APTS High-dimensional Statistics

Rajen D. Shah
r.shah@statslab.cam.ac.uk

July 11, 2019

Contents

1 Introduction **2**
1.1 High-dimensional data . . . 2
1.2 Classical statistics . . . 2
1.3 Maximum likelihood estimation . . . 3
1.4 Shortcomings of classical statistics . . . 4
1.5 Notation . . . 4

2 Ridge regression and kernel machines **5**
2.1 The SVD and PCA . . . 6
2.2 Cross-validation . . . 7
2.3 The kernel trick . . . 9
2.4 The representer theorem . . . 11

3 The Lasso **12**
3.1 The Lasso estimator . . . 12
3.2 Prediction error . . . 13
3.2.1 The event Ω . . . 14
3.3 Estimation error . . . 15
3.4 Subgradients and the KKT conditions . . . 18
3.5 Computation . . . 19
3.6 Extensions of the Lasso . . . 21
3.6.1 Generalised linear models . . . 21
3.6.2 Removing the bias of the Lasso . . . 21
3.6.3 The elastic net . . . 21
3.6.4 Group Lasso . . . 22
3.6.5 Fused Lasso . . . 22

4 Graphical modelling **22**
4.1 Conditional independence graphs 23
4.2 Gaussian graphical models 24
4.2.1 Neighbourhood selection 24
4.2.2 Schur complements and the precision matrix 25
4.2.3 The Graphical Lasso 25

5 High-dimensional inference **26**
5.1 Family-wise error rate control 27
5.2 The False Discovery Rate 29
5.3 Hypothesis testing in high-dimensional regression 30

1 Introduction

1.1 High-dimensional data

Over the last 25 years, the sorts of datasets that statisticians have been challenged to study have changed greatly. Where in the past, we were used to datasets with many observations with a few carefully chosen variables, we are now seeing datasets where the number of variables run into the thousands and even exceed the number of observations. For example, the field of genomics routinely encounters datasets where the number of variables or predictors can be in the tens of thousands or more. Classical statistical methods are often simply not applicable in these "high-dimensional" situations. Designing methods that can cope with these challenging settings has been and continues to be one of the most active areas of research in statistics. Before we dive into the details of these methods, it will be helpful to review some results from classical statistical theory to set the scene for the more modern methods to follow.

1.2 Classical statistics

Consider response–covariate pairs $(Y_i, \mathbf{x}_i) \in \mathbb{R} \times \mathbb{R}^p$, $i = 1, \ldots, n$. A *linear model* for the data assumes that it is generated according to

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is the vector of responses; $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the predictor matrix (or design matrix) with ith row $\mathbf{x}_i^T$; $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ represents random error; and $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ is the unknown vector of coefficients.

When $p \ll n$, a sensible way to estimate $\boldsymbol{\beta}^0$ is by ordinary least squares (OLS). This yields an estimator $\hat{\boldsymbol{\beta}}^{\text{OLS}}$ with

$$\hat{\boldsymbol{\beta}}^{\text{OLS}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \, \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{1.2}$$

provided $\mathbf{X}$ has full column rank.

Under the assumptions that $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ is unbiased and has variance $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. How small is this variance? The Gauss–Markov theorem states that OLS is the best linear unbiased estimator in this setting: for any other estimator $\tilde{\boldsymbol{\beta}}$ that is linear in $\mathbf{Y}$ (so $\tilde{\boldsymbol{\beta}} = \mathbf{A}\mathbf{Y}$ for some fixed matrix $\mathbf{A}$), we have

$$\mathrm{Var}_{\boldsymbol{\beta}^0,\sigma^2}(\tilde{\boldsymbol{\beta}}) - \mathrm{Var}_{\boldsymbol{\beta}^0,\sigma^2}(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}})$$

is positive semi-definite.

1.3 Maximum likelihood estimation

The method of least squares is just one way to construct as estimator. A more general technique is that of maximum likelihood estimation. Here given data $\mathbf{y} \in \mathbb{R}^n$ that we take as a realisation of a random variable $\mathbf{Y}$, we specify its density $f(\mathbf{y};\boldsymbol{\theta})$ up to some unknown vector of parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$, where Θ is the parameter space. The likelihood function is a function of $\boldsymbol{\theta}$ for each fixed $\mathbf{y}$ given by

$$L(\boldsymbol{\theta}) := L(\boldsymbol{\theta};\mathbf{y}) = c(\mathbf{y})f(\mathbf{y};\boldsymbol{\theta}),$$

where $c(\mathbf{y})$ is an arbitrary constant of proportionality. The maximum likelihood estimate of $\boldsymbol{\theta}$ maximises the likelihood, or equivalently it maximises the log-likelihood

$$\ell(\boldsymbol{\theta}) := \ell(\boldsymbol{\theta};\mathbf{y}) = \log f(\mathbf{y};\boldsymbol{\theta}) + \log(c(\mathbf{y})).$$

A very useful quantity in the context of maximum likelihood estimation is the *Fisher information* matrix with jkth $(1 \le j,k \le d)$ entry

$$i_{jk}(\boldsymbol{\theta}) := -\mathbb{E}_{\boldsymbol{\theta}}\left\{\frac{\partial^2}{\partial\theta_j\partial\theta_k}\ell(\boldsymbol{\theta})\right\}.$$

It can be thought of as a measure of how hard it is to estimate $\boldsymbol{\theta}$ when it is the true parameter value. The Cramér–Rao lower bound states that if $\tilde{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}$, then under regularity conditions,

$$\mathrm{Var}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}) - i^{-1}(\boldsymbol{\theta})$$

is positive semi-definite.

A remarkable fact about maximum likelihood estimators (MLEs) is that (under quite general conditions) they are asymptotically normally distributed, asymptotically unbiased and asymptotically achieve the Cramér–Rao lower bound.

Assume that the Fisher information matrix when there are n independent observations, $i^{(n)}(\boldsymbol{\theta})$ (where we have made the dependence on n explicit) satisfies $i^{(n)}(\boldsymbol{\theta})/n \to I(\boldsymbol{\theta})$ for some positive definite matrix I. Then denoting the maximum likelihood estimator of

$\boldsymbol{\theta}$ when there are n observations by $\hat{\boldsymbol{\theta}}^{(n)}$, under regularity conditions, as the number of observations $n \to \infty$ we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}_d(\mathbf{0}, I^{-1}(\boldsymbol{\theta})).$$

Returning to our linear model, if we assume in addition that $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$, then the log-likelihood for $(\boldsymbol{\beta}, \sigma^2)$ is

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2.$$

We see that the maximum likelihood estimate of $\boldsymbol{\beta}$ and OLS coincide. You can check that

$$i(\boldsymbol{\beta}, \sigma^2) = \begin{pmatrix} \sigma^{-2}\mathbf{X}^T\mathbf{X} & 0 \\ 0 & n\sigma^{-4}/2 \end{pmatrix}.$$

The general theory for MLEs would suggest that approximately

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \sim \mathcal{N}_p(0, n\sigma^2(\mathbf{X}^T\mathbf{X})^{-1});$$

in fact it is straightforward to show that this distributional result is exact.

1.4 Shortcomings of classical statistics

We have seen that the classical statistical methods of OLS and maximum likelihood estimation enjoy important optimality properties and come ready with a framework for performing inference. However, these methods do have their limitations.

- A crucial part of the optimality arguments for OLS and MLEs was *unbiasedness*. Do there exist biased methods whose variance is is reduced compared to OLS such that their overall prediction error is lower? Yes!—in fact the use of biased estimators is essential in dealing with settings where the number of parameters to be estimated is large compared to the number of observations.

- The asymptotic statements concerning MLEs may not be relevant in many practical settings. These statements concern what happens as $n \to \infty$ whilst p is kept fixed. In practice we often find that datasets with a large n also have a large p. Thus a more relevant asymptotic regime may be to consider p increasing with n at some rate; in these settings the optimality properties of MLEs will typically fail to hold.

1.5 Notation

Given $A, B \subseteq \{1, \ldots, p\}$, and $\mathbf{x} \in \mathbb{R}^p$, we will write $\mathbf{x}_A$ for the sub-vector of $\mathbf{x}$ formed from those components of $\mathbf{x}$ indexed by A. Similarly, we will write $\mathbf{M}_A$ for the submatrix of $\mathbf{M}$ formed from those columns of $\mathbf{M}$ indexed by A. Further, $\mathbf{M}_{A,B}$ will be the submatrix

of $\mathbf{M}$ formed from columns and rows indexed by A and B respectively. For example, $\mathbf{x}_{\{1,2\}} = (x_1, x_2)^T$, $\mathbf{M}_{\{1,2\}}$ is the matrix formed from the first two columns of $\mathbf{M}$, and $\mathbf{M}_{\{1,2\},\{1,2\}}$ is the top left 2×2 submatrix of $\mathbf{M}$.

In addition, when used in subscripts, we will use $-j$ and $-jk$ to denote $\{1, \ldots, p\} \setminus \{j\} := \{j\}^c$ and $\{1, \ldots, p\} \setminus \{j, k\} := \{j, k\}^c$ respectively. So for example, $\mathbf{M}_{-jk}$ is the submatrix of $\mathbf{M}$ that has columns j and k removed.

2 Ridge regression and kernel machines

Let us revisit the linear model

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta}^0 + \varepsilon_i$$

where $\mathbb{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$. For unbiased estimators of $\boldsymbol{\beta}^0$, their variance gives a way of comparing their quality in terms of squared error loss. For a potentially biased estimator, $\tilde{\boldsymbol{\beta}}$, the relevant quantity is

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\beta}^0,\sigma^2}\{(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T\} &= \mathbb{E}[\{\tilde{\boldsymbol{\beta}} - \mathbb{E}(\tilde{\boldsymbol{\beta}}) + \mathbb{E}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{\beta}^0\}\{\tilde{\boldsymbol{\beta}} - \mathbb{E}(\tilde{\boldsymbol{\beta}}) + \mathbb{E}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{\beta}^0\}^T] \\ &= \mathrm{Var}(\tilde{\boldsymbol{\beta}}) + \{\mathbb{E}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\}\{\mathbb{E}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\}^T,\end{aligned}$$

a sum of squared bias and variance terms. In this section and the next, we will explore two important methods for variance reduction based on different forms of penalisation: rather than forming estimators via optimising a least squares or log-likelihood term, we will introduce an additional penalty term that encourages estimates to be shrunk towards $\mathbf{0}$ in some sense. This will allow us to produce reliable estimators that work well when classical MLEs are infeasible, and in other situations can greatly out-perform the classical approaches.

Ridge regression [Hoerl and Kennard, 1970] performs shrinkage by solving the following optimisation problem

$$(\hat{\mu}_\lambda^{\mathrm{R}}, \hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}}) = \underset{(\mu,\boldsymbol{\beta})\in\mathbb{R}\times\mathbb{R}^p}{\arg\min} \{\|\mathbf{Y} - \mu\mathbf{1} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2\}.$$

Here $\mathbf{1}$ is an n-vector of 1's. We see that the usual OLS objective is penalised by an additional term proportional to $\|\boldsymbol{\beta}\|_2^2$. The parameter $\lambda \geq 0$, which controls the severity of the penalty and therefore the degree of the shrinkage towards $\mathbf{0}$, is known as a *regularisation parameter* or *tuning parameter*. Note we have explicitly included an intercept term which is not penalised. The reason for this is that were the variables to have their origins shifted so e.g. a variable representing temperature is given in units of Kelvin rather than Celsius, the fitted values would not change. However, $\mathbf{X}\hat{\boldsymbol{\beta}}$ is not invariant under scale transformations of the variables so it is standard practice to centre each column of $\mathbf{X}$ (hence making them orthogonal to the intercept term) and then scale them to have ℓ_2-norm $\sqrt{n}$.

The following lemma shows that after this standardisation of $\mathbf{X}$, $\hat{\mu}_\lambda^{\mathrm{R}} = \bar{\mathbf{Y}} := \sum_{i=1}^n Y_i/n$, so we may assume that $\sum_{i=1}^n Y_i = 0$ by replacing Y_i by $Y_i - \bar{\mathbf{Y}}$ and then we can remove μ from our objective function.

Lemma 1. *Suppose the columns of* $\mathbf{X}$ *have been centred. If a minimiser* $(\hat{\mu}, \hat{\boldsymbol{\beta}})$ *of*

$$\|\mathbf{Y} - \mu\mathbf{1} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + J(\boldsymbol{\beta}).$$

over $(\mu, \boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}^p$ *exists, then* $\hat{\mu} = \bar{\mathbf{Y}}$.

Once these modifications have been made, differentiating the resulting objective with respect to $\boldsymbol{\beta}$ yields

$$\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}.$$

In this form, we can see how the addition of the $\lambda\mathbf{I}$ term helps to stabilise the estimator. Note that when $\mathbf{X}$ does not have full column rank (such as in high-dimensional situations), $\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}}$ can still be computed. On the other hand, when $\mathbf{X}$ does have full column rank and assuming a linear model, we have the following theorem.

Theorem 2. *For* λ *sufficiently small (depending on* $\boldsymbol{\beta}^0$ *and* σ^2*),*

$$\mathbb{E}(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{\beta}^0)(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{\beta}^0)^T - \mathbb{E}(\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}} - \boldsymbol{\beta}^0)(\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}} - \boldsymbol{\beta}^0)^T$$

is positive definite.

At first sight this might appear to contradict the Gauss–Markov theorem. However, though ridge regression is a linear estimator, it is not unbiased. In order to better understand the performance of ridge regression, we will turn to one of the key matrix decompositions used in statistics.

2.1 The SVD and PCA

Recall that we can factorise any $\mathbf{X} \in \mathbb{R}^{n\times p}$ into its SVD

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

Here the $\mathbf{U} \in \mathbb{R}^{n\times n}$ and $\mathbf{V} \in \mathbb{R}^{p\times p}$ are orthogonal matrices and $\mathbf{D} \in \mathbb{R}^{n\times p}$ has $D_{11} \geq D_{22} \geq \cdots \geq D_{mm} \geq 0$ where $m := \min(n, p)$ and all other entries of $\mathbf{D}$ are zero.

When $n > p$, we can replace $\mathbf{U}$ by its first p columns and $\mathbf{D}$ by its first p rows to produce another version of the SVD (sometimes known as the thin SVD). Then $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ where $\mathbf{U} \in \mathbb{R}^{n\times p}$ has orthonormal columns (but is no longer square) and $\mathbf{D}$ is square and diagonal. There is an equivalent version for when $p > n$.

Let us take $\mathbf{X} \in \mathbb{R}^{n\times p}$ as our matrix of predictors and suppose $n \geq p$. Using the (thin) SVD we may write the fitted values from ridge regression as follows.

$$\begin{aligned}
\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^T(\mathbf{V}\mathbf{D}^2\mathbf{V}^T + \lambda\mathbf{I})^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{Y} \\
&= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{Y} \\
&= \sum_{j=1}^{p} \mathbf{U}_j \frac{D_{jj}^2}{D_{jj}^2 + \lambda}\mathbf{U}_j^T\mathbf{Y}.
\end{aligned}$$

Here we have used the notation (which we shall use throughout the course) that $\mathbf{U}_j$ is the jth column of $\mathbf{U}$. For comparison, the fitted values from OLS (when $\mathbf{X}$ has full column rank) are

$$\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{U}\mathbf{U}^T\mathbf{Y}.$$

Both OLS and ridge regression compute the coordinates of $\mathbf{Y}$ with respect to the columns of $\mathbf{U}$. Ridge regression then shrinks these coordinates by the factors $D_{jj}^2/(D_{jj}^2+\lambda)$; if D_{jj} is small, the amount of shrinkage will be larger.

To interpret this further, note that the SVD is intimately connected with Principal Component Analysis (PCA). Consider $\mathbf{v} \in \mathbb{R}^p$ with $\|\mathbf{v}\|_2 = 1$. Since the columns of $\mathbf{X}$ have had their means subtracted, the sample variance of $\mathbf{X}\mathbf{v} \in \mathbb{R}^n$, is

$$\frac{1}{n}\mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v} = \frac{1}{n}\mathbf{v}^T\mathbf{V}\mathbf{D}^2\mathbf{V}^T\mathbf{v}.$$

Writing $\mathbf{a} = \mathbf{V}^T\mathbf{v}$, so $\|\mathbf{a}\|_2 = 1$, we have

$$\frac{1}{n}\mathbf{v}^T\mathbf{V}\mathbf{D}^2\mathbf{V}^T\mathbf{v} = \frac{1}{n}\mathbf{a}^T\mathbf{D}^2\mathbf{a} = \frac{1}{n}\sum_j a_j^2 D_{jj}^2 \leq \frac{1}{n}D_{11}\sum_j a_j^2 = \frac{1}{n}D_{11}^2.$$

As $\|\mathbf{X}\mathbf{V}_1\|_2^2/n = D_{11}^2/n$, $\mathbf{V}_1$ determines the linear combination of the columns of $\mathbf{X}$ that has the largest sample variance, when the coefficients of the linear combination are constrained to have ℓ_2-norm 1. $\mathbf{X}\mathbf{V}_1 = D_{11}\mathbf{U}_1$ is known as the first principal component of $\mathbf{X}$. Subsequent principal components $D_{22}\mathbf{U}_2, \ldots, D_{pp}\mathbf{U}_p$ have maximum variance D_{jj}^2/n, subject to being orthogonal to all earlier ones.

Returning to ridge regression, we see that it shrinks $\mathbf{Y}$ most in the smaller principal components of $\mathbf{X}$. Thus it will work well when most of the signal is in the large principal components of $\mathbf{X}$. We now turn to the problem of choosing λ.

2.2 Cross-validation

Cross-validation is a general technique for selecting a good regression method from among several competing regression methods. We illustrate the principle with ridge regression, where we have a family of regression methods given by different λ values.

So far, we have considered the matrix of predictors $\mathbf{X}$ as fixed and non-random. However, in many cases, it makes sense to think of it as random. Let us assume that our data are i.i.d. pairs $(\mathbf{x}_i, \mathbf{Y}_i)$, $i = 1, \ldots, n$. Then ideally, we might want to pick a λ value such that

$$\mathbb{E}\{(Y^* - \mathbf{x}^{*T}\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}}(\mathbf{X}, \mathbf{Y}))^2 | \mathbf{X}, \mathbf{Y}\} \tag{2.1}$$

is minimised. Here $(\mathbf{x}^*, Y^*) \in \mathbb{R}^p \times \mathbb{R}$ is independent of $(\mathbf{X}, \mathbf{Y})$ and has the same distribution as $(\mathbf{x}_1, \mathbf{Y}_1)$, and we have made the dependence of $\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}}$ on the *training data* $(\mathbf{X}, \mathbf{Y})$ explicit. This λ is such that conditional on the original training data, it minimises the expected prediction error on a new observation drawn from the same distribution as the training data.

A less ambitious goal is to find a λ value to minimise the expected prediction error,

$$\mathbb{E}[\mathbb{E}\{(Y^* - \mathbf{x}^{*T}\hat{\boldsymbol{\beta}}^{\mathrm{R}}_{\lambda}(\mathbf{X}, \mathbf{Y}))^2 | \mathbf{X}, \mathbf{Y}\}] \tag{2.2}$$

where compared with (2.1), we have taken a further expectation over the training set.

We still have no way of computing (2.2) directly, but we can attempt to estimate it. The idea of v-fold cross-validation is to split the data into v groups or folds of roughly equal size Let $(\mathbf{X}^{(-k)}, \mathbf{Y}^{(-k)})$ be all the data except that in the kth fold, and let $A_k \subset \{1, \ldots, n\}$ be the observation indices corresponding to the kth fold. For each λ on a grid of values, we compute $\hat{\boldsymbol{\beta}}^{\mathrm{R}}_{\lambda}(\mathbf{X}^{(-k)}, \mathbf{Y}^{(-k)})$: the ridge regression estimate based on all the data except the kth fold. We choose the value of λ that minimises

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{k=1}^{v}\sum_{i \in A_k}\{Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{\mathrm{R}}_{\lambda}(\mathbf{X}^{(-k)}, \mathbf{Y}^{(-k)})\}^2. \tag{2.3}$$

Writing λ_{CV} for the minimiser, our final estimate of $\boldsymbol{\beta}^0$ can then be $\hat{\boldsymbol{\beta}}^{R}_{\lambda_{\mathrm{CV}}}(\mathbf{X}, \mathbf{Y})$.

Note that for each $i \in A_k$,

$$\mathbb{E}\{Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{\mathrm{R}}_{\lambda}(\mathbf{X}^{(-k)}, \mathbf{Y}^{(-k)})\}^2 = \mathbb{E}[\mathbb{E}\{Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{\mathrm{R}}_{\lambda}(\mathbf{X}^{(-k)}, \mathbf{Y}^{(-k)})\}^2 | \mathbf{X}^{(-k)}, \mathbf{Y}^{(-k)}]. \tag{2.4}$$

This is precisely the expected prediction error in (2.2) but with the training data $\mathbf{X}, \mathbf{Y}$ replaced with a training data set of smaller size. If all the folds have the same size, then $\mathrm{CV}(\lambda)$ is an average of n identically distributed quantities, each with expected value as in (2.4). However, the quantities being averaged are not independent as they share the same data.

Thus cross-validation gives a biased estimate of the expected prediction error. The amount of the bias depends on the size of the folds, the case when the $v = n$ giving the least bias—this is known as leave-one-out cross-validation. The quality of the estimate, though, may be worse as the quantities being averaged in (2.3) will be highly positively correlated. Typical choices of v are 5 or 10.

Cross-validation aims to allow us to choose the single best λ (or more generally regression procedure); we could instead aim to find the best weighted combination of regression procedures. Returning to our ridge regression example, suppose λ is restricted to a grid of values $\lambda_1 > \lambda_2 > \cdots > \lambda_L$. We can then minimise

$$\frac{1}{n}\sum_{k=1}^{v}\sum_{i \in A_k}\left\{Y_i - \sum_{l=1}^{L} w_l \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{\mathrm{R}}_{\lambda_l}(\mathbf{X}^{(-k)}, \mathbf{Y}^{(-k)})\right\}^2$$

over $w \in \mathbb{R}^L$ subject to $w_l \geq 0$ for all l. This is a non-negative least-squares optimisation, for which efficient algorithms are available. This sort of idea is known as *stacking* [Wolpert, 1992, Breiman, 1996] and it can often outperform cross-validation.

2.3 The kernel trick

The fitted values from ridge regression are

$$\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{2.5}$$

An alternative way of writing this is suggested by the following

$$\begin{aligned}
\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}) &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{X}^T \\
(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T &= \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1} \\
\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} &= \mathbf{X}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{Y}.
\end{aligned} \tag{2.6}$$

Two remarks are in order:

- Note while $\mathbf{X}^T\mathbf{X}$ is $p \times p$, $\mathbf{X}\mathbf{X}^T$ is $n \times n$. Computing fitted values using (2.5) would require roughly $O(np^2 + p^3)$ operations. If $p \gg n$ this could be extremely costly. However, our alternative formulation would only require roughly $O(n^2p + n^3)$ operations, which could be substantially smaller.

- We see that the fitted values of ridge regression depend only on inner products $\mathbf{K} = \mathbf{X}\mathbf{X}^T$ between observations (note $K_{ij} = \mathbf{x}_i^T\mathbf{x}_j$).

  Now suppose that we believe the signal depends quadratically on the predictors:

$$Y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \sum_{k,l} x_{ik}x_{il}\theta_{kl} + \varepsilon_i.$$

We can still use ridge regression provided we work with an enlarged set of predictors

$$x_{i1}, \ldots, x_{ip}, x_{i1}x_{i1}, \ldots, x_{i1}x_{ip}, x_{i2}x_{i1}, \ldots, x_{i2}x_{ip}, \ldots, x_{ip}x_{ip}. \tag{2.7}$$

This will give us $O(p^2)$ predictors. Our new approach to computing fitted values would therefore have complexity $O(n^2p^2 + n^3)$, which could still be rather costly if p is large.

However, rather than first creating all the additional predictors and then computing the new $\mathbf{K}$ matrix, we can attempt to directly compute $\mathbf{K}$. To this end consider

$$\begin{aligned}
\left(\frac{1}{2} + \mathbf{x}_i^T\mathbf{x}_j\right)^2 - \frac{1}{4} &= \left(\frac{1}{2} + \sum_k x_{ik}x_{jk}\right)^2 - \frac{1}{4} \\
&= \sum_k x_{ik}x_{jk} + \sum_{k,l} x_{ik}x_{il}x_{jk}x_{jl}.
\end{aligned} \tag{2.8}$$

Observe this amounts to an inner product between vectors of the form (2.7). Thus if we set

$$K_{ij} = (1/2 + \mathbf{x}_i^T\mathbf{x}_j)^2 - 1/4$$

and plug this into the formula for the fitted values, it is *exactly* as if we had performed ridge regression on an enlarged set of variables given by (2.7). Now computing $\mathbf{K}$ using

(2.8) would require only p operations per entry, so $O(n^2p)$ operations in total. It thus seems we have improved things by a factor of p using our new approach. This is a nice computational trick, but more importantly for us it serves to illustrate some general points. Since ridge regression only depends on inner products between observations, rather than fitting non-linear models by first mapping the original data $\mathbf{x}_i \in \mathbb{R}^p$ to $\phi(\mathbf{x}_i) \in \mathbb{R}^d$ (say) using some *feature map* ϕ (which could, for example introduce quadratic effects), we can instead try to directly compute $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$.

In fact, rather than thinking in terms of feature maps, we can instead try to think about an appropriate measure of similarity $k(\mathbf{x}_i, \mathbf{x}_j)$ between observations. It turns out that a necessary and sufficient property for such a similarity measure to have in order for it to correspond to an inner product of the form $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ may be formalised in the following definition.

Definition 1. Given a non-empty set $\mathcal{X}$ (the input space), a positive definite *kernel* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric map with the property that given any collection of m potential observation vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathcal{X}$, the matrix $\mathbf{K} \in \mathbb{R}^{m \times m}$ with entries

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$$

is positive semi-definite.

Modelling by directly picking a kernel with this positive-definiteness property is sometimes easier and more natural than thinking in terms of feature maps. Note that the input space $\mathcal{X}$ need not be $\mathbb{R}^d$ and so working with kernels gives us a way of performing regression on complex non-Euclidean data. Below are some examples of popular kernels.

Polynomial kernel. $k(\mathbf{x}, \mathbf{x}') = (a + \mathbf{x}^T\mathbf{x}')^r$.

Gaussian kernel.

$$k(x, x') = \exp\left(- \frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\sigma^2} \right).$$

For $\mathbf{x}$ close to $\mathbf{x}'$ it is large whilst for $\mathbf{x}$ far from $\mathbf{x}'$ the kernel quickly decays towards 0. The additional parameter σ^2 known as the *bandwidth* controls the speed of the decay to zero. Note it is less clear how one might find a corresponding feature map and indeed any feature map that represents this must be infinite dimensional.

Jaccard similarity kernel. Take $\mathcal{X}$ to be the set of all subsets of $\{1, \ldots, p\}$. For $x, x' \in \mathcal{X}$ with $x \cup x' \neq \emptyset$ define

$$k(x, x') = \frac{|x \cap x'|}{|x \cup x'|}$$

and if $x \cup x' = \emptyset$ then set $k(x, x') = 1$.

2.4 The representer theorem

Ridge regression is just one of many procedures that depends only on inner products between observation vectors. Consider mapping our original data $\mathbf{x}_i \in \mathbb{R}^p$ to $\phi(\mathbf{x}_i) \in \mathbb{R}^d$ (e.g. to include quadratic terms) and let $\mathbf{\Phi} \in \mathbb{R}^{n \times d}$ be the matrix with ith row $\phi(\mathbf{x}_i)$. Let positive definite kernel k be given by $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$. We now explain why any method involving optimising an objective of the form

$$c(\mathbf{Y}, \mathbf{\Phi}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2 \tag{2.9}$$

over $\boldsymbol{\beta} \in \mathbb{R}^d$ only depends on k provided it is predictions at new or existing data points that are of interest rather than the minimiser $\hat{\boldsymbol{\beta}}$ itself. Let $\mathbf{P} \in \mathbb{R}^{d \times d}$ be the orthogonal projection on to the row space of $\mathbf{\Phi}$ and note that $\mathbf{\Phi}\boldsymbol{\beta} = \mathbf{\Phi}\mathbf{P}\boldsymbol{\beta}$. Meanwhile

$$\|\boldsymbol{\beta}\|_2^2 = \|\mathbf{P}\boldsymbol{\beta}\|_2^2 + \|(\mathbf{I} - \mathbf{P})\boldsymbol{\beta}\|_2^2.$$

We conclude that any minimiser $\hat{\boldsymbol{\beta}}$ must satisfy $\hat{\boldsymbol{\beta}} = \mathbf{P}\hat{\boldsymbol{\beta}}$, that is $\hat{\boldsymbol{\beta}}$ must be in the row space of $\mathbf{\Phi}$. This means that we may write $\hat{\boldsymbol{\beta}} = \mathbf{\Phi}^T\hat{\boldsymbol{\alpha}}$ for some $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^n$. Let $\mathbf{K} \in \mathbb{R}^{n \times n}$ be the matrix with ijth entry $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, so $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^T$. Substituting $\boldsymbol{\beta} = \mathbf{\Phi}^T\boldsymbol{\alpha}$ into (2.9), we see that $\hat{\boldsymbol{\alpha}}$ minimises

$$c(\mathbf{Y}, \mathbf{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} \tag{2.10}$$

over $\boldsymbol{\alpha} \in \mathbb{R}^n$. Note that the predicted value at a new data point $\mathbf{x}$ is

$$\phi(\mathbf{x})^T\hat{\beta} = \phi(\mathbf{x})^T\mathbf{\Phi}^T\hat{\boldsymbol{\alpha}} = \sum_{i=1}^n k(\mathbf{x}, \mathbf{x}_i)\hat{\alpha}_i.$$

What is remarkable is that whilst the optimisation in (2.9) involves d variables, we have shown this is equivalent to (2.10) which involves n variables: this is a substantial simplification if $d \gg n$. In fact, this result, which is known as the *representer theorem* [Kimeldorf and Wahba, 1970, Schölkopf et al., 2001], can be generalised to the case where the inner product space that ϕ maps is infinite-dimensional.

A key application of the representer theorem is to the famous *support vector classifier* which is commonly used in the classification setting when the response is binary so $\mathbf{Y} \in \{-1, 1\}^n$. The optimisation problem can be expressed as

$$\sum_{i=1}^n (1 - Y_i\mathbf{x}_i^T\boldsymbol{\beta})_+ + \lambda\|\boldsymbol{\beta}\|_2^2,$$

which we see is of the form (2.9). Note here $(x)_+ = x\mathbb{1}_{\{x>0\}}$.

This section has been the briefest of introductions to the world of *kernel machines* which form an important class of machine learning methods. If you are interested in learning more, you can try the survey paper Hofmann et al. [2008], or Scholkopf and Smola [2001] for a more in depth treatment.

3 The Lasso

Let us revisit the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon}$ where $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$, $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. In many modern datasets, there are reasons to believe there are many more variables present than are necessary to explain the response. The MSPE of OLS is

$$\begin{aligned}\frac{1}{n}\mathbb{E}\|\mathbf{X}\boldsymbol{\beta}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}\|_2^2 &= \frac{1}{n}\mathbb{E}\{(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}})^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}})\} \\ &= \frac{1}{n}\mathbb{E}[\mathrm{tr}\{(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}})(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}})^T\mathbf{X}^T\mathbf{X}\}] \\ &= \frac{1}{n}\mathrm{tr}[\mathbb{E}\{(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}})(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}})^T\}\mathbf{X}^T\mathbf{X}] \\ &= \frac{1}{n}\mathrm{tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}})\mathbf{X}^T\mathbf{X}) = \frac{p}{n}\sigma^2.\end{aligned}$$

Let S be the set $S = \{k : \beta_k^0 \neq 0\}$ and suppose $s := |S| \ll p$. If we could identify S and then fit a linear model using just these variables, we'd obtain an MSPE of $\sigma^2 s/n$ which could be substantially smaller than $\sigma^2 p/n$. Furthermore, it can be shown that parameter estimates from the reduced model are more accurate. The smaller model would also be easier to interpret.

We now briefly review some classical model selection strategies.

Best subset regression. A natural approach to finding S is to consider all 2^p possible regression procedures each involving regressing the response on a different sets of explanatory variables $\mathbf{X}_M$ where M is a subset of $\{1, \ldots, p\}$. We can then pick the best regression procedure using cross-validation (say). For general design matrices, this involves an exhaustive search over all subsets, so this is not really feasible for $p > 50$.

Forward selection. This can be seen as a greedy way of performing best subsets regression. Given a target model size m (the tuning parameter), this works as follows.

1. Start by fitting an intercept only model.
2. Add to the current model the predictor variable that reduces the residual sum of squares the most.
3. Continue step 2 until m predictor variables have been selected.

3.1 The Lasso estimator

The ***L**east **a**bsolute **s**hrinkage and **s**election **o**perator (Lasso)* [Tibshirani, 1996] estimates $\boldsymbol{\beta}^0$ by $\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}}$, where $(\hat{\mu}^{\mathrm{L}}, \hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}})$ minimise

$$\frac{1}{2n}\|\mathbf{Y} - \mu\mathbf{1} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{3.1}$$

over $(\mu, \boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}^p$. Here $\|\boldsymbol{\beta}\|_1$ is the ℓ_1-norm of $\boldsymbol{\beta}$: $\|\boldsymbol{\beta}\|_1 = \sum_{k=1}^p |\beta_k|$.

Like ridge regression, $\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}}$ shrinks the OLS estimate towards the origin, but there is an important difference. The ℓ_1 penalty can force some of the estimated coefficients to be exactly 0. In this way the Lasso can perform simultaneous variable selection and parameter estimation. As we did with ridge regression, we can centre and scale the $\mathbf{X}$ matrix, and also centre $\mathbf{Y}$ and thus remove μ from the objective (see Lemma 1). Define

$$Q_\lambda(\boldsymbol{\beta}) = \frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{3.2}$$

Now any minimiser of $Q_\lambda(\boldsymbol{\beta})$ will also be a minimiser of

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \text{ subject to } \|\boldsymbol{\beta}\|_1 \leq \|\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}}\|_1.$$

Similarly, in the case of the ridge regression objective, we know that $\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}}$ minimises $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ subject to $\|\boldsymbol{\beta}\|_2 \leq \|\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}}\|_2$.

The contours of the OLS objective $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ are ellipsoids centred at $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$, while the contours of $\|\boldsymbol{\beta}\|_2^2$ are spheres centred at the origin, and the contours of $\|\boldsymbol{\beta}\|_1$ are 'diamonds' centred at $\mathbf{0}$.

The important point to note is that the ℓ_1 ball $\{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_1 \leq \|\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}}\|_1\}$ has corners where some of the components are zero, and it is likely that the OLS contours will intersect the ℓ_1 ball at such a corner.

3.2 Prediction error

A remarkable property of the Lasso is that even when $p \gg n$, it can still perform well in terms of prediction error. Suppose the columns of $\mathbf{X}$ have been centred and scaled (as we will always assume from now on unless stated otherwise) and a linear model holds:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1}. \tag{3.3}$$

Note we have already centred $\mathbf{Y}$. We will further assume that the ε_i are independent mean-zero sub-Gaussian random variables with common parameter σ. Note this includes as a special case $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$.

Theorem 3. *Let $\hat{\beta}$ be the Lasso solution when*

$$\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}.$$

With probability at least $1 - 2p^{-(A^2/2-1)}$

$$\frac{1}{n}\|\mathbf{X}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}})\|_2^2 \leq 4A\sigma\sqrt{\frac{\log(p)}{n}}\|\boldsymbol{\beta}^0\|_1.$$

Proof. From the definition of $\hat{\boldsymbol{\beta}}$ we have

$$\frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^0\|_2^2 + \lambda\|\boldsymbol{\beta}^0\|_1.$$

Rearranging,

$$\frac{1}{2n}\|\mathbf{X}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}})\|_2^2 \leq \frac{1}{n}\boldsymbol{\varepsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \lambda\|\boldsymbol{\beta}^0\|_1 - \lambda\|\hat{\boldsymbol{\beta}}\|_1.$$

Now $|\boldsymbol{\varepsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)| \leq \|\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1$ by Hölder's inequality. Let $\Omega = \{\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty/n \leq \lambda\}$. Lemma 6 below shows that $\mathbb{P}(\Omega) \geq 1 - 2p^{-(A^2/2-1)}$. Working on the event Ω, we obtain

$$\begin{aligned}
\frac{1}{2n}\|\mathbf{X}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}})\|_2^2 &\leq \lambda\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\|_1 + \lambda\|\boldsymbol{\beta}^0\|_1 - \lambda\|\hat{\boldsymbol{\beta}}\|_1, \\
\frac{1}{n}\|\mathbf{X}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}})\|_2^2 &\leq 4\lambda\|\boldsymbol{\beta}^0\|_1, \qquad \text{by the triangle inequality.}
\end{aligned}$$

□

3.2.1 The event Ω

The proof of Theorem 3 relies on a lower bound for the probability of the event Ω. A union bound gives

$$\begin{aligned}
\mathbb{P}(\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty/n > \lambda) &= \mathbb{P}(\cup_{j=1}^p |\mathbf{X}_j^T\boldsymbol{\varepsilon}|/n > \lambda) \\
&\leq \sum_{j=1}^p \mathbb{P}(|\mathbf{X}_j^T\boldsymbol{\varepsilon}|/n > \lambda).
\end{aligned}$$

Thus if we can obtain tail probability bounds on $|\mathbf{X}_j^T\boldsymbol{\varepsilon}|/n$, the argument above will give a lower bound for $\mathbb{P}(\Omega)$.

Recall a random variable is said to be *sub-Gaussian* with parameter σ if

$$\mathbb{E}e^{\alpha(W - \mathbb{E}W)} \leq e^{\alpha^2\sigma^2/2}$$

for all $\alpha \in \mathbb{R}$ (see Section 6 of the preliminary material). This property implies the tail bound

$$\mathbb{P}(W - \mathbb{E}W \geq t) \leq e^{-t^2/(2\sigma^2)}.$$

If $W \sim \mathcal{N}(0, \sigma^2)$, then W is sub-Gaussian with parameter σ. The sub-Gaussian class also includes bounded random variables:

Lemma 4 (Hoeffding's lemma)**.** *If W is mean-zero and takes values in $[a, b]$, then W is sub-Gaussian with parameter $(b-a)/2$.*

The following proposition shows that analogously to how a linear combination of jointly Gaussian random variables is Gaussian, a linear combination of sub-Gaussian random variables is also sub-Gaussian.

Proposition 5. *Let $(W_i)_{i=1}^n$ be a sequence of independent mean-zero sub-Gaussian random variables with parameters $(\sigma_i)_{i=1}^n$ and let $\boldsymbol{\gamma} \in \mathbb{R}^n$. Then $\boldsymbol{\gamma}^T \mathbf{W}$ is sub-Gaussian with parameter $\left(\sum_i \gamma_i^2 \sigma_i^2\right)^{1/2}$.*

Proof.

$$\begin{aligned}
\mathbb{E} \exp\left(\alpha \sum_{i=1}^n \gamma_i W_i\right) &= \prod_{i=1}^n \mathbb{E} \exp(\alpha \gamma_i W_i) \\
&\leq \prod_{i=1}^n \exp(\alpha^2 \gamma_i^2 \sigma_i^2 / 2) \\
&= \exp\left(\alpha^2 \sum_{i=1}^n \gamma_i^2 \sigma_i^2 / 2\right).
\end{aligned}$$

□

We are now in a position to obtain the probability bound required for Theorem 3.

Lemma 6. *Suppose $(\varepsilon_i)_{i=1}^n$ are independent, mean-zero and sub-Gaussian with common parameter σ. Let $\lambda = A\sigma\sqrt{\log(p)/n}$. Then*

$$\mathbb{P}(\|\mathbf{X}^T \boldsymbol{\varepsilon}\|_\infty / n \leq \lambda) \geq 1 - 2p^{-(A^2/2-1)}.$$

Proof.

$$\mathbb{P}(\|\mathbf{X}^T \boldsymbol{\varepsilon}\|_\infty / n > \lambda) \leq \sum_{j=1}^p \mathbb{P}(|\mathbf{X}_j^T \boldsymbol{\varepsilon}| / n > \lambda).$$

But $\pm \mathbf{X}_j^T \boldsymbol{\varepsilon}/n$ are both sub-Gaussian with parameter $(\sigma^2 \|\mathbf{X}_j\|_2^2 / n^2)^{1/2} = \sigma/\sqrt{n}$. Thus the RHS is at most

$$2p \exp(-A^2 \log(p)/2) = 2p^{1-A^2/2}.$$

□

3.3 Estimation error

Consider once more the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1}$ where the components of $\boldsymbol{\varepsilon}$ are independent mean-zero sub-Gaussian random variables with common parameter σ. Let S and $s = |S|$ be as in the previous section and define $N = S^c = \{1, \ldots, p\} \setminus S$. As we have noted before, in an artificial situation where S is known, we could apply OLS on $\mathbf{X}_S$ and have an MSPE of $\sigma^2 s/n$. Under a so-called *compatibility condition* on the design matrix, we can obtain a similar MSPE for the Lasso.

Definition 2. Given a matrix of predictors $\mathbf{X} \in \mathbb{R}^{n \times p}$ and support set S, define

$$\begin{aligned}\phi^2 &= \inf_{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta}_S \neq 0,\, \|\boldsymbol{\beta}_N\|_1 \leq 3\|\boldsymbol{\beta}_S\|_1} \frac{\frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}\|_2^2}{\frac{1}{s}\|\boldsymbol{\beta}_S\|_1^2} \\ &= \inf_{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}_S\|_1 = 1,\, \|\boldsymbol{\beta}_N\|_1 \leq 3} \frac{s}{n}\|\mathbf{X}_S\boldsymbol{\beta}_S - \mathbf{X}_N\boldsymbol{\beta}_N\|_2^2,\end{aligned}$$

and we take $\phi \geq 0$. The *compatibility condition* is that $\phi^2 > 0$.

Note that if $\mathbf{X}^T\mathbf{X}/n$ has minimum eigenvalue $c_{\min} > 0$ (so necessarily $p \leq n$), then $\phi^2 > c_{\min}$. Indeed by the Cauchy–Schwarz inequality,

$$\|\boldsymbol{\beta}_S\|_1 = \operatorname{sgn}(\boldsymbol{\beta}_S)^T\boldsymbol{\beta}_S \leq \sqrt{s}\|\boldsymbol{\beta}_S\|_2 \leq \sqrt{s}\|\boldsymbol{\beta}\|_2.$$

Thus

$$\phi^2 \geq \inf_{\beta \neq 0} \frac{\frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}\|_2^2}{\|\boldsymbol{\beta}\|_2^2} = c_{\min}.$$

Although in the high-dimensional setting we would have $c_{\min} = 0$, the fact that the infimum in the definition of ϕ^2 is over a restricted set of $\boldsymbol{\beta}$ can still allow ϕ^2 to be positive even in this case. For example, suppose the rows of $\mathbf{X}$ were drawn independently from a mean-zero sun-Gaussian distribution whose variance $\boldsymbol{\Sigma}$ has minimum eigenvalue greater than some constant $c > 0$. In this case it can be shown that $\mathbb{P}(\phi^2 > c/2) \to 1$ if $s\sqrt{\log(p)/n} \to 0$ (in fact stronger results are true for a Gaussian design [Raskutti et al., 2010]).

Theorem 7. *Suppose the compatibility condition holds and let $\hat{\boldsymbol{\beta}}$ be the Lasso solution with $\lambda = A\sigma\sqrt{\log(p)/n}$ for $A > 0$. Then with probability at least $1 - 2p^{-(A^2/8-1)}$, we have*

$$\frac{1}{n}\|\mathbf{X}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}})\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \leq \frac{12\lambda^2 s}{\phi^2} = \frac{12A^2\log(p)}{\phi^2}\frac{\sigma^2 s}{n}.$$

Proof. As in Theorem 3 we start with the "basic inequality":

$$\frac{1}{2n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{1}{n}\boldsymbol{\varepsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \lambda\|\boldsymbol{\beta}^0\|_1.$$

We work on the event $\Omega = \{2\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty/n \leq \lambda\}$ where after applying Hölder's inequality, we get

$$\frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2 + 2\lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 + 2\lambda\|\boldsymbol{\beta}^0\|_1. \tag{3.4}$$

Lemma 6 shows that $\mathbb{P}(\Omega) \geq 1 - 2p^{-(A^2/8-1)}$.

To motivate the rest of the proof, consider the following idea. We know

$$\frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2 \leq 3\lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1.$$

If we could get

$$3\lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \leq \frac{c\lambda}{\sqrt{n}}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2$$

for some constant $c > 0$, then we would have that $\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/n \leq c^2\lambda^2$ and also $3\lambda\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\|_1 \leq c^2\lambda^2$.

Returning to the actual proof, write $a = \|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/(n\lambda)$. Then from (3.4) we can derive the following string of inequalities:

$$\begin{aligned} a + 2(\|\hat{\boldsymbol{\beta}}_N\|_1 + \|\hat{\boldsymbol{\beta}}_S\|_1) &\leq \|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 + \|\hat{\boldsymbol{\beta}}_N\|_1 + 2\|\boldsymbol{\beta}_S^0\|_1 \\ a + \|\hat{\boldsymbol{\beta}}_N\|_1 &\leq \|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 + 2\|\boldsymbol{\beta}_S^0\|_1 - 2\|\hat{\boldsymbol{\beta}}_S\|_1 \\ a + \|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_N^0\|_1 &\leq 3\|\boldsymbol{\beta}_S^0 - \hat{\boldsymbol{\beta}}_S\|_1. \end{aligned}$$

Now, using the compatibility condition with $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} - \beta^0$ we have

$$\begin{aligned} \frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2 + \lambda\|\boldsymbol{\beta}_N^0 - \hat{\boldsymbol{\beta}}_N\|_1 &\leq 3\lambda\|\boldsymbol{\beta}_S^0 - \hat{\boldsymbol{\beta}}_S\|_1 \\ &\leq \frac{3\lambda}{\phi}\sqrt{\frac{s}{n}}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2. \end{aligned}$$

From this we get

$$\frac{1}{\sqrt{n}}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2 \leq \frac{3\lambda\sqrt{s}}{\phi},$$

which substituting into the RHS of the above gives

$$\frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2 + \lambda\|\boldsymbol{\beta}_N^0 - \hat{\boldsymbol{\beta}}_N\|_1 \leq 9\lambda^2 s/\phi^2.$$

Adding λ times the inequality

$$\|\boldsymbol{\beta}_S^0 - \hat{\boldsymbol{\beta}}_S\|_1 \leq 3\lambda s/\phi^2 \tag{3.5}$$

to both sides then gives the result. □

Variable screening. The estimation error results shows in particular that if components of $\boldsymbol{\beta}^0$ are sufficiently far away from 0, the corresponding Lasso coefficients will also be non-zero (and have the same sign). Thus the set of non-zeroes of the Lasso $\hat{S}_\lambda$ should contain the set of important predictors, though we may miss some variables in S that have small coefficients.

Corollary 8. *Consider the setup of Theorem 7. Let* $S_{imp} = \{j : |\beta_j^0| > 3\sigma s A\sqrt{\log(p)/n}/\phi^2\}$. *With probability* $1 - 2p^{-(A^2/8-1)}$

$$\hat{S}_\lambda \supset S_{imp}.$$

Proof. From (3.5) we know that on Ω we have (in particular) for $j \in S_{\text{imp}}$

$$|\beta_j^0 - \hat{\beta}_j| \leq 3\sigma s A\sqrt{\log(p)/n}/\phi^2$$

which implies $\text{sgn}(\hat{\beta}_j) = \text{sgn}(\beta_j^0)$. □

The conditions required for $\hat{S}_\lambda = S$ are somewhat stronger [Meinshausen and Bühlmann, 2006, Zhao and Yu, 2006, Wainwright, 2009].

Note that the choice of λ given in the results above involves σ which is typically unknown. In practice λ is usually selected by cross-validation (or some form of stacking may be used). However Belloni et al. [2011], Sun and Zhang [2012] show that by modifying the Lasso objective to remove the square on the least squares term $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ (known as the *square-root Lasso*), there is a universal choice of λ not depending on σ for which the above results hold.

3.4 Subgradients and the KKT conditions

Efficient computation of Lasso solutions makes use of a set of necessary and sufficient conditions for a vector to minimise the Lasso objective. These conditions are effectively zero gradient conditions, but take account of the fact that the Lasso objective is not differentiable at any point where any $\beta_j = 0$. In order to derive these conditions, we must introduce the notion of a *subgradeient*, which generalises the gradient to potentially non-differentiable but convex functions.

Definition 3. A vector $\mathbf{v} \in \mathbb{R}^d$ is a *subgradient* of a convex function $f : \mathbb{R}^d \to \mathbb{R}$ at $\mathbf{x}$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{v}^T(\mathbf{y} - \mathbf{x}) \qquad \text{for all } \mathbf{y} \in \mathbb{R}^d.$$

The set of subgradients of f at $\mathbf{x}$ is called the *subdifferential* of f at $\mathbf{x}$ and denoted $\partial f(\mathbf{x})$.

In order to make use of subgradients, we will require the following two facts:

Proposition 9. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, and suppose f is differentiable at $\mathbf{x}$. Then $\partial f(\mathbf{x}) = \{\partial f / \partial \mathbf{x}\}$.*

Proposition 10. *Let $f, g : \mathbb{R}^d \to \mathbb{R}$ be convex and let $\alpha > 0$. Then*

$$\partial(\alpha f)(\mathbf{x}) = \alpha \partial f(\mathbf{x}) = \{\alpha \mathbf{v} : \mathbf{v} \in \partial f(\mathbf{x})\},$$
$$\partial(f + g)(\mathbf{x}) = \partial f(\mathbf{x}) + \partial g(\mathbf{x}) = \{\mathbf{v} + \mathbf{w} : \mathbf{v} \in \partial f(\mathbf{x}),\, \mathbf{w} \in \partial g(\mathbf{x})\}.$$

The following easy (but key) result is often referred to in the statistical literature as the Karush–Kuhn–Tucker (KKT) conditions, though it is actually a much simplified version of them.

Proposition 11. $\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ *if and only if* $\mathbf{0} \in \partial f(\mathbf{x}^*)$.

Proof.

$$f(\mathbf{y}) \geq f(\mathbf{x}^*) \quad \text{for all } \mathbf{y} \in \mathbb{R}^d \Leftrightarrow f(\mathbf{y}) \geq f(\mathbf{x}^*) + \mathbf{0}^T(\mathbf{y} - \mathbf{x}) \quad \text{for all } \mathbf{y} \in \mathbb{R}^d$$
$$\Leftrightarrow \mathbf{0} \in \partial f(\mathbf{x}^*).$$

□

Let us now compute the subdifferential of the ℓ_1-norm. First note that $\|\cdot\|_1 : \mathbb{R}^d \to \mathbb{R}$ is convex. Indeed it is a norm so the triangle inequality gives $\|t\mathbf{x} + (1-t)\mathbf{y}\|_1 \leq t\|\mathbf{x}\|_1 + (1-t)\|\mathbf{y}\|_1$.

Proposition 12. *For $\mathbf{x} \in \mathbb{R}^d$ let $A = \{j : x_j \neq 0\}$. Then*

$$\partial\|\mathbf{x}\|_1 = \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_\infty \leq 1 \text{ and } \mathbf{v}_A = \operatorname{sgn}(\mathbf{x}_A)\}$$

Proof. For $j = 1, \ldots, d$, let

$$\begin{aligned} g_j : \mathbb{R}^d &\to \mathbb{R} \\ \mathbf{x} &\mapsto |x_j|. \end{aligned}$$

Then $\|\cdot\| = \sum_j g_j(\cdot)$ so by Proposition 10, $\partial\|\mathbf{x}\|_1 = \sum_j \partial g_j(\mathbf{x})$. When $\mathbf{x}$ has $x_j \neq 0$, g_j is differentiable at $\mathbf{x}$ so by Proposition 9 $\partial g_j(\mathbf{x}) = \{\operatorname{sgn}(x_j)e_j\}$ where e_j is the jth unit vector. When $x_j = 0$, if $\mathbf{v} \in \partial g_j(\mathbf{x})$ we must have

$$g_j(\mathbf{y}) \geq g_j(\mathbf{x}) + \mathbf{v}^T(\mathbf{y} - \mathbf{x}) \qquad \text{for all } \mathbf{y} \in \mathbb{R}^d,$$

so

$$|y_j| \geq \mathbf{v}^T(\mathbf{y} - \mathbf{x}) \qquad \text{for all } \mathbf{y} \in \mathbb{R}^d. \tag{3.6}$$

we claim that the above holds if and only if $v_j \in [-1, 1]$ and $\mathbf{v}_{-j} = 0$. For the 'if' direction, note that $\mathbf{v}^T(\mathbf{y} - \mathbf{x}) = v_j y_j \leq |y_j|$. Conversely, set $\mathbf{y}_{-j} = \mathbf{x}_{-j} + \mathbf{v}_{-j}$ and $y_j = 0$ in (3.6) to get $0 \geq \|\mathbf{v}_{-j}\|_2^2$, so $\mathbf{v}_{-j} = \mathbf{0}$. Then take $\mathbf{y}$ with $\mathbf{y}_{-j} = \mathbf{x}_{-j}$ to get $|y_j| \geq v_j y_j$ for all $y_j \in \mathbb{R}$, so $|v_j| \leq 1$. Forming the set sum of the subdifferentials then gives the result. □

Equipped with these tools from convex analysis, we can now fully characterise the solutions to the Lasso. We have that $\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}}$ is a Lasso solution if and only if $\mathbf{0} \in \partial Q_\lambda(\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}})$, which is equivalent to

$$\frac{1}{n}\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}}) = \lambda\hat{\boldsymbol{\nu}},$$

for $\hat{\boldsymbol{\nu}}$ with $\|\hat{\boldsymbol{\nu}}\|_\infty \leq 1$ and writing $\hat{S}_\lambda = \{k : \hat{\boldsymbol{\beta}}_{\lambda,k}^{\mathrm{L}} \neq 0\}$, $\hat{\boldsymbol{\nu}}_{\hat{S}_\lambda} = \operatorname{sgn}(\hat{\boldsymbol{\beta}}_{\lambda,\hat{S}_\lambda}^{\mathrm{L}})$.

3.5 Computation

One of the most efficient ways of computing Lasso solutions is to use a optimisation technique called *coordinate descent*. This is a quite general way of minimising a function $f : \mathbb{R}^d \to \mathbb{R}$ and works particularly well for functions of the form

$$f(\mathbf{x}) = g(\mathbf{x}) + \sum_{j=1}^{d} h_j(x_j)$$

where g is convex and differentiable and each $h_j : \mathbb{R} \to \mathbb{R}$ is convex (and so continuous). We start with an initial guess of the minimiser $\mathbf{x}^{(0)}$ (e.g. $\mathbf{x}^{(0)} = \mathbf{0}$) and repeat for $m = 1, 2, \ldots$

$$\begin{aligned} x_1^{(m)} &= \arg\min_{x_1 \in \mathbb{R}} f(x_1, x_2^{(m-1)}, \ldots, x_d^{(m-1)}) \\ x_2^{(m)} &= \arg\min_{x_2 \in \mathbb{R}} f(x_1^{(m)}, x_2, x_3^{(m-1)}, \ldots, x_d^{(m-1)}) \\ &\vdots \\ x_d^{(m)} &= \arg\min_{x_d \in \mathbb{R}} f(x_1^{(m)}, x_2^{(m)}, \ldots, x_{d-1}^{(m)}, x_d). \end{aligned}$$

Tseng [2001] proves that provided $A_0 = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^{(0)})\}$ is closed and bounded, then every converging subsequence of $\mathbf{x}^{(m)}$ will converge to a minimiser of f (which must exist since a continuous function on a closed bounded set attains its bounds). In particular this means that $f(\mathbf{x}^{(m)}) \to f(\mathbf{x}^*)$ where $\mathbf{x}^*$ is a minimiser of f. Moreover if $\mathbf{x}^*$ is the unique minimiser of f then $\mathbf{x}^{(m)} \to \mathbf{x}^*$.

Coordinate descent is particularly useful in the case of the Lasso because the coordinate-wise updates have closed-form expressions. Indeed

$$\arg\min_{\beta_j \in \mathbb{R}} \left\{ \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}_{-j}\mathbf{b} - \beta_j \mathbf{X}_j\|_2^2 + \lambda |\beta_j| \right\} = T_\lambda(\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}_{-j}\mathbf{b})/n)$$

where $T_\lambda(x) = \text{sgn}(x)(|x| - \lambda)_+$ is the so-called soft-thresholding operation. This result can easily be verified by checking that the given minimiser does satisfy the KKT conditions of the objective above.

We often want to solve the Lasso on a grid of λ values $\lambda_0 > \cdots > \lambda_L$ (for the purposes of cross-validation for example). To do this, we can first solve for λ_0, and then solve at subsequent grid points by using the solution at the previous grid points as an initial guess (known as a *warm start*). An active set strategy can further speed up computation. This works as follows: For $l = 1, \ldots, L$

1. Initialise $A_l = \{k : \hat{\boldsymbol{\beta}}^{\text{L}}_{\lambda_{l-1},k} \neq 0\}$.

2. Perform coordinate descent only on coordinates in A_l obtaining a solution $\hat{\boldsymbol{\beta}}$ (all components $\hat{\boldsymbol{\beta}}_k$ with $k \notin A_l$ are set to zero).

3. Let $V = \{k : |\mathbf{X}_k^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})|/n > \lambda_l\}$, the set of coordinates which violate the KKT conditions when $\hat{\boldsymbol{\beta}}$ is taken as a candidate solution.

4. If V is empty, we set $\hat{\boldsymbol{\beta}}^{\text{L}}_{\lambda_l} = \hat{\boldsymbol{\beta}}$. Else we update $A_l = A_l \cup V$ and return to 2.

Note that λ_0 may be taken as $\|\mathbf{X}^T\mathbf{Y}/n\|_\infty$ as for λ larger than this, $\hat{\boldsymbol{\beta}}^{\text{L}}_\lambda = \mathbf{0}$ (you can check that $\mathbf{0}$ satisfies the KKT conditions when λ larger than λ_0).

3.6 Extensions of the Lasso

3.6.1 Generalised linear models

We can add an ℓ_1 penalty to many other log-likelihoods besides that arising from the normal linear model. For ℓ_1-penalised generalised linear models, such as logistic regression, similar theoretical results to those we have obtained are available [van de Geer, 2008] and computations can proceed in a similar fashion to above [Friedman et al., 2009].

3.6.2 Removing the bias of the Lasso

One potential drawback of the Lasso is that the same shrinkage effect that sets many estimated coefficients exactly to zero also shrinks all non-zero estimated coefficients towards zero. One possible solution is to take $\hat{S}_\lambda = \{k : \hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda,k} \neq 0\}$ and then re-estimate $\boldsymbol{\beta}^0_{\hat{S}_\lambda}$ by OLS regression on $\mathbf{X}_{\hat{S}_\lambda}$. Another option is to re-estimate using the Lasso on $\mathbf{X}_{\hat{S}_\lambda}$; this procedure is known as the *relaxed Lasso* [Meinshausen, 2007].

The *adaptive Lasso* [Zou, 2006] takes an initial estimate of $\boldsymbol{\beta}^0$, $\hat{\boldsymbol{\beta}}^{\mathrm{init}}$ (e.g. from the Lasso) and then performs weighted Lasso regression:

$$\hat{\boldsymbol{\beta}}^{\mathrm{adapt}}_\lambda = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p : \boldsymbol{\beta}_{\hat{S}^c_{\mathrm{init}}}=0} \left\{ \frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{k\in\hat{S}_{\mathrm{init}}} \frac{|\beta_k|}{|\hat{\beta}^{\mathrm{init}}_k|} \right\},$$

where $\hat{S}_{\mathrm{init}} = \{k : \hat{\boldsymbol{\beta}}^{\mathrm{init}}_k \neq 0\}$.

Yet another approach involves using a family of non-convex penalty functions $p_{\lambda,\gamma} : [0,\infty) \to [0,\infty)$ and attempting to minimise

$$\frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{k=1}^p p_{\lambda,\gamma}(|\beta_k|).$$

A prominent example is the *minimax concave penalty* (MCP) [Zhang, 2010] which takes

$$p'_\lambda(u) = \left(\lambda - \frac{u}{\gamma}\right)_+.$$

One disadvantage of using a non-convex penalty is that there may be multiple local minima which can make optimisation problematic. However, typically if the non-convexity is not too severe, coordinate descent can produce reasonable results.

3.6.3 The elastic net

The *elastic net* [Zou and Hastie, 2005] uses a weighted combination of ridge and Lasso penalties:

$$\hat{\boldsymbol{\beta}}^{\mathrm{EN}}_{\lambda,\alpha} = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\{\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2\} \right\}.$$

Here $\alpha \in [0, 1]$ is an additional tuning parameter. The presence of the ℓ_2 penalty encourages coefficients of variables correlated to other variables with non-zero estimated coefficients to also be non-zero.

3.6.4 Group Lasso

The Lasso penalty encourages the estimated coefficients to be shrunk towards 0 and sometimes exactly to 0. Other penalty functions can be constructed to encourage different types of sparsity. Suppose we have a partition $G_1, \ldots, G_q$ of $\{1, \ldots, p\}$ (so $\cup_{k=1}^{q} G_k = \{1, \ldots, p\}$, $G_j \cap G_k = \emptyset$ for $j \neq k$). The *group Lasso* penalty [Yuan and Lin, 2006] is given by

$$\lambda \sum_{j=1}^{q} m_j \|\boldsymbol{\beta}_{G_j}\|_2.$$

The multipliers $m_j > 0$ serve to balance cases where the groups are of very different sizes; typically we choose $m_j = \sqrt{|G_j|}$. This penalty encourages either an entire group G to have $\hat{\boldsymbol{\beta}}_G = 0$ or $\hat{\beta}_k \neq 0$ for all $k \in G$. Such a property is useful when groups occur through coding for categorical predictors or when expanding predictors using basis functions [Ravikumar et al., 2007].

3.6.5 Fused Lasso

If there is a sense in which the coefficients are ordered, so β_j^0 is expected to be close to β_{j+1}^0, a *fused Lasso* penalty [Tibshirani et al., 2005] may be appropriate. This takes the form

$$\lambda_1 \sum_{j=1}^{p-1} |\beta_j - \beta_{j+1}| + \lambda_2 \|\boldsymbol{\beta}\|_1,$$

where the second term may be omitted depending on whether shrinkage towards 0 is desired. As an example, consider the simple setting where $Y_i = \mu_i^0 + \varepsilon_i$, and it is thought that the $(\mu_i^0)_{i=1}^n$ form a piecewise constant sequence. Then one option is to minimise over $\boldsymbol{\mu} \in \mathbb{R}^n$, the following objective

$$\frac{1}{n}\|\mathbf{Y} - \boldsymbol{\mu}\|_2^2 + \lambda \sum_{i=1}^{n-1} |\mu_i - \mu_{i+1}|.$$

4 Graphical modelling

So far we have considered the problem of relating a particular response to a large collection of explanatory variables. In some settings however, we do not have a distinguished response variable and instead we would like to better understand relationships between all the variables. We may, for example, wish to identify variables that are 'directly related' to each other in some sense. Trying to find pairs of variables that are independent and so

unlikely to be related to each other is not necessarily a good way to proceed as each variable may be correlated with a large number of variables without being directly related to them. A potentially better approach is to use *conditional independence*.

Definition 4. If $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are random vectors with a joint density $f_{\mathbf{XYZ}}$ then we say $\mathbf{X}$ is conditionally independent of $\mathbf{Y}$ given $\mathbf{Z}$, and write

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$$

if

$$f_{\mathbf{XY}|\mathbf{Z}}(\mathbf{x}, \mathbf{y}|\mathbf{z}) = f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{z}) f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z}).$$

Equivalently

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z} \iff f_{\mathbf{X}|\mathbf{YZ}}(\mathbf{x}|\mathbf{y}, \mathbf{z}) = f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{z}) :$$

the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$ and $\mathbf{Z}$ depends only on $\mathbf{Z}$.

4.1 Conditional independence graphs

Graphs provide a convenient way to visualise conditional independencies between random variables. A graph G is a pair (V, E) where V is a set and E is a subset of

$$\{\{x, y\} : x, y \in V,\, x \neq y\},$$

the set of unordered distinct pairs from V. We call the members of V vertices and members of E edges. In the graphs that we consider, we will always take V to be $\{1, \ldots, |V|\}$.

Given $j, k \in V$ a path of length m from j to k is a sequence $j = j_1, j_2, \ldots, j_m = k$ of distinct vertices such that $\{j_i, j_{i+1}\} \in E$, $i = 1, \ldots, m-1$.

Given a triple of subsets of vertices A, B, S, we say S separates A from B if every path from a vertex in A to a vertex in B contains a vertex in S.

Let $\mathbf{Z} = (Z_1, \ldots, Z_p)^T$ be a collection of random variables with joint law P and consider a graph $G = (V, E)$ where $V = \{1, \ldots, p\}$. We say that P satisfies the *pairwise Markov property* w.r.t. G if for any pair $j, k \in V$ with $j \neq k$ and $\{j, k\} \notin E$,

$$Z_j \perp\!\!\!\perp Z_k|\mathbf{Z}_{-jk}.$$

Note that the complete graph that has edges between every pair of vertices will satisfy the pairwise Markov property for any P. The minimal graph satisfying the pairwise Markov property w.r.t. a given P is called the *conditional independence graph (CIG)* for P. This does not have an edge between j and k if and only if $Z_j \perp\!\!\!\perp Z_k|\mathbf{Z}_{-jk}$.

We say P satisfies the *global Markov property* w.r.t. G if for any triple (A, B, S) of disjoint subsets of V such that S separates A from B, we have

$$\mathbf{Z}_A \perp\!\!\!\perp \mathbf{Z}_B|\mathbf{Z}_S.$$

Proposition 13 (See Lauritzen [1996] for example)**.** *If P has a positive density then if it satisfies the pairwise Markov property w.r.t. a graph $\mathcal{G}$, it also satisfies the global Markov property w.r.t. $\mathcal{G}$ and vice versa.*

4.2 Gaussian graphical models

Estimating the CIG given a sample drawn from P is a difficult task in general (indeed see Shah and Peters [2018] for an impossibility result on conditional independence testing). However, in the case where P is multivariate Gaussian, things simplify considerably as we shall see.

The following result on the conditional distributions of a multivariate normal (see the preliminary material for a proof) will be central to our discussion. Let $\mathbf{Z} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ positive definite. Note $\boldsymbol{\Sigma}_{A,A}$ is also positive definite for any A.

Proposition 14.

$$\mathbf{Z}_A|\mathbf{Z}_B = \mathbf{z}_B \sim \mathcal{N}_{|A|}(\boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_{B,B}^{-1}(\mathbf{z}_B - \boldsymbol{\mu}_B),\, \boldsymbol{\Sigma}_{A,A} - \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_{B,B}^{-1}\boldsymbol{\Sigma}_{B,A})$$

4.2.1 Neighbourhood selection

Specialising to the case where $A = \{k\}$ and $B = \{k\}^c$ we see that when conditioning on $\mathbf{Z}_{-k} = \mathbf{z}_{-k}$, we may write

$$Z_k = m_k + \mathbf{z}_{-k}^T \boldsymbol{\Sigma}_{-k,-k}^{-1}\boldsymbol{\Sigma}_{-k,k} + \varepsilon_k,$$

where

$$m_k = \mu_k - \boldsymbol{\Sigma}_{k,-k}\boldsymbol{\Sigma}_{-k,-k}^{-1}\boldsymbol{\mu}_{-k}$$
$$\varepsilon_k|\mathbf{Z}_{-k} = \mathbf{z}_{-k} \sim \mathcal{N}(0,\, \Sigma_{k,k} - \boldsymbol{\Sigma}_{k,-k}\boldsymbol{\Sigma}_{-k,-k}^{-1}\boldsymbol{\Sigma}_{-k,k}).$$

Note that if the jth element of the vector of coefficients $\boldsymbol{\Sigma}_{-k,-k}^{-1}\boldsymbol{\Sigma}_{-k,k}$ is zero, then the distribution of Z_k conditional on $\mathbf{Z}_{-k}$ will not depend at all on the jth component of $\mathbf{Z}_{-k}$. Then if that jth component was $Z_{j'}$, we would have that $Z_k|\mathbf{Z}_{-k} = \mathbf{z}_{-k}$ has the same distribution as $Z_k|\mathbf{Z}_{-j'k} = \mathbf{z}_{-j'k}$, so $Z_k \perp\!\!\!\perp Z_j|\mathbf{Z}_{-j'k}$.

Thus given $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{\text{i.i.d.}}{\sim} \mathbf{Z}$ and writing

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix},$$

we may estimate the coefficient vector $\boldsymbol{\Sigma}_{-k,-k}^{-1}\boldsymbol{\Sigma}_{-k,k}$ by regressing $\mathbf{X}_k$ on $\mathbf{X}_{-k}$ and including an intercept term.

The technique of *neighbourhood selection* [Meinshausen and Bühlmann, 2006] involves performing such a regression for each variable, using the Lasso. There are two options for populating our estimate of the CIG with edges based on the Lasso estimates. Writing $\hat{S}_k$ for the selected set of variables when regressing $\mathbf{X}_k$ on $\mathbf{X}_{-k}$, we can use the "OR" rule and put an edge between vertices j and k if and only if $k \in \hat{S}_j$ or $j \in \hat{S}^k$. An alternative is the "AND" rule where we put an edge between j and k if and only if $k \in \hat{S}_j$ and $j \in \hat{S}_k$.

Another popular approach to estimating the CIG works by estimating the precision matrix $\boldsymbol{\Sigma}^{-1}$.

4.2.2 Schur complements and the precision matrix

The following facts about blockwise inversion of matrices will help us to interpret the mean and variance in Proposition 14.

Proposition 15. *Let* $\mathbf{M} \in \mathbb{R}^{p\times p}$ *be a symmetric positive definite matrix and suppose*

$$\mathbf{M} = \begin{pmatrix} \mathbf{P} & \mathbf{Q}^T \\ \mathbf{Q} & \mathbf{R} \end{pmatrix}$$

with $\mathbf{P}$ *and* $\mathbf{R}$ *square matrices. The* Schur complement *of* $\mathbf{R}$ *is* $\mathbf{P} - \mathbf{Q}^T\mathbf{R}^{-1}\mathbf{Q} =: \mathbf{S}$. *We have that* $\mathbf{S}$ *is positive definite and*

$$\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{S}^{-1} & -\mathbf{S}^{-1}\mathbf{Q}^T\mathbf{R}^{-1} \\ -\mathbf{R}^{-1}\mathbf{Q}\mathbf{S}^{-1} & \mathbf{R}^{-1} + \mathbf{R}^{-1}\mathbf{Q}\mathbf{S}^{-1}\mathbf{Q}^T\mathbf{R}^{-1} \end{pmatrix}.$$

Set $\mathbf{M}$ to be $\mathbf{\Sigma}$ and let $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ be the precision matrix. Taking $P = \Sigma_{11}$ we see that

$$\mathbf{\Omega}_{-1,1} = -\mathbf{\Sigma}_{-1,-1}^{-1}\mathbf{\Sigma}_{-1,1}\Omega_{11}.$$

By symmetry we then also have $\mathbf{\Sigma}_{-k,-k}^{-1}\mathbf{\Sigma}_{-k,k} = -\Omega_{kk}^{-1}\mathbf{\Omega}_{-k,k}$ so

$$(\mathbf{\Sigma}_{-k,-k}^{-1}\mathbf{\Sigma}_{-k,k})_j = 0 \Leftrightarrow \begin{cases} \Omega_{j,k} = 0 \ \text{ for } j < k \\ \Omega_{j+1,k} = 0 \ \text{ for } j \geq k. \end{cases}$$

Thus

$$Z_k \perp\!\!\!\perp Z_j | \mathbf{Z}_{-jk} \Leftrightarrow \Omega_{jk} = 0.$$

This motivates another approach to estimating the CIG.

4.2.3 The Graphical Lasso

Recall that the density of $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ is

$$f(\mathbf{z}) = \frac{1}{(2\pi)^{p/2}\det(\mathbf{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})\right).$$

The log-likelihood of $(\boldsymbol{\mu}, \mathbf{\Sigma})$ based on an i.i.d. sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is

$$\ell(\boldsymbol{\mu}, \mathbf{\Omega}) = \frac{n}{2}\log\det(\mathbf{\Omega}) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\mathbf{\Omega}(\mathbf{x}_i - \boldsymbol{\mu}).$$

Write

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \qquad \mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{X}})(\mathbf{x}_i - \bar{\mathbf{X}})^T.$$

Then

$$\begin{aligned}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Omega}(\mathbf{x}_i-\boldsymbol{\mu}) &= \sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{X}}+\bar{\mathbf{X}}-\boldsymbol{\mu})^T\boldsymbol{\Omega}(\mathbf{x}_i-\bar{\mathbf{X}}+\bar{\mathbf{X}}-\boldsymbol{\mu})\\ &= \sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{X}})^T\boldsymbol{\Omega}(\mathbf{x}_i-\bar{\mathbf{X}})+n(\bar{\mathbf{X}}-\boldsymbol{\mu})^T\boldsymbol{\Omega}(\bar{\mathbf{X}}-\boldsymbol{\mu})\\ &\quad+2\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{X}})^T\boldsymbol{\Omega}(\bar{\mathbf{X}}-\boldsymbol{\mu}).\end{aligned}$$

Also,

$$\begin{aligned}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{X}})^T\boldsymbol{\Omega}(\mathbf{x}_i-\bar{\mathbf{X}}) &= \sum_{i=1}^{n}\mathrm{tr}\{(\mathbf{x}_i-\bar{\mathbf{X}})^T\boldsymbol{\Omega}(\mathbf{x}_i-\bar{\mathbf{X}})\}\\ &= \sum_{i=1}^{n}\mathrm{tr}\{(\mathbf{x}_i-\bar{\mathbf{X}})(\mathbf{x}_i-\bar{\mathbf{X}})^T\boldsymbol{\Omega}\}\\ &= n\mathrm{tr}(\mathbf{S}\boldsymbol{\Omega}).\end{aligned}$$

Thus

$$\ell(\boldsymbol{\mu},\boldsymbol{\Omega})=-\frac{n}{2}\{\mathrm{tr}(\mathbf{S}\boldsymbol{\Omega})-\log\det(\boldsymbol{\Omega})+(\bar{\mathbf{X}}-\boldsymbol{\mu})^T\boldsymbol{\Omega}(\bar{\mathbf{X}}-\boldsymbol{\mu})\}$$

and

$$\max_{\boldsymbol{\mu}\in\mathbb{R}^p}\ell(\boldsymbol{\mu},\boldsymbol{\Omega})=-\frac{n}{2}\{\mathrm{tr}(\mathbf{S}\boldsymbol{\Omega})-\log\det(\boldsymbol{\Omega})\}.$$

Hence the maximum likelihood estimate of Ω can be obtained by solving

$$\min_{\boldsymbol{\Omega}:\boldsymbol{\Omega}\succ\mathbf{0}}\{-\log\det(\boldsymbol{\Omega})+\mathrm{tr}(\mathbf{S}\boldsymbol{\Omega})\},$$

where $\boldsymbol{\Omega}\succ\mathbf{0}$ means $\boldsymbol{\Omega}$ is positive definite. One can show that the objective is convex and we are minimising over a convex set.

The *graphical Lasso* [Yuan and Lin, 2007, Friedman et al., 2008] penalises the log-likelihood for $\boldsymbol{\Omega}$ and solves

$$\min_{\boldsymbol{\Omega}:\boldsymbol{\Omega}\succ\mathbf{0}}\{-\log\det(\boldsymbol{\Omega})+\mathrm{tr}(\mathbf{S}\boldsymbol{\Omega})+\lambda\|\boldsymbol{\Omega}\|_1\},$$

where $\|\boldsymbol{\Omega}\|_1=\sum_{j,k}|\Omega_{jk}|$; this results in a sparse estimate of the precision matrix from which an estimate of the CIG can be constructed.

5 High-dimensional inference

In many modern applications, we may be interested in testing many hypotheses simultaneously. Suppose we are interested in testing null hypotheses $H_1,\ldots,H_m$ and H_i, $i\in I_0$

are the true null hypotheses with $|I_0| = m_0$ (we do not mention the alternative hypotheses explicitly). We will suppose we have available p-values $p_1, \ldots, p_m$ for each of the hypotheses so

$$\mathbb{P}(p_i \leq \alpha) \leq \alpha$$

for all $\alpha \in [0, 1]$, $i \in I_0$. Given a multiple testing procedure takes as input the vector of p-values, and outputs a subset $\mathcal{R} \subseteq \{1, \ldots, m\}$ of rejected hypotheses. Let $N = |\mathcal{R} \cap I_0|$ be the number of falsely rejected hypotheses, and let $R = |\mathcal{R}|$ be the number of rejections.

5.1 Family-wise error rate control

Traditional approaches to multiple testing have sought to control the familywise error rate (FWER) defined by

$$\text{FWER} = \mathbb{P}(N \geq 1)$$

at a prescribed level α; i.e. find procedures for which FWER $\leq \alpha$. The simplest such procedure is the *Bonferroni correction*, which rejects H_i if $p_i \leq \alpha/m$.

Theorem 16. *Using Bonferroni correction,*

$$\mathbb{P}(N \geq 1) \leq \mathbb{E}(N) \leq \frac{m_0 \alpha}{m} \leq \alpha.$$

Proof. The first inequality comes from Markov's inequality. Next

$$\begin{aligned} \mathbb{E}(N) &= \mathbb{E}\left(\sum_{i \in I_0} \mathbb{1}_{\{p_i \leq \alpha/m\}} \right) \\ &= \sum_{i \in I_0} \mathbb{P}(p_i \leq \alpha/m) \\ &\leq \frac{m_0 \alpha}{m}. \end{aligned}$$

□

A more sophisticated approach is the closed testing procedure. Given our family of hypotheses $\{H_i\}_{i=1}^m$, define the *closure* of this family to be

$$\{H_I : I \subseteq \{1, \ldots, m\},\ I \neq \emptyset\}$$

where $H_I = \cap_{i \in I} H_i$ is known as an *intersection hypothesis* (H_I is the hypothesis that all H_i $i \in I$ are true).

Suppose that for each I, we have an α-level test ϕ_I taking values in $\{0, 1\}$ for testing H_I (we reject if $\phi_I = 1$), so under H_I,

$$\mathbb{P}_{H_I}(\phi_I = 1) \leq \alpha.$$

The ϕ_I are known as *local tests*.

The *closed testing procedure* [Marcus et al., 1976] is defined as follows:

$$\text{Reject } H_I \text{ if and only if for all } J \supseteq I,$$
$$H_J \text{ is rejected by the local test } \phi_J.$$

Typically we only make use of the individual hypotheses that are rejected by the procedure i.e. those rejected H_I where I is a singleton.

We consider the case of 4 hypotheses as an example. Suppose the underlined hypotheses are rejected by the local tests.

$$\underline{H_{1234}}$$
$$\underline{H_{123}}\ \underline{H_{124}}\ \underline{H_{134}}\ \underline{H_{234}}$$
$$\underline{H_{12}}\ \underline{H_{13}}\ \underline{H_{14}}\ \underline{H_{23}}\ H_{24}\ H_{34}$$
$$\underline{H_1}\ \underline{H_2}\ H_3\ H_4$$

- Here H_1 is rejected be the closed testing procedure.
- H_2 is not rejected by the closed testing procedure as H_{24} is not rejected by the local test.
- H_{23} is rejected by the closed testing procedure.

Theorem 17. *The closed testing procedure makes no false rejections with probability* $1-\alpha$*. In particular it controls the FWER at level* α*.*

Proof. Assume I_0 is not empty (as otherwise no rejection can be false anyway). In order for the procedure to make a false rejection, we must have $\phi_{I_0} = 1$, but this will occur with probability at most α. □

Different choices for the local tests give rise to different testing procedures. *Holm's procedure* [Holm, 1979] takes ϕ_I to be the Bonferroni test i.e.

$$\phi_I = \begin{cases} 1 & \text{if } \min_{i\in I} p_i \leq \frac{\alpha}{|I|} \\ 0 & \text{otherwise.} \end{cases}$$

To understand what Holm's procedure does, let us order the p-values $p_1, \ldots, p_m$ as $p_{(1)} \leq \cdots \leq p_{(m)}$ with corresponding hypothesis tests $H_{(1)}, \ldots, H_{(m)}$, so (i) is the index of the ith smallest p-value. First consider under what circumstances $H_{(1)}$ is rejected. All subsets I containing (1) must have $\min_{i\in I} p_i = p_{(1)} \leq \alpha/|I|$. The minimum of the RHS occurs when I is the full set $\{1, \ldots, m\}$ so we conclude $H_{(1)}$ is rejected if and only if $p_{(1)} \leq \alpha/m$.

Similarly $H_{(2)}$ will be rejected when all I containing (2) have $\min_{i\in I} p_i \leq \alpha/|I|$. Thus we must certainly have $p_{(1)} \leq \alpha/m$, in which case we only need to ensure all I containing (2) but not (1) have $\min_{i\in I} p_i = p_{(2)} \leq \alpha/|I|$. The minimum of the RHS occurs when $I = \{1, \ldots, m\} \setminus \{1\}$, so we must have $p_{(2)} \leq \alpha/(m-1)$. Continuing this argument, we see that Holm's procedure amounts to the following:

Step 1. If $p_{(1)} \leq \alpha/m$ reject $H_{(1)}$, and go to step 2. Otherwise accept $H_{(1)}, \ldots, H_{(m)}$ and stop.

Step i. If $p_{(i)} \leq \alpha/(m-i+1)$, reject $H_{(i)}$ and go to step $i+1$. Otherwise accept $H_{(i)}, \ldots, H_{(m)}$.

Step m. If $p_{(m)} \leq \alpha$, reject $H_{(m)}$. Otherwise accept $H_{(m)}$.

The p-values are visited in ascending order and rejected until the first time a p-value exceeds a given critical value. This sort of approach is known (slightly confusingly) as a *step-down* procedure.

5.2 The False Discovery Rate

A different approach to multiple testing does not try to control the FWER, but instead attempts to control the *false discovery rate* (FDR) defined by

$$\begin{aligned} \text{FDR} &= \mathbb{E}(\text{FDP}) \\ \text{FDP} &= \frac{N}{\max(R, 1)}, \end{aligned}$$

where FDP is the *false discovery proportion*. Note the maximum in the denominator is to ensure division by zero does not occur. The FDR was introduced in Benjamini and Hochberg [1995], and it is now widely used across science, particularly biostatistics.

The *Benjamini–Hochberg procedure* attempts to control the FDR at level α and works as follows. Let

$$\hat{k} = \max\left\{ i : p_{(i)} \leq \frac{i\alpha}{m} \right\}.$$

Reject $H_{(1)}, \ldots, H_{(\hat{k})}$ (or perform no rejections if $\hat{k}$ is not defined).

Theorem 18. *Suppose that the p_i, $i \in I_0$ are independent, and independent of $\{p_i : i \notin I_0\}$. Then the Benjamini–Hochberg procedure controls the FDR at level α; in fact FDR $\leq \alpha m_0/m$.*

Proof. For each $i \in I_0$, let R_i denote the number of rejections we get by applying a modified Benjamini–Hochberg procedure to

$$p^{\setminus i} := \{p_1, p_2, \ldots, p_{i-1}, p_{i+1}, \ldots, p_m\}$$

with cutoff

$$\hat{k}_i = \max\left\{ j : p^{\setminus i}_{(j)} \leq \frac{\alpha(j+1)}{m} \right\},$$

where $p^{\setminus i}_{(j)}$ is the jth smallest p-value in the set $p^{\setminus i}$.

For $r = 1, \ldots, m$ and $i \in I_0$, note that

$$\begin{aligned}\left\{p_i \leq \frac{\alpha r}{m},\, R = r\right\} &= \left\{p_i \leq \frac{\alpha r}{m},\, p_{(r)} \leq \frac{\alpha r}{m},\, p_{(s)} > \frac{\alpha s}{m} \text{ for all } s > r\right\} \\ &= \left\{p_i \leq \frac{\alpha r}{m},\, p^{\setminus i}_{(r-1)} \leq \frac{\alpha r}{m},\, p^{\setminus i}_{(s-1)} > \frac{\alpha s}{m} \text{ for all } s > r\right\} \\ &= \left\{p_i \leq \frac{\alpha r}{m},\, R_i = r - 1\right\}.\end{aligned}$$

Thus

$$\begin{aligned}\mathrm{FDR} &= \mathbb{E}\left(\frac{N}{\max(R, 1)}\right) \\ &= \sum_{r=1}^{m} \mathbb{E}\left(\frac{N}{r} \mathbb{1}_{\{R=r\}}\right) \\ &= \sum_{r=1}^{m} \frac{1}{r} \mathbb{E}\left(\sum_{i \in I_0} \mathbb{1}_{\{p_i \leq \alpha r/m\}} \mathbb{1}_{\{R=r\}}\right) \\ &= \sum_{r=1}^{m} \frac{1}{r} \sum_{i \in I_0} \mathbb{P}(p_i \leq \alpha r/m,\, R = r) \\ &= \sum_{r=1}^{m} \frac{1}{r} \sum_{i \in I_0} \mathbb{P}(p_i \leq \alpha r/m) \mathbb{P}(R_i = r - 1) \\ &\leq \frac{\alpha}{m} \sum_{i \in I_0} \sum_{r=1}^{m} \mathbb{P}(R_i = r - 1) \\ &= \frac{\alpha m_0}{m}.\end{aligned}$$

□

5.3 Hypothesis testing in high-dimensional regression

Consider the normal linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon}$ where $\varepsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. In the low-dimensional setting, the fact that $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{\beta}^0 \sim \mathcal{N}_p(0, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$ allows us to perform hypothesis tests with $H_0 : \beta^0_j = 0$, for example.

One might hope that studying the distribution of $\hat{\boldsymbol{\beta}}^{\mathrm{L}}_\lambda - \boldsymbol{\beta}^0$ would enable us to do this in the high-dimensional setting when $p \gg n$. However, the distribution of $\hat{\boldsymbol{\beta}}^{\mathrm{L}}_\lambda - \boldsymbol{\beta}^0$ is intractable and depends delicately on the unknown $\boldsymbol{\beta}^0$, making it unsuitable as a basis performing formal hypothesis tests. Other approaches must therefore be used and below we review some of these (see also Dezeure et al. [2015] for a broader review).

Stability selection. *Stability selection* [Meinshausen and Bühlmann, 2010, Shah and Samworth, 2013] is a general technique for uncertainty quantification when performing

variable selection. Given a base variable selection procedure (e.g. $\hat{S}_\lambda$ from the Lasso, or the first q set of variables selected by forward selection) it advocates applying this to subsamples of the data of size $\lfloor n/2 \rfloor$. We can then compute the proportion of times $\hat{\pi}_j$ that the jth variable was selected across the different subsamples, for each j. We can take as our final selected set $\hat{S}_\tau = \{j : \hat{\pi}_j \geq \tau\}$ for some threshold τ. One advantage over simply taking the set of variables selected by the base selection procedure is that bounds are available on $\mathbb{E}(|\hat{S}_\tau \cap S^c|)$. However formal hypothesis tests are unavailable.

Sample splitting. The approach of Meinshausen et al. [2009] exploits the variable screening property of the Lasso (Corollary 8) that provided the true non-zero coefficients are sufficiently far from 0, the set of nonzeroes of the Lasso will contain the true set S. Specifically, first the data are split into two halves: $(\mathbf{Y}^{(1)}, \mathbf{X}^{(1)})$, $(\mathbf{Y}^{(2)}, \mathbf{X}^{(2)})$. The Lasso is applied on $(\mathbf{Y}^{(1)}, \mathbf{X}^{(1)})$ giving a selected set $\hat{S}$. Next OLS is applied on $(\mathbf{Y}^{(2)}, \mathbf{X}^{(2)}_{\hat{S}})$ and p-values are obtained for those variables in $\hat{S}$. The p-values for those variables in $\hat{S}^c$ are set to 1. This process is applied using different random splits of the data and the resulting collection of p-values are aggregated to give a single p-value for each variable. Though this works reasonably well in practice, a drawback is that the validity of the p-values requires a condition on the unknown $\boldsymbol{\beta}^0$. Given that the task is to perform inference for $\boldsymbol{\beta}^0$, such conditions are undesirable.

Recent advances. In the last few years, new methods have been introduced based on debiasing the Lasso that overcome the issue above [Zhang and Zhang, 2014, Van de Geer et al., 2014]. Extending and improving on these methods is currently a highly active area of research. We will try to give a flavour of some of these developments by outlining a simple method for performing hypothesis testing based on some material from Shah and Bühlmann [2018]. Consider testing $H_j : \beta^0_j \neq 0$ so under the null hypothesis

$$\mathbf{Y} = \mathbf{X}_{-j}\boldsymbol{\beta}^0_{-j} + \boldsymbol{\varepsilon}$$

(we will ignore the intercept term for simplicity). Further assume $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2\mathbf{I})$. Idea: if the null model is true, the residuals from regressing $\mathbf{Y}$ on $\mathbf{X}_{-j}$ using the Lasso should look roughly like $\boldsymbol{\varepsilon}$, and so their correlation with $\mathbf{X}_j$ should be small in absolute value. On the other hand, if $\beta^0_j \neq 0$, the residuals should contain some trace of $\beta^0_j\mathbf{X}_j$ and so perhaps correlation with $\mathbf{X}_j$ will be larger in absolute value.

Let $\hat{\boldsymbol{\beta}}$ be the Lasso estimate from regression of $\mathbf{Y}$ on $\mathbf{X}_{-j}$ with tuning parameter $A\sigma\sqrt{\log(p)/n}$. Now under the null, the residuals are $\mathbf{X}_{-j}(\boldsymbol{\beta}^0_{-j} - \hat{\boldsymbol{\beta}}) + \boldsymbol{\varepsilon}$. If we consider the dot product with $\mathbf{X}_j$, we get

$$\underbrace{\frac{1}{\sqrt{n}}\mathbf{X}_j^T\mathbf{X}_{-j}(\boldsymbol{\beta}^0_{-j} - \hat{\boldsymbol{\beta}}_{-j})}_{\text{negligible?}} + \underbrace{\frac{1}{\sqrt{n}}\mathbf{X}_j^T\boldsymbol{\varepsilon}}_{\sim\mathcal{N}(0,\sigma^2)}. \tag{5.1}$$

Unfortunately however, though we know from Theorem 7 that

$$\|\boldsymbol{\beta}^0_{-j} - \hat{\boldsymbol{\beta}}\|_1 \leq \frac{12A\sigma}{\phi^2} s\sqrt{\log(p)/n}$$

with high probability, $\|\mathbf{X}_{-j}^T \mathbf{X}_j\|_\infty / \sqrt{n}$ is the maximum of a sum of n terms and would typically be growing like $n/\sqrt{n} = \sqrt{n}$. We cannot therefore apply Hölder's inequality to conclude that the first term is negligible as desired. Instead consider defining $\mathbf{W}_j$ to be the residuals from regressing $\mathbf{X}_j$ on to $\mathbf{X}_{-j}$ using the Lasso with tuning parameter $B\sqrt{\log(p)/n}$ for some $B > 0$. The KKT conditions give us that

$$\frac{1}{n}\|\mathbf{X}_{-j}^T \mathbf{W}_j\|_\infty \leq B\sqrt{\log(p)/n}.$$

By replacing $\mathbf{X}_j$ with its decorrelated version $\mathbf{W}_j$ and choosing τ appropriately, we can ensure that the first term in (5.1) is negligible. Indeed by Hölder's inequality,

$$|\mathbf{W}_j^T \mathbf{X}_{-j}(\boldsymbol{\beta}^0_{-j} - \hat{\boldsymbol{\beta}})|/\|\mathbf{W}_j\|_2 \leq 12AB\sigma \log(p) s/\|\mathbf{W}_j\|_2$$

with high probability.

Thus under the null hypothesis, our test statistic $\mathbf{W}_j^T(\mathbf{Y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}})/\|\mathbf{W}_j\|_2$ is approximately $\mathcal{N}(0, \sigma^2)$ provided $\sigma \log(p) s/\|\mathbf{W}_j\|_2$ is small. Making use of the square-root Lasso gives a similar test statistic whose approximate null distribution is in fact standard normal thereby allowing for the computation of p-values. One can also show that the test has optimal power in a certain sense [Van de Geer et al., 2014, Shah and Bühlmann, 2018].

References

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.

L. Breiman. Stacked regressions. *Machine Learning*, 24:49–64, 1996.

R. Dezeure, P. Bühlmann, L. Meier, N. Meinshausen, et al. High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statistical Science*, 30(4):533–558, 2015.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1, 2009.

A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, pages 55–67, 1970.

T. Hofmann, B. Schölkopf, and A. Smola. Kernel methods in machine learning. *Annals of Statistics*, pages 1171–1220, 2008.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

G. S. Kimeldorf and G. Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2): 495–502, 1970.

S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

R. Marcus, P. Eric, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

N. Meinshausen. Relaxed lasso. *Computational Statistics and Data Analysis*, 52:374–393, 2007.

N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462, 2006.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B*, 72:417–473, 2010.

N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104:1671–1681, 2009.

G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

P. Ravikumar, H. Liu, J. D. Lafferty, and L. A. Wasserman. Spam: Sparse additive models. In *NIPS*, pages 1201–1208, 2007.

B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer, 2001.

R. D. Shah and P. Bühlmann. Goodness-of-fit tests for high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):113–135, 2018.

R. D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. *arXiv preprint arXiv:1804.07203*, 2018.

R. D. Shah and R. J. Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80, 2013.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

S. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36:614–645, 2008.

S. Van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42 (3):1166–1202, 2014.

M. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity. *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

D. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68:49–67, 2006.

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, pages 894–942, 2010.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *JRSS B*, 67:301–320, 2005.