the estimated test error is calculated as $\widehat{PE}_k = \left(y_k - \hat{f}^{-k}\left(\underline{x}_k\right)\right)^2$. The LOOCV estimate of the PE is

$$\begin{aligned} CV_n\left(\hat{f}\right) &= \frac{1}{n}\sum_{k=1}^{n}\widehat{PE}_k \\ &= \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \hat{f}^{-k}\left(\underline{x}_k\right)\right)^2. \end{aligned} \tag{3.3.3}$$

James *et al.* (2013:179-180) point out some advantages of this method over the validation set approach. Since each model is fitted on $n-1$ observations, nearly the entire data set, LOOCV will not overestimate the test error like the validation set approach and the estimate will be approximately unbiased. Furthermore, since there is no randomness in data splits, LOOCV will always yield the same results if repeated multiple times.

Another consideration is that LOOCV can be computationally expensive if n is large or if the model requires extensive calculation since the model has to be fit n times. If this is the case, K-fold CV could be more feasible since the model is fitted only K times. Although, for a least squares fit, the following formula can be used to obtain the LOOCV estimate of the PE (see James *et al.* (2013:180)),

$$CV_n\left(\hat{f}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}\left(\underline{x}_i\right)}{1-h_{ii}}\right)^2, \tag{3.3.4}$$

where $\hat{y}_i$ is the i-th least squares fitted values and h_{ii} is the i-th diagonal element of the projection matrix $\mathbf{H}$ in (2.1.8). The values h_{ii} are called leverages, they reveal the extent to which an observation y_i effects it own fit $\hat{y}_i$. They can also be used to identify outlying $\underline{x}_i$ observations (see Draper & Smith (1998:207)).

Generalized Cross-Validation

generalized cross-validation (GCV) is an approximation to LOOCV. For any linear fitting method where we have $\hat{f}\left(\underline{x}\right) = \hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$, under squared-error loss the GCV is given by (see Hastie *et al.* (2009:244-245)),

$$GCV\left(\hat{f}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}\left(\underline{x}_i\right)}{1-\operatorname{tr}\left(\mathbf{S}\right)/n}\right)^2, \tag{3.3.5}$$

since $s_{ii} \approx \sum_{i=1}^{n} s_{ii}/n = \operatorname{tr}\left(\mathbf{S}\right)/n$. We saw in (3.1.9) that $\operatorname{tr}\left(\mathbf{S}\right)$ is the effective degrees of freedom in the model. Thus, for a least squares fit the GCV is

$$GCV\left(\hat{f}\right) = \frac{RSS_p}{n\left(1-p/n\right)^2} = \frac{n}{\left(n-p\right)^2}RSS_p = \frac{n}{n-p}\hat{\sigma}^2.$$

Clarke *et al.* (2009:591-592) show that GCV can be seen as a weighted LOOCV. Furthermore, they show that both methods are asymptotically equivalent to Mallows' C_p and AIC. The GCV is also known as the prediction sum of squares statistic (PRESS), see Miller (2002:143-146).

Other Cross-Validation Methods

Delete-d CV removes d observations from the data and holds them out as a validation set. The model is fitted on the remaining $n - d$ observations and then used to predict the validation set. A major disadvantage is its computational load. When $d = 1$, it is identical to LOOCV. The method is mentioned in Clarke *et al.* (2009:590) who provide references. Clarke *et al.* (2009:598-599) also mention Monte Carlo CV (MCCV), Bayesian CV and median CV.

3.3.2 Bootstrap

The bootstrap is a resampling method that is used to assess accuracy. It can be used to find standard errors of an estimate when the distribution of the estimate is unknown. It can also be used to estimate prediction accuracy for model assessment.

Standard Errors

Suppose the training observations are denoted by $\underline{t}_i = (\underline{x}_i, y_i)$ for $i = 1, 2, \ldots, n$. The training set can then be denoted by $T = (\underline{t}_1, \underline{t}_2, \ldots, \underline{t}_n)$. A bootstrap sample $T^{*b} = (\underline{t}_1^{*b}, \underline{t}_2^{*b}, \ldots, \underline{t}_n^{*b})$ is a sample of size n drawn randomly from the training set by sampling with replacement. It can be interpreted as a random sample drawn from $\hat{F}$, the empirical distribution of T. The distribution $\hat{F}$ is an estimate of the probability distribution F of T. It is a discrete distribution that puts a probability of $1/n$ on each of the observations $\underline{t}_i$ for $i = 1, 2, \ldots, n$. We can estimate any function of the probability distribution by using the empirical distribution. Suppose we would like to determine the standard errors of an estimate $\underline{\hat{\theta}} = \hat{f}(\underline{x})$ which is given by $SE_F(\underline{\hat{\theta}})$. The ideal bootstrap estimate of $SE_F(\underline{\hat{\theta}})$ uses the empirical distribution instead of the unknown probability distribution to produce $SE_{\hat{F}}(\underline{\hat{\theta}}^*)$. The sample standard deviation of the bootstrap replications $(\underline{\hat{\theta}}^{*1}, \underline{\hat{\theta}}^{*2}, \ldots, \underline{\hat{\theta}}^{*B})$ is a consistent estimate of $SE_{\hat{F}}(\underline{\hat{\theta}}^*)$. The following algorithm demonstrates the process.

Algorithm 3.3.2 ***Bootstrap standard errors***

1. *For* $b = 1, 2, \ldots, B$

   (a) *Draw an independent bootstrap sample* $T^{*b} = (\underline{t}_1^{*b}, \underline{t}_2^{*b}, \ldots, \underline{t}_n^{*b})$

(b) Calculate the estimate using the bootstrap sample to obtain the bootstrap estimate

$$\underline{\hat{\theta}}^{*b} = \hat{f}\left(\underline{x}^{*b}\right)$$

2. *Calculate the sample mean of bootstrap estimates*

$$\bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\underline{\hat{\theta}}^{*b}$$

3. *A consistent estimate of the standard errors of* $\underline{\hat{\theta}}$ *is given by*

$$\widehat{SE}_B = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\underline{\hat{\theta}}^{*b} - \bar{\theta}^*\right)^2}$$

Efron & Tibshirani (1993:52) suggest that the number of replications needed to provide a good estimate of $SE_F\left(\underline{\hat{\theta}}\right)$ could be as little as $B = 50$ and seldomly exceeds $B = 200$. See Efron & Tibshirani (1993:45-56,105-117) for more information on the bootstrap.

Prediction Error

A simple approach for using the bootstrap to estimate PE is to let the training sample act as the test sample. Draw B bootstrap samples from the training data. For $b = 1, 2, \ldots, B$, estimate the model $\hat{f}^{*b}\left(\underline{x}^{*b}\right)$ from the b-th bootstrap sample, then use the model to predict the training observations and obtain an estimate of the PE

$$\widehat{PE}_b = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}^{*b}\left(\underline{x}_i\right)\right)^2.$$

The simple bootstrap estimate is then the average of these B estimates

$$\widehat{PE}_{boot} = \frac{1}{B}\sum_{b=1}^{B}\widehat{PE}_b.$$

So the training set is being used for predictions. But each model is fitted on a bootstrap sample which is sampled with replacement from the training set. So the data sets being used for estimation and prediction both contain some of the same observations. Thus this is not a good estimate of PE, it tends to underestimate the true PE and promote overfitted models.

Efron & Tibshirani (1993:247-252) suggest a more refined bootstrap estimate by estimating the optimism and adding it to the training error. In the simple bootstrap approach, for $b = 1, 2, \ldots, B$, compute the error of the fitted model. That is, the error when the model estimated from the b-th bootstrap sample is used to predict the b-th bootstrap sample itself. We can view this as the training error for the b-th bootstrap sample,

$$TE_b = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{*b} - \hat{f}^{*b}\left(\underline{x}_i^{*b}\right)\right)^2.$$

Because the model is predicted on the exact same observations used for estimation, TE_b should be lower than $\widehat{PE}_b$. An estimate of the optimism for the b-th bootstrap sample is then given by $\widehat{PE}_b - TE_b$. To obtain an estimate of the expected optimism, we take the average over the B bootstrap samples,

$$\hat{\omega} = \frac{1}{B}\sum_{b=1}^{B}\left\{\widehat{PE}_b - TE_b\right\}.$$

If $TE = RSS/n$ is the training error from the original training sample, then the refined bootstrap estimate of PE is given by

$$\widehat{PE}_{rboot} = TE + \hat{\omega}.$$

Efron & Tibshirani (1993:252-254) and Hastie *et al.* (2009:251) discuss an alternative approach. Similar to LOOCV, it involves predicting each observation in the training sample using only the bootstrap samples which do not contain that observation. Suppose $\mathcal{B}^{-i}$ is the set that indexes which bootstrap samples do not contain the i-th observation and B_{-i} is the number of these bootstrap samples. Then the leave-one-out bootstrap estimate is given by

$$\widehat{PE}^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{b\in\mathcal{B}^{-i}}\left(y_i - \hat{f}^{*b}\left(\underline{x}_i\right)\right)^2 \Big/ B_{-i}.$$

Although it does not overfit, it is biased due to the size of the training set, resulting in an estimate that is usually larger than the true PE. When drawing a bootstrap sample of size n, at each draw, an observation from the training set has probability $1/n$ of being selected and probability $1 - 1/n$ of not being selected. Since each draw is independent, the probability that the i-th observation does not appear in the b-th bootstrap sample is

$$P\left(\underline{t}_i \notin \left\{T^{*b}\right\}\right) = \prod_{i=1}^{n}(1 - 1/n) = (1 - 1/n)^n,$$

where $\{T^{*b}\}$ is the set of observations included in the b-th bootstrap sample. Using the following definition of the exponential function,

$$\exp(x) = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n,$$

we can approximate this probability by $\exp(-1) = 0.368$. So, the probability that the i-th observation appears at least once in the b-th bootstrap sample is

$$\begin{aligned} P\left(\underline{t}_i \in \{T^{*b}\}\right) &= 1 - P\left(\underline{t}_i \notin \{T^{*b}\}\right) \\ &= 1 - (1-1/n)^n \\ &\approx 1 - \exp(-1) \\ &= 0.632. \end{aligned}$$

Efron & Tibshirani (1993:252-254) make use of this probability and define an adjusted estimate of the optimism as

$$\hat{\omega}^{(.632)} = 0.632\left(\widehat{PE}^{(1)} - TE\right).$$

The .632 estimate of the PE is then

$$\begin{aligned} \widehat{PE}^{(.632)} &= TE + \hat{\omega}^{(.632)} \\ &= TE + 0.632\left(\widehat{PE}^{(1)} - TE\right) \\ &= 0.368\ TE\ + 0.632\ \widehat{PE}^{(1)}. \end{aligned}$$

This adjusted estimate lowers the value of $\widehat{PE}^{(1)}$ and is roughly unbiased for the true PE. However, this estimate does not work well in overfit situations (Hastie *et al.* (2009:251-252)). By estimating the rate of overfitting, a further alternative for improving $\widehat{PE}^{(.632)}$ can be derived (see Efron & Tibshirani (1997)).

Chapter 4

LASSO Methods

The LASSO is explored in this chapter, with an introductory look at its penalty function in Section 4.1.1. To further understand how the penalty operates, the orthogonal design is examined. A closed form solution can be obtained in orthogonal designs and such functions are known as thresholding functions. Section 4.1.2 compares the LASSO thresholding (soft-thresholding), with that of ridge regression and subset selection (hard thresholding). Further insight can be gained by exploring the geometry of the LASSO in the two predictor case. Section 4.1.3 looks at the norm balls formed by penalties constructed with ℓ_q-norms and shows how the LASSO is able to perform variable selection in contrast to ridge regression. The efficiency history of algorithms for computing the LASSO and an overview of those currently available are given in Section 4.1.4. Suggestions and approximations for finding the standard errors of LASSO estimates are discussed in Section 4.1.5, followed by a look at the consistency and asymptotic properties of the LASSO. Section 4.1.6 concludes the exploration of the LASSO with an approximation of the DF and ways to select tuning parameter. Methods for controlling the bias of the LASSO and improving on its selection consistency are covered in Section 4.2. The relaxed LASSO algorithm and the adaptive LASSO are discussed in Section 4.2.1 and Section 4.2.2, respectively. Further modifications of the LASSO to incorporate different structure between the predictors are presented in Section 4.3, including the fused LASSO for ordered predictors (Section 4.3.1) and LASSO methods for including grouped variables (Section 4.3.2). Chapter 4 by no means provides an exhaustive look at all the methods and adaptations available today. The research in this field has been explosive since the LASSOs introduction in 1996.

4.1 The LASSO

4.1.1 Estimation

The least angle selection and shrinkage operator (LASSO) was introduced by Tibshirani (1996). The formulation is similar to ridge regression,

$$\hat{\underline{\alpha}}^L = \arg\min_{\underline{\alpha}} \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 \text{ subject to } \|\underline{\alpha}\|_1 \leqslant t, \tag{4.1.1}$$

where $t > 0$. The difference is that the LASSO constraint uses the ℓ_1 norm $\|\underline{\alpha}\|_1$, whereas the ridge constraint uses the squared ℓ_2 norm $\|\underline{\alpha}\|^2$. As with ridge regression, the problem 4.1.1 is equivalent to a penalized regression, in this case

$$\hat{\underline{\alpha}}^L = \arg\min_{\underline{\alpha}} \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 + \lambda \|\underline{\alpha}\|_1, \tag{4.1.2}$$

where $\lambda \geqslant 0$ is selected so that $\|\underline{\alpha}\|_1 = t$. The LASSO penalty function,

$$P_L(\underline{\alpha}) = \lambda \|\underline{\alpha}\|_1 = \lambda \sum_{j=1}^{p} |\beta_j|, \tag{4.1.3}$$

is convex and is always positive. So both RSS and the penalty function are minimized and it is clear that the penalty function will be be a minimum when $\alpha_1, \alpha_2, \ldots, \alpha_p$ are all close to zero. Thus, the LASSO also shrinks the parameter estimates towards zero. However, the LASSO penalty is non-differentiable at zero and it has the effect of setting some parameter estimates exactly to zero. Hence, the LASSO fits a subset of variables, thereby performing variable selection as well as shrinkage. When $\lambda = 0$ the penalty has no effect and $\hat{\underline{\alpha}}^L = \hat{\underline{\alpha}}$, the LSE, but when $\lambda \to \infty$, the null model is obtained $\hat{\underline{\alpha}}^L = \underline{0}$. A convenience of the LASSO is that the tuning parameters are bounded. Osborne *et al.* (2000b) show that $\hat{\underline{\alpha}}^L = \underline{0}$ for all $\lambda \geqslant \lambda_{\max} = \left\|\mathbf{Z}^T\mathbf{v}\right\|_\infty = \max_j \left|\mathbf{z}_j^T\mathbf{v}\right|$ so that a search for the optimal value of λ can commence on the interval $(0, \lambda_{\max})$. The constrained problem can also provide a closed range of tuning parameters to consider. Let $t_0 = \|\hat{\underline{\alpha}}\|_1$, the ℓ_1 norm of the LSE, then $\hat{\underline{\alpha}}$ is obtained whenever $t \geqslant t_0$ and more shrinkage is applied as $t \to 0$ with the null model occurring at $t = 0$. So we can search for the optimal t on the interval $(0, t_0)$, or more conveniently, we can parameterize the LASSO by a tuning parameter $s = t/t_0 \in (0,1)$. Note that the relationship between λ and t is not strictly one-to-one because of this behaviour. Table 4.1.1 summarizes the many-to-one relationship of these parameters at the boundaries of the LASSO path.

	Boundary	**Constraint Parameter**	**Penalty Parameter**
$\hat{\underline{\alpha}}^L = \underline{0}$	Null model	$t = 0$	$\lambda \geqslant \left\|\mathbf{Z}^T\mathbf{v}\right\|_\infty$
$\hat{\underline{\alpha}}^L = \hat{\underline{\alpha}}$	Least squares model	$t \geqslant \|\hat{\underline{\alpha}}\|_1$	$\lambda = 0$

Table (4.1.1) Tuning parameters at the LASSO path boundaries. For the null model, $t = 0$ for many values of λ and for the least squares model, $\lambda = 0$ for many values of t.

Since

$$|\alpha_j| = \begin{cases} \alpha_j & \text{if } \alpha_j > 0 \\ 0 & \text{if } \alpha_j = 0 \\ -\alpha_j & \text{if } \alpha_j < 0 \end{cases}$$

The subgradient of $P_L(\underline{\alpha})$ for $\alpha_j > 0$ is λ and is $-\lambda$ for $\alpha_j < 0$ so that the subdifferential is given by $\lambda\underline{\varpi}$ where

$$\varpi_j \in \begin{cases} 1 & \text{if } \alpha_j > 0 \\ [-1,1] & \text{if } \alpha_j = 0 \\ -1 & \text{if } \alpha_j < 0 \end{cases}. \tag{4.1.4}$$

So the LASSO shrinks all parameters at a constant rate since ϖ_j is constant for all j and does not depend on $\underline{\alpha}$. Note that $\varpi_j = \text{sign}(\alpha_j)$ and we have that $\|\underline{\alpha}\|_1 = \underline{\varpi}^T\underline{\alpha}$.

4.1.2 Orthogonal Design

To compare the effect of the LASSOs shrinkage and selection with the effects of traditional methods, it is useful to look at an orthogonal design where a closed form solution can be obtained. When the predictors are mutually orthogonal, we have that

1. $\mathbf{1}^T\mathbf{x}_j = 0$ for all $j = 1, 2, \ldots, p$, and
2. $\mathbf{x}_j^T\mathbf{x}_k = 0$ for all $j \neq k$

Then, $\mathbf{X}^T\mathbf{X}$ is the diagonal matrix

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & \mathbf{1}^T\mathbf{x}_1 & \cdots & \mathbf{1}^T\mathbf{x}_p \\ \mathbf{1}^T\mathbf{x}_1 & \mathbf{x}_1^T\mathbf{x}_1 & \cdots & \mathbf{x}_1^T\mathbf{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}^T\mathbf{x}_p & \mathbf{x}_1^T\mathbf{x}_p & \cdots & \mathbf{x}_p^T\mathbf{x}_p \end{bmatrix} = \begin{bmatrix} n & 0 & \cdots & 0 \\ 0 & \mathbf{x}_1^T\mathbf{x}_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{x}_p^T\mathbf{x}_p \end{bmatrix}.$$

Thus,

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{bmatrix} 1/n & 0 & \cdots & 0 \\ 0 & 1/\mathbf{x}_1^T\mathbf{x}_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\mathbf{x}_p^T\mathbf{x}_p \end{bmatrix},$$

so the LSEs are given by

$$\hat{\beta}_j = \frac{\mathbf{x}_j^T\mathbf{y}}{\mathbf{x}_j^T\mathbf{x}_j} \text{ with } \text{var}\left(\hat{\beta}_j\right) = \frac{1}{\mathbf{x}_j^T\mathbf{x}_j}\sigma^2,$$

for $j = 1, 2, \ldots, p$. Thus, in the orthogonal design, $\hat{\beta}_j$ does not depend on any predictors other than X_j and remains unchanged if we set $\beta_k = 0$ for any $j \neq k$. Furthermore, Seber & Lee (2003:52) show that

$$RSS\left(\underline{\beta}\right) = \mathbf{v}^T\mathbf{v} - \sum_{j=1}^{p} \hat{\beta}_j^2 \mathbf{x}_j^T \mathbf{x}_j,$$

so that $RSS\left(\underline{\beta}\right)$ increases by exactly $\hat{\beta}_k^2 \mathbf{x}_k^T \mathbf{x}_k$ when $\hat{\beta}_k = 0$.

When the predictors are orthogonal and standardized then the design is orthonormal, since we have

1. $\mathbf{1}^T\mathbf{z}_j = 0$ for all $j = 1, 2, \ldots, p$, and
2. $\mathbf{z}_j^T\mathbf{z}_k = 0$ for all $j \neq k$, and
3. $\mathbf{z}_j^T\mathbf{z}_j = 1$.

Hence, $\mathbf{Z}$ is an orthonormal matrix with $\mathbf{Z}^T\mathbf{Z} = \mathbf{I}$. This can also be seen by noting that

$$\mathbf{Z}^T\mathbf{Z} = \mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & \mathbf{1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{1} \end{bmatrix} = \mathbf{I}_p,$$

since the correlation between two orthogonal vectors is zero, $r_{jk} = 0$. Therefore the standardized LSEs are,

$$\underline{\hat{\alpha}} = \mathbf{Z}^T\mathbf{v} \text{ or } \hat{\alpha}_j = \mathbf{z}_j^T\mathbf{v}$$

with

$$\text{var}\left(\hat{\alpha}_j\right) = \sigma^2.$$

See Draper & Smith (1998:165-167) or Seber & Lee (2003:51-53) for more information on orthogonal designs.

Since $\mathbf{Z}^T\mathbf{Z} = \mathbf{I}$ and $\underline{\hat{\alpha}} = \mathbf{Z}^T\mathbf{v}$, we can write RSS as

$$\begin{aligned} RSS\left(\underline{\alpha}\right) &= \left\|\mathbf{v} - \mathbf{Z}\underline{\alpha}\right\|^2 \\ &= \mathbf{v}^T\mathbf{v} - 2\underline{\alpha}^T\mathbf{Z}^T\mathbf{v} + \underline{\alpha}^T\mathbf{Z}^T\mathbf{Z}\underline{\alpha} \\ &= \mathbf{v}^T\mathbf{Z}^T\mathbf{Z}\mathbf{v} - 2\underline{\alpha}^T\underline{\hat{\alpha}} + \underline{\alpha}^T\underline{\alpha} \\ &= \underline{\hat{\alpha}}^T\underline{\hat{\alpha}} - 2\underline{\alpha}^T\underline{\hat{\alpha}} + \underline{\alpha}^T\underline{\alpha} \\ &= \left\|\underline{\hat{\alpha}} - \underline{\alpha}\right\|^2. \end{aligned}$$

Thus, minimizing the penalized RSS is equivalent to minimizing

$$h_\lambda\left(\underline{\alpha}\right) = \frac{1}{2}\left\|\underline{\hat{\alpha}} - \underline{\alpha}\right\|^2 + P_\lambda\left(\left|\underline{\alpha}\right|\right), \tag{4.1.5}$$

for any penalty function $P_\lambda\left(\left|\underline{\alpha}\right|\right)$ of $\underline{\alpha}$ depending on λ.

Using the LASSO penalty, the objective function is

$$\begin{aligned} h_L\left(\underline{\alpha}\right) &= \frac{1}{2}\sum_{j=1}^{p}\left(\hat{\alpha}_j - \alpha_j\right)^2 + \lambda\sum_{j=1}^{p}\left|\alpha_j\right| \\ &= \sum_{j=1}^{p}\left[\frac{1}{2}\left(\hat{\alpha}_j - \alpha_j\right)^2 + \lambda\left|\alpha_j\right|\right]. \end{aligned} \tag{4.1.6}$$

Since 4.1.6 is additively separable,

$$h_L\left(\underline{\alpha}\right) = \sum_{j=1}^{p} h_L\left(a_j\right),$$

we have that

$$\frac{\partial}{\partial\alpha_j}h_L\left(\underline{\alpha}\right) = \frac{d}{d\alpha_j}h_L\left(\alpha_j\right)$$

so that minimizing 4.1.6 with respect to $\underline{\alpha}$ is equivalent to p component-wise minimizations with respect to α_j for $j = 1, 2, \ldots, p$. Because of the $-\alpha_j\hat{\alpha}_j$ term in the objective function, we choose α_j to have the same sign as $\hat{\alpha}_j$ to preserve the formation of the problem.

1. Suppose that $\hat{\alpha}_j > 0$, then for $j = 1, 2, \ldots, p$ we must minimize

$$h_L\left(\alpha_j\right) = \frac{1}{2}\hat{\alpha}_j^2 - \hat{\alpha}_j\alpha_j + \frac{1}{2}\alpha_j^2 + \lambda\alpha_j, \tag{4.1.7}$$

since $|\alpha_j| = \alpha_j$ when $\alpha_j \geqslant 0$. The derivative of (4.1.7) is

$$h'_L(\alpha_j) = \alpha_j - \hat{\alpha}_j + \lambda = \alpha_j - (\hat{\alpha}_j - \lambda).$$

(a) If $|\hat{\alpha}_j| < \lambda$ then $-(\hat{\alpha}_j - \lambda) > 0$ so that $h'_L(\alpha_j) > 0$ for all $\alpha_j \geqslant 0$. Thus (4.1.7) is strictly increasing for all $\alpha_j \geqslant 0$ and $\hat{\alpha}_j^L = 0$.

(b) If $|\hat{\alpha}_j| \geqslant \lambda /$ then $-(\hat{\alpha}_j - \lambda) \leqslant 0$ and setting $h'_L(\alpha_j) = 0$ gives the solution

$$\hat{\alpha}_j^L = \hat{\alpha}_j - \lambda = \text{sign}(\hat{\alpha}_j)(|\hat{\alpha}_j| - \lambda). \tag{4.1.8}$$

2. Similarly, for $\hat{\alpha}_j < 0$ we must minimize

$$h_L(\alpha_j) = \frac{1}{2}\hat{\alpha}_j^2 - \hat{\alpha}_j\alpha_j + \frac{1}{2}\alpha_j^2 - \lambda\alpha_j, \tag{4.1.9}$$

since $|\alpha_j| = -\alpha_j$ when $\alpha_j \leqslant 0$. The derivative of (4.1.9) is

$$h'_L(\alpha_j) = \alpha_j - \hat{\alpha}_j - \lambda = \alpha_j - (\hat{\alpha}_j + \lambda).$$

(a) If $|\hat{\alpha}_j| < \lambda$ then $-(\hat{\alpha}_j + \lambda) < 0$ so that $h'_L(\alpha_j) < 0$ for all $\alpha_j \leqslant 0$. Thus (4.1.9) is strictly decreasing for all $\alpha_j \leqslant 0$ and $\hat{\alpha}_j^L = 0$.

(b) If $|\hat{\alpha}_j| \geqslant \lambda$ then $-(\hat{\alpha}_j + \lambda) \geqslant 0$ and setting $h'_L(\alpha_j) = 0$ gives the solution

$$\hat{\alpha}_j^L = \hat{\alpha}_j + \lambda = -(-\hat{\alpha}_j - \lambda) = \text{sign}(\hat{\alpha}_j)(|\hat{\alpha}_j| - \lambda). \tag{4.1.10}$$

Now, we have the solution in the common form (4.1.8) and (4.1.10) for $|\hat{\alpha}_j| \geqslant \lambda$. We can incorporate the solution for $|\hat{\alpha}_j| < \lambda$ by considering only the positive part of $(|\hat{\alpha}_j| - \lambda)$,

$$\hat{\alpha}_j^L = \text{sign}(\hat{\alpha}_j)(|\hat{\alpha}_j| - \lambda)\,\delta(|\hat{\alpha}_j| \geqslant \lambda) = \text{sign}(\hat{\alpha}_j)(|\hat{\alpha}_j| - \lambda)_+, \tag{4.1.11}$$

where

$$\delta(a \in \mathcal{A}) = \begin{cases} 1 & \text{if } a \in \mathcal{A} \\ 0 & \text{if } a \notin \mathcal{A} \end{cases} \quad \text{and} \quad (a)_+ = \begin{cases} a & \text{if } a > 0 \\ 0 & \text{if } a \leqslant 0 \end{cases}. \tag{4.1.12}$$

The solution (4.1.11) can also be written as

$$S\left(\hat{\alpha}_j, \lambda\right) = \begin{cases} \hat{\alpha}_j - \lambda & \text{if } \hat{\alpha}_j > \lambda \\ 0 & \text{if } \left|\hat{\alpha}_j\right| \leqslant \lambda \\ \hat{\alpha}_j + \lambda & \text{if } \hat{\alpha}_j > -\lambda \end{cases}, \tag{4.1.13}$$

which is known as the soft thresholding rule.

Subset selection can be viewed in a similar manner to ridge regression and the LASSO. The problem can be stated as finding a subset of $d < p$ parameter estimates which minimizes RSS,

$$\underline{\hat{\alpha}}^{SS} = \underset{\underline{\alpha}}{\arg\min} \ \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 \ \text{subject to} \sum_{j=1}^{p} \delta\left(\alpha_j \neq 0\right) \leqslant d, \tag{4.1.14}$$

It is clear from the constraint $\sum_{j=1}^{p} \delta\left(\alpha_j \neq 0\right)$ that the process is discrete. The problem is equivalent to the penalized regression

$$\underline{\hat{\alpha}}^{L} = \underset{\underline{\alpha}}{\arg\min} \ \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 + \frac{\lambda^2}{2} \|\underline{\alpha}\|_0, \tag{4.1.15}$$

where the penalty function contains the so-called ℓ_0-"norm" (She (2009)),

$$P_{SS}\left(\underline{\alpha}\right) = \frac{\lambda^2}{2} \|\underline{\alpha}\|_0 = \frac{\lambda^2}{2} \sum_{j=1}^{p} \delta\left(\alpha_j \neq 0\right) \tag{4.1.16}$$

and $\lambda \geqslant 0$ is chosen so that $\sum_{j=1}^{p} \delta\left(\alpha_j \neq 0\right) = d$ is the number of nonzero parameters. With σ^2 known, AIC and BIC correspond to this penalty with $\lambda^2/2 = 2\sigma^2/n$ and $\lambda^2/2 = \ln(n)\,\sigma^2$, respectively (see Bühlmann & van de Geer (2011:20)). Using (4.1.5), we need to minimize the following objective function for orthogonal designs,

$$h_{SS}\left(\alpha_j\right) = \frac{1}{2}\left(\hat{\alpha}_j - \alpha_j\right)^2 + \frac{\lambda^2}{2}\delta\left(\alpha_j \neq 0\right).$$

If $\alpha_j = 0$ then $h_{SS}(0) = \frac{1}{2}\hat{\alpha}_j^2$. If $\alpha_j \neq 0$ then $h_{SS}\left(\alpha_j\right) = \left(\hat{\alpha}_j - \alpha_j\right)^2 + \lambda^2/2$ and the minimum is obtained when $\alpha_j = \hat{\alpha}_j$ with $h_{SS}\left(\hat{\alpha}_j\right) = \lambda^2/2$. Thus, $\hat{\alpha}_j$ is the solution if

$$\begin{aligned} & h_{SS}\left(\hat{\alpha}_j\right) < h_{SS}(0) \\ \Leftrightarrow \quad & \tfrac{1}{2}\lambda^2 < \tfrac{1}{2}\hat{\alpha}_j^2 \\ \Leftrightarrow \quad & \left|\hat{\alpha}_j\right| > \lambda \end{aligned}$$

and otherwise the solution is 0. Therefore,

$$\hat{\alpha}_j^{SS} = \hat{\alpha}_j \delta\left(\left|\hat{\alpha}_j\right| > \lambda\right).$$

This is also known as the hard thresholding rule,

$$H\left(\hat{\alpha}_j, \lambda\right) = \begin{cases} \hat{\alpha}_j & \text{if } \left|\hat{\alpha}\right|_j > \lambda \\ 0 & \text{if } \left|\hat{\alpha}_j\right| \leqslant \lambda \end{cases}. \tag{4.1.17}$$

The ridge estimator is given in (2.3.5) as

$$\underline{\hat{\alpha}}^R = \left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{v}.$$

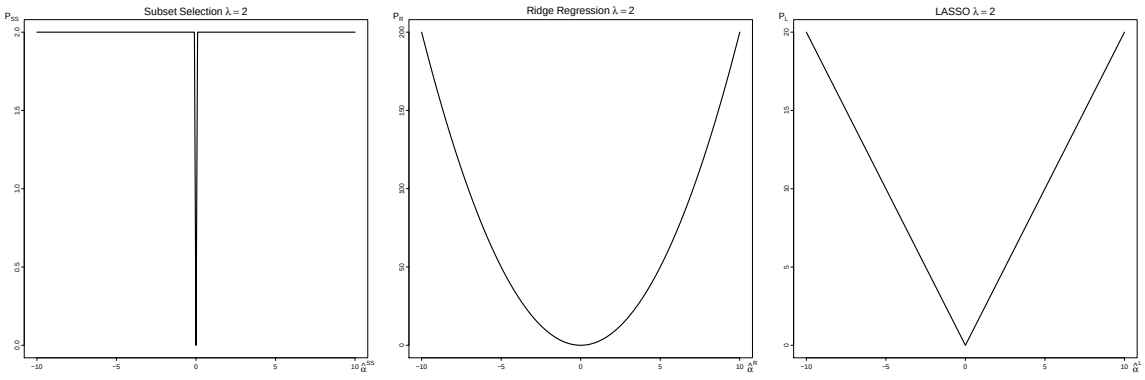
So for the orthogonal design we have $\underline{\hat{\alpha}}^R = (1+\lambda)^{-1}\,\underline{\hat{\alpha}}$ or

$$\hat{\alpha}_j^R = \frac{\hat{\alpha}_j}{1+\lambda}.$$

Donoho & Johnstone (1994) proposed using the functions (4.1.13) and (4.1.17) to denoise the estimates obtained when using wavelet transforms for function estimation. The estimates obtained in the orthogonal design can hence be viewed as thresholding functions. Table 4.1.2 summarizes the penalty functions and thresholding functions for subset selection, ridge regression and the LASSO. The penalty functions and thresholding functions for each method are shown in Figures 4.1.1 and 4.1.2, revealing how they shrink the LSEs when $\lambda = 2$. Subset selection does not perform any shrinkage and discretely sets parameters to zero. Conversely, ridge regression does not set any parameters to zero and the shrinkage is proportional to the size of the LSEs. The LASSO is a compromise between the two traditional methods, setting some parameters to zero and shrinking others, although the shrinkage is constant and does not depend on the size of the LSEs.

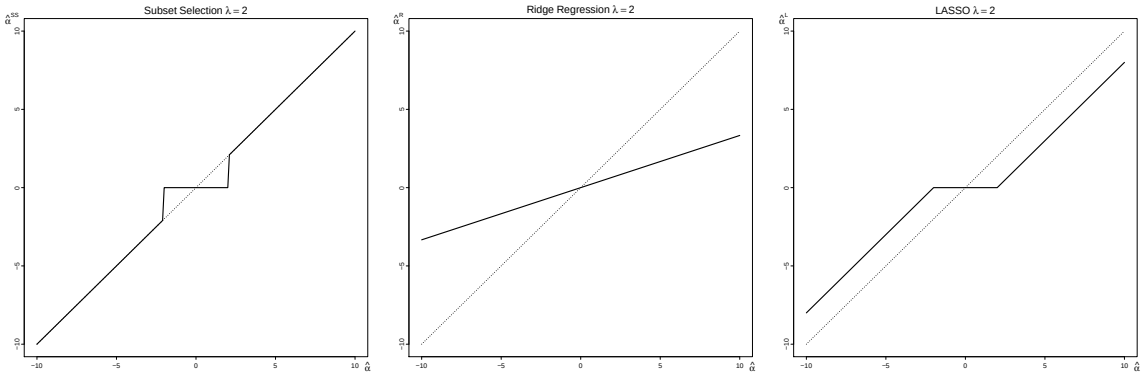
Method	Penalty Function	Thresholding Function
Subset Selection	$\lambda \sum \delta\left(\alpha_j \neq 0\right)$	$\hat{\alpha}_j \delta\left(\left\|\hat{\alpha}_j\right\| > \lambda\right)$
Ridge regression	$\lambda \sum \alpha_j^2$	$\hat{\alpha}_j / \left(1+\lambda\right)$
LASSO	$\lambda \sum \left\|\alpha_j\right\|$	$\text{sign}\left(\hat{\alpha}_j\right)\left(\left\|\hat{\alpha}_j\right\| - \lambda\right)_+$

Table (4.1.2) Penalty and thresholding functions for subset selection, ridge regression and LASSO



(a) (b) (c)

Figure (4.1.1) Penalty functions with $\lambda = 2$ for (a) subset selection, (b) ridge regression and (c) LASSO. The LASSO penalty is non-differentiable at zero.



(a) (b) (c)

Figure (4.1.2) Thresholding functions with $\lambda = 2$ for (a) subset selection, which is discrete with a jump to zero between $\hat{\alpha} \in [-\lambda, \lambda]$, (b) ridge regression, which shrinks estimates proportional to their size, and (c) the LASSO, which sets estimates to zero and applies constant shrinkage to nonzero estimates.

4.1.3 Geometry

To understand why the LASSO is able to perform variable selection, it is helpful to look at the geometric interpretation. The RSS can be written as

$$\begin{aligned} RSS(\underline{\alpha}) &= \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 \\ &= \|(\mathbf{v} - \mathbf{Z}\underline{\hat{\alpha}}) - (\mathbf{Z}\underline{\alpha} - \mathbf{Z}\underline{\hat{\alpha}})\|^2 \\ &= \|(\mathbf{v} - \mathbf{Z}\underline{\hat{\alpha}}) - \mathbf{Z}(\underline{\alpha} - \underline{\hat{\alpha}})\|^2 \\ &= \|\mathbf{v} - \mathbf{Z}\underline{\hat{\alpha}}\|^2 - 2(\underline{\alpha} - \underline{\hat{\alpha}})^T \mathbf{Z}^T (\mathbf{v} - \mathbf{Z}\underline{\hat{\alpha}}) + (\underline{\alpha} - \underline{\hat{\alpha}})^T \mathbf{Z}^T \mathbf{Z}(\underline{\alpha} - \underline{\hat{\alpha}}) \\ &= RSS(\underline{\hat{\alpha}}) + (\underline{\alpha} - \underline{\hat{\alpha}})^T \mathbf{Z}^T \mathbf{Z}(\underline{\alpha} - \underline{\hat{\alpha}}). \end{aligned}$$

since, from the normal equations,

$$\mathbf{Z}^T (\mathbf{v} - \mathbf{Z}\underline{\hat{\alpha}}) = \mathbf{0}.$$

Thus,

$$RSS(\underline{\alpha}) = (\underline{\alpha} - \underline{\hat{\alpha}})^T \mathbf{Z}^T \mathbf{Z}(\underline{\alpha} - \underline{\hat{\alpha}}) \tag{4.1.18}$$

up to an additive constant. So, RSS is a p-dimensional hypersurface in $\mathbb{R}^{p+1}$ space known as a quadric surface.

Consider the 3-dimensional case where $p = 2$. Then,

$$\mathbf{Z}^T \mathbf{Z} = \mathbf{R} = \begin{bmatrix} 1 & r_{12} \\ r_{11} & 1 \end{bmatrix}$$

with determinant

$$\left|\mathbf{Z}^T \mathbf{Z}\right| = 1 - r_{12}^2$$

and inverse

$$\left(\mathbf{Z}^T \mathbf{Z}\right)^{-1} = \frac{1}{1 - r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{11} & 1 \end{bmatrix}.$$

Furthermore, $\mathbf{v}^T\mathbf{v} = \sum_i (y_i - \bar{y})^2 = s_{yy}$ and

$$\begin{aligned}\mathbf{Z}^T\mathbf{v} &= \begin{bmatrix} \mathbf{z}_1^T\mathbf{v} \\ \mathbf{z}_2^T\mathbf{v} \end{bmatrix} \\ &= \begin{bmatrix} \sum_i (x_{i1} - \bar{x}_1)(y_i - \bar{y})/\sqrt{s_{11}} \\ \sum_i (x_{i2} - \bar{x}_2)(y_i - \bar{y})/\sqrt{s_{22}} \end{bmatrix} \\ &= \begin{bmatrix} r_{1y}\sqrt{s_{yy}} \\ r_{2y}\sqrt{s_{yy}} \end{bmatrix},\end{aligned}$$

where $r_{1y} = \sum_i (x_{i1} - \bar{x}_1)(y_i - \bar{y})/\sqrt{s_{11}}\sqrt{s_{yy}}$ is the sample correlation between X_1 and Y, and similarly r_{2y} is the sample correlation between X_2 and Y. Then the standardized LSEs are

$$\begin{aligned}\underline{\hat{\alpha}} &= \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{v} \\ &= \frac{1}{1 - r_{12}^2}\begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}\begin{bmatrix} r_{1y}\sqrt{s_{yy}} \\ r_{2y}\sqrt{s_{yy}} \end{bmatrix} \\ &= \frac{1}{1 - r_{12}^2}\begin{bmatrix} r_{1y}\sqrt{s_{yy}} - r_{12}r_{2y}\sqrt{s_{yy}} \\ r_{2y}\sqrt{s_{yy}} - r_{12}r_{1y}\sqrt{s_{yy}} \end{bmatrix}.\end{aligned}$$

That is,

$$\hat{\alpha}_1 = \frac{r_{1y}\sqrt{s_{yy}} - r_{12}r_{2y}\sqrt{s_{yy}}}{1 - r_{12}^2} \tag{4.1.19}$$

and

$$\hat{\alpha}_2 = \frac{r_{2y}\sqrt{s_{yy}} - r_{12}r_{1y}\sqrt{s_{yy}}}{1 - r_{12}^2}. \tag{4.1.20}$$

The RSS can be written as,

$$\begin{aligned}&RSS(\alpha_1, \alpha_2) \\ &= \underline{\alpha}^T\left(\mathbf{Z}^T\mathbf{Z}\right)\underline{\alpha} - 2\underline{\alpha}^T\mathbf{Z}^T\mathbf{v} + \mathbf{v}^T\mathbf{v} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \end{bmatrix}\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} - 2\begin{bmatrix} \alpha_1 & \alpha_2 \end{bmatrix}\begin{bmatrix} r_{1y}\sqrt{s_{yy}} \\ r_{2y}\sqrt{s_{yy}} \end{bmatrix} + s_{yy} \\ &= \alpha_1^2 + 2r_{12}\alpha_1\alpha_2 + \alpha_2^2 - 2r_{1y}\sqrt{s_{yy}}\alpha_1 - 2r_{2y}\sqrt{s_{yy}}\alpha_2 + s_{yy}.\end{aligned} \tag{4.1.21}$$

This is a quadratic equation in three variables and defines a quadric surface in $\mathbb{R}^3$. We can graph this surface with α_1 on the x-axis, α_2 on the y-axis and $RSS(\alpha_1, \alpha_2)$ on the z-axis. To determine the shape of the surface we look at the intersections that the surface makes with planes that are parallel to the coordinate axes. These curves are called the traces of the surface and are quadratic equations in two

variables called conic sections. Section A.4 contains details about conic sections used here.

Any plane parallel to the xz-plane is described by the equation $\alpha_1 = k$, where k is a constant. Setting $\alpha_1 = k$ in (4.1.21), we obtain the vertical traces,

$$\alpha_2^2 + 2\alpha_2\left(kr_{12} - r_{2y}\sqrt{s_{yy}}\right) + \left(k^2 - 2kr_{1y}\sqrt{s_{yy}} + s_{yy}\right) - RSS = 0.$$

This is a quadratic equation in terms of α_2 and RSS. The equation can be written as $\underline{\theta}^T\mathbf{A}\underline{\theta}$, where $\underline{\theta}^T = (\alpha_2, RSS, 1)$ and by (A.4.2),

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & kr_{12} - r_{2y}\sqrt{s_{yy}} \\ 0 & 0 & -1/2 \\ kr_{12} - r_{2y}\sqrt{s_{yy}} & -1/2 & k^2 - 2kr_{1y}\sqrt{s_{yy}} + s_{yy} \end{bmatrix},$$

with $|\mathbf{A}| = -1/4 \neq 0$, so the conic section is non-degenerate for all α_2. From (A.4.3), the discriminant is given by

$$\Delta = \begin{vmatrix} 1 & 0 \\ 0 & 0 \end{vmatrix} = 0,$$

so these traces are parabolas for all α_2. Since $\alpha_2^2 > 0$, the parabolas open upwards and the turning point is a minimum. By setting $\alpha_2 = k$ in (4.1.21), similar conclusions can be drawn about the vertical traces which are parallel to the yz-plane - they are always parabolas which open upward.

Setting $RSS(\alpha_1, \alpha_2) = k$, we obtain the horizontal traces of the surface. These curves are also called the contour lines of the function, each one is a curve along which the function has a constant value k. The contours of RSS are given by the equation

$$\alpha_1^2 + 2\alpha_1\alpha_2 r_{12} + \alpha_2^2 - 2\alpha_1 r_{1y}\sqrt{s_{yy}} - 2\alpha_2 r_{2y}\sqrt{s_{yy}} + \left(s_{yy} - k\right) = 0.$$

The equation can be as $\underline{\theta}^T\mathbf{A}\underline{\theta}$, where $\underline{\theta}^T = (\alpha_2, \alpha_2, 1)$ and by (A.4.2),

$$\mathbf{A} = \begin{bmatrix} 1 & r_{12} & -r_{1y}\sqrt{s_{yy}} \\ r_{12} & 1 & -r_{2y}\sqrt{s_{yy}} \\ -r_{1y}\sqrt{s_{yy}} & -r_{2y}\sqrt{s_{yy}} & s_{yy} - k \end{bmatrix},$$

with

$$\begin{aligned}|\mathbf{A}| &= \left(s_{yy}-k\right)+2r_{12}r_{1y}r_{2y}s_{yy}-r_{1y}^2s_{yy}-r_{2y}^2s_{yy}-r_{12}^2\left(s_{yy}-k\right)\\ &= \left(s_{yy}-k\right)\left(1-r_{12}^2\right)-s_{yy}\left(r_{1y}^2+r_{2y}^2-2r_{12}r_{1y}r_{2y}\right).\end{aligned}$$

If $r_{12}=1$ then

$$\begin{aligned}|\mathbf{A}| &= -s_{yy}\left(r_{1y}^2+r_{2y}^2-2r_{1y}r_{2y}\right)\\ &= -s_{yy}\left(r_{1y}-r_{2y}\right)^2\\ &= 0,\end{aligned}$$

since $r_{1y}=r_{2y}$ when $r_{12}=1$. If $r_{12}=-1$ then

$$\begin{aligned}|\mathbf{A}| &= -s_{yy}\left(r_{1y}^2+r_{2y}^2+2r_{1y}r_{2y}\right)\\ &= -s_{yy}\left(r_{1y}+r_{2y}\right)^2\\ &= 0,\end{aligned}$$

since $r_{1y}=-r_{2y}$ when $r_{12}=-1$. Thus, the conic will be degenerate when X_1 and X_2 are perfectly correlated (positively or negatively), so that $|r_{12}|=1$. When $|r_{12}|\in[0,1)$, the conic is non-degenrate and from (A.4.3), the discriminant is given by

$$\Delta=\begin{vmatrix}1 & r_{12}\\ r_{12} & 1\end{vmatrix}=1-r_{12}^2=\left|\mathbf{Z}^T\mathbf{Z}\right|.$$

Since $|r_{12}|\in[0,1)$ we have that $\Delta>0$ so that the contours are ellipses. The minimum value is at the center of the ellipses and the function gains height as the ellipses get larger. The center of the contours is given by (A.4.4) and turn out to be the LSEs:

$$\begin{aligned}&\left(\frac{r_{1y}\sqrt{s_{yy}}-r_{12}r_{2y}\sqrt{s_{yy}}}{1-r_{12}^2},\frac{r_{2y}\sqrt{s_{yy}}-r_{12}r_{1y}\sqrt{s_{yy}}}{1-r_{12}^2}\right)\\ &=\left(\hat{\alpha}_1,\hat{\alpha}_2^2\right)\text{ from (4.1.19) and (4.1.20).}\end{aligned}$$

The angle that the axes of the contours makes with the coordinate axes is determined by (A.4.5),

$$\cot 2\vartheta=\frac{1-1}{2r_{12}}=0$$

An angle that satisfies $\cot 2\theta = 0$ is $2\theta = \pm\,\pi/2$ so that the contour axes are at $\theta = \pm\,\pi/4 = \pm 45°$ from the coordinate axes.

Since the surface has parabolic vertical traces and elliptical horizontal traces, $RSS(\alpha_1, \alpha_2)$ defined by the equation (4.1.21) describes an elliptic paraboloid in $\mathbb{R}^3$ with the minimum turning point at the least squares solution. Solving the constrained problem, we seek the minimum point on the RSS quadric surface that lies within the feasible region described by the constraint. Figure 4.1.3 shows the contours of RSS for a linear model with two predictors, where the correlation between the predictors is set to $r_{12} = 0.75$ in the left panels and $r_{12} = -0.75$ in the right panels. The contours in red are values of the constraint, $\|\underline{\alpha}\|^2$ for ridge regression and $\|\underline{\alpha}\|_1$ for the LASSO. The lines are drawn at levels of the constraint where the ℓ_q-norm of the constrained parameters is a fraction, in $\mathcal{F} = \{0.2, 0.3, \ldots, 0.9\}$, of the ℓ_q-norm of the LSEs. That is, for ridge regression the value of the constraint surface at each contour line is such that $\|\underline{\alpha}\|^2 / \|\hat{\underline{\alpha}}\|^2 \in \mathcal{F}$. Similarly for the LASSO, the contour lines represent values of the constraint at which $s = \|\underline{\alpha}\|_1 / \|\hat{\underline{\alpha}}\|_1 \in \mathcal{F}$. The black dot in the center is the minimum point of the RSS function, the LSEs. The red triangles on each contour line of the constraint indicate where the constrained estimate would lie if that value of the constraint is used to estimate the model. For the LASSO, the estimates corresponding to the fractions 0.2, 0.3 and 0.4 lie on the x-axis, effectively setting $\hat{\alpha}_2 = 0$.

The penalty functions of ridge regression and the LASSO are both comprised of some ℓ_q-norm. ℓ_q-norms can be generalized to form a functional space known as ℓ_q-space and when combined with a vector space forms a normed vector space. In $\mathbb{R}^p$, a norm ball, or ℓ_q-ball with radius r and center $\underline{c} \in \mathbb{R}^p$ is convex and is given by $\left\{\underline{a} \in \mathbb{R}^p \,\middle|\, \|\underline{a} - \underline{c}\|_q \leqslant r\right\}$ (see Boyd & Vandenberghe (2004:30-31)). The ridge and LASSO constraints are therefore norm balls with $\underline{c} = \underline{0}$ and radius $r = \sqrt{\tau}$ and $r = t$, respectively. The LASSO norm ball has sharp corners at each point $\left\{\underline{\alpha} \in \mathbb{R}^p \,\middle|\, \alpha_i = 0, \alpha_{j\neq 1} = |t|\right\}$. If RSS touches the LASSO at one of the sharp corners where $\alpha_i = 0$ then that estimate is set to zero. Since the ridge norm ball is curved, it is unlikely that any estimates will be set to zero. The concept illustrated in Figure 4.1.3 can therefore be generalized to higher dimensions. RSS is the quadric surface described by equation (4.1.18) and the constraint region is an ℓ_q-ball, both in $\mathbb{R}^p$. The ℓ_1 and ℓ_2 balls are shown in Figure 4.1.4 in $\mathbb{R}^2$ and $\mathbb{R}^3$.

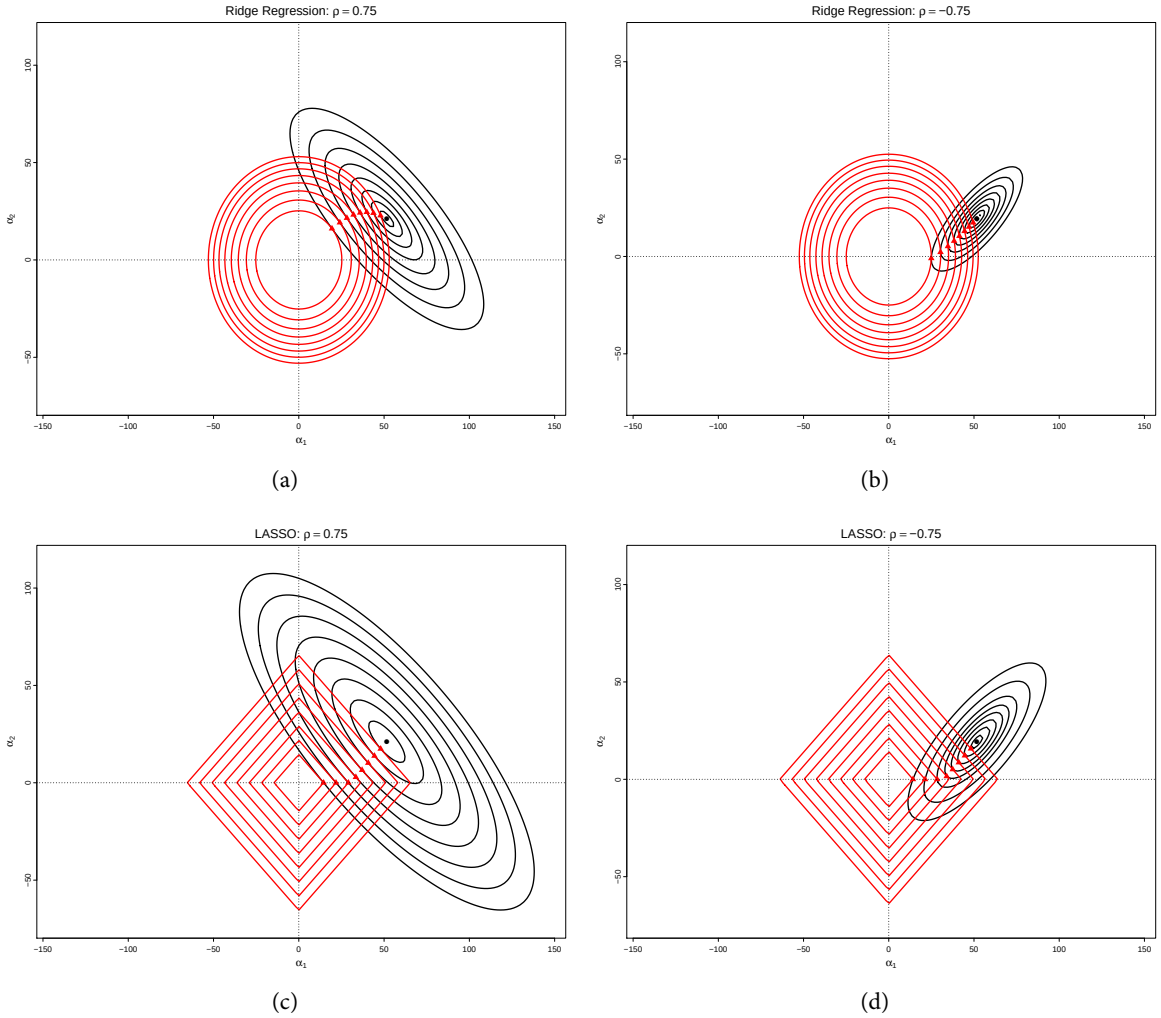


Figure (4.1.3) *RSS contours and constraint regions with $p = 2$ for (a)-(b) ridge regression and (c)-(d) LASSO with correlations of 0.75 (left) and -0.75 (right) between the two predictors. If t is chosen such that $s \leqslant 0.4$ then the minimum value of RSS that satisfies the constraint occurs at a corner of the constraint region so that $\hat{\alpha}_2^L = 0$.*

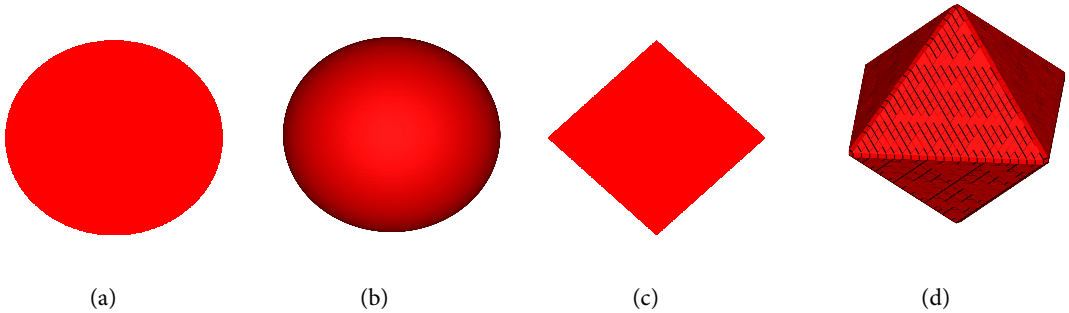


Figure (4.1.4) *Norm balls in $\mathbb{R}^2$ and $\mathbb{R}^3$ for (a)-(b) ridge regression and (c)-(d) LASSO. As the number of variables increases, the ℓ_1 norm ball has more sharp corners so that more estimates are likely to be set to zero.*

4.1.4 Computation

A number of algorithms now exist to solve the LASSO. Tibshirani (1996) proposed solving the problem using quadratic programming or the least squares with linear inequality constraints (LSI) algorithm. Efron *et al.* (2004) discovered that the path of the LASSO is piecewise linear and recognized a similarity between the least angle regression (LAR) algorithm and the LASSO. While quadratic programming requires operations of the order $O\left(n2^p\right)$, the LAR algorithm has the order of computation $O\left(np\min\left(n,p\right)\right)$. When $n > p$, the computational complexity is $O\left(np^2\right)$ which is the same as calculating the least squares estimate using the QR-decomposition (see Hastie *et al.* (2009:93)). Further computational efficiency is achieved by Friedman *et al.* (2007), who use the soft thresholding rule in a coordinate descent algorithm for the LASSO, requiring computations of order $O\left(np\right)$. These are the major advances for computing the LASSO and are discussed below.

Quadratic Programming and LSI

Tibshirani (1996) proposed two algorithms to solve the LASSO problem. The ℓ_1 norm of $\underline{\alpha}$ can be written in the form

$$\|\underline{\alpha}\|_1 = \sum_{j=1}^{p} \left|\alpha_j\right| = \sum_{j=1}^{p} \operatorname{sign}\left(\alpha_j\right)\alpha_j = \underline{\omega}^T\underline{\alpha},$$

where the elements of $\underline{\omega}$ are given by (4.1.4). The difficulty with this formulation is that $\underline{\omega}$ depends on the unknown parameters α_j. Alternatively, we could view the LASSO constraint as a system of inequalities considering all the possible signs of $\underline{\alpha}$. That is, $\omega\underline{\alpha} \leqslant t\mathbf{1}$, where the $o \times p$ matrix ω has rows in the form $\left(\pm1, \pm1, \ldots, \pm1\right)$. Then we can write the problem as

$$\underline{\hat{\alpha}}^L = \underset{\underline{\alpha}}{\arg\min} \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 \text{ subject to } \omega\underline{\alpha} \leqslant t\mathbf{1}. \tag{4.1.22}$$

This can easily be converted to the LSI problem which is stated as

$$\text{minimize } \|\mathbf{Z}\underline{\alpha} - \mathbf{v}\|^2 \text{ subject to } \mathbf{G}\underline{\alpha} \geqslant \mathbf{h},$$

where $\mathbf{G}$ is any $o \times p$ matrix and $\mathbf{h}$ is any $o \times 1$ vector. The problem is then converted into a form suitable for least distance programming (LDP) by using the SVD or QR-decomposition and is eventually solved by the nonnegative least squares (NNLS) algorithm, see Lawson & Hanson (1974:158-173) for details. However, this leads to $o = 2^p$ inequality constraints since the sign of each of the p parameters is either 1 or -1, so

the problem quickly becomes infeasible as p increases. Tibshirani (1996) noted that only some of the inequality constraints may be necessary and proposed an algorithm in which the constraints are used sequentially until convergence is reached. He uses the signs of the LSE as an initial ϖ and makes use of the LSI algorithm to compute the LASSO solution. While the solution is greater than t, the signs of the LASSO solution are added to ϖ and the process is continued until convergence. A drawback of this algorithm is that the LSE is required to get started.

The LASSO problem is also recognized as a quadratic programming problem. Since

$$RSS\left(\underline{\alpha}\right) = \left\|\mathbf{v} - \mathbf{Z}\underline{\alpha}\right\|^2 = \underline{\alpha}^T\mathbf{Z}^T\mathbf{Z}\underline{\alpha} - 2\mathbf{v}^T\mathbf{Z}\underline{\alpha} - \mathbf{v}^T\mathbf{v},$$

the problem is equivalent to the quadratic programming problem which is stated as

$$\text{minimize } \frac{1}{2}\underline{\alpha}^T\mathbf{Z}^T\mathbf{Z}\underline{\alpha} - \mathbf{v}^T\mathbf{Z}\underline{\alpha} - \mathbf{v}^T\mathbf{v} \text{ subject to } \mathbf{G}\underline{\alpha} \leqslant \mathbf{h},$$

where $\mathbf{G}$ is any $o \times p$ matrix, $\mathbf{h}$ is any $o \times 1$ vector and $\mathbf{Z}^T\mathbf{Z}$ is a symmetric positive semidefinite matrix (see Boyd & Vandenberghe (2004:152-153)). An alternative algorithm proposed by Tibshirani (1996) views the LASSO constraint via non-negative parameters and employs quadratic programming to solve the problem. Suppose that $\alpha_j = \alpha_j^+ - \alpha_j^-$, where

$$\alpha_j^+ = \begin{cases} \alpha_j & \text{if } \alpha_j > 0 \\ 0 & \text{if } \alpha_j \leqslant 0 \end{cases} \text{ and } \alpha_j^- = \begin{cases} 0 & \text{if } \alpha_j \geqslant 0 \\ -\alpha_j & \text{if } \alpha_j < 0 \end{cases}$$

for $j = 1, 2, \ldots, p$. These parameters are thus constrained to be non-negative, $\alpha_j^+ \geqslant 0$ and $\alpha_j^- \geqslant 0$ for $j = 1, 2, \ldots, p$ and the LASSO constraint becomes $\sum_{j=1}^p \alpha_j^+ + \sum_{j=1}^p \alpha_j^- \leqslant t$. The LASSO problem can then be written as

$$\hat{\underline{\alpha}}^* = \underset{\underline{\alpha}^*}{\arg\min} \left\|\mathbf{v} - \mathbf{Z}^*\underline{\alpha}^*\right\|^2 \text{ subject to } \mathbf{A}\underline{\alpha}^* \leqslant \mathbf{b},$$

where

$$\mathbf{Z}^* = \begin{bmatrix} \mathbf{Z} & -\mathbf{Z} \end{bmatrix}, \underline{\alpha}^* = \begin{bmatrix} \underline{\alpha}^+ \\ \underline{\alpha}^- \end{bmatrix}, \mathbf{A} = \begin{bmatrix} -\mathbf{I}_{2p} \\ \mathbf{1}_{2p}^T \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} \mathbf{0}_{2p} \\ t \end{bmatrix} \tag{4.1.23}$$

and $\hat{\alpha}_j^L = \hat{\alpha}_j^+ - \hat{\alpha}_j^-$. The problem then has $2p$ variables and $2p+1$ constraints and converges at a faster rate than the first algorithm. By solving the LASSO dual problem, Osborne *et al.* (2000b) developed a more efficient quadratic programming algorithm for computing the LASSO which can be used when $p > n$. Gong & Zhang (2011) use a projected Newton method to solve the dual problem.

Least Angle Regression

Efron *et al.* (2004) found that the solution path of the LASSO is piecewise linear and proposed a path following algorithm called LAR to find the entire path. When solving the LASSO problem, we usually compute an estimator $\hat{\underline{\alpha}}^L(\lambda)$ for many values of λ and perform a search for the best one. We know that λ attains its minimum at $\lambda_0 = 0$ and its maximum at $\lambda_{M-1} = \left\|\mathbf{Z}^T\mathbf{v}\right\|_\infty = \max_j \left|\mathbf{z}_j^T\mathbf{v}\right|$. Let $\lambda_k \in \{\lambda_0 < \lambda_1 < \cdots < \lambda_{M-1} < \lambda_M | \lambda_0 = 0, \lambda_M = \infty\}$ be the values of λ at which new predictors enter the model. Then the active set of estimates $\mathcal{A}_k = \left\{j \left| \hat{\alpha}_j(\lambda_k) \neq 0 \right.\right\}$ remains unchanged on each interval $[\lambda_k, \lambda_{k+1}]$ and it can be shown that

$$\hat{\underline{\alpha}}^L(\lambda) = \hat{\underline{\alpha}}^L(\lambda_k) + \underline{\eta}_k(\lambda - \lambda_k)$$

for $\lambda \in [\lambda_k, \lambda_{k+1}]$, $0 \leqslant k \leqslant M-1$ and $\underline{\eta}_k > 0$ (see equation (B.3.5)). Hence, $\hat{\underline{\alpha}}(\lambda)$ is linear on the interval $[\lambda_k, \lambda_{k+1}]$ and if we can identify the λ_k then we can compute the entire solution path using linear interpolation.

LAR proceeds in a fashion similar to forward selection (Section 2.2.3), beginning with the null model and adding one variable to the model at each step. However, LAR is less greedy and only adds a fraction of each coefficient to the model before moving on. It is based on examining the correlations between the predictors and the residual from the previous step. Beginning with the null model, the residual vector is the response. So at the first step, the variable most correlated with the response is added to the active set. The estimator is then moved in the direction of the least squares coefficient that is obtained when regressing the residuals on the active set. The estimator continues to move in that direction until another variable becomes equally correlated with the residual, that is, the residual bisects the angle between them. That variable is then added to the active set. The estimator moves in the direction of their joint least squares coefficient with the new residual until another variable becomes equally correlated with the residual. The residual is then the vector that has the smallest equal angle with all the predictors in the active set - this is where the names least angle regression comes from. The process continues in this way until the least squares model is reached in the final step. Thus, LAR is another method which performs shrinkage and selection. The formulas to update the estimates for each step are provided in the algorithm below.

Algorithm 4.1.1 ***Least angle regression***

Assume that $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_p$ *are linearly independent.*

1. *Initialize* $\hat{\underline{\alpha}} = \underline{0}$*, then the fitted model is* $\hat{\mu}_0 = \mathbf{0}$ *and the correlations between the predictors and the*

residual are $\hat{\underline{c}}_0 = \mathbf{Z}^T\mathbf{v}$.

2. *For* $k = 1, 2, \ldots, \min(n-1, p)$*:*

   (a) *Find the largest correlation,* $\hat{\rho}_k = \max_j \left|\hat{c}_{k,j}\right|$ *and determine the active set* $\mathcal{A}_k = \left\{j : \left|\hat{c}_{k,j}\right| = \hat{\rho}_k\right\}$.

   (b) *Determine the signs of the correlations* $s_j = \text{sign}\left(\hat{c}_{k,j}\right)$ *and form the matrix* $\mathbf{Z}_{\mathcal{A}_k}$ *with columns* $s_j\mathbf{z}_j$ *for* $j \in \mathcal{A}_k$.

   (c) *Calculate*

$$a_k = \left(\mathbf{1}_{|\mathcal{A}_k|}^T\left(\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k}\right)^{-1}\mathbf{1}_{|\mathcal{A}_k|}\right)^{-\frac{1}{2}} \text{ and } \underline{w}_{\mathcal{A}_k} = a_k\left(\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k}\right)^{-1}\mathbf{1}_{|\mathcal{A}_k|},$$

   Then

$$\mathbf{u}_{\mathcal{A}_k} = \mathbf{Z}_{\mathcal{A}_k}\underline{w}_{\mathcal{A}_k}$$

   is a unit vector that has equal angles with all the columns in $\mathbf{Z}_{\mathcal{A}_k}$ *such that* $\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{u}_{\mathcal{A}_k} = a_k\mathbf{1}_{|\mathcal{A}_k|}$ *and* $\left\|\mathbf{u}_{\mathcal{A}_k}\right\|^2 = 1$. *Let*

$$\underline{b} = \mathbf{Z}^T\mathbf{u}_{\mathcal{A}_k}.$$

   i. If $\text{rank}\left(\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k}\right) = |\mathcal{A}_k|$ *then* $\hat{\varsigma} = \hat{\rho}_k/a_k$

   ii. Else if $\text{rank}\left(\mathbf{Z}_{\mathcal{A}_k}^T\mathbf{Z}_{\mathcal{A}_k}\right) < |\mathcal{A}_k|$ *then*

$$\hat{\varsigma} = \min_{j\in\mathcal{A}_k^c}{}^{+}\left\{\frac{\hat{\rho}_k - \hat{c}_{k,j}}{a_k - b_j}, \frac{\hat{\rho}_k + \hat{c}_{k,j}}{a_k + b_j}\right\},$$

   where $\min^+$ *means that only the positive elements are considered for the minimum.*

   (d) *Update the estimates, fitted model and the correlations between the predictors and the residual,*

$$\hat{\underline{\alpha}}_{\mathcal{A}_k} = \hat{\underline{\alpha}}_{\mathcal{A}_{k-1}} + \hat{\varsigma}\underline{w}_{\mathcal{A}_k},$$

$$\hat{\mu}_k = \hat{\mu}_{k-1} + \hat{\varsigma}\mathbf{u}_{\mathcal{A}_k}$$

$$\hat{\underline{c}}_k = \mathbf{Z}^T\left(\mathbf{v} - \mathbf{Z}^T\hat{\underline{\alpha}}_{\mathcal{A}_k}\right) = \hat{\underline{c}}_{k-1} - \hat{\varsigma}\underline{b}.$$

Efron *et al.* (2004) shows that the LAR algorithm can be modified to fit forward stagewise regression (mentioned in Section 2.2.5) and the LASSO. A connection between LASSO and LAR can be made by looking at the KKT optimality conditions of the LASSO, see Section B.3. For the LASSO, equation (B.3.4) shows that

the active variables satisfy

$$\mathbf{z}_j^T (\mathbf{v} - \mathbf{Z}\underline{\alpha}) = \text{sign}(\alpha_j)\,\lambda.$$

Efron *et al.* (2004) show this with a geometrical approach. At any step of the LAR algorithm, the active variables have equal maximum correlation with the residuals. Denote the correlations at this step by

$$\mathbf{Z}^T (\mathbf{v} - \mathbf{Z}\underline{\alpha}) = \underline{c}$$

and the maximum correlation by

$$\max_j |c_j| = \rho,$$

then

$$\mathbf{z}_j^T (\mathbf{v} - \mathbf{Z}\underline{\alpha}) = \text{sign}(c_j)\,\rho \text{ for } j \in \mathcal{A}.$$

So if $\text{sign}(\alpha_j) = \text{sign}(c_j) = s_j$ then $\lambda = \rho$. To enforce this sign restriction in the LAR algorithm, a predictor can be removed from the active set if it changes sign. Efron *et al.* (2004) suggest the following modification to solve the LASSO problem:

Algorithm 4.1.2 ***LAR-LASSO***

1. *Apply LAR algorithm up to step 2c(ii).*
2. *Step 2c(iii) added to LAR: Calculate*

$$\varsigma_j = \frac{-\alpha_j}{s_j w_j} \text{ for } j \in \mathcal{A} \text{ and } \tilde{\varsigma} = \min_{j \in \mathcal{A}}{}^{+} \{\varsigma_j\}.$$

   (a) *If $\tilde{\varsigma}_j > \hat{\varsigma}_j$ then go to step 2d.*

   (b) *Else if $\tilde{\varsigma}_j < \hat{\varsigma}_j$ then remove j from the active set and compute the next direction without it,*

$$\underline{\hat{\alpha}}_{\mathcal{A}_k} = \underline{\hat{\alpha}}_{\mathcal{A}_{k-1}} + \tilde{\varsigma}\underline{w}_{\mathcal{A}_k},$$

$$\mathcal{A}_k = \mathcal{A}_{k-1} - j.$$

By inspection of Algorithm 4.1.1, we see that LAR is a descent algorithm to optimize $\hat{\mu} = \mathbf{Z}\underline{\hat{\alpha}}$ with the search direction determined by the unit equiangular vector $\mathbf{u}$ and the step length by $\hat{\varsigma}$. Efron *et al.*

(2004) explain that LAR requires $O\left(p^3 + np^2\right)$ computations when $p < n$, which is the same order of computation when fitting least squares via the Cholesky decomposition. The LASSO modification can cost up to an additional $O\left(p^2\right)$ operations for every variable that must be dropped. When $p \gg n$, LAR stops after $n-1$ variables are in the model ($n-1$ because of the centering) with a cost of $O\left(n^3\right)$. However, Rosset & Zhu (2007) generalize the LAR algorithm for LASSO (LAR-LASSO) algorithm for use with other "almost quadratic" loss functions and the ℓ_1 penalty. They state that the number of steps used on average is $M = O\left(\min\left(n, p\right)\right)$ so that the overall computational expense of the algorithm is $O\left(np\min\left(n, p\right)\right)$. A related method is presented by Osborne *et al.* (2000a), an homotopy algorithm which also follows the piecewise linear path of the LASSO (see Nocedal & Wright (1999:304-310) for information about homotopy methods). Their algorithm is closely related to the LAR-LASSO algorithm, although it is somewhat indirect and lacks the transparency of LAR.

Coordinate Descent

The coordinate descent algorithm can be described as the steepest descent algorithm in the ℓ_1 norm (see Boyd & Vandenberghe (2004:475-484), a summary of descent methods is provided in Definition A.3.5). The problem is given by

$$\text{minimize } l\left(\underline{\alpha}\right),$$

where $l\left(\underline{\alpha}\right)$ is convex and twice differentiable. The search direction is

$$\Delta\underline{\alpha} = -\frac{\partial l\left(\underline{\alpha}\right)}{\partial \alpha_j},$$

where

$$j = \arg\max_j \left|\frac{\partial l\left(\underline{\alpha}\right)}{\partial \alpha_j}\right|.$$

Thus, at each step of the algorithm we update only the coordinate of $\underline{\alpha}$ corresponding to the coordinate of the gradient for which $\left|\nabla l\left(\underline{\alpha}\right)_j\right| = \left\|\nabla l\left(\underline{\alpha}\right)\right\|_\infty$. Selectively updating the coordinates in this way is a greedy version of coordinate descent. An alternative is to cycle through the coordinates and update each of them in turn.

The problem is an unconstrained minimization so we can use coordinate descent to solve the La-

grangian problem (4.1.2), which is equivalent to minimizing

$$l\left(\underline{\alpha}\right) = \frac{1}{2}\left\|\mathbf{v}-\sum_j \alpha_j \mathbf{z}_j\right\|^2 + \lambda \sum_j \left|\alpha_j\right|. \tag{4.1.24}$$

Although $l\left(\underline{\alpha}\right)$ is convex, it is not differentiable. However, coordinate descent can still be applied to a function

$$f\left(\underline{\alpha}\right) = g\left(\underline{\alpha}\right) + \sum_j h_j\left(\alpha_j\right), \tag{4.1.25}$$

where $g\left(\underline{\alpha}\right)$ is convex and differentiable and $h_j\left(\alpha_j\right)$ are convex but not necessarily differentiable. This is true because the nondifferentiable part $\sum_j h_j\left(\alpha_j\right)$ is additively separable. The first order condition for convex functions is

$$f\left(\underline{\alpha}\right) - f\left(\underline{\hat{\alpha}}\right) \geqslant \nabla f\left(\underline{\hat{\alpha}}\right)^T \left(\underline{\alpha} - \underline{\hat{\alpha}}\right)$$

and $\underline{\hat{\alpha}}$ is optimal if $\nabla f\left(\underline{\hat{\alpha}}\right)^T \left(\underline{\alpha} - \underline{\hat{\alpha}}\right) = \underline{0}$. For f in the form (4.1.25), we have

$$\begin{aligned} f\left(\underline{\alpha}\right) - f\left(\underline{\hat{\alpha}}\right) &\geqslant \nabla g\left(\underline{\hat{\alpha}}\right)^T \left(\underline{\alpha} - \underline{\hat{\alpha}}\right) + \sum_j \left[h_j\left(\alpha_j\right) - h_j\left(\hat{\alpha}_j\right)\right] \\ &= \sum_j \left[\nabla g\left(\underline{\hat{\alpha}}\right)_j \left(\alpha_j - \hat{\alpha}_j\right) + h_j\left(\alpha_j\right) - h_j\left(\hat{\alpha}_j\right)\right] \\ &\geqslant \underline{0}. \end{aligned}$$

Convergence of coordinate descent algorithms applied to functions such as (4.1.25) is proved by Tseng (2001). Thus, we can apply coordinate descent to (4.1.24) since the LASSO penalty is separable, $\left\|\underline{\alpha}\right\|_1 = \sum_j \left|\alpha_j\right|$.

If we fit a univariate model, that is, $j = 1$ in (4.1.24), then

$$l\left(\alpha_j\right) = \frac{1}{2}\left\|\mathbf{v}-\alpha_j \mathbf{z}_j\right\|^2 + \lambda \left|\alpha_j\right|$$

and similarly to the orthogonal design (Section 4.1.2), a closed form solution can be obtained. When $\alpha_j > 0$ then

$$\begin{aligned} & l'\left(\alpha_j\right) = 0 \\ \Leftrightarrow \quad & -\mathbf{z}_j^T\left(\mathbf{v}-\alpha_j \mathbf{z}_j\right) + \lambda = 0 \\ \Leftrightarrow \quad & \alpha_j = \mathbf{z}_j^T \mathbf{v} - \lambda \text{ since } \mathbf{z}_j^T \mathbf{z}_j = 1. \end{aligned}$$

Since $\alpha_j > 0$, we must have $\mathbf{z}_j^T\mathbf{v} - \lambda > 0$ or $\mathbf{z}_j^T\mathbf{v} > \lambda$. Similarly, suppose $\alpha_j < 0$, then

$$\begin{aligned} & l'\left(\alpha_j\right) = 0 \\ \Leftrightarrow \quad & -\mathbf{z}_j^T\left(\mathbf{v} - \alpha_j\mathbf{z}_j\right) - \lambda = 0 \\ \Leftrightarrow \quad & \alpha_j = \mathbf{z}_j^T\mathbf{v} + \lambda \text{ since } \mathbf{z}_j^T\mathbf{z}_j = 1, \end{aligned}$$

but $\alpha_j < 0$, so we must have $\mathbf{z}_j^T\mathbf{v} - \lambda < 0$ or $\mathbf{z}_j^T\mathbf{v} < -\lambda$. The solution is therefore the soft thresholding rule

$$\hat{\alpha}_j^L = \text{sign}\left(\alpha_j\right)\left(\left|\mathbf{z}_j^T\mathbf{v}\right| - \lambda\right)_+ = S\left(\mathbf{z}_j^T\mathbf{v}, \lambda\right). \tag{4.1.26}$$

Note that $\mathbf{z}_j^T\mathbf{v}$ is the univariate LSE so that the solution (4.1.26) is identical to Equation (4.1.11). Hence, if we use this rule to update the coordinates then we are making the assumption that the predictors are orthogonal.

To use coordinate descent, we need to include all the predictors in the model but when we update the j-th coordinate, the other $k \neq j$ coordinates are held fixed. The objective function (4.1.24) is written as

$$l\left(\alpha_j\right) = \frac{1}{2}\left\|\mathbf{v} - \alpha_j\mathbf{z}_j - \mathbf{Z}_{-j}\underline{\alpha}_{-j}\right\|^2 + \lambda\left|\alpha_j\right| + \lambda\left\|\underline{\alpha}_{-j}\right\|_1, \tag{4.1.27}$$

where $\mathbf{Z}_{-j}$ is the predictor matrix excluding the j-th variable and $\underline{\alpha}_{-j}$ is the parameter vector excluding the j-th parameter and is held fixed. Then

$$\begin{aligned} l'\left(\alpha_j\right) &= -\mathbf{z}_j^T\left(\mathbf{v} - \alpha_j\mathbf{z}_j - \mathbf{Z}_{-j}\underline{\alpha}_{-j}\right) + \text{sign}\left(\alpha_j\right)\lambda \\ &= \alpha_j\mathbf{z}_j^T\mathbf{z}_j - \mathbf{z}_j^T\left(\mathbf{v} - \mathbf{Z}_{-j}\underline{\alpha}_{-j}\right) + \text{sign}\left(\alpha_j\right)\lambda \\ &= \alpha_j - \mathbf{z}_j^T\left(\mathbf{v} - \mathbf{Z}_{-j}\underline{\alpha}_{-j}\right) + \text{sign}\left(\alpha_j\right)\lambda \\ &= \alpha_j - \mathbf{z}_j^T\mathbf{r}_{-j} + \text{sign}\left(\alpha_j\right)\lambda, \end{aligned}$$

where $\mathbf{r}_{-j} = \mathbf{v} - \mathbf{Z}_{-j}\underline{\alpha}_{-j}$ is the residual vector excluding the j-th variable. By the same logic as above, the solution to (4.1.27) is therefore the soft thresholding rule,

$$\breve{\alpha}_j = \text{sign}\left(\alpha_j\right)\left(\left|\mathbf{z}_j^T\mathbf{r}_{-j}\right| - \lambda\right)_+ = S\left(\mathbf{z}_j^T\mathbf{r}_{-j}, \lambda\right).$$

Note that $\mathbf{z}_j^T\mathbf{r}_{-j}$ is the univariate LSE when we regress $\mathbf{r}_{-j}$ on $\mathbf{z}_j$.

Friedman *et al.* (2007) used this result to develop a pathwise cyclical coordinate descent algorithm for the LASSO. At each iteration, the algorithm cycles through the variables for $j = 1, 2, \ldots, p$, incrementing the j-th estimate by $\tilde{\alpha}_j$ while the others are held fixed. The algorithm is applied to a sequence of λ values, $\lambda_1 > \lambda_2 > \cdots > \lambda_M$, using each solution as a warm start for the next problem. The algorithm is outlined below:

Algorithm 4.1.3 ***Pathwise coordinate descent***

1. *Start with an initial estimate* $\underline{\breve{\alpha}}^{(0)}$*, possibly the univariate estimates (4.1.26).*
2. *For* $i = 1, 2, \ldots, M$*: Set* $k = 1$*. Until* $\|\mathbf{r}\|^2$ *converges, repeat:*
   (a) *For* $j = 1, 2, \ldots, p$*:*
      i. *Calculate* $\mathbf{r} = \mathbf{v} - \mathbf{Z}\underline{\breve{\alpha}}$
      ii. *Update* $\breve{\alpha}_j^{(k)} = S\left(\breve{\alpha}_j^{(k-1)} + \mathbf{z}_j^T\mathbf{r}, \lambda_i\right)$
      iii. $k = k + 1$
   (b) *On convergence,* $\underline{\hat{\alpha}}_i^L = \underline{\breve{\alpha}}$ *is the estimate corresponding to* λ_i*.Reset* $\underline{\breve{\alpha}}^{(0)} = \underline{\hat{\alpha}}_i^L$ *and go back to step 2.*

Coordinate descent methods provide a massive improvement in the computational efficiency of the LASSO. The idea was first suggested by Fu (1998) who called it a shooting algorithm. Daubechies (2004) revisited the problem with an algorithm known as iterative shrinkage thresholding algorithm (ISTA). Beck & Teboulle (2009) improves the convergence rate with the fast iterative shrinkage thresholding algorithm (FISTA). These algorithms did not receive much attention until the contribution by Friedman *et al.* (2007). He also derived the thresholding functions for using coordinate descent for the nonnegative garrote, elastic net, group LASSO and fused LASSO, among other methods not mentioned in this paper. Wu & Lange (2008) compare a greedy coordinate descent algorithm with the cyclic version, as well as with LAR-LASSO. Their numerical studies show that the cyclic version has faster convergence than the greedy version for squared error loss. They also suggest that both methods are more robust and faster than LAR-LASSO. She (2009) discuss a class of thresholding-based iterative selection procedures (TISP. Bradley *et al.* (2011) propose a parallel coordinate descent algorithm which they call ShotGun (as opposed to shooting).

Other algorithms

These are the major advances of computational efficiency in the history of the LASSO but there are many more algorithms available. Grandvalet (1998) use the EM algorithm. Zhao & Yu (2007) propose a boosting algorithm which approximates the path of the LASSO using forward and backward steps. A boosting algorithm was also proposed by Bühlmann & Yu (2006), who use information criteria, including AIC, BIC, final prediction error (FPE) and MDL, as a stopping criterion. Schmidt *et al.* (2007) propose two algorithms, one based on a smooth approximation of the LASSO penalty, the other a gradient projection method. Wang & Leng (2007) use a least squares approximation (LSA). Park & Hastie (2007) compute the LASSO path using the predictor-corrector convex optimization method. Schmidt *et al.* (2009) provide an empirical study comparing a number of algorithms, including cyclical coordinate descent, steepest descent, sub-gradient descent, EM algorithm, grafting, log-barrier method, interior-point method, sequential quadratic programming, some smooth approximation methods, projection methods and other descent methods. Furthermore, the LASSO is a special case of every shrinkage method so that algorithms developed to solve these methods may also be used to compute the LASSO estimate.

4.1.5 Properties of LASSO Estimates

Standard Errors

Unfortunately the LASSO does not have an explicit solution like ridge regression since the LASSO penalty is non-differentiable at zero. A consequence is that the standard errors of the estimates and the predictions are not readily obtainable. One approach is to estimated the standard errors using the bootstrap by selecting the best tuning parameter for each bootstrap sample. Tibshirani (1996) states that holding t fixed during bootstrapping is equivalent to the subset selection situation where the best subset is first selected and then least squares standard errors for that subset are used.

Tibshirani (1996) also proposed an approximation formula for standard errors based on ridge regression. The LASSO penalty can be written $\sum |\alpha_j| = \sum \alpha_j^2 \big/ |\alpha_j| = \underline{\alpha}^T \mathbf{W}^- \underline{\alpha}$, where $\mathbf{W} = \text{diag}\left(|\alpha_j|\right)$. Now,

$$\mathbf{W}^- \underline{\alpha} = \alpha_j \big/ |\alpha_j| = \underline{\omega} = \begin{cases} 1 & \text{if } \alpha_j > 0 \\ 0 & \text{if } \alpha_j = 0 \\ -1 & \text{if } \alpha_j < 0 \end{cases}$$

from (4.1.4). So, the subdifferential of the LASSO Lagrangian (4.1.2) is given by

$$\begin{aligned} & -\mathbf{Z}^T\left(\mathbf{v}-\mathbf{Z}\underline{\alpha}\right)+\lambda\underline{\varpi} \\ &= -\mathbf{Z}^T\left(\mathbf{v}-\mathbf{Z}\underline{\alpha}\right)+\lambda\mathbf{W}^-\underline{\alpha}. \end{aligned} \tag{4.1.28}$$

Thus, if $\underline{\breve{\alpha}}$ is a solution to (4.1.2), then we must have

$$\begin{aligned} & \underline{0} = -\mathbf{Z}^T\mathbf{v}+\mathbf{Z}^T\mathbf{Z}\underline{\breve{\alpha}}+\lambda\mathbf{W}^-\underline{\breve{\alpha}} \\ \Leftrightarrow\quad & \left(\mathbf{Z}^T\mathbf{Z}+\lambda\mathbf{W}^-\right)\underline{\breve{\alpha}} = \mathbf{Z}^T\mathbf{v} \\ \Leftrightarrow\quad & \underline{\breve{\alpha}} = \left(\mathbf{Z}^T\mathbf{Z}+\lambda\mathbf{W}^-\right)^{-1}\mathbf{Z}^T\mathbf{v}, \end{aligned} \tag{4.1.29}$$

and the covariance matrix can be approximated by

$$\operatorname{var}\left(\underline{\breve{\alpha}}\right) \approx \hat{\sigma}^2\left(\mathbf{Z}^T\mathbf{Z}+\lambda\mathbf{W}^-\right)^{-1}\mathbf{Z}^T\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}+\lambda\mathbf{W}^-\right)^{-1}. \tag{4.1.30}$$

However, zero estimates will have approximated variance of zero.

Osborne *et al.* (2000b) provide an improved approximation for which zero estimates have positive standard errors. As above, the subdifferential of the LASSO Lagrangian (4.1.2) is given by (4.1.28). Thus, if $\underline{\breve{\alpha}}$ is a solution to (4.1.2), then we must have

$$\underline{0} = -\mathbf{Z}^T\breve{\mathbf{r}}+\lambda\underline{\varpi}, \tag{4.1.31}$$

where $\breve{\mathbf{r}} = \mathbf{v}-\mathbf{Z}\underline{\breve{\alpha}}$. Now $\underline{\varpi}^T\underline{\breve{\alpha}} = \left\|\underline{\breve{\alpha}}\right\|_1$, so that

$$\lambda = \frac{\breve{\mathbf{r}}^T\mathbf{Z}\underline{\breve{\alpha}}}{\left\|\underline{\breve{\alpha}}\right\|_1} \tag{4.1.32}$$

satisfies (4.1.31),

$$\begin{aligned} & \lambda\underline{\varpi} = \mathbf{Z}^T\breve{\mathbf{r}} \\ \Leftrightarrow\quad & \lambda\underline{\varpi}^T\underline{\breve{\alpha}} = \left(\mathbf{Z}^T\breve{\mathbf{r}}\right)^T\underline{\breve{\alpha}} \\ \Leftrightarrow\quad & \lambda = \breve{\mathbf{r}}^T\mathbf{Z}\underline{\breve{\alpha}}/\left\|\underline{\breve{\alpha}}\right\|_1. \end{aligned}$$

Also $\|\underline{\omega}\|_\infty = 1$, so that (4.1.31) yields

$$\begin{aligned} & \lambda\underline{\omega} = \mathbf{Z}^T\check{\mathbf{r}} \\ \Leftrightarrow \quad & \|\lambda\underline{\omega}\|_\infty = \left\|\mathbf{Z}^T\check{\mathbf{r}}\right\|_\infty \\ \Leftrightarrow \quad & \lambda\,\|\underline{\omega}\|_\infty = \left\|\mathbf{Z}^T\check{\mathbf{r}}\right\|_\infty \\ \Leftrightarrow \quad & \lambda = \left\|\mathbf{Z}^T\check{\mathbf{r}}\right\|_\infty, \end{aligned}$$

and consequently, we can write

$$\underline{\omega} = \frac{\mathbf{Z}^T\check{\mathbf{r}}}{\|\mathbf{Z}^T\check{\mathbf{r}}\|_\infty}. \tag{4.1.33}$$

Again, from (4.1.31) we have

$$\begin{aligned} & \underline{0} = -\mathbf{Z}^T\mathbf{v} + \mathbf{Z}^T\mathbf{Z}\check{\underline{\alpha}} + \lambda\underline{\omega} \\ \Leftrightarrow \quad & \mathbf{Z}^T\mathbf{v} = \mathbf{Z}^T\mathbf{Z}\check{\underline{\alpha}} + \lambda\underline{\omega} \\ \Leftrightarrow \quad & \mathbf{Z}^T\mathbf{v} = \mathbf{Z}^T\mathbf{Z}\check{\underline{\alpha}} + \frac{\mathbf{Z}^T\left(\check{\mathbf{r}}\check{\mathbf{r}}^T\right)\mathbf{Z}\check{\underline{\alpha}}}{\|\mathbf{Z}^T\check{\mathbf{r}}\|_\infty\,\|\check{\underline{\alpha}}\|_1} \text{ from (4.1.32) and (4.1.33)} \\ \Leftrightarrow \quad & \mathbf{Z}^T\mathbf{v} = \left(\mathbf{Z}^T\mathbf{Z} + \mathbf{W}\right)\check{\underline{\alpha}} \\ \Leftrightarrow \quad & \check{\underline{\alpha}} = \left(\mathbf{Z}^T\mathbf{Z} + \mathbf{W}\right)^{-1}\mathbf{Z}^T\mathbf{v}, \end{aligned}$$

where $\mathbf{W} = \dfrac{\mathbf{Z}^T\left(\check{\mathbf{r}}\check{\mathbf{r}}^T\right)\mathbf{Z}}{\|\mathbf{Z}^T\check{\mathbf{r}}\|_\infty\,\|\check{\underline{\alpha}}\|_1}$ and $\text{rank}\left(\mathbf{W}\right) = 1$. Note that,

$$\mathbf{Z}^T\mathbf{Z} + \mathbf{W} = \mathbf{Z}^T\left(\mathbf{I}_n + \frac{\left(\check{\mathbf{r}}\check{\mathbf{r}}^T\right)}{\|\mathbf{Z}^T\check{\mathbf{r}}\|_\infty\,\|\check{\underline{\alpha}}\|_1}\right)\mathbf{Z}$$

so that $\text{rank}\left(\mathbf{Z}^T\mathbf{Z} + \mathbf{W}\right) = \text{rank}\left(\mathbf{Z}\right) = \text{rank}\left(\mathbf{Z}^T\mathbf{Z}\right)$. Therefore, if $\mathbf{Z}^T\mathbf{Z}$ has full rank then $\left(\mathbf{Z}^T\mathbf{Z} + \mathbf{W}\right)$ is invertible. The variance can be approximated by

$$\text{var}\left(\check{\underline{\alpha}}\right) \approx \hat{\sigma}^2\left(\mathbf{Z}^T\mathbf{Z} + \mathbf{W}\right)^{-1}\mathbf{Z}^T\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \mathbf{W}\right)^{-1}. \tag{4.1.34}$$

Although their approximation might be an improvement over the one by Tibshirani (1996), they note that the estimates may be far from Gaussian so that standard errors are perhaps not an appropriate measure of uncertainty.

However, Lockhart *et al.* (2014) have succeeded in developing a significance test for coefficients as

they enter the LASSO path. Consider using the LAR-LASSO algorithm. At the beginning of the k-th step, the active set is given by $\mathcal{A}_k$. Suppose that the j-th variable enters next at λ_{k+1} so that

$$\underline{\hat{\alpha}}^L(\lambda_{k+1}) = \underline{\hat{\alpha}}^L(\lambda_k) + \hat{\varsigma}\underline{w}_{\mathcal{A}_{k+1}}$$

where $\mathcal{A}_{k+1} = \mathcal{A}_k \cup \{j\}$. Let, $\underline{\hat{\alpha}}^L_{\mathcal{A}_k}(\lambda_{k+1})$ be the solution at λ_{k+1} using only the active set $\mathcal{A}_k$,

$$\underline{\hat{\alpha}}^L_{\mathcal{A}_k}(\lambda_{k+1}) = \underset{\underline{\alpha}_{\mathcal{A}_k}}{\arg\min} \left\|\mathbf{v} - \mathbf{Z}_{\mathcal{A}_k}\underline{\alpha}_{\mathcal{A}_k}\right\|^2 + \lambda_{k+1}\left\|\underline{\alpha}_{\mathcal{A}_k}\right\|_1.$$

The full effect of the j-th variable on the fit is therefore given by

$$\mathbf{Z}\underline{\hat{\alpha}}^L(\lambda_{k+1}) - \mathbf{Z}_{\mathcal{A}_k}\underline{\hat{\alpha}}^L_{\mathcal{A}_k}(\lambda_{k+1}).$$

To assess the importance of the j-th variable in the model $\mathcal{A}_{k+1}$, they define the covariance test statistic

$$T_k = \left(\mathbf{v}^T\mathbf{Z}\underline{\hat{\alpha}}^L(\lambda_{k+1}) - \mathbf{v}^T\mathbf{Z}_{\mathcal{A}_k}\underline{\hat{\alpha}}^L_{\mathcal{A}_k}(\lambda_{k+1})\right)/\sigma^2$$

and they show that, under the null hypothesis that all the relevant variable are in the model

$$H_0 : \mathcal{A}_k \supseteq \mathcal{D},$$

the distribution of T_k is asymptotically the standard exponential distribution,

$$T_k \xrightarrow{d} Exp(1).$$

Hence, based on this distribution, a p-value can be calculated at each step of the algorithm. These p-values can then be used to decide when to stop adding variables to the model. That is, the hypothesis tests relate only to a step in the path of the LASSO.

Near Minimax Optimality

In the orthogonal design, Donoho & Johnstone (1994) proved the near-minimax optimality of the LASSO. Section 4.1.2 revealed that the loss function is given by $\left\|\underline{\hat{\alpha}} - \underline{\alpha}\right\|^2$ for orthogonal designs, so that the risk

function is given by

$$R\left(\underline{\hat{\alpha}}, \underline{\alpha}\right) = E\left\|\underline{\hat{\alpha}} - \underline{\alpha}\right\|^2,$$

which we recognize as the MSE. Consider

$$\hat{\alpha}_j = \alpha_j + \sigma\epsilon_j \text{ with } \epsilon_j \sim N\left(0,1\right),$$

note that $\text{var}\left(\hat{\alpha}_j\right) = \sigma^2$ for orthogonal designs. Suppose that we select a subset of the $\hat{\alpha}_j$ using a diagonal linear projection $\mu\hat{\alpha}_j$ with $\mu \in \{0,1\}$. If $\mu = 0$ then the risk is α_j^2 and if $\mu = 1$ then the risk is σ^2. Hence, the ideal risk is $R\left(\text{ideal}\right) = \min\left(\alpha_j^2, \sigma^2\right)$ so that the ideal projection includes those α_j which are greater than the noise level, $\mu = \delta\left(\left|a_j\right| > \sigma\right)$. Donoho & Johnstone (1994) introduced the universal threshold $\lambda^* = \sigma\sqrt{2\ln\left(n\right)}$ and showed that the soft thresholding function $\hat{\mu}_S^* = \text{sign}\left(\hat{\alpha}_j\right)\left(\left|\hat{\alpha}_j\right| - \lambda\right)_+$ yields the risk

$$R_S\left(\underline{\hat{\alpha}}^*, \underline{\alpha}\right) \leqslant 2\ln p\left(R\left(\text{ideal}\right) + \sigma^2\right).$$

They show that

$$\inf_{\hat{\alpha}} \sup \frac{R_S\left(\underline{\hat{\alpha}}^*, \underline{\alpha}\right)}{1 + R\left(\text{ideal}\right)} \sim 2\ln\left(n\right)$$

so that the estimator is near minimax optimal because the ideal risk is achieved up to a factor of at most $2\ln\left(n\right)$, a sharp minimax bound. They also showed that the hard thresholding function $\hat{\mu}_H^* = \hat{\alpha}_j\delta\left(\left|\hat{\alpha}_j\right| > \sigma\lambda_l\right)$, yields the risk

$$R_H\left(\underline{\hat{\alpha}}^*, \underline{\alpha}\right) \leqslant \lambda_l\left(R\left(\text{ideal}\right) + \sigma^2\right)$$

where $\left(1 - l\right)\ln\left(\ln\left(p\right)\right) \leqslant \lambda_l^2 - 2\ln\left(p\right) \leqslant o\left(\ln\left(p\right)\right)$ for some $l > 0$. This shows that the LASSO and subset selection have the same asymptotic performance for orthogonal designs.

Persistence

Greenshtein & Ritov (2004) showed that, in high dimensional settings $p \gg n$, in particular $p\left(n\right) = O\left(n^a\right)$ with $a > 1$ so that p increases with n, the LASSO is consistent for prediction. Consider the predic-

tion error

$$\begin{aligned} PE_F\left(f\left(X,\underline{\beta}\right)\right) &= E_F\left(y - f\left(X;\underline{\beta}\right)\right)^2 \\ &= E_F\left(y - \textstyle\sum_{j=1}^{p} X_j\beta_j\right)^2 \\ &= \underline{\nu}^T\Sigma_F\underline{\nu}, \end{aligned}$$

where X is random, F is the distribution of (X, Y), $\underline{\nu}^T = \left(-1, \underline{\beta}\right)$ and Σ_F has typical element $E_F\left(X_jX_k\right)$ for $j, k = 0, 1, \ldots, p$ with $X_0 = Y$. Suppose that $f\left(X, \underline{\beta}^*\right)$ is the ideal predictor,

$$\underline{\beta}^* = \arg\min_{\underline{\beta}} PE_F\left(f\left(X,\underline{\beta}\right)\right).$$

Let $\hat{F}$ be the empirical distribution of (X, Y) and

$$\begin{aligned} PE_{\hat{F}}\left(f\left(X,\underline{\beta}\right)\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f\left(X_i;\underline{\beta}\right)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \textstyle\sum_{j=1}^{p} X_{ij}\beta_j\right)^2 \\ &= \underline{\nu}^T\Sigma_{\hat{F}}\underline{\nu}, \end{aligned}$$

where $\Sigma_{\hat{F}}$ has typical element $\sum_{i=1}^{n} X_{ij}X_{ik}/n$ for $j, k = 0, 1, \ldots, p$. They show that, if one of the following regularity conditions hold:

1. The variances of X_iX_j are bounded and their moment generating functions have bounded third order derivatives in the neighbourhood of 0, or
2. $E_F\left(Y^2\right) < C$ and $P\left(\left|X_j\right| < L\right) = 1$ for $j = 1, 2, \ldots, p$ for finite constants C and L.

Then,

$$\underline{\hat{\beta}}_n = \arg\min_{\left\{\underline{\beta}\middle|\left\|\underline{\beta}\right\|_1 \le t(n)\right\}} PE_{\hat{F}}\left(f\left(X,\underline{\beta}\right)\right)$$

with

$$t(n) = o\left(\left(n/\ln n\right)^{1/4}\right)$$

is a persistent sequence of procedures, such that

$$PE_{\hat{F}}\left(f\left(X,\underline{\hat{\beta}}_n\right)\right) - PE_F\left(f\left(X,\underline{\beta}^*\right)\right) \xrightarrow{p} 0 \qquad (4.1.35)$$

They call this property persistence. If F is Gaussian, they relax the order of the constraint to $t(n) = o\left(\sqrt{n/\ln n}\right)$. In practice, the persistence rate $t(n)$ is not known and they suggest testing estimators resulting from different constraints on a test set. A result of persistence is that if the condition on $t(n)$ holds, then there is no harm in using the LASSO to search through the entire set of predictors. That is, prior screening of a smaller subset of variables will not significantly improve the LASSO, it is an effective method to find the optimal predictors in high dimensions. Under additional assumptions, they show that procedures like subset selection, which select a subset of size $k(n)$ variables are persistent if $k(n) = o\left(\sqrt{n/\ln}\right)$. Greenshtein (2006) extend the results to more general loss functions. They show that, under the assumptions:

1. $f\left(X, \underline{\beta}\right) = \left(y - \sum_{j=1}^{p} X_j \beta_j\right)$ is bounded and uniformly continuous in $\sum_{j=1}^{p} X_j \beta_j$, uniformly in y,
2. X_j is bounded, $\left|X_j\right| < M$ for $j = 1, 2, \ldots, p$ for a finite constants M,
3. $\underline{\beta}^*$ has $k(n) = o(n/\ln n)$ non-zero elements (sparsity rate), and
4. $\left\|\underline{\beta}^*\right\|$ is bounded,

then

$$\underline{\hat{\beta}}_n = \underset{\left\{\underline{\beta} \middle\| \underline{\beta}\right\|_1 \leq \sqrt{k(n)}\}}{\arg\min} PE_{\hat{F}}\left(f\left(X, \underline{\beta}\right)\right) \tag{4.1.36}$$

is a persistent sequence of procedures, such that (4.1.35) holds. Furthermore, (4.1.36) is persistent even without the assumption on the sparsity rate - a property they call self consistency. They demonstrate that ridge regression, or any bridge estimates with $q > 1$, are not persistent and that there is not much improvement when using bridge estimates with $q \in [0, 1)$. Further studies on the persistence of the LASSO are provided by Bunea *et al.* (2007) and Bartlett *et al.* (2012).

Bühlmann & van de Geer (2011:13-14,23-24,101-108) give a similar result for fixed designs when $p \gg n$, assuming that the model is exactly the linear model with true parameter vector $\underline{\alpha}$,

$$\mathbf{v} = \mathbf{Z}\underline{\alpha} + \varepsilon,$$

and the errors are Gaussian

$$\varepsilon \sim N\left(\mathbf{0}, \sigma^2 \mathbf{I}\right).$$

In a finite sense, they prove that if

$$\lambda = \sqrt{2\hat{\sigma}^2 \left(t^2 + 2\ln p\right)/n},$$

for some estimate of σ^2 and $t > 0$, then

$$2\left\|\mathbf{Z}\left(\underline{\hat{\alpha}}^L - \underline{\alpha}\right)\right\|^2 \Big/ n \leqslant 3\lambda \left\|\underline{\alpha}\right\|_1$$

with probability

$$1 - 2\exp\left(-t^2/2\right) - P\left(\hat{\sigma} \leqslant \sigma\right).$$

Asymptotically, under the assumptions on sparsity,

$$\left\|\underline{\alpha}\right\|_1 = o\left(\sqrt{n/\ln p}\right) \tag{4.1.37}$$

and the shrinkage parameter,

$$\lambda = \lambda_n \asymp \sqrt{\log p/n}, \tag{4.1.38}$$

it holds that the LASSO is consistent for prediction,

$$MSE\left[f\left(Z, \underline{\hat{\alpha}}_n^L\right)\right] = o_p\left(1\right) \text{ as } n \to \infty,$$

where

$$\begin{aligned} MSE\left[f\left(Z, \underline{\hat{\alpha}}_n^L\right)\right] &= \left(\underline{\hat{\alpha}}_n^L - \underline{\alpha}\right)^T \Sigma \left(\underline{\hat{\alpha}}_n^L - \underline{\alpha}\right) \\ &= \left\|\mathbf{Z}\left(\underline{\hat{\alpha}}_n^L - \underline{\alpha}\right)\right\|^2 \Big/ n, \end{aligned}$$

with $\Sigma = \mathbf{Z}^T\mathbf{Z}/n$. Although consistency is established, they show that the rate of convergence is slow,

$$\left\|\mathbf{Z}\left(\underline{\hat{\alpha}}_n^L - \underline{\alpha}\right)\right\|^2 \Big/ n = O_p\left(\left\|\underline{\alpha}\right\|_1 \sqrt{\ln p/n}\right), \tag{4.1.39}$$

so that prediction consistency is attained if $\left\|\underline{\alpha}\right\|_1 \ll \sqrt{n/\ln p}$. An oracle inequality improves the convergence rate considerably. An additional assumption is necessary, called the compatibility condition. Let

$\mathcal{D} = \left\{ j | \alpha_j \neq 0 \right\}$ be the true active set of variables of size $|\mathcal{D}| = d$. The compatibility constant is given by

$$v_c^2 (M, \mathcal{D}) = \min \left\{ \frac{d \underline{\alpha}^T \Sigma \underline{\alpha}}{\|\underline{\alpha}_{\mathcal{D}}\|_1^2} \middle| \|\underline{\alpha}_{\mathcal{D}^c}\|_1 \leqslant M \|\underline{\alpha}_{\mathcal{D}}\|_1 \right\}, \tag{4.1.40}$$

where $M > 1$. For λ as above we require $M = 3$ and λ must be adjusted accordingly for any change of M. The compatibility condition is satisfied for the set $\mathcal{D}$ if $v_c (M, \mathcal{D}) > 0$. In practice, $\mathcal{D}$ is not known but if its size d is known then the condition can be checked for all subsets $\mathcal{S} \subset \{1, 2, \ldots, p\}$ with $|\mathcal{S}| = d$. Under the assumptions the compatibility condition, an oracle inequality is given by

$$\left\| \mathbf{Z} \left(\hat{\underline{\alpha}}^L - \underline{\alpha} \right) \right\|^2 \Big/ n + \lambda \left\| \hat{\underline{\alpha}}^L - \underline{\alpha} \right\|_1 \leqslant 4\lambda^2 d / v_c^2. \tag{4.1.41}$$

The inequality shows the bound on prediction error,

$$\left\| \mathbf{Z} \left(\hat{\underline{\alpha}}^L - \underline{\alpha} \right) \right\|^2 \Big/ n \leqslant 4\lambda^2 d / v_c^2,$$

and asymptotically gives the convergence rate

$$\left\| \mathbf{Z} \left(\hat{\underline{\alpha}}_n^L - \underline{\alpha} \right) \right\|^2 \Big/ n = O_p \left(v_c^{-2} d \ln p / n \right), \tag{4.1.42}$$

which is optimal up to the $\ln p$ factor (and the compatibility constant v_c^{-2}) since using least squares with the correct subset would have the rate $O(d/n)$. They remark that the situation can be generalized for non-Gaussian errors and extend the results to the case when the true model is not exactly linear (pages 108-114) and the case when the predictors are random (pages 150-156).

Estimation Consistency

Knight & Fu (2000) study the asymptotic properties of bridges estimates for fixed designs where p does not vary with n, under the mild regularity conditions,

$$\Sigma_n = \frac{1}{n} \sum_{i=1}^{n} \underline{x}_i \underline{x}_i^T \to \Sigma \text{ and } \frac{1}{n} \max_{i \leqslant i \leqslant n} \underline{x}_i^T \underline{x}_i \to 0. \tag{4.1.43}$$

They showed that, for the LASSO,

1. If Σ is nonsingular and $\lambda_n / n \to \lambda_0 \geqslant 0$ (i.e. $\lambda_n = o(n)$), then $\hat{\underline{\alpha}}^L$ is consistent, $\hat{\underline{\alpha}}_n^L \xrightarrow{p} \arg\min V_1 (\underline{u})$,

where

$$V_1(\underline{u}) = \left[(\underline{u}-\underline{\alpha})^T \Sigma (\underline{u}-\underline{\alpha}) + \lambda_0 \|\underline{u}\|_1\right].$$

2. If Σ is nonsingular and $\lambda_n/\sqrt{n} \to \lambda_0 \geqslant 0$ (i.e. $\lambda_n = O(\sqrt{n})$), then $\hat{\underline{\alpha}}^L$ is asymptotically normal, $\sqrt{n}\left(\hat{\underline{\alpha}}_n^L - \underline{\alpha}\right) \xrightarrow{d} \arg\min V_2(\underline{u})$, where

$$V_2(\underline{u}) = -2\underline{u}^T\mathbf{W} + \underline{u}^T\Sigma\underline{u} + \lambda_0 \sum_{j=1}^{p}\left[u_j \operatorname{sign}(\alpha_j)\,\delta(\alpha_j \neq 0) + |u_j|\,\delta(\alpha_j = 0)\right]$$

with $\mathbf{W} \sim N(\mathbf{0}, \sigma^2\Sigma)$.

3. Suppose Σ_n is nearly singular but tends to a singular matrix Σ, and assume that $a_n(\Sigma_n - \Sigma) \to \mathbf{D}$, where $\mathbf{D}$ is nonsingular and a_n is a sequence tending to ∞. If $\lambda_n/\sqrt{n/a_n} \to \lambda_0 \geqslant 0$, (i.e. $\lambda_n = o\sqrt{n}$) then $\left(\sqrt{n/a_n}\right)\left(\hat{\underline{\alpha}}_n^L - \underline{\alpha}\right) \xrightarrow{d} \arg\min\{V_3(\underline{u}) | \Sigma\underline{u} = \underline{0}\}$, where

$$V_3(\underline{u}) = -2\underline{u}^T\mathbf{W} + \underline{u}^T\mathbf{D}\underline{u} + \lambda_0 \sum_{j=1}^{p}\left[u_j \operatorname{sign}(\alpha_j)\,\delta(\alpha_j \neq 0) + |u_j|\,\delta(\alpha_j = 0)\right]$$

with $\mathbf{W} \sim N(\mathbf{0}, \operatorname{var}(\mathbf{W}))$ and $\operatorname{var}(\mathbf{W})$ is such that $\operatorname{var}\left(\underline{u}^T\mathbf{W}\right) = \underline{u}^T\mathbf{D}\underline{u} > 0$ for all $\underline{u} > 0$ which satisfies $\Sigma\underline{u} = \underline{0}$. Thus the rate of convergence is slower than when Σ is nonsingular

Bühlmann & van de Geer (2011:14-17,135-137) show that the LASSO is consistent for estimation in the high dimensional setting with $p \gg n$. Under the assumptions above on sparsity (4.1.37), the shrinkage parameter (4.1.38) and the compatibility condition (4.1.40),

$$\left\|\hat{\underline{\alpha}}^L - \underline{\alpha}\right\|_1 \xrightarrow{p} 0,$$

and following from (4.1.41), a bound on the ℓ_1 error is

$$\left\|\hat{\underline{\alpha}}^L - \underline{\alpha}\right\|_1 \leqslant 4\lambda d/v_c^2$$

so that the rate of convergence is

$$\left\|\hat{\underline{\alpha}}^L - \underline{\alpha}\right\|_1 = O_p\left(v_c^{-2} d\sqrt{\ln p/n}\right).$$

For the usual ℓ_2 error, a stronger assumption is necessary, called the restricted eigenvalue condition. Let $\mathcal{U}$ be any set such that $\mathcal{D} \subset \mathcal{U}$ with $|\mathcal{U}| = u \geqslant d$.The set $\mathcal{U} \setminus \mathcal{D} = \mathcal{U} \cap \mathcal{D}^c$ is the relative complement of $\mathcal{U}$ with respect to $\mathcal{D}$, that is, the elements that are in the set $\mathcal{U}$ but not in the set $\mathcal{D}$. The restricted eigenvalue is given by

$$v\left(M, \mathcal{D}, u\right) = \min\left\{ \frac{\underline{\alpha}^T \Sigma \underline{\alpha}}{\|\underline{\alpha}_{\mathcal{U}}\|_2^2} \middle| \underline{\alpha} \in \mathcal{R}\left(M, \mathcal{D}, \mathcal{U}\right), \mathcal{U} \supset \mathcal{D}, \ |\mathcal{U}| = u \right\}, \tag{4.1.44}$$

where

$$\mathcal{R}\left(M, \mathcal{D}, \mathcal{U}\right) = \left\{ \|\underline{\alpha}_{\mathcal{D}^c}\|_1 \leqslant M \|\underline{\alpha}_{\mathcal{D}}\|_1, \|\underline{\alpha}_{\mathcal{U}}\|_\infty \leqslant \min_{j \in \mathcal{U} \setminus \mathcal{D}} |\alpha_j| \right\}$$

with $M > 1$ and $\min_{j \in \mathcal{U} \setminus \mathcal{D}} |\alpha_j| = \infty$ if $\mathcal{U} = \mathcal{D}$. If $v\left(M, \mathcal{D}, u\right) > 0$, then the restricted eigenvalue condition is satisfied. The condition is stricter than the compatibility condition and it holds that $v\left(M, \mathcal{D}\right) \leqslant v_c\left(M, \mathcal{D}, u\right)$ for all $u \geqslant d$. Note that the restricted eigenvalue condition is related to a lower bound on the eigenvalues of Σ, since

$$e_{\min}\left(\Sigma\right) \|\underline{\alpha}\|^2 \leqslant \underline{\alpha}^T \Sigma \underline{\alpha} \leqslant e_{\max}\left(\Sigma\right) \|\underline{\alpha}\|^2,$$

where $e_{\min}\left(\Sigma\right)$ and $e_{\max}\left(\Sigma\right)$ are the smallest and largest eigenvalues of Σ, respectively. Recall (from Section 2.1.7) that eigenvalues of Σ that are close to zero are indicative of high levels of collinearity. See Bühlmann & van de Geer (2011:156-177) and Bickel *et al.* (2009) for more information about compatibility and restricted eigenvalue conditions. Under the restricted eigenvalue condition,

$$\left\|\hat{\underline{\alpha}}^L - \underline{\alpha}\right\|_2 \xrightarrow{p} 0,$$

and the rate of convergence is

$$\left\|\hat{\underline{\alpha}}^L - \underline{\alpha}\right\|_2 = O_p\left(v^{-2}\sqrt{d \ln p/n}\right).$$

A direct consequence is that the LASSO can be used for variable screening, identification of the important variables. Denote the important variables, those with large effects, by

$$\mathcal{D}^D = \left\{ j \middle| |\alpha_j| \geqslant D \right\},$$

for some $D > 0$. With $\lambda_n \asymp \sqrt{\log p/n}$, asymptotic results are obtained under different assumptions:

- Under the compatibility condition, $\left\|\hat{\underline{\alpha}}^L - \underline{\alpha}\right\|_1 \leqslant a_n = O\left(d\sqrt{\ln p/n}\right)$ so that

  $\lim_{n \to \infty} P\left(\widehat{\mathcal{D}}\left(\lambda_n\right) \supset \mathcal{D}^{D_n} \middle| D_n > a_n\right) \to 1.$

- Under the restricted eigenvalue condition, $\left\|\hat{\underline{\alpha}}^L - \underline{\alpha}\right\| \leqslant b_n = O\left(\sqrt{d \ln p/n}\right)$ so that $\lim_{n\to\infty} P\left(\widehat{\mathcal{D}}(\lambda_n) \supset \mathcal{D}^{D_n} \middle| D_n > b_n\right) \to 1$.
- If $\mathcal{D}^{D_n} = \mathcal{D}$, then the beta-min condition described below holds and

$$\lim_{n\to\infty} P\left(\widehat{\mathcal{D}}(\lambda) \supseteq \mathcal{D}\right) \to 1. \tag{4.1.45}$$

Variable screening is used purely to reduce the dimension of the problem and is performed prior to a variable selection and/or estimation method. Other methods for variable screening are sure independence screening (SIS) proposed by Fan & Lv (2008), safe feature elimination (SAFE) by Ghaoui *et al.* (2012), and the strong rules discussed in Tibshirani *et al.* (2012).

Variable Selection Consistency

Using the LASSO for variable selection, we can use the LAR-LASSO algorithm to compute all LASSO models of different sizes over the path of $\lambda \in \left[0, \left\|\mathbf{Z}^T\mathbf{v}\right\|_\infty\right]$. Let the set of LASSO subsets be denoted by $\widehat{\mathcal{L}} = \left\{\widehat{\mathcal{D}}_k(\lambda) \middle| k = 0, 1, \ldots, M\right\}$, where $\widehat{\mathcal{D}}_k(\lambda) = \left\{j \middle| \hat{\alpha}_j^L(\lambda) \neq 0\right\}$, $\left|\widehat{\mathcal{D}}_k(\lambda)\right| = k$ and $M = O(\min(n, p))$. Since the active set is unchanged between steps, $\widehat{\mathcal{L}}$ contains all the possible subsets that can be selected by the LASSO so that there are a total of $\left|\widehat{\mathcal{L}}\right| = O(\min(n, p))$ models to consider. Comparison with Table 2.2.1 shows that variable selection with the LASSO is far more efficient than any of the subset selection methods. The question is whether the correct subset $\mathcal{D}$ is contained in $\widehat{\mathcal{L}}$ and if so, how do we select the value of λ to identify it. The problems we encounter when using the LASSO for variable selection are:

1. The shrinkage parameter λ must be larger for selection than prediction.
2. Small nonzero parameters cannot be detected consistently.
3. High correlations between predictors leads to poor selection performance.

Leng *et al.* (2006) showed that the LASSO is generally not consistent for variable selection when λ is chosen for prediction accuracy. Meinshausen & Bühlmann (2006) proved a similar result for $p = O(n^a)$ with $a > 1$. Their work is in the context of neighbourhood selection for Gaussian graphical models but can be interpreted as variable selection for linear regression with Gaussian variables. They show that consistent neighbourhood selection with the LASSO is possible (under a number of conditions) if λ is

chosen to be larger than the prediction optimal value. That is, λ_n should decay at a slower rate than $O\left(\sqrt{n}\right)$ given by Knight & Fu (2000). Corresponding with the results above, the shrinkage parameter should satisfy

$$\lambda_n \gg \sqrt{\log p/n}$$

A lower bound on the nonzero parameters is necessary to overcome the LASSOs inability to detect small variables. Bühlmann & van de Geer (2011:21,24) call such a restriction the beta-min condition, which must satisfy

$$\alpha_{\min} = \min_{j\in\mathcal{D}} \left|\alpha_j\right| \gg \upsilon^{-2}\sqrt{d \ln p/n} \tag{4.1.46}$$

The condition is discussed further in Bühlmann & van de Geer (2011:187-188) where they relate it to the signal to noise ratio and the minimal eigenvalue of Σ_{11}.

A very strict condition, known as the irrepresentable condition, is needed to address the correlations between predictors. Coincidentally, a number of researchers independently discovered similar results: the LASSO is selection consistent,

$$\lim_{n\to\infty} P\left(\widehat{\mathcal{D}}\left(\lambda_n\right) = \mathcal{D}\right) = 1, \tag{4.1.47}$$

if and (almost) only if

$$\left\|\Sigma_{21}\Sigma_{11}^{-1}\operatorname{sign}\left(\underline{\alpha}_{\mathcal{D}}\right)\right\|_{\infty} \leqslant 1-\epsilon \text{ for } \epsilon > 0, \tag{4.1.48}$$

where $\Sigma = \mathbf{Z}^T\mathbf{Z}/n$ is partitioned as

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

so that

- Σ_{11} is the $d \times d$ nonsingular matrix with elements $\left(\Sigma_{jk}\right)_{j,k\in\mathcal{D}}$,
- Σ_{21} is the $(p-d)\times d$ matrix with elements $\left(\Sigma_{jk}\right)_{j\notin\mathcal{D},k\in\mathcal{D}}$,
- Σ_{12} is the $d\times(p-d)$ matrix Σ_{21}^T
- Σ_{22} is the $(p-d)\times(p-d)$ matrix with elements $\left(\Sigma_{jk}\right)_{j,k\notin\mathcal{D}}$.

Meinshausen & Bühlmann (2006) proposed the neighbourhood stability condition which is equivalent to (4.1.48) and showed that the condition cannot be relaxed. Zou (2006) proved that it is a necessary condition for selection consistency. Yuan & Lin (2007) showed that it is necessary and sufficient

for the LASSO to be path consistent, a term they use to describe simultaneous consistence in estimation and in variable selection. Zhao & Yu (2006) coined the term irrepresentable condition and showed that it is almost necessary and sufficient for sign consistency. With sign consistency, the signs of the nonzero estimates should match the signs of true parameters in addition to distinguishing them from the zero parameters. Let

$$\text{sign}_0\left(\alpha_j\right) = \begin{cases} 1 & \text{if } \alpha_j > 0 \\ -1 & \text{if } \alpha_j < 0 \\ 0 & \text{if } \alpha_j = 0 \end{cases}$$

then sign consistency implies

$$\lim_{n\to\infty} P\left(\text{sign}_0\left(\hat{\underline{\alpha}}\left(\lambda_n\right)\right) = \text{sign}_0\left(\underline{\alpha}\right)\right) = 1.$$

Note that the irrepresentable condition is similar to applying a constraint on the regression coefficients obtained when regressing the irrelevant parameters on the relevant ones,

$$\left\|\Sigma_{11}^{-1}\Sigma_{12}\right\|_\infty = \left\|\left(\mathbf{Z}_{\mathcal{D}}^T\mathbf{Z}_{\mathcal{D}}\right)^{-1}\mathbf{Z}_{\mathcal{D}}^T\mathbf{Z}_{\mathcal{D}^c}\right\|_\infty \leqslant 1-\epsilon.$$

In other words, the relevant variables should not be too highly correlated with the irrelevant ones. Usually, a restriction on the minimum eigenvalues of $\mathbf{Z}_{\mathcal{D}}^T\mathbf{Z}_{\mathcal{D}}$, $e_{\min}\left(\Sigma_{11}\right)$, is also necessary so that the level of collinearity among the relevant variables is not too high.

Some examples where the irrepresentable condition holds are provided by Zhao & Yu (2006):

- *Constant positive correlation:* $\Sigma_{jj} = 1$ for $j = 1, 2, \ldots, p$ and $\Sigma_{jk} = \rho$ for $j \neq k$, where $0 \leqslant \rho \leqslant 1/(1+cd)$ for any $c > 0$.
- *Bounded correlation:* $\Sigma_{jj} = 1$ for $j = 1, 2, \ldots, p$ and $\Sigma_{jk} = \rho$ for $j \neq k$, where $|\rho| \leqslant c/(2d-1)$ for any $0 \leqslant c < 1$.
- *Power decay correlation:* $\Sigma_{jk} = \rho^{|j-k|}$ for $j, k = 1, 2, \ldots, p$, where $|\rho| < 1$.
- *Block designs:* Consider designs such as

$$\Sigma = \begin{bmatrix} A_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & A_k \end{bmatrix}$$

where $\underline{\alpha} = \left(\underline{\alpha}_1, \underline{\alpha}_2, \ldots, \underline{\alpha}_k\right)$ with $\underline{\alpha}_j$ corresponding to block A_j. The strong irrepresentable condition holds if there is a common $\underline{\epsilon}$ such that the strong irrepresentable condition holds for all A_j and $\underline{\alpha}_j$.

- Orthogonal designs: $\Sigma = I_p$
- General designs where $d = 1$
- General designs where $p = 2$

Wainwright (2009) calls condition (4.1.48) mutual incoherence. He provides results for sparse models, $|\mathcal{D}| = d \ll p$, with a general scaling on n, p and d, where $p = p(n)$ and $d = d(n)$ are allowed to grow as the number of observations grow. For sign consistency,

$$\lim_{n\to\infty} P\left(\text{sign}_0\left(\hat{\underline{\alpha}}\left(\lambda_n\right)\right) = \text{sign}_0\left(\underline{\alpha}\right)\right) = 1 - 4\exp\left(-cn\lambda_n^2\right) \to 1,$$

for some $c > 0$, the following conditions are sufficient:

1. $\left\|\Sigma_{21}\Sigma_{11}^{-1}\,\text{sign}\left(\underline{\beta}_{\mathcal{D}}\right)\right\|_\infty \leqslant 1 - \epsilon$ for $\epsilon \in (0,1]$
2. $e_{\min}\left(\Sigma_{11}\right) \geqslant c_{\min}$ for some $c_{\min} > 0$
3. $\lambda_n > \frac{2}{\epsilon}\sqrt{2\sigma^2 \ln p/n}$
4. $\alpha_{\min} > g\left(\lambda_n\right) = \lambda_n \left\|\Sigma_{11}^{-1}\right\|_\infty + 4\sigma\lambda_n/\sqrt{c_{\min}}$

Consequences of these conditions on estimation are:

- $\lim_{n\to\infty} P\left(\widehat{\mathcal{D}}\left(\lambda_n\right) \subseteq \mathcal{D}\right) = 1 - 4\exp\left(-c_1 n\lambda_n^2\right) \to 1$
- $\left\|\hat{\underline{\alpha}}_{\mathcal{D}} - \underline{\alpha}_{\mathcal{D}}\right\|_\infty \leqslant g\left(\lambda_n\right)$
- Assuming that $\left\|\Sigma_{11}\right\|_\infty = O(1)$ and choosing $\lambda_n = O\left(\sqrt{\ln p/n}\right)$it follows that $\left\|\hat{\underline{\alpha}}_{\mathcal{D}} - \underline{\alpha}_{\mathcal{D}}\right\|_2 = O\left(\lambda_n\sqrt{d}\right) = O\left(\sqrt{d\ln p/n}\right)$

These results tie in nicely with the results presented above by Bühlmann & van de Geer (2011). This is because the irrepresentable condition implies the restricted eigenvalue condition so that under conditions above, the LASSO achieves consistent selection and consistent estimation. Further details about the irrepresentable condition and how it relates to the restricted eigenvalue conditions can be found in

Bühlmann & van de Geer (2011:189-203). The results obtained by Zhao & Yu (2006) are a special case where the scaling of (n, p, d) is such that p is exponentially larger than n. With $p = O\left(\exp\left(n^{c_3}\right)\right)$ and $d = O\left(n^{c_1}\right)$, the shrinkage parameter and beta-min conditions are $\lambda_n^2 = 1/n^{1-c_4}$ and $\alpha_{\min}^2 = 1/n^{1-c_2}$ with $0 \leqslant c_1 < c_2 < 1$ and $0 \leqslant c_3 < c_4 < c_2 - c_1$ so that consistency is achieved with probability $1 - \exp\left(-cn^{c_4}\right) \to 1$. Wainwright (2009) remarks that the sparsity constraint is very strong in this case since $d/p \approx n^{c_1} \exp\left(-n^{c_3}\right)$ disappears quickly. He shows further that sign consistency is not achieved, in particular

$$P\left(\text{sign}_0\left(\underline{\hat{\alpha}}\left(\lambda_n\right)\right) = \text{sign}_0\left(\underline{\alpha}\right)\right) \leqslant 1/2,$$

if either

1. the irrepresentable condition is violated, $\left\|\Sigma_{21}\Sigma_{11}^{-1}\,\text{sign}\left(\underline{\beta}_{\mathcal{D}}\right)\right\|_\infty = 1 + \epsilon$ for $\epsilon \in (0, 1]$, or
2. $\alpha_j \in \left(0, \left|h_j\left(\lambda_n\right)\right|\right)$ for any $j \in \mathcal{D}$ where $h_j\left(\lambda_n\right) = \lambda_n \underline{e}_j^T \Sigma_{11}^{-1}\,\text{sign}\left(\underline{\alpha}_{\mathcal{D}}\right)$ and $\underline{e}_j$ is the j-th coordinate vector.

Thus, the irrepresentable condition is necessary for sign consistency. Furthermore, the matrix Σ_{11} must be well conditioned so that none of the elements in Σ_{11}^{-1} are too large and $\alpha_{\min}$ must not decay faster than λ_n. Wainwright (2009) extends these results for random Gaussian predictors and provides a threshold for the number of observations, depending on the scaling of the parameters (p, d), that are sufficient for consistent selection with the LASSO. For further insights concerning LASSO selection, see Zhang & Huang (2008), Candès & Plan (2009) and Meinshausen & Yu (2009).

Summary

In orthogonal designs, the LASSO thresholding function is near minimax optimal. More generally, when the true model is sparse and the tuning parameter is selected to minimize the squared error loss, the LASSO displays impressive properties for estimation and prediction. It has superior prediction performance, even when $p \gg n$, and although convergence may be slow, an oracle inequality can be established under the weak compatibility condition to improve the convergence rate. For fixed p, parameters are estimated consistently and achieve asymptotic normality. Under the slightly stronger restricted eigenvalue condition, the estimates are also consistent when $p \gg n$ and p grows with n. If additionally there are no small nonzero parameters, the beta-min condition is satisfied and the LASSO can be utilized for variable screening. However, variable selection with the LASSO requires more restrictive conditions. If the tuning

parameter is selected to be larger than the prediction optimal value, consistent selection is attained only under both the beta-min condition and the irrepresentable condition. Bühlmann & van de Geer (2011:24) provide a convenient summary of the conditions required for different purposes of the LASSO, it is shown in Table 4.1.3.

Purpose	**Conditions**	
	Design	*Parameters*
Prediction, slow (4.1.39)	None	None
Prediction, fast (4.1.42)	Compatibility (4.1.40)	None
Variable screening (4.1.45)	Restricted eigenvalue (4.1.44)	Beta-min (4.1.46)
Variable selection (4.1.47)	Irrepresentable (4.1.48)	Beta-min (4.1.46)

Table (4.1.3) Conditions for consistency when using the LASSO for different purposes

4.1.6 Model Selection

The performance of the LASSO relies heavily on the choice of tuning parameter to select the optimal model. For prediction purposes, the squared error loss is minimized using either cross-validation methods or information criteria. A drawback of using information criteria is that the model DF must be known. However, recent studies have shown surprisingly that the LASSO uses DF equal to the number of nonzero parameters in the model. Selection of the tuning parameter for variable selection is more difficult since the prediction optimal value is inconsistent for selection. Recent advances have been made to stabilize the selection and usually entail some form of resampling, like multiple sample splitting or bootstrapping.

Prediction

When using the LASSO for prediction, the tuning parameter can be found using K-fold cross-validation. For each K the model is estimated at each value of the tuning parameter and then used to predict the hold out sample. The value of the tuning parameter yielding the lowest CV error is selected as the best one and the final model is fitted using that value. Cross-validation can be performed by optimizing over either λ, t or s, the latter providing a convenient choice since it must lie on the interval $[0,1]$. Homrighausen & McDonald (2013) show that the LASSO is persistent when k-fold cross-validation is used for selection of the tuning parameter and Homrighausen & Mcdonald (2014) show persistence for LOOCV.

Cross-validation can become computationally expensive depending on the dimension of the data and

the algorithm used to find the LASSO solution. As an alternative, GCV or information criteria can be used. The difficulty therein lies in determining the DF. The LASSO estimate is nonlinear and cannot be written as a linear combination of the response, hence (3.1.10) can't be used. Tibshirani (1996) approximated the DF by making use of the linear approximation (4.1.29). Then

$$df\left(\hat{f}_L\right) \approx \text{tr}\left[\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{W}^-\right)^{-1}\mathbf{Z}^T\right]$$

and he suggested using GCV along with this approximation. He also proposes an approximation based on Stein's unbiased risk estimation (SURE), which assumes normality. See Seber & Lee (2003:420-422) for information about Stein shrinkage and Tibshirani (1996) for the approximation. The estimate (3.1.9) is actually based on SURE theory, and is used in further studies concerning DF. Efron *et al.* (2004) used the theory to find the DF of the LASSO and showed that the C_p statistic using this DF is an unbiased estimate of the prediction error. For LAR they show that, at the k-th step, the DF is approximated by the step number k. However, the LAR-LASSO algorithm can have more steps than the number of variables. Interestingly, it turns out that the DF is well approximated by the number of nonzero predictors in the model,

$$df\left(\hat{f}_L\right) \approx E\left|\widehat{\mathcal{D}}\left(\lambda\right)\right|.$$

Although a price is paid in DF for the adaptive fitting of the LASSO, the DF that are saved due to shrinking the estimates appears to balances out. Zou *et al.* (2007) developed the theory further and conclude that the number of nonzero predictors is an unbiased and consistent estimate of the DF. They use the estimate $df\left(\hat{f}_L\right)$ to construct the statistics C_p, AIC and BIC. Furthermore, they prove that the optimal value of λ is one of the transition points in the LASSO path. That is, at one of the LAR steps when a new variable joins the active set. These results only hold if the predictor matrix has full column rank. Tibshirani & Taylor (2012) and Dossal *et al.* (2013) generalized the result (independently) so that the full rank assumption is not needed. Their estimates can therefore be used when $p > n$, even when the LASSO solution might not be unique. Tibshirani & Taylor (2012) prove that the DF is given by the rank of the active predictors,

$$df\left(\hat{f}_L\right) = E\left[\text{rank}\left(\mathbf{Z}_{\widehat{\mathcal{D}}}\right)\right].$$

The result by Dossal *et al.* (2013) is similar,

$$df\left(\hat{f}_L\right) = E\left|\widehat{\mathcal{D}}^*\left(\lambda\right)\right|,$$

where $\widehat{\mathcal{D}}^*\left(\lambda\right) = \left\{j|\hat{\alpha}_j^* \neq 0\right\}$ and $\underline{\hat{\alpha}}^*$ is a solution such that $\mathbf{Z}_{\widehat{\mathcal{D}}^*}$ has full rank. They show that $\widehat{\mathcal{D}}^*\left(\lambda\right)$ is the minimum size of all active sets of LASSO solutions. Since each of these studies are based on SURE, the response is assumed to be Gaussian.

Selection

Use of the LASSO for variable selection requires a larger shrinkage parameter. Cross-validation does not lead to consistent selection. The 1 SE rule might improve selection performance but there is no theoretical justification for its use with the LASSO.

Wang *et al.* (2009) constructed a modified *BIC* criterion,

$$BIC_{\mathcal{D}} = \ln\left(\hat{\sigma}_{\mathcal{D}}^2\right) + \left|\mathcal{D}\right|\frac{\ln\left(n\right)}{n}C_n,$$

where $\hat{\sigma}_{\mathcal{D}}^2 = \left\|\mathbf{v} - \mathbf{Z}_{\mathcal{D}}\underline{\hat{\alpha}}_{\mathcal{D}}\right\|^2 / n$ and $C_n > 0$ is a positive constant. The ordinary *BIC* is obtained when $C_n = 1$. They prove that it is consistent for model selection for fixed p and when p grows with n. However, $BIC_{\mathcal{D}}$ cannot be used when $p > n$ since $\hat{\sigma}_{\mathcal{D}}^2$ becomes 0 and $\ln\left(\hat{\sigma}_{\mathcal{D}}^2\right)$ is undefined. They use the value of $C_n = \ln\left(\ln p\right)$ in their studies where p varies with n. Chand (2012) show that $C_n = \sqrt{n}/p$ leads to consistent selection. A more general information criterion is proposed by Fan & Tang (2013), the generalized information criteria GIC, which can be used when $p \gg n$.

Sun *et al.* (2013) propose selecting the tuning parameter by variable selection stability. They make use of the kappa coefficient, which measures the agreement between two independent sets,

$$\kappa\left(\mathcal{A}_1, \mathcal{A}_2\right) = \frac{P\left(\text{obs}\right) - P\left(\text{chance}\right)}{P\left(\text{chance}\right)},$$

where the relative observed agreement is

$$P\left(\text{obs}\right) = \frac{1}{p}\left[\left|\mathcal{A}_1 \cap \mathcal{A}_2\right| + \left|\mathcal{A}_1^c \cap \mathcal{A}_2^c\right|\right]$$

and the probability of chance agreement is

$$P(\text{chance}) = \frac{1}{p^2}\left[|\mathcal{A}_1 \cap \mathcal{A}_2| + |\mathcal{A}_1 \cap \mathcal{A}_2^c|\right]\left[|\mathcal{A}_1 \cap \mathcal{A}_2| + |\mathcal{A}_1^c \cap \mathcal{A}_2|\right] + \frac{1}{p^2}\left[|\mathcal{A}_1 \cap \mathcal{A}_2^c| + |\mathcal{A}_1^c \cap \mathcal{A}_2^c|\right]\left[|\mathcal{A}_1^c \cap \mathcal{A}_2| + |\mathcal{A}_1^c \cap \mathcal{A}_2^c|\right].$$

The coefficient lies on the interval $\kappa \in [-1, 1]$, $\kappa = -1$ for complete disagreement and $\kappa = 1$ for complete agreement between the two sets. By applying the same model to two different data sets, two active sets are obtained for the same model and the agreement between them can be measured by the kappa coefficient. The set of variables chosen by a particular method should not vary much for samples drawn from the same population. Thus, the kappa coefficient can be used as a measure of variable selection performance and the tuning parameter can be selected by maximizing the kappa coefficient. Suppose the training observations are denoted by $\underline{t}_i = \left(\underline{x}_i, y_i\right)$ for $i = 1, 2, \ldots, n$, then the training set can then be denoted by $T = \left(\underline{t}_1, \underline{t}_2, \ldots, \underline{t}_n\right)$. They propose the procedure below.

Algorithm 4.1.4 ***Variable selection stability***

1. *For* $b = 1, 2, \ldots, B$*:*

   (a) *Randomly split the training data into two equally sized samples* $T_1^{*b} = \left(\underline{t}_1^{*b}, \underline{t}_2^{*b}, \ldots, \underline{t}_m^{*b}\right)$ *and* $T_2^{*b} = \left(\underline{t}_{m+1}^{*b}, \underline{t}_{m+2}^{*b}, \ldots, \underline{t}_{2m}^{*b}\right)$

   (b) *Calculate the* LASSO *path for each sample* $\widehat{\mathcal{L}}_1^{*b} = \left\{\widehat{\mathcal{D}}_1^{*b}(\lambda_k)\right\}$ *and* $\widehat{\mathcal{L}}_2^{*b} = \left\{\widehat{\mathcal{D}}_2^{*b}(\lambda_k)\right\}$

   (c) *For* $k = 1, 2, \ldots M$*: Calculate the variable selection stability*

$$\hat{s}^{*b}(\lambda_k) = \kappa^{*b}\left(\widehat{\mathcal{D}}_1^{*b}(\lambda_k), \widehat{\mathcal{D}}_2^{*b}(\lambda_k)\right)$$

2. *For* $k = 1, 2, \ldots M$*: Calculate the average variable selection stability*

$$\hat{s}(\lambda_k) = \frac{1}{B}\sum_{b=1}^{B} \hat{s}^{*b}(\lambda_k)$$

3. *Calculate*

$$\hat{s}_{\max} = \max_{\lambda_k}\left\{\hat{s}(\lambda_k)\right\}$$

and select the optimal $\hat{\lambda}$ *for variable selection as the one obtaining the upper* $1-\theta_n$ *percentile of* $\hat{s}^*(\lambda_k)$

$$\hat{\lambda} = \min\left\{\lambda_k \left| \frac{\hat{s}(\lambda_k)}{\hat{s}_{\max}} \geqslant 1-\theta_n \right.\right\}.$$

Sun *et al.* (2013) recommend using a small value of $\lim_{n\to\infty}\theta_n \to 0$ and remark that their studies were performed using $\theta_n = 0.1$. However, they state that, while θ_n varies in a certain range, it has little effect on the selection performance. They prove that the method leads to consistent variable selection for fixed p and when p is allowed to grow with n. Fang *et al.* (2013) propose combining the performance of variable selection and prediction by a criterion they call prediction and stability selection (PASS),

$$PASS(\lambda) = \frac{\sum_{b=1}^{B}\kappa^{*b}\left(\widehat{\mathcal{D}}_1^{*b}(\lambda_k), \widehat{\mathcal{D}}_2^{*b}(\lambda_k)\right)}{\sum_{b=1}^{B} CV\left(Z_1^{*b}, Z_2^{*b}|\lambda_k\right)}.$$

Hence, the criterion is the ratio of the average kappa coefficient to the cross-validation error. They also show that this criterion is consistent.

Roberts & Nowak (2014) propose the percentile-lasso, a method that repeatedly performs K-fold cross-validation. The idea is to stabilize the variability due to different fold allocations. K-fold CV is applied for M repetitions and the θ-th percentile of the vector $\left(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_M\right)$ is used as the shrinkage parameter. They find that $M \geqslant 10$ is necessary but suggest using $M = 100$ if an efficient algorithm is used to compute the LASSO. They further suggest using $\theta = 0.95$ or alternatively, for a range of θ, find the subset obtained by the LASSO with $\lambda = \hat{\lambda}(\theta)$ and fit the model using least squares. $\hat{\theta}$ is then selected as the value corresponding to the re-estimated model with the lowest error. They show by simulations that the number of true nonzero parameters selected decreases by a negligible amount, while the number of false inclusions and the variability of the model size $\hat{d}$ is greatly reduced.

Bühlmann & van de Geer (2011:339-385) discuss two methods of variable selection using the LASSO that do not directly choose a value of the tuning parameter for model selection. The first is stability selection which was introduced by Meinshausen & Bühlmann (2010). Subsampling or bootstrapping is used for the selection of variables. Instead of selecting the model for a value of λ, the selected model is chosen such that the probability of its selection over the subsamples exceeds some threshold value. They remark that the selection is more stable with this approach and that the error rate of false positives is

controlled. However, a rather strong assumption, called the exchangeability condition, is required for its application. The second method requires weaker assumptions and was proposed by Meinshausen (2009). The data is repeatedly split into two samples of roughly equal size. The LASSO is used to select the variables with the first sample. The chosen subset is then estimated using least squares on the second sample and the p-values are recorded for each active variable (corrected for multiple testing) and set to one for each inactive variables. After the multi sample splitting, the p-values for each variable are aggregated based on quantiles. Variables can then be selected for the final model if their aggregated p-values exceed a pre-determined threshold.

4.2 Two-stage LASSO Methods

The LASSO shrinkage is constant - small parameters are set to zero but the other parameters are shrunk at a constant rate regardless of their size. As a result, large parameters can be overshrunk which causes the bias of the LASSO estimate to increase. A problem with the LASSO is that it relies on the use of one tuning parameter for both selection and estimation. Selecting the tuning parameter with cross-validation or information criteria yields a model that is optimal for prediction. The value of the shrinkage parameter λ is often small in this case and too many variables are included in the model. Using the 1 SE rule with cross-validation can improve the variable selection properties of the LASSO. In this case, the value of λ is usually much larger so that more variables are set to zero. However, the larger λ also shrinks the other parameters more and the large bias results in poor prediction accuracy. Although the LASSO overfits the model when a prediction-optimal value of λ is used, the resulting model contains the true subset of variables with a high probability. This is suggestive of using the LASSO in a two-stage design which utilizes more than one tuning parameter.

4.2.1 Relaxed LASSO

Meinshausen (2007) proposed the relaxed LASSO as a two-stage design to control the bias. The motivation was to find a method with low computational complexity and good asymptotic properties in a high-dimensional setting where $p \gg n$. The LASSO, ridge regression and subset selection are special cases of bridge estimators suggested by Frank & Friedman (1993),

$$\underline{\hat{\alpha}}^B = \arg\min_{\underline{\alpha}} \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 + \lambda \|\underline{\alpha}\|_\gamma^\gamma, \tag{4.2.1}$$

for $\lambda \geqslant 0$ and $\gamma \geqslant 0$. The idea is to use two tuning parameters where λ controls the amount of shrinkage and γ controls the rotation of the estimates with respect to the coordinate axes. Fu (1998) developed an algorithm to solve for bridge estimates with $\gamma > 1$ using a modified Newton-Raphson method. The bridge penalty,

$$P_B(\underline{\alpha}) = \lambda \|\underline{\alpha}\|_\gamma^\gamma = \lambda \left(\ell_\gamma(\underline{\alpha})\right)^\gamma, \tag{4.2.2}$$

includes the following special cases by using different ℓ_γ-norms:

- $\ell_0(\underline{\alpha}) = \sum_{j=1}^p \delta\left(\alpha_j \neq 0\right)$ corresponds to subset selection
- $\ell_1(\underline{\alpha}) = \sum_{j=1}^p |\alpha_j|$ corresponds to the LASSO penalty
- $\ell_2(\underline{\alpha}) = \left(\sum_{j=1}^p \alpha_j^2\right)^{\frac{1}{2}}$ corresponds to the ridge penalty

The study by Knight & Fu (2000) on the limiting distributions of bridge estimates shows that when $\gamma > 1$, the amount of shrinkage increases as the size of the parameter increases so that the bias of large parameters will be very high. In contrast, they showed that nonzero parameters, including large parameters, are estimated without bias while zero parameters are shrunk to zero when $\gamma < 1$. In the high dimensional setting convergence rates are much faster when $\gamma < 1$. With $\gamma > 1$, convergence is slow as p increases. Meinshausen (2007) shows that for the LASSO ($\gamma = 1$), the rate of convergence is also slow and with a constant rate of shrinkage, large parameters are still biased to a degree. Therefore, a procedure such as bridge estimation with $\gamma < 1$ is desired. However, these procedures lack the low computational complexity experienced when $\gamma \geqslant 1$ since the calculations involve minimizing a concave penalty function, which can be difficult especially when $p \gg n$.

Meinshausen (2007) shows that the relaxed LASSO achieves low bias, fast convergence and low computational expense. Let $\widehat{\mathcal{D}} = \left\{j : \hat{\alpha}_j^L \neq 0\right\}$ be the nonzero subset variables obtained by the LASSO estimate, then the relaxed LASSO is given by

$$\underline{\hat{\alpha}}_{\widehat{\mathcal{D}}}^{RL} = \underset{\underline{\alpha}_{\widehat{\mathcal{D}}}}{\arg\min} \left\|\mathbf{v} - \mathbf{Z}_{\widehat{\mathcal{D}}}\underline{\alpha}_{\widehat{\mathcal{D}}}\right\|^2 + \phi\lambda \|\underline{\alpha}\|_1, \tag{4.2.3}$$

where $\lambda \geqslant 0$ is the shrinkage parameter, $\phi \in (0,1]$ is the relaxation parameter and $\underline{\hat{\alpha}}_{\widehat{\mathcal{D}}^c}^{RL} = \underline{0}$. The penalty function

$$P_{RL}(\underline{\alpha}) = \phi\lambda \|\underline{\alpha}\|_1 \tag{4.2.4}$$

is convex and is identical to the LASSO penalty except for the factor ϕ which decreases, or relaxes, the amount of shrinkage applied by λ. In this case, the shrinkage parameter λ controls variable selection in the first stage and the relaxation parameter ϕ controls estimation in the second stage by relaxing the amount of shrinkage. When $\phi = 1$, the shrinkage is not relaxed and the LASSO estimate is obtained. The smaller ϕ is, the less shrinkage is applied. As a result, the relaxed LASSO often selects sparser models than the LASSO and can yield better prediction accuracy. The case when $\phi = 0$ is defined as the limit when $\phi \to 0$ and corresponds to the LAR algorithm used with least squares (LAR-OLS) hybrid method in which estimation is carried out using least squares. Meinshausen (2007) proves that the relaxed LASSO outperforms both the LAR-OLS hybrid and the LASSO. Furthermore, they show that variable selection is consistent when choosing the tuning parameters by optimizing for prediction. The relaxed LASSO is shown to have the following asymptotic properties:

1. For known values of (λ, ϕ), the convergence rate is fast and independent of p
   $\inf MSE\left(f_{RL}\left(\underline{z}\right)\right) = O_p\left(1/n\right)$.
2. Selection of (λ, ϕ) by K-fold CV:
   (a) The convergence is near the optimal rate with known (λ, ϕ),
   $MSE\left(\hat{f}_{RL}\left(\underline{z}\right)\right) = O_p\left(\ln\left(n\right)^2/n\right)$.
   (b) Consistent variable selection (shown by simulations).

Computation of the relaxed LASSO is performed using a modification of the LAR-LASSO algorithm. The basic idea is that the entire LASSO path is calculated first using LAR-LASSO. Then for each model produced by LAR-LASSO, the LASSO is applied again (using LAR-LASSO), but this time using the relaxed penalty and only the predictors in that particular model. In this way, the entire path of the relaxed LASSO is computed and the optimal tuning parameters λ and ϕ can be selected simultaneously using cross-validation. The final algorithm interpolates the path from the initial LAR-LASSO fit to find the solutions at no extra cost. However, if the extrapolations cross zero for a particular model then LAR-LASSO must be applied again for that model. In the worst case, with all extrapolations crossing zero, the computational complexity increases to the order $O\left(np\min\left(n, p\right)^2\right)$.

Algorithm 4.2.1 ***Relaxed LASSO***

1. *Compute all* LASSO *solutions using the* LAR-LASSO *algorithm and name them* $\underline{\hat{\alpha}}_1^L, \underline{\hat{\alpha}}_2^L, \ldots, \underline{\hat{\alpha}}_o^L$ *with corresponding shrinkage parameters* $\lambda_1, \lambda_2, \ldots, \lambda_o$ *and nonzero variables* $\widehat{\mathcal{D}}_1, \widehat{\mathcal{D}}_2, \ldots, \widehat{\mathcal{D}}_o$ *, where* $o = \min(n, p)$.

2. *For each* $k = 1, 2, \ldots, o$, *compute the directions in which the solutions lie,*

   $\underline{e}_k = \left(\underline{\hat{\alpha}}_k^L - \underline{\hat{\alpha}}_{k-1}^L\right) / \left(\lambda_{k-1} - \lambda_k\right)$. *Let* $\underline{\hat{\alpha}}_k^* = \underline{\hat{\alpha}}_k^L + \lambda_k \underline{e}_k$.

3. *If any* $\text{sign}\left(\hat{\alpha}_{k,j}^*\right) \neq \text{sign}\left(\hat{\alpha}_{k,j}^L\right)$ *for* $j = 1, 2, \ldots, p$,

   (a) *compute all relaxed* LASSO *solutions using the* LAR-LASSO *algorithm with the subset of variables* $\widehat{\mathcal{D}}_k$ *and varying the tuning parameter between 0 and* λ_k.

   (b) *Else interpolate the relaxed* LASSO *solutions between* $\underline{\hat{\alpha}}_{k-1}^L$ *(equivalent to* $\phi = 1$*) and* $\underline{\hat{\alpha}}_k^*$ *(equivalent to* $\phi = 0$*).*

4.2.2 Adaptive LASSO

The adaptive LASSO proposed by Zou (2006) uses a positive weighting factors to directly control the bias by shrink parameters by different amounts. The adaptive LASSO estimator is given by

$$\underline{\hat{\alpha}}^{AL} = \underset{\underline{\alpha}}{\arg\min} \ \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 + \lambda \sum_{j=1}^{p} w_j \left|\alpha_j\right|, \tag{4.2.5}$$

where $\lambda \geqslant 0$ is the shrinkage parameter $w_j > 0$ are the weights. Adaptively selected weights are defined as $\hat{w}_j = 1 \Big/ \left|\hat{\alpha}_j^{init}\right|^{\zeta}$, where $\zeta > 0$ and $\hat{\alpha}_j^{init}$ is an initial estimate. Thus, the amount of shrinkage applied to a parameter is inversely proportional to its size - large parameters are shrunk less and small parameters are shrunk more. The penalty function,

$$P_{AL}\left(\underline{\alpha}\right) = \lambda \sum_{j=1}^{p} w_j \left|\alpha_j\right| \tag{4.2.6}$$

is convex and is the same as the LASSO penalty when $w_j = 1$. Suppose that the true set of relevant variables are the first d variables, $\mathcal{D} = \{1, 2, \ldots, d\}$ then

$$\frac{1}{n}\mathbf{Z}^T\mathbf{Z} = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

where Σ_{11} is the $d \times d$ covariance matrix of the relevant variables. Zou (2006) prove that the adaptive LASSO has the following properties:

1. **Near-minimax optimality**

   In the orthogonal design, the adaptive LASSO thresholding function (shown in Section 4.2.3) attains the near minimax risk. If $\lambda^* = (2\ln n)^{(1+\zeta)/2}$ and $\sigma^2 = 1$ then

   $$R_{AL}\left(\underline{\hat{\alpha}}^*, \underline{\alpha}\right) \leqslant \left(2\ln p + 5 + 4\zeta^{-1}\right)\left(R\left(\text{ideal}\right) + 1\Big/\sqrt{4\pi\ln\left(n\right)}\right).$$

2. **Oracle properties**

   Suppose that $\hat{\alpha}_j^{init}$ is a $\sqrt{n}$-consistent estimator, $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\zeta-1)/2} \to \infty$. Alternatively, the condition can be relaxed so that $\hat{\alpha}_j^{init}$ is an a_n-consistent estimator, $\lambda_n = o\left(\sqrt{n}\right)$ and $\alpha_n^{\zeta}\lambda_n n \to \infty$. Then the adaptive LASSO is

   (a) consistent for variable selection, $\lim_{n\to\infty} P\left(\widehat{\mathcal{D}}_n = \mathcal{D}\right) = 1$.

   (b) asymptotically normal, $\sqrt{n}\left(\underline{\hat{\alpha}}_{\widehat{\mathcal{D}}}^{AL} - \underline{\alpha}_{\mathcal{D}}\right) \xrightarrow{d} N\left(\underline{0}, \sigma^2\Sigma_{11}^{-1}\right)$.

See Huang *et al.* (2008) and Lin *et al.* (2009) for properties of the adaptive LASSO in high dimensional settings.

The adaptive LASSO can be stated as a LASSO problem and solved using the LAR-LASSO algorithm at no extra cost.

Algorithm 4.2.2 ***Adaptive LASSO***

1. *Let* $\mathbf{Z}^*$ *be the matrix with columns* $\mathbf{z}_j^* = \mathbf{z}_j/\hat{w}_j$ *for* $j = 1, 2, \ldots, p$
2. *Solve the* LASSO *problem* $\underline{\hat{\alpha}}^* = \underset{\underline{\alpha}}{\arg\min}\ \|\mathbf{v} - \mathbf{Z}^*\underline{\alpha}\|^2 + \lambda\|\underline{\alpha}\|_1$
3. *Compute the adaptive* LASSO *solution* $\hat{\alpha}_j^{AL} = \hat{\alpha}_j^*/\hat{w}_j$ *for* $j = 1, 2, \ldots, p$.

For a given value of ζ, the LAR-LASSO algorithm fits the entire path and cross-validation can be used to select the best value of ζ. This can be repeated over a grid of ζ values and the one giving the minimum CV error is used. Zou (2006) recommends using the LSE to calculate the adaptive weights. In the presence of collinearity or if the design matrix does not have full rank, he recommends using the best fitting ridge regression estimate.

Qian & Yang (2013) propose a method called standard error adjusted LASSO (SEA-LASSO) to improve the performance when using the LSE as an initial estimate in the presence of collinearity. Let $s_1, s_2, \ldots, s_p$ be the standard errors of the LSEs, then the weights are calculated as $w_j = s_j/\hat{\alpha}_j^{\zeta}$. They note that when Σ is ill-conditioned, the LSE can be poor, with true nonzero parameters having estimates far from zero. Since these estimates are not as small as they should be, the weights in the adaptive LASSO will not be sufficiently large and the estimates will not be penalized enough. Such estimates are unstable and usually present with inflated variances due to collinearity. Thus they suggest that multiplying the weights by the standard errors of the estimates will improve the regularization, if the model is sparse. Furthermore, they show that SEA-LASSO has the same theoretical properties of the adaptive LASSO. If the model is not sparse, they propose a two-stage method called NSEA-LASSO. These methods are recommended when the condition number is large, in particular $\kappa_2(\Sigma) \geqslant 10$.

Bühlmann & van de Geer (2011:25) suggest using the LASSO estimate as an initial estimate when $p \gg n$ and calculating the weights with $\zeta = 1$. The LASSO is applied in the first and second stage, each time selecting the shrinkage parameter λ for optimal prediction.

Algorithm 4.2.3 ***Adaptive LASSO with LASSO initial estimate***

1. *Calculate the* LASSO *solution* $\underline{\hat{\alpha}}^L$ *selecting* λ *by* CV *for optimal prediction and let* $\widehat{\mathcal{D}} = \left\{j : \hat{\alpha}_j^L \neq 0\right\}$
2. *Calculate the adaptive* LASSO *solution with* $\hat{\alpha}_{\widehat{\mathcal{D}}^c}^{AL} = 0$ *and*

$$\underline{\hat{\alpha}}_{\widehat{\mathcal{D}}}^{AL} = \underset{\underline{\alpha}_{\widehat{\mathcal{D}}}}{\arg\min} \left\|\mathbf{v} - \mathbf{Z}_{\widehat{\mathcal{D}}}\underline{\alpha}_{\widehat{\mathcal{D}}}\right\|^2 + \lambda^* \sum_{j\in\widehat{\mathcal{D}}} \frac{|\alpha_j|}{\left|\hat{\alpha}_j^L\right|},$$

again selecting λ^* *by* CV *for optimal prediction*.

Applying the LASSO again in a second stage can simultaneously reduce the number of irrelevant variables included in the first stage and estimate the nonzero parameters with less bias. The computational burden of cross-validation is eased by sequentially optimizing over a single parameter instead of optimizing over two parameters simultaneously. If LAR-LASSO is used in both steps, the computational cost is of order $O\left(np\min(n,p)^2\right)$ when simultaneously optimizing over two parameters. In contrast, the cost is of order $O(2np\min(n,p))$ when optimizing twice over one parameter.

Bühlmann & van de Geer (2011:30) discuss a further generalization of the idea by applying the LASSO in multiple stages.

Algorithm 4.2.4 ***Multi-stage adaptive LASSO***

1. *Initialize the weights* $\hat{w}_j^{(0)} = 1$ *for* $j = 1, 2, \ldots, p$ *and initialize the set of nonzero parameters* $\widehat{\mathcal{D}}_{(0)} = \{j : 1, 2, \ldots, p\}$

2. *For* $k = 1, 2, \ldots, K$*:*

   (a) *Calculate the multi-stage adaptive LASSO estimate as* $\underline{\hat{\alpha}}^{ML}_{\widehat{\mathcal{D}}^c_{(k-1)}} = 0$ *and*

$$\underline{\hat{\alpha}}^{ML}_{\widehat{\mathcal{D}}_{(k-1)}} = \underset{\underline{\alpha}_{\widehat{\mathcal{D}}_{(k-1)}}}{\arg\min} \left\| \mathbf{v} - \mathbf{Z}_{\widehat{\mathcal{D}}_{(k-1)}} \underline{\alpha}_{\widehat{\mathcal{D}}_{(k-1)}} \right\|^2 + \lambda_{(k)} \sum_{j \in \widehat{\mathcal{D}}_{(k-1)}} \hat{w}_j^{(k-1)} \left| \alpha_j \right|, \tag{4.2.7}$$

   selecting $\lambda_{(k)}$ *for optimal prediction.*

   (b) *Update the set of nonzero parameters* $\widehat{\mathcal{D}}_{(k)} = \left\{ j : \hat{\alpha}_j^{ML} \neq 0 \right\}$

   (c) *Update the weights* $\hat{w}_j^{(k)} = 1/\left| \hat{\alpha}_j^{ML} \right|$ *for* $j \in \widehat{\mathcal{D}}_{(k)}$

The sparsity increases at each step of the multi-stage LASSO. Since the shrinkage parameters are selected sequentially, the computational complexity is of the order $O\left(Knp \min\left(n, p\right)\right)$ if LAR-LASSO is used at each step. Bühlmann & van de Geer (2011:32-33) state that the multi-stage adaptive LASSO is an approximation to the concave bridge estimates with $0 < \gamma < 1$. Furthermore, they relate the computation of the multi-stage LASSO to that of SCAD using iterative local linear approximations.

Nonnegative Garrote

Around the same time that the LASSO was proposed, Breiman (1995) proposed the nonnegative garrote as a stable, scale invariant method which could be used as an alternative to variable selection methods and ridge regression. This method does not standardize the data but instead scales each least squares estimate directly by a nonnegative constant. Let $\underline{\hat{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T$ be the least squares estimate. The nonnegative garrote estimate is given by

$$\underline{\hat{\beta}}^{NG} = \left(c_1\hat{\beta}_1, c_2\hat{\beta}_2, , \ldots, c_p\hat{\beta}_p\right)^T,$$

where $\underline{c} = \left(c_1, c_2, \ldots, c_p\right)^T$ is the solution to the penalized regression

$$\underset{\mathbf{c}}{\arg\min} \left\| \mathbf{v} - \sum_{j=1}^{p} c_j \hat{\beta}_j \mathbf{x}_j \right\|^2 + \lambda \sum_{j=1}^{p} c_j, \tag{4.2.8}$$

where $\lambda \geqslant 0$ is the shrinkage parameter and $c_j \geqslant 0$ for $j = 1, 2, \ldots, p$. A different amount of shrinkage is placed on each least squares estimate. Small least squares estimates, which are possibly redundant, are shrunk more than large ones. Like the LASSO, the problem can be written as a quadratic programming problem (see Seber & Lee (2003:426)). Breiman (1995) adapted the NNLS algorithm of Lawson & Hanson (1974:158-173) to solve the problem. He recognized that NNLS handled the nonnegative constraint $\underline{c} \geqslant \underline{0}$ and incorporated a barrier method to include the constraint $\sum_{j=1}^{p} c_j \leqslant \tau$. A drawback of the nonnegative garrote is its dependence on the LSEs. In situations where the LSEs perform poorly, the nonnegative garrote will likely suffer the same consequences. Furthermore, the nonnegative garrote cannot be used when $p > n$ because of its reliance on the LSEs. Yuan & Lin (2007) propose generalizing the problem to use the estimates from other methods, such as ridge regression, the LASSO or the elastic net, as initial estimates in the nonnegative garrote. They show that, similar to the LASSO, the entire path of the nonnegative garrote solution is piecewise-linear. Furthermore, they provide an efficient algorithm for computing this path, which alleviates any problems that the procedure might have regarding computational cost.

Zou (2006) shows that the adaptive LASSO is similar to the nonnegative garrote when $\gamma = 1$ and $\hat{\alpha}_j^{init}$ are the least squares estimates. In this case, the adaptive LASSO minimizes

$$\begin{aligned}
&\left\|\mathbf{v} - \sum \alpha_j \mathbf{z}_j\right\|^2 + \lambda \sum \hat{w}_j \left|\alpha_j\right| \\
&= \left\|\mathbf{v} - \sum \alpha_j \mathbf{z}_j\right\|^2 + \lambda \sum \left|\alpha_j\right| / \left|\hat{\alpha}_j\right| \\
&= \left\|\mathbf{v} - \sum c_j \hat{\alpha}_j \mathbf{z}_j\right\|^2 + \lambda \sum \left|c_j\right|,
\end{aligned}$$

which is similar to the nonnegative garrote since $\hat{\alpha}_j^{NG} = c_j \hat{\alpha}_j$. Zou (2006) states that when adding the constraint $\alpha_j \hat{\alpha}_j \geqslant 0$, the problems are equivalent.

4.2.3 Orthogonal Design

Using Equation (4.1.5), closed form estimates can be obtained for the two-stage methods under an orthogonal design. These estimates are called thresholding functions and are summarized in Table 4.2.1, along with penalty functions for each method.

The penalty and thresholding functions for bridge estimates are depicted in Figure 4.2.1. When $\gamma \geqslant 2$, the penalty function is convex and the thresholding function shrinks parameters proportional to their size. As γ increases beyond 2, large parameters are biased more towards zero and small parameters not

Method	Penalty Function	Thresholding Function
Bridge Estimates	$\lambda \sum \lvert\alpha_j\rvert^{\gamma}$	$\hat{\alpha}_j - \text{sign}\left(\alpha_j\right) \lambda\gamma \lvert\alpha_j\rvert^{\gamma-1}$
Relaxed LASSO	$\lambda\phi \sum \lvert\alpha_j\rvert$	$\text{sign}\left(\hat{\alpha}_j\right)\left(\lvert\hat{\alpha}_j\rvert - \phi\lambda\right)\delta\left(\lvert\hat{\alpha}_j\rvert \geqslant \lambda\right)$
Adaptive LASSO	$\lambda \sum \lvert\alpha_j\rvert \Big/ \lvert\hat{\alpha}_j\rvert^{\zeta}$	$\text{sign}\left(\hat{\alpha}_j\right)\left(\lvert\hat{\alpha}_j\rvert - \lambda \Big/ \lvert\hat{\alpha}_j\rvert^{\zeta}\right)_+$

Table (4.2.1) Penalty and thresholding functions for bridge estimates and two-stage methods

shrunk. When $1 \leqslant \gamma < 2$, the penalty function is still convex but we see the discontinuity at zero for $\gamma = 1$. The thresholding function shows that large parameters are shrunk slightly less and small parameters are only set to zero when $\gamma = 1$. When $0 < \gamma < 1$, the penalty function is concave and the shrinkage of the thresholding function is inversely proportional to the size of the parameters. Here, with $\lambda = 4$ and $\gamma = 0.25$ or $\gamma = 0.5$, large parameters remain fairly untouched by the shrinkage.

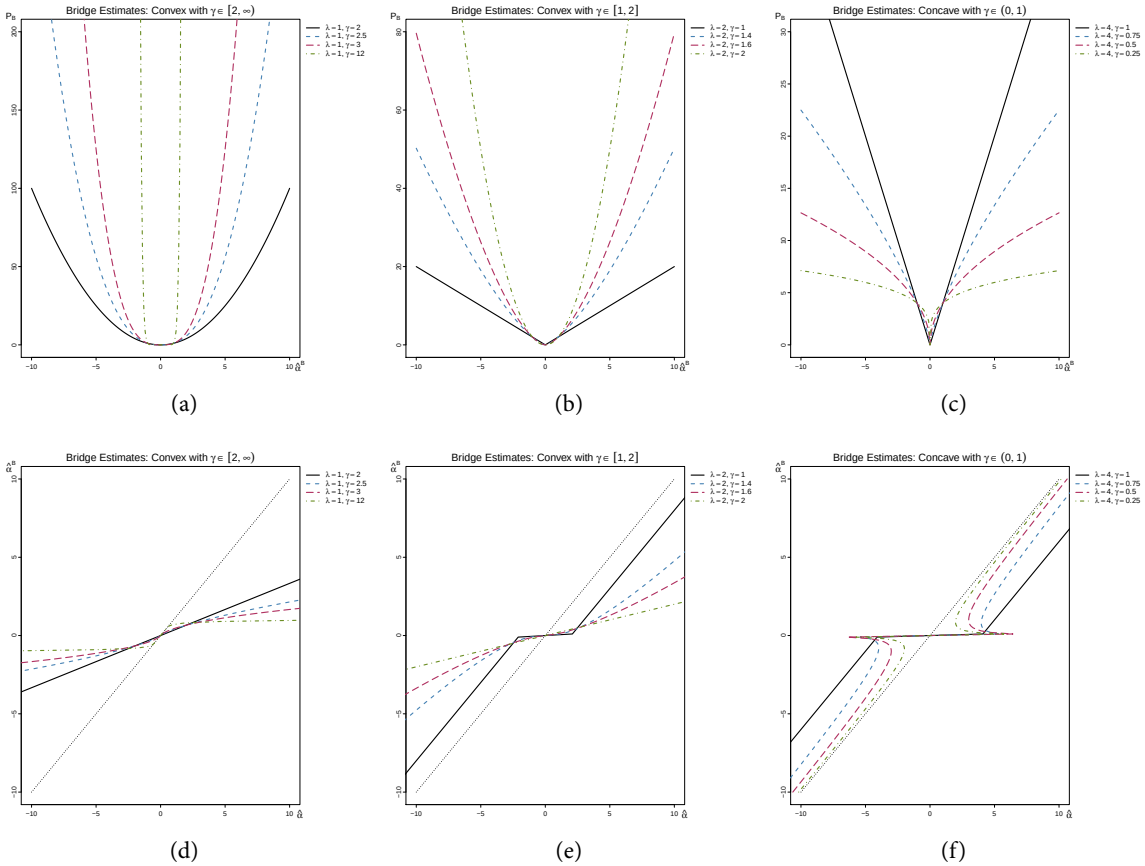


Figure (4.2.1) Penalty functions (a)-(c) and thresholding functions (d)-(f) for bridge estimates at various values of λ and γ. The penalty functions are convex when $\gamma \geqslant 1$ and discontinuous at zero when $\gamma \leqslant 1$. When the penalty function is discontinuous, the thresholding function sets parameters to zero. The shrinkage of nonzero estimates is proportional to their size when $\gamma > 1$, inversely proportional when $\gamma < 1$ and constant when $\gamma = 1$.

The functions for the relaxed LASSO and adaptive LASSO are shown in Figure 4.2.2. The adaptive LASSO penalty appears to be convex and when $\xi = 0.5$, it is very similar to the bridge penalty with $\gamma = 0.5$. The relaxed LASSO is convex for all $\phi \in (0,1)$. The thresholding function of the relaxed LASSO is similar to the LASSO threshold but it is clear that less shrinking is applied. The adaptive LASSO threshold seems to mimic the behaviour of concave bridge penalties quite well, large parameters are shrunk very little and small parameters are shrunk slightly less than the bridge penalties.

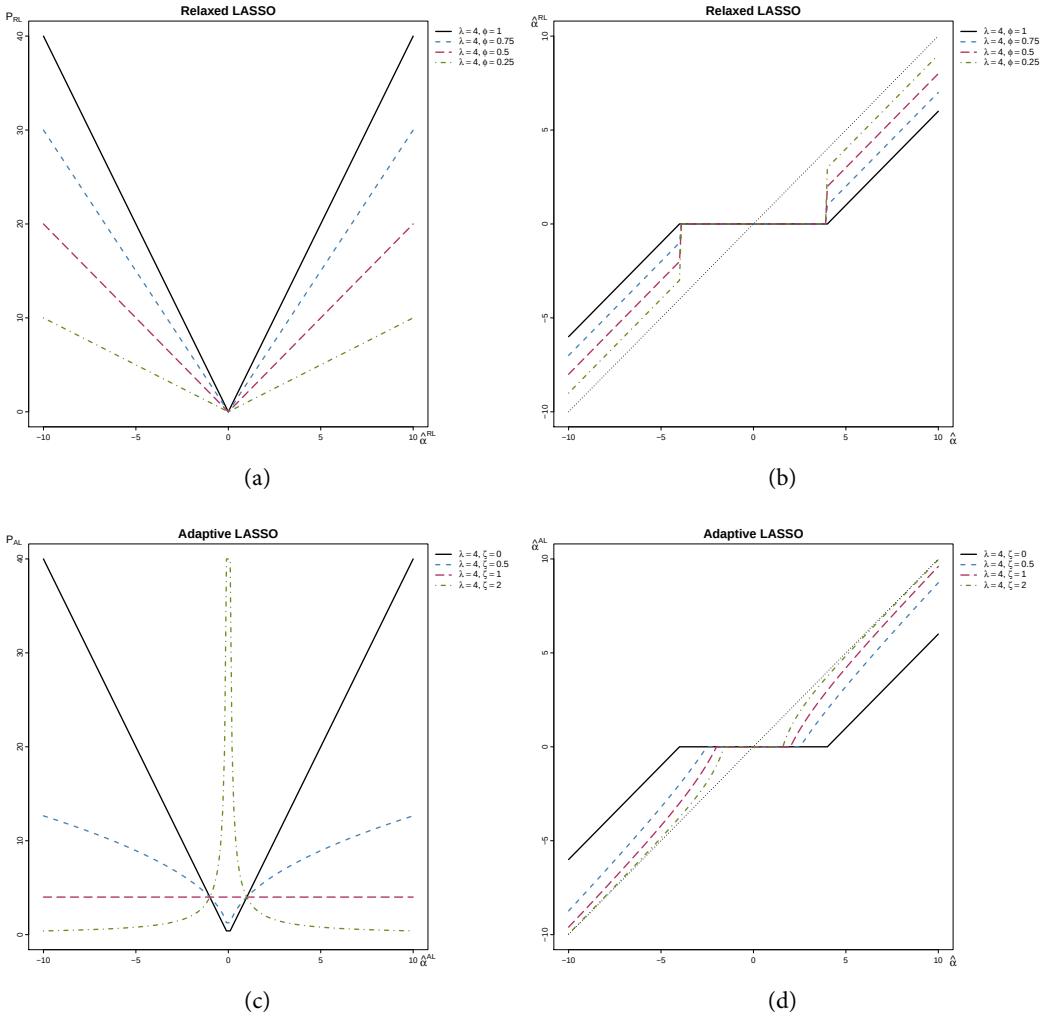


Figure (4.2.2) Penalty and thresholding functions for (a)-(b) relaxed LASSO and (c)-(d) adaptive LASSO. The ordinary LASSO is depicted by the solid black line. The shrinkage mimics the concave bridge penalties in Figure 4.2.1 and the bias is reduced.

4.2.4 Other Methods for Controlling Bias

The LAR-OLS hybrid is discussed by Efron *et al.* (2004). In this procedure the LAR algorithm is used solely for selection purposes. The solution path is calculated and after identifying the optimal model, least squares is used to estimate it. Bühlmann & van de Geer (2011:33) propose a similar strategy to the LAR-OLS hybrid method. Using CV to select the prediction optimal value of λ, the LASSO estimate $\hat{\underline{\alpha}}^R$ is obtained. A thresholding rule is then applied to select all estimates that are greater than some $\iota > 0$,

$$\hat{\alpha}_j^{thresh} = \hat{\alpha}_j^R \delta\left(\left|\hat{\alpha}_j^R\right| > \iota\right).$$

Least squares is then used to fit the model using the subset of variables $\widehat{\mathcal{D}} = \left\{j : \hat{\alpha}_j^{thresh} \neq 0\right\}$. Cross-validation is used to select the best thresholding parameter ι by calculating the LSE for different $\widehat{\mathcal{D}}$ which result from varying ι. While the method is similar to LAR-OLS hybrid, it includes an addition thresholding stage which improves the performance. In fact, Bühlmann & van de Geer (2011:210-215) show that thresholding the LASSO has similar properties to the adaptive LASSO for prediction and variable selection.

Zhang & Huang (2008) discuss using an initial estimate which is ℓ_∞-consistent and then applying either the adaptive LASSO, the nonnegative garrote or hard thresholding to obtain the final estimate. They show that these estimates are consistent for variable selection and estimation even in ill-conditioned designs.

4.3 Modified LASSO Methods

The LASSO must be modified in order to incorporate different structures among the predictor variables. Clarke *et al.* (2009:606) and Hastie *et al.* (2009:661) point out that when a group of highly correlated variables is present, the LASSO tends to randomly select one of them in order to deal with collinearity. Tibshirani *et al.* (2005) proposed the fused LASSO to handle predictors that can be ordered in a meaningful way. A number of modifications have been developed to handle groups of variables. The group LASSO methods are designed to overcome this by considering whole groups for inclusion in the model instead of individual variables. Some of these methods are discussed briefly in Section 4.3.2. The group LASSO developed by Yuan & Lin (2006) using a quadratic norm in the penalty function to produce sparsity between groups and not within groups. Zhao *et al.* (2009) generalize the method with CAP, where groups may be overlapped and can be specified in such a way as to preserve hierarchy. Huang *et al.* (2009) proposed

another generalization called the group bridge which performs selection at the group level and within groups.

4.3.1 Fused LASSO

Tibshirani *et al.* (2005) proposed the fused LASSO for ordered predictors. They provide two examples when this situation occurs. The first is protein mass spectroscopy in which the intensity is observed for many time-of-flight values j for each blood serum i. Here the predictors are ordered *a priori* by time-of-flight values. The second example is gene expression data from a microarray. In this case the ordering of the variables is unknown and must be estimated from the data. Correlated genes can be placed next to each other after estimating their order using, for example, a clustering algorithm.

The fused LASSO estimate is given by

$$\hat{\underline{\alpha}}^{FL} = \arg\min_{\underline{\alpha}} \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 \text{ subject to } \sum_{j=1}^{p} |\alpha_j| \leqslant t \text{ and } \sum_{j=2}^{p} |\alpha_j - \alpha_{j-1}| \leqslant u, \tag{4.3.1}$$

where $t \geqslant 0$ and $u \geqslant 0$ are tuning parameters. The ℓ_1-norm imposed on the difference between adjacent parameters encourages nearby variables to have similar coefficients, while the LASSO penalty promotes sparsity between the coefficients. When the order of the predictors is unknown, the constraint on the differences can be modified as

$$\sum_{j} |\alpha_j - \alpha_{k(j)}| \leqslant u,$$

where $k(j)$ is the the variable closest to the j-th variable in terms of some similarity measure such as a distance function or correlation index. The problem can be written in the Lagrangian form

$$\hat{\underline{\alpha}}^{FL} = \arg\min_{\underline{\alpha}} \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 + \lambda \left[\|\underline{\alpha}\| + \psi \sum_{j=2}^{p} |\alpha_j - \alpha_{j-1}| \right].$$

Two approaches are suggested for computing the solution. First, quadratic programming can be used by including nonnegative variables as in (4.1.23). Let $\theta_1 = 1$ and $\theta_j = \alpha_j - \alpha_{j-1}$ for $j > 1$, then the nonnegative variables $\alpha_j^+, \alpha_j^- \geqslant 0$ and $\theta_j^+, \theta_j^- \geqslant 0$ are introduced such that $\alpha_j = \alpha_j^+ - \alpha_j^-$ and $\theta_j = \theta_j^+ - \theta_j^-$. The

problem is then equivalent to

$$\begin{aligned}
\text{minimize} \quad & \tfrac{1}{2}\left\|\mathbf{v}-\mathbf{Z}\underline{\alpha}\right\|^2 \\
\text{subject to} \quad & \alpha_j = \alpha_j^+ - \alpha_j^- \text{ for } j = 1, 2, \ldots, p \\
& \theta_j = \theta_j^+ - \theta_j^- \text{ for } j = 1, 2, \ldots, p \\
& \alpha_j^+, \alpha_j^-, \theta_j^+, \theta_j^- \geqslant 0, \text{for } j = 1, 2, \ldots, p \\
& \textstyle\sum_{j=1}^{p}\left(\alpha_j^+ + \alpha_j^-\right) \leqslant t \\
& \textstyle\sum_{j=2}^{p}\left(\theta_j^+ + \theta_j^-\right) \leqslant u
\end{aligned}$$

which includes $6p$ constraints and p variables. Let $\mathbf{L}$ be a $p \times p$ matrix with elements $l_{ii} = 1$, $l_{i+1,i} = -1$ and $l_{ij} = 0$, then $\underline{\theta} = \mathbf{L}\underline{\alpha}$. The equality constraints can then be written as

$$\begin{bmatrix} \mathbf{L} & \mathbf{0} & \mathbf{0} & -\mathbf{I}_p & \mathbf{I}_p \\ \mathbf{I}_p & -\mathbf{I}_p & \mathbf{I}_p & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \underline{\alpha} \\ \underline{\alpha}^+ \\ \underline{\alpha}^- \\ \underline{\theta}^+ \\ \underline{\theta}^- \end{bmatrix} = \begin{bmatrix} \underline{0} \\ \underline{0} \end{bmatrix}$$

and the inequality constraints can be written as

$$\begin{bmatrix} -\mathbf{I}_p & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_p & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{I}_p \\ \mathbf{1}_p^T & \mathbf{1}_p^T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_0^T & \mathbf{1}_0^T \end{bmatrix} \begin{bmatrix} \underline{\alpha}^+ \\ \underline{\alpha}^- \\ \underline{\theta}^+ \\ \underline{\theta}^- \end{bmatrix} = \begin{bmatrix} \underline{0} \\ \underline{0} \\ 0 \\ 0 \\ t \\ u \end{bmatrix}$$

where $\mathbf{1}_0^T$ is a vector of 1s with the first element set to zero since $\theta_1 = 1$. In the second approach, they set $\mathbf{Z}^* = \mathbf{Z}\mathbf{L}^{-1}$, fit the LASSO model using LAR-LASSO to obtain $\hat{\underline{\alpha}}^*$ and then compute $\hat{\underline{\alpha}}^{FL} = \hat{\underline{\alpha}}^*\mathbf{L}^{-1}$.

Tibshirani & Taylor (2011) extended the idea of enforcing structural or geometrical constraints and proposed the generalized LASSO,

$$\hat{\underline{\alpha}}^{DL} = \underset{\underline{\alpha}}{\arg\min} \left\|\mathbf{v}-\mathbf{Z}\underline{\alpha}\right\|^2 + \lambda \left\|\mathbf{D}\underline{\alpha}\right\|_1. \tag{4.3.2}$$

Some special cases of the generalized LASSO are listed here.

- LASSO: $\mathbf{D} = \mathbf{I}$

- 1-dimensional fused LASSO: $\mathbf{X} = \mathbf{I}$,

$$\underset{(n-1)\times n}{\mathbf{D}} = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots \\ 0 & -1 & 1 & 0 & \cdots \\ 0 & 0 & -1 & 1 & \cdots \\ \vdots & & & & \end{bmatrix}$$

- 2-dimensional fused LASSO: $\mathbf{X} = \mathbf{I}$,each row of $\mathbf{D}$ is given by $\mathbf{D}_i = (0, 0, \ldots, -1, \ldots, 1, \ldots 0,)$ The 2d-fused LASSO penalty is $\lambda \sum_{j,k\epsilon\mathcal{D}} |\alpha_j - \alpha_k|$. Hence the -1 entries in $\mathbf{D}_i$ correspond to the j-th variable and the 1 entries correspond to the k-th variable.

- Linear trend filtering: $\mathbf{X} = \mathbf{I}$,

$$\underset{(n-1)\times n}{\mathbf{D}} = \begin{bmatrix} -1 & 2 & -1 & 0 & \cdots \\ 0 & -1 & 2 & -1 & \cdots \\ 0 & 0 & -1 & 2 & \cdots \\ \vdots & & & & \end{bmatrix}$$

- Wavelet smoothing: $\mathbf{X} = \mathbf{I}$, $\mathbf{D} = \mathbf{W}^T$ where the columns of $\mathbf{W}$ are orthogonal wavelet basis.

4.3.2 Group LASSO

Group LASSO methods are an extension of the LASSO method to select known groups of variables called factors. Examples where factors would be of interest include dummy variables, polynomial functions, nonparametric basis functions and genes in the same molecular pathway. Categorical variables can be included in the linear model by deriving dummy variables which correspond to the levels of the variable. If the categorical variable is a good predictor, we would usually want to include all levels in the model. When there is no significant difference between two levels, the usual approach would be to combine levels together rather than exclude them. To capture any curvature in the data, polynomial terms and interactions can be included in the linear model. If a polynomial term or interaction is included in the model, we would usually include any associated lower order terms to maintain hierarchy and facilitate interpretation. Preserving hierarchical order helps to prevent shifts in the data from reparameterizing the model. Similarly, we might like to include groups of nonparametric basis functions or groups of genes. The levels or terms in these factors are often highly correlated. The variables in each group have a combined effect on the response variable and relate to one measured variable, so we would like to either include or exclude the whole group instead of the individual derived variables.

The group LASSO was first introduced by Bakin (1999) and later developed by Yuan & Lin (2006). Consider fitting a linear model including g groups. Let $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_g$ denote the subset of variables in each group and let the number of variables in the k-th group be $|\mathcal{G}_k| = p_k$, where $p_1 + p_2 + \cdots + p_g = p$. Then we have the linear model

$$\mathbf{v} = \sum_{k=1}^{g} \mathbf{Z}_{\mathcal{G}_k} \underline{\alpha}_{\mathcal{G}_k} + \varepsilon. \tag{4.3.3}$$

The usual linear model where individual variables are considered occurs when $p_1 = p_2 = \cdots = p_g = 1$. For the group LASSO, the response variable and predictors are all centered to have mean 0 and each group is orthonormalized so that $\mathbf{Z}_{\mathcal{G}_k}^T \mathbf{Z}_{\mathcal{G}_k} = \mathbf{I}_{p_k}$ for $k = 1, 2, \ldots, g$. Applying the LASSO directly to (4.3.3) is problematic. The selected model often includes too many groups since it is based on the effects that individual variables have on the response instead of the effects that the groups have. The model will also depend on how the groups are othonormalized and may include different subsets of factors if any of the groups are reparameterized. The group LASSO extends the LASSO method to handle effects at the group level and is invariable to the parameterization of the groups.Bakin (1999) defines the group LASSO estimate as

$$\hat{\underline{\alpha}}^{GL} = \arg\min_{\underline{\alpha}} \left\| \mathbf{v} - \sum_{k=1}^{g} \mathbf{Z}_{\mathcal{G}_k} \underline{\alpha}_{\mathcal{G}_k} \right\|^2 + \lambda \sum_{k=1}^{g} \left\| \underline{\alpha}_{\mathcal{G}_k} \right\|_{Q_k}, \tag{4.3.4}$$

where $\lambda \geqslant 0$ is the shrinkage parameter and $\left\| \underline{\alpha}_{\mathcal{G}_k} \right\|_{Q_k} = \left(\underline{\alpha}_{\mathcal{G}_k}^T \mathbf{Q}_k \underline{\alpha}_{\mathcal{G}_k} \right)^{\frac{1}{2}} = \left\| \mathbf{Q}_k^{\frac{1}{2}} \underline{\alpha}_{\mathcal{G}_k} \right\|_2$ is the Q_k-quadratic norm where $\mathbf{Q}_k$ is a symmetric positive definite matrix (see Boyd & Vandenberghe (2004:635) for more about quadratic norms). Yuan & Lin (2006) choose $\mathbf{Q}_j = p_k \mathbf{I}_{p_k}$ so that the penalty function is

$$P_{GL}(\underline{\alpha}) = \lambda \sum_{k=1}^{g} p_k \left\| \underline{\alpha}_{\mathcal{G}_k} \right\|_2 = \lambda \sum_{k=1}^{g} p_k \sqrt{\underline{\alpha}_{\mathcal{G}_k}^T \underline{\alpha}_{\mathcal{G}_k}}. \tag{4.3.5}$$

When $p_k = 1$, the penalty function is the same as the LASSO penalty since $\left\| \alpha_{\mathcal{G}_k} \right\| = \sqrt{\alpha_{\mathcal{G}_k}^2} = \left| \alpha_{\mathcal{G}_k} \right|$. When $p_k > 1$, the penalty function is similar to the ridge regression penalty. The group LASSO promotes sparsity between groups but not within groups. Yuan & Lin (2006) extends the LAR algorithm to handle groups of variables but they show that the group LASSO is not necessarily piecewise linear. Furthermore, they modify the LAR algorithm to calculate the group estimator for the nonnegative garrote, which guaranteed to be piecewise linear. To solve (4.3.4), they use an extension of the shooting algorithm by Fu (1998).

Zhao *et al.* (2009) proposed CAP as a more generalized method to incorporate grouped variables into

the design. The CAP estimator is given

$$\underline{\hat{\alpha}}^{C} = \arg\min_{\underline{\alpha}} \left\| \mathbf{v} - \sum_{k=1}^{g} \mathbf{Z}_{\mathcal{G}_k} \underline{\alpha}_{\mathcal{G}_k} \right\|^2 + \lambda \sum_{k=1}^{g} \left| \left\| \underline{\alpha}_{\mathcal{G}_k} \right\|_{q_k} \right|^{\gamma}, \tag{4.3.6}$$

where $\lambda \geqslant 0$ is the shrinkage parameter, $\gamma \geqslant 1$ and $q_k \geqslant 1$ for all $k = 1, 2, \ldots, g$ control the grouping. The ℓ_γ-norm is the overall norm and controls the relationship between the groups, whereas the ℓ_{q_k}-norm controls the relationship between the variables in the k-th group. The CAP penalty includes a composition of ℓ_q-norms,

$$P_C(\underline{\alpha}) = \lambda \left[\ell_\gamma \left(\ell_{q_k} \left(\underline{\alpha}_{\mathcal{G}_k} \right) \right) \right]^{\gamma} = \lambda \left\| \left\| \underline{\alpha}_{\mathcal{G}_k} \right\|_{q_k} \right\|_{\gamma}^{\gamma} \tag{4.3.7}$$

and can almost be seen as a two-stage penalty, first applying shrinkage between the variables in each group and then applying shrinkage among the groups. The flexible penalty allows for different penalties on each group so that the norm most appropriate for the structure in a specific group can applied just to that group. When $\gamma = 1$ and $q_k = 2$ for all k, the group LASSO is obtained. When $\gamma = 1$, the bridge parameter is not strictly necessary since norms are always positive. However, viewing the penalty in this light does help to interpret why sparsity is promoted at the group level - the LASSO is being applied between groups. The bridge parameter γ controls the directions in which we believe the true parameters are aligned with respect to the coordinate axes. Table 4.3.1 summarizes the directions that are favoured for different intervals of γ. Sparse solutions only occur for $\gamma \leqslant 1$, when the estimates are likely to lie on the axes. The estimates move further away from the axes as γ increases and their sizes get closer together. The ℓ_q-norm balls in Figure 4.3.1 illustrate the concept visually - the point of intersection with RSS is mostly likely to occur wherever there are sharp points or edges on the norm ball. The norm balls are for values of γ, from left to right, $\gamma = 0.5, 1, 1.5, 2, 3, \infty$.

γ **Interval**	**Favoured Direction**
$[0, 1]$	on the axes
$(1, 2)$	close to the axes
2	none
$(2, \infty)$	along diagonals

Table (4.3.1) Directions favoured by bridge estimates. Estimates are set to zero when they lie on the axes and are equally sized when they lie along the diagonals

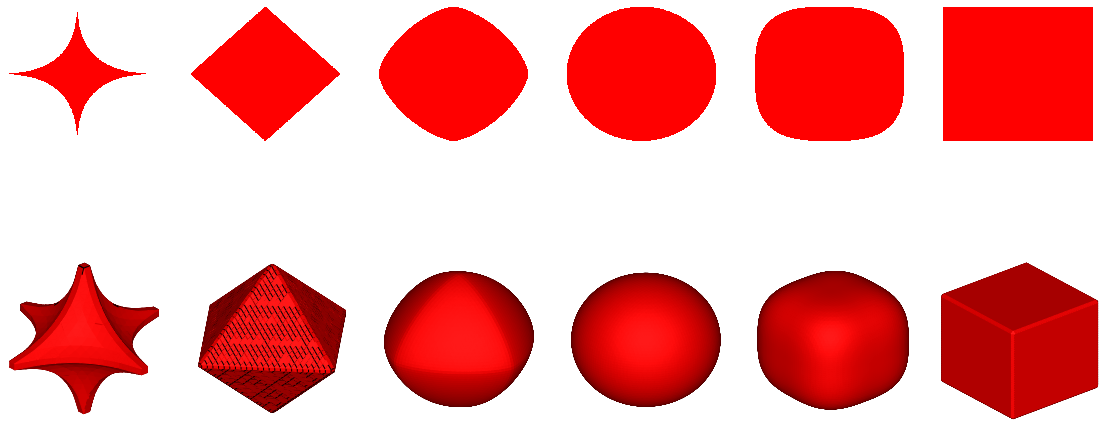


Figure (4.3.1) Norm balls for bridge estimates in $\mathbb{R}^2$ (top) and $\mathbb{R}^3$ (bottom). From left to right, $\gamma = 0.5$, 1 (LASSO), 1.5, 2 (ridge), 3 and ∞. The first two figures on the left have protruding points on the axes which encourage sparsity among the estimates. The last two figures on the right have protruding points on the diagonals which encourage equality of estimates.

Zhao *et al.* (2009) consider $\gamma = 1$ and $q_k > 1$ for all k. Hence, sparsity is encouraged between groups but not within groups so that complete groups are considered for selection. In particular, they suggest setting $q_k = \infty$ for all k, thus promoting equally sized coefficients within groups. It is noted that the different group sizes are irrelevant and have no effect on selection or estimation. Zhao *et al.* (2009) also outline a way of performing hierarchical selection by defining the groups so that they are nested and overlapping. They provide path algorithms to solve the general case $q_k > 1$, the case when $q_k = \infty$ (which is more efficient), and the overlapping group case.

Huang *et al.* (2009) proposed a similar idea to CAP, called the group bridge. However, they allow for selection not only between groups but also within groups simultaneously. This is called bi-level selection, if a group is selected, variables within that group can be discarded. The group bridge is given by

$$\underline{\hat{\alpha}}^{GB} = \arg\min_{\underline{\alpha}} \left\| \mathbf{v} - \sum_{k=1}^{g} \mathbf{Z}_{\mathcal{G}_k} \underline{\alpha}_{\mathcal{G}_k} \right\|^2 + \lambda \sum_{k=1}^{g} c_k \left\| \underline{\alpha}_{\mathcal{G}_k} \right\|_1^{\gamma}, \tag{4.3.8}$$

where $\lambda \geqslant 0$ is the shrinkage parameter. The c_j are constants that can be used to adjust for different sizes of the groups and a suggested value is $c_j \propto p_k^{1-\gamma}$. The bridge parameter $\gamma \in (0,1)$ is applied to the ℓ_1 norm

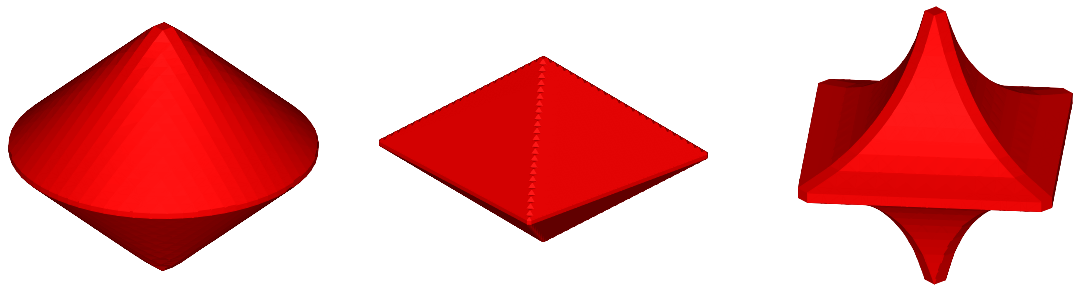
of each group, resulting in the concave penalty function

$$P_{GB}(\underline{\alpha}) = \lambda \sum_{k=1}^{g} c_k \left\| \underline{\alpha}_{\mathcal{G}_k} \right\|_1^{\gamma}. \tag{4.3.9}$$

The penalty function is the ordinary bridge penalty when $p_k = 1$. When $\gamma = 1$, the LASSO penalty is obtained. Here, the LASSO is used as the within group penalty and the bridge parameter is used as the overall penalty.

4.3.3 Geometry

The norm balls shown in Figure 4.3.2 clarify how the group norms and overall norms affect the parameters. The parameters on the x and y axes correspond to a group of two predictors, while the parameter on the z axis corresponds to an individual predictor, or a group of size 1. Thus, in these plots, the group norms act horizontally within the group and the overall norm acts vertically between the groups.



(a) (b) (c)

Figure (4.3.2) Norm balls for (a) group LASSO, (b) CAP, and (c) group bridge penalties in $\mathbb{R}^3$. The parameters on the x and y axes are in a group of size two and the parameter on the z axis corresponds to an individual variable (group of size 1). Each norm ball promotes sparsity between groups. The group LASSO includes entire groups, CAP encourages equality of parameters within each group and the group bridge allows for sparsity within groups.

In each case, sparsity is induced between groups - by the sharp points of the ℓ_1 norm for CAP and group LASSO, or by the concave ℓ_γ-norm with $\gamma \in (0,1)$ for group bridge. However, only the group bridge penalty allows for sparsity within groups where the ℓ_1 norm has its sharp points on the axes. Although the CAP penalty has sharp points on the xy-plane, these are the corners of the ℓ_∞-norm lying along the diagonals of each quadrant and not on the axes. Rather than setting any estimates within the group to zero, they are encouraged to be equally sized. The group LASSO has the curved ℓ_2 norm acting within

groups, where no direction is favoured and the possibility of sparsity is very low.

The fused LASSO penalty is a combination of two ℓ_1 penalties, one on the parameters and one on the difference between adjacent parameters. Thus, the norm balls for the fused LASSO are shown in Figure 5.1.2 (in $\mathbb{R}^2$ and $\mathbb{R}^3$) for comparison with those of the combined penalties.

4.3.4 Hierarchy

The hierarchy principal is to include any main effects that are associated with higher order effects. Statisticians argue that hierarchical structure is necessary to avoid any reparameterization of the model if the data shifts location. As a simple example, consider fitting a model including a squared term and its main effect,

$$Y = \beta_0 + \beta_1 X^2 + \beta_2 X \tag{4.3.10}$$

or fitting the model excluding the main effect,

$$Y = \beta_0 + \beta_1 X^2. \tag{4.3.11}$$

Suppose that X were to shift in location, say to $X + a$. Then model (4.3.10) is unaffected by the change,

$$\begin{aligned} Y &= \beta_0 + \beta_1 (X + a)^2 + \beta_2 (X + a) \\ &= \left(\beta_0 + a^2\beta_1 + a\beta_2\right) + \beta_1 X^2 + (\beta_2 + 2a\beta_1) X \\ &= \beta_0^* + \beta_1 X^2 + \beta_2^* X. \end{aligned}$$

However, this is not the case for model (4.3.11),

$$\begin{aligned} Y &= \beta_0 + \beta_1 (X + a)^2 \\ &= \left(\beta_0 + a^2\beta_1\right) + \beta_1 X^2 + 2a\beta_1 X \\ &= \beta_0^* + \beta_1 X^2 + \beta_2^* X. \end{aligned}$$

The main effect thus reappears when the data shifts and predictions using model (4.3.11) will not include a parameter for it. Geometrically, removal of the X term means that the quadratic curve is symmetric about $x = 0$ and has its turning point at $x = 0$. Similarly, if X_1^2 and X_2^2 are included in the model then $X_1 X_2$ should also be included. Omitting the $X_1 X_2$ term assumes that the quadric surface is aligned with

the coordinate axis and any rotation of the surface will reintroduce the term. See Faraway (2005:130-131).

As mentioned above, the CAP estimator (4.3.6) can be used to enforce hierarchy by creating overlapping groups in a specific way. Alternatively, Bien *et al.* (2013) focus on a two-way interaction model, including pairwise interactions between variables,

$$Y = \beta_0 + \sum_j X_j \beta_j + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} X_j X_k + \varepsilon. \tag{4.3.12}$$

The response vector $\mathbf{y}$ is centered to form the vector $\mathbf{v}$. The $n \times p$ predictor matrix $\mathbf{X}$ is first centered and scaled to produce the matrix $\mathbf{Z}_1$. The $n \times p(p-1)$ matrix $\mathbf{Z}_2$ is calculated with columns $Z_j Z_k$ for $j \neq k$ and the columns of $\mathbf{Z}_2$ are then centered. The $p \times 1$ parameter vector corresponding to $\mathbf{Z}_1$ is $\underline{\alpha}$ and the $p \times p$ parameter matrix corresponding to $\mathbf{Z}_2$ is Θ with $\Theta_{jj} = 0$ and $\Theta^T = \Theta$. Let $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2]$ and $\underline{\theta}^T = \left(\underline{\alpha}^T, \text{vec}(\Theta)^T/2\right)$. The problem is then given by

$$\hat{\underline{\alpha}}^H = \underset{\underline{\alpha}}{\arg\min} \ \|\mathbf{v} - \mathbf{Z}\underline{\theta}\|^2 + \lambda \|\underline{\alpha}\|_1 + \frac{\lambda}{2} \|\Theta\|_1$$

$$\text{subject to } \Theta^T = \Theta, \ \left\|\Theta_j\right\|_1 \leqslant \left|\alpha_j\right| \ \text{ for } j = 1, 2, \ldots, p$$

where $\|\Theta\|_1 = \sum_{j \neq k} \left|\Theta_{jk}\right|$ and Θ_j is the j-th row (or column) of Θ. If $\hat{\Theta}_{jk} \neq 0$ then $\left\|\Theta_j\right\|_1 > 0$ and $\|\Theta_k\|_1 > 0$ so that $\hat{\alpha}_j, \hat{\alpha}_k \neq 0$. Thus, hierarchy is maintained by design.

Chapter 5

Other Shrinkage Methods

This chapter provides a brief look at other shrinkage methods currently available. The EN(Section 5.1.1) and OSCAR (Section 5.1.2) combine the LASSO penalty with the ℓ_2 norm and ℓ_∞ norm, respectively. These penalties are capable of including groups of correlated predictors in the model. Finally, the concave penalties of SCAD (Section 5.2.1) and MCP (Section 5.2.2) are mentioned. These penalties produce nearly unbiased estimates which are consistent and efficient, despite being concave and non-differentiable and are the key focus of many researchers today.

5.1 Combined Penalties

Ridge regression often outperforms the LASSO when there are high pairwise correlations between groups of predictors. In this situation, the LASSO tends to randomly select only one of the predictors in the group. For predictive purposes that is often satisfactory, but there may be instances when identifying the whole group is of importance. Zou & Hastie (2005) uncovered the reason behind the grouping effect of ridge regression and found a way to combine the ridge and LASSO penalties without overshrinking parameters. The result is the elastic net (EN), which is capable of including groups of correlated variables while promoting sparsity. Bondell & Reich (2008) proposed a similar idea, OSCAR, which combines the LASSO penalty with a pairwise max norm. The result is a penalty function which allows multiple groups of differing magnitudes to be identified. These combined penalties were later modified to have the oracle property: the adaptive EN by Zou & Zhang (2009) and PACS by Sharma *et al.* (2013).

The EN is also an attractive method to use when there are many relevant predictors in the high dimensional setting with $p \gg n$. Osborne *et al.* (2000b) show that the LASSO selects at most $\min(n, p)$ variables, so that when $p > n$ it cannot select more than n variables. In contrast, the EN can potentially select all p variables.

5.1.1 Elastic Net

The elastic net (EN) was introduced by Zou & Hastie (2005) as a combination of ridge regression and the LASSO. The naive EN estimator is given by

$$\hat{\underline{\alpha}}^{NE} = \underset{\underline{\alpha}}{\arg\min} \ \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 \ \text{subject to} \ (1-\psi)\|\underline{\alpha}\|_1 + \psi\|\underline{\alpha}\|^2 \leqslant \tau, \tag{5.1.1}$$

where $\tau > 0$ is the tuning parameter controlling the size of the constraint and $\psi \in (0,1)$ controls the weighting of the ℓ_1 and ℓ_2 norms. When $\psi = 0$, the constraint becomes the LASSO and when $\psi = 1$ it becomes the ridge regression constraint. Like the LASSO, the constraint is non-differentiable at 0 and it has the ability to produce sparse solutions by setting parameter estimates exactly to zero. Like ridge regression, the constraint is strictly convex for all $\psi > 0$. Zou & Hastie (2005) state that the strict convexity allows the EN to include groups of highly correlated predictors if their effects are equal in size. In the extreme case when the predictors are exactly identical, their parameter estimates will be identical. The LASSO constraint is convex, but not strictly convex. Thus, the LASSO does not have this grouping effect and in the case of identical predictors it will not have a unique solution. Problem (5.1.1) is equivalent to the penalized regression,

$$\begin{aligned}\hat{\underline{\alpha}}^{NE} &= \underset{\underline{\alpha}}{\arg\min} \ \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 + \lambda\left\{(1-\psi)\|\underline{\alpha}\|_1 + \psi\|\underline{\alpha}\|^2\right\} \\ &= \underset{\underline{\alpha}}{\arg\min} \ \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 + \lambda_1\|\underline{\alpha}\|_1 + \lambda_2\|\underline{\alpha}\|^2, \end{aligned} \tag{5.1.2}$$

where $\psi = \lambda_2/(\lambda_1 + \lambda_2)$ and $\lambda = \lambda_1 + \lambda_2$, with $\lambda_1 \geqslant 0$ and setting $\lambda_2 > 0$ ensures strict convexity. So the penalty function is

$$P_E(\underline{\alpha}) = \lambda_1\|\underline{\alpha}\|_1 + \lambda_2\|\underline{\alpha}\|^2. \tag{5.1.3}$$

In this form we see that the ridge penalty can be obtained by setting $\lambda_1 = 0$ and the LASSO penalty is obtained by setting $\lambda_2 = 0$.

The naive EN can be written as a LASSO problem and solved in the same fashion. Let

$$\underset{(n+p)\times p}{\mathbf{Z}^*} = \frac{1}{\sqrt{1+\lambda_2}}\begin{bmatrix}\mathbf{Z} \\ \sqrt{\lambda_2}\mathbf{I}_p\end{bmatrix} \text{ and } \underset{(n+p)\times 1}{\mathbf{v}^*} = \begin{bmatrix}\mathbf{v} \\ \mathbf{0}\end{bmatrix}, \tag{5.1.4}$$

and let $\lambda^* = \lambda_1 / \sqrt{1+\lambda_2}$ and $\underline{\alpha}^* = \sqrt{1+\lambda_2}\underline{\alpha}$. Then

$$\underline{\hat{\alpha}}^* = \underset{\underline{\alpha}^*}{\arg\min} \left\| \mathbf{v}^* - \mathbf{Z}^* \underline{\alpha}^* \right\|^2 + \lambda^* \left\| \underline{\alpha}^* \right\|_1 \tag{5.1.5}$$

is the solution of the augmented problem. The equivalence of (5.1.2) and (5.1.5) can easily be verified by substitution,

$$\begin{aligned}
&\left\| \mathbf{v}^* - \mathbf{Z}^* \underline{\alpha}^* \right\|^2 + \lambda^* \left\| \underline{\alpha}^* \right\|_1 \\
&= \left\| \begin{bmatrix} \mathbf{v} \\ \mathbf{0} \end{bmatrix} - \frac{1}{\sqrt{1+\lambda_2}} \begin{bmatrix} \mathbf{Z} \\ \sqrt{\lambda_2}\mathbf{I}_p \end{bmatrix} \sqrt{1+\lambda_2}\underline{\alpha} \right\|^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} \left\| \sqrt{1+\lambda_2}\underline{\alpha} \right\|_1 \\
&= \left\| \begin{bmatrix} \mathbf{v} - \mathbf{Z}\underline{\alpha} \\ \sqrt{\lambda_2}\underline{\alpha} \end{bmatrix} \right\|^2 + \lambda_1 \left\| \underline{\alpha} \right\|_1 \\
&= \left\| \mathbf{v} - \mathbf{Z}\underline{\alpha} \right\|^2 + \left\| \sqrt{\lambda_2}\underline{\alpha} \right\|^2 + \lambda_1 \left\| \underline{\alpha} \right\|_1 \\
&= \left\| \mathbf{v} - \mathbf{Z}\underline{\alpha} \right\|^2 + \lambda_2 \left\| \underline{\alpha} \right\|^2 + \lambda_1 \left\| \underline{\alpha} \right\|_1 .
\end{aligned}$$

The naive EN solution is then

$$\underline{\hat{\alpha}}^{NE} = \frac{\underline{\hat{\alpha}}^*}{\sqrt{1+\lambda_2}}.$$

Note the similarity of the augmented matrices (5.1.4) with the augmented ridge problem (2.3.8). The naive EN differs only by the factor of $1/\sqrt{1+\lambda_2}$ in the $\mathbf{Z}^*$ matrix. The augmented matrix $\mathbf{Z}^*$ has $n+p$ rows and rank $(\mathbf{Z}^*) = p$ so the naive EN could potentially include all p variables in the model, even when $p \gg n$.

Zou & Hastie (2005) realized that the naive EN appears to double the amount of shrinkage, which inflates the bias without any further reduction in variance. The problem is that both the ridge and LASSO penalties attempt to shrink the estimates. Using (2.3.7), the form of the ridge estimate with shrinkage parameter λ_2 is

$$\left(\mathbf{Z}^T\mathbf{Z} + \lambda_2\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{v} = \frac{1}{1+\lambda_2} \begin{bmatrix} 1 & \frac{r_{12}}{1+\lambda_2} & \cdots & \frac{r_{1p}}{1+\lambda_2} \\ \frac{r_{12}}{1+\lambda_2} & 1 & \cdots & \frac{r_{2p}}{1+\lambda_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{r_{1p}}{1+\lambda_2} & \frac{r_{2p}}{1+\lambda_2} & \cdots & 1 \end{bmatrix}^{-1} \mathbf{Z}^T\mathbf{v}.$$

Zou & Hastie (2005) suggest that decorrelation, shrinking the correlations by $1/(1+\lambda_2)$, is the cause of the grouping effect in ridge regression. However, they argue that the direct shrinkage factor of $1/(1+\lambda_2)$ is not needed by the EN since the LASSO shrinkage effectively controls the variance in addition to promoting

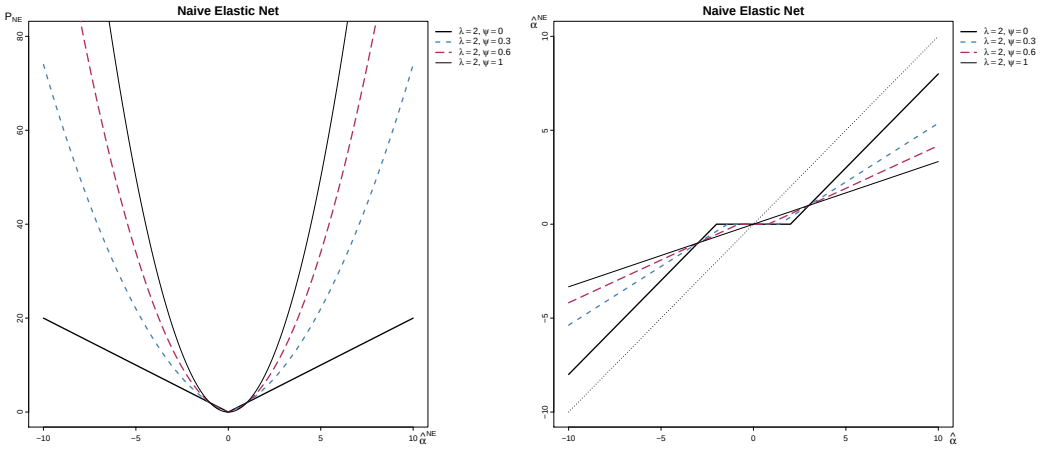
sparsity. The corrected EN estimate is therefore scaled to undo the extra shrinkage,

$$\hat{\underline{\alpha}}^{E} = (1+\lambda_2)\,\hat{\underline{\alpha}}^{NE} = \sqrt{1+\lambda_2}\hat{\underline{\alpha}}^{*}.$$

A further motivation for the correction factor is seen from the orthogonal design. In this case,

$$\hat{\alpha}_j^{NE} = \frac{\operatorname{sign}\left(\hat{\alpha}_j\right)\left(\left|\hat{\alpha}_j\right| - \lambda_1\right)_+}{(1+\lambda_2)}.$$

For large LSEs, the naive EN threshold has substantial bias. Applying the correction factor $(1+\lambda_2)$, the EN threshold is identical to the LASSO threshold and achieves near minimax optimality. Figure 5.1.1 the penalty function and thresholding function for the naive elastic net. The penalty function is convex but the thresholding has very large bias, shrinking parameters nearly as much as ridge regression.



(a) (b)

Figure (5.1.1) Penalty and thresholding functions for the naive EN. The LASSO is represented by the thick black curve and ridge regression by the thin black curve. Except that the naive EN penalty is discontinuous at zero when $\psi < 1$, the functions look similar to the bridge functions in Figure 4.2.1 with $\gamma \in [1,2]$. Large estimates are subject to larger shrinkage than the LASSO. After applying the correction factor, the EN thresholding function is identical to the LASSO.

The EN estimate is thus given by

$$\begin{aligned}\hat{\underline{\alpha}}^{E} &= \arg\min_{\underline{\alpha}}\ (1+\lambda_2)\left\{\|\mathbf{v}-\mathbf{Z}\underline{\alpha}\|^2 + \lambda_1\|\underline{\alpha}\|_1 + \lambda_2\|\underline{\alpha}\|^2\right\} \\ &= \arg\min_{\underline{\alpha}}\ \underline{\alpha}^T\left(\frac{\mathbf{Z}^T\mathbf{Z}+\lambda_2\mathbf{I}}{1+\lambda_2}\right)\underline{\alpha} - 2\mathbf{v}^T\mathbf{Z}\underline{\alpha} + \lambda_1\|\underline{\alpha}\|_1, \end{aligned} \tag{5.1.6}$$

where the LASSO is obtained by setting $\lambda_2 = 0$. So the EN can be seen as stabilizing the LASSO. The parameter λ_1 controls the amount of shrinkage and selection whereas the λ_2 parameter controls the amount grouping.

As with the LASSO, the EN can be solved efficiently by using a modification of the LAR algorithm, LAR-EN, to calculate the entire solution path for fixed λ_2. However, the EN has two tuning parameters (λ_1, λ_2) which must be estimated. For a grid of λ_2 values, the algorithm provides the entire solution path for each λ_2. Then λ_1 can be selected using K-fold cross-validation and the value of λ_2 giving the lowest cross-validation error is selected. The computational cost increases with p but is still manageable even when $p \gg n$, although early stopping rules can be used to lessen the computational load. Since $\underline{\hat{\alpha}}^E \propto \underline{\hat{\alpha}}^*$ from the augmented LASSO problem, we could also parameterize the EN by (λ_2, t) or (λ_2, s), where t is the ℓ_1 norm of the coefficients and $s = t/t_0 = \left\|\underline{\hat{\alpha}}^E\right\|_1 / \|\underline{\hat{\alpha}}\|_1$.

Adaptive Elastic Net

The adaptive LASSO and the EN improve the LASSO in different ways. The adaptive LASSO controls the bias by shrinking larger parameters less, while the EN handles collinearity by incorporating the ridge penalty. The EN can be also extended to shrink parameters by different amounts. The adaptive EN is a combination of the adaptive LASSO and the EN and enjoys the good properties of both methods. It is given by

$$\underline{\hat{\alpha}}^{AE} = \underset{\underline{\alpha}}{\arg\min}\ (1+\lambda_2)\left\{\|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 + \lambda_1^* \sum_{j=1}^{p} w_j \left|\alpha_j\right| + \lambda_2 \|\underline{\alpha}\|^2\right\}, \tag{5.1.7}$$

where $w_j > 0$ are weights. Setting $\lambda_1 = 0$ we obtain the ridge penalty and setting $\lambda_2 = 0$ we obtain the adaptive LASSO penalty. Zou & Zhang (2009) suggest fitting the EN model to obtain $\underline{\hat{\alpha}}^E$ and calculating the weights as

$$\hat{w}_j = \left(1/\left|\hat{\alpha}_j^E\right|\right)^{\zeta},$$

where $\hat{\alpha}_j^E \neq 0$. Let $\widehat{\mathcal{D}} = \left\{j : \hat{\alpha}_j^E \neq 0\right\}$, then the adaptive EN estimates can be calculated as $\underline{\hat{\alpha}}^{AE}_{\widehat{\mathcal{D}}^c} = 0$ and

$$\underline{\hat{\alpha}}^{AE}_{\widehat{\mathcal{D}}} = \underset{\underline{\alpha}}{\arg\min}\ (1+\lambda_2)\left\{\left\|\mathbf{v} - \mathbf{Z}_{\widehat{\mathcal{D}}}\underline{\alpha}_{\widehat{\mathcal{D}}}\right\|^2 + \lambda_1^* \sum_{j\in\widehat{\mathcal{D}}} \hat{w}_j \left|\alpha_j\right| + \lambda_2 \left\|\underline{\alpha}_{\widehat{\mathcal{D}}}\right\|^2\right\}.$$

The same tuning parameter λ_2 can be used for the EN and adaptive EN since it has the same contribution in both methods but λ_1 and λ_1^* are likely to differ. We can obtain the solution by using the LAR-EN algorithm.

Let $\mathbf{Z}^*_{\widehat{\mathcal{D}}}$ be the matrix with columns $\mathbf{z}^*_j = \mathbf{z}_j/\hat{w}_j$ for $j \in \widehat{\mathcal{D}}$. Then we can formulate the problem as

$$\underline{\hat{\alpha}}^* = \arg\min_{\underline{\alpha}} \left\|\mathbf{v} - \mathbf{Z}^*_{\widehat{\mathcal{D}}}\underline{\alpha}_{\widehat{\mathcal{D}}}\right\|^2 + \lambda_1^* \left\|\underline{\alpha}_{\widehat{\mathcal{D}}}\right\|_1 + \lambda_2 \left\|\underline{\alpha}_{\widehat{\mathcal{D}}}\right\|^2,$$

and $\hat{\alpha}_j^{AE} = (1+\lambda_2)\,\hat{\alpha}_j^* \Big/ \hat{w}_j$ for $j = 1, 2, \ldots, p$. See Zou & Zhang (2009) for details.

Friedman (2012) proposed the generalized elastic net penalty,

$$P_{GE}(\underline{\alpha}) = \sum_{j=1}^{p} \ln\left((1-\psi)\left|\alpha_j\right| + \psi\right),$$

where $0 < \psi < 1$. The penalty is concave and approaches the LASSO as $\psi \to 1$ and approaches subset selection as $\psi \to 0$. So this penalty provides a bridge between subset selection and the LASSO, while the elastic net provides a bridge between the LASSO and ridge regression. Together the elastic net family encompass the same range of penalties as bridge estimates. However, it is shown that the elastic net penalties are more stable than bridge penalties.

5.1.2 OSCAR

The octagonal shrinkage and clustering algorithm for regression (OSCAR) penalty was introduced by Bondell & Reich (2008). It is similar to the EN since it is also a combination of two norms, but in this case, the ℓ_1-norm and the pairwise ℓ_∞-norm of the parameters. The estimate is given by

$$\underline{\hat{\alpha}}^O = \arg\min_{\underline{\alpha}} \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 \text{ subject to } \sum_{j=1}^{p}\left|\alpha_j\right| + \psi\sum_{j<k}\max\left\{\left|\alpha_j\right|, \left|\alpha_k\right|\right\} \leqslant \tau, \tag{5.1.8}$$

where $\tau > 0$ controls the size of the constraint and $\psi \geqslant 0$ controls the extent of the pairwise ℓ_∞-norm. Similarly to the EN, the ℓ_1 norm controls the variance and promotes sparsity, while the ℓ_∞-norm promotes equality of parameter estimates. Thus, the OSCAR penalty is also capable of including groups of highly correlated variables by setting the estimates of parameters within a group to be equal. The pairwise ℓ_∞-norm is used instead of the overall ℓ_∞-norm so that multiple groups of variables with different magnitudes can be included, the latter would only allow one group with the largest magnitude to be included. The OSCAR penalty also bears some resemblance to the fused LASSO, which imposes a pairwise ℓ_1 norm in combination with the regular ℓ_1 norm. However, OSCAR considers any pairs of variables and not only adjacent ones. The LASSO is obtained by setting $\psi = 0$, which results in sparsity but no grouping. Letting

$\psi \to \infty$ results in the grouping effect without any sparsity. Problem (5.1.8) is equivalent to the penalized regression

$$\begin{aligned}\hat{\underline{\alpha}}^{O} &= \arg\min_{\underline{\alpha}} \ \|\mathbf{v}-\mathbf{Z}\underline{\alpha}\|^2 + \lambda\left\{\|\underline{\alpha}\|_1 + \psi\sum_{j<k}\max\left\{|\alpha_j|,|\alpha_k|\right\}\right\} \\ &= \arg\min_{\underline{\alpha}} \ \|\mathbf{v}-\mathbf{Z}\underline{\alpha}\|^2 + \lambda_1\|\underline{\alpha}\|_1 + \lambda_2\sum_{j<k}\max\left\{|\alpha_j|,|\alpha_k|\right\},\end{aligned} \tag{5.1.9}$$

where $\lambda \geqslant 0$ and the penalty function is given by

$$P_O(\underline{\alpha}) = \lambda_1\|\underline{\alpha}\|_1 + \lambda_2\ell_\infty\left(\alpha_j,\alpha_k\right). \tag{5.1.10}$$

The penalty function is convex and to solve the problem, Bondell & Reich (2008) use the $2p$ variables corresponding to the nonnegative parameters α_j^+ and α_j^- such that $\alpha_j = \alpha_j^+ - \alpha_j^-$. They also introduce $p(p-1)/2$ variables l_{jk} for the pairwise maxima for $1 \leqslant j \leqslant k \leqslant p$. The problem can then be written as a quadratic programming problem with $\left(p^2-3p\right)/2$ variables and p^2+p+1 linear constraints. Since the quadratic programming is of order p^2, it can be computationally expensive for large p. The problem is stated as

$$\begin{aligned}\text{minimize} \quad & \tfrac{1}{2}\left\|\mathbf{v}-\textstyle\sum_{j=1}^{p}\mathbf{Z}_j\left(\alpha_j^+-\alpha_j^-\right)\right\|^2 \\ \text{subject to} \quad & \textstyle\sum_{j=1}^{p}\left(\alpha_j^++\alpha_j^-\right)+\psi\sum_{j<k}l_{jk} \leqslant \tau \\ & l_{jk} \geqslant \alpha_j^++\alpha_j^- \text{ for } 1 \leqslant j \leqslant k \leqslant p \\ & l_{jk} \geqslant \alpha_k^++\alpha_k^- \text{ for } 1 \leqslant j \leqslant k \leqslant p \\ & \alpha_j^+ \geqslant 0, \text{ for all } j=1,2,\ldots,p \\ & \alpha_j^- \geqslant 0 \text{ for all } j=1,2,\ldots,p\end{aligned}$$

Wu *et al.* (2009) proposed a similar penalty given by

$$P_\infty(\underline{\alpha}) = (1-\psi)\|\underline{\alpha}\|_1 + \psi\|\underline{\alpha}\|_\infty,$$

which also includes features of sparsity and grouping. They show that the penalty is piecewise linear and provide a homotopy algorithm for its solution.

PACS

Sharma *et al.* (2013) generalize the OSCAR penalty by including weights on the ℓ_1 norm. However, their pairwise absolute clustering and sparsity (PACS) penalty includes weighted sums and differences of pairs of coefficients instead of the pairwise ℓ_∞-norm. The estimate is given by

$$\hat{\underline{\alpha}}^P = \underset{\underline{\alpha}}{\arg\min} \ \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 + \lambda \left[\sum_{j=1}^{p} w_j |\alpha_j| + \sum_{j<k} w_{jk(-)} |\alpha_k - \alpha_j| + \sum_{j<k} w_{jk(+)} |\alpha_j + \alpha_k| \right], \tag{5.1.11}$$

where $\lambda \geqslant 0$ is the tuning parameter and w_j, $w_{jk(-)}$ and $w_{jk(+)}$ are nonnegative weights. Sharma *et al.* (2013) show that OSCAR is a special case of PACS by noting that

$$\max\left\{|\alpha_j|, |\alpha_k|\right\} = \frac{1}{2}\left\{|\alpha_k - \alpha_j| + |\alpha_j + \alpha_k|\right\}.$$

Then the OSCAR estimate is given by

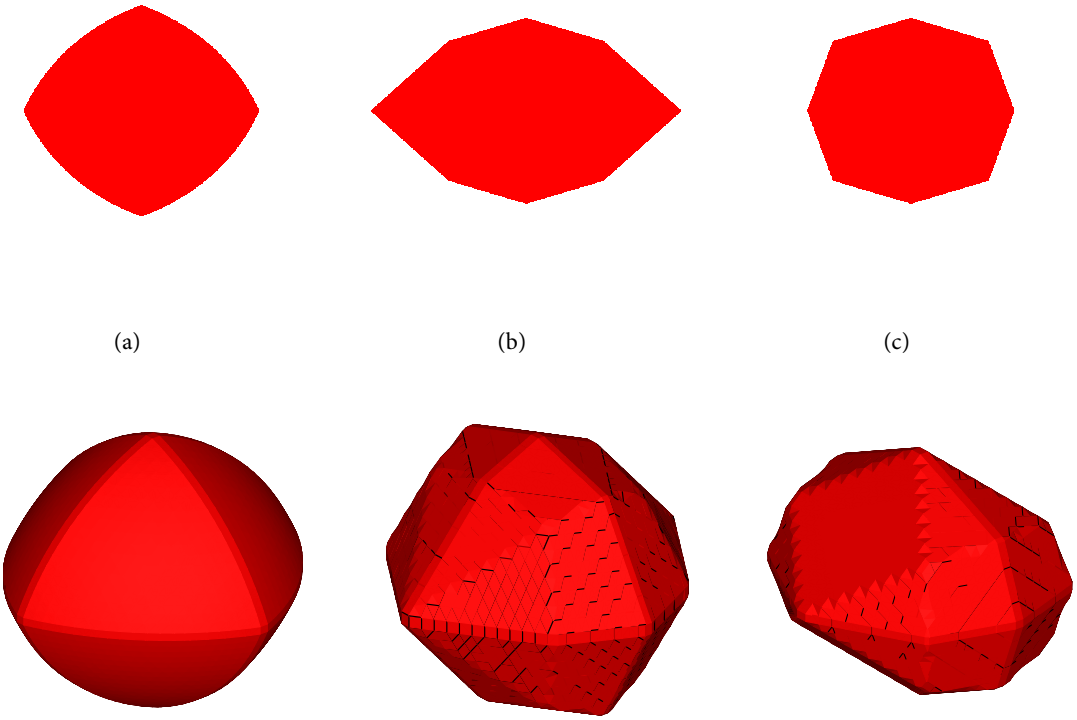
$$\hat{\underline{\alpha}}^O = \underset{\underline{\alpha}}{\arg\min} \ \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 + \lambda \left[\psi \sum_{j=1}^{p} |\alpha_j| + \frac{1}{2}(1-\psi) \sum_{j<k} |\alpha_k - \alpha_j| + \frac{1}{2}(1-\psi) \sum_{j<k} |\alpha_j + \alpha_k| \right],$$

where $\psi \in [0,1]$. They also point out that the ridge penalty can be formulated as $2(p-1)\sum_{j=1}^{p} \alpha_j^2 = \sum_{j\neq k}\left[(\alpha_j - \alpha_k)^2 + (\alpha_j + \alpha_k)^2\right]$. Four different approaches for calculating the weights are discussed by Sharma *et al.* (2013). In particular, the adaptive weights are given by $\hat{w}_j = |\hat{\alpha}_j|^{-\zeta}$, $\hat{w}_{jk(-)} = |\hat{\alpha}_k - \hat{\alpha}_j|^{-\zeta}$ and $\hat{w}_{jk(+)} = |\hat{\alpha}_k - \hat{\alpha}_j|^{-\zeta}$ for $\zeta > 0$, where $\hat{\alpha}_j$ are any consistent estimates such as the LSEs. Sharma *et al.* (2013) also develop a local quadratic approximations (LQA) algorithm to compute the solution more efficiently than quadratic programming.

5.1.3 Geometry

Presented here are the norm balls for EN, OSCAR and the fused LASSO in Figure 5.1.2. The elastic net is curved like ridge regression but has points on the axes due to its LASSO characteristic. The effect of the curves meeting at a point on each axis is to bulge the curve outward along the diagonals of each quadrant. Parameters are thus encouraged to fall either on a point and be set to zero, or near the diagonals and have the same size as other parameters. Adjusting the ψ parameter lower yields sharper points and adjusting it upward increases the curvature. The OSCAR norm ball is similar but, instead of having points only on the axes, it has pointy edges in every direction. This allows for multiple groups of equally sized parameters

within groups and different sized parameters between groups. The octagonal shape is the reason for its name. The fused LASSO looks very similar to OSCAR in two dimensions since the pairwise ℓ_1 norm acts on adjacent parameters. The difference can be seen by examining the three dimensional ball. The fused LASSO produces a rather strange looking surface with pointy edges along the xy-plane and the yz-plane but flat LASSO-like diamonds along the xz-plane.



(a) (b) (c)

(d) (e) (f)

Figure (5.1.2) Norm balls in $\mathbb{R}^2$ and $\mathbb{R}^3$ for (a),(d) elastic net, (b), (e) OSCAR, and (c),(f) fused LASSO, all with $\psi = 0.5$. Each norm ball promotes sparsity with protruding points on the axes and encourages estimates of equal size with protruding points on the diagonals. OSCAR allows for multiple groups with different sized estimates between groups and fused LASSO sets estimates equal only for adjacent variables.

5.2 Concave Penalties

Bridge penalties with $0 < \gamma < 1$ are concave functions but they have the appeal of decreasing the amount of shrinkage as the size of the parameter increases. Furthermore, Knight & Fu (2000) show that these penalties have the oracle properties (Definition A.3.8). This is the idea behind the concave penalty functions,

large parameters are penalized less so that the resulting estimates are nearly unbiased. In particular, this is achieved by placing a constant bound on the penalty function. Fan & Li (2001) proposed SCAD, which was the first shrinkage method having the oracle property. Although the adaptive LASSO is oracle, the bias may decrease at a faster rate with SCAD. MCP, proposed by Zhang (2010), follows a similar approach but penalizes smaller parameters less. Despite having concave penalties which are also non-differentiable at zero, they both provide efficient algorithms for computing the solution, even in high dimensional settings when $p \geqslant n$.

5.2.1 SCAD

The smoothly clipped absolute deviation (SCAD) estimate proposed by Fan & Li (2001) solves the penalized regression

$$\hat{\underline{\alpha}}^S = \underset{\underline{\alpha}}{\arg\min} \; \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 + \sum_{j=1}^{p} P_S\left(\alpha_j\right), \tag{5.2.1}$$

with penalty function

$$P_S\left(\alpha_j\right) = \begin{cases} \lambda\left|\alpha_j\right| & \text{if } \left|\alpha_j\right| \leqslant \lambda \\ \dfrac{-\left(\alpha_j^2 - 2\xi\lambda\left|\alpha_j\right| + \lambda^2\right)}{2\left(\xi - 1\right)} & \text{if } \lambda < \left|\alpha_j\right| \leqslant \xi\lambda \\ \left(\xi + 1\right)\lambda^2/2 & \text{if } \left|\alpha_j\right| > \xi\lambda, \end{cases} \tag{5.2.2}$$

where $\xi > 2$ and $\lambda \geqslant 0$ are tuning parameters. The penalty function $P_S\left(\alpha_j\right)$ is a symmetric quadratic spline function with knots at λ and $\xi\lambda$. It applies a different amount of shrinkage to parameters based on their size. For small parameters, the penalty is equal to the LASSO penalty and the shrinkage is constant. While the LASSO penalty remains linear for all parameters, the SCAD penalty becomes quadratic for moderately sized parameters and starts applying less shrinkage as the size of the parameter grows. For large parameters the penalty is constant and little or no shrinkage is applied. Thus, the ξ parameter effectively controls the region in which parameters are almost unpenalized. As a result the SCAD estimate will have less bias than the LASSO when there are large parameters in the model. The rate of shrinkage is clear when examining the derivative of the penalty function,

$$P_S'\left(\alpha_j\right) = \begin{cases} \operatorname{sign}\left(a_j\right)\lambda & \text{if } \left|\alpha_j\right| \leqslant \lambda \\ \operatorname{sign}\left(a_j\right)\left(\xi\lambda - \left|\alpha_j\right|\right)/\left(\xi - 1\right) & \text{if } \lambda < \left|\alpha_j\right| \leqslant \xi\lambda \\ 0 & \text{if } \left|\alpha_j\right| > \xi\lambda. \end{cases}$$

Since the penalty function is the LASSO penalty for $|\alpha_j| \leqslant \lambda$ it is also non-differentiable at zero so that SCAD produces sparse models. Fan & Li (2001) show that scad has the oracle properties and also provide a sandwich formula for calculating standard errors.

A difficulty with SCAD is that the penalty function is concave, which complicates its computation. Despite the concavity, Fan & Li (2001) propose an algorithm using LQA to solve the problem. Zou & Li (2008) developed an algorithm based on local linear approximation (LLA). Breheny & Huang (2011) developed a coordinate descent algorithm to compute the SCAD solution. Clarke *et al.* (2009:611-615) discuss the the LQA algorithm, along with three other algorithms which have been developed to calculate the SCAD estimate. Cross-validation or GCV can be used to estimate the tuning parameters ξ and λ. Although, searching over a two-dimensional grid of values can easily become computationally expensive and they recommend fixing $\xi = 3.7$ since they find it to be similar to GCV.

5.2.2 MCP

Zhang (2010) introduced the minimax concave penalty (MCP) which also has a concave penalty function. The penalized regression takes the form

$$\hat{\underline{\alpha}}^M = \arg\min_{\underline{\alpha}} \left\{ \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\| + \sum_{j=1}^{p} P_M\left(\alpha_j\right) \right\}, \tag{5.2.3}$$

and the penalty function is defined as

$$P_M\left(\alpha_j\right) = \begin{cases} \lambda|\alpha_j| - \alpha_j^2 \big/ 2\xi & \text{if } |\alpha_j| \leqslant \xi\lambda \\ \xi\lambda^2 \big/ 2 & \text{if } |\alpha_j| > \xi\lambda, \end{cases} \tag{5.2.4}$$

where $\xi > 1$ and $\lambda \geqslant 0$ are tuning parameters. The LASSO penalty is obtained when $\xi \to \infty$. While the limiting distributions of concave bridge penalties obtained by Knight & Fu (2000) show that large parameters are estimated with less bias, they also show that small nonzero parameters are not estimated consistently but are instead set to zero. The MCP attempts to correct the problem and applies less shrinkage to smaller parameters. The penalty is thus a an improvement of the SCAD penalty, since the bias is slightly lower and the accuracy of variable selection is improved. Its derivative is given by

$$P'_M\left(\alpha_j\right) = \begin{cases} \lambda - \text{sign}\left(\hat{\alpha}_j\right)|\alpha_j| \big/ \xi & \text{if } |\alpha_j| \leqslant \xi\lambda \\ 0 & \text{if } |\alpha_j| > \xi\lambda, \end{cases}$$

and is also non-differentiable at zero so that parameters can be set to zero. Zhang (2010) propose the penalized linear unbiased selection (PLUS) algorithm to find the solution and the coordinate descent algorithm by Breheny & Huang (2011) also solves the MCP problem.

5.2.3 Orthogonal Design

The penalty and thresholding functions for SCAD and MCP are summarized in Table 5.2.1 and displayed in Figure 5.2.1. Both penalty functions are convex and look almost identical. The shrinkage is also similar, both penalties set small parameters to zero and leave large parameters untouched. Let's call the region in between the shrinkage region. The difference between the penalties is the rate at which parameters in the shrinkage region are shrunk. The SCAD penalty has a steeper gradient than MCP so that smaller parameters will be shrunk towards zero at a faster rate.

Method	**Penalty Function**	**Thresholding Function**
SCAD	$\lambda\lvert\alpha_j\rvert$ if $\lvert\alpha_j\rvert \leqslant \lambda$ $\frac{-\left(\alpha_j^2-2\xi\lambda\lvert\alpha_j\rvert+\lambda^2\right)}{2(\xi-1)}$ if $\lambda < \lvert\alpha_j\rvert \leqslant \xi\lambda$ $(\xi+1)\lambda^2/2$ if $\lvert\alpha_j\rvert > \xi\lambda$	$\operatorname{sign}\left(\hat{\alpha}_j\right)\left(\lvert\hat{\alpha}_j\rvert-\lambda\right)_+$ if $\lvert\hat{\alpha}_j\rvert \leqslant 2\lambda$ $\frac{(\xi-1)\hat{\alpha}_j-\operatorname{sign}(\hat{\alpha}_j)\xi\lambda}{(\xi-2)}$ if $2\lambda < \lvert\hat{\alpha}_j\rvert \leqslant \xi\lambda$ $\hat{\alpha}_j$ if $\lvert\hat{\alpha}_j\rvert > \xi\lambda$
MCP	$\lambda\lvert\alpha_j\rvert - \lvert\alpha_j\rvert^2/2\xi$ if $\lvert\alpha_j\rvert < \xi\lambda$ $\xi\lambda^2/2$ if $\lvert\alpha_j\rvert \geqslant \xi\lambda$	$\frac{\operatorname{sign}(\hat{\alpha}_j)\left(\lvert\hat{\alpha}_j\rvert-\lambda\right)_+}{1-1/\xi}$ if $\lvert\hat{\alpha}_j\rvert \leqslant \xi\lambda$ $\hat{\alpha}_j$ if $\lvert\hat{\alpha}_j\rvert > \xi\lambda$

Table (5.2.1) Penalty and thresholding functions for concave penalty methods

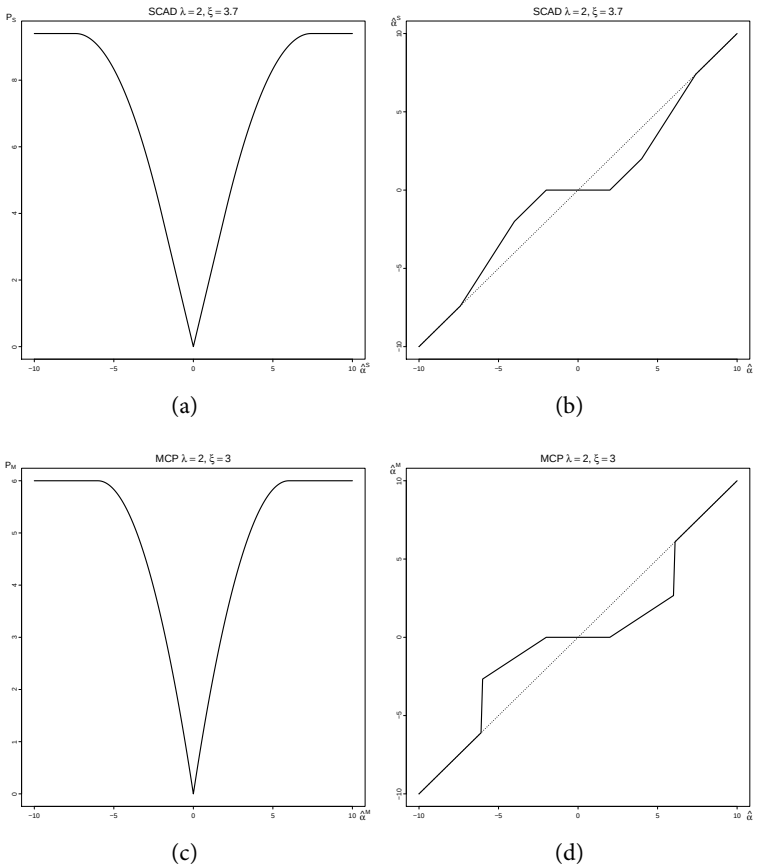


Figure (5.2.1) Penalty and thresholding functions for (a)-(b) SCAD and (c)-(d) MCP at the recommend value of ξ, 3.7 for SCAD and 3 for MCP. Large parameters are not subject to shrinkage so that the estimate is not biased. The functions look similar, but MCP shrinks smaller parameters less than SCAD.

Chapter 6

Simulation Studies

Simulation studies are performed to support the theory and identify scenarios in which the LASSO performs well. Section 6.1 explains how the performance of each method is assessed. The LASSO is compared with ridge regression and subset selection in Section 6.2. A study of the prediction accuracy along the pathways of these methods are explored and it is shown how the bias variance trade-off influences the quality of the final models. Furthermore, a number of information criteria are used to select the final model and the performance in terms of prediction and selection is assessed. It is also shown how the DF is heavily underestimated in the subset selection case when using the number of nonzero variables as an approximation. The LASSO is shown to be a good competitor for both prediction and variable selection. Section 6.3 studies the consistency of the LASSO and a number of other shrinkage methods.

6.1 Performance Measures

In the simulation studies below, regression models are estimated for each method over a grid of some complexity measure θ. In some cases, the complexity parameter indexes the entire solution path from the null model to the full least squares model. A set of coefficients $\underline{\hat{\beta}}_\theta$, and hence a set of models $\hat{f}_\theta(X) = X\underline{\hat{\beta}}_\theta$, are estimated on the training sample for each method. Each study is repeated on $N = 100$ samples and the best models $\hat{f}_{\hat{\theta}}(X)$, chosen using either CV or information criteria, are recorded for each sample. Performance measures are calculated for each iteration of the process, producing a sample of size N of each measure. Sample statistics can then be derived by analysing the distributions of these measures over the N repetitions. The measures below were calculated to compare the performance of different methods.

6.1.1 Estimation Accuracy

To assess the accuracy of estimation, the parameter estimates of the best model $\underline{\hat{\beta}}_{\hat{\theta}}$ are recorded for each sample. The statistics below can be calculated to compare their efficiency.

1. The mean of each estimate,

$$\bar{\beta}_j\left(\hat{\theta}_i\right) = \sum_{i=1}^{N} \hat{\beta}_j\left(\hat{\theta}_i\right) \Big/ N,$$

and their variances

$$\text{var}\left(\bar{\beta}_j\left(\hat{\theta}_i\right)\right) = \text{var}\left(\hat{\beta}_j\left(\hat{\theta}_i\right)\right) / N$$

2. The variance of each estimate,

$$\text{var}\left(\hat{\beta}_j\left(\hat{\theta}_i\right)\right) = \sum_{i=1}^{N} \left(\hat{\beta}_j\left(\hat{\theta}_i\right) - \bar{\beta}_j\left(\hat{\theta}_i\right)\right)^2 \Big/ (N-1)$$

3. The bias of each estimate,

$$\text{Bias}\left(\hat{\beta}_j\left(\hat{\theta}_i\right)\right) = \bar{\beta}_j\left(\hat{\theta}_i\right) - \beta_j$$

4. The MSE of each estimate,

$$MSE\left(\hat{\beta}_j\left(\hat{\theta}_i\right)\right) = \text{var}\left(\hat{\beta}_j\left(\hat{\theta}_i\right)\right) + \text{Bias}\left(\hat{\beta}_j\left(\hat{\theta}_i\right)\right)^2$$

By summing the statistics in 2-4, we obtain the total variance, bias and MSE, respectively. In particular, the total MSE can be used to compare the overall efficiency of estimation.

6.1.2 Prediction Accuracy

Prediction accuracy is assessed by calculating the mean squared error of the predictions $\hat{f}_{\hat{\theta}}(X)$ for each sample. There are a number of ways that this can be done.

1. Since the true parameter vector and covariance matrix of the predictor variables are known, we can calculate the true MSE of prediction directly,

$$MSE\left(\hat{f}_{\hat{\theta}}(X)\right) = \left(\underline{\hat{\beta}}_{\hat{\theta}} - \underline{\beta}\right)^T \Sigma \left(\underline{\hat{\beta}}_{\hat{\theta}} - \underline{\beta}\right). \tag{6.1.1}$$

The accuracy can then be assessed by calculating either,

(a) the sample mean of the MSEs and its standard error, which is the sample variance of the MSEs divided by N, or

(b) if the distribution is skewed, the sample median of the MSEs is a better statistic. The standard error of the median is calculated using the bootstrap with $B = 200$ bootstrap replications. For each $b = 1, 2, \ldots, B$, a sample of N MSEs is drawn, with replacement, and the median is calculated each time. The sample standard deviation over the B medians is an estimate of its standard error.

2. Alternatively, a test set can be used and the predictions $\hat{f}_{\hat{\theta}}(X_t)$ on this set are recorded for each training sample. The statistics described in Section 6.1.1 for the parameter estimates, namely the mean, bias, variance and MSE can be calculated in the same way for the predictions. This MSE should be similar to that obtained using equation (6.1.1).

3. A further alternative is to use an estimate of the prediction error, such as,

   (a) the test error, $TE\left(\hat{f}_{\hat{\theta}}(X)\right) = \sum_{i=1}^{m}\left(y_{t,i} - \hat{f}_{\hat{\theta}}(x_{t,i})\right)^2 \Big/ m$ for a test sample of size m. The average over the 100 samples is equal to the expected prediction error, or

   (b) the information criteria C_p, AIC and BIC, or

   (c) the cross-validation error.

   Note that these methods include the irreducible error σ^2, that is, the estimate should be similar to $MSE\left(\hat{f}_{\hat{\theta}}(X)\right) + \sigma^2$. Therefore, results are usually reported relative to σ^2. That is, for an estimate of PE, $\widehat{PE}$, results are reported for $\widehat{PE}\left(\hat{f}_{\hat{\theta}}(X)\right)\Big/\sigma^2$.

6.1.3 Variable Selection

Variable selection can be assessed by looking at the estimated parameters which are included in the best models. The measures below are useful for assessing selection performance.

1. An indicator of whether each estimate is included in the selected subset

2. The number of parameter estimates, further split by

   (a) the number of correct or incorrect nonzero parameter estimates

   (b) the number of correct or incorrect zero parameter estimates

3. Indicators of whether the true model is:

(a) selected as the best subset,

(b) a subset of the selected variables, or

(c) contained in the solution path of the method. That is, the true model could be obtained by selecting a different value of the complexity parameter $\hat{\theta}$.

The number of incorrect nonzero estimates corresponds to a type I error of the hypothesis $H_0 : \beta = 0$ and the number of incorrect zero estimates corresponds to a type II error. The measures are all averaged over the N samples. Since measures 3a - 3b, and the indicator in 1, record either a success or a failure, the sum over the N samples has a binomial distribution and the average is an estimate of the probability of success. That is, the probability of the method selecting the correct model is estimated by the proportion of times the correct model is selected out of the N times the model is fitted. Similarly, we can estimate the probability that the correct model is a subset of the selected model or if it lies in the solution path of the method and the inclusion probability of each parameter can also be estimated.

6.2 Selection and Prediction

This simulation study analyses the selection and prediction performance of the LASSO in comparison with the traditional methods. Model selection is also examined by comparing the performance when using the information criteria and CV methods described in Sections 3.2 and 3.3 and some of the methods described in Section 4.1.6.

6.2.1 Data

The data generating process is given by

$$Y = X\beta + \varepsilon \text{ where } X \sim N(0,1) \text{ and } \varepsilon \sim N\left(0, \sigma^2\right).$$

The predictor variables are related by a power decay correlation structure $\text{corr}\left(X_i, X_j\right) = \rho^{|i-j|}$. Since the predictors have unit variance, the matrix of predictor observations are distributed as $\mathbf{X} \sim N(\mathbf{0}, \Sigma)$ with $\Sigma_{ij} = \rho^{|i-j|}$. The true relationship between the response and the predictors is given by $E(\mathbf{y}) = \mathbf{X}\underline{\beta}$ where $\underline{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. That is,

$$f(X) = \sum_{j=1}^{8} X_j \beta_j,$$

and the correct model is given by

$$f(X) = X_1\beta_1 + X_2\beta_2 + X_5\beta_5.$$

This example was studied in the original LASSO paper by Tibshirani (1996) and appears in a number of studies, including but not limited to Fan & Li (2001), Zou & Hastie (2005), Zou (2006), Yuan & Lin (2007) and Bondell & Reich (2008), under various scenarios. In this study the sample size is $n = 25$, quite a small sample affording about 3 observations for each parameter. The correlation is varied by considering $\rho \in \{0, 0.5, 0.9\}$ to test the effect of collinearity. The effect of noise in the data is tested by using $\sigma \in \{1, 3, 6\}$. Estimation, selection and prediction is carried out by generating $N = 100$ samples from this process. For the generated data, the average condition number of the predictor correlation matrix and the average signal to noise ratio (SNR) are shown in Table 6.2.1. The condition number is given by

$$\kappa_2(\Sigma) = \sqrt{\frac{\max_j e_j(\Sigma)}{\min_j e_j(\Sigma)}},$$

where $e_j(\Sigma)$ are the eigenvalues of Σ, and the SNR, is given by

$$SNR = \frac{\|f(\mathbf{X})\|_2}{\sigma^2}.$$

When $\rho = 0.9$, the large condition number indicates that high levels of collinearity are present. The signal to noise ratio is very low when $\sigma = 6$, making significant effects harder to find.

		SNR		
ρ	$\kappa_2(\Sigma)$	$\sigma = 1$	$\sigma = 3$	$\sigma = 6$
0	2.690	19.654	2.184	0.546
0.5	4.585	23.213	2.579	0.645
0.9	16.751	30.131	3.348	0.837

Table (6.2.1) Average condition number and SNR for generated data

6.2.2 Estimation and Model Selection

The model is estimated using forward selection, ridge regression and the LASSO. The LAR algorithm is used for computation of the LASSO. Forward selection is indexed by the number of nonzero variables $\theta = p$, ridge regression by the shrinkage parameter $\theta = \lambda$, and the LASSO by the ℓ_1 fraction, $\theta = s = \|\underline{\alpha}\|_1 / \|\underline{\hat{\alpha}}\|_1 \in$

$[0,1]$, where $\hat{\alpha}$ is the LSE. For every model selection procedure, the optimal tuning parameter is searched for over a fixed grid of values. For comparison, least squares models and oracle least squares (that is, least squares using only the true nonzero parameters) are also fitted.

Model selection is carried out using information criteria and CV on the training sample. An independent validation sample, selected to be the same size as the training sample, is also used for comparison. The model $\hat{f}_\theta(X)$ is used to predict the response for the observations in the validation set and the expected test error is estimated by $\sum \left(y_v - \hat{f}_\theta(x_v)\right)^2 \Big/ n$. Suppose that the model with the lowest value occurs at $\hat{\theta}$, then $\hat{f}_{\hat{\theta}}(X)$ is selected as the best model. The information criteria used were C_p (3.2.5), AIC (3.2.13) and BIC (3.2.16). CV methods used include 5-fold CV, 10-fold CV, LOOCV and GCV. All of these methods attempt to estimate the expected test error. Thus, to see how well they perform, each model is used to predict an independent test sample of size $m = 200$ and the test error $\sum \left(y_t - \hat{f}_\theta(x_t)\right)^2 / m$ is recorded. The training error is also recorded to examine the extent of its optimism.

The expected test error is supposedly equal to the true PE. Since the data generating process is known, we can verify this by comparing the average test error over the N samples with the true prediction error given by $PE = MSE + \sigma^2$, where MSE is calculated using equation (6.1.1). We can also show how the prediction error is composed of the irreducible error, squared bias and variance by collecting the predicted values $\hat{f}_{\hat{\theta}}(x_t)$ for each sample. Collecting these values for each value of model complexity, we can show how these measures are related to complexity. The fitted model $\hat{f}_{\hat{\theta}}(x)$ on the training set is also collected for each value of model complexity over the N samples in order to estimate the effective DF using equation (3.1.9).

Criteria used for variable selection are also investigated. K-fold CV is applied with the 1 SE rule for $K = 5, 10$. In addition, the modified BIC, percentile CV and kappa selection methods discussed in Section 4.1.6 are applied. For both the percentile CV and kappa coefficient, 20 repetitions were used.

6.2.3 Results

Solution path

Figures 6.2.1, 6.2.2 and 6.2.3 show the prediction error, estimates thereof and its decomposition into the squared bias, variance and σ^2 for forward selection, ridge regression and the LASSO, respectively for $\rho = 0.5$ and $\sigma = 3$. The plots are all increasing with model complexity from left to right. Forward selection is

indexed by the number of variables p. The LASSO models were fitted over a grid of $s \in [0,1]$ with a step size 0.01. Ridge regression models were fitted over a grid of $\lambda \in [0,50]$ with step size of 0.2. The difficulty with the selection of the ridge tuning parameter is that there is no upper bound since estimates are not set exactly to zero. The plots for ridge regression are indexed by $-\ln(\lambda)$ so that model complexity is in the same direction as the other methods. Of course, the least squares model at $\lambda = 0$ is therefore not included.

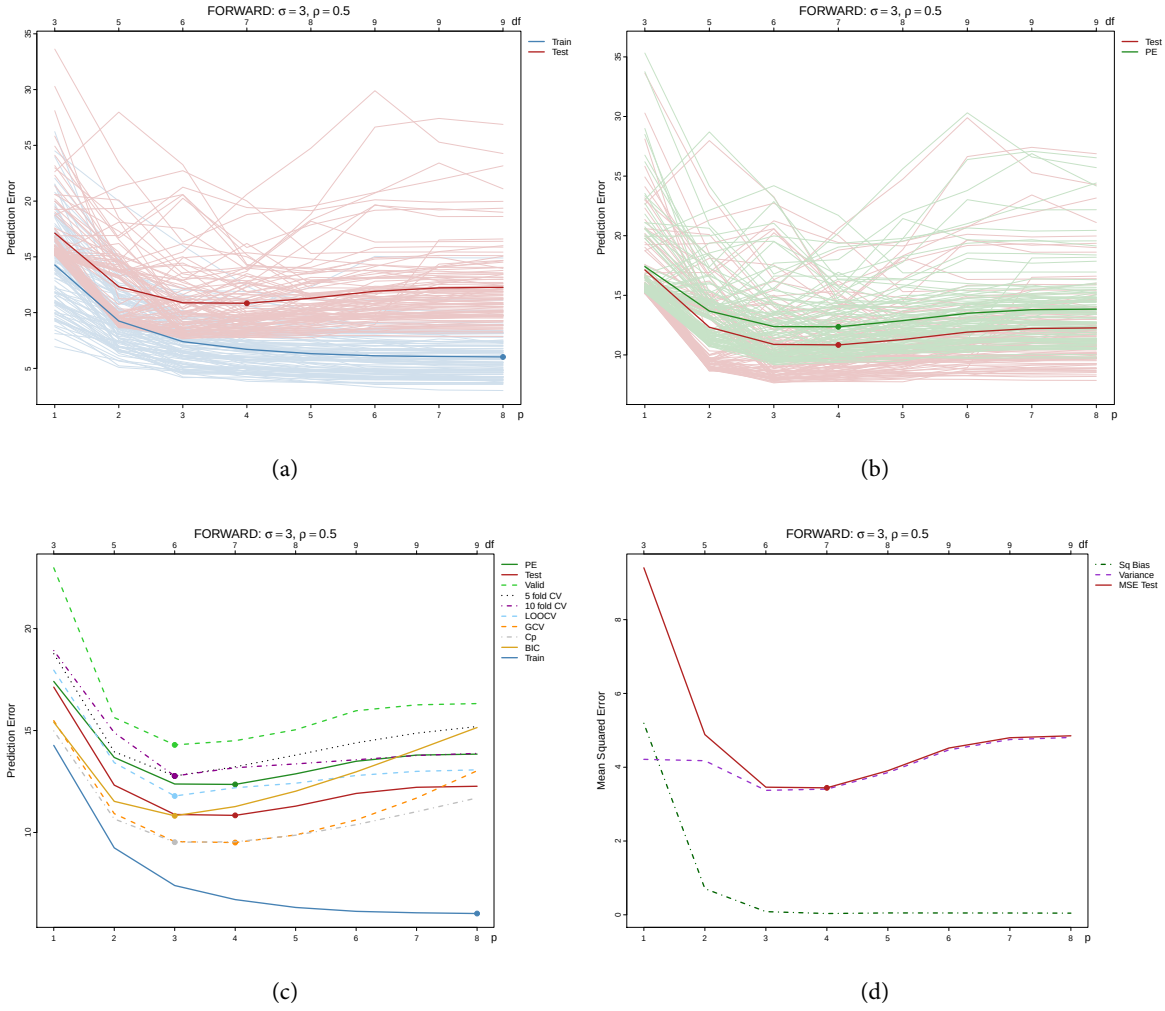


Figure (6.2.1) Prediction error for forward selection when $\sigma = 3$ and $\rho = 0.5$. PE is plotted against the number of variables p and the effective DF is shown on the top axes. The top panels show the PE for each sample (light curves) and the average over the 100 samples (thick curves), with the test error and training error in (a) and the test error and true PE in (b). Estimates of PE are shown in (c) and (d) displays the decomposition of MSE into the squared bias and variance of the predictions, where $PE = MSE + \sigma^2$

For each method, the training error and test error comparisons are shown in panel *(a)*. The lighter curves represent the error for each sample and the average is shown by the thicker curves. The variation

in the lighter curves reveals that the training error has high variance at low model complexity and the variance decreases as we fit the model harder. In contrast, the variability of the test error increases with complexity. The average training error decreases steadily, while the test error begins to increase as we start overfitting. The difference between the two thick curves shows the average optimism of the training error. Similarly, panel *(b)* displays the test error and the true PE for each sample along with the averages. To be clear, the true PE is calculated as described in point 1 of Section 6.1.2, where $PE = MSE + \sigma^2$ and the test error is calculated as described in point 3a. In each case, the curves are similar in terms of their values and their shape. The average curves and the position of their minimum value are almost identical, with the test error only slightly under estimating the PE. The instability of forward selection due to its discrete nature can be seen by the high variability and wild behaviour of the PE. Ridge regression and the LASSO display smooth curves with stabilized variance and less erratic behaviour.

Averages of PE estimates are shown in panel *(c)*. While some estimates may fail to accurately predict the correct value PE, they mostly perform well in identifying the position of the minimum value. This makes them well suited for model selection and generally a test set is preferred for model assessment. The validation set usually provides an over estimate of PE but still manages to identify the minimum position adequately. The C_p and GCV are almost identical, they tend to under estimate PE and select a slightly more complex model. AIC also showed similar results for the model selection since $AIC \propto C_p$, but because of the scale difference its value does not approach the true PE so it is not shown. BIC often has its minimum at a model with lower complexity and then increases dramatically as the model complexity increases. LOOCV, 10-fold CV and 5-fold CV are very similar to each other. They perform exceptionally well as estimates of PE, outperforming even the test error, and also excel in identifying the optimal model. The performance, in terms of prediction and selection, for the best models chosen using each of these criteria are shown in the next subsection.

The decomposition of PE is shown in panel *(d)*. The solid red curve is the MSE, which is obtained by adding the two dashed curves representing the squared bias and the variance of the predictions. The values were calculated as described in point 2 of Section 6.1.2. The difference between the MSE in *(d)* and the PE in *(b)* is an amount of about $\sigma^2 = 9$. This confirms that PE is composed of the irreducible error, squared bias and variance. The plots show clearly that models with low complexity have high bias and low variance, while more complex models have low bias and high variance.

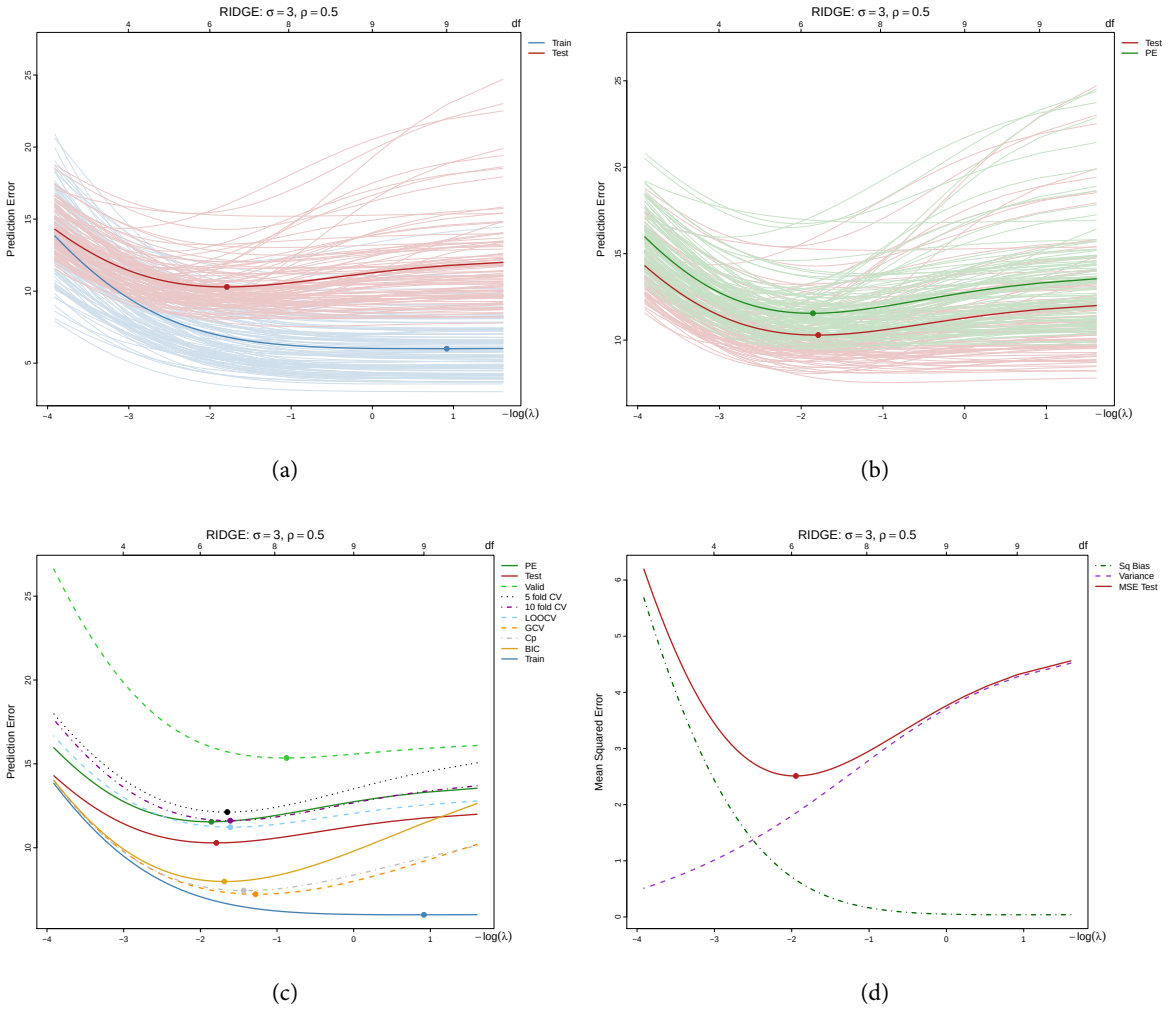


Figure (6.2.2) Prediction error for ridge regression when $\sigma = 3$ and $\rho = 0.5$. PE is plotted against the $-ln(\lambda)$ and the effective DF is shown on the top axes. The top panels show the PE for each sample (light curves) and the average over the 100 samples (thick curves), with the test error and training error in (a) and the test error and true PE in (b). Estimates of PE are shown in (c) and (d) displays the decomposition of MSE into the squared bias and variance of the predictions, where $PE = MSE + \sigma^2$

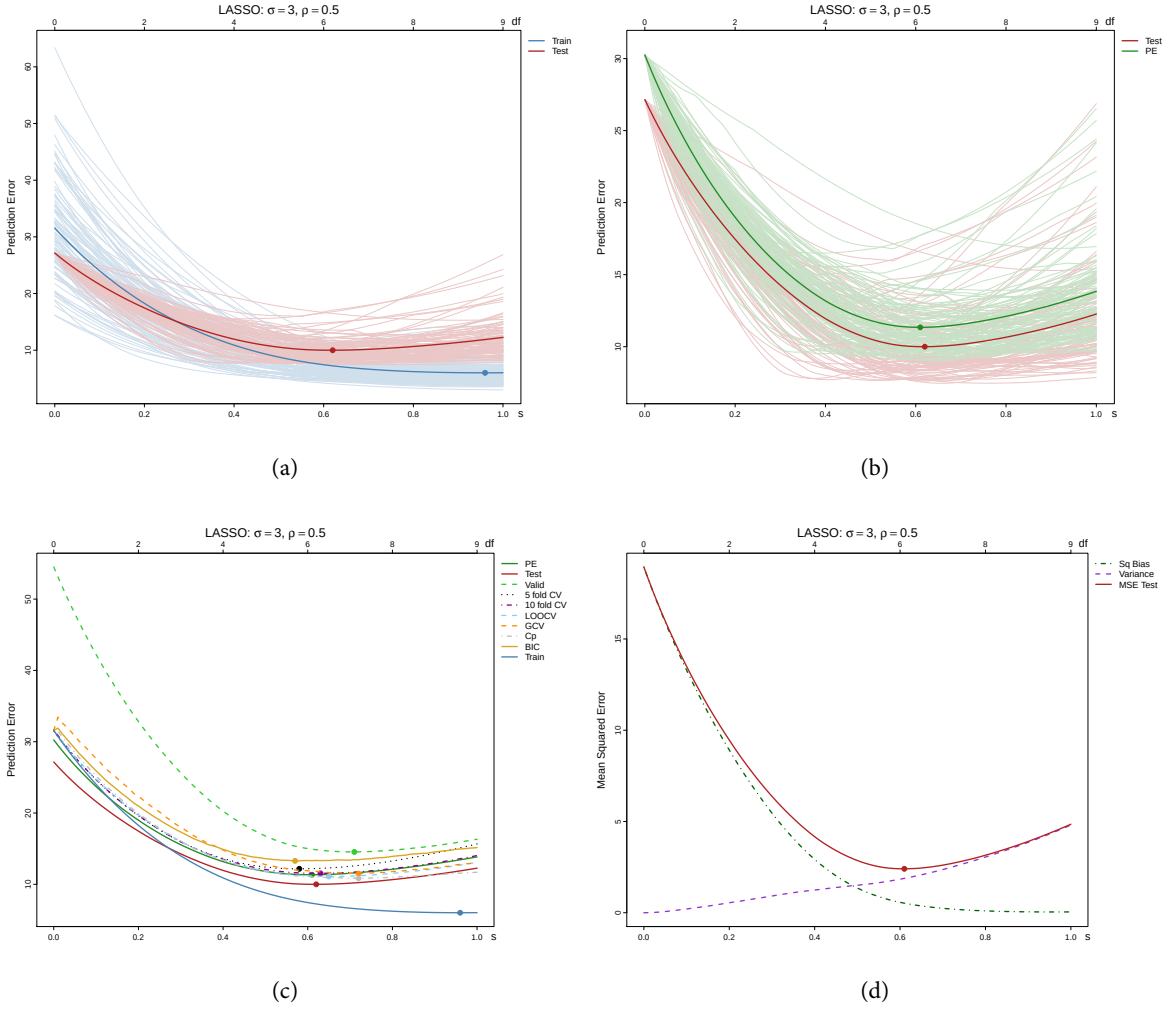
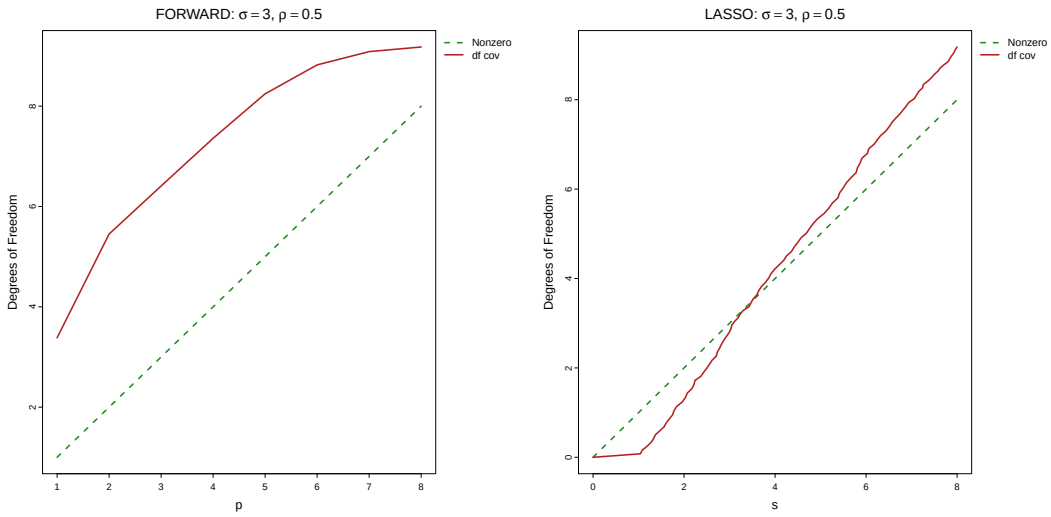


Figure (6.2.3) Prediction error for LASSO when $\sigma = 3$ and $\rho = 0.5$. PE is plotted against s and the effective DF is shown on the top axes. The top panels show the PE for each sample (light curves) and the average over the 100 samples (thick curves), with the test error and training error in (a) and the test error and true PE in (b). Estimates of PE are shown in (c) and (d) displays the decomposition of MSE into the squared bias and variance of the predictions, where $PE = MSE + \sigma^2$

Here we see again that forward selection has larger variance than ridge regression or the LASSO. However, the estimation bias of ridge regression and the LASSO due to the constrained model space is clear, with the bias increasing substantially as the constraint region is reduced. The LASSO has larger bias than ridge regression because the radius of the constraint region is a lot smaller, $\|\underline{\alpha}\|_1 \leqslant \|\underline{\alpha}\|^2$. Ridge regression also has lower bias since all the variables are retained in the model. While the LASSO drops variables from the model, the prediction vs selection dilemma is clear - selecting a smaller model comes at the cost of increased bias. Forward selection is only biased when the model is underfitted (that is, $p = 1, 2$) and displays no bias for the true model and any of the overfitted models. This makes sense since each model is the best fitting least squares model of that size. However, more DF than a least squares fit have been used to identify the best fitting models. The secondary x-axis is the average DF for each model, estimated using the covariance formula (3.1.9). It appears that on average, about 2 to 3 DF are spent by the adaptive search of forward selection. Figure 6.2.4 show the average DF in comparison with the number of nonzero variables for forward selection and the LASSO when $\sigma = 3$ and $\rho = 0.5$. The number of nonzero variables under estimates the DF for forward selection, while it is a close approximation of the DF of the LASSO. The LASSO also performs an adaptive search, however, the DF saved by the shrinkage of estimates balances out with the DF spent on the search.



(a) (b)

Figure (6.2.4) Comparison of DF and the number of nonzero variables for (a) forward selection and (b) the LASSO. The number of nonzero variables is a good approximation of the DF of the LASSO but heavily under estimates the DF of forward selection.

Although the plots above are all shown for $\sigma = 3$ and $\rho = 0.5$, similar results were observed for each

method, for all combinations of σ and ρ. The effects of increasing the noise and/or the correlation on the bias and the variance of each method is shown in Table 6.2.2. The maximum variance and squared bias over the solution path of each method is shown, where the null model and full model are not included. The variance and squared bias of least squares and oracle least squares is shown for comparison. For all methods, the variance increases with σ but is little affected by increasing ρ. This verifies that the predictions do not suffer from collinearity among the predictor variables. In particular, the full least squares model has identical variance and squared bias for all ρ. The bias of the LASSO increases substantially as ρ increases, while the bias of ridge regression decreases.

		Variance			Squared Bias		
Method	ρ	$\sigma = 1$	$\sigma = 3$	$\sigma = 6$	$\sigma = 1$	$\sigma = 3$	$\sigma = 6$
FORWARD	0	2.533	4.643	19.072	4.944	4.145	3.616
	0.5	1.873	4.750	18.868	6.403	5.191	3.171
	0.9	2.555	4.735	18.405	1.674	1.048	0.465
LASSO	0	0.509	4.700	18.890	14.098	14.022	13.985
	0.5	0.510	4.705	18.911	18.460	18.302	18.084
	0.9	0.746	4.739	18.981	35.446	34.658	33.254
RIDGE	0	0.546	4.627	18.512	6.790	6.719	6.620
	0.5	0.501	4.520	18.094	5.822	5.683	5.483
	0.9	0.407	3.651	14.609	4.019	3.869	3.658
OLS	0	0.534	4.807	19.227	0.005	0.045	0.180
	0.5	0.534	4.807	19.227	0.005	0.045	0.180
	0.9	0.534	4.807	19.227	0.005	0.045	0.180
ORACLE	0	0.149	1.342	5.368	0.001	0.005	0.021
	0.5	0.142	1.281	5.122	0.002	0.014	0.058
	0.9	0.171	1.536	6.146	0.002	0.021	0.084

Table (6.2.2) Maximum variance and squared bias of predictions along the path of each method. The null model and the full model (least squares) are excluded from the path of forward selection, ridge regression and the LASSO. The variance and squared bias for least squares and oracle least squares is shown for comparison.

For each method, the minimum MSE occurs at the best balance of bias and variance and indicates which models should be selected. The models selected by ridge regression and the LASSO have lower variance and slightly more bias than the best forward selection model. Table 6.2.3 shows the variance and squared bias at the minimum MSE as σ and ρ are varied. Forward selection appears to have lower variance when $\sigma = 1$. As σ increases, ridge regression and the LASSO tend to select models with lower complexity. The increase in variance shifts the minimum MSE towards the larger bias where the DF is lower. The LASSOs

increase in bias, as ρ increases, appears to have little effect on the selected model. Conversely, the DF for forward selection increases as σ or ρ are increased. The number of variables and DF at the minimum MSE are shown in Table 6.2.4. The entire situation is seen clearly when presented graphically as in Figure 6.2.5.

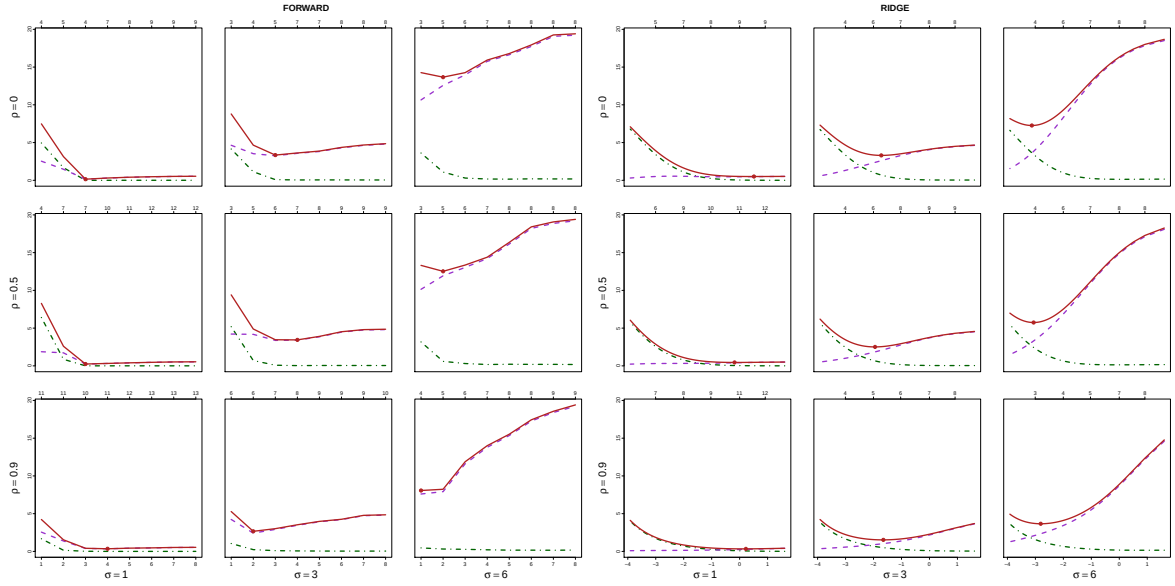
		Variance			Squared Bias		
ρ	Method	$\sigma = 1$	$\sigma = 3$	$\sigma = 6$	$\sigma = 1$	$\sigma = 3$	$\sigma = 6$
0	FORWARD	0.149	3.255	12.570	0.001	0.090	1.093
	LASSO	0.278	2.142	4.450	0.055	0.553	3.055
	RIDGE	0.488	2.602	3.624	0.015	0.699	3.639
0.5	FORWARD	0.242	3.407	11.927	0.004	0.035	0.597
	LASSO	0.269	1.883	4.577	0.044	0.523	2.268
	RIDGE	0.409	1.852	3.301	0.044	0.659	2.442
0.9	FORWARD	0.347	2.427	7.605	0.004	0.236	0.465
	LASSO	0.387	2.199	4.385	0.016	0.160	0.783
	RIDGE	0.274	1.033	2.342	0.063	0.497	1.322

Table (6.2.3) Variance and squared bias of predictions at the minimum MSE

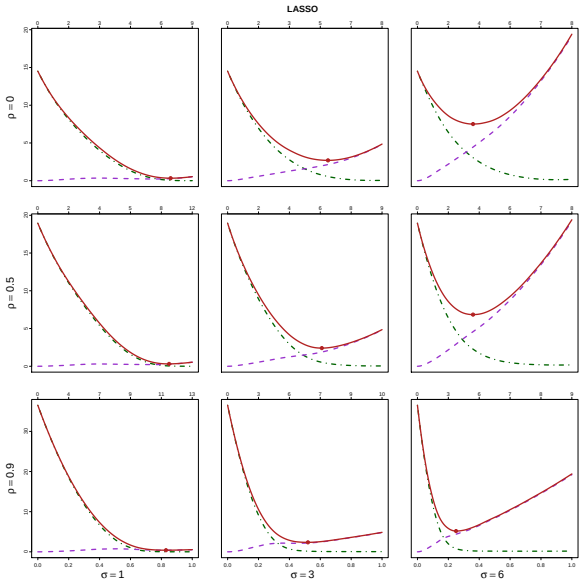
		Number Variables			df		
Method	ρ	$\sigma = 1$	$\sigma = 3$	$\sigma = 6$	$\sigma = 1$	$\sigma = 3$	$\sigma = 6$
FORWARD	0	3.00	3.00	2.00	4.43	5.26	4.96
	0.5	3.00	4.00	2.00	6.82	7.36	5.24
	0.9	4.00	2.00	1.00	10.98	5.93	3.98
LASSO	0	5.70	5.50	4.16	6.87	5.47	3.63
	0.5	5.70	5.26	4.13	9.21	5.69	3.82
	0.9	6.65	5.60	3.92	11.73	6.82	3.91
RIDGE	0	8.00	8.00	8.00	9.06	6.49	3.94
	0.5	8.00	8.00	8.00	11.04	6.39	3.93
	0.9	8.00	8.00	8.00	11.14	5.27	3.10

Table (6.2.4) Number of variables and estimated degrees of freedom at the minimum MSE

The average coefficient profiles and their probabilities of inclusion are shown in Figure 6.2.6 and Figure 6.2.7. From the coefficient profiles, we can see how ridge regression shrinks parameters proportionally to their size. The LASSO shrinkage is constant, all parameters are shrunken equally, but additional bias is introduced since they are shrunk all the way to zero. In this example, it is clear from the coefficient profiles which of the variables are important. In other situations, particularly when $p > n$, examining the inclusion probabilities can clarify which variables should be included. Supposedly, selecting the variables based on these probabilities can improve the variable selection properties of the LASSO.

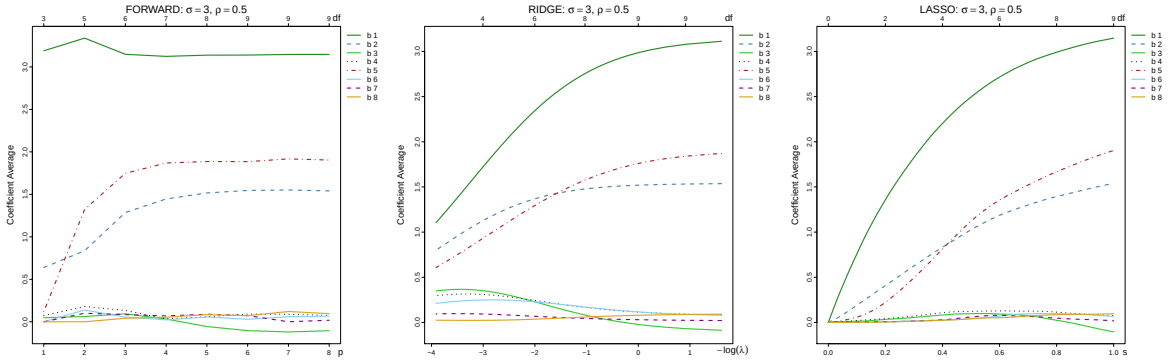


(a) (b)



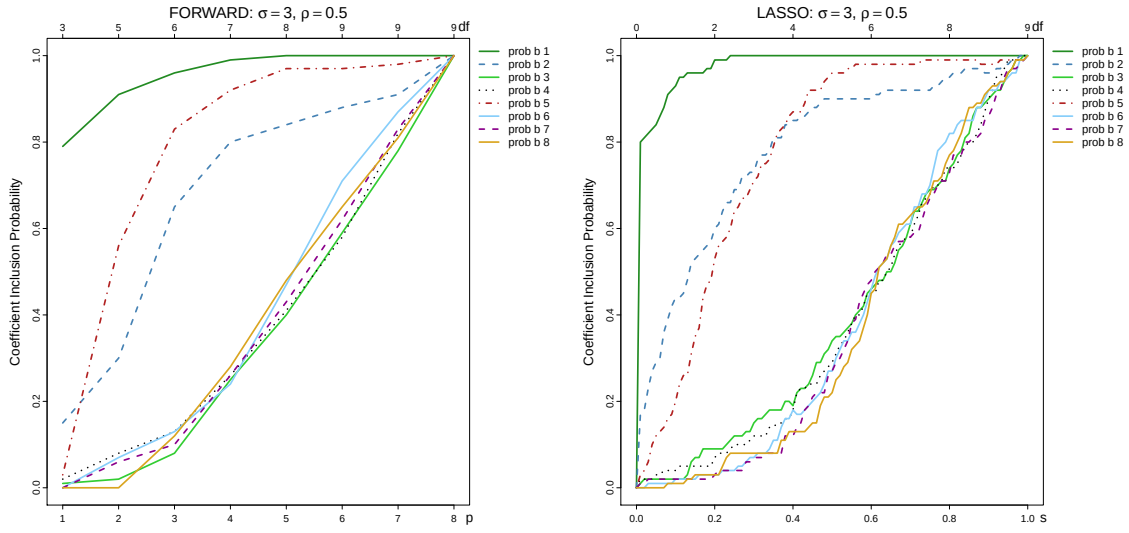
(c)

Figure (6.2.5) *MSE, squared bias and variance as σ and ρ are varied for (a) forward selection, (b) ridge regression and (c) the LASSO*



(a) (b) (c)

Figure (6.2.6) Average coefficient profiles for forward selection, ridge regression and the LASSO



(a) (b)

Figure (6.2.7) Coefficient inclusion probabilities for forward selection and the LASSO

Model Selection

Figures 6.2.8, 6.2.9, 6.2.10 show how each method performs when a particular CV method or information criterion is used to select the best model from the solution path. The box and whisker plots show the distribution of the MSE of predictions. For forward selection and the LASSO, the probability of selecting the correct model is shown simultaneously with the use of a colour scale, where darker shades indicate higher probabilities. The figures are shown for $\sigma = 3$ and $\rho = 0.5$ and the discussion below describes how results differ as σ and ρ are varied.

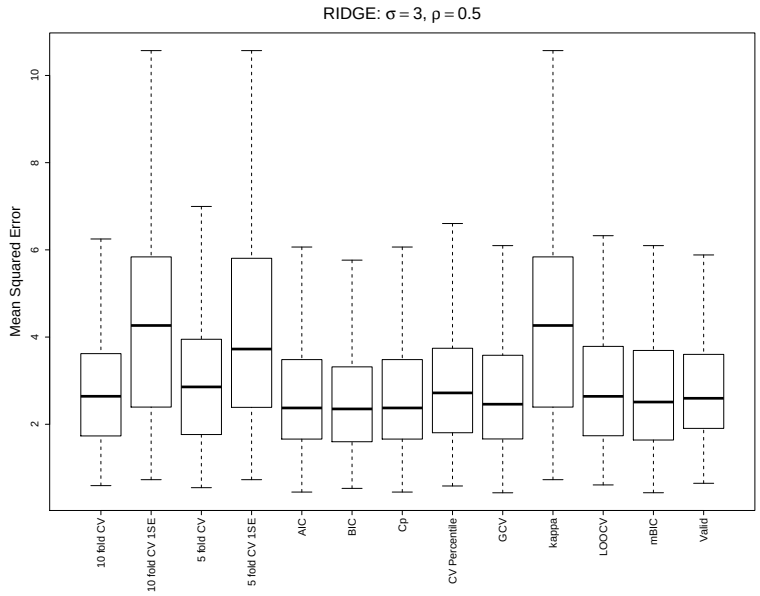


Figure (6.2.8) Model selection for ridge regression: MSE when using CV and information criteria

For each method, K-fold CV, LOOCV, GCV, AIC and C_p are all equivalent, with GCV, AIC and C_p almost identical. Although the validation set approach has the best MSE with the lowest variation, these methods offer a satisfactory alternative. They select the best model for prediction but do not do very well for variable selection. For forward selection, using K-fold CV with the 1 SE rule works particularly well when the noise level is low, showing excellent variable selection and low PE. The percentile CV method and kappa coefficient also perform well in both aspects. The BIC is slightly more variable than these two methods and has slightly lower selection accuracy, although it does seem less affected by the increase in noise. For the LASSO, the BIC and the percentile CV method perform similarly, showing good prediction accuracy and improved variable selection. The 1 SE rule improves the variable selection further and the kappa coefficient even more so. However, the resulting models using these methods are very biased and highly variable. The 1 SE rule appears to worsen as the noise is increased, while the kappa coefficient seems

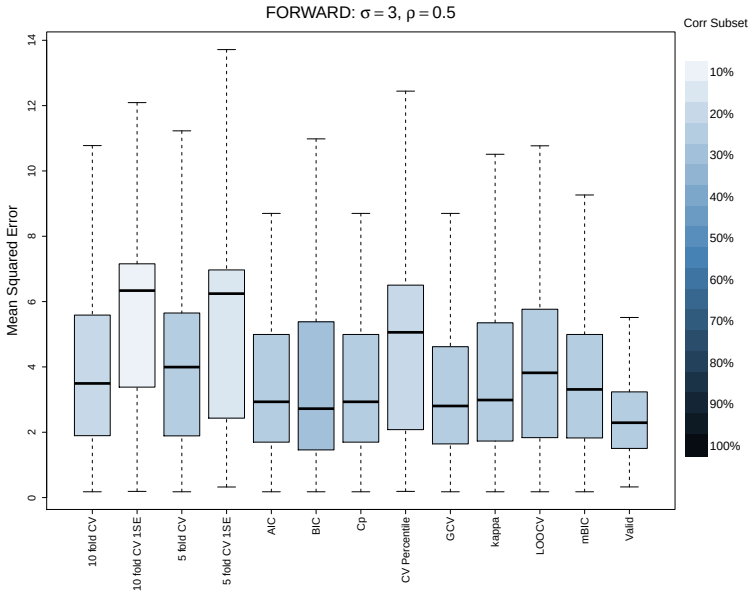


Figure (6.2.9) Model selection for forward selection: MSE and probability of selecting the correct model when using CV and information criteria

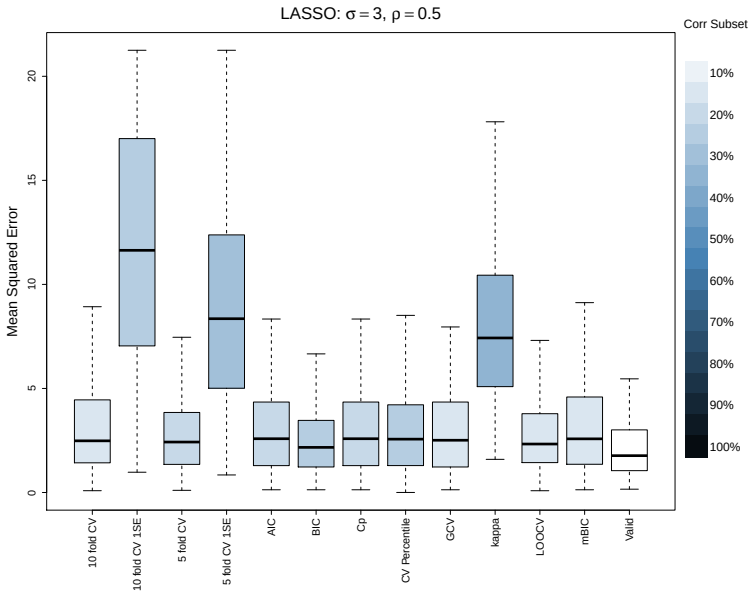


Figure (6.2.10) Model selection for the LASSO: MSE and probability of selecting the correct model when using CV and information criteria

more susceptible to increases in correlation. The modified BIC does not appear to offer any improvement.

Performance

Prediction

The performance of each method is examined when using either 5-fold CV or the kappa coefficient for model selection. Table 6.2.5 shows the median MSE of predictions and the probability of selecting the correct model for each method as σ and ρ are varied. Least squares and oracle least squares are included for comparison. The accuracy of the median MSE is assessed by bootstrap standard errors, calculated as described in Section 6.1.2. The MSE and probability of selecting the correct model is also presented graphically in Figure 6.2.11 for $\sigma = 3$ and $\rho = 0.5$.

As expected, oracle least squares has the highest prediction accuracy in all scenarios, with the MSE generally increasing slightly as σ and ρ increase. The larger MSE and variance of least squares is due to overfitting and it is clear that the predictions are not affected by collinearity. When using CV, the LASSO, forward selection and ridge regression all have lower MSE than least squares in all scenarios except one, the low noise orthogonal design, where ridge regression performs poorly. Forward selection performs the best when $\sigma = 1$ and $\rho = 0, 0.5$, with high prediction accuracy and good selection performance, but suffers higher MSE and high variance in other scenarios, with variance larger than least squares in some cases. The LASSO is a close competitor when $\sigma = 1$, but really shines when $\sigma = 3$. In this case, it has high prediction accuracy and selection performance nearly as good as forward selection. Also, for orthogonal designs, the MSE for the LASSO does appear to be close to that of oracle least squares (near minimax optimality). Ridge regression performs best in the high noise scenario when $\sigma = 6$ and handles collinearity superbly, outperforming other methods when $\rho = 0.9$ (except in the low noise case). Model selection using the kappa coefficient improves the performance of forward selection when $\sigma = 1, 3$, increasing the probability of selecting the correct model and resulting in lower MSE as well as low variance which is smaller than that of least squares. The LASSOs selection performance is remarkably improved when using the kappa coefficient for $\sigma = 1, 3$, such that it surpasses that of forward selection. However, the MSE increases substantially due to large bias. The kappa coefficient is not suitable for ridge regression, which seems to have selected the full (least squares) model.

		$\sigma = 1$		$\sigma = 3$		$\sigma = 6$	
ρ	Method	Median MSE	PCS	Median MSE	PCS	Median MSE	PCS
				5-fold *CV*			
0	FORWARD	0.174 (0.022)	0.70	4.042 (0.549)	0.20	13.391 (1.265)	0.03
	LASSO	0.320 (0.028)	0.21	2.817 (0.362)	0.14	10.261 (0.915)	0.04
	RIDGE	0.489 (0.036)	0.00	3.554 (0.301)	0.00	8.802 (0.461)	0.00
0.5	FORWARD	0.174 (0.022)	0.55	3.995 (0.496)	0.22	10.893 (0.996)	0.05
	LASSO	0.306 (0.026)	0.14	2.425 (0.274)	0.16	7.804 (0.598)	0.06
	RIDGE	0.405 (0.032)	0.00	2.857 (0.189)	0.00	7.523 (0.472)	0.00
0.9	FORWARD	0.336 (0.044)	0.30	3.542 (0.407)	0.03	7.613 (0.982)	0.00
	LASSO	0.330 (0.022)	0.07	2.127 (0.131)	0.08	5.818 (0.745)	0.04
	RIDGE	0.334 (0.022)	0.00	1.720 (0.104)	0.00	4.480 (0.448)	0.00
				κ Coefficient			
0	FORWARD	0.146 (0.019)	0.84	3.051 (0.249)	0.28	13.416 (1.350)	0.05
	LASSO	2.307 (0.145)	0.84	6.204 (0.252)	0.32	8.654 (0.762)	0.04
	RIDGE	0.474 (0.048)	0.00	4.266 (0.420)	0.00	17.062 (1.705)	0.00
0.5	FORWARD	0.136 (0.018)	0.78	2.986 (0.327)	0.20	13.220 (1.023)	0.02
	LASSO	2.724 (0.140)	0.83	7.426 (0.333)	0.32	9.201 (0.399)	0.04
	RIDGE	0.474 (0.047)	0.00	4.266 (0.412)	0.00	17.062 (1.642)	0.00
0.9	FORWARD	0.256 (0.031)	0.37	2.623 (0.321)	0.03	10.696 (0.948)	0.00
	LASSO	8.930 (0.599)	0.26	12.889 (1.013)	0.03	9.144 (1.379)	0.02
	RIDGE	0.474 (0.045)	0.00	4.266 (0.418)	0.00	17.062 (1.666)	0.00
				Least Squares			
0	OLS	0.474 (0.045)	0	4.266 (0.447)	0	17.062 (1.756)	0
0.5		0.474 (0.046)	0	4.266 (0.429)	0	17.062 (1.661)	0
0.9		0.474 (0.045)	0	4.266 (0.414)	0	17.062 (1.707)	0
0	Oracle	0.115 (0.014)	1	1.038 (0.117)	1	4.153 (0.485)	1
0.5		0.107 (0.010)	1	0.965 (0.087)	1	3.860 (0.350)	1
0.9		0.157 (0.017)	1	1.410 (0.152)	1	5.638 (0.544)	1

Table (6.2.5) Median MSE (with bootstrap standard errors) and probability of selecting the correct subset when using CV and the kappa coefficient for model selection. Least squares and oracle least squares are shown for comparison.

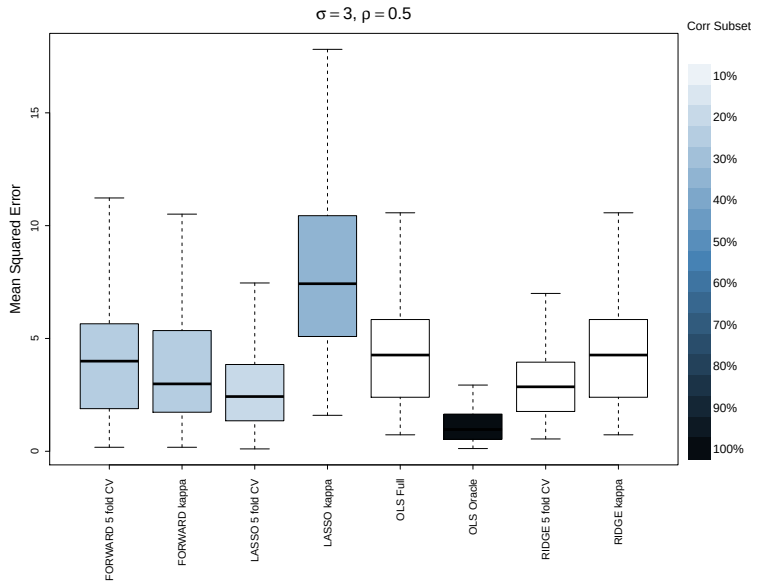


Figure (6.2.11) Comparison of prediction and selection performance between different methods: MSE and probability of selecting the correct model as σ and ρ vary

Estimation

The variance and squared bias of the parameter estimates is shown in Table 6.2.6. The effect of collinearity on the least squares estimates is seen clearly by the inflated variance. The LASSO, forward selection and ridge regression estimates all have much lower variance than the LSEs when using either CV or the kappa coefficient. When using CV, forward selection has the lowest bias in all scenarios. It also has the lowest variance when $\sigma = 1$ and $\rho = 0, 0.5$ but in other scenarios it is highly variable, compared to the LASSO and ridge regression. The LASSO appears to have the lowest variance in most of these scenarios. It also has highest bias in most of the scenarios, although the bias is still within an acceptable range. Using the kappa coefficient with forward selection and the LASSO, in almost every scenario, the variance is reduced and the bias is increased - the LASSOs bias increasing considerably.

Figures 6.2.12 and 6.2.13 examine the distributions of the parameter estimates. For each parameter, a box and whisker plot depicts the range of the estimates value, along with a colour scale showing its inclusion probability, for each method. Histograms with the normal probability density function are also shown for each method. Results are shown for $\sigma = 3$ and $\rho = 0.5$. Every method includes β_1, the largest parameter, with high probability (tending to 1). Forward selection and least squares have the largest variability. When using kappa, the LASSO over shrinks β_1 so that it is biased toward zero. Examining the histograms, each method displays a fairly normal distribution for β_1. Only LASSO using CV includes β_2,

		$\sigma = 1$		$\sigma = 3$		$\sigma = 6$	
ρ	Method	Variance	Sq Bias	Variance	Sq Bias	Variance	Sq Bias
				5-fold *CV*			
0	FORWARD	0.249	0.002	4.616	0.155	15.100	0.666
	LASSO	0.293	0.093	2.563	0.998	7.855	3.760
	RIDGE	0.515	0.013	3.234	0.589	6.444	2.674
0.5	FORWARD	0.432	0.003	5.502	0.240	17.791	0.621
	LASSO	0.437	0.062	2.792	0.751	8.210	2.517
	RIDGE	0.713	0.039	3.551	0.619	9.062	1.911
0.9	FORWARD	3.121	0.060	21.092	1.211	56.139	2.388
	LASSO	1.826	0.294	11.024	1.665	16.432	3.861
	RIDGE	2.533	0.405	8.360	2.011	24.339	2.622
				κ Coefficient			
0	FORWARD	0.199	0.001	3.513	0.120	15.797	0.281
	LASSO	0.374	2.204	1.100	5.167	5.192	5.234
	RIDGE	0.538	0.005	4.844	0.048	19.376	0.192
0.5	FORWARD	0.233	0.002	5.583	0.251	22.198	0.362
	LASSO	0.391	2.139	1.214	5.100	3.642	5.270
	RIDGE	0.839	0.007	7.548	0.063	30.191	0.253
0.9	FORWARD	2.940	0.136	20.818	0.667	97.168	2.011
	LASSO	1.678	5.546	2.865	7.221	16.175	5.338
	RIDGE	4.454	0.035	40.088	0.312	160.352	1.246
				Least Squares			
0	OLS	0.538	0.005	4.844	0.048	19.376	0.191
0.5		0.839	0.007	7.548	0.063	30.191	0.253
0.9		4.454	0.035	40.088	0.312	160.352	1.246
0	Oracle	0.147	0.001	1.326	0.006	5.302	0.022
0.5		0.163	0.001	1.469	0.013	5.878	0.052
0.9		0.781	0.003	7.024	0.026	28.096	0.104

Table (6.2.6) Total variance and squared bias of parameter estimates when using CV and the kappa coefficient for model selection

the smallest nonzero parameter, with high probability. The inclusion probabilities are significantly lower for forward selection and LASSO using kappa, and the histograms for these methods also display a significant deviation from the normal distribution. Similarly to β_1, forward selection has large variance and LASSO using kappa results with β_2 biased towards zero. β_5 is also included by LASSO using CV with high probability. The inclusion probabilities are somewhat lower for forward selection and LASSO using kappa, although slightly higher than for β_2. These methods show the same behaviour as for the other nonzero parameters, forward selection is highly variable and LASSO using kappa is heavily biased towards zero. Neither forward selection nor the LASSO exhibit normal distributions for β_5. While it is not clear from the plots, LASSO using CV selects each of the zero parameters with probability of approximately 0.4; forward selection using CV selects them with probability of approximately 0.2; and using kappa, both LASSO and forward selection select them with low probability. The zero parameter estimates are not normally distributed when using LASSO or forward selection.

Variable Selection

Figure 6.2.14 shows the variable selection performance of forward selection and the LASSO when $\sigma = 3$ and $\rho = 0.5$. The average number of nonzero parameter estimates that are included in the selected models are shown in panel *(a)*. The estimates that are correctly estimated as nonzero are coloured in green and those incorrectly estimated as nonzero are coloured in red. The LASSO using CV always overfits the model and usually includes more variables than the other methods. In this scenario, none of the methods perform exceptionally well in terms of variable selection. When $\sigma = 1$ and $\rho = 0, 0.5$, all three nonzero parameters are correctly estimated as nonzero for all methods, and the use of kappa yields the lowest false inclusion rate.

Panel *(b)* displays the probabilities of selecting the correct model, including the correct model in the selected subset of variables, and containing the correct model in the solution path. In this scenario, the correct model is included in the selected subsets with high probability when using CV for the LASSO. The same is true when $\sigma = 1$. However, for $\sigma = 1$, forward selection performs variable selection exceptionally well, and surprisingly, the true model does lie in the local search path of forward selection with high probability. Selection performance is very poor for the high noise scenarios when $\sigma = 6$, but LASSO does appear to perform slightly better than forward selection.

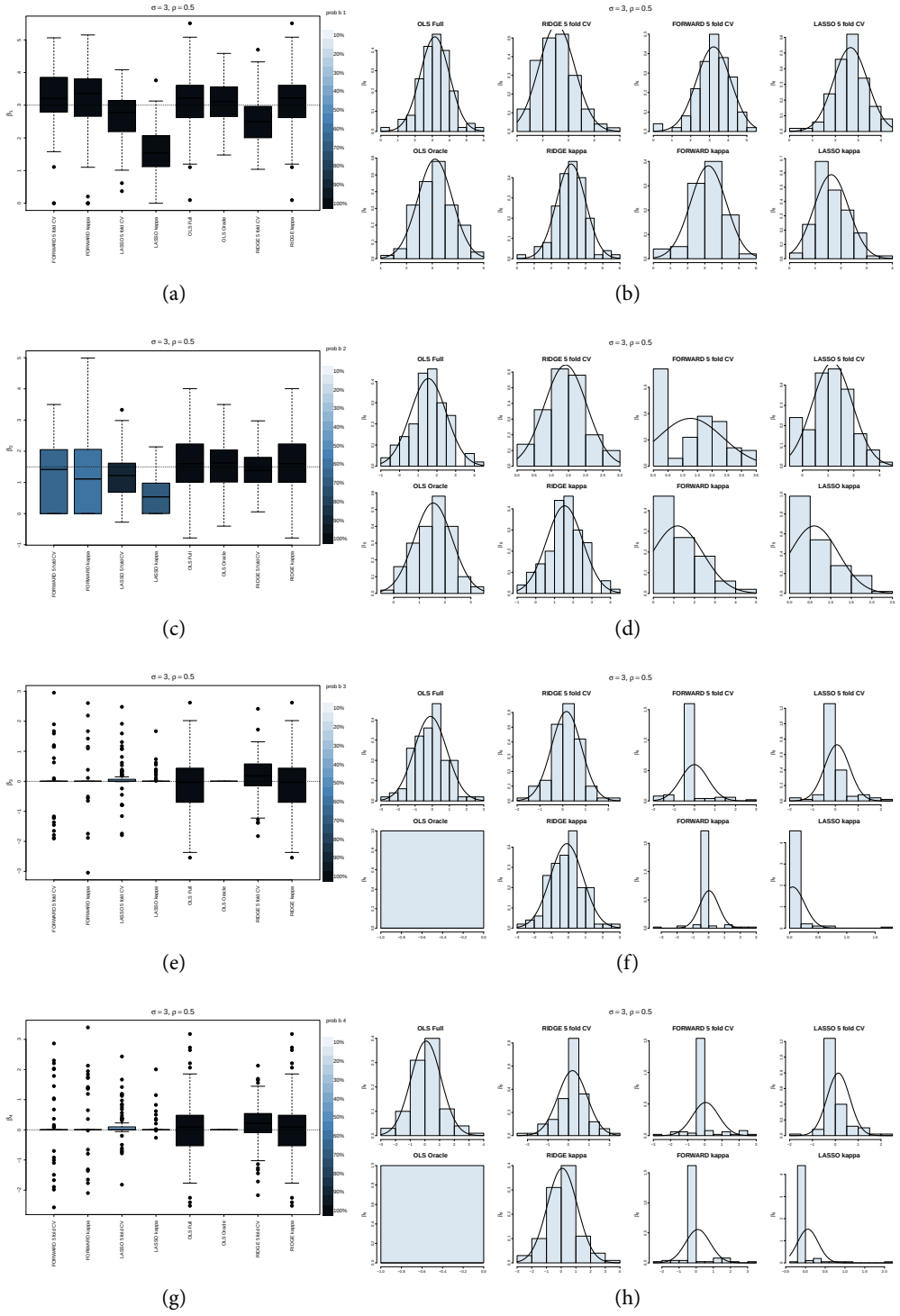


Figure (6.2.12) Comparison of parameter estimates $\beta_1 - \beta_4$ between different methods: value of estimates and their inclusion probabilities, along with histograms of their distributions for $\sigma = 3$ and $\rho = 0.5$

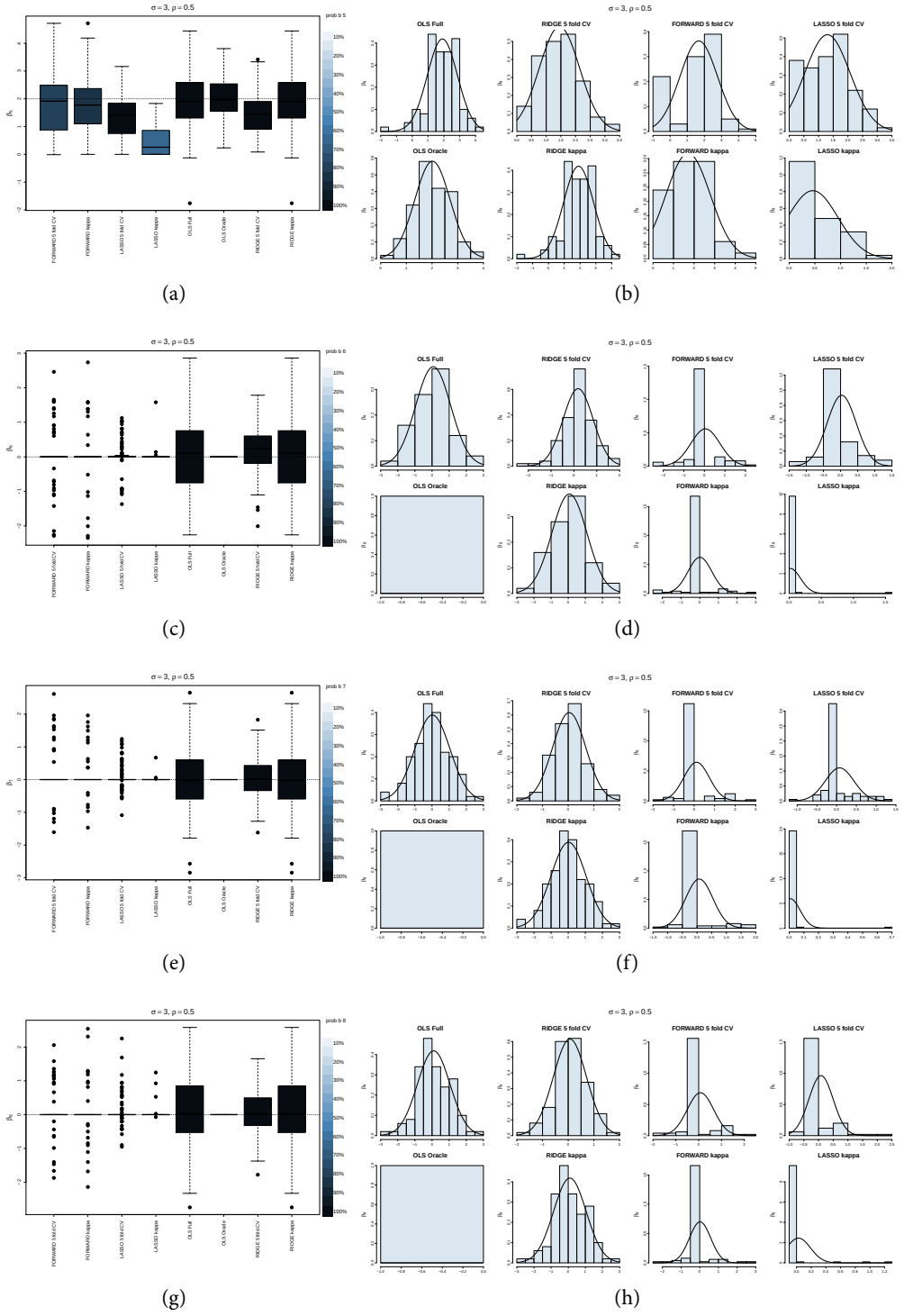
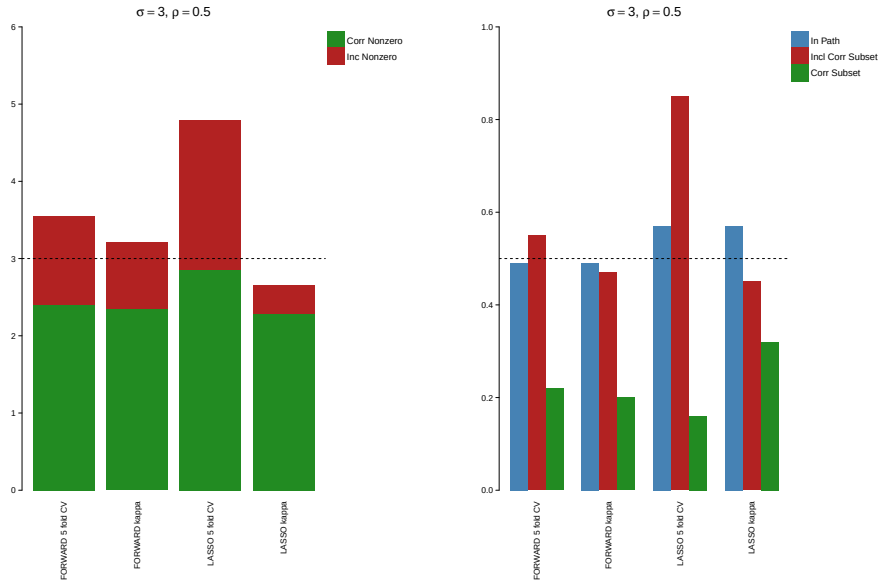


Figure (6.2.13) Comparison of parameter estimates $\beta_5 - \beta_8$ between different methods: value of estimates and their inclusion probabilities, along with histograms of their distributions for $\sigma = 3$ and $\rho = 0.5$



(a) (b)

Figure (6.2.14) Variable selection performance of forward selection and LASSO: the number of nonzero parameters is shown in (a) and the probabilities of selecting the correct model, containing the correct model and having the correct model in the solution path are shown in (b)

6.3 Oracle Procedures

This simulation study analyses the consistency, in terms of variable selection estimation and prediction, of the LASSO, some two-stage LASSO methods and some of the other shrinkage methods. Small sample results can be compared with the previous simulation study.

6.3.1 Data

The data generating process is similar to Section 6.2. The error variance is fixed at $\sigma = 3$ and the parameter vector is given by $\underline{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. Consistency is tested by allowing the sample size to grow, $n \in \{25, 50, 75, \ldots, 500\}$. The correlation is fixed at $\rho = 0.5$ but, in order to examine the capability of each method, different correlation structures are used:

- Power decay correlation, or AR(1) covariance structures, which will be denoted as AR: $\Sigma_{jk} = \rho^{|j-k|}$ for $j, k = 1, 2, \ldots, p$
- Constant positive correlation, or compound symmetry, which will be denoted as CS: $\Sigma_{jj} = 1$ for $j = 1, 2, \ldots, p$ and $\Sigma_{jk} = \rho$ for $j \neq k$
- A disturbed orthogonal design, which will be denoted as IR: $\Sigma_{jj} = 1$ for $j = 1, 2, \ldots, p$ and $\Sigma_{jk} = 0$

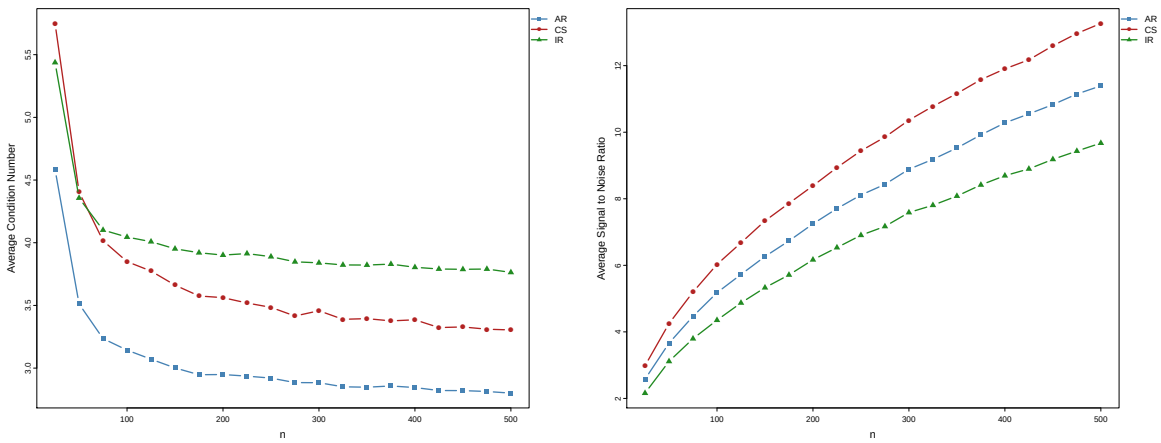
for $j \neq k$, except for $\Sigma_{jk} = \rho$, where $j \in \mathcal{D}$, the set of true nonzero parameters $\mathcal{D} = \{1, 2, 5\}$, and $k \in \mathcal{A}$ for some set of parameters $\mathcal{A} \subset \mathcal{D}^c$. Here, the set is chosen as $\mathcal{A} = \{3\}$.

According to Zhao & Yu (2006), the first two types of correlation satisfy the irrepresentable condition in equation (4.1.48),

$$\left\|\Sigma_{21}\Sigma_{11}^{-1}\,\text{sign}\,(\underline{\alpha}_{\mathcal{D}})\right\|_{\infty} \leqslant 1 - \epsilon \text{ for } \epsilon > 0.$$

The third correlation structure is constructed to violate the condition. Assuming an orthogonal design for the active set of variables (those indexed by $\mathcal{D}$, we have $\Sigma_{11}^{-1} = \Sigma_{11} = \mathbf{I}_d$ and since $\alpha_j \geqslant 0$ for all $j = 1, 2, \ldots, p$, the condition does not hold when $\|\Sigma_{21}\|_{\infty} > 1 - \epsilon$ for $\epsilon > 0$. Now, $\|\Sigma_{21}\|_{\infty} = \left\|\mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{D}}\right\|_{\infty} = \mathbf{z}_j^T\mathbf{Z}_{\mathcal{D}}$ such that $j = \arg\max_j \left|\mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{D}}\right|$. Therefore, allowing one of the irrelevant predictor variables to have strong correlations with all the variables in the active set will present a correlation structure dissatisfying the irrepresentable condition.

$N = 100$ training samples are simulated for each combination of n and the three correlation structures. Figure 6.3.1 shows the average SNR and condition number of the correlation matrix for the generated data. Both measures improve as the sample size increases and for all scenarios, the SNR and condition number are within a suitable range.



(a) (b)

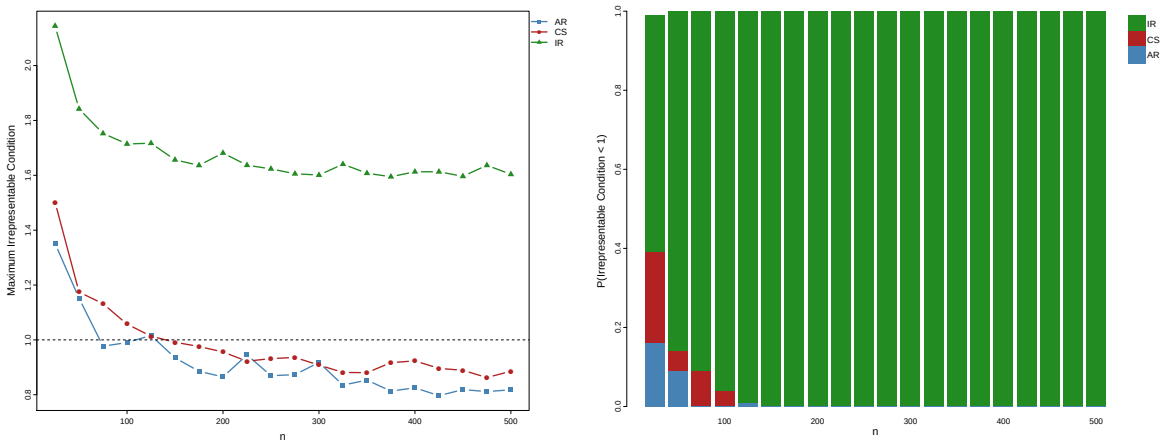
Figure (6.3.1) Condition numbers and signal to noise ratio for the generated data

The average of the compatibility condition and the restricted eigenvalue condition are given in Table 6.3.1 for the generated data. The compatibility condition holds in each case but the restricted eigenvalue condition does not hold for the IR correlation structure. The irrepresentable condition for the

generated data is depicted in Figure 6.3.2. Panel (a) shows the maximum value of the irrepresentable condition over the 100 samples. The probability that the condition is not satisfied is shown in panel (b), where the bars have been overlayed instead of stacked. For the AR and CS correlation structures, the condition does not hold when $n \leqslant 100$, but with low probability. For instance, when $n = 25$, the probability of a false irrepresentable condition is about 0.4 for CS structures and for AR structures the probability is under 0.2. For the IR structure, the condition is not met with probability tending to 1 (as designed).

	Condition	
Correlation Structure	Compatibility	Restricted Eigenvalue
AR	1.509	1.393
CS	2.041	1.885
IR	1.083	1.000

Table (6.3.1) Compatibility and restricted eigenvalue conditions for the generated data



(a) (b)

Figure (6.3.2) Irrepresentable condition for the generated data. The maximum value of the condition over the 100 samples is shown in (a) and the probability that the condition does not hold is shown in (b), where the bars have been overlayed rather than stacked.

6.3.2 Estimation and Model Selection

The model is estimated using least squares, oracle least squares with

$$\hat{f}(X) = X_1\beta_1 + X_2\beta_2 + X_5\beta_5,$$

the LASSO, relaxed LASSO, adaptive LASSO, SEA-LASSO, two-stage LASSO (using LASSO initial estimates for the adaptive LASSO), EN, adaptive EN, MCP and SCAD.

In each case, model selection is performed using 10-fold CV over a fixed grid of one tuning parameter and, where applicable, the second tuning parameter is held fixed. For methods excluding the two-stage LASSO and adaptive EN, which sequentially cross-validate over one tuning parameter, the kappa coefficient with 20 repetitions is also used to select the tuning parameter. Table 6.3.2 shows the tuning parameters that are considered in the study. For the LASSO and EN methods, the LAR algorithm was used with the primary tuning parameter selected over a fixed grid of $s = \|\underline{\alpha}\|_1 / \|\underline{\hat{\alpha}}\|_1 \in [0,1]$ with increments of 0.01. The relaxed LASSO is an exception, where the first p steps of the LAR algorithm is used instead. For the concave penalties of SCAD and MCP, the coordinate descent method by Breheny & Huang (2011) is used and the primary tuning parameter is selected over a grid of 100 λ values, equally spaced on the logarithmic scale.

	Tuning Parameter	
Method	Primary	Secondary
LASSO	$s \in [0,1]$	
Relaxed LASSO	LAR step $\in [1, p]$	$\phi = 0.3$
Adaptive LASSO	$s \in [0,1]$	$\zeta = 1$
SEA-LASSO	$s \in [0,1]$	$\zeta = 1$
Two-stage LASSO	$s \in [0,1]$	
EN	$s \in [0,1]$	$\lambda_2 = 0.5$
Adaptive EN	$s \in [0,1]$	$\lambda_2 = 0.5$
SCAD	$\lambda \in [e^{1.5}, e^{-6}]$	$\xi = 3.7$
MCP	$\lambda \in [e^{1.5}, e^{-6}]$	$\xi = 3$

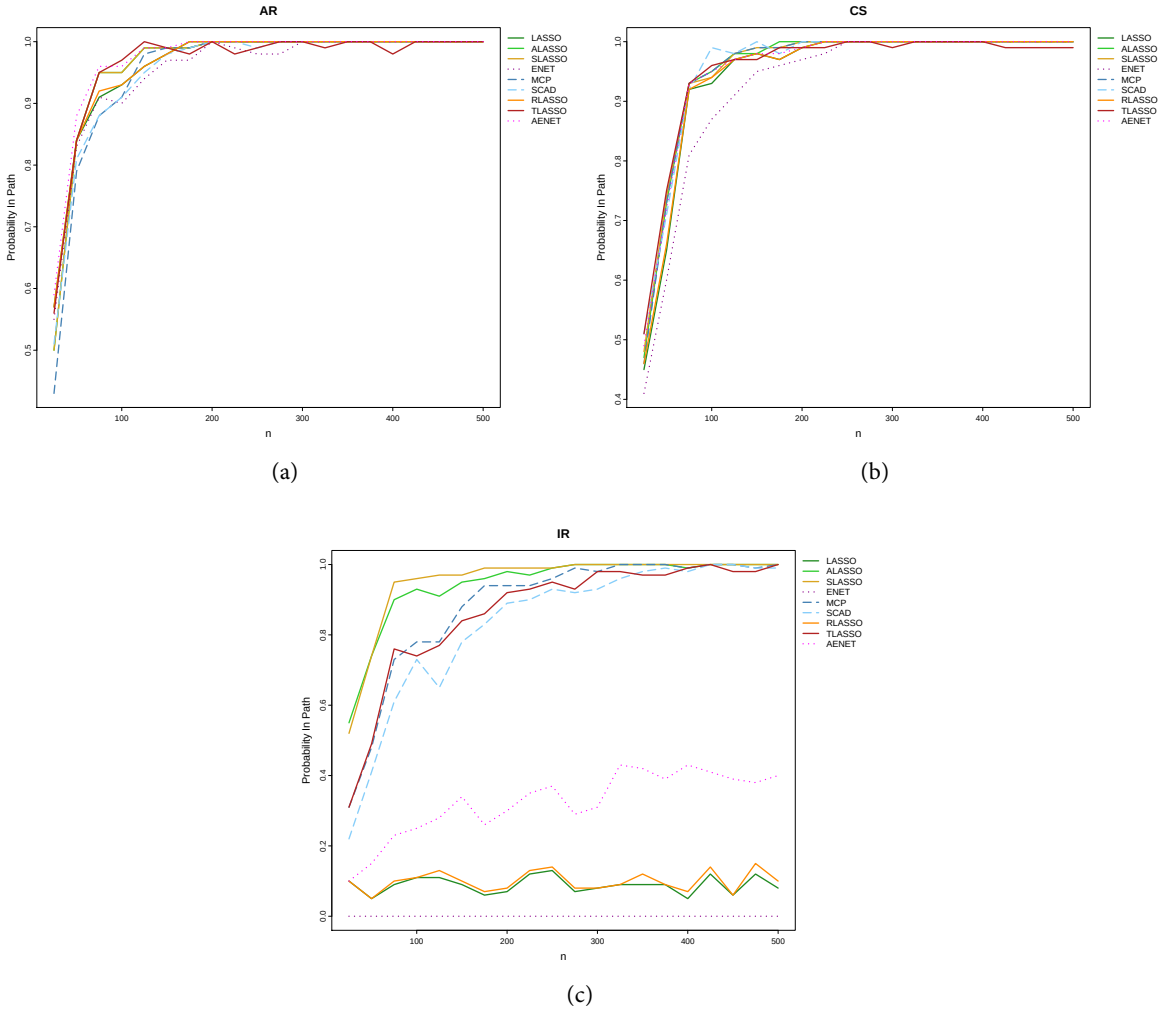
Table (6.3.2) Tuning parameters used for model selection

The variable selection performance measures described in Section 6.1.3 are calculated to determine selection consistency. Estimation consistency is tested by calculating the MSE of the parameter estimates as detailed in Section 6.1.1. Furthermore, persistence is assessed by calculating MSE using equation (6.1.1), along with its median and bootstrap standard errors.

6.3.3 Results

Consistency

Figure 6.3.3 reveals which methods are path consistent for each correlation structure. A method is termed path consistent if the correct model lies its solution path with probability tending to 1 as $n \to \infty$. The AR and CS correlation structures are shown in panels (a) and (b), respectively, where it is clear that all the methods are path consistent. In panel (c), we see that the LASSO is not path consistent when the irrepresentable condition does not hold. The EN, adaptive EN and relaxed LASSO also do not appear to be path consistent. However, it can only be said with certainty that they are not path consistent when holding the secondary tuning parameter fixed at the chosen value. It is possible that they may perform better when fixing the parameter at an alternative value or if it is chosen adaptively for each sample.



(a) (b)

(c)

Figure (6.3.3) Probability that the correct model is in the solution path

It is comforting to know that the correct solution exists within the path of a method but the question of whether we are able to recover that solution still remains. For this study, the tuning parameter is selected using either 10-fold CV or the kappa coefficient. The probability of selecting the correct model is shown in Figure 6.3.4 for each correlation structure with the models selected by CV on the left and those selected by the kappa coefficient on the right.

When using CV, only the adaptive EN, when applied to an AR correlation structure, appears to be consistent for variable selection. However, the concave penalties of MCP and SCAD approach a high probability of 0.8 for each correlation structure. Furthermore, they appear to be strictly increasing so that, if the sample size were to grow larger, they could indeed attain a probability of 1. The same argument could hold for the adaptive EN and the CS design. The relaxed LASSO also appears to be increasing for AR and CS designs, albeit at a slower rate. The LASSO and two-stage LASSO show no improvement in terms of variable selection as n increases. The same can be said for the EN, adaptive LASSO and SEA-LASSO, although these methods could possibly improve by adjusting the secondary tuning parameter.

Using the kappa coefficient for model selection, all methods appear to be selection consistent under AR and CS designs. However, when the irrepresentable condition is not met, only the adaptive LASSO, SEA-LASSO, MCP and SCAD achieve consistent selection, while the LASSO, relaxed LASSO and EN fail hopelessly.

Although not shown here, it is worth noting that, for all correlation structures and whether CV or the kappa coefficient are used, every method includes the correct model in the selected model with probability tending to one. That is, the correct model is a subset of the chosen model, so that these methods can be used for variable screening. The only exception is the adaptive EN, where the probability of including the correct model is nearly identical to the probability of selecting the correct model - that is, the adaptive EN does not overfit the model.

Figure 6.3.5 shows the MSE of the parameter estimates for each correlation structure when using CV and the kappa coefficient for model selection. When using CV, the MSE tends to zero as $n \to \infty$, indicating that the parameter estimates are consistent. The EN and adaptive EN are exceptions, although for AR and CS designs, the EN might approach 0 with a larger sample size. When using the kappa coefficient, only the relaxed LASSO, adaptive LASSO and SEA-LASSO are consistent for estimation. When the irrepresentable condition does not hold, the MSE of the MCP and SCAD estimates also appears to approach 0.

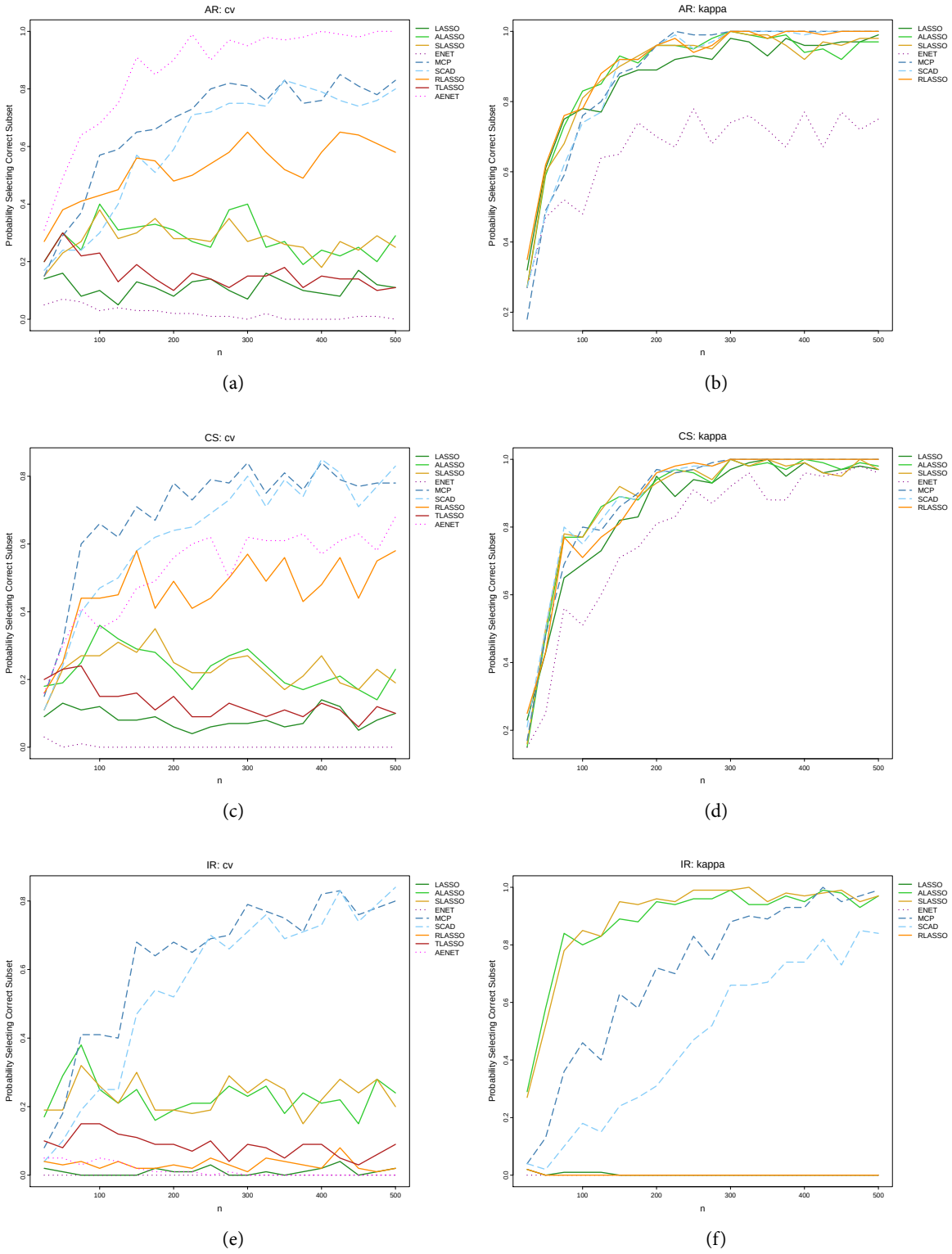


Figure (6.3.4) Probability of selecting the correct model

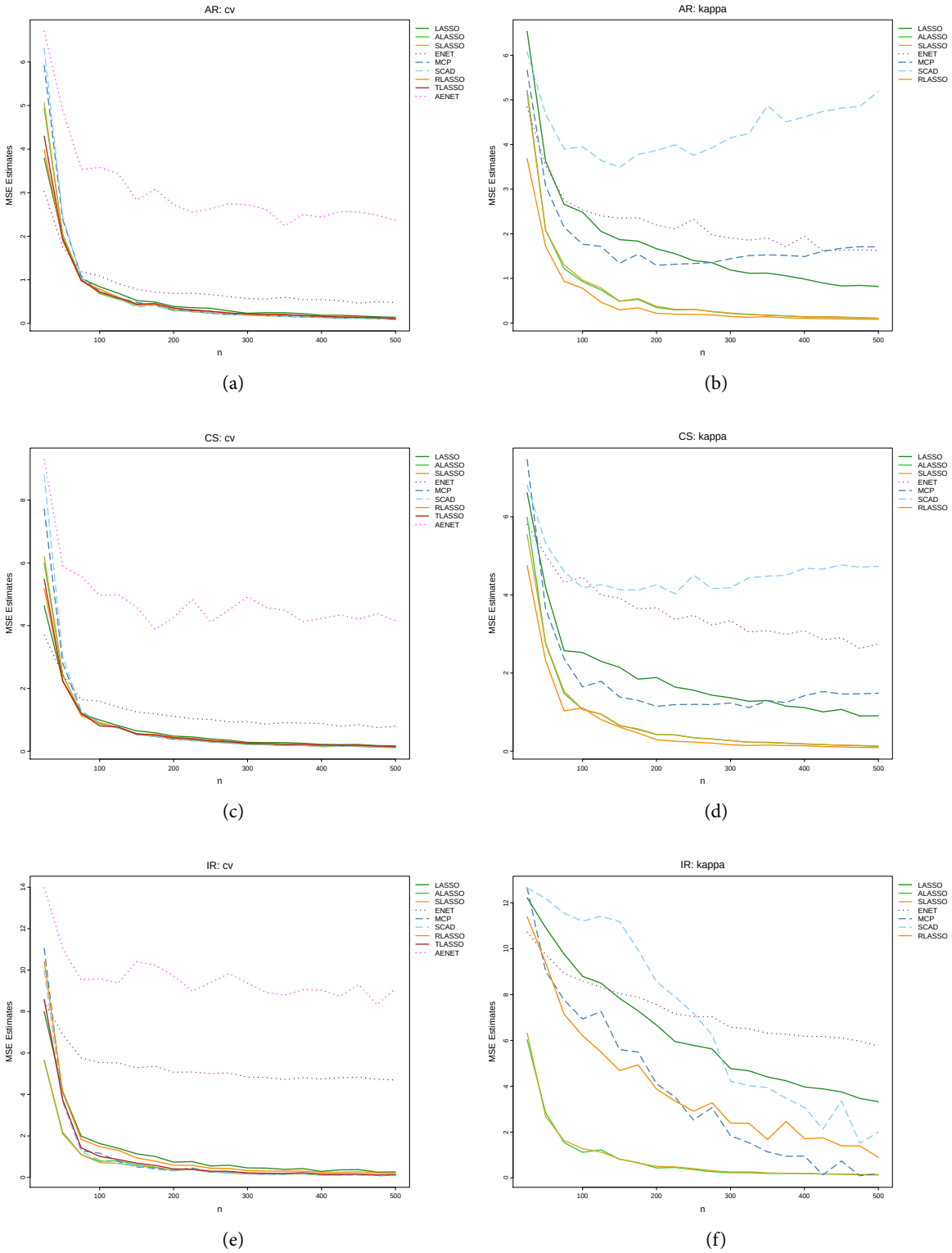


Figure (6.3.5) *MSE of parameter estimates*

Persistence, or consistent prediction, can be shown in a similar way by plotting the MSE of the predictions against the sample size. The results are very similar to those obtained for estimation and the figures are therefore omitted. For CV, the curves are almost identical to estimation, all methods appear to be persistent except the EN and adaptive EN. Similar results are also seen when using the kappa coefficient, except that MCP appears to be persistent for any correlation design.

The kappa coefficient therefore appears to be an excellent model selection criteria for the adaptive LASSO, SEA-LASSO and relaxed LASSO, yielding consistent prediction, estimation and variable selection (except for relaxed LASSO and IR structures). The relaxed LASSO also appears to enjoy consistent prediction, estimation and selection (except for IR designs), when using CV for model selection. MCP and SCAD appear to perform better in all aspects when using CV for model selection. The EN methods do not achieve consistence in all aspects using either CV or the kappa coefficient and perhaps it would be worth investigating the use of a different secondary tuning parameter. The LASSO achieves consistent estimation and prediction when using CV and achieves consistent selection when using the kappa coefficient. Neither of these two model selection criteria are capable of yielding all-round consistency for the LASSO.

Small Sample Results

The results shown below are for the AR correlation structure when $n = 25$ and the data is identical to that in Section 6.2. Table 6.3.3 shows the total variance and squared bias of the parameter estimates when using 5-fold CV and the kappa coefficient for model selection. When using 10-fold CV, the variance of the LASSO estimate is slightly higher and the bias slightly lower than when using 5-fold CV. It is clear that the two-stage LASSO methods and the concave penalties control the bias better than the LASSO, they all have lower bias than the LASSO regardless of whether CV or the kappa coefficient is used. Although they also have higher variance than the LASSO, they all display lower variability than the least squares model. The EN is the only method which has lower variance than the LASSO but its bias is also larger than the LASSO. The adaptive EN performs very poorly, with large bias and large variance.

Figure 6.3.6 shows the distribution of the MSE of predictions and the median MSE along with bootstrap standard errors are shown in Table 6.3.4. The probability of selecting the correct model is also shown in both the figure and the table. The LASSO appears to be quite competitive in terms of prediction accuracy when using CV, the median MSE is among the lowest and it has the lowest variance. The EN and relaxed

	10-fold CV		κ Coefficient	
Method	Variance	Squared Bias	Variance	Squared Bias
LASSO	3.243	0.5504	1.183	5.3504
RLASSO	3.683	0.3063	3.073	0.6085
ALASSO	4.592	0.3492	3.187	2.0228
SLASSO	4.710	0.3427	2.808	2.3222
TLASSO	3.980	0.3174		
ENET	2.212	0.8164	1.368	3.4698
AENET	5.294	1.4086		
MCP	5.673	0.2490	4.710	0.9512
SCAD	6.112	0.2021	3.232	2.8406
	Full		Oracle	
OLS	7.548	0.06334	1.469	0.01287

Table (6.3.3) Variance and squared bias of parameter estimates

LASSO also perform very well in terms of prediction. While the adaptive EN has the highest probability of selecting the correct model, there is no gain in fitting the smaller model since its median MSE and its variance is larger than that of the least squares model. MCP and SCAD also perform no better than least squares. Using the kappa coefficient does increase the probability of selecting the correct model for each method, but in most cases, the MSE and its variance are increased so that least squares is a more attractive option. However, it does work remarkably well with the relaxed LASSO, not only improving its variable selection performance but also the MSE, with only a slight increase in variance. The relaxed LASSO outperforms forward selection in terms of prediction and selection.

	10-fold CV		κ Coefficient	
Method	Median MSE	PCS	Median MSE	PCS
LASSO	2.70 (0.19)	0.14	7.94 (0.39)	0.32
RLASSO	2.63 (0.39)	0.27	2.52 (0.55)	0.35
ALASSO	2.99 (0.28)	0.20	4.43 (0.42)	0.27
SLASSO	2.96 (0.46)	0.15	4.15 (0.29)	0.27
TLASSO	2.84 (0.33)	0.20		
ENET	2.12 (0.22)	0.05	4.63 (0.42)	0.27
AENET	5.69 (0.42)	0.31		
MCP	3.77 (0.45)	0.15	4.00 (0.63)	0.18
SCAD	4.21 (0.47)	0.17	6.63 (0.52)	0.27
	Full		Oracle	
OLS	4.266 (0.421)	0	0.965 (0.081)	1

Table (6.3.4) Median MSE and probability of selecting the correct model

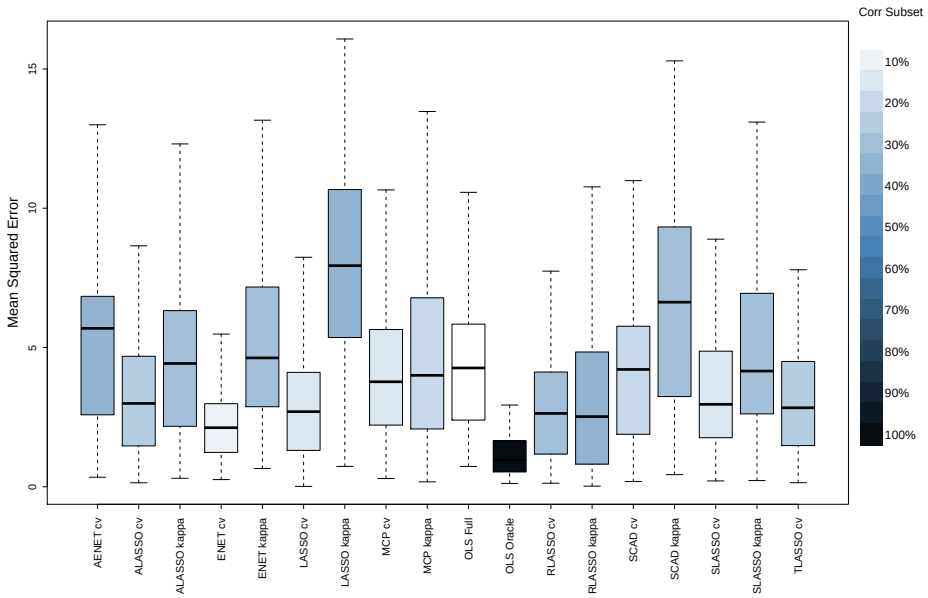


Figure (6.3.6) *MSE of predictions and probability of selecting the correct model*

Figures 6.3.7 and 6.3.8 take a closer look at the variable selection performance of each method. The LASSO and EN both include the correct model in the selected models with high probability. However, they both tend to overfit the model (the EN more so than the LASSO), resulting in a large number of false inclusions and a low probability of selecting the correct model. Except for the adaptive EN each method overfits the model when using CV and underfits the model when using the kappa coefficient. MCP performs especially poor with this data, the true model lying in its solution path only about 40% of the time. The relaxed LASSO performs the best in terms of selection.

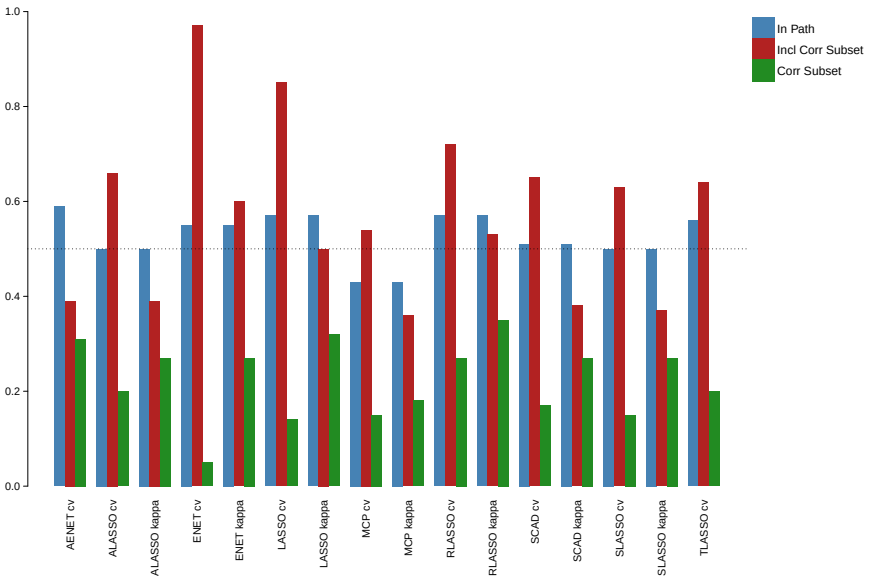


Figure (6.3.7) Variable selection performance: probability of selecting the correct model, including the correct model and having the correct model in the solution path

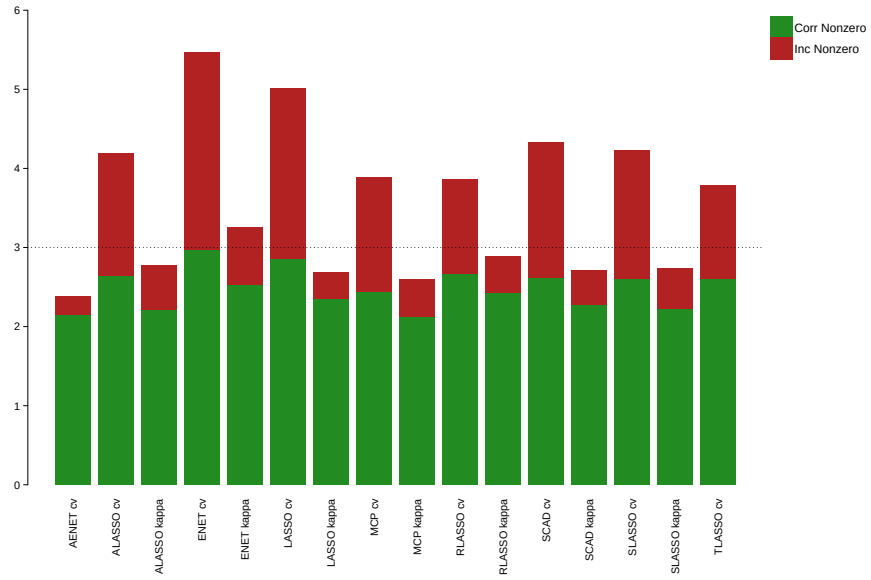


Figure (6.3.8) Number of nonzero parameters included

Chapter 7

Application

7.1 Data

An analysis is performed on the diabetes data set from Efron *et al.* (2004) available in the R package `lars` or at `http://web.stanford.edu/~hastie/Papers/LARS`. A study was conducted on 442 diabetes patients. Baseline measurements were taken at the beginning of the study and a quantitative measure of disease progression was recorded one year after the baseline. Ten baseline variables were recorded, including age, gender, body mass index, average blood pressure, as well as six blood serum measurements. These variables are used as predictor variables in a regression model with the measure of disease progression as the response variable. The model can be useful in determining which factors promote the progression of the disease as well as predicting disease progression for future patients using their baseline measurements. The data is presented visually in Figure 7.1.1. The blood serum measurements appear to have moderate to high correlations with each other. In particular, the variables tc, ldl, tch and ltg form a group of variables with high pairwise correlations, with substantial correlation between ldl and tc. Of the remaining blood serum measurements, hdl is very highly correlated with tch and moderately with ltg, and glu has moderate correlations with all blood serum measurements. The average blood pressure, more formally the mean arterial pressure (MAP) has moderate correlations with ltg and glu, and bmi with map, hdl, tch, ltg and glu. Variables hdl and ltg are significantly different between males and females, while map and glu are moderately related to age. Collinearity is definitely a problem with this data set, with a condition number of the correlation matrix being 21.68. The response has the largest correlation with bmi, followed closely by ltg, and has moderate correlations with map, hdl, tch and glu. All variables appear to be normally distributed, apart from bmi and tch which appear to be right-tailed, or skewed to the left.

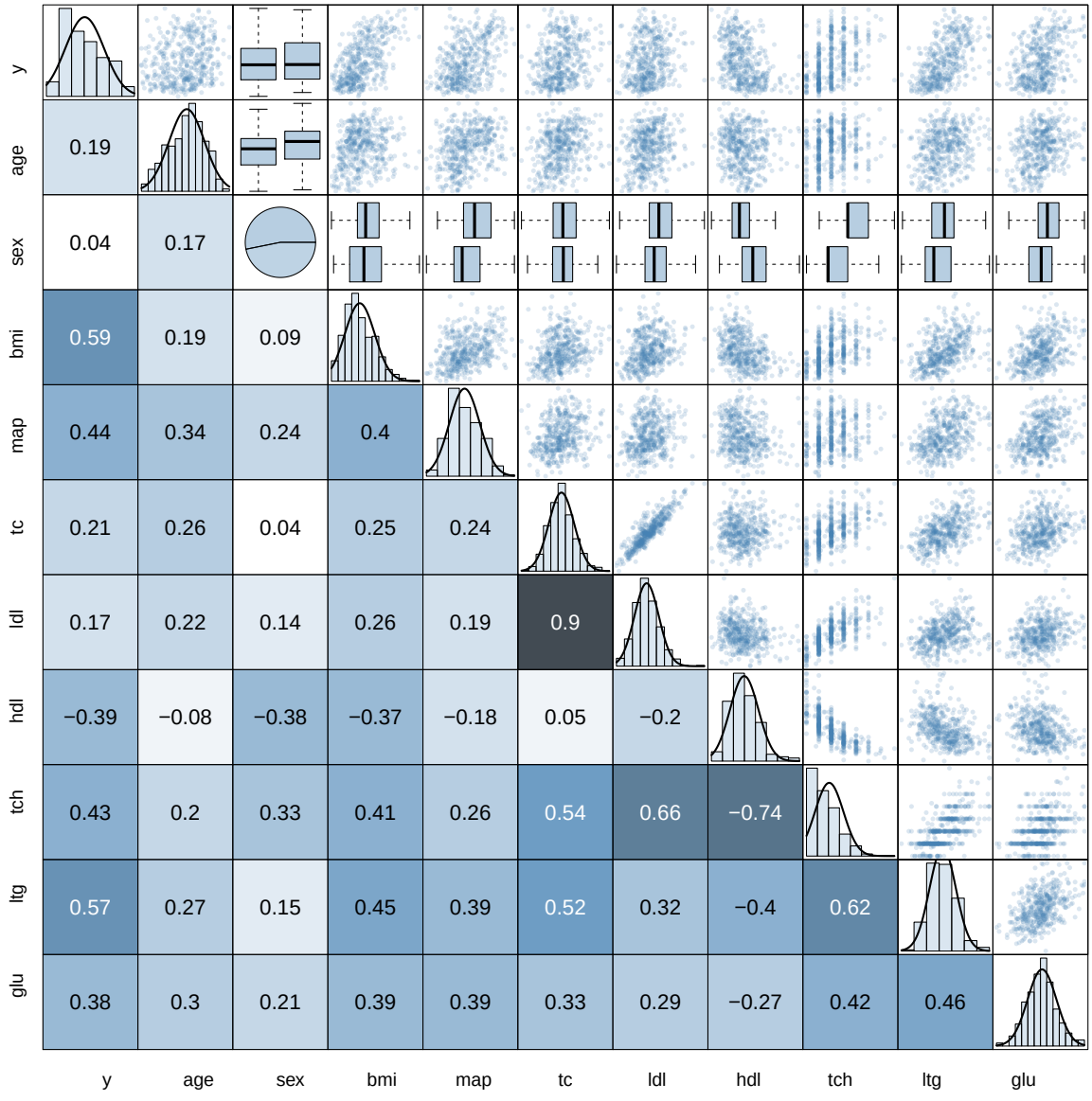


Figure (7.1.1) Visual presentation of diabetes data set

7.2 Estimation, Model Selection and Prediction

The data is split roughly 75-25%, into a training set of size 332 and a test set of size 110 using simple random sampling without replacement. The model is estimated on the training set using all the methods examined in Chapter 6, including least squares, forward selection, ridge regression, the LASSO, relaxed LASSO, adaptive LASSO, SEA-LASSO, two-stage LASSO, EN, adaptive EN, MCP and SCAD.

Model selection is performed by resampling data from the training set using 10-fold cross-validation, simultaneously optimizing over two tuning parameters where relevant. The primary tuning parameter (the shrinkage parameter) is selected in a similar fashion as in Chapter 6 for each method, except for the LASSO, where λ is now selected as one of the values at which new predictors enter the LAR-LASSO algorithm. The values considered for the secondary tuning parameter, where necessary, are shown in Table 7.2.1. The λ_2 parameter for the adaptive EN is set to the value that is chosen by CV for the EN. As before, the primary tuning parameters for the adaptive EN and the two-stage LASSO are chosen sequentially using CV. For all other methods, the model is also selected using the kappa coefficient, where the secondary tuning parameter is fixed at its cross-validated value for easy comparison. Finally, the selected models are used to predict the observations in the test set and the generalizability of each model is assessed via the test error.

	Tuning Parameter	
Method	Primary	Secondary
Forward selection	$p \in [1, 10]$	
Ridge regression	$\lambda \in [0, 50]$	
LASSO	$\lambda \in [2, 835]$	
Relaxed LASSO	LAR step $\in [1, 10]$	$\phi \in \{0, 0.1, \ldots, 1\}$
Adaptive LASSO	$s \in [0, 1]$	$\zeta \in \{0.5, 1, 2\}$
SEA-LASSO	$s \in [0, 1]$	$\zeta \in \{0.5, 1, 2\}$
Two-stage LASSO	$s \in [0, 1]$	
EN	$s \in [0, 1]$	$\lambda_2 \in \{0.1, 0.5, 1, 5, 10\}$
Adaptive EN	$s \in [0, 1]$	$\lambda_2 = 0.1$
SCAD	$\lambda \in [0.05, 46]$	$\xi \in \{2.5, 2.6, \ldots, 3.5\}$
MCP	$\lambda \in [0.05, 46]$	$\xi \in \{3, 3.1, \ldots, 4\}$

Table (7.2.1) Tuning parameters considered for the diabetes data

7.3 Results

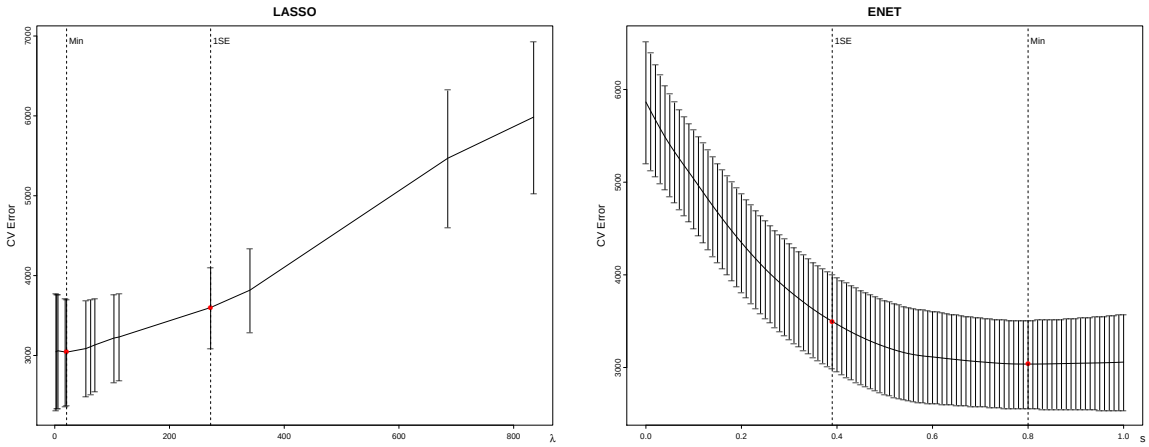
Table 7.3.1 shows the values of the tuning parameters selected by CV and the kappa coefficient for each method. The tuning parameters chosen by CV in the initial round of the two-stage LASSO and the adaptive EN are shown too. The initial models selected are similar to those selected by CV for the LASSO and EN, respectively. Also included in the table is the ℓ_1 fraction, the ratio of the ℓ_1 norm of the parameter vector to that of the least squares model. It is usually denoted by s but is not done so here in order to avoid confusion with the tuning parameter s. Examining the ratios, we see that the LASSO and EN models chosen by CV all shrink the size of the parameter vector by about 50% and they are shrunk even more when using the kappa coefficient.

	Secondary		Primary				ℓ_1 Fraction	
Method	Type	CV	Type	CV	Initial	κ	CV	κ
FORWARD			p	6.00		5.00	0.82	0.54
RIDGE			λ	20.80		0.00	0.56	1.00
LASSO			λ	20.61		112.03	0.51	0.33
RLASSO	ϕ	0.30	LAR step	9.00		5.00	0.52	0.37
ALASSO	ζ	0.50	s	0.52		0.39	0.53	0.43
SLASSO	ζ	0.50	s	0.32		0.12	0.50	0.29
TLASSO			s	0.76	0.50		0.54	
ENET	λ_2	0.1	s	0.8		0.52	0.48	0.31
AENET	λ_2	0.1	s	0.13	0.84		0.55	
MCP	ξ	2.90	λ	2.62		4.91	0.64	0.57
SCAD	ξ	3.70	λ	1.98		4.91	0.64	0.50

Table (7.3.1) Tuning parameters selected for the diabetes data

The CV error curves are shown in Figure 7.3.1, including standard error bars, for the LASSO and the EN. The position of the minimum CV error and the position of the CV error within 1 SE of the minimum are indicated by the dotted black lines on the plot. For each method, the model selected by the kappa coefficient lies somewhere in between the model selected using the minimum and the model selected using the 1 SE rule. Also, for each method, the variance of the CV error is quite large and it could be beneficial to use 5-fold CV instead.

The order in which variables are included in the forward selection and LAR algorithms is shown in Table 7.3.2. In most cases, bmi is the first variable entered since it has the largest correlation with the response. The adaptive LASSO scales the data by the squared size of the least squares coefficients. The least



(a) (b)

Figure (7.3.1) CV curves for the diabetes data

squares coefficient of ltg is much larger than that of bmi and causes ltg to have a higher correlation with the response. The SEA-LASSO also takes the standard errors of the least squares coefficients into account so that bmi, which has a smaller standard error than ltg, remains the most correlated with the response. The two-stage LASSO and adaptive EN begin with the subsets selected for the LASSO and EN, respectively, so that not all of the variables are considered in their paths.

Step	FORWARD	LASSO	ALASSO	SLASSO	TLASSO	ENET	AENET
1	bmi	bmi	ltg	bmi	bmi	bmi	bmi
2	ltg	ltg	bmi	ltg	ltg	ltg	ltg
3	tc	map	tc	map	map	map	map
4	map	hdl	ldl	sex	hdl	tch	hdl
5	sex	sex	map	tc	tc	hdl	sex
6	ldl	ldl	sex	ldl	sex	glu	ldl
7	tch	tc	tch	tch	glu	age	tc
8	hdl	-ldl	hdl	glu		tc	glu
9	glu	glu	glu	hdl		ldl	tch
10	age	age	age	age		sex	
11		tch					
12		ldl					

Table (7.3.2) Order in which variables are included in the forward selection and LAR algorithms

The coefficient profiles for some of the methods are shown in Figure 7.3.2. The discrete nature of forward selection can be seen by the dramatic changes in the active coefficient values whenever a new predictor enters the model. For ridge regression, we see the proportional shrinkage, where larger coef-

ficients are shrunk more than smaller ones. The least squares coefficients for tc and ltg are the largest and it is these parameters that are shrunk the most. As seen before, the kappa coefficient does not work well for ridge regression and selects the least squares model. The LASSO displays a more constant kind of shrinkage across variables, allowing smaller parameters to be set to zero quickly. CV selects a value of $\lambda = 20.8$ for the shrinkage parameter which is equivalent to an ℓ_1 fraction of 0.51 and the model contains 7 parameters. We can expect this model to be overfitted but to include the true set of relevant predictors with high probability. The kappa coefficient selects a larger value of the shrinkage parameter, $\lambda = 112$, in order to set more coefficients equal to zero. The model only contains 4 parameters and has an ℓ_1 fraction of 0.33. The kappa coefficient selects the same model for the EN as the LASSO. The models are similar when using CV, except that the EN also included ldl - probably because it has high pairwise correlations with tc and ltg. Comparing the path of the LASSO and adaptive LASSO, it is clear that the adaptive LASSO applies a different amount of shrinkage to each parameter. The path for the SEA-LASSO looks similar to the adaptive LASSO and is not shown. The path for MCP is also shown, the large parameters of ltg and bmi are not shrunk as much as in the LASSO and the small parameters of age and glu are not set to zero as quickly. The SCAD path is similar to MCP and CV selects exactly the same model for both.

The standardized coefficients are shown in Table 7.3.3 for least squares and the LASSO models selected by CV and kappa. The standard errors and p-values are shown for least squares. The variables sex, bmi, map, tc and ltg are significant at the 5% level. The standard errors are shown for the LASSOmodels, calculated using the approximations given by Tibshirani (1996) and Osborne *et al.* (2000b). The standard errors using the Osborne approximation are quite similar to the least squares standard errors. The Tibshirani standard errors are substantially lower than least squares for variables tc, ldl, hdl and ltg. Which are correct is debatable and it would have been beneficial to include bootstrap standard errors for comparison. However, from the simulation studies it was seen that the LASSO estimates had lower variance than the least squares estimates.

The standardized coefficients for all methods are shown in Table 7.3.4 for model selection by 10-fold CV and in Table 7.3.5 for model selection using the kappa coefficient. Ridge regression is excluded from the discussion of variable selection since it retains all of the variables. The variables that have a significant effect in the least squares mode (sex, bmi, map, tc, ltg) are indicated with bold headers. When using CV, all models include these variables except the adaptive EN, where tc is not included. In addition to these variables, hdl is included for all methods except forward selection. LASSO, SEA-LASSO and EN include glu;

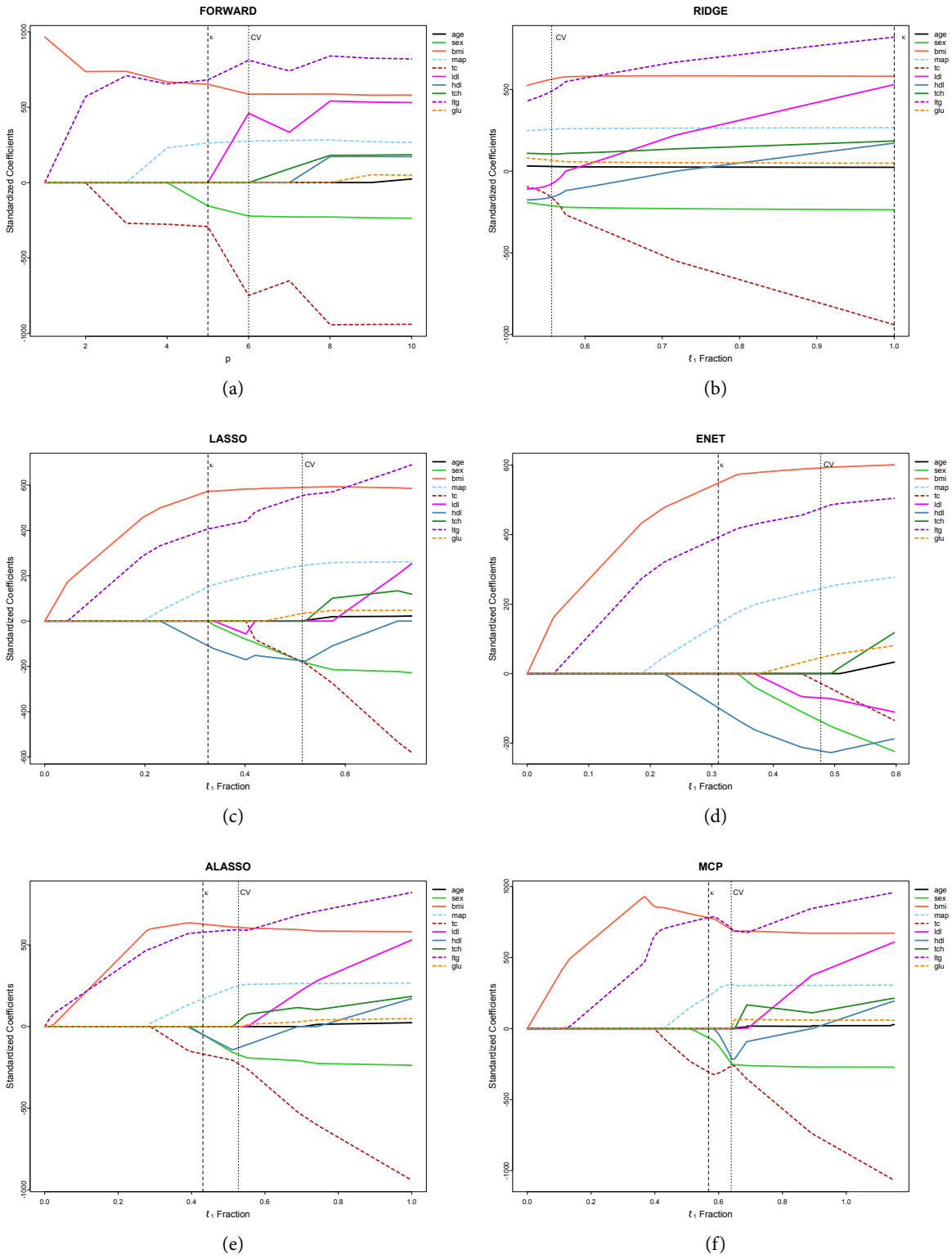


Figure (7.3.2) Coefficient profiles for the diabetes data

	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu
					Least Squares					
Est	24	−237	580	266	−941	531	172	185	821	49
SE	60	62	67	66	410	329	210	158	170	66
p-value	0.732	0.001	0.000	0.000	0.046	0.166	0.474	0.313	0.000	0.539
					LASSO using 10-fold CV					
Est	0	−179	590	245	−181	0	−175	0	553	34
SE Tibs	59	53	62	57	52	102	67	123	73	34
SE Osb	60	60	65	64	400	327	206	158	168	66
					LASSO using the κ coefficient					
Est	0	0	573	151	0	0	−109	0	407	0
SE Tib	59	60	50	31	145	158	14	94	49	65
SE Osb	60	53	66	62	406	335	212	160	167	66

Table (7.3.3) Standardized parameter estimates (Est) and standard error estimates (SE) for least squares and the LASSO. For the LASSO, the standard errors are calculated using both the Tibshirani approximation (SE Tib) and the Osborne approximation (SE Osb).

forward selection, EN and adaptive EN include ldl; and only adaptive LASSO includes tch. When using kappa, four models include the least squares significant predictors, forward selection, adaptive LASSO, SCAD and MCP, although adaptive LASSO and SCAD also select hdl. The other four methods, LASSO, relaxed LASSO, SEA-LASSO and EN all select the model including only bmi, map, hdl and ltg. The exception here is SEA-LASSO, which does not select hdl.

	age	**sex**	**bmi**	**map**	**tc**	ldl	hdl	tch	**ltg**	glu	Test Error
FORWARD	0	−222	587	276	−750	460	0	0	812	0	2968
RIDGE	29	−212	563	256	−162	−77	−159	105	490	66	2916
LASSO	0	−179	590	245	−181	0	−175	0	553	34	2934
RLASSO	0	−183	596	257	−188	0	−180	0	569	0	2951
ALASSO	0	−172	608	248	−227	0	−130	31	592	0	2980
SLASSO	0	−168	620	253	−160	0	−130	0	555	7	2950
TLASSO	0	−202	604	268	−216	0	−181	0	595	0	2970
ENET	0	−137	592	245	−28	−70	−222	0	476	46	2880
AENET	0	−196	668	278	0	−131	−266	0	540	0	2947
MCP	0	−247	694	315	−258	0	−217	0	702	0	3077
SCAD	0	−247	694	315	−258	0	−217	0	702	0	3077

Table (7.3.4) Standardized coefficients selected when using CV

Figure 7.3.3 shows the test error when using the selected models to make predictions on the observations in the test data set. The least squares test error is 2954 and is indicated by the dashed horizontal line

	age	**sex**	**bmi**	**map**	**tc**	ldl	hdl	tch	**ltg**	glu	Test Error
FORWARD	0	−155	652	263	−292	0	0	0	682	0	3126
RIDGE	24	−237	580	266	−941	531	172	185	821	49	2951
LASSO	0	0	573	151	0	0	−109	0	407	0	3043
RLASSO	0	0	609	203	0	0	−162	0	443	0	2958
ALASSO	0	−48	627	169	−168	0	−49	0	578	0	3067
SLASSO	0	0	679	133	0	0	0	0	289	0	3307
ENET	0	0	548	143	0	0	−98	0	392	0	3077
MCP	0	−65	782	228	−306	0	0	0	782	0	3289
SCAD	0	−16	816	115	−192	0	−4	0	741	0	3309

Table (7.3.5) Standardized coefficients selected when using kappa

in the figure. The test error for each method is also shown in Tables 7.3.4 and 7.3.5. When using CV the following methods have lower test error than least squares: ridge, LASSO, relaxed LASSO, SEA-LASSO, adaptive EN. Only ridge regression has lower test error than least squares when using the kappa coefficient.

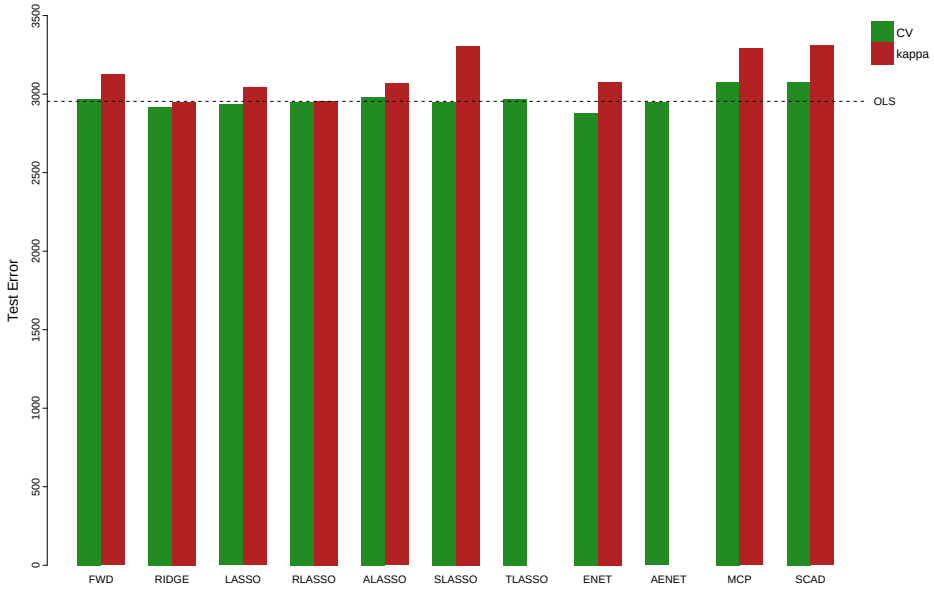


Figure (7.3.3) Test error when using CV and kappa

I would recommend using the LASSO model chosen by CV. Other than ridge regression and the EN, it has the lowest test error and should yield a more accurate prediction of disease progression than least squares. Although ridge regression and EN have lower test error, the LASSO produces a sparser model and helps to narrow down the risk factors associated with the progression of diabetes. Furthermore, the LASSO model should contain the correct model with high probability so we can be confident that an important risk factor has not been falsely excluded.

Chapter 8

Conclusion

The LASSO and related methods provide an elegant class of methods which simultaneously perform variable selection and estimation with superb performance when the underlying model is sparse. Each LASSO model is delivered with an interesting geometrical interpretation and its entire pathway can be produced which aids in the interpretability of a data set. With state of the art algorithms for efficient computation and model selection procedures, the LASSO can be applied to high dimensional data with ease. The LASSO is shown to have excellent prediction accuracy, consistent estimation and is suitable for variable selection under certain conditions. Since the predictions are not sensitive to collinearity, weaker conditions are necessary for persistence than for consistent variable selection. Where these conditions are not met, one of the two-stage LASSO methods or concave penalties can be used. Modified LASSO methods or combined penalties allow for more flexibility by incorporating different structures between predictor variables.

The hexagonal operator for regression with shrinkage and equality selection (HORSES) is a modified LASSO method by Jang *et al.* (2013) which is not mentioned above. Similar to the fused LASSO, it is a combination of two ℓ_1 penalties. However, the second penalty is applied to all the pairwise differences of the coefficients instead of only the adjacent ones. The result is a hexagonal shaped penalty function which has a natural grouping effect such as the effect experienced with the combined penalties. Two other combined penalties also worth noting are Mnet proposed by Huang *et al.* (2010) and the sparse Laplacian shrinkage (SLS) estimator proposed by Huang *et al.* (2011). The former is a combination of the ridge and MCP penalties, while the latter combines the MCP penalty with a Laplacian quadratic penalty.

There are a number of group penalties, not mentioned above, which can be employed. The concave penalties have been adapted to perform bi-level selection. Wang *et al.* (2007) proposed the group SCAD, Breheny & Huang (2009) proposed the group MCP and Breheny & Huang (2014) developed descent algorithms for their solutions. Other bi-level selection procedures include the sparse-group LASSO (Simon *et al.* (2013)) and the group exponential lasso (Breheny (2014)). For comparisons of group penal-

ties, see Yang (2011) and Huang *et al.* (2012). Also worth mentioning, is the hierarchical group-lasso proposed by Lim & Hastie (2014) which handles interactions in the group LASSO.

The LASSO has also been modified for other purposes than to incorporate different structures between the predictors. He (2011) proposed a model which can incorporate prior information into the LASSO by using a set of linear constraints. The model proposed by Li (2012) utilizes a mean-shift to allow for simultaneous outlier detection and variable selection with the LASSO.

The LASSO can also be applied outside the scope of the general linear model. Turlach *et al.* (2005) provide an extension of the LASSO to multiple response regression. The LASSO has also been extended to generalized linear models (GLMs), where the errors follow a distribution from the exponential family and the linear model is related to the response via a link function. GLMs are discussed by Tibshirani (1996), who provides an example using logistic regression (binomial distribution). Zhao (2008) also discusses using the LASSO for logistic regression. Park & Hastie (2007) and Friedman *et al.* (2010) propose path algorithms for the solution path of ℓ_1 regularized GLMs. An extension to survival models is covered by Tibshirani (1997), who uses the LASSO for the Cox proportional hazards model. Wang *et al.* (2007) extends LASSO to least absolute deviations (LAD) estimation, where the parameters are estimated using the ℓ_1 loss function. Gaussian graphical models using the LASSO are discussed by Meinshausen & Bühlmann (2006), Witten *et al.* (2011) and Mazumder & Hastie (2012).

There are also some extensions of the LASSO for nonparametric methods, including regression splines (Osborne *et al.* (1998), Rosset & Zhu (2007)), the support vector machine (SVM) and kernel smoothers (Roth (2004)), and wavelet analysis (Donoho & Johnstone (1994), Donoho (1995), Antoniadis (1997), Donoho & Johnstone (1998), Sardy *et al.* (1999)). Furthermore, Sun (1999) discusses using the LASSO for neural networks.

The LASSO solution has an alternative interpretation as the Bayesian posterior mode with double-exponential (Laplace) priors on the regression parameters (Tibshirani (1996)). Park & Casella (2008) discuss the Bayesian LASSO and derive Bayesian interval estimates. Armagan & Zaretzki (2010), Kyung *et al.* (2010) and Lykou & Ntzoufras (2012) also approach the problem from a Bayesian perspective.

Further developments may still be necessary before these methods have mainstream appeal. Few advances have been made concerning statistical inferences for the models produced. Better standard

errors of LASSO estimates and derivation of confidence intervals remain a topic for further research. The significance test described by Lockhart *et al.* (2014) is a step in the right direction but can only be used as a stopping rule for the LAR-LASSO algorithm. Testing the overall significance of a predictor, the goodness of fit of the model and methods for multiple testing still need to be uncovered.

Appendix A

Definitions and Theorems

A.1 Vectors and Matrices

Definition A.1.1 ℓ_q**-norm**

The ℓ_q-norm of a $p \times 1$ vector $\underline{a}$ is given by

$$\ell_q(\underline{a}) = \|\underline{a}\|_q = \left(\sum_{j=1}^{p} |a_j|^q \right)^{\frac{1}{q}},$$

where $q \geqslant 1$, and has the following properties:

1. $\|\underline{a}\|_q \geqslant 0$ *for all* $\underline{a} \in \mathbb{R}^p$ *(nonnegative)*
2. $\|\underline{a}\|_q = 0 \Leftrightarrow \underline{a} = \underline{0}$ *(definite)*
3. $\|s\underline{a}\|_q = |s| \, \|\underline{a}\|_q$ *for all* $\underline{a} \in \mathbb{R}^p, s \in \mathbb{R}$ *(homogenous)*
4. $\|\underline{a} + \underline{b}\|_q \leqslant \|\underline{a}\|_q + \|\underline{b}\|_q$ *for all* $\underline{a}, \underline{b} \in \mathbb{R}^p$ *(subadditive)*

Gentle (2007:16-18) or Boyd & Vandenberghe (2004:633-637) can be consulted for more information about norms. Some notes:

- The ℓ_2 norm corresponds to the usual Euclidean norm and the subscript is normally omitted.
- The max norm, also called the Chebyshev norm, is given by

$$\ell_\infty(\underline{a}) = \lim_{q \to \infty} \|\underline{a}\|_q = \max\left\{ |a_1|, |a_2|, \ldots, |a_p| \right\}.$$

- $\|\underline{a}\|_q$ is a measure of length or size, $\|\underline{a} - \underline{b}\|_q$ is a measure of distance.
- When $q \in [0, 1)$

- $\|\underline{a}\|_q$ is homogeneous but it is not subadditive.
- $\|\underline{a}\|_q^q$ is subadditive but it is not homogeneous.

Definition A.1.2 ***Generalized inverse***

A generalized inverse of an $n \times p$ matrix $\mathbf{A}$ is defined as the $p \times n$ matrix $\mathbf{A}^-$ such that

$$\mathbf{A} = \mathbf{A}\mathbf{A}^-\mathbf{A}.$$

$\mathbf{A}^-$ is not unique, unless $\mathbf{A}$ is square and has full rank then $\mathbf{A}^- = \mathbf{A}^{-1}$.

See Searle (1971:1-7) for a discussion and for methods of computing the generalized inverse.

Theorem A.1.1 ***Generalized inverse of Gramian matrix***

The generalized inverse of a Gramian matrix, $\left(\mathbf{A}^T\mathbf{A}\right)^-$, has all of the following properties:

1. $\left(\left(\mathbf{A}^T\mathbf{A}\right)^-\right)^T$ *is also a generalized inverse of* $\mathbf{A}^T\mathbf{A}$
2. $\left(\mathbf{A}^T\mathbf{A}\right)^-\mathbf{A}^T$ *is a generalized inverse of* $\mathbf{A}$
3. $\mathbf{A}\left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^-\mathbf{A}^T$ *is invariant to the choice of* $\left(\mathbf{A}^T\mathbf{A}\right)^-$
4. $\mathbf{A}\left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^-\mathbf{A}^T$ *is always symmetric regardless of the choice of* $\left(\mathbf{A}^T\mathbf{A}\right)^-$

See Searle (1971:20) for a proof of the theorem.

Definition A.1.3 ***Moore-Penrose inverse***

The Moore-Penrose inverse is a generalized inverse of $\mathbf{A}$ that satisfies all of the following conditions:

1. $\mathbf{A} = \mathbf{A}\mathbf{A}^-\mathbf{A}$
2. $\mathbf{A}^- = \mathbf{A}^-\mathbf{A}\mathbf{A}^-$
3. $\left(\mathbf{A}\mathbf{A}^-\right)^T = \mathbf{A}\mathbf{A}^-$
4. $\left(\mathbf{A}^-\mathbf{A}\right)^T = \mathbf{A}^-\mathbf{A}$

The Moore-Penrose inverse exists for any matrix. It is unique and will be denoted by $\mathbf{A}^+$

See Searle (1971:16-18) or Gentle (2007:102-103) for a proof of the existence and uniqueness of the Moore-Penrose inverse.

Definition A.1.4 ***Four fundamental subspaces of a matrix***

The four fundamental subspaces of an $n \times p$ *matrix* $\mathbf{A}$ *with* $\text{rank}(\mathbf{A}) = r$ *are:*

1. *The column space, also known as the image or range, is the subspace spanned by the columns of* $\mathbf{A}$ *and is given by*

$$\mathcal{C}(\mathbf{A}) = \left\{\mathbf{a} \in \mathbb{R}^n : \mathbf{A}\underline{b} = \mathbf{a} \text{ for all } \underline{b} \in \mathbb{R}^p\right\}$$

   The set of linearly independent columns of $\mathbf{A}$ *is a basis for the column space of* $\mathbf{A}$ *and* $\dim(\mathcal{C}(\mathbf{A})) = r$.

2. *The row space is the subspace spanned by the rows of* $\mathbf{A}$ *and is the column space of* $\mathbf{A}^T$,

$$\mathcal{C}\left(\mathbf{A}^T\right) = \left\{\underline{b} \in \mathbb{R}^p : \mathbf{A}^T\mathbf{a} = \underline{b} \text{ for all } \mathbf{a} \in \mathbb{R}^n\right\}$$

   The set of linearly independent rows of $\mathbf{A}$ *is a basis for the row space of* $\mathbf{A}$ *and* $\dim\left(\mathcal{C}\left(\mathbf{A}^T\right)\right) = r$.

3. *The null space, also known as the kernel, is given by*

$$\mathcal{N}(\mathbf{A}) = \left\{\underline{b} \in \mathbb{R}^p : \mathbf{A}\underline{b} = \mathbf{0}\right\}$$

   The dimension of $\mathcal{N}(\mathbf{A})$ *is called the nullity of* $\mathbf{A}$ *and is given by* $\dim(\mathcal{N}(\mathbf{A})) = p - r$.

4. *The left null space, also known as the cokernel, is the null space of* $\mathbf{A}^T$

$$\mathcal{N}\left(\mathbf{A}^T\right) = \left\{\mathbf{a} \in \mathbb{R}^n : \mathbf{A}^T\mathbf{a} = \underline{0}\right\}$$

   and $\dim\left(\mathcal{N}\left(\mathbf{A}^T\right)\right) = n - r$.

These definitions can be found in any text on linear algebra, see Messer (1994:245-254) or Strang (2006:115-121). The left null space of $\mathbf{A}$ is the orthogonal complement of the column space of $\mathbf{A}$ which is denoted by $\mathcal{C}^{\perp}(\mathbf{A}) = \mathcal{N}\left(\mathbf{A}^T\right)$.

Theorem A.1.2 ***Expectation of quadratic forms***

Let v be an $n \times 1$ vector and let $\mathbf{A}$ be an $n \times n$ symmetric matrix. If $E(\mathbf{v}) = \mu$ and $\text{var}(\mathbf{v}) = \Sigma$ then

$$E\left(\mathbf{v}^T\mathbf{A}\mathbf{v}\right) = \text{tr}\left(\mathbf{A}\Sigma\right) + \mu^T\mathbf{A}\mu.$$

See Searle (1971:55) or Seber & Lee (2003:9) for a proof.

Theorem A.1.3 ***Inverse of a partitioned matrix***

A nonsingular matrix $\mathbf{A}$ partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

where $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are nonsingular, has inverse

$$\begin{aligned} \mathbf{A}^{-1} &= \begin{bmatrix} \mathbf{A}_{11}^{-1} + \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21} & -\mathbf{B}_{12}\mathbf{B}_{22}^{-1} \\ -\mathbf{B}_{22}^{-1}\mathbf{B}_{21} & \mathbf{B}_{22}^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{C}_{11}^{-1} & -\mathbf{C}_{11}^{-1}\mathbf{C}_{12} \\ -\mathbf{C}_{21}\mathbf{C}_{11}^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12} \end{bmatrix}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{B}_{22} &= \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}, \\ \mathbf{B}_{12} &= \mathbf{A}_{11}^{-1}\mathbf{A}_{12}, \\ \mathbf{B}_{21} &= \mathbf{A}_{21}\mathbf{A}_{11}^{-1} \end{aligned}$$

and

$$\begin{aligned} \mathbf{C}_{11} &= \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}, \\ \mathbf{C}_{12} &= \mathbf{A}_{12}\mathbf{A}_{22}^{-1}, \\ \mathbf{C}_{21} &= \mathbf{A}_{22}^{-1}\mathbf{A}_{21}. \end{aligned}$$

The result is given in Seber & Lee (2003:466) and can easily be proved by showing that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. Gentle (2007:101) provides a similar result for nonsingular matrices and Searle (1971:27) provides a result specifically for symmetric matrices.

Definition A.1.5 ***Singular value decomposition (SVD)***

If $\mathbf{A}$ *is an* $n \times p$ *matrix with* $\text{rank}(\mathbf{A}) = r$*, it has the SVD*

$$\mathbf{A} = \mathbf{UDV}^T,$$

where $\mathbf{U}$ *is an* $n \times n$ *orthogonal matrix,* $\mathbf{V}$ *is a* $p \times p$ *orthogonal matrix and* $\mathbf{D}$ *is an* $n \times p$ *diagonal matrix (with* $\min(n, p)$ *diagonal elements and zeroes elsewhere). The nonnegative diagonal elements of* $\mathbf{D}$ *are the singular values of* $\mathbf{A}$*, with* $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_r > 0$*. If any* $r < \min(n, p)$ *then* $d_{r+1} = d_{r+2} = \cdots = d_{\min(n,p)} = 0$ *and*

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{D}_r = \text{diag}(d_1, d_2 \ldots, d_r)$*. The columns of* $\mathbf{U}$ *span the column space of* $\mathbf{A}$ *and the columns of* $\mathbf{V}$ *span the row space of* $\mathbf{A}$*.*

See Gentle (2007:127-128). The above representation always holds but alternative representations are given as follows:

- If $n > p$ then $\mathbf{U}$ is an $n \times p$ matrix with orthogonal columns and $\mathbf{D}$ is a $p \times p$ diagonal matrix,
- If $n < p$ then $\mathbf{V}$ is a $p \times n$ matrix with orthogonal columns and $\mathbf{D}$ is a $p \times p$ diagonal matrix.

Definition A.1.6 ***Spectral decomposition***

If the $p \times p$ *matrix* $\mathbf{A}$ *is symmetric, it has the spectral decomposition*

$$\mathbf{A} = \mathbf{VEV}^T,$$

where $\mathbf{E}$ *is a diagonal matrix whose diagonal elements are the eigenvalues of* $\mathbf{A}$ *and the columns of* $\mathbf{V}$ *are the eigenvectors of* $\mathbf{A}$ *that are chosen to be orthonormal so that* $\mathbf{VV}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}$*..*

- Using the SVD, $\mathbf{A}^T\mathbf{A} = \mathbf{VDU}^T\mathbf{UDV}^T = \mathbf{VD}^2\mathbf{V}^T$. But $\mathbf{A}^T\mathbf{A}$ is symmetric and has the spectral decomposition above. Thus $\mathbf{E} = \mathbf{D}^2$, the eigenvalues of $\mathbf{A}^T\mathbf{A}$ are the squared singular values of $\mathbf{A}$, $e_i\left(\mathbf{A}^T\mathbf{A}\right) = d_i^2(\mathbf{A}) \Leftrightarrow d_i(\mathbf{A}) = \sqrt{e_i(\mathbf{A}^T\mathbf{A})}$.
- Similarly, if $\mathbf{A}$ is a symmetric matrix then $d_i(\mathbf{A}) = |e_i(\mathbf{A})|$.

Definition A.1.7 ***QR decomposition***

If $\mathbf{A}$ *is an* $n \times p$ *matrix, it has the QR decomposition*

$$\mathbf{A} = \mathbf{QR}.$$

1. *If* $n = p$ *then* $\mathbf{Q}$ *is an* $n \times n$ *orthogonal matrix and* $\mathbf{R}$ *is an* $n \times n$ *upper triangular matrix,*
2. *If* $n > p$ *then* $\mathbf{Q}$ *is an* $n \times n$ *orthogonal matrix and* $\mathbf{R}$ *is the* $n \times p$*matrix* $(\mathbf{R}_1, \mathbf{0})^T$ *where* $\mathbf{R}_1$ *is a* $p \times p$ *upper triangular matrix,*
3. *If* $n < p$ *then* $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2)$ *and* $\mathbf{R} = (\mathbf{R}_1, \mathbf{R}_2)$ *where* $\mathbf{Q}_1$ *is an* $n \times p$ *matrix with orthogonal columns and* $\mathbf{R}_1$ *is a* $p \times p$ *upper triangular matrix. In this case,* $\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1$.

See Gentle (2007:188-189).

A.2 Estimators

Definition A.2.1 ***Estimable functions***

Consider the linear model $\mathbf{y} = \mathbf{X}\underline{\beta}+\varepsilon$ *with* $E(\varepsilon) = \mathbf{0}$ *and* $\text{var}(\varepsilon) = \sigma^2\mathbf{I}$. *A linear function of* $\underline{\beta}$ *given by* $\underline{a}^T\underline{\beta}$ *is estimable if any of these equivalent conditions hold:*

1. $\underline{a}^T\underline{\beta} = E(\mathbf{t}^T\mathbf{y})$ *for any vector* $\mathbf{t}$
2. $\underline{a} = \mathbf{X}^T\mathbf{t}$ *for any vector* $\mathbf{t}$
3. $\underline{a} \in \mathcal{C}(\mathbf{X}^T)$

When $\mathbf{X}$ *has full column rank,* $\underline{a}^T\underline{\beta}$ *is estimable for any* $\underline{a} \in \mathbb{R}^{p+1}$.

See Searle (1971:180-188) or Shao (1999:148-150) for more details.

Definition A.2.2 ***Best linear unbiased estimator (BLUE)***

A linear estimate $\mathbf{c}^T\mathbf{y}$ *is the* BLUE *of* $\underline{a}^T\underline{\beta}$ *if it is unbiased and has the lowest variance among all linear unbiased estimates. That is, both these conditions are satisfied:*

1. $E(\mathbf{c}^T\mathbf{y}) = \underline{a}^T\underline{\beta}$
2. $\text{var}(\mathbf{c}^T\mathbf{y}) \leqslant \text{var}(\mathbf{d}^T\mathbf{y})$ *for any other unbiased linear estimate* $\mathbf{d}^T\mathbf{y}$.

Theorem A.2.1 ***Gauss-Markov Theorem***

Consider the linear model $\mathbf{y} = \mathbf{X}\underline{\beta}+\varepsilon$ *with* $E(\varepsilon) = \mathbf{0}$ *and* $\text{var}(\varepsilon) = \sigma^2\mathbf{I}$. *If* $\underline{a}^T\underline{\beta}$ *is an estimable function then the LSE* $\underline{a}^T\underline{\hat{\beta}}$ *is the BLUE of* $\underline{a}^T\underline{\beta}$.

See Gentle (2007:234-235), Seber & Lee (2003:42-43) or Shao (1999:155) for a proof.

Definition A.2.3 ***Uniformly minimum variance unbiased estimator (UMVUE)***

An estimate $g(\mathbf{y})$ *is the UMVUE of* $\underline{a}^T\underline{\beta}$ *if it is unbiased and has the lowest variance among all unbiased estimates. That is, both these conditions are satisfied:*

1. $E(g(\mathbf{y})) = \underline{a}^T\underline{\beta}$
2. $\text{var}(g(\mathbf{y})) \leqslant \text{var}(h(\mathbf{y}))$ *for any other unbiased estimate* $h(\mathbf{y})$.

See Spanos (1989:232-244), Shao (1999:127-139) or Casella & Berger (2002:334-348) for more information about unbiased estimates.

Theorem A.2.2 ***Properties of LSEs under normality***

Consider the linear model $\mathbf{y} = \mathbf{X}\underline{\beta}+\varepsilon$ *with* $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

1. *If* $\underline{a}^T\underline{\beta}$ *is an estimable function then the LSE* $\underline{a}^T\underline{\hat{\beta}}$ *is the UMVUE of* $\underline{a}^T\underline{\beta}$.
2. *If* $\text{rank}(\mathbf{X}) = r$ *then* $\hat{\sigma}^2 = RSS\left(\underline{\hat{\beta}}\right)\Big/(n-r)$ *is the UMVUE of* σ^2

See Shao (1999:152-154) for a proof.

Definition A.2.4 ***Sampling properties of an estimator***

A sample estimate of a parameter θ *calculated from observed data* a *is given by* $\hat{\theta}(a)$. *The estimator is a function of the random variable* A *and is given by* $\hat{\theta}(A)$. *The estimator is denoted in short by* $\hat{\theta}$ *and has the following finite sample properties:*

1. *The bias of the estimator is* $B(\hat{\theta}) = E(\hat{\theta}) - \theta$
2. *The variance of the estimator is* $\text{var}(\hat{\theta}) = E\left[\hat{\theta} - E(\hat{\theta})\right]^2$
3. *The MSE of an estimator is* $MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \text{var}(\hat{\theta}) + B(\hat{\theta})^2$.

See Casella & Berger (2002:330). A way to decompose the MSE into the squared bias and variance of the estimate is shown in Section B.1.1.

Theorem A.2.3 ***Cramér-Rao lower bound***

The variance of any estimate of θ is bound by the Cramér-Rao lower bound, which is given by

$$CR\left(\theta\right) = \frac{\left[\frac{d}{d\theta}E\left(\theta\right)\right]^2}{I_n\left(\theta\right)} = \frac{\left[1 + B'\left(\theta\right)\right]^2}{I_n\left(\theta\right)},$$

where $I_n\left(\theta\right) = E\left[\frac{\partial}{\partial\theta}\ln L\left(\theta\right)\right]^2$ *is called the Fisher information of the sample and* $L\left(\theta\right)$ *is the likelihood function. That is,* $\text{var}\left(\hat{\theta}\right) \geqslant CR\left(\theta\right)$ *for all estimates* $\hat{\theta}$.

See Spanos (1989:237-241), Shao (1999:135-138) and Casella & Berger (2002:335-338) for a proof and more information. Some notes:

- By the definition of bias, $E\left(\hat{\theta}\right) = \theta + B\left(\hat{\theta}\right)$ so that $\frac{d}{d\theta}E\left(\theta\right) = 1 + B'\left(\theta\right)$.
- Since $MSE\left(\hat{\theta}\right) = \text{var}\left(\hat{\theta}\right) + B\left(\hat{\theta}\right)^2$, the MSE is bounded by $MSE\left(\hat{\theta}\right) \geqslant CR\left(\theta\right) + B\left(\theta\right)^2$.
- An unbiased estimator is said to be fully efficient of its variance equals the Cramér-Rao lower bound which is simplified to $I_n\left(\theta\right)^{-1}$. Such an estimator is also the UMVUE.

Definition A.2.5 ***Convergence of random variables***

Suppose $A_n = A_1, A_2, \ldots, A_n$ *is a sequence of random variables.*

1. *The sequence converges to A in probability, denoted by* $A_n \xrightarrow{p} A$, *if* $\lim_{n\to\infty} P\left(|A_n - A| < \epsilon\right) = 1$ *or equivalently* $\lim_{n\to\infty} P\left(|A_n - A| \geqslant \epsilon\right) = 0$ *for every* $\epsilon > 0$.
2. *The sequence converges to A in* ℓ_q, *or in q-th moment, denoted by* $A_n \xrightarrow{\ell_q} A$,

   if $\lim_{n\to\infty} E\left(|A_n - A|^q\right) = 0$ *for* $q > 0$.
3. *The sequence converges to A almost surely, denoted by* $A_n \xrightarrow{a.s.} A$, *if* $P\left(\lim_{n\to\infty} A_n = A\right) = 1$.
4. *The sequence converges to A in distribution, denoted by* $A_n \xrightarrow{d} A$, *if* $\lim_{n\to\infty} F_{A_n}\left(a\right) = F_A\left(a\right)$ *for all a where the cumulative distribution function* $F_A\left(a\right) = P\left(A \leqslant a\right)$ *is continuous.*

See Casella & Berger (2002:232-245) or Shao (1999:38-41) for more information about convergence.

Definition A.2.6 ***Order of a sequence***

Suppose A_n is a sequence of random variables, b_n and c_n are two sequences of real numbers and d_n is a sequence of positive real numbers.

1. *The sequence b_n is at most of order c_n, denoted by $b_n = O(c_n)$, if $\lim_{n\to\infty} \frac{|b_n|}{c_n} < \epsilon$ for some $\epsilon > 0$.*
2. *The sequence b_n is of smaller order than c_n, denoted by $b_n = o(c_n)$, if $\lim_{n\to\infty} \frac{b_n}{c_n} = 0$.*
3. *The sequence A_n is at most of order d_n in probability, denoted by $A_n = O_p(d_n)$, if $\frac{A_n}{d_n} \xrightarrow{p} a_n$ for some sequence $a_n = O(1)$.*
4. *The sequence A_n is of smaller order than d_n in probability, denoted by $A_n = o_p(d_n)$, if $\frac{A_n}{d_n} \xrightarrow{p} 0$.*

See Spanos (1989:195-198) and Shao (1999:42).

Definition A.2.7 ***Asymptotic properties of an estimator***

Suppose $\hat{\theta}_n = \hat{\theta}(A_1, A_2, \ldots, A_n)$ is a sequence of estimators. The estimator $\hat{\theta}$ has the following asymptotic properties:

1. *Asymptotic accuracy:*
   (a) *An estimator is consistent if $\hat{\theta}_n \xrightarrow{p} \theta$*
   (b) *An estimator is ℓ_q-consistent if $\hat{\theta}_n \xrightarrow{\ell_q} \theta$.*
   (c) *An estimator is strongly consistent if $\hat{\theta}_n \xrightarrow{a.s.} \theta$*
   (d) *An estimator is a_n-consistent if $a_n\left|\hat{\theta}_n - \theta\right| = O_p(1)$, where a_n is a sequence of positive constants.*
2. *Asymptotic normality:*
   an estimator is asymptotically normal if $\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N\left(0, \tilde{\text{var}}(\theta)\right)$, where $\tilde{\text{var}}(\theta) > 0$ is the asymptotic variance of θ.
3. *Asymptotic efficience:*
   an asymptotically normal estimator is asymptotically efficient if $\tilde{\text{var}}(\theta) = I_\infty(\theta)^{-1}$, where $I_\infty(\theta) = \lim_{n\to\infty}\left(\frac{1}{n} I_n(\theta)\right)$. That is, the asymptotic variance, $\tilde{\text{var}}(\theta) = \lim_{n\to\infty} \text{var}\left(\hat{\theta}_n\right)$ equals the limit of the Cramér-Rao lower bound.

4. *Asymptotic bias:*

   an estimator is asymptotically unbiased if $\text{var}\left(\hat{\theta}_n\right) = O\left(1/n\right)$ *and* $\lim_{n\to\infty} \sqrt{n}\left(\hat{\theta}_n - \theta\right) = 0$.

A note on (1b), the following are equivalent:

- $\hat{\theta}$ is ℓ_2 consistent (for $q = 2$) if $\lim_{n\to\infty} E\left(\left|\hat{\theta}_n - \theta\right|^2\right) = 0$.
- The estimator is consistent in MSE since $MSE\left(\hat{\theta}_n\right) = E\left(\left|\hat{\theta}_n - \theta\right|^2\right)$.

The same is true for (1a):

- $\hat{\theta}$ is consistent if $\lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| \geqslant \epsilon\right) = 0$.
- The estimator is consistent in MSE since $P\left(\left|\hat{\theta}_n - \theta\right| \geqslant \epsilon\right) \leqslant E\left(\hat{\theta}_n - \theta\right)^2 \Big/ \epsilon^2$ by Chebyshev's inequality (see Shao (1999:51)) and $MSE\left(\hat{\theta}_n\right) = E\left(\hat{\theta}_n - \theta\right)^2$.

Since $MSE\left(\hat{\theta}_n\right) = \text{var}\left(\hat{\theta}_n\right) + B\left(\hat{\theta}_n\right)^2$, $\hat{\theta}$ is consistent in MSE if both

- $\lim_{n\to\infty} \text{var}\left(\hat{\theta}_n\right) = 0$ and
- $\lim_{n\to\infty} B\left(\hat{\theta}_n\right) = 0$.

See Spanos (1989:244-247), Shao (1999:102-109) or Casella & Berger (2002:467-473) for more information.

Definition A.2.8 ***Coefficient of determination***

The coefficient of determination is given by

$$R^2 = \frac{\sum\left(\hat{y}_i - y\right)^2}{\sum\left(y_i - y\right)^2}.$$

R^2 *measures the proportion of the total variation in* v *that is explained by the model* $\hat{Y}$ *and its range is* $0 \leqslant R^2 \leqslant 1$. *It can also be written as*

$$R^2 = 1 - \frac{\sum\left(y_i - \hat{y}_i\right)^2}{\sum\left(y_i - y\right)^2} = 1 - \frac{RSS\left(\hat{\underline{\beta}}\right)}{\mathbf{v}^T\mathbf{v}}.$$

See Draper & Smith (1998:33). Seber & Lee (2003:110-112) show that $R^2 = \text{corr}\left(Y, \hat{Y}\right)$ and when $p = 1$ for a straight line fit then $R^2 = \text{corr}\left(X, Y\right)$.

A.3 Optimization

Definition A.3.1 ***Convex functions***

1. *$\mathcal{A}$ is a convex set if the line segment between any two points in a $\mathcal{A}$ also lies in $\mathcal{A}$. That is, for all $\underline{a}, \underline{b} \in \mathcal{A}$ and $c \in (0,1)$,*

$$c\underline{a} + (1-c)\,\underline{b} \in \mathcal{A}.$$

2. *f is a convex function if its domain is a convex set, say $\mathcal{A}$, and the line segment between any two points on the function lies above the function. That is, for all $\underline{a}, \underline{b} \in \mathcal{A}$ and $c \in (0,1)$,*

$$f\left(c\underline{a} + (1-c)\,\underline{b}\right) \leqslant cf\left(\underline{a}\right) + (1-c)\,f\left(\underline{b}\right).$$

3. *f is a strictly convex function if for all $\underline{a}, \underline{b} \in \mathcal{A}$ and $c \in (0,1)$,*

$$f\left(c\underline{a} + (1-c)\,\underline{b}\right) < cf\left(\underline{a}\right) + (1-c)\,f\left(\underline{b}\right).$$

See Boyd & Vandenberghe (2004:23-25,67-68) for more information.

Definition A.3.2 ***First and second order conditions***

1. *If f is differentiable then f is convex if and only if the domain of f is the convex set $\mathcal{A}$ and for all $\underline{a}, \underline{b} \in \mathcal{A}$,*

$$f\left(\underline{b}\right) \geqslant f\left(\underline{a}\right) + \nabla f\left(\underline{a}\right)^T\left(\underline{b} - \underline{a}\right).$$

For strict convexity, strict inequality is required, $f\left(\underline{b}\right) > f\left(\underline{a}\right) + \nabla f\left(\underline{a}\right)^T\left(\underline{b} - \underline{a}\right)$.

2. *If f is twice differentiable then f is convex if and only if the domain of f is the convex set $\mathcal{A}$ and for all $\underline{a} \in \mathcal{A}$, $\nabla^2 f\left(\underline{a}\right)$ is positive semidefinite. For strict convexity, $\nabla^2 f\left(\underline{a}\right)$ must be positive definite.*

See Boyd & Vandenberghe (2004:69-71) for more information.

Definition A.3.3 ***Optimization terminology***

For any optimization problem

$$\begin{array}{ll} \textit{minimize} & f(\underline{\theta}) \\ \textit{subject to} & g_i(\underline{\theta}) \leqslant 0, i = 1, 2, \ldots, I \\ & h_j(\underline{\theta}) = 0, j = 1, 2, \ldots, J, \end{array} \tag{A.3.1}$$

1. *The domain of the problem is the set of points $\mathcal{A}$ for which the objective function f and all the constraint functions g and h are satisfied.*

2. *A point $\underline{\theta} \in \mathcal{A}$ is feasible if it satisfies all the constraints $g_i(\underline{\theta}) \leqslant 0, h_j(\underline{\theta}) = 0 \forall i, j$. The set of all feasible points is called the feasible set $\mathcal{F}$.*

3. *If $\underline{\theta} \in \mathcal{F}$ and $g_i(\underline{\theta}) = 0$ then the inequality constraint $g_i(\underline{\theta}) \leqslant 0$ is active.*

4. *If $\underline{\theta} \in \mathcal{F}$ and $g_i(\underline{\theta}) < 0$ then the inequality constraint $g_i(\underline{\theta}) \leqslant 0$ is inactive.*

5. *The optimal value of the problem is $opt = \inf\left\{ f(\underline{\theta}) | \, g_i(\underline{\theta}) \leqslant 0, h_j(\underline{\theta}) = 0 \forall i, j \right\}$.*

6. *$\hat{\underline{\theta}}$ is an optimal point if $\hat{\underline{\theta}} \in \mathcal{F}$ and $f\left(\hat{\underline{\theta}}\right) = opt$. This is also known as the globally optimal point.*

7. *$\hat{\underline{\theta}}$ is sub-optimal if $\hat{\underline{\theta}} \in \mathcal{F}$ and $f\left(\hat{\underline{\theta}}\right) = opt + \epsilon$ for some $\epsilon > 0$.*

8. *$\hat{\underline{\theta}}$ is locally optimal if $\hat{\underline{\theta}} \in \mathcal{F}$ and*

   $f\left(\hat{\underline{\theta}}\right) = \inf\left\{ f(\underline{\theta}_1) | \, g_i(\underline{\theta}_1) \leqslant 0, h_j(\underline{\theta}_1) = 0, \|\underline{\theta}_1 - \underline{\theta}\| \leqslant \epsilon \forall i, j \text{ and some } \epsilon > 0 \right\}$.

See Boyd & Vandenberghe (2004:127-129) for more information. The problem (A.3.1) is the standard form of the optimization problem. Boyd & Vandenberghe (2004:129-135) discuss problems that are equivalent to (A.3.1), including maximization, change of variables, function transformations, slack variables, eliminating or introducing equality constraints, sequentially optimizing over some variables, including the objective function as a constraint and including a constraint implicitly in the objective function.

Definition A.3.4 ***Convex optimization***

If $f, g_1, \ldots, g_I$ are convex and $h_1, \ldots, h_J$ are affine then the optimization problem (A.3.1) is convex and the following hold:

1. *Any locally optimal point is also globally optimal.*

2. $\hat{\underline{\theta}}$ *is optimal if and only if* $\hat{\underline{\theta}} \in \mathcal{F}$ *and* $\nabla f\left(\hat{\underline{\theta}}\right)^T \left(\underline{\theta} - \hat{\underline{\theta}}\right) \geqslant 0$ *for all* $\underline{\theta} \in \mathcal{F}$.

3. *For unconstrained problems, a necessary and sufficient condition for* $\hat{\underline{\theta}}$ *to be optimal is* $\nabla f\left(\hat{\underline{\theta}}\right) = \underline{0}$.

See Boyd & Vandenberghe (2004:136-144) for more information about convex optimization.

Definition A.3.5 ***Descent method***

A descent method solves the optimization problem

$$\text{minimize } f\left(\underline{\theta}\right),$$

where f *is convex and twice differentiable. The algorithm can be used if there is no closed form for the necessary and sufficient optimality condition*

$$\nabla f\left(\hat{\underline{\theta}}\right) = \underline{0}.$$

A general descent algorithm produces a minimizing sequence by iteratively updating

$$\underline{\theta}^{(k+1)} = \underline{\theta}^{(k)} + s^{(k)} \Delta\underline{\theta}^{(k)},$$

where $\Delta\underline{\theta}^{(k)}$ *is the search direction and* $s^{(k)} > 0$ *is the step length. The search direction is chosen so that* f *descends,*

$$f\left(\underline{\theta}^{(k+1)}\right) < f\left(\underline{\theta}^{(k)}\right)$$

unless $\left\|\nabla f\left(\underline{\theta}^{(k)}\right)\right\| \leqslant \epsilon$ *for small* $\epsilon > 0$ *so that* $\underline{\theta}^{(k)}$ *is optimal. Since convexity implies that* $f\left(\underline{\theta}\right) \geqslant f\left(\hat{\underline{\theta}}\right)$ *when* $\nabla f\left(\hat{\underline{\theta}}\right)^T \left(\underline{\theta} - \hat{\underline{\theta}}\right) \geqslant 0$, *the search direction must satisfy*

$$\nabla f\left(\underline{\theta}^{(k)}\right)^T \Delta\underline{\theta}^{(k)} < 0.$$

See Boyd & Vandenberghe (2004:463-484) for more information about descent methods.

Given a starting point $\underline{\theta}^{(0)}$, the algorithm repeatedly performs the following steps until convergence of the gradient:

1. Find the search direction $\Delta\underline{\theta}$
   - *Gradient descent:* $\Delta\underline{\theta}_{gd} = -\nabla f(\underline{\theta})$
   - *Normalized steepest descent:* $\Delta\underline{\theta}_{nsd} = \arg\min_v \left\{ \nabla f(\underline{\theta})^T v \middle| \|v\| \leqslant 1 \right\}$ for any norm $\|\cdot\|$. Thus, $\Delta\underline{\theta}_{nsd}$ extends $\underline{\theta}$ by the greatest distance in the direction of $-\nabla f(\underline{\theta})$ while remaining in the unit ball of $\|\cdot\|$.
   - *Steepest descent:* $\Delta\underline{\theta}_{sd} = \Delta\underline{\theta}_{nsd} \|\nabla f(\underline{\theta})\|_*$ where $\|\cdot\|_*$ is the dual norm which is given by $\|\underline{a}\|_* = \sup\left\{ \left|\underline{a}^T \underline{b}\right| \middle| \|\underline{b}\| \leqslant 1 \right\}$ such that $\|\underline{a}\| \|\underline{b}\|_* \geqslant \underline{a}^T \underline{b}$
2. Line search - find s which minimizes f along $\{\underline{\theta} + t\Delta s | \geqslant 0\}$
   - *Exact:* $s = \arg\min_{t \geqslant 0} f(\underline{\theta} + t\Delta\underline{\theta})$
   - *Backtracking:* $s = sb$ while $f(\underline{\theta} + s\Delta\underline{\theta}) > f(\underline{\theta}) + as\nabla f(\underline{\theta})^T \Delta\underline{\theta}$, where $a \in (0, 0.5)$, $b \in (0,1)$ and we begin with $s = 1$
3. Update $\underline{\theta}_+ = \underline{\theta} + s\Delta\underline{\theta}$

Note that for any ℓ_q-norm, the dual norm is the ℓ_r-norm where $1/q + 1/r = 1$. The ℓ_1 norm is the dual norm of the ℓ_∞-norm and conversely, the ℓ_∞-norm is the dual norm of the ℓ_1 norm. For more information, see Boyd & Vandenberghe (2004:637).

Definition A.3.6 ***Karush-Kuhn-Tucker conditions***

For any optimization problem (A.3.1), the Lagrangian function is given by

$$l\left(\underline{\theta}, \underline{\eta}, \underline{\mu}\right) = f(\underline{\theta}) + \sum_{i=1}^{I} \eta_i g_i(\underline{\theta}) + \sum_{j=1}^{J} \mu_j h_j(\underline{\theta}),$$

where η_i and μ_i are called dual variables or Lagrange multipliers. If $f, g_1, \ldots, g_I, h_1, \ldots, h_J$ are differentiable then the following conditions are necessary for $\hat{\underline{\theta}}$ and $\left(\hat{\underline{\eta}}, \hat{\underline{\mu}}\right)$ to be optimal:

1. $g_i\left(\hat{\underline{\theta}}\right) \leqslant 0, i = 1, 2, \ldots, I,$
2. $\hat{\eta}_i \geqslant 0, i = 1, 2, \ldots, J,$

3. $\hat{\eta}_i g_i\left(\hat{\underline{\theta}}\right) = 0, , i = 1, 2, \ldots, J,$

4. $\nabla l\left(\hat{\underline{\theta}}, \hat{\underline{\eta}}, \hat{\underline{\mu}}\right) = \nabla f\left(\hat{\underline{\theta}}\right) + \sum_{i=1}^{I} \hat{\eta}_i \nabla g_i\left(\hat{\underline{\theta}}\right) + \sum_{j=1}^{J} \hat{\mu}_j \nabla h_j\left(\hat{\underline{\theta}}\right) = 0,$

5. $h_j\left(\hat{\underline{\theta}}\right) = 0, i = 1, 2, \ldots, J.$

These are called the Karush-Kuhn-Tucker (KKT) conditions. If the problem is convex then the KKT conditions are also sufficient for optimality.

See Boyd & Vandenberghe (2004:243-244) for further details.

Definition A.3.7 ***Minimax optimality***

An estimator $\hat{\underline{\theta}}_1$ is minimax optimal for $\underline{\theta}$ if it minimizes the maximum MSE. That is, for all $\hat{\underline{\theta}}_2$,

$$\sup_{\theta} MSE\left(\hat{\underline{\theta}}_1\right) \leqslant \sup_{\theta} MSE\left(\hat{\underline{\theta}}_2\right).$$

In other words, a minimax estimator performs the best possible in the worst case. See Rao (1973:341) and Shao (1999:223).

Definition A.3.8 ***Oracle properties***

Suppose the true regression parameters $\underline{\alpha}$ include a set of important effects indexed by $\mathcal{D} = \left\{j : \alpha_j \neq 0\right\}$. Then $\underline{\alpha} = \left(\underline{\alpha}_{\mathcal{D}}, \underline{\alpha}_{\mathcal{D}^c}\right)^T = \left(\underline{\alpha}_{\mathcal{A}}, \underline{0}\right)^T$ and the true model is given by

$$V_i = \sum_{j \in \mathcal{A}} Z_{ij} \alpha_j + \varepsilon_i.$$

Furthermore, suppose that $\text{var}\left(\underline{\alpha}_{\mathcal{D}}\right) = \Sigma$. *An oracle variable selection method for linear models is able to:*

1. *Select the correct parameters consistently, that is,* $\lim_{n \to \infty} P\left(\widehat{\mathcal{D}}_n = \mathcal{D}\right) = 1$

2. *Estimate the nonzero parameters efficiently, that is,* $\sqrt{n}\left(\hat{\underline{\alpha}}_{\mathcal{A}} - \underline{\alpha}_{\mathcal{A}}\right) \xrightarrow{d} N\left(\underline{0}, \Sigma\right).$

See Fan & Li (2001) and Zou (2006) for more information.

A.4 Geometry

Definition A.4.1 ***Conic sections***

Conic sections are surfaces defined by the intersections of a quadric surface with the coordinate planes. A conic section in two variables θ_1 and θ_2 has the general form

$$a\theta_1^2 + 2b\theta_1\theta_2 + c\theta_2^2 + 2d\theta_1 + 2e\theta_2 + f = 0. \tag{A.4.1}$$

The equation can be written as a quadratic form

$$\underline{\theta}^T \mathbf{A} \underline{\theta} = 0,$$

where $\underline{\theta}^T = (\theta_1, \theta_2, 1)$ *and*

$$\mathbf{A} = \begin{bmatrix} a & b & d \\ b & c & e \\ d & e & f \end{bmatrix}. \tag{A.4.2}$$

If $|\mathbf{A}| = 0$ *then the conic section is degenerate, otherwise it is not degenerate. The top* 2×2 *partition of* $\mathbf{A}$ *can be used to write the equation as*

$$\begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + 2dx + 2ev + f = 0.$$

The determinant of this partition is known as the discriminant,

$$\triangle = \begin{vmatrix} a & b \\ b & c \end{vmatrix} = ac - b^2. \tag{A.4.3}$$

The shape of the conic section is determined by the discriminant. In the non-degenrate case:

- *if* $\triangle = 0$ *then the conic is a parabola,*
- *if* $\triangle > 0$ *then the conic is an ellipse,*
- *if* $\triangle < 0$ *then the conic is an hyperbola.*

The center of the conic is the point where the gradient of the quadratic form is zero, and is given by

$$\left(\frac{be - cd}{ac - b^2}, \frac{bd - ae}{ac - b^2} \right). \tag{A.4.4}$$

If $b \neq 0$, then the axes of the conic section are not parallel to the coordinate axes. The angle ϑ that the axes of the conic section makes with the coordinate axes can be found by rotating the system to eliminate the $b\theta_1\theta_2$-term and is determined by

$$\cot 2\vartheta = \frac{a - c}{2b}. \tag{A.4.5}$$

More information about quadric surfaces and conic sections can be found in Siceloff *et al.* (1922) or any text on analytic geometry.

Appendix B

Calculations

B.1 Estimation and Prediction Accuracy

B.1.1 Mean Squared Error

The MSE of an estimator $\hat{\beta}$ is a measure of how well it estimates the true β and is given by $MSE\left(\hat{\beta}\right) = E\left(\hat{\beta}-\beta\right)^2$. We can write

$$\hat{\beta}-\beta = \left[\hat{\beta}-E\left(\hat{\beta}\right)\right]-\left[E\left(\hat{\beta}\right)-\beta\right].$$

Squaring both sides gives

$$\begin{aligned}&\left(\hat{\beta}-\beta\right)^2\\&=\left[\hat{\beta}-E\left(\hat{\beta}\right)\right]^2+\left[E\left(\hat{\beta}\right)-\beta\right]^2-2\left[\hat{\beta}-E\left(\hat{\beta}\right)\right]\left[E\left(\hat{\beta}\right)-\beta\right],\end{aligned}$$

and taking expectations gives

$$\begin{aligned}&E\left[\hat{\beta}-E\left(\hat{\beta}\right)\right]\left[E\left(\hat{\beta}\right)-\beta\right]\\&=\left[E\left(\hat{\beta}\right)-E\left(\hat{\beta}\right)\right]\left[E\left(\hat{\beta}\right)-\beta\right]=0,\end{aligned}$$

so that

$$\begin{aligned}MSE\left(\hat{\beta}\right)&=E\left(\hat{\beta}-\beta\right)^2\\&=E\left[\hat{\beta}-E\left(\hat{\beta}\right)\right]^2+\left[E\left(\hat{\beta}\right)-\beta\right]^2\\&=\text{var}\left(\hat{\beta}\right)+B\left(\hat{\beta}\right)^2.\end{aligned}$$

Thus the MSE of an estimate is a trade-off between its variance and its bias.

The result can easily be generalized for a vector of estimates $\underline{\hat{\beta}}$. We have $MSE\left(\underline{\hat{\beta}}\right) = E\left\|\underline{\hat{\beta}}-\underline{\beta}\right\|^2$. We

can write

$$\underline{\hat{\beta}} - \underline{\beta} = \left[\underline{\hat{\beta}} - E\left(\underline{\hat{\beta}}\right)\right] - \left[E\left(\underline{\hat{\beta}}\right) - \underline{\beta}\right].$$

Taking squared norms on both sides gives

$$\begin{aligned}&\left\|\underline{\hat{\beta}} - \underline{\beta}\right\|^2 \\ &= \left\|\underline{\hat{\beta}} - E\left(\underline{\hat{\beta}}\right)\right\|^2 + \left\|E\left(\underline{\hat{\beta}}\right) - \underline{\beta}\right\|^2 - 2\left[\underline{\hat{\beta}} - E\left(\underline{\hat{\beta}}\right)\right]^T \left[E\left(\underline{\hat{\beta}}\right) - \underline{\beta}\right],\end{aligned}$$

and taking expectations gives

$$\begin{aligned}&E\left[\underline{\hat{\beta}} - E\left(\underline{\hat{\beta}}\right)\right]^T \left[E\left(\underline{\hat{\beta}}\right) - \underline{\beta}\right] \\ &= \left[E\left(\underline{\hat{\beta}}\right) - E\left(\underline{\hat{\beta}}\right)\right]^T \left[E\left(\underline{\hat{\beta}}\right) - \underline{\beta}\right] = \underline{0},\end{aligned}$$

so that

$$\begin{aligned}MSE\left(\underline{\hat{\beta}}\right) &= E\left\|\underline{\hat{\beta}} - \underline{\beta}\right\|^2 \\ &= E\left\|\underline{\hat{\beta}} - E\left(\underline{\hat{\beta}}\right)\right\|^2 + \left\|E\left(\underline{\hat{\beta}}\right) - \underline{\beta}\right\|^2,\end{aligned}$$

where $E\left\|\underline{\hat{\beta}} - E\left(\underline{\hat{\beta}}\right)\right\|^2$ is the total variance of $\underline{\hat{\beta}}$. We can show this by noting that

$$\begin{aligned}&E\left\|\underline{\hat{\beta}} - E\left(\underline{\hat{\beta}}\right)\right\|^2 \\ &= E\left[\left(\underline{\hat{\beta}} - E\left(\underline{\hat{\beta}}\right)^2\right)^T \left(\underline{\hat{\beta}} - E\left(\underline{\hat{\beta}}\right)^2\right)\right] \\ &= E\left\{\text{tr}\left[\left(\underline{\hat{\beta}} - E\left(\underline{\hat{\beta}}\right)^2\right)^T \left(\underline{\hat{\beta}} - E\left(\underline{\hat{\beta}}\right)^2\right)\right]\right\} \\ &= E\left\{\text{tr}\left[\left(\underline{\hat{\beta}} - E\left(\underline{\hat{\beta}}\right)^2\right)\left(\underline{\hat{\beta}} - E\left(\underline{\hat{\beta}}\right)^2\right)^T\right]\right\} \\ &= \text{tr}\left\{E\left[\left(\underline{\hat{\beta}} - E\left(\underline{\hat{\beta}}\right)^2\right)\left(\underline{\hat{\beta}} - E\left(\underline{\hat{\beta}}\right)^2\right)^T\right]\right\} \\ &= \text{tr}\left[\text{var}\left(\underline{\hat{\beta}}\right)\right].\end{aligned}$$

Thus,

$$\begin{aligned} MSE\left(\underline{\hat{\beta}}\right) &= \text{tr}\left[\text{var}\left(\underline{\hat{\beta}}\right)\right] + \left\|E\left(\underline{\hat{\beta}}\right) - \underline{\beta}\right\|^2 \\ &= \sum_{j=1}^{p} \text{var}\left(\hat{\beta}_j\right) + \sum_{j=1}^{p} B\left(\hat{\beta}_j\right)^2 \\ &= \sum_{j=1}^{p} MSE\left(\hat{\beta}_j\right). \end{aligned}$$

B.1.2 Prediction Error

Individual Observation

Suppose we would like to predict a new response at a new observation $(\underline{x}_0, y_0)$, where y_0 has the same probability structure as the elements of y: $E(y_0) = f(\underline{x}_0)$, $\text{var}(y_0) = \sigma^2$ and $\text{cov}(y_0, y_i) = 0$. The expected PE of the predictor $\hat{f}(\underline{x}_0)$ is a measure of how well it predicts the new response and is given by $PE\left(\hat{f}(\underline{x}_0)\right) = E\left[y_0 - \hat{f}(\underline{x}_0)\right]^2$. We can write

$$\begin{aligned} y_0 - \hat{f}(\underline{x}_0) &= \left[y_0 - E(y_0)\right] - \left[\hat{f}(\underline{x}_0) - E(y_0)\right] \\ &= \left[y_0 - E(y_0)\right] - \left[\hat{f}(\underline{x}_0) - f(\underline{x}_0)\right], \end{aligned}$$

since $E(y_0) = f(\underline{x}_0)$. Squaring both sides gives

$$\begin{aligned} &\left[y_0 - \hat{f}(\underline{x}_0)\right]^2 \\ &= \left[y_0 - E(y_0)\right]^2 + \left[\hat{f}(\underline{x}_0) - f(\underline{x}_0)\right]^2 - 2\left[y_0 - E(y_0)\right]^T \left[\hat{f}(\underline{x}_0) - f(\underline{x}_0)\right], \end{aligned}$$

and taking expectations gives

$$\begin{aligned} &E\left[y_0 - E(y_0)\right]^T \left[\hat{f}(\underline{x}_0) - f(\underline{x}_0)\right] \\ &= \left[E(y_0) - E(y_0)\right]^T \left[E\left(\hat{f}(\underline{x}_0)\right) - f(\underline{x}_0)\right] = 0, \end{aligned}$$

so that

$$\begin{aligned} PE\left(\hat{f}\left(\underline{x}_0\right)\right) &= E\left[y_0 - \hat{f}\left(\underline{x}_0\right)\right]^2 \\ &= E\left[y_0 - E\left(y_0\right)\right]^2 + E\left[\hat{f}\left(\underline{x}_0\right) - f\left(\underline{x}_0\right)\right]^2 \\ &= \text{var}\left(y_0\right) + MSE\left(\hat{f}\left(\underline{x}_0\right)\right) \\ &= \sigma^2 + MSE\left(\hat{f}\left(\underline{x}_0\right)\right). \end{aligned}$$

Thus, the PE is also a trade-off between the variance and bias but it includes an additional irreducible variance σ^2 to account for the variation in the data.

Similarly, if we would like to predict m new observations, $\left(\underline{x}_{0,1}, y_{0,1}\right), \left(\underline{x}_{0,2}, y_{0,2}\right), \ldots, \left(\underline{x}_{0,m}, y_{0,m}\right)$, then the PE is given by $PE\left(\hat{f}\left(\mathbf{X}_0\right)\right) = E\left\|\mathbf{y}_0 - \hat{f}\left(\mathbf{X}_0\right)\right\|^2$. We can write

$$\begin{aligned} \mathbf{y}_0 - \hat{f}\left(\mathbf{X}_0\right) &= \left[\mathbf{y}_0 - E\left(\mathbf{y}_0\right)\right] - \left[E\left(\mathbf{y}_0\right) - \hat{f}\left(\mathbf{X}_0\right)\right] \\ &= \left[\mathbf{y}_0 - E\left(\mathbf{y}_0\right)\right] - \left[f\left(\mathbf{X}_0\right) - \hat{f}\left(\mathbf{X}_0\right)\right], \end{aligned}$$

since $E\left(\mathbf{y}_0\right) = f\left(\mathbf{X}_0\right)$. Taking squared norms on both sides gives

$$\begin{aligned} &\left\|\mathbf{y}_0 - \hat{f}\left(\mathbf{X}_0\right)\right\|^2 \\ &= \left\|\mathbf{y}_0 - E\left(\mathbf{y}_0\right)\right\|^2 + \left\|f\left(\mathbf{X}_0\right) - \hat{f}\left(\mathbf{X}_0\right)\right\|^2 + 2\left[\mathbf{y}_0 - E\left(\mathbf{y}_0\right)\right]^T\left[f\left(\mathbf{X}_0\right) - \hat{f}\left(\mathbf{X}_0\right)\right], \end{aligned}$$

and taking expectations gives

$$\begin{aligned} &E\left[\mathbf{y}_0 - E\left(\mathbf{y}_0\right)\right]^T\left[f\left(\mathbf{X}_0\right) - \hat{f}\left(\mathbf{X}_0\right)\right] \\ &= \left[E\left(\mathbf{y}_0\right) - E\left(\mathbf{y}_0\right)\right]^T\left[f\left(\mathbf{X}_0\right) - E\left(\hat{f}\left(\mathbf{X}_0\right)\right)\right] = \mathbf{0}, \end{aligned}$$

so that

$$\begin{aligned} PE\left(\hat{f}\left(\mathbf{X}_0\right)\right) &= E\left\|\mathbf{y}_0 - \hat{f}\left(\mathbf{X}_0\right)\right\|^2 \\ &= E\left\|\mathbf{y}_0 - E\left(\mathbf{y}_0\right)\right\|^2 + E\left\|f\left(\mathbf{X}_0\right) - \hat{f}\left(\mathbf{X}_0\right)\right\|^2 \\ &= \operatorname{tr}\left[\operatorname{var}\left(\mathbf{y}_0\right)\right] + MSE\left(\hat{f}\left(\mathbf{X}_0\right)\right) \\ &= m\sigma^2 + \sum_{i=1}^{m} MSE\left(\hat{f}\underline{x}_{0,i}\right) \\ &= \sum_{i=1}^{m}\left[\sigma^2 + MSE\left(\hat{f}\underline{x}_{0,i}\right)\right] \\ &= \sum_{i=1}^{m} PE\left(\hat{f}\underline{x}_{0,i}\right), \end{aligned}$$

since

$$\begin{aligned} & E\left\|\mathbf{y}_0 - E\left(\mathbf{y}_0\right)\right\|^2 \\ &= E\left[\left(\mathbf{y}_0 - E\left(\mathbf{y}_0\right)\right)^T\left(\mathbf{y}_0 - E\left(\mathbf{y}_0\right)\right)\right] \\ &= E\left\{\operatorname{tr}\left[\left(\mathbf{y}_0 - E\left(\mathbf{y}_0\right)\right)^T\left(\mathbf{y}_0 - E\left(\mathbf{y}_0\right)\right)\right]\right\} \\ &= E\left\{\operatorname{tr}\left[\left(\mathbf{y}_0 - E\left(\mathbf{y}_0\right)\right)\left(\mathbf{y}_0 - E\left(\mathbf{y}_0\right)\right)^T\right]\right\} \\ &= \operatorname{tr}\left\{E\left[\left(\mathbf{y}_0 - E\left(\mathbf{y}_0\right)\right)\left(\mathbf{y}_0 - E\left(\mathbf{y}_0\right)\right)^T\right]\right\} \\ &= \operatorname{tr}\left[\operatorname{var}\left(\mathbf{y}_0\right)\right]. \end{aligned}$$

B.1.3 Optimism

The expected optimism of the training sample is given by $\omega = E\left[PE_{in}\left(\hat{f}\left(\underline{x}\right)\right)\right] - E\left[TE\left(\hat{f}\left(\underline{x}\right)\right)\right]$. The in-sample error is

$$PE_{in}\left(\hat{f}\left(\underline{x}\right)\right) = \frac{1}{n}\sum_{i=1}^{n} E_{y_0}\left(y_{0,i} - \hat{f}\left(\underline{x}_i\right)\right)^2.$$

We can write

$$y_{0,i} - \hat{f}\left(\underline{x}_i\right) = \left[y_{0,i} - f\left(\underline{x}_i\right)\right] - \left[\hat{f}\left(\underline{x}_i\right) - E\left(\hat{f}\left(\underline{x}_i\right)\right)\right] - \left[E\left(\hat{f}\left(\underline{x}_i\right)\right) - f\left(\underline{x}_i\right)\right].$$

Squaring both sides gives

$$\begin{aligned}&\left[y_{0,i}-\hat{f}\left(\underline{x}_i\right)\right]^2\\&=\left[y_{0,i}-f\left(\underline{x}_i\right)\right]^2+\left[\hat{f}\left(\underline{x}_i\right)-f\left(\underline{x}_i\right)\right]^2-2\left[y_{0,i}-f\left(\underline{x}_i\right)\right]^T\left[\hat{f}\left(\underline{x}_i\right)-f\left(\underline{x}_i\right)\right].\end{aligned}$$

Taking expectations over y_0 gives

$$\begin{aligned}&E_{y_0}\left[y_{0,i}-f\left(\underline{x}_i\right)\right]^T\left[\hat{f}\left(\underline{x}_i\right)-f\left(\underline{x}_i\right)\right]\\&=\left[E_{y_0}\left(y_{0,i}\right)-E_{y_0}\left(y_{0,i}\right)\right]^T\left[\hat{f}\left(\underline{x}_i\right)-f\left(\underline{x}_i\right)\right]=0,\end{aligned}$$

since $E_{y_0}\left(y_{0,i}\right)=f\left(\underline{x}_i\right)$. Thus,

$$\begin{aligned}&E_{y_0}\left[y_{0,i}-\hat{f}\left(\underline{x}_i\right)\right]^2\\&=E_{y_0}\left[y_{0,i}-E_{y_0}\left(y_{0,i}\right)\right]^2+\left[\hat{f}\left(\underline{x}_i\right)-f\left(\underline{x}_i\right)\right]^2\\&=\text{var}\left(y_{0,i}\right)+\left[\hat{f}\left(\underline{x}_i\right)-f\left(\underline{x}_i\right)\right]^2,\end{aligned}$$

so that

$$\begin{aligned}&E_{\mathbf{y}}E_{y_0}\left[y_{0,i}-\hat{f}\left(\underline{x}_i\right)\right]^2\\&=\text{var}\left(y_{0,i}\right)+E\left[\hat{f}\left(\underline{x}_i\right)-f\left(\underline{x}_i\right)\right]^2\\&=\text{var}\left(y_{0,i}\right)+MSE\left(\hat{f}\left(\underline{x}_i\right)\right)\\&=\sigma^2+MSE\left(\hat{f}\left(\underline{x}_i\right)\right).\end{aligned}$$

The training error is

$$TE\left(\hat{f}\left(\underline{x}\right)\right)=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{f}\left(\underline{x}_i\right)\right)^2.$$

We can write

$$y_i-\hat{f}\left(\underline{x}_i\right)=\left[y_i-f\left(\underline{x}_i\right)\right]-\left[\hat{f}\left(\underline{x}_i\right)-E\left(\hat{f}\left(\underline{x}_i\right)\right)\right]-\left[E\left(\hat{f}\left(\underline{x}_i\right)\right)-f\left(\underline{x}_i\right)\right].$$

Squaring both sides gives

$$\begin{aligned}
&\left[y_i - \hat{f}\left(\underline{x}_i\right)\right]^2 \\
&= \left[y_i - f\left(\underline{x}_i\right)\right]^2 + \left[\hat{f}\left(\underline{x}_i\right) - E\left(\hat{f}\left(\underline{x}_i\right)\right)\right]^2 + \left[E\left(\hat{f}\left(\underline{x}_i\right)\right) - f\left(\underline{x}_i\right)\right]^2 \\
&- 2\left[y_i - f\left(\underline{x}_i\right)\right]\left[\hat{f}\left(\underline{x}_i\right) - E\left(\hat{f}\left(\underline{x}_i\right)\right)\right] \\
&- 2\left[y_i - f\left(\underline{x}_i\right)\right]\left[E\left(\hat{f}\left(\underline{x}_i\right)\right) - f\left(\underline{x}_i\right)\right] \\
&- 2\left[\hat{f}\left(\underline{x}_i\right) - E\left(\hat{f}\left(\underline{x}_i\right)\right)\right]\left[E\left(\hat{f}\left(\underline{x}_i\right)\right) - f\left(\underline{x}_i\right)\right]
\end{aligned}$$

Since $E\left(y_i\right) = f\left(\underline{x}_i\right)$, taking expectations gives

$$\begin{aligned}
&E\left[y_i - f\left(\underline{x}_i\right)\right]\left[E\left(\hat{f}\left(\underline{x}_i\right)\right) - f\left(\underline{x}_i\right)\right] \\
&= \left[E\left(y_i\right) - E\left(y_i\right)\right]\left[E\left(\hat{f}\left(\underline{x}_i\right)\right) - f\left(\underline{x}_i\right)\right] = 0
\end{aligned}$$

and

$$\begin{aligned}
&E\left[\hat{f}\left(\underline{x}_i\right) - E\left(\hat{f}\left(\underline{x}_i\right)\right)\right]\left[E\left(\hat{f}\left(\underline{x}_i\right)\right) - f\left(\underline{x}_i\right)\right] \\
&= \left[E\left(\hat{f}\left(\underline{x}_i\right)\right) - E\left(\hat{f}\left(\underline{x}_i\right)\right)\right]\left[E\left(\hat{f}\left(\underline{x}_i\right)\right) - f\left(\underline{x}_i\right)\right] = 0
\end{aligned}$$

so that

$$\begin{aligned}
&E\left[y_i - \hat{f}\left(\underline{x}_i\right)\right]^2 \\
&= E\left[y_i - E\left(y_i\right)\right]^2 + E\left[\hat{f}\left(\underline{x}_i\right) - E\left(\hat{f}\left(\underline{x}_i\right)\right)\right]^2 + E\left[E\left(\hat{f}\left(\underline{x}_i\right)\right) - f\left(\underline{x}_i\right)\right]^2 \\
&- 2E\left[y_i - E\left(y_i\right)\right]\left[\hat{f}\left(\underline{x}_i\right) - E\left(\hat{f}\left(\underline{x}_i\right)\right)\right] \\
&= \operatorname{var}\left(y_i\right) + \operatorname{var}\left(\hat{f}\left(\underline{x}_i\right)\right) + B\left(\hat{f}\left(\underline{x}_i\right)\right)^2 - 2\operatorname{cov}\left(y_i, \hat{f}\left(\underline{x}_i\right)\right) \\
&= \sigma^2 + MSE\left(\hat{f}\left(\underline{x}_i\right)\right) - 2\operatorname{cov}\left(y_i, \hat{f}\left(\underline{x}_i\right)\right).
\end{aligned}$$

Thus,

$$\begin{aligned}\omega &= E\left[PE_{in}\left(\hat{f}\left(\underline{x}\right)\right)\right] - E\left[TE\left(\hat{f}\left(\underline{x}\right)\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E_{\mathbf{y}}E_{y_0}\left(y_{0,i} - \hat{f}\left(\underline{x}_i\right)\right)^2 - \frac{1}{n}\sum_{i=1}^{n} E_{\mathbf{y}}\left(y_i - \hat{f}\left(\underline{x}_i\right)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\sigma^2 + MSE\left(\hat{f}\left(\underline{x}_i\right)\right)\right] \\ &- \frac{1}{n}\sum_{i=1}^{n}\left[\sigma^2 + MSE\left(\hat{f}\left(\underline{x}_i\right)\right) - 2\operatorname{cov}\left(y_i, \hat{f}\left(\underline{x}_i\right)\right)\right] \\ &= \frac{2}{n}\sum_{i=1}^{n}\operatorname{cov}\left(y_i, \hat{f}\left(\underline{x}_i\right)\right).\end{aligned}$$

B.2 Overfitting and Underfitting

B.2.1 Overfitting

Estimation

Suppose that the true model includes only the predictors $X_1, X_2, \ldots, X_d$, so that $\underline{\beta}^T = \left(\underline{\beta}_{\mathcal{D}}^T, \underline{0}^T\right)$. Let $\mathbf{X} = (\mathbf{X}_{\mathcal{D}}, \mathbf{X}_{\mathcal{D}^c})$, where $\mathcal{D} = \{j : 0, 1, \ldots, d\}$ and $\mathcal{D}^c = \{j : d+1, \ldots, p\}$ so that $\mathbf{X}_{\mathcal{D}}$ is the first $d+1$ columns of $\mathbf{X}$ (d predictors plus the intercept) and $\mathbf{X}_{\mathcal{D}^c}$ is the last $p-d$ columns of $\mathbf{X}$. Assume that both $\mathbf{X}_{\mathcal{D}}$ and $\mathbf{X}_{\mathcal{D}^c}$ have full column rank. Similarly, partition $\underline{\beta}^T = \left(\underline{\beta}_{\mathcal{D}}^T, \underline{\beta}_{\mathcal{D}^c}^T\right)$. Thus the true model is given by

$$E\left(\mathbf{y}\right) = f^{true}\left(\mathbf{X}\right) = \mathbf{X}_{\mathcal{D}}\underline{\beta}_{\mathcal{D}}. \tag{B.2.1}$$

Then the estimate of the true model $\left(\hat{\underline{\beta}}^{true}\right)^T = \left(\left(\hat{\underline{\beta}}_{\mathcal{D}}^{true}\right)^T, \underline{0}^T\right)$ has MSE

$$\begin{aligned}MSE\left(\hat{\underline{\beta}}^{true}\right) &= \operatorname{tr}\left[\operatorname{var}\left(\hat{\underline{\beta}}^{true}\right)\right] + \left\|E\left(\hat{\underline{\beta}}^{true}\right) - \underline{\beta}\right\|^2 \\ &= \operatorname{tr}\left[\sigma^2\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\right] + \left\|\left(\underline{\beta}_{\mathcal{D}}^T, \underline{0}^T\right) - \left(\underline{\beta}_{\mathcal{D}}^T, \underline{0}^T\right)\right\|^2 \\ &= \sigma^2\operatorname{tr}\left[\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\right].\end{aligned}$$

The following is an examination of results mentioned in Seber & Lee (2003:230-231). Suppose we overfit the model by including the predictors $X_1, X_2, \ldots, X_p$,

$$\hat{f}\left(\mathbf{X}\right) = \mathbf{X}\hat{\underline{\beta}} = \mathbf{X}_{\mathcal{D}}\hat{\underline{\beta}}_{\mathcal{D}} + \mathbf{X}_{\mathcal{D}^c}\hat{\underline{\beta}}_{\mathcal{D}^c}. \tag{B.2.2}$$

We have

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} \mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}} & \mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}^c} \\ \mathbf{X}_{\mathcal{D}^c}^T\mathbf{X}_{\mathcal{D}} & \mathbf{X}_{\mathcal{D}^c}^T\mathbf{X}_{\mathcal{D}^c} \end{bmatrix},$$

and we can invert this matrix using Theorem A.1.3 to obtain

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{bmatrix} \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1} + \mathbf{B}\mathbf{M}^{-1}\mathbf{B}^T & -\mathbf{B}\mathbf{M}^{-1} \\ -\mathbf{M}^{-1}\mathbf{B}^T & \mathbf{M}^{-1} \end{bmatrix},$$

where

$$\mathbf{B} = \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}^c} \tag{B.2.3}$$

and

$$\begin{aligned} \mathbf{M} &= \mathbf{X}_{\mathcal{D}^c}^T\mathbf{X}_{\mathcal{D}^c} - \mathbf{X}_{\mathcal{D}^c}^T\left\{\mathbf{X}_{\mathcal{D}}\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\mathbf{X}_{\mathcal{D}}^T\right\}\mathbf{X}_{\mathcal{D}^c} \\ &= \mathbf{X}_{\mathcal{D}^c}^T\mathbf{X}_{\mathcal{D}^c} - \mathbf{X}_{\mathcal{D}^c}^T\mathbf{H}_{\mathcal{D}}\mathbf{X}_{\mathcal{D}^c} \\ &= \mathbf{X}_{\mathcal{D}^c}^T\left(\mathbf{I} - \mathbf{H}_{\mathcal{D}}\right)\mathbf{X}_{\mathcal{D}^c} \\ &= \mathbf{E}^T\mathbf{E}, \end{aligned}$$

where

$$\mathbf{H}_{\mathcal{D}} = \mathbf{X}_{\mathcal{D}}\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\mathbf{X}_{\mathcal{D}}^T$$

and

$$\mathbf{E} = \left(\mathbf{I} - \mathbf{H}_{\mathcal{D}}\right)\mathbf{X}_{\mathcal{D}^c}.$$

(Notice that $\mathbf{B}$ is the estimate obtained when regressing $\mathbf{X}_{\mathcal{D}^c}$ on $\mathbf{X}_{\mathcal{D}}$, $\mathbf{E}$ is the residual matrix and $\mathbf{M}$ is the minimum RSS from that regression). Thus, the LSEs are

$$\begin{aligned} \underline{\hat{\beta}} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\ &= \begin{bmatrix} \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1} + \mathbf{B}\mathbf{M}^{-1}\mathbf{B}^T & -\mathbf{B}\mathbf{M}^{-1} \\ -\mathbf{M}^{-1}\mathbf{B}^T & \mathbf{M}^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{X}_{\mathcal{D}}^T\mathbf{y} \\ \mathbf{X}_{\mathcal{D}^c}^T\mathbf{y} \end{bmatrix}. \end{aligned}$$

That is,

$$\begin{aligned}
\hat{\underline{\beta}}_{\mathcal{D}^c} &= \mathbf{M}^{-1}\mathbf{X}_{\mathcal{D}^c}^T\mathbf{y} - \mathbf{M}^{-1}\mathbf{B}^T\mathbf{X}_{\mathcal{D}}^T\mathbf{y} && \text{(B.2.4)}\\
&= \mathbf{M}^{-1}\mathbf{X}_{\mathcal{D}^c}^T\mathbf{y} - \mathbf{M}^{-1}\mathbf{X}_{\mathcal{D}^c}^T\left\{\mathbf{X}_{\mathcal{D}}\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\mathbf{X}_{\mathcal{D}}^T\right\}\mathbf{y}\\
&= \mathbf{M}^{-1}\mathbf{X}_{\mathcal{D}^c}^T\mathbf{y} - \mathbf{M}^{-1}\mathbf{X}_{\mathcal{D}^c}^T\mathbf{H}_{\mathcal{D}}\mathbf{y}\\
&= \mathbf{M}^{-1}\mathbf{X}_{\mathcal{D}^c}^T\left(\mathbf{I} - \mathbf{H}_{\mathcal{D}}\right)\mathbf{y} && \text{(B.2.5)}\\
&= \left(\mathbf{E}^T\mathbf{E}\right)^{-1}\mathbf{E}^T\mathbf{y}
\end{aligned}$$

and

$$\begin{aligned}
\hat{\underline{\beta}}_{\mathcal{D}} &= \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{y} + \mathbf{B}\mathbf{M}^{-1}\mathbf{B}^T\mathbf{X}_{\mathcal{D}}^T\mathbf{y} - \mathbf{B}\mathbf{M}^{-1}\mathbf{X}_{\mathcal{D}^c}^T\mathbf{y}\\
&= \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{y} - \mathbf{B}\left(\mathbf{M}^{-1}\mathbf{X}_{\mathcal{D}^c}^T\mathbf{y} - \mathbf{M}^{-1}\mathbf{B}^T\mathbf{X}_{\mathcal{D}}^T\mathbf{y}\right)\\
&= \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{y} - \mathbf{B}\hat{\underline{\beta}}_{\mathcal{D}^c} \text{ from (B.2.4)} && \text{(B.2.6)}\\
&= \hat{\underline{\beta}}_{-1}^{true} - \mathbf{B}\hat{\underline{\beta}}_{\mathcal{D}^c}.
\end{aligned}$$

The expected values of the estimates are

$$\begin{aligned}
E\left(\hat{\underline{\beta}}_{\mathcal{D}^c}\right) &= E\left[\mathbf{M}^{-1}\mathbf{X}_{\mathcal{D}^c}^T\left(\mathbf{I} - \mathbf{H}_{\mathcal{D}}\right)\mathbf{y}\right]\\
&= \mathbf{M}^{-1}\mathbf{X}_{\mathcal{D}^c}^T\left(\mathbf{I} - \mathbf{H}_{\mathcal{D}}\right)E\left(\mathbf{y}\right)\\
&= \mathbf{M}^{-1}\mathbf{X}_{\mathcal{D}^c}^T\left(\mathbf{I} - \mathbf{H}_{\mathcal{D}}\right)\mathbf{X}_{\mathcal{D}}\underline{\beta}_{\mathcal{D}} \text{ from (B.2.1)}\\
&= \underline{0} \text{ since } \left(\mathbf{I} - \mathbf{H}_{\mathcal{D}}\right)\mathbf{X}_{\mathcal{D}} = \mathbf{0} && \text{(B.2.7)}
\end{aligned}$$

and

$$\begin{aligned}
E\left(\hat{\underline{\beta}}_{\mathcal{D}}\right) &= E\left(\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{y} - \mathbf{B}\hat{\underline{\beta}}_{\mathcal{D}^c}\right)\\
&= \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\mathbf{X}_{\mathcal{D}}^TE\left(\mathbf{y}\right) - \mathbf{B}E\left(\hat{\underline{\beta}}_{\mathcal{D}^c}\right)\\
&= \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\underline{\beta}_{\mathcal{D}} - \underline{0} \text{ from (B.2.1) and (B.2.7)}\\
&= \underline{\beta}_{\mathcal{D}}, && \text{(B.2.8)}
\end{aligned}$$

and the variance-covariance matrix of the estimates is

$$\begin{aligned}\text{var}\left(\hat{\underline{\beta}}\right) &= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\ &= \sigma^2\begin{bmatrix}\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}+\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^T & -\mathbf{B}\mathbf{M}^{-1} \\ -\mathbf{M}^{-1}\mathbf{B}^T & \mathbf{M}^{-1}\end{bmatrix}.\end{aligned} \tag{B.2.9}$$

So the MSE of $\hat{\underline{\beta}}$ is

$$\begin{aligned}MSE\left(\hat{\underline{\beta}}\right) &= \text{tr}\left[\text{var}\left(\hat{\underline{\beta}}\right)\right]+\left\|E\left(\hat{\underline{\beta}}\right)-\underline{\beta}\right\|^{\mathcal{D}^c} \\ &= \text{tr}\left[\sigma^{\mathcal{D}^c}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]+\left\|\left(\underline{\beta}_{\mathcal{D}},\underline{0}\right)^T-\left(\underline{\beta}_{\mathcal{D}},\underline{0}\right)^T\right\|^{\mathcal{D}^c} \\ &= \sigma^2\left(\text{tr}\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\right)+\text{tr}\left(\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^T\right)+\text{tr}\left(\mathbf{M}^{-1}\right) \\ &= \text{var}\left(\hat{\underline{\beta}}^{true}\right)+ \\ &\sigma^2\,\text{tr}\left[\mathbf{X}_{\mathcal{D}^c}^T\left(\mathbf{X}_{\mathcal{D}}\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-\mathcal{D}^c}\mathbf{X}_{\mathcal{D}}^T\right)\mathbf{X}_{\mathcal{D}^c}\left(\mathbf{X}_{\mathcal{D}^c}^T\left(\mathbf{I}-\mathbf{H}_{\mathcal{D}}\right)\mathbf{X}_{\mathcal{D}^c}\right)^{-1}\right]+ \\ &\sigma^2\,\text{tr}\left[\left(\mathbf{X}_{\mathcal{D}^c}^T\left(\mathbf{I}-\mathbf{H}_{\mathcal{D}}\right)\mathbf{X}_{\mathcal{D}^c}\right)^{-1}\right] \\ &> MSE\left(\hat{\underline{\beta}}^{true}\right).\end{aligned} \tag{B.2.10}$$

Prediction

The prediction of a new response y_0 at $\underline{x}_0 = \left(\underline{x}_{0,\mathcal{D}}, \underline{x}_{0,\mathcal{D}^c}\right)$ using the overfitted model is

$$\hat{f}\left(\underline{x}_0\right) = \underline{x}_0^T\hat{\underline{\beta}} = \underline{x}_{0,\mathcal{D}}^T\hat{\underline{\beta}}_{\mathcal{D}} + \underline{x}_{0,\mathcal{D}^c}^T\hat{\underline{\beta}}_{\mathcal{D}^c}$$

with

$$\begin{aligned}E\left(\hat{f}\left(\underline{x}_0\right)\right) &= E\left(\underline{x}_0^T\hat{\underline{\beta}}\right) \\ &= \underline{x}_{0,\mathcal{D}}^T E\left(\hat{\underline{\beta}}_{\mathcal{D}}\right)+\underline{x}_{0,\mathcal{D}^c}^T E\left(\hat{\underline{\beta}}_{\mathcal{D}^c}\right) \\ &= \underline{x}_{0,\mathcal{D}}^T\underline{\beta}_{\mathcal{D}} \text{ from (B.2.8) and (B.2.7)},\end{aligned}$$

$$
\begin{aligned}
B\left(\hat{f}\left(\underline{x}_0\right)\right) &= E\left(\hat{f}\left(\underline{x}_0^T\right)\right) - f^{true}\left(\underline{x}_0^T\right) \\
&= \underline{x}_{0,\mathcal{D}}^T \underline{\beta}_{\mathcal{D}} - \underline{x}_{0,\mathcal{D}}^T \underline{\beta}_{\mathcal{D}} \text{ from (B.2.1)} \\
&= 0 \\
&= B\left(\hat{f}^{true}\left(\underline{x}_0\right)\right)
\end{aligned}
$$

and

$$
\begin{aligned}
&\text{var}\left(\hat{f}\left(\underline{x}_0\right)\right) \\
&= \text{var}\left(\underline{x}_0^T \hat{\underline{\beta}}\right) \\
&= \sigma^2 \underline{x}_0^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \underline{x}_0 \\
&= \sigma^2 \begin{bmatrix} \underline{x}_{0,\mathcal{D}}^T & \underline{x}_{0,\mathcal{D}^c}^T \end{bmatrix} \begin{bmatrix} \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1} + \mathbf{B}\mathbf{M}^{-1}\mathbf{B}^T & -\mathbf{B}\mathbf{M}^{-1} \\ -\mathbf{M}^{-1}\mathbf{B}^T & \mathbf{M}^{-1} \end{bmatrix} \begin{bmatrix} \underline{x}_{0,\mathcal{D}} \\ \underline{x}_{0,\mathcal{D}^c} \end{bmatrix} \text{ from (B.2.9)} \\
&= \sigma^2 \Big[\underline{x}_{0,\mathcal{D}}^T \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1} \underline{x}_{0,\mathcal{D}} + \underline{x}_{0,\mathcal{D}}^T \mathbf{B}\mathbf{M}^{-1}\mathbf{B}^T \underline{x}_{0,\mathcal{D}} \\
&- \underline{x}_{0,\mathcal{D}^c}^T \mathbf{M}^{-1}\mathbf{B}^T \underline{x}_{0,\mathcal{D}} - \underline{x}_{0,\mathcal{D}}^T \mathbf{B}\mathbf{M}^{-1} \underline{x}_{0,\mathcal{D}^c} + \underline{x}_{0,\mathcal{D}^c}^T \mathbf{M}^{-1} \underline{x}_{0,\mathcal{D}^c} \Big] \\
&= \sigma^2 \underline{x}_{0,\mathcal{D}}^T \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1} \underline{x}_{0,\mathcal{D}} + \sigma^2 \left(\mathbf{B}^T \underline{x}_{0,\mathcal{D}} - \underline{x}_{0,\mathcal{D}^c}\right)^T \mathbf{M}^{-1} \left(\mathbf{B}^T \underline{x}_{0,\mathcal{D}} - \underline{x}_{0,\mathcal{D}^c}\right) \\
&= \sigma^2 \underline{x}_{0,\mathcal{D}}^T \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1} \underline{x}_{0,\mathcal{D}} + \sigma^2 \underline{d}^T \mathbf{M}^{-1} \underline{d} \\
&= \text{var}\left(\hat{f}^{true}\left(\underline{x}_0\right)\right) + \sigma^2 \underline{d}^T \mathbf{M}^{-1} \underline{d} \\
&> \text{var}\left(\hat{f}^{true}\left(\underline{x}_0\right)\right),
\end{aligned}
$$

where

$$
\underline{d} = \mathbf{B}^T \underline{x}_{0,\mathcal{D}} - \underline{x}_{0,\mathcal{D}^c}. \tag{B.2.11}
$$

So the expected PE is,

$$
\begin{aligned}
PE\left(\hat{f}\left(\underline{x}_0\right)\right) &= \sigma^2 + MSE\left(\hat{f}\left(\underline{x}_0\right)\right) \\
&= \sigma^2 + \text{var}\left(\hat{f}\left(\underline{x}_0\right)\right) + B\left(\hat{f}\left(\underline{x}_0\right)\right)^2 \\
&= \sigma^2 + \sigma^2 \underline{x}_{0,\mathcal{D}}^T \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1} \underline{x}_{0,\mathcal{D}} + \sigma^2 \underline{d}^T \mathbf{M}^{-1} \underline{d} \qquad \text{(B.2.12)} \\
&= PE\left(\hat{f}^{true}\left(\underline{x}_0\right)\right) + \sigma^2 \underline{d}^T \mathbf{M}^{-1} \underline{d} \\
&> PE\left(\hat{f}^{true}\left(\underline{x}_0\right)\right).
\end{aligned}
$$

B.2.2 Underfitting

Estimation

In contrast to Section B.2.1, suppose that the true model is given by

$$E(\mathbf{y}) = f^{true}(\mathbf{X}) = \mathbf{X}_{\mathcal{D}}\underline{\beta}_{\mathcal{D}} + \mathbf{X}_{\mathcal{D}^c}\underline{\beta}_{\mathcal{D}^c}, \tag{B.2.13}$$

where $\mathcal{D} = \{j : 0, 1, \ldots, d\}$ and $\mathcal{D}^c = \{j : d+1, \ldots, p\}$, and the model is underfitted using only d predictors,

$$\hat{f}(\mathbf{X}) = \mathbf{X}_{\mathcal{D}}\underline{\beta}_{\mathcal{D}}.$$

The LSE is $\underline{\hat{\beta}}^T = \left(\underline{\hat{\beta}}_{\mathcal{D}}^T, \underline{0}^T\right)$ where

$$\underline{\hat{\beta}}_{\mathcal{D}} = \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{y}$$

with

$$\begin{aligned} E\left(\underline{\hat{\beta}}_{\mathcal{D}}\right) &= \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\mathbf{X}_{\mathcal{D}}^T E(\mathbf{y}) \\ &= \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\mathbf{X}_{\mathcal{D}}^T\left(\mathbf{X}_{\mathcal{D}}\underline{\beta}_{\mathcal{D}} + \mathbf{X}_{\mathcal{D}^c}\underline{\beta}_{\mathcal{D}^c}\right) \text{ from (B.2.13)} \\ &= \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\underline{\beta}_{\widehat{\mathcal{D}}} + \left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}^c}\underline{\beta}_{\mathcal{D}^c} \\ &= \underline{\beta}_{\mathcal{D}} + \mathbf{B}\underline{\beta}_{\mathcal{D}^c} \text{ from (B.2.3)} \end{aligned} \tag{B.2.14}$$

and

$$\text{var}\left(\underline{\hat{\beta}}_{\mathcal{D}}\right) = \sigma^2\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}. \tag{B.2.15}$$

Equation (B.2.14) shows that the estimate is biased. Comparing equations (B.2.9) and (B.2.15) shows that $\text{var}\left(\underline{\hat{\beta}}\right) \leqslant \text{var}\left(\underline{\hat{\beta}}^{true}\right)$. The MSE is

$$\begin{aligned} MSE\left(\underline{\hat{\beta}}\right) &= \text{tr}\left[\text{var}\left(\underline{\hat{\beta}}\right)\right] + \left\|E\left(\underline{\hat{\beta}}\right) - \underline{\beta}\right\|^2 \\ &= \text{tr}\left[\sigma^2\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\right] + \left\|\left(\underline{\beta}_{\mathcal{D}} + \mathbf{B}\underline{\beta}_{\mathcal{D}^c}, \underline{\hat{\beta}}\right)^T - \left(\underline{\beta}_{\mathcal{D}}, \underline{\beta}_{\mathcal{D}^c}\right)^T\right\|^2 \\ &= \sigma^2\,\text{tr}\left(\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\right) + \left\|\left(\mathbf{B}\underline{\beta}_{\mathcal{D}^c}, -\underline{\beta}_{\widehat{\mathcal{D}}^c}\right)^T\right\|^2 \\ &= \sigma^2\,\text{tr}\left(\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\right) + \underline{\beta}_{\mathcal{D}^c}^T\left(\mathbf{B}^T\mathbf{B} + \mathbf{I}\right)\underline{\beta}_{\widehat{\mathcal{D}}^c}. \end{aligned}$$

See Draper & Smith (1998:235-241) and Seber & Lee (2003:228-230) for more information.

Prediction

The bias of the prediction at a new observation $\underline{x}_0 = \left(\underline{x}_{0,\mathcal{D}}, \underline{x}_{0,\mathcal{D}^c}\right)$ is

$$\begin{aligned} B\left(\hat{f}\left(\underline{x}_0\right)\right) &= E\left(\hat{f}\left(\underline{x}_0\right)\right) - f^{true}\left(\underline{x}_0\right) \\ &= \underline{x}_{0,\mathcal{D}}^T\left(\underline{\beta}_{\mathcal{D}} + \mathbf{B}\underline{\beta}_{\widehat{\mathcal{D}^c}}\right) - \left(\underline{x}_{0,\mathcal{D}}^T\underline{\beta}_{\mathcal{D}} + \underline{x}_{0,\mathcal{D}^c}^T\underline{\beta}_{\mathcal{D}^c}\right) \\ &= \left(\mathbf{B}^T\underline{x}_{0,\mathcal{D}} - \underline{x}_{0,\mathcal{D}^c}\right)^T\underline{\beta}_{\mathcal{D}^c} \\ &= \underline{d}^T\underline{\beta}_{\mathcal{D}^c} \text{ from (B.2.11)} \end{aligned}$$

and the variance is

$$\begin{aligned} \text{var}\left(\hat{f}\left(\underline{x}_0\right)\right) &= \text{var}\left(\underline{x}_{0,\mathcal{D}}^T\hat{\underline{\beta}}_{\mathcal{D}}\right) \\ &= \sigma^2\underline{x}_{0,\mathcal{D}}^T\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\underline{x}_{0,\mathcal{D}}, \end{aligned}$$

so that the expected PE is

$$\begin{aligned} PE\left(\hat{f}\left(\underline{x}_0\right)\right) &= \sigma^2 + MSE\left(\hat{f}\left(\underline{x}_0\right)\right) \\ &= \sigma^2 + \text{var}\left(\hat{f}\left(\underline{x}_0\right)\right) + B\left(\hat{f}\left(\underline{x}_0\right)\right)^2 \\ &= \sigma^2 + \sigma^2\underline{x}_{0,\mathcal{D}}^T\left(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\right)^{-1}\underline{x}_{0,\mathcal{D}} + \left(\underline{d}^T\underline{\beta}_{\mathcal{D}^c}\right)^2. \end{aligned}$$

B.3 LASSO and LAR

Write $\alpha_j = \alpha_j^+ - \alpha_j^-$, where $\alpha_j^+, \alpha_j^- \geqslant 0$ such that

$$\alpha_j^+ = \begin{cases} \alpha_j \text{ if } \alpha_j > 0 \\ 0 \text{ if } \alpha_j \leqslant 0 \end{cases} \text{ and } \alpha_j^- = \begin{cases} 0 \text{ if } \alpha_j \geqslant 0 \\ -\alpha_j \text{ if } \alpha_j < 0 \end{cases}. \tag{B.3.1}$$

Then the LASSO constraint becomes $\sum|\alpha_j| = \sum\left|\alpha_j^+ - \alpha_j^-\right| = \sum\left(\alpha_j^+ + \alpha_j^-\right)$ and we must solve

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{v} - \mathbf{Z}\underline{\alpha}\|^2 \\ \text{subject to} \quad & \textstyle\sum_j\left(\alpha_j^+ + \alpha_j^-\right) \leqslant t \\ & -\alpha_j^+ \leqslant 0 \text{ for } j = 1, 2, \ldots, p \\ & -\alpha_j^- \leqslant 0 \text{ for } j = 1, 2, \ldots, p \end{aligned}$$

The Lagrangian of the problem is given by

$$l(\underline{\alpha}) = RSS(\underline{\alpha}) + \lambda \sum \left(\alpha_j^+ + \alpha_j^- - t\right) - \sum \lambda_j^+ \alpha_j^+ - \sum \lambda_j^- \alpha_j^-$$

and the KKT optimality conditions are (by Definition A.3.6):

1. $\sum_j \left(\alpha_j^+ + \alpha_j^- - t\right) \leqslant 0, -\alpha_j^+, -\alpha_j^- \leqslant 0 \forall j = 1, 2, \ldots, p$
2. $\lambda, \lambda_j^+, \lambda_j^- \geqslant 0 \forall j = 1, 2, \ldots, p$
3. (a) $\lambda \sum_j \left(\alpha_j^+ + \alpha_j^- - t\right) = 0,$
   (b) $-\lambda_j^+ \alpha_j^+ = 0, j = 1, 2, \ldots, p$
   (c) $-\lambda_j^- \alpha_j^- = 0, j = 1, 2, \ldots, p$
4. (a) $\frac{\partial}{\partial \alpha_j^+} l(\underline{\alpha}) = \nabla RSS(\underline{\alpha})_j + \lambda - \lambda_j^+ = 0$
   (b) $\frac{\partial}{\partial \alpha_j^-} l(\underline{\alpha}) = -\nabla RSS(\underline{\alpha})_j + \lambda - \lambda_j^- = 0$

The conditions imply that

$$\begin{aligned} & \nabla RSS(\underline{\alpha})_j + \lambda - \lambda_j^+ = 0 \quad \text{by cond. 4a} \\ \Leftrightarrow \quad & -\nabla RSS(\underline{\alpha})_j = \lambda - \lambda_j^+ \\ \Leftrightarrow \quad & -\nabla RSS(\underline{\alpha})_j \leqslant \lambda \quad \text{by cond. 2} \end{aligned} \tag{B.3.2}$$

and

$$\begin{aligned} & -\nabla RSS(\underline{\alpha})_j + \lambda - \lambda_j^+ = 0 \quad \text{by cond. 4b} \\ \Leftrightarrow \quad & \nabla RSS(\underline{\alpha})_j = \lambda - \lambda_j^- \\ \Leftrightarrow \quad & \nabla RSS(\underline{\alpha})_j \leqslant \lambda \quad \text{by cond. 2,} \end{aligned} \tag{B.3.3}$$

hence,

$$\left|\nabla RSS(\underline{\alpha})_j\right| \leqslant \lambda.$$

A Suppose that $\lambda = 0$. Then $\nabla RSS(\underline{\alpha})_j = 0 \forall j$.

B Suppose that $\lambda > 0$ and $\alpha_j > 0$. Then

1 $\alpha_j^+ > 0$ by eq. B.3.1

2 $\alpha_j^- = 0$ by eq. B.3.1

3 $\lambda_j^+ = 0$ by cond. 3b and point 2a

4 $\nabla RSS\left(\underline{\alpha}\right)_j = -\lambda < 0$ by eq. B.3.2 and point 2c

C Suppose that $\lambda > 0$ and $\alpha_j < 0$. Then

1 $\alpha_j^+ = 0$ by eq. B.3.1

2 $\alpha_j^- > 0$ by eq. B.3.1

3 $\lambda_j^- = 0$ by cond. 3c and point 3b

4 $\nabla RSS\left(\underline{\alpha}\right)_j = \lambda > 0$ by eq. B.3.3 and point 3c

Therefore, for any active predictor with $\alpha_j \neq 0$, we have that

$$\nabla RSS\left(\underline{\alpha}\right)_j = \begin{cases} -\lambda \text{ if } \alpha_j > 0 \\ \lambda \text{ if } \alpha_j < 0 \end{cases}$$

and since $\nabla RSS\left(\underline{\alpha}\right)_j = -\mathbf{z}_j^T\left(\mathbf{v} - \mathbf{Z}\underline{\alpha}\right)$,

$$\mathbf{z}_j^T\left(\mathbf{v} - \mathbf{Z}\underline{\alpha}\right) = \text{sign}\left(\alpha_j\right)\lambda. \tag{B.3.4}$$

So λ is related to the correlation between the j-th predictor and the residuals by equation (B.3.4).

Suppose that $\mathcal{A}\left(\lambda\right) = \left\{j : \hat{\alpha}_j^L\left(\lambda\right) \neq 0\right\}$ is the active set of variables and $\mathcal{A}\left(\lambda\right)$ does not change on the interval $\lambda \in [\lambda_0, \lambda_1]$. The estimates for the non-active set are zero, $\underline{\hat{\alpha}}_{\mathcal{A}^c}^L\left(\lambda\right) = 0 \; \forall \lambda \in [\lambda_0, \lambda_1]$. Let $\hat{\omega}_j = \text{sign}\left(\hat{\alpha}_j^L\left(\lambda\right)\right)$. For the active set, we have from equation (B.3.4) that

$$\begin{aligned}
& \mathbf{Z}_{\mathcal{A}}^T\left(\mathbf{v} - \mathbf{Z}\underline{\hat{\alpha}}^L\left(\lambda\right)\right) = \underline{\hat{\omega}}_{\mathcal{A}}\lambda \\
\Leftrightarrow \; & \mathbf{Z}_{\mathcal{A}}^T\left(\mathbf{v} - \mathbf{Z}_{\mathcal{A}}\underline{\hat{\alpha}}_{\mathcal{A}}^L\left(\lambda\right)\right) = \underline{\hat{\omega}}_{\mathcal{A}}\lambda \text{ since } \underline{\hat{\alpha}}_{\mathcal{A}^c}^L\left(\lambda\right) = 0 \\
\Leftrightarrow \; & \mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{A}}\underline{\hat{\alpha}}_{\mathcal{A}}^L\left(\lambda\right) = \mathbf{Z}_{\mathcal{A}}^T\mathbf{v} - \underline{\hat{\omega}}_{\mathcal{A}}\lambda \\
\Leftrightarrow \; & \underline{\hat{\alpha}}_{\mathcal{A}}^L\left(\lambda\right) = \left(\mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{A}}\right)^{-1}\left(\mathbf{Z}_{\mathcal{A}}^T\mathbf{v} - \underline{\hat{\omega}}_{\mathcal{A}}\lambda\right) \text{ if } \mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{A}}^T \text{ is positive definite} \\
\Leftrightarrow \; & \underline{\hat{\alpha}}_{\mathcal{A}}\left(\lambda\right) = \underline{\hat{\alpha}}_{\mathcal{A}}^L\left(\lambda\right) + \lambda\underline{\eta},
\end{aligned}$$

where $\underline{\hat{\alpha}}_{\mathcal{A}}(\lambda) = \left(\mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{A}}\right)^{-1}\mathbf{Z}_{\mathcal{A}}^T\mathbf{v}$ is the LSE for the active set and $\underline{\eta} = \left(\mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{A}}\right)^{-1}\underline{\hat{\omega}}_{\mathcal{A}}$. If the active set is unchanged on $\lambda \in [\lambda_0, \lambda_1]$, then the LSEs are identical for all $\lambda \in [\lambda_0, \lambda_1]$,

$$\begin{aligned} & \underline{\hat{\alpha}}_{\mathcal{A}}(\lambda) = \underline{\hat{\alpha}}_{\mathcal{A}}(\lambda_0) \\ \Leftrightarrow \quad & \underline{\hat{\alpha}}^L(\lambda) + \lambda\underline{\eta} = \underline{\hat{\alpha}}^L(\lambda_0) + \lambda_0\underline{\eta} \\ \Leftrightarrow \quad & \underline{\hat{\alpha}}^L(\lambda) = \underline{\hat{\alpha}}^L(\lambda_0) - (\lambda - \lambda_0)\,\underline{\eta}. \end{aligned} \tag{B.3.5}$$

Thus, $\underline{\hat{\alpha}}^L(\lambda)$ is linear as λ ranges from λ_0 to λ_1.

Appendix C

R Packages

C.1 Subset Selection Methods

The linear regression model can be estimated via least squares in R using the `lm` function. The `summary.lm` method includes the measures σ, R^2 and adjusted R^2. The `extractAIC` function can be used to obtain the AIC (equivalent to C_p for linear models) and BIC. There are a number of R functions available for performing subset selection. Some of these functions are described below and summarized in Table C.1.1.

The path of forward selection and backward elimination can be followed manually by applying the `add1` and `drop1` functions, respectively, to an `lm` object and examining the RSS values for each variable. The desired variable can then be added or removed from the model using the `update` function. The `step` function performs these procedures automatically by repeatedly using `add1` or `drop1` and chooses the best model based on either AIC or BIC. Similar functions are available in the `MASS` package, namely `addterm`, `dropterm` and `stepAIC`.

The `regsubsets` function in the `leaps` package also performs forward selection and backward elimination. In addition, it can apply an exhaustive search using a branch and bound algorithm called leaps where it returns the `nbest` models for each size of subsets. Although the function only provides the variable subsets and does not estimate parameters, the estimates and their covariance matrix can be computed using the `coef` and `vcov` functions. The measures RSS, R^2, adjusted R^2, AIC and BIC are provided for each subset.

The `bestglm` package also makes use of the leaps algorithm when fitting linear models, it provides the best subsets of each size and chooses the best model based on some criteria. The selection criteria available for selecting the best model are AIC and various forms of BIC including BIC, BIC_g and BIC_q. In addition, various forms of CV can be used including $LOOCV$, delete-d CV, K-fold CV and adjusted K-fold CV. When K-fold CV is used, the standard errors of the CV estimates are also provided so the 1 SE rule can be used. See McLeod & Xu (2010) for a detailed description of the selection criteria.

The `glmulti` package is also available for subset selection. It can perform an exhaustive search and also includes a genetic algorithm which can handle larger numbers of variables more efficiently. The selection criteria available for choosing the best model are AIC, the small-sample corrected version $AICc$, other variants $QAIC$ and $QAICc$, and BIC. See Calcagno & de Mazancourt (2010) for more details.

The `subselect` package offers subset selection methods by using one of four different algorithms for selecting the best subsets: an adaptation of the leaps algorithm, a simulated annealing algorithm, a genetic algorithm and a modified local search algorithm. A number of coefficients corresponding to test statistics are available as selection criteria, in particular, the Wald statistic is used for linear models. See Cadima *et al.* (2012) for further details.

Package	Function	Methods	Selection Criteria
`stats`	`add1`	Forward selection	AIC
	`drop1`	Backward elimination	BIC
`MASS`	`addterm`	Forward selection	AIC
	`dropterm`	Backward elimination	BIC
`leaps`	`regsubsets`	Forward selection	adjusted R^2
		Backward elimination	AIC
		Leaps algorithm	BIC
`bestglm`	`bestglm`	Forward selection	AIC
		Backward elimination	BIC, BIC_g, BIC_q
		Leaps algorithm	$LOOCV$, delete-d CV
			K-fold CV, adjusted K-fold CV
`glmulti`	`glmulti`	Leaps algorithm	AIC, AIC_c
		Exhaustive screening	$QAIC$, $QAIC_c$
		Genetic algorithm	BIC
`subselect`	`eleaps`	Adapted leaps algorithm	Wald statistic
	`genetic`	Genetic algorithm	
	`anneal`	Simulated annealing algorithm	
	`improve`	Modified local search algorithm	

Table (C.1.1) Subset selection methods in R

C.2 Shrinkage Methods

R packages are also available for shrinkage methods. Some of these functions are described below and summarized in Table C.2.1. The `MASS` package has a function to perform ridge regression, `lm.ridge`. The

tuning parameter is specified by λ and the GCV statistic for each value of λ is output.

The `lasso2` package is based on Osborne *et al.* (2000b). The `l1ce` function fits the ℓ_1 constrained linear model using their algorithm on the LASSO and its dual. The package also contains functions to calculate the deviance (which is RSS for linear models), the GCV and the covariance matrix of the coefficients. The `summary.l1ce` function calculates standard errors of the coefficients as shown in Equation (4.1.34). The tuning parameter is given by t from the constrained problem. Alternatively, $s = t/t_0$ can be used, where t_0 is the ℓ_1 norm of the least squares estimates. This is beneficial when selecting the tuning parameter since the range is closed, $s \in [0,1]$. The entire path of the LASSO can be computed by specifying a sequence from 0 to 1, producing a `l1celist` object and a `plot.l1celist` method is available to plot this path.

The `lars` package fits the entire LASSO solution simultaneously for all values of s via the LAR algorithm by Efron *et al.* (2004). The `lars` function fits the LASSO piecewise linear path without specifying the tuning parameter. The `plot.lars` method can plot the path against either s or the effective degrees of freedom. The `summary.lars` function gives the RSS and C_p statistic at each step in the path. A plot of the C_p statistic can also be drawn with the `plot.lars` method. The package provides a function for performing K-fold CV which can be used to select the optimal tuning parameter, in this case, s. The `cv.lars` function calculates standard errors of the CV estimates so that the 1 SE rule can be used to choose $\hat{s}$. When using a `lars` object to make predictions or estimate coefficients with `predict.lars`, the tuning parameter can be specified by either s, t, or λ.

The `glmpath` function in the `glmpath` package also computes the path of the LASSO and is based on the predictor-corrector algorithm by Park & Hastie (2007). The tuning parameter is not specified when fitting the path, as with the `lars` function. The output of the function provides, for each step of the algorithm, the degrees of freedom, the tuning parameter λ, the deviance, AIC and BIC. The `plot.glmpath` method can plot the coefficient paths, AIC or BIC against either t or λ. Either of these two tuning parameters can be used in the `cv.glmpath` function, to select the optimal value via K-fold CV, and in the `predict.glmpath` function. In `cv.glmpath`, they are specified as a fraction of the maximum (corresponding to least squares) but in `predict.glmpath` you can choose whether to specify them as a fraction or not.

The `enet` function in the `elasticnet` package is based on the LAR algorithm for elastic net (LAR-EN), discussed in Zou & Hastie (2005). The `lambda` argument to the function is the tuning parameter λ_2 corresponding to the ridge penalty, setting `lambda=0` fits the LASSO model, otherwise the EN is fitted. Like the `lars`

function, the tuning parameter for the LASSO penalty is not specified to fit the path. The `plot.enet` method plots the coefficient paths against either s, t, or the LASSO tuning parameter λ_1. Either of these three parameters can be used in `predict.enet` to make predictions or estimate coefficients. The `cv.enet` function can also compute the K-fold CV estimates and standard errors for values of either s, t or λ_1. When fitting the EN, the value of t is still interpreted as the ℓ_1 norm of the coefficients.

The `penalized` package fits combinations of the ℓ_1 and ℓ_2 penalties with arguments `lambda1` and `lambda2` for the tuning parameters, respectively. Therefore it is capable of fitting ridge regression, LASSO and EN models. Altenatively, by specifying `fusedl=TRUE`, the fused LASSO can be fitted where `lambda2` is then used for the penalty on the differences of parameters. The package includes functions for fitting the model, plots, predictions and CV.

The EN model can also be fit using the `glmnet` package. The `glmnet` function fits the path of the EN using the cyclical coordinate descent algorithm by Friedman *et al.* (2010). The LASSO and ridge regression models can also be fit using this function. The `alpha` argument controls which penalty function is used: when `alpha=0` the ridge penalty is used, when `alpha=1` the LASSO penalty is used and the EN penalty is obtained when `alpha` is between 0 and 1. The tuning parameter is given by λ from the Lagrangian form of the problem and it is best to supply a sequence so that prior estimates are ued for the warm start. The `plot.glmnet` method plots the coefficient paths against either t, $\ln(\lambda)$, or the percentage of deviance explained by the model. The package also provides functions to calculate the deviance and to perform K-fold CV. The `cv.glmnet` function includes standard errors for the CV estimates, it also specifies which value of λ corresponds to the minimum CV and the largest value of λ such that the CV lies within one standard error of the minimum. These values are also indicated in the plot produced by the `plot.cv.glmnet` method, which graphs the CV curve along with its standard errors against $\ln(\lambda)$. The `cv.glmnet` function does not validate the `alpha` parameter. To do so, the folds need to be prespecified and used repeatedly in `cv.glmnet` for different values of alpha.

The `ncvreg` package is based on Breheny & Huang (2011) and contains functions to apply the concave penalty functions of SCAD and MCP. The `ncvreg` function fits the path using a coordinate descent algorithm. It can also fit the LASSO model and has an argument `alpha` to include an optional ridge penalty. When `alpha=1` no ridge penalty is added. A ridge regression would correspond to `alpha=0` but this is not supported, `alpha` can be set close to zero. Functions are available for K-fold CV, prediction, estimating coefficents and

plotting. In all functions, the tuning parameter is specified by λ and the second parameter has default 3.7 for SCAD and 3 otherwise. The `plot.ncvreg` method shades the nonconvex region.

The `lqa` package was developed by Ulbricht (2010) as part of his PhD Thesis. The `lqa` function uses a modified LQA algorithm and can be used for a specified value of the tuning parameter λ. Among the methods we have discussed, the following penalty functions can be used: ridge, LASSO, bridge, adaptive LASSO, fused LASSO, EN, OSCAR and SCAD. Other penalty functions available are the weighted fusion and a number of correlation based penalties developed in his Thesis. There is a function corresponding to each penalty function which creates an object of class `penalty` to be used in `lqa`. The `cv.lqa` function can be used to choose the optimal value for the tuning parameter λ by searching over a grid of values. Penalty functions parameterized with multiple tuning parameters are supported, up to 3 tuning parameters can be validated simultaneously. A validation set can be supplied for this purpose, otherwise K-fold CV is performed. A number of loss functions can be used for validation: RSS, AIC, BIC, GCV and deviance (which is just RSS for linear models). The function returns the optimal tuning parameter as well as the `lqa` model using the optimal tuning parameter. A `cv.nng` function is available for the nonnegative garrote model. There is also a `plot.lqa` function which plots the coefficient path against a tuning parameter. For multiple tuning parameters, fixed values must be supplied for all tuning parameters except the one being plotted against. A `pedict.lqa` function is also available for predicting new data.

Other packages for penalized regression include `relaxo` and `relaxnet`, which implement the relaxed LASSO, `sealasso` performs the SEA-LASSO, `genlasso` has functions for the generalized LASSO, `grplasso` is available for the group LASSO and `SGL` for the sparse-group LASSO, `hierNet` implements the hierarchical LASSO and `glinternet` the hierarchical group-LASSO. The PLUS algorithm for MCP is provided in the `plus` package. The `grpreg` package is based on the descent algorithms by Breheny & Huang (2014) and implements a number of group penalties including the group LASSO, group bridge, group MCP, group SCAD and the group exponential lasso. Furthermore, variable screening with SIS can be performed using the `SIS` package. The kappa coefficient and PASS methods, for selecting the tuning parameter for variable selection purposes, can be applied using the `pass` package. The `covTest` package performs the significance test discussed by Lockhart *et al.* (2014). Other packages which can be consulted include `lassoshooting`, `lassogrp` and `parcor`. For Gaussian graphical models, see `glasso` .

Package	Function	Methods	Selection Criteria	Tuning Parameter
MASS	lm.ridge	Ridge regression	GCV	λ
lasso2	l1ce	LASSO	RSS, GCV	s, t
lars	lars	LASSO	RSS, C_p	s
			K-fold CV	
glmpath	glmpath	LASSO	RSS, AIC, BIC	s, t, λ
			K-fold CV	
elasticnet	enet	LASSO	RSS, C_p	s, t, λ
		EN	K-fold CV	
penalized	penalized	Ridge regression	K-fold CV	λ
		LASSO		
		EN		
		Fused LASSO		
glmnet	glmnet	Ridge regression	RSS	λ
		LASSO	K-fold CV	
		EN		
ncvreg	ncvreg	LASSO	RSS	λ
		SCAD	K-fold CV	
		MCP		
lqa	lqa	Ridge regression	K-fold CV	λ
		Nonnegative Garrote	Validation set	
		LASSO	GCV	
		Adaptive LASSO	RSS, AIC, BIC	
		Bridge		
		EN		
		OSCAR		
		SCAD		

Table (C.2.1) Shrinkage methods in R

Bibliography

Antoniadis, A. 1997. Wavelets in statistics: a review. *Journal of the Italian Statistical Society*, *2*:97–130.

Armagan, A. and Zaretzki, R.L. 2010. Model selection via adaptive shrinkage with t priors. *Computational Statistics*, 25(3):441–461.

Bakin, S. 1999. *Adaptive regression and model selection in data mining problems*. PhD Thesis, Australian National University, Canberra, Australia.

Bartlett, P.L., Mendelson, S. and Neeman, J. 2012. L1-regularized linear regression: Persistence and oracle inequalities. *Probability Theory and Related Fields*, 154(1-2):193–244.

Beck, A. and Teboulle, M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

Bickel, P., Ritov, Y. and Tsybakov, A. 2009. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, *37*:1705–1732.

Bien, J., Taylor, J. and Tibshirani, R. 2013. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141.

Bondell, H.D. and Reich, B.J. 2008. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64(1):115–23.

Boyd, S. and Vandenberghe, L. 2004. *Convex optimization*. New York, NY, USA: Cambridge University Press.

Bradley, J.K., Kyrola, A., Bickson, D. and Guestrin, C. 2011. Parallel coordinate descent for L1-regularized loss minimization. In: Getoor, L. and Scheffer, T. (eds.) *ICML 2011: Proceedings of the 28th international conference on machine learning*, 1998:321–328, ACM.

Breheny, P. 2014. The group exponential lasso for bi-level variable selection. *Technical report*, University of Iowa.

Breheny, P. and Huang, J. 2009. Penalized methods for bi-level variable selection. *Statistics and its interface*, 2(3):369–380.

Breheny, P. and Huang, J. 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232–253.

Breheny, P. and Huang, J. 2014. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*:1–23.

Breiman, L. 1995. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Breiman, L., Friedman, J.H., Ohlsen, R.A. and Stone, C.J. 1984. *Classification and regression trees*. Boca Raton, FL, USA: Chapman & Hall/CRC.

Bühlmann, P. and van de Geer, S. 2011. *Statistics for high-dimensional data: Methods, theory and applications*. (Springer Series in Statistics). Heidelberg, Germany: Springer.

Bühlmann, P. and Yu, B. 2006. Sparse boosting. *Journal of Machine Learning Research*, *7*:1001–1024.

Bunea, F., Tsybakov, A. and Wegkamp, M. 2007. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, *1*:169–194.

Burnham, K.P. and Anderson, D.R. 2002. *Model selection and multimodel inference: a practical information-theoretic approach*. 2nd ed. New York, NY, USA: Springer.

Cadima, J., Cerdeira, J.O., Silva, P.D. and Minhoto, M. 2012. *The subselect R package*. An R package vignette available at http://cran.r-project.org/web/packages/subselect/vignettes/subselect.pdf.

Calcagno, V. and de Mazancourt, C. 2010. glmulti: An R package for easy automated model selection with (generalized) linear models. *Journal of Statistical Software*, 34(12):1–29.

Candès, E. and Plan, Y. 2009. Near-ideal model selection by l1 minimization. *The Annals of Statistics*, 37(5):2145–2177.

Casella, G. and Berger, R.L. 2002. *Statistical inference*. 2nd ed. Pacific Grove, CA, USA: Thomson Learning.

Chand, S. 2012. On tuning parameter selection of lasso-type methods - a monte carlo study. In: *IBCAST 2012: Proceedings of the 9th international Bhurban conference on applied sciences & technology*:120–129.

Clarke, B., Fokoué, E. and Zhang, H.H. 2009. *Principles and theory for data mining and machine learning*. (Springer Series in Statistics). New York, NY, USA: Springer.

Daubechies, I. 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on pure and applied Mathematics*, 57(11):1413–1457.

Donoho, D. 1995. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627.

Donoho, D. and Johnstone, J. 1994. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Donoho, D.L. and Johnstone, I.M. 1998. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879–921.

Dossal, C., Kachour, M., Fadili, J., Peyre, G. and Chesneau, C. 2013. The degrees of freedom of the LASSO for general design matrix. *Statistica Sinica*, *23*:809–828.

Draper, N.R. and Smith, H. 1998. *Applied regression analysis*. 3rd ed. (Wiley Series in Probability and Statistics). New York, NY, USA: John Wiley & Sons.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. 2004. Least angle regression. *The Annals of Statistics*, 32(2):407–451.

Efron, B. and Tibshirani, R. 1997. Improvements on cross-validation : The . 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.

Efron, B. and Tibshirani, R.J. 1993. *An introduction to the bootstrap*. (Monographs on Statistics and Applied Probability: 57). New York, NY, USA: Chapman and Hall.

Fan, J. and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. 2008. Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, Y. and Tang, C.Y. 2013. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):531–552.

Fang, Y., Wang, J. and Sun, W. 2013. A note on selection stability: combining stability and prediction. *Technical report*, New York University, University of Illinois and Purdue University.

Faraway, J.J. 2005. *Linear models with R*. (Texts in Statistical Science Series). Boca Raton, FL, USA: Chapman & Hall/CRC.

Frank, I.E. and Friedman, J.H. 1993. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Friedman, J. 2012. Fast sparse regression and classification. *International Journal of Forecasting*, *1*.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. 2007. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T. and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Fu, W.J. 1998. Penalized regressions: The bridge versus the LASSO. *Journal of Computational and Graphical Statistics*, 7(3):397–416.

Gentle, J.E. 2007. *Matrix algebra: Theory, computations, and applications in statistics*. (Springer Texts in Statistics). New York, NY, USA: Springer.

Ghaoui, L., Viallon, V. and Rabbani, T. 2012. Safe feature elimination boosting LASSO and sparse supervised learning problems. *Pacific Journal of Optimization*, 8(4):667–698.

Gong, P. and Zhang, C. 2011. A fast dual projected newton method for L1-regularized least squares. In: Walsh, T. (ed.) *IJCAI 2011: Proceedings of the 22nd international joint conference on artificial intelligence*:1275–1280, IJCAI/AAAI.

Grandvalet, Y. 1998. Least absolute shrinkage is equivalent to quadratic penalization. In: Niklasson, L., Bodén, M. and Ziemke, T. (eds.) *ICANN 1998: Proceedings of the 8th international conference on artificial neural networks*:201–206, Springer.

Greenshtein, E. 2006. Best subset selection, persistence in high-dimensional statistical learning and optimization under L1 constraint. *The Annals of Statistics*, 34(5):2367–2386.

Greenshtein, E. and Ritov, Y. 2004. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988.

Hastie, T., Tibshirani, R. and Friedman, J. 2009. *The elements of statistical learning: Data mining, inference, and prediction*. 2nd ed. (Springer Series in Statistics). New York, NY, USA: Springer.

He, T. 2011. *Lasso and general L1-regularized regression under linear equality and inequality constraints*. Phd, Purdue University, Indiana.

Hoerl, A.E. and Kennard, R.W. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Homrighausen, D. and McDonald, D. 2013. The lasso, persistence, and cross-validation. In: *ICML 2013: Proceedings of the 30th international conference on machine learning*:103–1039.

Homrighausen, D. and Mcdonald, D.J. 2014. Leave-one-out cross-validation is risk consistent for lasso. *Machine Learning*, *97*:65–78.

Huang, J., Breheny, P. and Ma, S. 2012. A selective review of group selection in high-dimensional models. *Statistical Science*, 27(4):481–499.

Huang, J., Breheny, P., Ma, S. and Zhang, C. 2010. The Mnet method for variable selection. Technical Report 402, The University of Iowa.

Huang, J., Ma, S., Li, H. and Zhang, C.H. 2011. The sparse Laplacian shrinkage estimator for high-dimensional regression. *The Annals of Statistics*, 39(4):2021–2046.

Huang, J., Ma, S., Xie, H. and Zhang, C.H. 2009. A group bridge approach for variable selection. *Biometrika*, 96(2):339–355.

Huang, J., Ma, S. and Zhang, C.h. 2008. Adaptive LASSO for sparse high-dimensional regression models. *Statistica Sinica*, 18(4):1603–1618.

James, G., Witten, D., Hastie, T. and Tibshirani, R. 2013. *An introduction to statistical learning: with applications in R*. (Springer Texts in Statistics: 103). New York, NY, USA: Springer.

Jang, W., Lim, J., Lazar, N., Loh, J. and Yu, D. 2013. Regression shrinkage and grouping of highly correlated predictors with HORSES. *arXiv preprint arXiv:1302.0256*.

Johnson, R.A. and Wichern, D.W. 2007. *Applied multivariate statistical analysis*. 6th ed. Upper Saddle River, NJ, USA: Pearson Prentice Hall.

Knight, K. and Fu, W. 2000. Asymptotics for LASSO-type estimators. *The Annals of Statistics*, 28(5):1356–1378.

Kyung, M., Gill, J., Ghosh, M. and Casella, G. 2010. Penalized regression, standard errors, and bayesian LASSOs. *Bayesian Analysis*, 5(2):369–411.

Lawson, C.L. and Hanson, R.J. 1974. *Solving least square problems*. (Prentice-Hall Series in Automatic Computing). Englewood Cliffs, NJ, USA: Prentice-Hall.

Leng, C., Lin, Y. and Wahba, G. 2006. A note on the LASSO and related procedures in model selection. *Statistica Sinica*, 16(4):1273–1284.

Li, W. 2012. *Simultaneous variable selection and outlier detection using LASSO with applications to aircraft landing data analysis*. Phd, Rutgers, The State University of New Jersey, New Jersey.

Lim, M. and Hastie, T. 2014. Learning interactions through hierarchical group-lasso regularization. *arXiv preprint arXiv:1308.2719*:1–35.

Lin, Z., Xiang, Y. and Zhang, C. 2009. Adaptive LASSO in high-dimensional settings. *Journal of Non-parametric Statistics*, 21(6):683–696.

Lockhart, B.R., Taylor, J., Tibshirani, R.J. and Tibshirani, R. 2014. A significance test for the LASSO. *The Annals of Statistics*, 42(2):413–468.

Lykou, A. and Ntzoufras, I. 2012. On bayesian LASSO variable selection and the specification of the shrinkage parameter. *Statistics and Computing*, 23(3):361–390.

Mazumder, R. and Hastie, T. 2012. The graphical LASSO: New insights and alternatives. *Electronic Journal of Statistics*, *6*:2125–2149.

McLeod, A.I. and Xu, C. 2010. *bestglm: Best subset GLM*. An R package vignette available at http://cran.r-project.org/web/packages/bestglm/vignettes/bestglm.pdf.

Meinshausen, N. 2007. Relaxed LASSO. *Computational Statistics & Data Analysis*, 52(1):374–393.

Meinshausen, N. 2009. P-values for high-dimensional regression. *Journal of the American Statistical Association*, *104*:1671–1681.

Meinshausen, N. and Bühlmann, P. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*:417–473.

Meinshausen, N. and Bühlmann, P. 2006. High-dimensional graphs and variable selection with the LASSO. *The Annals of Statistics*, 34(3):1436–1462.

Meinshausen, N. and Yu, B. 2009. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270.

Messer, R. 1994. *Linear algebra: Gateway to mathematics*. Glenview, IL, USA: HarperCollins College Publishers.

Miller, A. 2002. *Subset selection in regression*. 2nd ed. (Monographs on Statistics and Applied Probability: 95). Boca Raton, FL, USA: Chapman & Hall/CRC.

Nocedal, J. and Wright, S.J. 1999. *Numerical optimization*. (Springer Series in Operations Research). New York, NY, USA: Springer.

Osborne, M.R., Presnell, B. and Turlach, B.A. 1998. Knot selection for regression splines via the LASSO. In: *Dimension reduction, computational complexity, and information*:44–49, Interface Foundation of North America.

Osborne, M.R., Presnell, B. and Turlach, B.A. 2000a. A new approach to variable selection in least squares. *IMA Journal of Numerical Analysis*, 20(3):389–403.

Osborne, M.R., Presnell, B. and Turlach, B.A. 2000b. On the LASSO and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337.

Park, M.Y. and Hastie, T. 2007. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677.

Park, T. and Casella, G. 2008. The bayesian LASSO. *Journal of the American Statistical Association*, 103(482):681–686.

Qian, W. and Yang, Y. 2013. Model selection via standard error adjusted adaptive LASSO. *Annals of the Institute of Statistical Mathematics*, 65(2):295–318.

Rao, C.R. 1973. *Linear statistical inference and its applications*. 2nd ed. (Wiley Series in Probability and Statistics). New York, NY, USA: John W.

Roberts, S. and Nowak, G. 2014. Stabilizing the lasso against cross-validation variability. *Computational Statistics & Data Analysis*, *70*:198–211.

Rosset, S. and Zhu, J. 2007. Piecewise linear regularized solution paths. *The Annals of Statistics*, 35(3):1012–1030.

Roth, V. 2004. The generalized LASSO. *IEEE Transactions on Neural Networks*, 15(1):16–28.

Sardy, S., Percival, D. and Bruce, A. 1999. Wavelet shrinkage for unequally spaced data. *Statistics and Computing*, *9*:65–75.

Schmidt, M., Fung, G. and Rosales, R. 2007. Fast optimization methods for L1 regularization: A comparative study and two new approaches. In: Kok, J.N., Koronacki, J., Mantaras, R.L., Matwin, S., Mladenič, D. and Skowron, A. (eds.) *ECML 2007: Proceedings of the 18th European conference on machine learning*:286–297, Springer.

Schmidt, M., Fung, G. and Rosales, R. 2009. Optimization methods for L1-regularization. Technical Report TR-2009-19, University of British Columbia.

Searle, S.R. 1971. *Linear models*. (Wiley Series in Probability and Statistics). New York, NY, USA: John Wiley & Sons.

Seber, G.A.F. and Lee, A.J. 2003. *Linear regression analysis*. 2nd ed. (Wiley Series in Probability and Statistics). Hoboken, NJ, USA: John Wiley & Sons.

Shao, J. 1997. An asymptotic theory for linear model selection. *Statistica Sinica*, *7*:221–264.

Shao, J. 1999. *Mathematical statistics*. (Springer Texts in Statistics). New York, NY, USA: Springer.

Sharma, D.B., Bondell, H.D. and Zhang, H.H. 2013. Consistent group identification and variable selection in regression with correlated predictors. *Journal of Computational and Graphical Statistics*, 22(2):319–340.

She, Y. 2009. Thresholding-based iterative selection procedures for model selection and shrinkage. *Electronic Journal of Statistics*, *3*:384–415.

Siceloff, L.P., Wentworth, G. and Smith, D.E. 1922. *Analytic geometry*. (Wentworth-Smith Mathematical Series). Boston, MA, USA: Ginn and Company.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. 2013. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Spanos, A. 1989. *Statistical foundations of econometric modelling*. Cambridge University Press.

Strang, G. 2006. *Linear algebra and its applications*. 4th ed. Belmont, CA, USA: Brooks Cole.

Sun, W., Wang, J. and Fang, Y. 2013. Consistent selection of tuning parameters via variable selection stability. *Journal of Machine Learning Research*, *14*:3419–3440.

Sun, X. 1999. *The Lasso and its implementation for neural networks*. Phd, University of Toronto, Canada.

Tibshirani, R. 1996. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288.

Tibshirani, R. 1997. The LASSO method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J. and Tibshirani, R.J. 2012. Strong rules for discarding predictors in LASSO-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. 2005. Sparsity and smoothness via the fused LASSO. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tibshirani, R.J. and Taylor, J. 2011. The solution path of the generalized LASSO. *The Annals of Statistics*, 39(3):1335–1371.

Tibshirani, R.J. and Taylor, J. 2012. Degrees of freedom in LASSO problems. *The Annals of Statistics*, 40(2):1198–1232.

Tseng, P. 2001. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494.

Turlach, B.A., Venables, W.N. and Wright, S.J. 2005. Simultaneous variable selection. *Technometrics*, 47(3):349–363.

Ulbricht, J. 2010. *Variable selection in generalized linear models*. PhD Thesis, Ludwig Maximilian University, Munich, Germany.

Wainwright, M.J. 2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using L1 - constrained quadratic programming (LASSO). *IEEE Transactions on Information Theory*, 55(5):2183–2202.

Wang, H. and Leng, C. 2007. Unified LASSO estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048.

Wang, H., Li, B. and Leng, C. 2009. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683.

Wang, H., Li, G. and Jiang, G. 2007. Robust regression shrinkage and consistent variable selection through the LAD-LASSO. *Journal of Business & Economic Statistics*, 25(3):347–355.

Wang, L., Chen, G. and Li, H. 2007. Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–94.

Witten, D.M., Friedman, J.H. and Simon, N. 2011. New insights and faster computations for the graphical LASSO. *Journal of Computational and Graphical Statistics*, 20(4):892–900.

Wolpert, D. and Macready, W. 1997. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82.

Wu, S., Shen, X. and Geyer, C.J. 2009. Adaptive regularization using the entire solution surface. *Biometrika*, 96(3):513–527.

Wu, T.T. and Lange, K. 2008. Coordinate descent algorithms for LASSO penalized regression. *The Annals of Applied Statistics*, 2(1):224–244.

Yang, X.D. 2011. Statistical methods for variable selection in the context of high-dimensional data: LASSO and extensions. Msc, McMaster University, Canada.

Yuan, M. and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Yuan, M. and Lin, Y. 2007. On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):143–161.

Zhang, C.H. 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, C.H. and Huang, J. 2008. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1597.

Zhao, P., Rocha, G. and Yu, B. 2009. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497.

Zhao, P. and Yu, B. 2006. On model selection consistency of LASSO. *Journal of Machine Learning Research*, *7*:2541–2563.

Zhao, P. and Yu, B. 2007. Stagewise LASSO. *Journal of Machine Learning Research*, *8*:2701–2726.

Zhao, X. 2008. LASSO and its applications. Msc, University of Minnesota, Duluth.

Zou, H. 2006. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Zou, H., Hastie, T. and Tibshirani, R. 2007. On the "degrees of freedom" of the LASSO. *The Annals of Statistics*, 35(5):2173–2192.

Zou, H. and Li, R. 2008. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533.

Zou, H. and Zhang, H.H. 2009. On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4):1733–1751.