Modern Statistical Methods

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

Course webpage:
http://www.statslab.cam.ac.uk/~rds37/modern_stat_methods.html

In this course we will study a selection of important modern statistical methods. This selection is heavily biased towards my own interests, but I hope it will nevertheless give you a flavour of some of the most important recent methodological developments in statistics.

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 can run into the thousands and greatly exceed the number of observations. For example, with microarray data, we typically have gene expression values measured for several thousands of genes, but only for a few hundred tissue samples. The classical statistical methods are often simply not applicable in these "high-dimensional" situations.

The course is divided into 4 chapters (of unequal size). Our first chapter will start by introducing ridge regression, a simple generalisation of ordinary least squares. Our study of this will lead us to some beautiful connections with functional analysis and ultimately one of the most successful and flexible classes of learning algorithms: kernel machines.

The second chapter concerns the Lasso and its extensions. The Lasso has been at the centre of much of the developments that have occurred in high-dimensional statistics, and will allow us to perform regression in the seemingly hopeless situation when the number of parameters we are trying to estimate is larger than the number of observations.

Where the previous chapters consider methods for relating a particular response variable to a potentially large collection of (explanatory) variables, in the third chapter, we will study how to infer relationships between the variables themselves. We will see that a fruitful way of formalising the idea of variables being related is via conditional independence, and we investigate ways of inferring conditional depedendencies in high-dimensional data.

Statistics is not only about developing methods that can predict well in the presence of noise, but also about assessing the uncertainty in our predictions and estimates. In the final chapter we will tackle the problem of how to handle performing thousands of hypothesis tests at the same time and more generally the task of quantifying uncertainty in high-dimensional settings.

Before we begin the main content of the course, we will briefly review two key classical statistical methods: ordinary least squares and maximum likelihood estimation. This will help to set the scene and provide a warm-up for the modern methods to come later.

Classical statistics

Ordinary least squares

Imagine data are available in the form of observations $(Y_i, x_i) \in \mathbb{R} \times \mathbb{R}^p$, $i = 1, \ldots, n$, and the aim is to infer a simple *regression function* relating the average value of a *response*, Y_i, and a collection of *predictors* or *variables*, x_i. This is an example of regression analysis, one of the most important tasks in statistics.

A *linear model* for the data assumes that it is generated according to

$$Y = X\beta^0 + \varepsilon, \tag{1}$$

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

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

$$\hat{\beta}^{\mathrm{OLS}} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 = (X^T X)^{-1} X^T Y, \tag{2}$$

provided X has full column rank.

Under the assumptions that (i) $\mathbb{E}(\varepsilon_i) = 0$ and (ii) $\mathrm{Var}(\varepsilon) = \sigma^2 I$, we have that:

- $\mathbb{E}_{\beta^0, \sigma^2}(\hat{\beta}^{\mathrm{OLS}}) = \mathbb{E}\{(X^T X)^{-1} X^T (X\beta^0 + \varepsilon)\} = \beta^0$.
- $\mathrm{Var}_{\beta^0, \sigma^2}(\hat{\beta}^{\mathrm{OLS}}) = (X^T X)^{-1} X^T \mathrm{Var}(\varepsilon) X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$.

The Gauss–Markov theorem states that OLS is the best linear unbiased estimator in our setting: for any other estimator $\tilde{\beta}$ that is linear in Y (so $\tilde{\beta} = AY$ for some fixed matrix A), we have

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

is positive semi-definite.

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 $y \in \mathbb{R}^n$ that we take as a realisation of a random variable Y, we specify its density $f(y; \theta)$ up to some unknown vector of parameters $\theta \in \Theta \subseteq \mathbb{R}^d$, where Θ is the parameter space. The likelihood function is a function of θ for each fixed y given by

$$L(\theta) := L(\theta; y) = c(y) f(y; \theta),$$

where $c(y)$ is an arbitrary constant of proportionality. The maximum likelihood estimate of θ maximises the likelihood, or equivalently it maximises the log-likelihood

$$\ell(\theta) := \ell(\theta; y) = \log f(y; \theta) + \log(c(y)).$$

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

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

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

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

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 observations, $i^{(n)}(\theta)$ (where we have made the dependence on n explicit) satisfies $i^{(n)}(\theta)/n \to I(\theta)$ for some positive definite matrix I. Then denoting the maximum likelihood estimator of θ when there are n observations by $\hat{\theta}^{(n)}$, under regularity conditions, as the number of observations $n \to \infty$ we have

$$\sqrt{n}(\hat{\theta}^{(n)} - \theta) \xrightarrow{d} N_d(0, I^{-1}(\theta)).$$

Returning to our linear model, if we assume in addition that $\varepsilon_i \sim N(0, \sigma^2)$, then the log-likelihood for (β, σ^2) is

$$\ell(\beta, \sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2.$$

We see that the maximum likelihood estimate of β and OLS coincide. It is easy to check that

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

The general theory for MLEs would suggest that approximately $\sqrt{n}(\hat{\beta} - \beta) \sim N_p(0, n\sigma^2 (X^T X)^{-1})$; in fact it is straight-forward to show that this distributional result is exact.

Contents

Chapter 1

Kernel machines

Let us revisit the linear model with

$$Y_i = x_i^T \beta^0 + \varepsilon_i.$$

For unbiased estimators of β^0, their variance gives a way of comparing their quality in terms of squared error loss. For a potentially biased estimator, $\tilde{\beta}$, the relevant quantity is the mean-squared error (MSE),

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

a sum of squared bias and variance terms. 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. In the first two chapters 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 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 outperform the classical approaches.

1.1 Ridge regression

One way to reduce the variance of $\hat{\beta}^{\mathrm{OLS}}$ is to shrink the estimated coefficients towards 0. *Ridge regression* [Hoerl and Kennard, 1970] does this by solving the following optimisation problem

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

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 $\|\beta\|_2^2$. The parameter $\lambda \geq 0$, which controls the severity of the penalty and therefore the degree of the shrinkage towards 0, is known as a *regularisation parameter* or *tuning parameter*. 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, $X\hat{\beta}$ is not invariant under scale transformations of the variables so it is standard practice to centre each column of X (hence making them orthogonal to the intercept term) and then scale them to have ℓ_2-norm $\sqrt{n}$.

It is straightforward to show that after this standardisation of X, $\hat{\mu}_\lambda^{\mathrm{R}} = \bar{Y} := \sum_{i=1}^n Y_i/n$, so we may assume that $\sum_{i=1}^n Y_i = 0$ by replacing Y_i by $Y_i - \bar{Y}$ and then we can remove μ from our objective function. In this case

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

In this form, we can see how the addition of the λI term helps to stabilise the estimator. Note that when X does not have full column rank (such as in high-dimensional situations), we can still compute this estimator. On the other hand, when X does have full column rank, we have the following theorem.

Theorem 1. *For λ sufficiently small (depending on β^0 and σ^2),*

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

is positive definite.

Proof. First we compute the bias of $\hat{\beta}_\lambda^{\mathrm{R}}$. We drop the subscript λ and superscript R for convenience.

$$\begin{aligned}
\mathbb{E}(\hat{\beta}) - \beta^0 &= (X^T X + \lambda I)^{-1} X^T X \beta^0 - \beta^0 \\
&= (X^T X + \lambda I)^{-1} (X^T X + \lambda I - \lambda I)\beta^0 - \beta^0 \\
&= -\lambda (X^T X + \lambda I)^{-1} \beta^0.
\end{aligned}$$

Now we look at the variance of $\hat{\beta}$.

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= \mathbb{E}\{(X^T X + \lambda I)^{-1} X^T \varepsilon\}\{(X^T X + \lambda I)^{-1} X^T \varepsilon\}^T \\
&= \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}.
\end{aligned}$$

Thus $\mathbb{E}(\hat{\beta}^{\mathrm{OLS}} - \beta^0)(\hat{\beta}^{\mathrm{OLS}} - \beta^0)^T - \mathbb{E}(\hat{\beta} - \beta^0)(\hat{\beta} - \beta^0)^T$ is equal to

$$\sigma^2 (X^T X)^{-1} - \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1} - \lambda^2 (X^T X + \lambda I)^{-1} \beta^0 \beta^{0T} (X^T X + \lambda I)^{-1}.$$

After some simplification, we see that this is equal to

$$\lambda (X^T X + \lambda I)^{-1} [\sigma^2 \{2I + \lambda (X^T X)^{-1}\} - \lambda \beta^0 \beta^{0T}] (X^T X + \lambda I)^{-1}.$$

Thus $\mathbb{E}(\hat{\beta}^{\mathrm{OLS}} - \beta^0)(\hat{\beta}^{\mathrm{OLS}} - \beta^0)^T - \mathbb{E}(\hat{\beta} - \beta^0)(\hat{\beta} - \beta^0)^T$ is positive definite for $\lambda > 0$ if and only if

$$\sigma^2\{2I + \lambda(X^TX)^{-1}\} - \lambda\beta^0\beta^{0T}$$

is positive definite, which is true for $\lambda > 0$ sufficiently small (we can take $0 < \lambda < 2\sigma^2/\|\beta^0\|_2^2$). □

The theorem says that $\hat{\beta}^{\mathrm{R}}_\lambda$ outperforms $\hat{\beta}^{\mathrm{OLS}}$ provided λ is chosen appropriately. To be able to use ridge regression effectively, we need a way of selecting a good λ—we will come to this very shortly. What the theorem doesn't really tell us is in what situations we expect ridge regression to perform well. To understand that, we will turn to one of the key matrix decompositions used in statistics, the singular value decomposition (SVD).

1.1.1 Connection to principal components analysis

The singular value decomposition (SVD) is a generalisation of an eigendecomposition of a square matrix. We can factorise any $X \in \mathbb{R}^{n\times p}$ into its SVD

$$X = UDV^T.$$

Here the $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{p\times p}$ are orthogonal matrices and $D \in \mathbb{R}^{n\times p}$ has $D_{11} \geq D_{22} \geq \cdots \geq D_{mm} \geq 0$ where $m := \min(n, p)$ and all other entries of D are zero. To compute such a decomposition requires $O(np\min(n, p))$ operations. The rth columns of U and V are known as the rth left and right singular vectors of X respectively, and D_{rr} is the rth singular value.

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

Let us take $X \in \mathbb{R}^{n\times p}$ 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.

$$\begin{aligned} X\hat{\beta}^{\mathrm{R}}_\lambda &= X(X^TX + \lambda I)^{-1}X^TY \\ &= UDV^T(VD^2V^T + \lambda I)^{-1}VDU^TY \\ &= UD(D^2 + \lambda I)^{-1}DU^TY \\ &= \sum_{j=1}^{p} U_j \frac{D_{jj}^2}{D_{jj}^2 + \lambda} U_j^TY. \end{aligned}$$

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

$$X\hat{\beta}^{\mathrm{OLS}} = X(X^TX)^{-1}X^TY = UU^TY.$$

Both OLS and ridge regression compute the coordinates of Y with respect to the columns of 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 Components Analysis (PCA). Consider $v \in \mathbb{R}^p$ with $\|v\|_2 = 1$. Since the columns of X have had their means subtracted, the sample variance of $Xv \in \mathbb{R}^n$, is

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

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

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

As $\|XV_1\|_2^2/n = D_{11}^2/n$, V_1 determines the linear combination of the columns of X which has the largest sample variance, when the coefficients of the linear combination are constrained to have ℓ_2-norm 1. $XV_1 = D_{11}U_1$ is known as the first principal component of X. Subsequent principal components $D_{22}U_2, \ldots, D_{pp}U_p$ have maximum variance D_{jj}^2/n, subject to being orthogonal to all earlier ones—see example sheet 1 for details.

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

1.2 v-fold cross-validation

Cross-validation is a general technique for selecting a good regression method from among several competing regression methods $\{\hat{f}_\lambda\}_{\lambda \in \Lambda}$. Here the $\hat{f}_\lambda$ take in as arguments *training data* (X, Y) and vector of predictors $x \in \mathbb{R}^p$, and output prediction $\hat{f}_\lambda(x; X, Y) \in \mathbb{R}$.

So far, we have considered the matrix of predictors 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 (x_i, Y_i), $i = 1, \ldots, n$. Then ideally, we might want to pick a λ value such that

$$\mathbb{E}\{(Y^* - \hat{f}_\lambda(x^*; X, Y))^2 | X, Y\} \tag{1.1}$$

is minimised. Here $(x^*, Y^*) \in \mathbb{R}^p \times \mathbb{R}$ is independent of (X, Y) and has the same distribution as (x_1, Y_1). This λ is such that conditional on the original training data (X, Y), 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,

$$\mathbb{E}[\mathbb{E}\{(Y^* - \hat{f}_\lambda(x^*; X, Y))^2 | X, Y\}] \tag{1.2}$$

where compared with (1.1), we have taken a further expectation over the training set.

We still have no way of computing (1.2) 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: $(X^{(1)}, Y^{(1)}), \ldots, (X^{(v)}, Y^{(v)})$. Let $(X^{(-k)}, Y^{(-k)})$ be all the data except that in the kth fold. For each λ on a grid of values, we compute $\hat{f}_\lambda(\cdot; X^{(-k)}, Y^{(-k)})$: the fitted regression function based on all the data except the kth fold. Writing $\kappa(i)$ for the fold to which (x_i, Y_i) belongs, we choose the value of λ that minimises

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{f}_\lambda(x_i; X^{(-\kappa(i))}, Y^{(-\kappa(i))})\}^2. \tag{1.3}$$

Writing λ_{CV} for the minimiser, our final regression function is given by $\hat{f}_{\lambda_{\mathrm{CV}}}(\cdot; X, Y)$.

Note that for each i,

$$\mathbb{E}\{Y_i - \hat{f}_\lambda(x_i; X^{(-\kappa(i))}, Y^{(-\kappa(i))})\}^2 = \mathbb{E}[\mathbb{E}\{Y_i - \hat{f}_\lambda(x_i; X^{(-\kappa(i))}, Y^{(-\kappa(i))})\}^2 | X^{(-\kappa(i))}, Y^{(-\kappa(i))}]. \tag{1.4}$$

This is precisely the expected prediction error in (1.2) but with the training data X, 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 (1.4). 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 (1.3) will tend to be more positively correlated. Common 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. Suppose $\Lambda = \{\lambda_1, \ldots, \lambda_L\}$. We can then minimise

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \sum_{l=1}^{L} w_l \hat{f}_{\lambda_l}(x_i; X^{(-\kappa(i))}, Y^{(-\kappa(i))})\right\}^2$$

over $w \in \mathbb{R}^L$ subject to $w_l \geq 0$ for all l, to give $\hat{w}$. This is a non-negative least-squares optimisation, for which efficient algorithms are available. We may then take the final regression function to be $\sum_{l=1}^{L} \hat{w}_l \hat{f}_{\lambda_l}(\cdot; X, Y)$. This is sometimes known as *stacking* [Wolpert, 1992, Breiman, 1996] and it can often outperform cross-validation.

1.3 The kernel trick

The fitted values from ridge regression are

$$X(X^TX + \lambda I)^{-1}X^TY. \tag{1.5}$$

An alternative way of writing this is suggested by the following

$$\begin{aligned} X^T(XX^T + \lambda I) &= (X^TX + \lambda I)X^T \\ (X^TX + \lambda I)^{-1}X^T &= X^T(XX^T + \lambda I)^{-1} \\ X(X^TX + \lambda I)^{-1}X^TY &= XX^T(XX^T + \lambda I)^{-1}Y. \end{aligned} \tag{1.6}$$

Two remarks are in order:

- Note while X^TX is $p \times p$, XX^T is $n \times n$. Computing fitted values using (1.5) 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 $K = XX^T$ between observations (note $K_{ij} = x_i^Tx_j$).

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

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

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

$$x_{i1}, \ldots, x_{ip}, x_{i1}x_{i1}, \ldots, x_{i1}x_{ip}, x_{i2}x_{i1}, \ldots, x_{i2}x_{ip}, \ldots, x_{ip}x_{ip}. \tag{1.7}$$

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 be rather costly if p is large.

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

$$\begin{aligned} (1/2 + x_i^Tx_j)^2 - 1/4 &= \left(\frac{1}{2} + \sum_k x_{ik}x_{jk}\right)^2 - \frac{1}{4} \\ &= \sum_k x_{ik}x_{jk} + \sum_{k,l} x_{ik}x_{il}x_{jk}x_{jl}. \end{aligned}$$

Observe this amounts to an inner product between vectors of the form (1.7) Thus if we set

$$K_{ij} = (1/2 + x_i^Tx_j)^2 - 1/4 \tag{1.8}$$

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 (1.7). Now computing K using (1.8) would require only $O(p)$ 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 $x_i \in \mathbb{R}^p$ to $\phi(x_i) \in \mathbb{R}^d$ (say) using some *feature map* ϕ (which could, for example introduce quadratic effects), we can instead try to directly compute $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$.
- In fact rather than thinking in terms of feature maps, we can instead try to think about an appropriate measure of similarity $k(x_i, x_j)$ between observations. Modelling in this fashion is sometimes much easier.

We will now formalise and extend what we have learnt with this example.

1.4 Kernels

We have seen how a model with quadratic effects can be fitted very efficiently by replacing the inner product matrix (known as the *Gram matrix*) XX^T in (1.6) with the matrix in (1.8). It is then natural to ask what other non-linear models can be fitted efficiently using this sort of approach.

We will not answer this question directly, but instead we will try to understand the sorts of similarity measures k that can be represented as inner products between transformations of the original data.

That is, we will study the similarity measures $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ from the input space $\mathcal{X}$ to $\mathbb{R}$ for which there exists a *feature map* $\phi : \mathcal{X} \to \mathcal{H}$ where $\mathcal{H}$ is some (real) inner product space with

$$k(x, x') = \langle \phi(x), \phi(x') \rangle. \tag{1.9}$$

Recall that an inner product space is a real vector space $\mathcal{H}$ endowed with a map $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ that obeys the following properties.

(i) Symmetry: $\langle u, v \rangle = \langle v, u \rangle$.

(ii) Linearity: for $a, b \in \mathbb{R}$ $\langle au + bw, v \rangle = a\langle u, v \rangle + b\langle w, v \rangle$.

(iii) Positive-definiteness: $\langle u, u \rangle \geq 0$ with equality if and only if $u = 0$.

Definition 1. A *positive definite kernel* or more simply a *kernel* (for brevity) k is a symmetric map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which for all $n \in \mathbb{N}$ and all $x_1, \ldots, x_n \in \mathcal{X}$, the matrix K with entries

$$K_{ij} = k(x_i, x_j)$$

is positive semi-definite.

A kernel is a little like an inner product, but need not be bilinear in general. However, a form of the Cauchy–Schwarz inequality does hold for kernels.

Proposition 2.

$$k(x, x')^2 \leq k(x, x)k(x', x').$$

Proof. The matrix

$$\begin{pmatrix} k(x,x) & k(x,x') \\ k(x',x) & k(x',x') \end{pmatrix}$$

must be positive semi-definite so in particular its determinant must be non-negative. □

First we show that any inner product of feature maps will give rise to a kernel.

Proposition 3. *k defined by $k(x,x') = \langle \phi(x), \phi(x') \rangle$ is a kernel.*

Proof. Let $x_1, \dots, x_n \in \mathcal{X}$, $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ and consider

$$\begin{aligned} \sum_{i,j} \alpha_i k(x_i, x_j) \alpha_j &= \sum_{i,j} \alpha_i \langle \phi(x_i), \phi(x_j) \rangle \alpha_j \\ &= \left\langle \sum_i \alpha_i \phi(x_i), \sum_j \alpha_j \phi(x_j) \right\rangle \geq 0. \end{aligned}$$ □

Showing that every kernel admits a representation of the form (1.9) is more involved, and we delay this until after we have studied some examples.

1.4.1 Examples of kernels

Proposition 4. *Suppose $k_1, k_2, \dots$ are kernels.*

(i) If $\alpha_1, \alpha_2 \geq 0$ then $\alpha_1 k_1 + \alpha_2 k_2$ is a kernel. If $\lim_{m\to\infty} k_m(x,x') =: k(x,x')$ exists for all $x, x' \in \mathcal{X}$, then k is a kernel.

(ii) The pointwise product $k = k_1 k_2$ is a kernel.

Linear kernel. $k(x,x') = x^T x'$.

Polynomial kernel. $k(x,x') = (1 + x^T x')^d$. To show this is a kernel, we can simply note that $1 + x^T x'$ gives a kernel owing to the fact that 1 is a kernel and (i) of Proposition 4. Next (ii) and induction shows that k as defined above is a kernel.

Gaussian kernel. The highly popular Gaussian kernel is defined by

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

For x close to x' it is large whilst for x far from x' the kernel quickly decays towards 0. The additional parameter $\sigma^2 > 0$, sometimes 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.

To show that it is a kernel first decompose $\|x-x'\|_2^2 = \|x\|_2^2 + \|x'\|_2^2 - 2x^T x'$. Note that by Proposition 3,

$$k_1(x,x') = \exp\left(-\frac{\|x\|_2^2}{2\sigma^2}\right)\exp\left(-\frac{\|x'\|_2^2}{2\sigma^2}\right)$$

is a kernel. Next writing

$$k_2(x,x') = \exp(x^T x'/\sigma^2) = \sum_{r=0}^{\infty}\frac{(x^T x'/\sigma^2)^r}{r!}$$

and using (i) of Proposition 4 shows that k_2 is a kernel. Finally observing that $k = k_1 k_2$ and using (ii) shows that the Gaussian kernel is indeed a kernel.

Sobolev kernel. Take $\mathcal{X}$ to be $[0,1]$ and let $k(x,x') = \min(x,x')$. Note this is the covariance function of Brownian motion so it must be positive definite.

Jaccard similarity kernel. Take $\mathcal{X}$ to be the set of all subsets of $\{1,\dots,p\}$. 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 $k(x,x') = 1$. Showing that this is a kernel is left to the example sheet.

1.4.2 Reproducing kernel Hilbert spaces

Theorem 5. *For every kernel k there exists a feature map ϕ taking values in some inner product space $\mathcal{H}$ such that*

$$k(x,x') = \langle \phi(x), \phi(x') \rangle. \tag{1.10}$$

Proof. We will take $\mathcal{H}$ to be the vector space of functions of the form

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i), \tag{1.11}$$

where $n \in \mathbb{N}$, $x_i \in \mathcal{X}$ and $\alpha_i \in \mathbb{R}$. Our feature map $\phi : \mathcal{X} \to \mathcal{H}$ will be

$$\phi(x) = k(\cdot, x). \tag{1.12}$$

We now define an inner product on $\mathcal{H}$. If f is given by (1.11) and

$$g(\cdot) = \sum_{j=1}^{m} \beta_j k(\cdot, x'_j) \tag{1.13}$$

we define their inner product to be

$$\langle f, g\rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j). \tag{1.14}$$

We need to check this is well-defined as the representations of f and g in (1.11) and (1.13) need not be unique. To this end, note that

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j) = \sum_{i=1}^{n} \alpha_i g(x_i) = \sum_{j=1}^{m} \beta_j f(x'_j). \tag{1.15}$$

The first equality shows that the inner product does not depend on the particular expansion of g whilst the second equality shows that it also does not depend on the expansion of f. Thus the inner product is well-defined.

First we check that with ϕ defined as in (1.12) we do have relationship (1.10). Observe that

$$\langle k(\cdot, x), f\rangle = \sum_{i=1}^{n} \alpha_i k(x_i, x) = f(x), \tag{1.16}$$

so in particular we have

$$\langle \phi(x), \phi(x')\rangle = \langle k(\cdot, x), k(\cdot, x')\rangle = k(x, x').$$

It remains to show that it is indeed an inner product. It is clearly symmetric and (1.15) shows linearity. We now need to show positive definiteness.

First note that

$$\langle f, f\rangle = \sum_{i,j} \alpha_i k(x_i, x_j)\alpha_j \geq 0 \tag{1.17}$$

by positive definiteness of the kernel. Now from (1.16),

$$f(x)^2 = (\langle k(\cdot, x), f\rangle)^2.$$

If we could use the Cauchy–Schwarz inequality on the right-hand side, we would have

$$f(x)^2 \leq \langle k(\cdot, x), k(\cdot, x)\rangle\langle f, f\rangle, \tag{1.18}$$

which would show that if $\langle f, f\rangle = 0$ then necessarily $f = 0$; the final property we need to show that $\langle \cdot, \cdot\rangle$ is an inner product. However, in order to use the traditional Cauchy–Schwarz inequality we need to first know we're dealing with an inner product, which is precisely what we're trying to show!

Although we haven't shown that $\langle \cdot, \cdot\rangle$ is an inner product, we do have enough information to show that it is itself a kernel. We may then appeal to Proposition 2 to obtain (1.18). With this in mind, we argue as follows. Given functions $f_1, \ldots, f_m$ and coefficients $\gamma_1, \ldots, \gamma_m \in \mathbb{R}$, we have

$$\sum_{i,j} \gamma_i \langle f_i, f_j\rangle \gamma_j = \left\langle \sum_i \gamma_i f_i, \sum_j \gamma_j f_j \right\rangle \geq 0$$

where we have used linearity and (1.17), showing that it is a kernel. □

To further discuss the space $\mathcal{H}$ we recall some facts from analysis. Any inner product space $\mathcal{B}$ is also a normed space: for $f \in \mathcal{B}$ we may define $\|f\|_{\mathcal{B}}^2 := \langle f, f \rangle_{\mathcal{B}}$. Recall that a Cauchy sequence $(f_m)_{m=1}^{\infty}$ in $\mathcal{B}$ has $\|f_m - f_n\|_{\mathcal{B}} \to 0$ as $n, m \to \infty$. A normed space where every Cauchy sequence has a limit (in the space) is called *complete*, and a complete inner product space is called a *Hilbert space*.

Hilbert spaces may be thought of as the (potentially) infinite-dimensional analogues of finite-dimensional Euclidean spaces. For later use we note that if V is a closed subspace of a Hilbert space $\mathcal{B}$, then any $f \in \mathcal{B}$ has a decomposition $f = u + v$ with $u \in V$ and

$$v \in V^{\perp} := \{v \in \mathcal{B} : \langle v, z \rangle_{\mathcal{B}} = 0 \text{ for all } z \in V\}.$$

By adding the limits of Cauchy sequences to $\mathcal{H}$ (from Theorem 5) we can make $\mathcal{H}$ a Hilbert space. Indeed, note that if $(f_m)_{m=1}^{\infty} \in \mathcal{H}$ is Cauchy, then since by (1.18) we have

$$|f_m(x) - f_n(x)| \leq \sqrt{k(x,x)} \|f_m - f_n\|_{\mathcal{H}},$$

we may define function $f^* : \mathcal{X} \to \mathbb{R}$ by $f^*(x) = \lim_{m \to \infty} f_m(x)$. We can check that all such f^* can be added to $\mathcal{H}$ to create a Hilbert space.

In fact, the completion of $\mathcal{H}$ is a special type of Hilbert space known as a *reproducing kernel Hilbert space* (RKHS). Since it is the completion of $\mathcal{H}$ that will me of most use to us in what follows, with a slight abuse of notation, we will refer to this space as $\mathcal{H}$.

Definition 2. A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ is a *reproducing kernel Hilbert space* (RKHS) if for all $x \in \mathcal{X}$, there exists $k_x \in \mathcal{H}$ such that

$$f(x) = \langle k_x, f \rangle \qquad \text{for all } f \in \mathcal{B}.$$

The function

$$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\to \mathbb{R} \\ (x, x') &\mapsto \langle k_x, k_{x'} \rangle = k_{x'}(x) \end{aligned}$$

is known as the *reproducing kernel* of $\mathcal{H}$.

By Proposition 3 the reproducing kernel of any RKHS is a (positive definite) kernel, and Theorem 5 shows that to any kernel k is associated an RKHS that has reproducing kernel k; exercise 14 in Example Sheet 1 shows that this RKHS is unique.

Examples

Linear kernel. Here $\mathcal{H} = \{f : f(x) = \beta^T x, \beta \in \mathbb{R}^p\}$ and if $f(x) = \beta^T x$ then $\|f\|_{\mathcal{H}}^2 = \|\beta\|_2^2$.

Sobolev kernel. It can be shown that $\mathcal{H}$ is roughly the space of continuous functions $f : [0,1] \to \mathbb{R}$ with $f(0) = 0$ that are differentiable almost everywhere, and for which $\int_0^1 f'(x)^2 dx < \infty$. It contains the class of Lipschitz functions (functions $f : [0,1] \to \mathbb{R}$ for which there exists some L with $|f(x) - f(y)| \leq L|x - y|$ for all $x, y \in [0,1]$) that are 0 at the origin. It may be shown that the norm is

$$\left(\int_0^1 f'(x)^2 dx \right)^{1/2}.$$

Though the construction of the RKHS from a kernel is explicit, it can be challenging to understand precisely the space and the form of the norm.

1.4.3 The representer theorem

To recap, what we have shown so far is that replacing the matrix XX^T in the definition of an algorithm by K derived form a positive definite kernel is essentially equivalent to running the same algorithm on some mapping of the original data, though with the modification that instances of $x_i^T x_j$ become $\langle \phi(x_i), \phi(x_j) \rangle$.

But what exactly is the optimisation problem we are solving when performing kernel ridge regression? Clearly it is determined by the kernel or equivalently by the RKHS. Note we know that an alternative way of writing the usual ridge regression optimisation is

$$\arg\min_{f \in \mathcal{H}} \left\{ \sum_{i=1}^{n} \{Y_i - f(x_i)\}^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\} \tag{1.19}$$

where $\mathcal{H}$ is the RKHS corresponding to the linear kernel. The following theorem shows in particular that kernel ridge regression (i.e. ridge regression replacing XX^T with K) with kernel k is equivalent to the above with $\mathcal{H}$ now being the RKHS corresponding to k.

Theorem 6 (Representer theorem, [Kimeldorf and Wahba, 1970, Schölkopf et al., 2001])**.** *Let $c : \mathbb{R}^n \times \mathcal{X}^n \times \mathbb{R}^n \to \mathbb{R}$ be an arbitrary loss function, and let $J : [0, \infty) \to \mathbb{R}$ be strictly increasing. Let $x_1, \ldots, x_n \in \mathcal{X}$, $Y \in \mathbb{R}^n$. Finally, let $f \in \mathcal{H}$ where $\mathcal{H}$ is an RKHS with reproducing kernel k, and let $K_{ij} = k(x_i, x_j)$ $i, j = 1, \ldots, n$. Then $\hat{f}$ minimises*

$$Q_1(f) := c(Y, x_1, \ldots, x_n, f(x_1), \ldots, f(x_n)) + J(\|f\|_{\mathcal{H}}^2)$$

over $f \in \mathcal{H}$ iff. $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$ and $\hat{\alpha} \in \mathbb{R}^n$ minimises Q_2 over $\alpha \in \mathbb{R}^n$ where

$$Q_2(\alpha) = c(Y, x_1, \ldots, x_n, K\alpha) + J(\alpha^T K \alpha).$$

Proof. Suppose $\hat{f}$ minimises Q_1. Note that $V := \text{span}\{k(\cdot, x_1), \ldots, k(\cdot, x_n)\}$ is a closed subspace of $\mathcal{H}$. Thus we may write $\hat{f} = u + v$ where $u \in V$ and $v \in V^\perp$. Then

$$\hat{f}(x_i) = \langle k(\cdot, x_i), u + v \rangle = \langle k(\cdot, x_i), u \rangle = u(x_i).$$

Meanwhile, by Pythagoras' theorem we have $J(\|\hat{f}\|_{\mathcal{H}}^2) = J(\|u\|_{\mathcal{H}}^2 + \|v\|_{\mathcal{H}}^2) \geq J(\|u\|_{\mathcal{H}}^2)$ with equality iff. $v = 0$. Thus by optimality of $\hat{f}$, $v = 0$, so $\hat{f}(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ for $\alpha \in \mathbb{R}^n$. Now observe that if $\hat{f}$ takes this form, then $\|\hat{f}\|_{\mathcal{H}}^2 = \alpha^T K \alpha$, so $Q_1(\hat{f}) = Q_2(\alpha)$. Then by optimality of $\hat{f}$, we have that α must minimise Q_2.

Now suppose $\hat{\alpha}$ minimises Q_2 and $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$. Note that $Q_1(\hat{f}) = Q_2(\hat{\alpha})$. If $\tilde{f} \in \mathcal{H}$ has $Q_1(\tilde{f}) \leq Q_1(\hat{f})$, by the argument above, writing $\tilde{f} = u + v$ with $u \in V$, $v \in V^\perp$, we know that $Q_1(u) \leq Q_1(\tilde{f})$. But by optimality of $\hat{\alpha}$ we have $Q_1(\hat{f}) \leq Q_1(u)$, so $Q_1(\hat{f}) = Q_1(\tilde{f})$. □

Consider the result specialised the ridge regression objective. We see that (1.19) is essentially equivalent to minimising

$$\|Y - K\alpha\|_2^2 + \lambda \alpha^T K \alpha,$$

and you may check (see example sheet 1) that the minimiser $\hat{\alpha}$ satisfies $K\hat{\alpha} = K(K + \lambda I)^{-1} Y$. Thus (1.19) is indeed an alternative way of expressing kernel ridge regression.

Viewing the result in the opposite direction gives a more "sensational" perspective. If you had set out trying to minimise Q_1, it might appear completely hopeless as $\mathcal{H}$ could be infinite-dimensional. However, somewhat remarkably we see that this reduces to finding the coefficients $\hat{\alpha}_i$ which solve the simple(r) optimisation problem Q_2.

The result also tells us how to form predictions: given a new observation x, our prediction for $f(x)$ is

$$\hat{f}(x) = \sum_{i=1}^n \hat{\alpha}_i k(x, x_i).$$

1.5 Kernel ridge regression

We have seen how the kernel trick allows us to solve a potentially infinite-dimensional version of ridge regression. This may seem impressive, but ultimately we should judge kernel ridge regression on its statistical properties e.g. predictive performance. Consider a setting where

$$Y_i = f^0(x_i) + \varepsilon_i, \qquad \mathbb{E}(\varepsilon) = 0,\ \mathrm{Var}(\varepsilon) = \sigma^2 I.$$

We shall assume that $f^0 \in \mathcal{H}$ where $\mathcal{H}$ is an RKHS with reproducing kernel k. By scaling σ^2, we may assume $\|f^0\|_{\mathcal{H}} \leq 1$. Let K be the kernel matrix $K_{ij} = k(x_i, x_j)$ with eigenvalues $d_1 \geq d_2 \geq \cdots \geq d_n \geq 0$. We will see that the predictive performance depends delicately on these eigenvalues.

Let $\hat{f}_\lambda$ be the estimated regression function from kernel ridge regression with kernel k:

$$\hat{f}_\lambda = \operatorname*{arg\,min}_{f \in \mathcal{H}} \left\{ \sum_{i=1}^n \{Y_i - f(x_i)\}^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}.$$

Theorem 7. *The mean squared prediction error (MSPE) may be bounded above in the following way:*

$$\frac{1}{n}\mathbb{E}\Big\{\sum_{i=1}^{n}\{f^0(x_i)-\hat{f}_\lambda(x_i)\}^2\Big\} \leq \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{d_i^2}{(d_i+\lambda)^2}+\frac{\lambda}{4n} \tag{1.20}$$
$$\leq \frac{\sigma^2}{n}\frac{1}{\lambda}\sum_{i=1}^{n}\min(d_i/4,\lambda)+\frac{\lambda}{4n}.$$

Proof. We know from the representer theorem that

$$\Big(\hat{f}_\lambda(x_1),\ldots,\hat{f}_\lambda(x_n)\Big)^T = K(K+\lambda I)^{-1}Y.$$

You will show on the example sheet that

$$\Big(f^0(x_1),\ldots,f^0(x_n)\Big)^T = K\alpha,$$

for some $\alpha\in\mathbb{R}^n$, and moreover that $\|f^0\|_{\mathcal{H}}^2\geq\alpha^TK\alpha$. Let the eigendecomposition of K be given by $K=UDU^T$ with $D_{ii}=d_i$ and define $\theta=U^TK\alpha$. We see that n times the LHS of (1.20) is

$$\begin{aligned}\mathbb{E}\|K(K+\lambda I)^{-1}(U\theta+\varepsilon)-U\theta\|_2^2 &= \mathbb{E}\|DU^T(UDU^T+\lambda I)^{-1}(U\theta+\varepsilon)-\theta\|_2^2\\ &= \mathbb{E}\|D(D+\lambda I)^{-1}(\theta+U^T\varepsilon)-\theta\|_2^2\\ &= \|\{D(D+\lambda I)^{-1}-I\}\theta\|_2^2+\mathbb{E}\|D(D+\lambda I)^{-1}U^T\varepsilon\|_2^2.\end{aligned}$$

To compute the second term, we use the 'trace trick':

$$\begin{aligned}\mathbb{E}\|D(D+\lambda I)^{-1}U^T\varepsilon\|_2^2 &= \mathbb{E}[\{D(D+\lambda I)^{-1}U^T\varepsilon\}^TD(D+\lambda I)^{-1}U^T\varepsilon]\\ &= \mathbb{E}[\mathrm{tr}\{D(D+\lambda I)^{-1}U^T\varepsilon\varepsilon^TUD(D+\lambda I)^{-1}\}]\\ &= \sigma^2\mathrm{tr}\{D(D+\lambda I)^{-1}D(D+\lambda I)^{-1}\}\\ &= \sigma^2\sum_{i=1}^{n}\frac{d_i^2}{(d_i+\lambda)^2}.\end{aligned}$$

For the first term, we have

$$\|\{D(D+\lambda I)^{-1}-I\}\theta\|_2^2=\sum_{i=1}^{n}\frac{\lambda^2\theta_i^2}{(d_i+\lambda)^2}.$$

Now as $\theta=DU^T\alpha$, note that $\theta_i=0$ when $d_i=0$. Let D^+ be the diagonal matrix with ith diagonal entry equal to D_{ii}^{-1} if $D_{ii}>0$ and 0 otherwise. Then

$$\sum_{i:d_i>0}\frac{\theta_i^2}{d_i}=\|\sqrt{D^+}\theta\|_2^2=\alpha^TKUD^+U^TK\alpha=\alpha^TUDD^+DU^T\alpha=\alpha^TK\alpha\leq 1.$$

By Hölder's inequality we have

$$\sum_{i=1}^{n} \frac{\lambda^2 \theta_i^2}{(d_i+\lambda)^2} = \sum_{i:d_i>0} \frac{\theta_i^2}{d_i} \frac{d_i \lambda^2}{(d_i+\lambda)^2} \leq \max_{i=1,\ldots,n} \frac{d_i \lambda^2}{(d_i+\lambda)^2} \leq \lambda/4,$$

using the inequality $(a+b)^2 \geq 4ab$ in the final line. Finally note that

$$\frac{d_i^2}{(d_i+\lambda)^2} \leq \min\{1, d_i^2/(4d_i\lambda)\} = \min(\lambda, d_i/4)/\lambda. \qquad \square$$

To interpret this result further, it will be helpful to express it in terms of $\hat{\mu}_i := d_i/n$ (the eigenvalues of K/n) and $\gamma := \lambda/n$. We have

$$\frac{1}{n}\mathbb{E}\bigg\{ \sum_{i=1}^{n} \{f^0(x_i) - \hat{f}_{n\gamma}(x_i)\}^2 \bigg\} \leq \frac{\sigma^2}{\gamma_n} \frac{1}{n} \sum_{i=1}^{n} \min(\hat{\mu}_i/4, \gamma) + \gamma/4 =: \delta_n(\gamma). \tag{1.21}$$

Here we have treated the x_i as fixed, but we could equally well think of them as random. Consider a setup where the x_i are i.i.d. and independent of ε. If we take a further expectation on the RHS of (1.21), our result still holds true (the $\hat{\mu}_i$ are random in this setting). Ideally we would like to then replace $\mathbb{E}\min(\hat{\mu}_i/4, \gamma)$ with a quantity more directly related to the kernel k.

Mercer's theorem is helpful in this regard. This guarantees (under some mild conditions) an eigendecomposition for kernels, which are somewhat like infinite-dimensional analogues of symmetric positive semi-definite matrices. Under certain technical conditions, we may write

$$k(x, x') = \sum_{j=1}^{\infty} \mu_j e_j(x) e_j(x') \tag{1.22}$$

where writing $p(x)$ for the density of each x_i, the eigenfunctions e_j and corresponding eigenvalues $\mu_j \geq 0$ satisfy $\sum_{j=1}^{\infty} \mu_j < \infty$ and obey the integral equation

$$\mu_j e_j(x') = \int_{\mathcal{X}} k(x, x') e_j(x) p(x) dx.$$

The e_j form an orthonormal basis of $\mathcal{H}$ in the sense that

$$\int_{\mathcal{X}} e_k(x) e_j(x) p(x) dx = \mathbb{1}_{\{k=j\}}.$$

One can show that for all $\gamma > 0$,

$$\mathbb{E}\bigg(\frac{1}{n} \sum_{i=1}^{n} \min(\hat{\mu}_i/4, \gamma) \bigg) \leq \frac{1}{n} \sum_{j=1}^{\infty} \min(\mu_j/4, \gamma).$$

Theorem 8. *Provided the eigendecomposition* (1.22) *holds, there exists* γ_n *such that for fixed* $\sigma^2 > 0$*,*

$$\frac{1}{n}\mathbb{E}\left\{ \sum_{i=1}^{n} \{f^0(x_i) - \hat{f}_{\gamma_n}(x_i)\}^2 \right\} = o(n^{-1/2}).$$

Proof. Let $\phi : [0, \infty) \to [0, \infty)$ be given by

$$\phi(\gamma) := \sum_{j=1}^{\infty} \min(\mu_j, \gamma).$$

Observe that ϕ is increasing and as $\sum_{j=1}^{\infty} \mu_j < \infty$, $\lim_{\gamma \downarrow 0} \phi(\gamma) = 0$ (indeed, note that given an arbitrary $\epsilon > 0$, there exists M such that $\sum_{j=M}^{\infty} \mu_j \leq \epsilon$, but then $\phi(\gamma) \leq M\gamma + \epsilon \to \epsilon$ as $\gamma \downarrow 0$). Let $\gamma_n = n^{-1/2}\sqrt{\phi(n^{-1/2})}$ so $\gamma_n = o(n^{-1/2})$. Thus for n sufficiently large $\phi(\gamma_n) \leq \phi(n^{-1/2})$, whence for such n we have

$$\begin{aligned} \inf_{\gamma > 0} \{\phi(\gamma)/(n\gamma) + \gamma\} &\leq \frac{\phi(\gamma_n)}{n\gamma_n} + \gamma_n \\ &\leq 2\sqrt{\phi(n^{-1/2})}/\sqrt{n} = o(n^{-1/2}). \end{aligned}$$

□

Sobolev kernel. When k is the Sobolev kernel and $p(x)$ is the uniform density on $[0, 1]$, one can show that the eigenvalues satisfy

$$\mu_j/4 = \frac{1}{\pi^2(2j-1)^2}.$$

Thus

$$\begin{aligned} \sum_{i=1}^{\infty} \min(\mu_i/4, \gamma_n) &\leq \frac{\gamma_n}{2}\left(\frac{1}{\sqrt{\pi^2\gamma_n}} + 1 \right) + \frac{1}{\pi^2} \int_{\{(\pi^2\gamma_n)^{-1/2}+1\}/2}^{\infty} \frac{1}{(2x-1)^2} dx \\ &= \sqrt{\gamma_n}/\pi + \gamma_n/2 = O(\sqrt{\gamma_n}) \end{aligned}$$

as $\gamma_n \to 0$. Putting things together, we see that

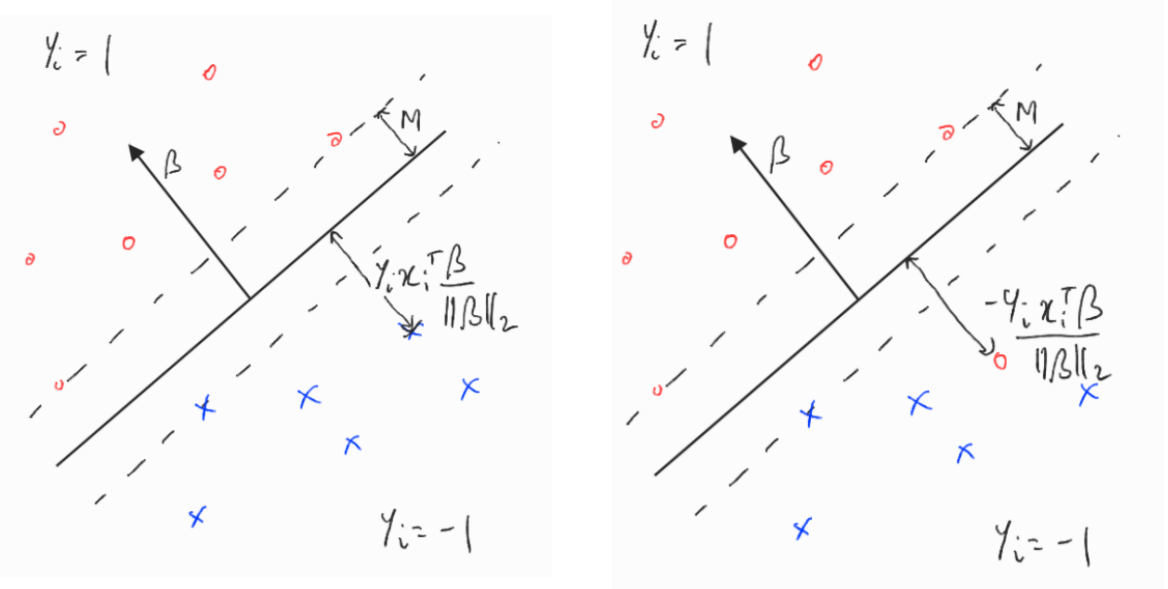
$$\mathbb{E}\delta_n(\gamma_n) = O\left(\frac{\sigma^2}{n\gamma_n^{1/2}} + \gamma_n \right).$$

Thus an optimal $\gamma_n \sim (\sigma^2/n)^{2/3}$ gives an error rate of order $(\sigma^2/n)^{2/3}$.

1.6 Other kernel machines

Thus far we have we have only considered applying the kernel trick to ridge regression, which as we have seen has attractive theoretical properties as a regression method. However the kernel trick and the representer theorem are much more generally applicable. In settings where the Y_i are not continuous but are in $\{-1, 1\}$ (e.g. labels for spam and ham, fraud and not fraud etc.), popular approaches include kernel logistic regression and the support vector machine (SVM) [Cortes and Vapnik, 1995].

1.6.1 The support vector machine



Consider first the simple case where the data in the two classes $\{x_i\}_{i:Y_i=1}$ and $\{x_i\}_{i:Y_i=-1}$ are separable by a hyperplane through the origin, so there exists $\beta \in \mathbb{R}^p$ with $\|\beta\|_2 = 1$ such that $Y_i \beta^T x_i > 0$ for all i. Note β would then be a unit normal vector to a plane that separates the two classes.

There may be an infinite number of planes that separate the classes, in which case it seems sensible to use the plane that maximises the margin between the two classes. Consider therefore the following optimisation problem,

$$\max_{\beta \in \mathbb{R}^p, M>0} M$$
$$\text{subject to } Y_i x_i^T \beta / \|\beta\|_2 \geq M,\ i = 1, \ldots, n, \tag{1.23}$$

or equivalently,

$$\min_{\beta \in \mathbb{R}^p, M>0} 1/M^2 \qquad \text{subject to (1.23).}$$

Note that by normalising β above we need not impose the constraint that $\|\beta\|_2 = 1$.

Suppose now that the classes are not separable. One way to handle this is to replace the constraint $Y_i x_i^T \beta / \|\beta\|_2 \geq M$ with a penalty for how far over the margin boundary x_i is. We would like the penalty to be zero if x_i is on the correct side of the boundary (i.e. when $Y_i x_i^T \beta / \|\beta\|_2 \geq M$), and should be equal to the distance over the boundary, $M - Y_i x_i^T \beta / \|\beta\|_2$ otherwise. It will in fact be more convenient to penalise according to $1 - Y_i x_i^T \beta / (\|\beta\|_2 M)$ in the latter case, which is the distance measured in units of M. This penalty is invariant to β undergoing any positive scaling, so we may set $\|\beta\|_2 = 1/M$, thus eliminating M from the objective function. Adding the penalty we then arrive at

$$\arg\min_{\beta \in \mathbb{R}^p} \|\beta\|_2^2 + \lambda \sum_{i=1}^{n} (1 - Y_i x_i^T \beta)_+,$$

where $(\cdot)_+$ denotes the positive part. Replacing λ with $1/\lambda$ we can write the objective in the more familiar-looking form

$$\underset{\beta\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n}(1 - Y_i x_i^T\beta)_+ + \lambda\|\beta\|_2^2.$$

Thus far we have restricted ourselves to hyperplanes through the origin but we would more generally want to consider any translate of these i.e. any hyperplane. This can be achieved by allowing ourselves to translate the x_i by an arbitrary vector b, giving

$$\underset{\beta\in\mathbb{R}^p, b\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n}(1 - Y_i(x_i - b)^T\beta)_+ + \lambda\|\beta\|_2^2,$$

or equivalently

$$(\hat{\mu}, \hat{\beta}) = \underset{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n}\{1 - Y_i(x_i^T\beta + \mu)\}_+ + \lambda\|\beta\|_2^2. \tag{1.24}$$

This final objective defines the *support vector classifier*; given a new observation x predictions are obtained via $\text{sgn}(\hat{\mu} + x^T\hat{\beta})$.

Note that the objective in (1.24) may be re-written as

$$(\hat{\mu}, \hat{f}) = \underset{(\mu,f)\in\mathbb{R}\times\mathcal{H}}{\arg\min} \sum_{i=1}^{n}[1 - Y_i\{f(x_i) + \mu\}]_+ + \lambda\|f\|_{\mathcal{H}}^2, \tag{1.25}$$

where $\mathcal{H}$ is the RKHS corresponding to the linear kernel. The representer theorem (more specifically the variant in question 10 of example sheet 1) shows that (1.25) for an arbitrary RKHS with kernel k and kernel matrix K is equivalent to the *support vector machine*

$$(\hat{\mu}, \hat{\alpha}) = \underset{(\mu,\alpha)\in\mathbb{R}\times\mathbb{R}^n}{\arg\min} \sum_{i=1}^{n}[1 - Y_i\{K_i^T\alpha + \mu\}]_+ + \lambda\alpha^T K\alpha.$$

Predictions at a new x are given by

$$\text{sgn}\left(\hat{\mu} + \sum_{i=1}^{n}\hat{\alpha}_i k(x, x_i)\right).$$

1.6.2 Logistic regression

Recall that standard logistic regression may be motivated by assuming

$$\log\left(\frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = -1)}\right) = x_i^T\beta^0$$

and picking $\hat{\beta}$ to maximise the log-likelihood. This leads to (see example sheet) the following optimisation problem:

$$\arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \log\{1 + \exp(-Y_i x_i^T \beta)\}.$$

The 'kernelised' version is given by

$$\arg\min_{f \in \mathcal{H}} \left\{ \sum_{i=1}^{n} \log[1 + \exp\{-Y_i f(x_i)\}] + \lambda \|f\|_{\mathcal{H}}^2 \right\},$$

where $\mathcal{H}$ is an RKHS. As in the case of the SVM, the representer theorem gives a finite-dimensional optimisation that is equivalent to the above.

1.7 Large-scale kernel machines

We introduced the kernel trick as a computational device that avoided performing calculations in a high or infinite dimensional feature space and, in the case of kernel ridge regression reduced computation down to forming the $n \times n$ matrix K and then inverting $K + \lambda I$. This can be a huge saving, but when n is very large, this can present serious computational difficulties. Even if p is small, the $O(n^3)$ cost of inverting $K + \lambda I$ may cause problems. What's worse, the fitted regression function is a sum over n terms:

$$\hat{f}(\cdot) = \sum_{i=1}^{n} \hat{\alpha}_i k(x_i, \cdot).$$

Even to evaluate a prediction at a single new observation requires $O(n)$ computations unless $\hat{\alpha}$ is sparse.

In recent years, there has been great interest in speeding up computations for kernel machines. We will discuss one exciting approach based on random feature expansions. Given a kernel k, the key idea is to develop a random map

$$\hat{\phi} : \mathcal{X} \to \mathbb{R}^b$$

with b small such that $\mathbb{E}\{\hat{\phi}(x)^T \hat{\phi}(x')\} = k(x, x')$. In a sense we are trying to reverse the kernel trick by approximating the kernel using a random feature map. To increase the quality of the approximation of the kernel, we can consider

$$x \mapsto \frac{1}{\sqrt{L}} (\hat{\phi}_1(x), \ldots, \hat{\phi}_L(x)) \in \mathbb{R}^{Lb}$$

with each $(\hat{\phi}_l(x))_{l=1}^L$ being i.i.d. for each x. Let Φ be the matrix with ith row given by $(\hat{\phi}_1(x_i), \ldots, \hat{\phi}_L(x_i))/\sqrt{L}$. We may then run our learning algorithm replacing the initial

matrix of predictors X with Φ. For example, when performing ridge regression, we can compute

$$(\Phi^T\Phi + \lambda I)^{-1}\Phi^T Y,$$

which would require $O(nL^2b^2 + L^3b^3)$ operations: a cost linear in n. Predicting a new observation would cost $O(Lb)$.

The work of Rahimi and Recht [2007] proposes a construction of such a random mapping $\hat{\phi}$ for shift-invariant kernels, that is kernels for which there exists a function h with $k(x, x') = h(x - x')$ for all $x, x' \in \mathcal{X} = \mathbb{R}^p$. A useful property of such kernels is given by Bochner's theorem.

Theorem 9 (Bochner's theorem)**.** *Let $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be a continuous kernel. Then k is shift-invariant if and only if there exists some $c > 0$ and distribution F on $\mathbb{R}^p$ such that when $W \sim F$*

$$k(x, x') = c\mathbb{E}e^{i(x-x')^T W} = c\mathbb{E}\cos((x - x')^T W).$$

To make use of this theorem, first observe the following. Let $u \sim U[-\pi, \pi]$, $x, y \in \mathbb{R}$. Then

$$2\mathbb{E}\cos(x + u)\cos(y + u) = 2\mathbb{E}\{(\cos x \cos u - \sin x \sin u)(\cos y \cos u - \sin y \sin u)\}.$$

Now as $u \overset{d}{=} -u$, $\mathbb{E}\cos u \sin u = \mathbb{E}\cos(-u)\sin(-u) = -\mathbb{E}\cos u \sin u = 0$. Also of course $\cos^2 u + \sin^2 u = 1$ so $\mathbb{E}\cos^2 u = \mathbb{E}\sin^2 u = 1/2$. Thus

$$2\mathbb{E}\cos(x + u)\cos(y + u) = \cos x \cos y + \sin x \sin y = \cos(x - y).$$

Given a shift-invariant kernel k with associated distribution F, suppose $W \sim F$ and let $u \sim U[-\pi, \pi]$ independently. Define

$$\hat{\phi}(x) = \sqrt{2c}\cos(W^T x + u).$$

Then

$$\begin{aligned}\mathbb{E}\hat{\phi}(x)\hat{\phi}(x') &= 2c\mathbb{E}[\mathbb{E}\{\cos(W^T x + u)\cos(W^T x' + u)|W\}] \\ &= c\mathbb{E}\cos((x - x')^T W) = k(x, x').\end{aligned}$$

As a concrete example of this approach, let us take the Gaussian kernel $k(x, x') = \exp\{-\|x - x'\|_2^2/(2\sigma^2)\}$. Note that if $W \sim N(0, \sigma^{-2}I)$, it has characteristic function $\mathbb{E}(e^{it^T W}) = e^{-\|t\|_2^2/(2\sigma^2)}$ so we may take $\hat{\phi}(x) = \sqrt{2}\cos(W^T x + u)$.

Chapter 2

The Lasso and beyond

2.1 Model selection

Let us revisit the linear model $Y = X\beta^0 + \varepsilon$ where $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2 I$. In many modern datasets, there are reasons to believe there are many more variables present than are necessary to explain the response. Let S be the set $S = \{k : \beta_k^0 \neq 0\}$ and suppose $s := |S| \ll p$.

The MSPE of OLS is

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

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 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.

2.2 The Lasso estimator

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

$$\frac{1}{2n}\|Y - \mu\mathbf{1} - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{2.1}$$

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

Like ridge regression, $\hat{\beta}^{\mathrm{L}}_\lambda$ 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 X matrix, and also centre Y and thus remove μ from the objective. Define

$$Q_\lambda(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1. \tag{2.2}$$

Now the minimiser(s) of $Q_\lambda(\beta)$ will also be the minimiser(s) of

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

Similarly, with the Ridge regression objective, we know that $\hat{\beta}^{\mathrm{R}}_\lambda$ minimises $\|Y - X\beta\|_2^2$ subject to $\|\beta\|_2 \leq \|\hat{\beta}^{\mathrm{R}}_\lambda\|_2$.

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

The important point to note is that the ℓ_1 ball $\{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq \|\hat{\beta}^{\mathrm{L}}_\lambda\|_1\}$ 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.

2.2.1 Prediction error of the Lasso (slow rate)

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 X have been centred and scaled (as we will always assume from now on unless stated otherwise) and assume the normal linear model (where we have already centred Y),

$$Y = X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbf{1} \tag{2.3}$$

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

Theorem 10. *Let $\hat{\beta}$ be any Lasso solution when*

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

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

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

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

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

Rearranging,

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

Now $|\varepsilon^T X(\hat{\beta} - \beta^0)| \leq \|X^T\varepsilon\|_\infty\|\hat{\beta} - \beta^0\|_1$. Let $\Omega = \{\|X^T\varepsilon\|_\infty/n \leq \lambda\}$. Lemma 14 below shows that $\mathbb{P}(\Omega) \geq 1 - 2p^{-(A^2/2-1)}$. Working on the event Ω, we obtain

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

□

2.2.2 Concentration inequalities I

The proof of Theorem 10 relies on a lower bound for the probability of the event Ω. A union bound gives

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

Now $X_j^T\varepsilon/n \sim N(0, \sigma^2/n)$, so if we obtain a bound on the tail probabilities of normal distributions, the argument above will give a bound for $\mathbb{P}(\Omega)$.

Motivated by the need to bound normal tail probabilities, we will briefly discuss the topic of *concentration inequalities* that provide such bounds for much wider classes of random variables. Concentration inequalities are vital for the study of many modern algorithms and in our case here, they will reveal that the attractive properties of the Lasso presented in Theorem 10 hold true for a variety of non-normal errors.

We begin our discussion with the simplest tail bound, *Markov's inequality*, which states that given a non-negative random variable W,

$$\mathbb{P}(W \geq t) \leq \frac{\mathbb{E}(W)}{t}.$$

This immediately implies that given a strictly increasing function $\varphi : \mathbb{R} \to [0, \infty)$ and any random variable W,

$$\mathbb{P}(W \geq t) = \mathbb{P}\{\varphi(W) \geq \varphi(t)\} \leq \frac{\mathbb{E}(\varphi(W))}{\varphi(t)}.$$

Applying this with $\varphi(t) = e^{\alpha t}$ ($\alpha > 0$) yields the so-called *Chernoff bound*:

$$\mathbb{P}(W \geq t) \leq \inf_{\alpha > 0} e^{-\alpha t}\mathbb{E}e^{\alpha W}.$$

Consider the case when $W \sim N(0, \sigma^2)$. Recall that

$$\mathbb{E}e^{\alpha W} = e^{\alpha^2\sigma^2/2}. \tag{2.4}$$

Thus

$$\mathbb{P}(W \geq t) \leq \inf_{\alpha > 0} e^{\alpha^2\sigma^2/2 - \alpha t} = e^{-t^2/(2\sigma^2)}.$$

Note that to arrive at this bound, all we required was (an upper bound on) the moment generating function (mgf) of W (2.4).

Sub-Gaussian variables

Definition 3. We say a random variable W is *sub-Gaussian* if there exists $\sigma > 0$ such that

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

for all $\alpha \in \mathbb{R}$. We then say that W is *sub-Gaussian with parameter* σ.

Proposition 11 (Sub-Gaussian tail bound)**.** *If W is sub-Gaussian with parameter σ then*

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

As well as Gaussian random variables, the sub-Gaussian class includes bounded random variables.

Lemma 12 (Hoeffding's lemma)**.** *If W 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 13. *Let $(W_i)_{i=1}^n$ be a sequence of independent sub-Gaussian random variables with parameters $(\sigma_i)_{i=1}^n$ and let $\gamma \in \mathbb{R}^n$. Then $\gamma^T W$ is sub-Gaussian with parameter $\left(\sum_i \gamma_i^2 \sigma_i^2\right)^{1/2}$.*

Proof. Wlog, we may assume $\mathbb{E}W_i = 0$ for all i. We have

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

□

We can now prove a more general version of the probability bound required for Theorem 10.

Lemma 14. *Suppose $(\varepsilon_i)_{i=1}^n$ are independent, mean-zero and sub-Gaussian with common parameter σ. Note that this includes $\varepsilon \sim N_n(0, \sigma^2 I)$. Let $\lambda = A\sigma\sqrt{\log(p)/n}$. Then*

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

Proof.

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

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

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

□

2.2.3 Some facts from optimisation theory and convex analysis

In order to study the Lasso in detail, it will be helpful to review some basic facts from optimisation and convex analysis.

Convexity

A set $C \subseteq \mathbb{R}^d$ is *convex* if

$$x, y \in C \Rightarrow (1-t)x + ty \in C \qquad \text{for all } t \in (0,1).$$

Given $C \subseteq \mathbb{R}^d$, We say a function $f : C \to \mathbb{R}$ is *convex* if C is convex and

$$f\big((1-t)x + ty\big) \leq (1-t)f(x) + tf(y)$$

for all $x, y \in C$ and $t \in (0,1)$. It is *strictly convex* if the inequality is strict for all $x, y \in C$ with $x \neq y$. In the following, $C \subseteq \mathbb{R}^d$ is a convex set.

Proposition 15. *(i) Let $f_1, \ldots, f_m : C \to \mathbb{R}$ be convex functions. Then if $c_1, \ldots, c_m \geq 0$, $c_1 f_1 + \cdots + c_m f_m : C \to \mathbb{R}$ is a convex function.*

(ii) If $f : C \to \mathbb{R}$, and $A : \mathbb{R}^m \to \mathbb{R}^d$ is an affine function (so $A(x) = Mx + b$ for $M \in \mathbb{R}^{d \times m}$ and $b \in \mathbb{R}^d$) then $g : D \to \mathbb{R}$, where $D = \{x \in \mathbb{R}^m : A(x) \in C\}$ given by $g(x) = f(A(x))$ is convex.

(iii) If $f : C \to \mathbb{R}$ is convex with C open and f is twice continuously differentiable on C, then

(a) f is convex iff. its Hessian $H(x)$ is positive semi-definite for all $x \in C$,

(b) f is strictly convex if $H(x)$ is positive definite for all $x \in C$.

The Lagrangian method

Consider an optimisation problem of the form

$$\text{minimise } f(x), \text{ subject to } g(x) = 0, \qquad x \in C \subseteq \mathbb{R}^d, \tag{2.5}$$

where $g : C \to \mathbb{R}^b$. Suppose the optimal value is $c^* \in \mathbb{R}$. The Lagrangian for this problem is defined as

$$L(x, \theta) = f(x) + \theta^T g(x)$$

where $\theta \in \mathbb{R}^b$. Note that

$$\inf_{x \in C} L(x, \theta) \leq \inf_{x \in C : g(x) = 0} L(x, \theta) = c^*$$

for all θ. The Lagrangian method involves finding a $\theta = \theta^*$ such that the minimising $x = x^*$ on the LHS satisfies $g(x^*) = 0$. This x^* must then be a minimiser in the original problem (2.5).

Subgradients

Definition 4. Given convex $C \subseteq \mathbb{R}^d$, a vector $v \in \mathbb{R}^d$ is a *subgradient* of a convex function $f : C \to \mathbb{R}$ at x if

$$f(y) \geq f(x) + v^T(y - x) \qquad \text{for all } \; y \in C.$$

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

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

Proposition 16. *Let $f : C \to \mathbb{R}$ be convex, and suppose f is differentiable at $x \in \text{int}(C)$. Then $\partial f(x) = \{\nabla f(x)\}$.*

Proposition 17. *Let $f, g : C \to \mathbb{R}$ be convex with $\text{int}(C) \neq \emptyset$ and let $\alpha > 0$. Then*

$$\begin{aligned} \partial(\alpha f)(x) &= \alpha \partial f(x) = \{\alpha v : v \in \partial f(x)\}, \\ \partial(f + g)(x) &= \partial f(x) + \partial g(x) = \{v + w : v \in \partial f(x),\, w \in \partial g(x)\}. \end{aligned} \tag{2.6}$$

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 18. *Given convex $f : C \to \mathbb{R}$, $x^* \in \underset{x \in C}{\arg\min}\, f(x)$ if and only if $0 \in \partial f(x^*)$.*

Proof.

$$\begin{aligned} f(y) \geq f(x^*) \quad \text{for all } \; y \in C &\Leftrightarrow f(y) \geq f(x^*) + 0^T(y - x) \quad \text{for all } \; y \in C \\ &\Leftrightarrow 0 \in \partial f(x^*). \end{aligned}$$

□

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 $\|tx + (1-t)y\|_1 \leq t\|x\|_1 + (1-t)\|y\|_1$. We introduce some notation that will be helpful here and throughout the rest of the course.

For $x \in \mathbb{R}^d$ and $A = \{k_1, \ldots, k_m\} \subseteq \{1, \ldots, d\}$ with $k_1 < \cdots < k_m$, by x_A we will mean $(x_{k_1}, \ldots, x_{k_m})^T$. Similarly if X has d columns we will write X_A for the matrix

$$X_A = (X_{k_1} \cdots X_{k_m}).$$

Further in this context, by A^c, we will mean $\{1, \ldots, d\} \setminus A$. Additionally, when in subscripts we will use the shorthand $-j = \{j\}^c$ and $-jk = \{j, k\}^c$. Note these column and component extraction operations will always be considered to have taken place first before any further operations on the matrix, so for example $X_A^T = (X_A)^T$. Finally, define

$$\text{sgn}(x_1) = \begin{cases} -1 & \text{if } x_1 < 0 \\ 0 & \text{if } x_1 = 0 \\ 1 & \text{if } x_1 > 0, \end{cases}$$

and

$$\text{sgn}(x) = (\text{sgn}(x_1), \ldots, \text{sgn}(x_d))^T.$$

Proposition 19. *For $x \in \mathbb{R}^d$ let $A = \{j : x_j \neq 0\}$. Then*

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

Proof. With a view to applying (2.6) For $j = 1, \ldots, d$, let

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

Then $\|\cdot\| = \sum_j g_j(\cdot)$ so by Proposition 17, $\partial\|x\|_1 = \sum_j \partial g_j(x)$. When x has $x_j \neq 0$, g_j is differentiable at x so by Proposition 16 $\partial g_j(x) = \{\operatorname{sgn}(x_j)e_j\}$ where e_j is the jth unit vector. When $x_j = 0$, if $v \in \partial g_j(x)$ we must have

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

so

$$|y_j| \geq v^T(y - x) \qquad \text{for all } y \in \mathbb{R}^d. \tag{2.7}$$

we claim that the above holds iff. $v_j \in [-1, 1]$ and $v_{-j} = 0$. For the 'if' direction, note that $v^T(y - x) = v_j y_j \leq |y_j|$. Conversely, set $y_{-j} = x_{-j} + v_{-j}$ and $y_j = 0$ in (2.7) to get $0 \geq \|v_{-j}\|_2^2$, so $v_{-j} = 0$. Then take y with $y_{-j} = x_{-j}$ and $y_j = \operatorname{sgn}(v_j)$ to get $1 \geq |v_j|$. Forming the set sum of the subdifferentials as in (2.6) then gives the result. □

2.2.4 Lasso solutions

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

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

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

Lasso solutions need not be unique (e.g. if X has duplicate columns), though for most reasonable design matrices, Lasso solutions will be unique. We will often tacitly assume Lasso solutions are unique in the statement of our theoretical results. It is however straightforward to show that the Lasso fitted values are unique.

Proposition 20. *Fix $\lambda \geq 0$ and suppose $\beta^{(1)}$ and $\beta^{(2)}$ are two Lasso solutions. Then $X\beta^{(1)} = X\beta^{(2)}$.*

Proof. Suppose $\beta^{(1)}$ and $\beta^{(2)}$ both give an optimal objective value of c^*. Now by strict convexity of $\|\cdot\|_2^2$,

$$\|Y - X\beta^{(1)}/2 - X\beta^{(2)}/2\|_2^2 \leq \|Y - X\beta^{(1)}\|_2^2/2 + \|Y - X\beta^{(2)}\|_2^2/2,$$

with equality if and only if $X\beta^{(1)} = X\beta^{(2)}$. Since $\|\cdot\|_1$ is also convex, we see that

$$\begin{aligned}
c^* &\leq Q_\lambda(\beta^{(1)}/2 + \beta^{(2)}/2) \\
&= \|Y - X\beta^{(1)}/2 - X\beta^{(2)}/2\|_2^2/(2n) + \lambda\|\beta^{(1)}/2 + \beta^{(2)}/2\|_1 \\
&\leq \|Y - X\beta^{(1)}\|_2^2/(4n) + \|Y - X\beta^{(2)}\|_2^2/(4n) + \lambda\|\beta^{(1)}/2 + \beta^{(2)}/2\|_1 \\
&\leq \{\|Y - X\beta^{(1)}\|_2^2/(4n) + \lambda\|\beta^{(1)}\|_1/2\} + \{\|Y - X\beta^{(2)}\|_2^2/(4n) + \lambda\|\beta^{(2)}\|_1/2\} \\
&= Q(\beta^{(1)})/2 + Q(\beta^{(2)})/2 = c^*.
\end{aligned}$$

Equality must prevail throughout this chain of inequalities, so $X\beta^{(1)} = X\beta^{(2)}$. □

Define the *equicorrelation set* $\hat{E}_\lambda$ to be the set of k such that

$$\frac{1}{n}|X_k^T(Y - X\hat{\beta}_\lambda^{\mathrm{L}})| = \lambda.$$

Note that $\hat{E}_\lambda$ is well-defined since it only depends on the fitted values, which (as we have just shown) are unique. By the KKT conditions, the equicorrelation set contains the set of non-zeroes of all Lasso solutions. Note that if $\mathrm{rank}(X_{\hat{E}_\lambda}) = |\hat{E}_\lambda|$ then the Lasso solution must be unique: indeed if $\beta^{(1)}$ and $\beta^{(2)}$ are two Lasso solutions, then as

$$X_{\hat{E}_\lambda}(\beta^{(1)}_{\hat{E}_\lambda} - \beta^{(2)}_{\hat{E}_\lambda}) = 0$$

by linear independence of the columns of $X_{\hat{E}_\lambda}$, $\beta^{(1)}_{\hat{E}_\lambda} = \beta^{(2)}_{\hat{E}_\lambda}$.

2.2.5 Variable selection

Consider now the "noiseless" version of the high-dimensional linear model (2.3), $Y = X\beta^0$. The case with noise can be dealt with by similar arguments to those we will use below when we work on an event that $\|X^T\varepsilon\|_\infty/n$ is small (see example sheet).

Let $S = \{k : \beta_k^0 \neq 0\}$, $N = \{1, \ldots, p\} \setminus S$ and assume wlog that $S = \{1, \ldots, s\}$, and also that $\mathrm{rank}(X_S) = s$.

Theorem 21. *Let $\lambda > 0$ and define $\Delta = X_N^T X_S (X_S^T X_S)^{-1}\mathrm{sgn}(\beta_S^0)$. If $\|\Delta\|_\infty \leq 1$ and for $k \in S$,*

$$|\beta_k^0| > \lambda|\mathrm{sgn}(\beta_S^0)^T[\{\tfrac{1}{n}X_S^T X_S\}^{-1}]_k|, \tag{2.8}$$

then there exists a Lasso solution $\hat{\beta}_\lambda^{\mathrm{L}}$ with $\mathrm{sgn}(\hat{\beta}_\lambda^{\mathrm{L}}) = \mathrm{sgn}(\beta^0)$. As a partial converse, if there exists a Lasso solution $\hat{\beta}_\lambda^{\mathrm{L}}$ with $\mathrm{sgn}(\hat{\beta}_\lambda^{\mathrm{L}}) = \mathrm{sgn}(\beta^0)$, then $\|\Delta\|_\infty \leq 1$.

Remark 1. We can interpret $\|\Delta\|_\infty$ as the maximum in absolute value over $k \in N$ of the dot product of $\mathrm{sgn}(\beta_S^0)$ and $(X_S^T X_S)^{-1} X_S^T X_k$, the coefficient vector obtained by regressing X_k on X_S. The condition $\|\Delta\|_\infty \leq 1$ is known as the irrepresentable condition.

Proof. Fix $\lambda > 0$ and write $\hat{\beta} = \hat{\beta}^{\mathrm{L}}_{\lambda}$ and $\hat{S} = \{k : \hat{\beta}_k \neq 0\}$ for convenience. The KKT conditions for the Lasso give

$$\frac{1}{n} X^T X(\beta^0 - \hat{\beta}) = \lambda \hat{\nu}$$

where $\|\hat{\nu}\|_\infty \leq 1$ and $\hat{\nu}_{\hat{S}} = \mathrm{sgn}(\hat{\beta}_{\hat{S}})$. We can expand this into

$$\frac{1}{n} \begin{pmatrix} X_S^T X_S & X_S^T X_N \\ X_N^T X_S & X_N^T X_N \end{pmatrix} \begin{pmatrix} \beta_S^0 - \hat{\beta}_S \\ -\hat{\beta}_N \end{pmatrix} = \lambda \begin{pmatrix} \hat{\nu}_S \\ \hat{\nu}_N \end{pmatrix}. \tag{2.9}$$

We prove the converse first. If $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^0)$ then $\hat{\nu}_S = \mathrm{sgn}(\beta_S^0)$ and $\hat{\beta}_N = 0$. The top block of (2.9) gives

$$\beta_S^0 - \hat{\beta}_S = \lambda(\tfrac{1}{n} X_S^T X_S)^{-1} \mathrm{sgn}(\beta_S^0).$$

Substituting this into the bottom block, we get

$$\lambda \tfrac{1}{n} X_N^T X_S (\tfrac{1}{n} X_S^T X_S)^{-1} \mathrm{sgn}(\beta_S^0) = \lambda \hat{\nu}_N.$$

Thus as $\|\hat{\nu}_N\|_\infty \leq 1$, we have $\|\Delta\|_\infty \leq 1$.

For the positive statement, we need to find a $\hat{\beta}$ and $\hat{\nu}$ such that $\mathrm{sgn}(\hat{\beta}_S) = \mathrm{sgn}(\beta_S^0)$ and $\hat{\beta}_N = 0$, for which the KKT conditions hold. We claim that taking

$$\begin{aligned} (\hat{\beta}_S,\, \hat{\beta}_N) &= (\beta_S^0 - \lambda(\tfrac{1}{n} X_S^T X_S)^{-1} \mathrm{sgn}(\beta_S^0),\, 0) \\ (\hat{\nu}_S,\, \hat{\nu}_N) &= (\mathrm{sgn}(\beta_S^0),\, \Delta) \end{aligned}$$

satisfies (2.9). We only need to check that $\mathrm{sgn}(\beta_S^0) = \mathrm{sgn}(\hat{\beta}_S)$, but this follows from (2.8). □

2.2.6 Prediction and estimation

Consider once more the model $Y = X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbf{1}$ where the components of ε are independent mean-zero sub-Gaussian random variables with common parameter σ. Let S, s and N be defined as in the previous section. As we have noted before, in an artificial situation where S is known, we could apply OLS on 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 5. Given a matrix of predictors $X \in \mathbb{R}^{n \times p}$ and support set $S \neq \emptyset$, define the *compatibility factor*

$$\phi^2 = \inf_{\delta \in \mathbb{R}^p : \delta_S \neq 0,\, \|\delta_N\|_1 \leq 3\|\delta_S\|_1} \frac{\frac{1}{n}\|X\delta\|_2^2}{\frac{1}{s}\|\delta_S\|_1^2},$$

where $s = |S|$ and we take $\phi \geq 0$. The *compatibility condition* is that $\phi^2 > 0$.

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

$$\|\delta_S\|_1 = \mathrm{sgn}(\delta_S)^T \delta_S \leq \sqrt{s}\|\delta_S\|_2 \leq \sqrt{s}\|\delta\|_2.$$

Thus

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

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 δ can still allow ϕ^2 to be positive even in this case, as we discuss after the presentation of the theorem.

Theorem 22. *Suppose the compatibility condition holds and let $\lambda^* = A\sigma\sqrt{\log(p)/n}$ for $A > 2\sqrt{2}$. Then with probability at least $1 - 2p^{-(A^2/8-1)}$, we have that for all $\lambda \geq \lambda^*$,*

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta}^{\mathrm{L}}_\lambda)\|_2^2 + \lambda\|\beta^0 - \hat{\beta}^{\mathrm{L}}_\lambda\|_1 \leq \frac{16\lambda^2 s}{\phi^2}.$$

In particular, if $\tilde{\beta}$ is the Lasso estimate with $\lambda = \lambda^*$, then

$$\frac{1}{n}\|X(\beta^0 - \tilde{\beta})\|_2^2 \leq \frac{16A^2\log(p)}{\phi^2}\frac{\sigma^2 s}{n}, \qquad \textit{and} \qquad \|\beta^0 - \tilde{\beta}\|_1 \leq \frac{16A\sigma s}{\phi^2}\sqrt{\frac{\log p}{n}}.$$

Proof. Let us fix $\lambda \geq \lambda^*$, and write $\hat{\beta} = \hat{\beta}^{\mathrm{L}}_\lambda$. As in Theorem 10 we start with the "basic inequality":

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

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

$$\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 + 2\lambda\|\hat{\beta}\|_1 \leq \lambda\|\hat{\beta} - \beta^0\|_1 + 2\lambda\|\beta^0\|_1. \tag{2.10}$$

Lemma 14 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

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

If we could obtain

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

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

Returning to the actual proof, write $a = \|X(\hat{\beta} - \beta^0)\|_2^2/(n\lambda)$. Then from (2.10) we can derive the following string of inequalities:

$$\begin{aligned}
a + 2(\|\hat{\beta}_N\|_1 + \|\hat{\beta}_S\|_1) &\leq \|\hat{\beta}_S - \beta^0_S\|_1 + \|\hat{\beta}_N\|_1 + 2\|\beta^0_S\|_1 \\
a + \|\hat{\beta}_N\|_1 &\leq \|\hat{\beta}_S - \beta^0_S\|_1 + 2\|\beta^0_S\|_1 - 2\|\hat{\beta}_S\|_1 \\
a + \|\hat{\beta}_N - \beta^0_N\|_1 &\leq 3\|\beta^0_S - \hat{\beta}_S\|_1 \\
a + \|\hat{\beta} - \beta^0\|_1 &\leq 4\|\beta^0_S - \hat{\beta}_S\|_1,
\end{aligned}$$

the final inequality coming from adding $\|\beta_S^0 - \hat{\beta}_S\|_1$ to both sides.

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

$$\begin{aligned}\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 + \lambda\|\beta^0 - \hat{\beta}\|_1 &\leq 4\lambda\|\beta_S^0 - \hat{\beta}_S\|_1 \\ &\leq \frac{4\lambda}{\phi}\sqrt{\frac{s}{n}}\|X(\hat{\beta} - \beta^0)\|_2. \end{aligned} \tag{2.11}$$

From this we get

$$\frac{1}{\sqrt{n}}\|X(\hat{\beta} - \beta^0)\|_2 \leq \frac{4\lambda\sqrt{s}}{\phi},$$

and substituting this into the RHS of (2.11) gives the result. □

2.2.7 The compatibility condition

How strong is the compatibility condition? In order to answer this question, we shall think of X as random and try to understand what conditions on the population covariance matrix $\Sigma^0 := \mathbb{E}(X^T X/n)$ imply that X satisfies a compatibility condition with high probability. To this end let us define

$$\phi_\Sigma^2(S) = \inf_{\delta:\|\delta_S\|_1 \neq 0, \|\delta_N\|_1 \leq 3\|\delta_S\|_1} \frac{\delta^T\Sigma\delta}{\|\delta_S\|_1^2/|S|} = |S| \inf_{\delta:\|\delta_S\|_1=1, \|\delta_N\|_1 \leq 3} \delta^T\Sigma\delta,$$

where $\Sigma \in \mathbb{R}^{p\times p}$. Note then our $\phi^2 = \phi_{\hat{\Sigma}}^2(S)$ where $\hat{\Sigma} := X^T X/n$ and S is the support set of β^0. The following result shows that if Σ is close to a matrix Θ for which $\phi_\Theta^2(S) > 0$, then also $\phi_\Sigma^2(S) > 0$.

Lemma 23. *Suppose* $\phi_\Theta^2(S) > 0$ *and* $\max_{jk}|\Sigma_{jk} - \Theta_{jk}| \leq \phi_\Theta^2(S)/(32|S|)$. *Then* $\phi_\Sigma^2(S) \geq \phi_\Theta^2(S)/2$.

Proof. In the following we suppress dependence on S and write $s := |S|$. Let $\mathcal{B} := \{\delta : \|\delta_S\|_1 = 1, \|\delta_N\|_1 \leq 3\}$. Take $\delta \in \mathcal{B}$. The we have

$$s\delta^T\Sigma\delta = s\delta^T\Theta\delta - s\delta^T(\Theta - \Sigma)\delta \geq \phi_\Theta^2 - s|\delta^T(\Sigma - \Theta)\delta|.$$

Furthermore,

$$\begin{aligned}|\delta^T(\Theta - \Sigma)\delta| &\leq \|\delta\|_1\|(\Theta - \Sigma)\delta\|_\infty \ \text{(Hölder)} \\ &\leq \frac{\phi_\Theta^2}{32s}\|\delta\|_1^2 \ \text{(Hölder again)}\end{aligned}$$

and $\|\delta\|_1 = \|\delta_N\|_1 + \|\delta_S\|_1 \leq 4$. Thus

$$s\delta^T\Sigma\delta \geq \phi_\Theta^2 - \phi_\Theta^2/2 = \phi_\Theta^2/2.$$

Taking infima over $\delta \in \mathcal{B}$ gives the result. □

We would like to apply the result above with $\Theta = \Sigma^0$, and use it to argue that if Σ^0 satisfies the compatibility condition, then so will $\hat{\Sigma}$ with high probability. In order to do this, we need to argue that the event that $\max_{jk}|\hat{\Sigma}_{jk} - \Sigma_{jk}^0|$ is small occurs with high probability. We can obtain such a result with the aid of concentration inequalities.

2.2.8 Concentration inequalities II

When trying to understand the concentration properties of $\hat{\Sigma}_{jk}$, it will be helpful to have a tail bound for a product of sub-Gaussian random variables. Bernstein's inequality, which applies to random variables satisfying the condition below, is helpful in this regard.

Definition 6 (Bernstein's condition)**.** We say that the random variable W satisfies Bernstein's condition with parameter (σ, b) where $\sigma, b > 0$ if

$$\mathbb{E}(|W - \mathbb{E}W|^k) \leq \frac{1}{2}k!\sigma^2 b^{k-2} \qquad \text{for } k = 2, 3, \ldots.$$

Proposition 24 (Bernstein's inequality)**.** *Let $W_1, W_2, \ldots$ be independent random variables with $\mathbb{E}(W_i) = \mu$. Suppose each W_i satisfies Bernstein's condition with parameter (σ, b). Then*

$$\mathbb{E}(e^{\alpha(W_i - \mu)}) \leq \exp\left(\frac{\alpha^2\sigma^2/2}{1 - b|\alpha|}\right) \qquad \textit{for all } |\alpha| < 1/b$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} W_i - \mu \geq t\right) \leq \exp\left(-\frac{nt^2}{2(\sigma^2 + bt)}\right) \qquad \textit{for all } t > 0.$$

Proof. Fix i and let $W = W_i$. We have

$$\begin{aligned}
\mathbb{E}(e^{\alpha(W-\mu)}) &= \mathbb{E}\left(1 + \alpha(W - \mu) + \sum_{k=2}^{\infty} \frac{\alpha^k (W - \mu)^k}{k!}\right) \\
&\leq \mathbb{E}\left(1 + \sum_{k=2}^{\infty} \frac{|\alpha|^k |W - \mu|^k}{k!}\right) \\
&= 1 + \sum_{k=2}^{\infty} |\alpha|^k \frac{\mathbb{E}(|W - \mu|^k)}{k!} \\
&\leq 1 + \frac{\sigma^2\alpha^2}{2} \sum_{k=2}^{\infty} |\alpha|^{k-2} b^{k-2} \\
&= 1 + \frac{\sigma^2\alpha^2}{2} \frac{1}{1 - |\alpha|b} \leq \exp\left(\frac{\alpha^2\sigma^2/2}{1 - b|\alpha|}\right),
\end{aligned}$$

provided $|\alpha| < 1/b$ and using the inequality $e^u \geq 1 + u$ in the final line. For the probability bound, first note that

$$\begin{aligned}
\mathbb{E}\exp\left(\sum_{i=1}^{n} \alpha(W_i - \mu)/n\right) &= \prod_{i=1}^{n} \mathbb{E}\exp\{\alpha(W_i - \mu)/n\} \\
&\leq \exp\left(n\frac{(\alpha/n)^2\sigma^2/2}{1 - b|\alpha/n|}\right)
\end{aligned}$$

for $|\alpha|/n < 1/b$. Then we use the Chernoff method, though without minimising over $\alpha > 0$: instead we set $\alpha/n = t/(bt + \sigma^2) \in (0, 1/b)$. □

Lemma 25. *Let W, Z be mean-zero and sub-Gaussian with parameters σ_W and σ_Z respectively. Then the product WZ satisfies Bernstein's condition with parameter $(8\sigma_W\sigma_Z, 4\sigma_W\sigma_Z)$.*

Proof. In order to use Bernstein's inequality (Proposition 24) we first obtain bounds on the moments of W and Z. Note that $W^{2k} = \int_0^\infty \mathbb{1}_{\{x<W^{2k}\}}dx$. Thus by Fubini's theorem

$$\begin{aligned}
\mathbb{E}(W^{2k}) &= \int_0^\infty \mathbb{P}(W^{2k} > x)dx \\
&= 2k\int_0^\infty t^{2k-1}\mathbb{P}(|W| > t)dt \qquad \text{substituting } t^{2k} = x \\
&\leq 4k\int_0^\infty t^{2k-1}\exp\{-t^2/(2\sigma_W^2)\}dt \qquad \text{by Proposition 11} \\
&= 4k\sigma_W^2\int_0^\infty (2\sigma_W^2 x)^{k-1}e^{-x}dx \qquad \text{substituting } t^2/(2\sigma_W^2) = x \\
&= 2^{k+1}\sigma_W^{2k}k!.
\end{aligned}$$

Next note that for any random variable Y,

$$\begin{aligned}
\mathbb{E}|Y - \mathbb{E}Y|^k &= 2^k\mathbb{E}|Y/2 - \mathbb{E}Y/2|^k \\
&\leq 2^{k-1}(\mathbb{E}|Y|^k + |\mathbb{E}Y|^k) \qquad \text{by Jensen's inequality applied to } t \mapsto |t|^k, \\
&\leq 2^k\mathbb{E}|Y|^k.
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{E}(|WZ - \mathbb{E}WZ|^k) &\leq 2^k\mathbb{E}|WZ|^k \\
&\leq 2^k(\mathbb{E}W^{2k})^{1/2}(\mathbb{E}Z^{2k})^{1/2} \qquad \text{by Cauchy–Schwarz} \\
&\leq 2^k 2^{k+1}\sigma_W^k\sigma_Z^k k! \\
&= \frac{k!}{2}(8\sigma_W\sigma_Z)^2(4\sigma_W\sigma_Z)^{k-2}.
\end{aligned}$$

□

2.2.9 Random design

We now show that we can expect the compatibility condition to hold with high probability. To make the result more readily interpretable, we shall state it in an asymptotic framework. Imagine a sequence of design matrices with n and p growing, each with their own compatibility condition. We will however suppress the asymptotic regime in the notation.

Theorem 26. *Suppose the rows of X are i.i.d. and each entry of X is mean-zero sub-Gaussian with parameter v. Let $\hat{\Sigma} := X^TX/n$ and $\Sigma^0 := \mathbb{E}(\hat{\Sigma})$. Suppose $s\sqrt{\log(p)/n} \to 0$ (and $s, p, n > 1$) as $n \to \infty$. Let*

$$\begin{aligned}
\phi^2_{\hat{\Sigma},s} &= \min_{S:|S|=s} \phi^2_{\hat{\Sigma}}(S) \\
\phi^2_{\Sigma^0,s} &= \min_{S:|S|=s} \phi^2_{\Sigma^0}(S),
\end{aligned}$$

and suppose $\phi^2_{\Sigma^0,s} > c$ for a constant $c > 0$. Then $\mathbb{P}(\phi^2_{\hat{\Sigma},s} \geq \phi^2_{\Sigma^0,s}/2) \to 1$ as $n \to \infty$.

Proof. In view of Lemma 23, we need only show that

$$\mathbb{P}(\max_{jk} |\hat{\Sigma}_{jk} - \Sigma^0_{jk}| \geq \phi^2_{\Sigma^0,s}/(32s)) \to 0.$$

Let $t = \phi^2_{\Sigma^0,s}/(32s)$. By a union bound and then Lemma 25 we have

$$\begin{aligned}
\mathbb{P}(\max_{jk} |\hat{\Sigma}_{jk} - \Sigma^0_{jk}| \geq \phi^2_{\Sigma^0,s}/(32s)) &< p^2 \max_{jk} \mathbb{P}\Big(\Big| \sum_{i=1}^n X_{ij}X_{ik}/n - \Sigma^0_{jk} \Big| \geq t \Big) \\
&\leq 2\exp\left(-\frac{nt^2}{2(64v^4 + 4v^2 t))} + 2\log p \right) \\
&\leq 2\exp(-c'n/s^2 + 2\log p) \qquad (2.12)
\end{aligned}$$

for a constant $c' > 0$. To justify the last line, observe that any constant $C > 0$

$$\frac{(\phi^2_{\Sigma^0,s})^2}{C + \phi^2_{\Sigma^0,s}/s} \geq \min_{u \geq c} \frac{u^2}{C+u} = \min_{u \geq c} u\left(1 - \frac{C}{u+C}\right) = c\left(1 - \frac{C}{c+C}\right) > 0.$$

Thus returning to (2.12), we see that this tends to zero as $\log p = o(n/s^2)$, which completes the proof. □

Corollary 27. *Consider the setup of Theorem 26 and suppose additionally that the rows of X are independent with distribution $N_p(0, \Sigma^0)$. Suppose the diagonal entries of Σ^0 are bounded above and the minimum eigenvalue of Σ^0, $c_{\min}$ is bounded away from 0. Then $\mathbb{P}(\phi^2_{\hat{\Sigma},s} \geq c_{\min}/2) \to 1$.*

2.2.10 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

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

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 $x^{(0)}$ (e.g. $x^{(0)} = 0$) 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 $A_0 = \{x : f(x) \leq f(x^{(0)})\}$ is compact, then every converging subsequence of $x^{(m)}$ will converge to a minimiser of f.

Corollary 28. *Suppose A^0 is compact. Then*

(i) There exists a minimiser of f, x^* and $f(x^{(m)}) \to f(x^*)$.

(ii) If x^* is the unique minimiser of f then $x^{(m)} \to x^*$.

**Proof*.* Function f is continuous and so attains its infimum on the compact set A_0. Suppose $f(x^{(m)}) \nrightarrow f(x^*)$. Then there exists $\epsilon > 0$ and a subsequence $(x^{(m_j)})_{j=0}^{\infty}$ such that $f(x^{(m_j)}) \geq f(x^*) + \epsilon$ for all j. Note that since $f(x^{(m)}) \leq f(x^{(m-1)})$, we know that $x^{(m)} \in A_0$ for all m. Thus if A_0 is compact then any subsequence of $(x^{(m)})_{m=0}^{\infty}$ has a further subsequence that converges by the Bolzano–Weierstrass theorem. Let $\tilde{x}$ be the limit of the converging subsequence of $(x^{(m_j)})_{j=0}^{\infty}$. Then $f(\tilde{x}) \geq f(x^*) + \epsilon$, contradicting the result of Tseng [2001]. Thus (i) holds. The proof of (ii) is similar. □

We can replace individual coordinates by blocks of coordinates and the same result holds. That is if $x = (x_1, \ldots, x_B)$ where now $x_b \in \mathbb{R}^{d_b}$ and

$$f(x) = g(x) + \sum_{b=1}^{B} h_b(x_b)$$

with g convex and differentiable and each $h_b : \mathbb{R}^{d_b} \to \mathbb{R}$ convex, then block coordinate descent can be used.

One of the reasons that coordinate descent is so effective for solving the Lasso is that the coordinatewise optimisations are very simple and have closed form solutions. To see this, suppose at the mth iteration we are optimising for variable k. Let us write

$$R := Y - \sum_{j=1}^{k-1} X_j \hat{\beta}_j^{(m)} - \sum_{j=k+1}^{p} X_j \hat{\beta}_j^{(m-1)}.$$

We have that

$$\hat{\beta}_k^{(m)} = \arg\min_{\beta \in \mathbb{R}} \left\{ \frac{1}{2n} \|R - \beta X_k\|_2^2 + \lambda |\beta| \right\}$$

A minimiser $\hat{\beta}_k^{(m)}$ is characterised by the subgradient optimality condition:

$$-\frac{1}{n} X_k^T R + \hat{\beta}_k^{(m)} + \lambda \hat{\nu} = 0,$$

where $\hat{\nu} \in [-1, 1]$ and if $\hat{\beta}_k^{(m)} \neq 0$, $\hat{\nu} = \operatorname{sgn}(\hat{\beta}_k^{(m)})$. Rearranging, we have

$$\hat{\beta}_k^{(m)} = \frac{1}{n} X_k^T R - \lambda \hat{\nu},$$

and we may check that this is satisfied by

$$\hat{\beta}_k^{(m)} = S_\lambda(X_k^T R/n),$$

where $S_t(u) := \text{sgn}(u)(|u| - t)_+$ is the *soft-thresholding* operator. Note that $\hat{\beta}_k^{(m)}$ is the unique minimiser as the coordinatewise objective is strictly convex.

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 $A_l = \{k : \hat{\beta}^{\text{L}}_{\lambda_{l-1},k} \neq 0\}$.
2. Perform coordinate descent only on coordinates in A_l obtaining a solution $\hat{\beta}$ (all components $\hat{\beta}_k$ with $k \notin A_l$ are set to zero).
3. Let $V = \{k : |X_k^T(Y - X\hat{\beta})|/n > \lambda_l\}$, the set of coordinates which violate the KKT conditions when $\hat{\beta}$ is taken as a candidate solution.
4. If V is empty, we set $\hat{\beta}^{\text{L}}_{\lambda_l} = \hat{\beta}$. Else we update $A_l = A_l \cup V$ and return to 2.

2.3 Extensions of the Lasso

2.3.1 The square-root Lasso

Consider the normal linear model

$$Y = X\beta^0 + \varepsilon, \tag{2.13}$$

where $\varepsilon \sim N_n(0, \sigma^2 I)$. A misgiving one might have about the theoretical results on the Lasso is that they rely on knowledge of the unknown σ in that the λ concerned takes the form $A\sigma\sqrt{\log(p)/n}$.

Now given a Lasso estimate $\hat{\beta}^{\text{L}}_\lambda$ for β^0, a sensible estimate of σ is given by

$$\hat{\sigma}^{\text{L}}_\lambda := \frac{1}{\sqrt{n}}\|Y - X\hat{\beta}^{\text{L}}_\lambda\|_2.$$

Indeed, this is supported by the following general result.

Lemma 29. *Suppose $Y = f + \varepsilon$ where error $\varepsilon \in \mathbb{R}^n$ is such that $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with $\mathbb{E}\varepsilon_1 = 0$ and $\text{Var}\varepsilon_1 = \sigma^2 > 0$. Consider an asymptotic regime where n increases, and suppose $\hat{f} \in \mathbb{R}^n$ is an estimate of f such that there exists sequence $a_n \to 0$ with*

$$\mathbb{P}(\|\hat{f} - f\|_2^2/n \leq a_n) \to 1.$$

Then setting

$$\hat{\sigma} := \frac{1}{\sqrt{n}}\|Y - \hat{f}\|_2,$$

there exists sequence $b_n \to 0$ *such that*

$$\mathbb{P}\left(\left|\frac{\sigma}{\hat{\sigma}} - 1\right| \leq b_n\right) \to 1.$$

Given this estimate, a sensible tuning parameter to choose for the Lasso would be $A\hat{\sigma}^{\mathrm{L}}_{\lambda}\sqrt{\log(p)/n}$, which would lead to a new estimate of β^0, which could then in turn give a new estimate of σ. Iterating this process amounts to performing a block coordinate descent optimisation (alternating between optimising over β and σ) of the following convex objective function,

$$Q^{\mathrm{sq}}_{\gamma}(\beta, \sigma) := \frac{1}{2n\sigma}\|Y - X\beta\|_2^2 + \frac{\sigma}{2} + \gamma\|\beta\|_1,$$

with $\gamma = A\sqrt{\log(p)/n}$ and initial value for σ given by $\hat{\sigma}^{\mathrm{L}}_{\lambda}$. Corollary 28 indicates that this will lead to a minimiser of Q^{sq}_{γ}. However, a more direct route to the minimiser is offered by the so-called *square-root Lasso* $\hat{\beta}^{\mathrm{sq}}_{\gamma}$ [Belloni et al., 2011, Sun and Zhang, 2012], which minimises

$$\frac{1}{\sqrt{n}}\|Y - X\beta\|_2 + \gamma\|\beta\|_1. \tag{2.14}$$

Note that the display above is equal to $\min_{\sigma>0} Q^{\mathrm{sq}}_{\gamma}(\beta, \sigma)$ provided $Y \neq X\beta$.

The square-root Lasso is not really a new estimator for β^0, but rather a particular reparametrisation of the Lasso solution path, as may be deduced by comparing its KKT conditions with those of the (regular) Lasso. Let us write $\hat{\sigma}^{\mathrm{sq}}_{\gamma} := \|Y - X\hat{\beta}^{\mathrm{sq}}_{\gamma}\|_2/\sqrt{n}$. We have, provided $\hat{\sigma}^{\mathrm{sq}}_{\gamma} > 0$, that

$$\frac{1}{n\hat{\sigma}^{\mathrm{sq}}_{\gamma}}X^T(Y - X\hat{\beta}^{\mathrm{sq}}_{\gamma}) = \gamma\hat{\nu}^{\mathrm{sq}}_{\gamma}, \qquad \frac{1}{n}X^T(Y - X\hat{\beta}^{\mathrm{L}}_{\lambda}) = \lambda\hat{\nu}^{\mathrm{L}}_{\lambda},$$

where $\|\hat{\nu}^{\mathrm{sq}}_{\gamma}\|_\infty \leq 1$, and $\hat{\nu}^{\mathrm{sq}}_{\gamma}$ agrees in sign with $\hat{\beta}^{\mathrm{sq}}_{\gamma}$ on its active set (and similarly for $\hat{\nu}^{\mathrm{L}}_{\lambda}$). Thus any Lasso solution $\hat{\beta}^{\mathrm{L}}_{\lambda}$ is a square-root Lasso solution $\hat{\beta}^{\mathrm{sq}}_{\gamma}$ with $\gamma = \lambda/\hat{\sigma}^{\mathrm{L}}_{\lambda}$, provided $\hat{\sigma}^{\mathrm{L}}_{\lambda} > 0$. Conversely, any $\hat{\beta}^{\mathrm{sq}}_{\gamma}$ is equal to a Lasso solution $\hat{\beta}^{\mathrm{L}}_{\lambda}$ with $\lambda = \gamma\hat{\sigma}^{\mathrm{sq}}_{\gamma}$, provided $\hat{\sigma}^{\mathrm{sq}}_{\gamma} > 0$.

The next result formalises the way in which the square-root Lasso does not rely on knowledge of σ.

Theorem 30. *Consider the normal linear model* (2.13) *and let* $\hat{\beta}$ *be a square-root Lasso estimate with* $\gamma = B\sqrt{\log(p)/n}$ *where* $B > 2\sqrt{2}$. *Consider an asymptotic regime where* $n, p \to \infty$, $s\log(p)/n \to 0$ *and the compatibility factor* ϕ^2 *is bounded away from* 0. *Then with probability tending to 1,*

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq \frac{17B^2\log p}{\phi^2}\frac{s\sigma^2}{n} \qquad \textit{and} \qquad \|\beta^0 - \hat{\beta}\|_1 \leq \frac{17B\sigma s}{\phi^2}\sqrt{\frac{\log p}{n}}.$$

Proof. We would like to argue that $\hat{\beta}$ is a Lasso solution with $\lambda \approx A\sigma\sqrt{\log(p)/n}$, and then apply Theorem 22. This relies on $\hat{\sigma} := \hat{\sigma}^{\mathrm{sq}}_{\gamma}$ satisfying $\hat{\sigma} \approx \sigma$. For this, we might like to apply Lemma 29, but this requires $\hat{\beta}$ having good prediction properties, which is among the properties we are trying to show!

To circumvent these difficulties, we will find Lasso solutions with tuning parameters $\lambda_1 < \lambda_2$ for which the corresponding square-root Lasso tuning parameters satisfy

$$\gamma_1 := \frac{\lambda_1}{\hat{\sigma}^{\mathrm{L}}_{\lambda_1}} \leq \gamma \leq \frac{\lambda_2}{\hat{\sigma}^{\mathrm{L}}_{\lambda_2}} =: \gamma_2.$$

Given such λ_1 and λ_2, we claim that

$$\lambda_1 \leq \hat{\sigma}\gamma \leq \lambda_2, \tag{2.15}$$

so $\hat{\beta}$ is a Lasso solution with a tuning parameter sandwiched between λ_1 and λ_2. Indeed, $\lambda_1 = \gamma_1\hat{\sigma}^{\mathrm{sq}}_{\gamma_1}$ as $\hat{\sigma}^{\mathrm{sq}}_{\gamma_1} = \hat{\sigma}^{\mathrm{L}}_{\lambda_1}$, and since $\gamma_1 \leq \gamma$ by assumption, we have $\hat{\sigma}^{\mathrm{sq}}_{\gamma_1} \leq \hat{\sigma}$. To show the final point, observe that writing $\tilde{\beta} := \hat{\beta}^{\mathrm{L}}_{\lambda_1} = \hat{\beta}^{\mathrm{sq}}_{\gamma_1}$,

$$\begin{aligned} \frac{1}{\sqrt{n}}\|Y - X\tilde{\beta}\|_2 + \gamma_1\|\tilde{\beta}\|_1 &\leq \frac{1}{\sqrt{n}}\|Y - X\hat{\beta}\|_2 + \gamma_1\|\hat{\beta}\|_1 \\ \frac{1}{\sqrt{n}}\|Y - X\hat{\beta}\|_2 + \gamma\|\hat{\beta}\|_1 &\leq \frac{1}{\sqrt{n}}\|Y - X\tilde{\beta}\|_2 + \gamma\|\tilde{\beta}\|_1. \end{aligned} \tag{2.16}$$

These imply

$$\gamma_1\|\tilde{\beta}\|_1 + \gamma\|\hat{\beta}\|_1 \leq \gamma_1\|\hat{\beta}\|_1 + \gamma\|\tilde{\beta}\|_1,$$

so

$$(\gamma - \gamma_1)(\|\tilde{\beta}\|_1 - \|\hat{\beta}\|_1) \geq 0$$

from which we deduce that $\|\tilde{\beta}\|_1 \geq \|\hat{\beta}\|_1$. Substituting this in (2.16), we see that $\hat{\sigma}^{\mathrm{sq}}_{\gamma_1} \leq \hat{\sigma}$ as required. Thus $\lambda_1 \leq \hat{\sigma}\gamma$, and similarly $\hat{\sigma}\gamma \leq \lambda_2$.

Now let us take $\lambda_j = \sigma A_j\sqrt{\log(p)/n}$ for $2\sqrt{2} < A_1 < B < A_2$ where $17B^2 \geq 16A_2^2$ (which we note also implies $17B \geq 16A$). Now we know that (see example sheet) there exists a sequence $a_n \to 0$ such that on a sequence of events $\Omega^{(1)}_n$ with $\mathbb{P}(\Omega^{(1)}_n) \to 1$,

$$1 - a_n \leq \frac{\sigma}{\hat{\sigma}^{\mathrm{L}}_{\lambda_j}} \leq 1 + a_n,$$

for $j = 1, 2$. Thus on $\Omega^{(1)}_n$,

$$\gamma_1 \leq (1 + a_n)A_1\sqrt{\frac{\log p}{n}} < \gamma \qquad \text{and} \qquad \gamma_2 \geq (1 - a_n)A_2\sqrt{\frac{\log p}{n}} > \gamma,$$

for n sufficiently large, and so (2.15) holds for such n. Applying Theorem 22 with $\lambda^* = \lambda_1$, we see that on a sequence of events $\Omega^{(2)}_n$ with $\mathbb{P}(\Omega^{(2)}_n) \to 1$, we have

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta}^{\mathrm{L}}_{\lambda})\|_2^2 + \lambda\|\beta^0 - \hat{\beta}^{\mathrm{L}}_{\lambda}\|_1 \leq \frac{16s\lambda^2}{\phi^2}$$

for all $\lambda \geq \lambda_1$. Then on $\Omega_n^{(1)} \cap \Omega_n^{(2)}$ (which has $\mathbb{P}(\Omega_n^{(1)} \cap \Omega_n^{(2)}) \to 1$), additionally (2.15) holds, and so

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 + \hat{\sigma}\gamma\|\beta^0 - \hat{\beta}\|_1 \leq \frac{16s\hat{\sigma}^2\gamma^2}{\phi^2}.$$

The result follows from noting that on $\Omega_n^{(1)}$, $\hat{\sigma}\gamma \leq \lambda_2 \leq 17\sigma B\sqrt{\log(p)/n}/16$ and $\lambda_2^2 \leq 17\sigma^2 B^2 \log(p)/(16n)$. □

2.3.2 Other loss functions

We can add an ℓ_1 penalty to many other log-likelihoods, or more generally other loss functions besides the squared-error loss that arises from the normal linear model. For Lasso-penalised generalised linear models, such as logistic regression, similar theoretical results to those we have obtained are available and computations can proceed in a similar fashion to above.

2.3.3 Structural penalties

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.

Group Lasso

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

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

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 $\hat{\beta}_G = 0$ 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.

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

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

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 $\mu \in \mathbb{R}^n$, the following objective

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

2.3.4 Reducing 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 $\hat{S}_\lambda = \{k : \hat{\beta}^{\mathrm{L}}_{\lambda,k} \neq 0\}$ and then re-estimate $\beta^0_{\hat{S}_\lambda}$ by OLS regression on $X_{\hat{S}_\lambda}$.

Another option is to re-estimate using the Lasso on $X_{\hat{S}_\lambda}$; this procedure is known as the *relaxed Lasso* [Meinshausen, 2007]. The *adaptive Lasso* [Zou, 2006] takes an initial estimate of β^0, $\hat{\beta}^{\mathrm{init}}$ (e.g. from the Lasso) and then performs weighted Lasso regression:

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

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

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

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

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.

Chapter 3

Graphical models

So far we have considered the problem of relating a particular response to a potentially 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 our measured variables. One simple way of formalising the relatedness between variables is to measure their correlation, or (marginal) dependence. However, this does not always lead to the most interpretable results as many pairs of variables may exhibit dependence with one another without there being a very meaningful relationship between them.

3.1 Conditional independence

Conditional dependence is often a better property on which to base our desired notion of relatedness, and is defined as follows.

Definition 7. If X, Y and Z are random vectors with a joint density f_{XYZ} (w.r.t. a product measure μ) then we say X is *conditionally independent* of Y given Z, and write

$$X \perp\!\!\!\perp Y \mid Z$$

if

$$f_{XY|Z}(x, y|z) = f_{X|Z}(x|z) f_{Y|Z}(y|z) \qquad \text{whenever } f_Z(z) > 0,$$

and if not we say X and Y are *conditionally dependent* given Z and write $X \not\perp\!\!\!\perp Y \mid Z$. Equivalently,

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

for some (integrable) function m whenever $f_{YZ}(y, z) > 0$, and moreover this m will then be the conditional density $f_{X|Z}$.

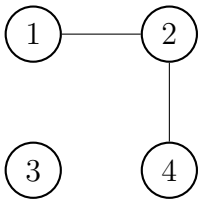
The interpretation of the conditional independence $X \perp\!\!\!\perp Y|Z$ is that 'knowing Z renders X irrelevant for learning about Y'.

3.1.1 Conditional independence graphs

It is convenient to represent the conditional independence relationships between variables through *graphs*.

Definition 8. An *undirected graph* is a pair $\mathcal{G} = (V, E)$ where V is a set of *vertices* or *nodes* and $E \subseteq \{\{j, k\} : j \neq k \text{ and } j, k \in V\}$ is a set of edges.

Example 3.1.1. We can represent the undirected graph on 4 vertices with edge set $E = \{\{1, 2\}, \{2, 4\}\}$ as follows.



Let $Z = (Z_1, \ldots, Z_p)^T$ be a collection of random variables. The graphs we consider will always have $V = \{1, \ldots, p\}$ so V indexes the random variables. We use the following notation: $-k$ and $-jk$ when in subscripts denote the sets $\{1, \ldots, p\} \setminus \{k\}$ and $\{1, \ldots, p\} \setminus \{j, k\}$ respectively.

Definition 9. The *conditional independence graph* (CIG) of a distribution P on $\mathbb{R}^p$ is the graph $\mathcal{G} = (V, E)$ where given $Z \sim P$,

$$\{j, k\} \in E \iff Z_j \not\perp\!\!\!\perp Z_k \,|\, Z_{-jk}.$$

3.1.2 *Conditional independence graphs and causality*

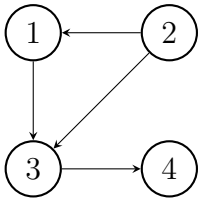
In the following subsection, which is *non-examinable*, we explain how edges in a conditional independence graph may be thought of as having a causal basis; in order to do this, we will need to introduce some further concepts.

Definition 10. A *graph* is a pair $\mathcal{G} = (V, E)$ where $E \subseteq \{(j, k) : j \neq k \text{ and } j, k \in V\}$. Let $j, k \in V$.

If $(j, k) \in E$, we write $j \to k$.

- If $(j, k) \in E$ and $(j, k) \notin E$, we say edge (j, k) is *directed* and write $j \to k$. A graph where all edges are directed is called a *directed graph*.
- A *directed cycle* is a collection of vertices $j_1, \ldots, j_m$ with $j_1 \to j_2, \ldots, j_{m-1} \to j_m, j_m \to j_1$.
- A *directed acyclic graph (DAG)* is a directed graph with no cycles.
- Say j is a *parent* of k and k is a *child* of j if $j \to k$. The set of parents of k will be denoted $\text{pa}(k)$.

Example 3.1.2. We can represent the directed graph on 4 vertices with $E = \{(1,3),(2,1),(3,4),(2,3)\}$ as follows.



Here $\text{pa}(3) = \{1,2\}$. As there are no directed cycles, this is a DAG.

We will make use of the following property of DAGs.

Proposition 31. *Given a DAG $\mathcal{G}$ with vertex set $V = \{1,\dots,p\}$, we say that a permutation π of V is a* topological (or causal) ordering *of the variables if for all $j \in V$,*

$$\text{if } k \in \text{pa}(j) \text{ then } \pi(k) < \pi(j)$$

Every DAG has a topological ordering.

Proof. We use induction on the number of nodes p. Clearly the result is true when $p = 1$.

Now we show that in any DAG, we can find a node with no parents. Pick any node and move to one of its parents, if possible. Then move to one of the new node's parents, and continue in this fashion. This procedure must terminate since no node can be visited twice, or we would have found a cycle. The final node we visit must therefore have no parents, which we call a source node.

Suppose then that $p \geq 2$, and we know that all DAGs with $p-1$ nodes have a topological ordering. Find a source s (wlog $s = p$) and form a new DAG $\tilde{\mathcal{G}}$ with $p-1$ nodes by removing the source (and all edges emanating from it). Note we keep the labelling of the nodes in this new DAG the same. This smaller DAG must have a topological order $\tilde{\pi}$. A topological ordering π for our original DAG is then given by $\pi(s) = 1$ and $\pi(k) = \tilde{\pi}(k)+1$ for $k \neq s$. □

Having established the notion of a DAG, we can introduce structural equation models, which provide a formalism for understanding causal relationships.

Definition 11. A *structural equation model (SEM)* $\mathcal{S}$ for a distribution P on $\mathbb{R}^p$ is a collection of p equations where random vector $Z \in \mathbb{R}^p$ defined via

$$Z_k = h_k(Z_{P_k}, \varepsilon_k), \qquad k = 1,\dots,p \tag{3.1}$$

is such that $Z \sim P$. Here

- $\varepsilon_1,\dots,\varepsilon_p$ are all independent random variables;
- $P_k \subseteq \{1,\dots,p\} \setminus \{k\}$ are such that the graph with edges given by P_k being $\text{pa}(k)$ is a DAG; if $P_k = \emptyset$, then the RHS of (3.1) is understood to be $h_k(\varepsilon_k)$.

Note that the condition that the graph concerned is a DAG ensures that Z is well-defined given ε. Indeed, one can generate the components of Z according to the topological ordering π of the DAG: assuming without loss of generality that $\pi^{-1}(j) = j$ for all j, we would generate Z via

$$Z_1 = h_1(\varepsilon_1), Z_2 = h_2(Z_{\mathrm{pa}(2)}, \varepsilon_2), \ldots, Z_p = h_p(Z_{\mathrm{pa}(p)}, \varepsilon_p).$$

Note that, for each j, $Z_{\mathrm{pa}(j)}$ will be generated before Z_j. In this way the SEM (3.1) is a recipe for for generating draws from the distribution P. The significance of this SEM is that were it to represent the true mechanism by which data with distribution P were generated, we can reason about how the distribution might change if we intervene on the system by, for example, setting a particular variable k to a given value a. We can expect that the new distribution will be determined by the modified SEM where the equation $Z_k = h_k(Z_{P_k}, \varepsilon_k)$ is replaced by $Z_k = a$, with the remaining equations unchanged. In this sense, an SEM can encode causal relationships between variables.

In general, there may be many SEMs that give rise to a given joint distribution P. However, we do have the following result connecting conditional independence graphs and structural equation models which formalises the sense in which the edges of a conditional independence graph may be thought of as having some causal basis.

Proposition 32. *Let $Z \in \mathbb{R}^p$ have density f. If $Z_j \not\perp\!\!\!\perp Z_k \,|\, Z_{-jk}$, then the DAG associated with any SEM for f must have that either $j \to k$, $k \to j$, or there exists $v \in \{1, \ldots, p\} \setminus \{j, k\}$ with $j \to v \leftarrow k$.*

Proof. Suppose we do not have any of $j \to k$, $k \to j$, or $j \to v \leftarrow k$ for some $v \in \{1, \ldots, p\} \setminus \{j, k\}$, so $k \notin \mathrm{pa}(j)$, $j \notin \mathrm{pa}(k)$ and $\mathrm{pa}(j) \cap \mathrm{pa}(k) = \emptyset$. Without loss of generality, suppose that for a topological ordering π we have $\pi^{-1}(l) = l$ for all l. Then

$$\begin{aligned} f(z_1, \ldots, z_p) &= f_{Z_1}(z_1) \prod_{l=2}^{p} f_{Z_l \,|\, Z_1, \ldots, Z_{l-1}}(z_l \,|\, z_1, \ldots, z_{l-1}) \\ &= f_{Z_1}(z_1) \prod_{l=2}^{p} f_{Z_l \,|\, Z_{\mathrm{pa}(l)}}(z_l \,|\, z_{\mathrm{pa}(l)}), \end{aligned}$$

as $1, \ldots, l-1$ come before l in the topological order. Now note that the RHS may be expressed as $g(z_j, z_{-jk}) \times h(z_k, z_{-jk})$ for functions g, h as for any l, none of the arguments of $f_{Z_l \,|\, Z_{\mathrm{pa}(l)}}(z_l \,|\, z_{\mathrm{pa}(l)})$ involve both z_j and z_k. Thus $Z_j \perp\!\!\!\perp Z_k \,|\, Z_{-jk}$ (see example sheet) as required. □

3.2 Gaussian graphical models

We have seen that the CIG may be a useful way of formalising whether variables are related to one another, particularly if one is interested in causal relationships. Estimating the CIG given samples from P is however a difficult task in general, though in the case

where P is multivariate Gaussian, things simplify considerably as we shall see. We begin with some notation. For a matrix $M \in \mathbb{R}^{p\times p}$, and sets $A, B \subseteq \{1,\dots,p\}$, let $M_{A,B}$ be the $|A| \times |B|$ submatrix of M consisting of those rows and columns of M indexed by the sets A and B respectively. The submatrix extraction operation is always performed first (so e.g. $M^T_{k,-k} = (M_{k,-k})^T$).

3.2.1 Normal conditionals

Now let $Z \sim N_p(\mu, \Sigma)$ with Σ positive definite. Note $\Sigma_{A,A}$ is also positive definite for any A.

Proposition 33.

$$Z_A|Z_B = z_B \sim N_{|A|}(\mu_A + \Sigma_{A,B}\Sigma^{-1}_{B,B}(z_B - \mu_B),\ \Sigma_{A,A} - \Sigma_{A,B}\Sigma^{-1}_{B,B}\Sigma_{B,A})$$

Proof. Idea: write $Z_A = MZ_B + (Z_A - MZ_B)$ with matrix $M \in \mathbb{R}^{|A|\times|B|}$ such that $Z_A - MZ_B$ and Z_B are independent, i.e. such that

$$\mathrm{Cov}(Z_B,\ Z_A - MZ_B) = \Sigma_{B,A} - \Sigma_{B,B}M^T = 0.$$

This occurs when we take $M^T = \Sigma^{-1}_{B,B}\Sigma_{B,A}$. Because $Z_A - MZ_B$ and Z_B are independent, the distribution of $Z_A - MZ_B$ conditional on $Z_B = z_B$ is equal to its unconditional distribution. Now

$$\begin{aligned}
\mathbb{E}(Z_A - MZ_B) &= \mu_A - \Sigma_{A,B}\Sigma^{-1}_{B,B}\mu_B \\
\mathrm{Var}(Z_A - MZ_B) &= \Sigma_{A,A} + \Sigma_{A,B}\Sigma^{-1}_{B,B}\Sigma_{B,B}\Sigma^{-1}_{B,B}\Sigma_{B,A} - 2\Sigma_{A,B}\Sigma^{-1}_{B,B}\Sigma_{B,A} \\
&= \Sigma_{A,A} - \Sigma_{A,B}\Sigma^{-1}_{B,B}\Sigma_{B,A}.
\end{aligned}$$

Since MZ_B is a function of Z_B and $Z_A - MZ_B$ is normally distributed, we have the result. □

3.2.2 Nodewise regression

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

$$Z_k = m_k + z^T_{-k}\Sigma^{-1}_{-k,-k}\Sigma_{-k,k} + \varepsilon_k,$$

where

$$\begin{aligned}
m_k &= \mu_k - \Sigma_{k,-k}\Sigma^{-1}_{-k,-k}\mu_{-k} \\
\varepsilon_k|Z_{-k} = z_{-k} &\sim N(0,\ \Sigma_{k,k} - \Sigma_{k,-k}\Sigma^{-1}_{-k,-k}\Sigma_{-k,k}).
\end{aligned}$$

Note that if the jth element of the vector of coefficients $\Sigma^{-1}_{-k,-k}\Sigma_{-k,k}$ is zero, then the distribution of Z_k conditional on Z_{-k} will not depend at all on the jth component of Z_{-k}.

Then if that jth component was $Z_{j'}$, we would have that $Z_k|Z_{-k} = z_{-k}$ has the same distribution as $Z_k|Z_{-j'k} = z_{-j'k}$, so $Z_k \perp\!\!\!\perp Z_j|Z_{-j'k}$.

Thus given $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} Z$ and writing

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix},$$

we may estimate the coefficient vector $\Sigma_{-k,-k}^{-1}\Sigma_{-k,k}$ by regressing X_k on $X_{\{k\}^c}$ 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 X_k on $X_{\{k\}^c}$, 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 exploiting a connection between the CIG and the precision matrix, as we explain below.

3.2.3 The precision matrix and conditional independence

The following facts about blockwise inversion of matrices will help us to interpret the mean and variance in Proposition 33.

Proposition 34. *Let $M \in \mathbb{R}^{p\times p}$ be a symmetric positive definite matrix and suppose*

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

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

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

Furthermore $det(M) = det(S)det(R)$.

Let $\Omega = \Sigma^{-1}$ be the precision matrix. Note that $\Sigma_{k,k} - \Sigma_{k,-k}\Sigma_{-k,-k}^{-1}\Sigma_{-k,k} = \Omega_{kk}^{-1}$, and more generally that $\mathrm{Var}(Z_A|Z_{A^c}) = \Omega_{A,A}^{-1}$. Also, we see that $\Sigma_{-k,-k}^{-1}\Sigma_{-k,k} = -\Omega_{kk}^{-1}\Omega_{-k,k}$, so

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

Thus

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

This motivates another approach to estimating the CIG.

3.2.4 The Graphical Lasso

Recall that the density of $N_p(\mu, \Sigma)$ is

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

The log-likelihood of (μ, Σ) based on an i.i.d. sample $x_1, \ldots, x_n$ is

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

Write

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

Then

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

Also,

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

Thus

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

and

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

Hence the maximum likelihood estimate of Ω, $\hat{\Omega}^{ML}$ can be obtained by solving

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

where $\Omega \succ 0$ means Ω is positive definite. One can show that the objective is convex and $\{\Omega : \Omega \succ 0\}$ is convex. As

$$\frac{\partial}{\partial \Omega_{jk}} \log\det(\Omega) = (\Omega^{-1})_{kj} = (\Omega^{-1})_{jk},$$
$$\frac{\partial}{\partial \Omega_{jk}} \mathrm{tr}(S\Omega) = S_{kj} = S_{jk},$$

if X has full column rank so S is positive definite, $\hat{\Omega}^{ML} = S^{-1}$.

The *graphical Lasso* [Yuan and Lin, 2007] penalises the log-likelihood for Ω and solves

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

where $\|\Omega\|_1 = \sum_{j,k} |\Omega_{jk}|$; this results in a sparse estimate of the precision matrix from which an estimate of the CIG can be constructed. Often the $\|\Omega\|_1$ is modified such that the diagonal elements are not penalised.

Chapter 4

High-dimensional inference

Statistics is not just about providing estimates of quantities of interest: we would also like to quantify uncertainty about these estimates. For instance, we would like to provide confidence intervals for regression coefficients in high-dimensional linear regression models, or test for the presence of a given edge in a conditional independence graph. These are challenging questions, but in the last several years there have been some breakthroughs on these fronts. In the next sections, we will aim to cover the main ideas in these exciting developments. Finally we will consider the important question of how to aggregate the outcomes of large numbers of hypothesis tests, a task which is referred to as multiple testing, and which continues to be an area of active research.

4.1 The debiased Lasso

Consider the normal linear model $Y = X\beta^0 + \varepsilon$ where $\varepsilon \sim N_n(0, \sigma^2 I)$. In the low-dimensional setting where X has full column rank, the fact that $\hat{\beta}^{\mathrm{OLS}} - \beta^0 \sim N_p(0, \sigma^2(X^T X)^{-1})$ allows us to form confidence intervals for components of β_j^0 and perform hypothesis tests for null $H_0 : \beta_j^0 = 0$, for example.

One might hope that studying the distribution of $\hat{\beta}_\lambda^{\mathrm{L}} - \beta^0$ would enable us to perform these tasks in the high-dimensional setting when $p \gg n$. However, the distribution of $\hat{\beta}_\lambda^{\mathrm{L}} - \beta^0$ is intractable and due to the bias of the Lasso, it depends delicately on the unknown β^0, making it unsuitable as a basis for constructing confidence intervals or hypothesis tests.

Whilst several methods have been proposed over the years, typically they have involved placing conditions on the unknown β^0, other than the usual assumption of sparsity. Given that the task is to perform inference for β^0, such conditions are undesirable. The recently developed *debiased Lasso* [Zhang and Zhang, 2014, Van de Geer et al., 2014] cleverly avoids these issues by attempting to remove enough of the bias of the Lasso to allow for a Gaussian limiting distribution. The construction may be motivated as follows.

Suppose, for the time being, that $X \in \mathbb{R}^{n \times p}$ is low-dimensional (i.e. $p < n$) and moreover has full column rank. Let $R_j \in \mathbb{R}^n$ be the vector of residuals from regressing $X_j \in \mathbb{R}^n$ onto $X_{-j} \in \mathbb{R}^{n \times (p-1)}$ using OLS. Then an alternative representation of the

ordinary least squares coefficients is given by $\hat{\beta}_j^{\text{OLS}} = R_j^T Y / R_j^T X_j$, and moreover[1]

$$\frac{R_j^T X_j}{\sigma \|R_j\|_2}(\hat{\beta}_j^{\text{OLS}} - \beta_j^0) \sim N(0,1).$$

This may be seen as a consequence of the fact that for any $\beta \in \mathbb{R}^p$,

$$\beta_j = \frac{R_j^T X \beta}{R_j^T X_j}. \tag{4.1}$$

Indeed, R_j is orthogonal to the column space of X_{-j} and so $R_j^T X\beta = R_j^T X_j \beta_j$. We may exploit this fact to remove the bias of any given estimator $\tilde{\beta}$ of β^0 as follows:

$$\tilde{\beta}_j + \frac{R_j^T(Y - X\tilde{\beta})}{R_j^T X_j} = \tilde{\beta}_j + \underbrace{\frac{R_j^T X(\beta^0 - \tilde{\beta})}{R_j^T X_j}}_{=\beta_j^0 - \tilde{\beta}_j \text{ by (4.1)}} + \underbrace{\frac{R_j^T \varepsilon}{R_j^T X_j}}_{\sim N(0, \sigma^2 \|R_j\|_2^2/(R_j^T X_j)^2)}. \tag{4.2}$$

This debiasing procedure is of no real use in low dimensions as the above simply recovers $\hat{\beta}_j^{\text{OLS}}$ given any initial $\tilde{\beta}$. However for high-dimensional settings, we may replace all OLS regressions in (4.2) with (square-root) Lasso regressions to give a genuinely new estimate of β_j^0, and this is known as the *debiased Lasso*. We will study a version of the debiased Lasso using square-root Lasso regressions given by

$$\hat{b}_j := \hat{\beta}_j^{\text{sq}} + \frac{(X_j - X_{-j}\hat{\theta}_j^{\text{sq}})^T (Y - X\hat{\beta}^{\text{sq}})}{(X_j - X_{-j}\hat{\theta}_j^{\text{sq}})^T X_j},$$

where

$$\hat{\beta}^{\text{sq}} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{\sqrt{n}} \|Y - X\beta\|_2 + \gamma \|\beta\|_1 \right\},$$
$$\hat{\theta}_j^{\text{sq}} := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^{p-1}} \left\{ \frac{1}{\sqrt{n}} \|X_j - X_{-j}\theta\|_2 + \gamma \|\theta\|_1 \right\},$$

and we have suppressed the dependence of the square-root Lasso estimates on the tuning parameter γ.

Theorem 35. *Let $\gamma = A\sqrt{\log p/n}$. Suppose $X_j - X_{-j}\hat{\theta}_j^{\text{sq}} \neq 0$ for all $j = 1, \ldots, p$. Then under the normal linear model $Y = X\beta^0 + \varepsilon$ with $\varepsilon \sim N_n(0, \sigma^2 I)$, we have that for each j,*

$$\frac{(X_j - X_{-j}\hat{\theta}_j^{\text{sq}})^T X_j}{\sigma \|X_j - X_{-j}\hat{\theta}_j^{\text{sq}}\|_2}(\hat{b}_j - \beta_j^0) = \delta_j + \zeta_j,$$

where $\zeta \in \mathbb{R}^p$ is mean-zero and Gaussian with $\text{Var}(\zeta_j) = 1$ and $\delta \in \mathbb{R}^p$ satisfies

$$\|\delta\|_\infty \leq A\sqrt{\log p} \|\beta^0 - \hat{\beta}^{\text{sq}}\|_1 / \sigma.$$

[1]Note that $R_j^T X_j = \|R_j\|_2^2$ in the case of OLS residuals, so some of the formulae may be simplified. However, such an equality does not hold when considering Lasso residuals, so in order to maintain the analogy with the debiased Lasso, we have left the $R_j^T X_j$ terms unsimplified.

Proof. Now

$$Y - X\hat{\beta}^{\mathrm{sq}} = X_j(\beta_j^0 - \hat{\beta}_j^{\mathrm{sq}}) + X_{-j}(\beta_{-j}^0 - \hat{\beta}_{-j}^{\mathrm{sq}}) + \varepsilon.$$

Note that

$$\frac{(X_j - X_{-j}\hat{\theta}^{\mathrm{sq}})^T X_j(\beta_j^0 - \hat{\beta}_j^{\mathrm{sq}})}{(X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}})^T X_j} = \beta_j^0 - \hat{\beta}_j^{\mathrm{sq}},$$

so

$$\frac{(X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}})^T X_j}{\sigma\|X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}}\|_2}(\hat{b}_j - \beta_j^0) = \underbrace{\frac{(X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}})^T X_{-j}(\beta_{-j}^0 - \hat{\beta}_{-j}^{\mathrm{sq}})}{\sigma\|X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}}\|_2}}_{=:\delta_j} + \underbrace{\frac{(X_j - X_{-j}\hat{\theta}^{\mathrm{sq}})^T \varepsilon}{\sigma\|X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}}\|_2}}_{=:\zeta_j}.$$

Note that ζ is thus mean-zero, Gaussian, and has $\mathrm{Var}(\zeta_j) = 1$. Meanwhile

$$\begin{aligned}
\sigma|\delta_j| &= \left|(\beta_{-j}^0 - \hat{\beta}_{-j}^{\mathrm{sq}})^T \frac{X_{-j}^T(X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}})}{\|X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}}\|_2}\right| \\
&\leq \sqrt{n}\|\beta_{-j}^0 - \hat{\beta}_{-j}^{\mathrm{sq}}\|_1 \frac{\|X_{-j}^T(X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}})\|_\infty}{\|X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}}\|_2} \\
&\leq A\sqrt{\log p}\|\beta^0 - \hat{\beta}^{\mathrm{sq}}\|_1,
\end{aligned}$$

applying Hölder's inequality, and appealing to the KKT condtions of the square-root Lasso. □

Theorem 30 shows that under reasonable conditions we can expect that $\|\beta^0 - \hat{\beta}^{\mathrm{sq}}\|_1 \leq$ const. $\times\ \sigma s\sqrt{\log(p)/n}$ where s is the sparsity level, and so δ will be negligible in an asymptotic regime where $s\log(p)/\sqrt{n} \to 0$.

Theorem 35 suggests the following approximate $(1-\alpha)$-level confidence interval for β_j^0:

$$\hat{C}_j := \left[\hat{b}_j - z_{\alpha/2}\frac{\hat{\sigma}\|X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}}\|_2}{X_j^T(X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}})},\ \hat{b}_j + z_{\alpha/2}\frac{\hat{\sigma}\|X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}}\|_2}{X_j^T(X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}})}\right],$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ point of a standard Gaussian distribution and

$$\hat{\sigma} := \frac{1}{\sqrt{n}}\|Y - X\hat{\beta}^{\mathrm{sq}}\|_2.$$

One outstanding question regarding Theorem 35 is how large we can expect the factor

$$\frac{X_j^T(X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}})}{\|X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}}\|_2} \tag{4.3}$$

to be; the magnitude of this determines the size of β_j^0 we can hope to detect with the debiased Lasso. Note the low-dimensional analogue is $X_j^T R_j / \|R_j\|_2 = \|R_j\|_2$. To understand this quantity better, consider a random design setting where we have a normal linear model $X_j = X_{-j}\theta_j^0 + \xi$ with $\theta_j^0 \in \mathbb{R}^{p-1}$ and $\xi \sim N_n(0, v^2 I)$. Then Lemma 29 suggests $\|R_j\|_2/\sqrt{n} \approx v$ and so we can expect that approximately we have

$$\sqrt{n}(\hat{\beta}_j^{\mathrm{OLS}} - \beta_j^0) \dot{\sim} N(0, \sigma^2/v^2).$$

Turning to the high-dimensional setting, we can show (see example sheet) that the quantity in (4.3) divided by $\sqrt{n}$ is in fact equal to

$$\frac{1}{\sqrt{n}}\|X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}}\|_2 + \gamma\|\hat{\theta}_j^{\mathrm{sq}}\|_1$$

provided $\|X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}}\|_2 \neq 0$. If θ_j^0 is sufficiently sparse, we can expect $\gamma\|\hat{\theta}_j^{\mathrm{sq}}\|_1 \approx \gamma\|\theta_j^0\|_1 \approx 0$, and $\|X_j - X_{-j}\hat{\theta}_j^{\mathrm{sq}}\|_2/\sqrt{n} \approx v$ owing to Lemma 29, thus giving the analogous result that approximately

$$\sqrt{n}(\hat{b}_j - \beta_j^0) \dot{\sim} N(0, \sigma^2/v^2).$$

4.2 Basic asymptotic statistics

Definition 12. Let $W_1, W_2, \ldots$ and W be real-valued random variables.

- We say the W_n *converge in distribution* to W with distribution function F, and write $W_n \xrightarrow{d} W$, if for all $t \in \mathbb{R}$ at which F is continuous,

$$\mathbb{P}(W_n \leq t) \to F(t) \qquad \text{as } n \to \infty.$$

- We say the W_n *converge in probability* to W and write $W_n \xrightarrow{p} W$ if for all $\epsilon > 0$,

$$\mathbb{P}(|W_n - W| > \epsilon) \to 0.$$

One can show that if $W_n \xrightarrow{p} W$, then $W_n \xrightarrow{d} W$ (so convergence in probability is a stronger notion of convergence). However, they coincide if $W = c \in \mathbb{R}$ is deterministic i.e. if $W_n \xrightarrow{d} c$, then $W_n \xrightarrow{p} c$.

Lemma 36. *Let the sequence W_n and W be as above. Then $W_n \xrightarrow{p} W$ if and only if there exists a sequence $a_n \to 0$ such that*

$$\mathbb{P}(|W_n - W| \leq a_n) \to 1.$$

**Proof*.* Suppose $W_n \xrightarrow{p} W$. We know that given any $\epsilon, \delta > 0$, there exists $N(\epsilon, \delta)$ such that for all $n \geq N(\epsilon, \delta)$,

$$\mathbb{P}(|W_n - W| > \epsilon) \leq \delta.$$

We may take $(N(1/j, 1/j))_{j=1}^{\infty}$ to be an increasing sequence with $N(1,1) = 1$, and then define for each $n \in \mathbb{N}$,

$$a_n := \frac{1}{\max\{j : N(1/j, 1/j) \leq n\}}.$$

Then $a_n \to 0$. Moreover, if $a_n = 1/j$ for some j, then $n \geq N(1/j, 1/j)$, so

$$\mathbb{P}(|W_n - W| > a_n) = \mathbb{P}(|W_n - W| > 1/j) \leq 1/j = a_n \leq a_n.$$

The other direction is clear. □

Lemma 37 (Weak law of large numbers (WLLN))**.** *If $W_1, W_2, \ldots$ are i.i.d. real-valued random variables and $\mathbb{E}(W_1) = \mu < \infty$, then as $n \to \infty$,*

$$\frac{1}{n}\sum_{i=1}^{n} W_i \xrightarrow{p} \mu.$$

Proof. We only prove this with the additional assumption that $\mathrm{Var}(W_1) < \infty$. Given $\epsilon > 0$, by Markov's inequality,

$$\mathbb{P}\left(\left(\frac{1}{n}\sum_{i=1}^{n} W_i - \mu\right)^2 > \epsilon\right) \leq \frac{1}{\epsilon}\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} W_i\right) = \epsilon^{-1}\mathrm{Var}(W_1)/n \to 0.$$

□

Theorem 38 (Central limit theorem (CLT))**.** *Consider the setup of Lemma 37. We have*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} W_i - \mu\right) \xrightarrow{d} N(0,1).$$

Theorem 39 (Continuous mapping theorem (CMT))**.** *Suppose the sequence of random variables $(W_n)_{n=1}^{\infty}$ is such that $W_n \xrightarrow{p} W$. Let $f : \mathbb{R} \to \mathbb{R}$ be continuous at every point in a set C with $\mathbb{P}(W \in C) = 1$. Then $f(W_n) \xrightarrow{p} f(W)$.*

Lemma 40 (Slutsky's lemma)**.** *Let $(U_n)_{n=1}^{\infty}$ and $(W_n)_{n=1}^{\infty}$ be sequences of random variables where $U_n \xrightarrow{d} U$ and $W_n \xrightarrow{p} c$ for random variable $U \in \mathbb{R}$ and deterministic $c \in \mathbb{R}$. Then*

1. *$U_n + W_n \xrightarrow{d} U + c$,*
2. *$U_n W_n \xrightarrow{d} Uc$,*
3. *$U_n / W_n \xrightarrow{d} U/c$ if $c \neq 0$.*

Proof of Lemma 29

Recall our setup that $Y = f + \varepsilon$ where error $\varepsilon \in \mathbb{R}^n$ is such that $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with $\mathbb{E}\varepsilon_1 = 0$ and $\mathrm{Var}\varepsilon_1 = \sigma^2 > 0$. We consider an asymptotic regime where n increases, and suppose $\hat{f} \in \mathbb{R}^n$ is an estimate of f such that there exists sequence $a_n \to 0$ with

$$\mathbb{P}(\|\hat{f} - f\|_2^2/n \leq a_n) \to 1.$$

Then setting

$$\hat{\sigma} := \frac{1}{\sqrt{n}}\|Y - \hat{f}\|_2,$$

we claim there exists sequence $b_n \to 0$ such that

$$\mathbb{P}\left(\left|\frac{\sigma}{\hat{\sigma}} - 1\right| \leq b_n\right) \to 1.$$

Proof. From Lemma 36, we see that the assumption of Lemma 29 is equivalent to $\|\hat{f} - f\|_2^2/n \xrightarrow{p} 0$. Similarly, the conclusion is that $\sigma/\hat{\sigma} \xrightarrow{p} 1$. Moreover, applying Slutsky's lemma with $U_n = \sigma$, $W_n = \hat{\sigma}$ and noting that convergence in distribution implies convergence in probability when the limit is deterministic, we see that this follows from $\hat{\sigma} \xrightarrow{p} \sigma$, which we now show.

We have

$$\hat{\sigma}^2 = \underbrace{\frac{1}{n}\|f - \hat{f}\|_2^2}_{\xrightarrow{p}0} + \underbrace{\frac{1}{n}\|\varepsilon\|_2^2}_{\xrightarrow{p}\sigma^2 \text{ by WLLN}} + \frac{2}{n}\varepsilon^T(f - \hat{f}). \tag{4.4}$$

Now by the Cauchy–Schwarz inequality,

$$\frac{1}{n}|\varepsilon^T(f - \hat{f})| \leq \frac{1}{\sqrt{n}}\|\varepsilon\|_2 \frac{1}{\sqrt{n}}\|f - \hat{f}\|_2 \xrightarrow{p} 0,$$

by the WLLN, the CMT, and Slutsky's lemma. Thus returning to (4.4) and applying Slutsky once more, we see that $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ whence (by the CMT) $\hat{\sigma} \xrightarrow{p} \sigma$ as required. □

4.3 Conditional independence testing

The debiased Lasso may be viewed as part of a wider theme involving using modern regression methods (in our case the Lasso) within methods for classical inferential problems (in our case forming a confidence interval for a regression coefficient). The use of such flexible regressions can often substantially weaken the assumptions required by more classical techniques. One important problem that can benefit from such an approach is that of testing conditional independence.

Suppose $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^n$ and $Z \in \mathbb{R}^{n \times p}$ have components or rows respectively that are i.i.d. Consider the first observation $(x_1, y_1, z_1) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^p$. One way of connecting regression with the problem of conditional independence testing is via the following observation:

$$x_1 \perp\!\!\!\perp y_1 \,|\, z_1 \implies \mathbb{E}[\{x_1 - \mathbb{E}(x_1 \,|\, z_1)\}\{y_1 - \mathbb{E}(y_1 \,|\, z_1)\}] = 0 \tag{4.5}$$

provided $\mathbb{E}x_1^2, \mathbb{E}y_1^2 < \infty$. In other words, the population level residuals from regressing each of x_1 and y_1 on z_1 are uncorrelated. To see this, observe that in fact as $\mathbb{E}(x_1 y_1 \,|\, z_1) = \mathbb{E}(x_1 \,|\, z_1)\mathbb{E}(y_1 \,|\, z_1)$, under conditional independence, we have

$$\mathbb{E}[\{x_1 - \mathbb{E}(x_1 \,|\, z_1)\}\{y_1 - \mathbb{E}(y_1 \,|\, z_1)\} \,|\, z_1] = 0,$$

so (4.5) follows from the tower property. The relationship in (4.5) suggests regressing each of X and Y on Z, and constructing a test statistic based on the products of the residuals. Suppose that

$$x_1 = f(z_1) + \varepsilon_1 \qquad \text{and} \qquad y_1 = g(z_1) + \xi_1, \tag{4.6}$$

where $f(z_1) := \mathbb{E}(x_1 \,|\, z_1)$, $g(z_1) := \mathbb{E}(y_1 \,|\, z_1)$, and so under conditional independence $\mathbb{E}(\varepsilon_1 \xi_1) = 0$. Suppose additionally that $\varepsilon_1 \perp\!\!\!\perp z_1$ and $\xi_1 \perp\!\!\!\perp z_1$, so when $x_1 \perp\!\!\!\perp y_1 \,|\, z_1$,

$$\mathrm{Var}(\varepsilon_1 \xi_1) = \mathbb{E}\{\mathbb{E}(\varepsilon_1^2 \xi_1^2 \,|\, z_1)\} = \mathbb{E}\{\mathbb{E}(\varepsilon_1^2 \,|\, z_1)\mathbb{E}(\xi_1^2 \,|\, z_1)\} = \mathrm{Var}(\varepsilon_1)\mathrm{Var}(\xi_1).$$

Given fitted regression functions $\hat{f}$ and $\hat{g}$ from regressing X and Y respectively on Z, consider

$$\tau_N := \frac{1}{n} \sum_{i=1}^n \{x_i - \hat{f}(z_i)\}\{y_i - \hat{g}(z_i)\}.$$

Define $\varepsilon_i := x_i - f(z_i)$ and $\xi_i := y_i - g(z_i)$. If f and g are estimated sufficiently well by $\hat{f}$ and $\hat{g}$, then the ith summand above will be close to $\varepsilon_i \xi_i$. Under the null these quantities are mean-zero and i.i.d., so the CLT suggests we can expect $\sqrt{n}\tau_N$ to have a Gaussian distribution. To obtain a standard normal limit, we should normalise by the square root of $\mathrm{Var}(\sqrt{n}\tau_N) \approx \mathrm{Var}(\varepsilon_1 \xi_1) = \mathrm{Var}(\varepsilon_1)\mathrm{Var}(\xi_1)$. Lemma 29 suggests we could estimate this using

$$\tau_D^2 := \left(\frac{1}{n} \sum_{i=1}^n \{x_i - \hat{f}(z_i)\}^2 \right) \left(\frac{1}{n} \sum_{i=1}^n \{y_i - \hat{g}(z_i)\}^2 \right),$$

and so take as our final test statistic

$$T := \sqrt{n} \frac{\tau_N}{\tau_D}.$$

The result below formalises our intuition that provided

$$\mathrm{MSPE}_f := \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n \{f(z_i) - \hat{f}(z_i)\}^2 \right) \qquad \text{and} \qquad \mathrm{MSPE}_g := \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n \{g(z_i) - \hat{g}(z_i)\}^2 \right)$$

decay sufficiently fast, the test statistic will have an asymptotic standard normal distribution under the null.

Theorem 41. *Consider model* (4.6) *and suppose* $0 < \mathrm{Var}(\varepsilon_1) < \infty$ *and* $0 < \mathrm{Var}(\xi_1) < \infty$. *Suppose* $\mathrm{MSPE}_f \to 0$, $\mathrm{MSPE}_g \to 0$ *and* $\mathrm{MSPE}_f\mathrm{MSPE}_g = o(n^{-1})$. *Then under the null that* $x_1 \perp\!\!\!\perp y_1 \,|\, z_1$, *we have* $T \xrightarrow{d} N(0,1)$.

Proof. Let us consider τ_N first. Writing $x_i = f(z_i) + \varepsilon_i$ and similarly for y_i, we have

$$\begin{aligned} n\tau_N &= \sum_{i=1}^n \{f(z_i) - \hat{f}(z_i) + \varepsilon_i\}\{g(z_i) - \hat{g}(z_i) + \xi_i\} \\ &= \sum_{i=1}^n \{f(z_i) - \hat{f}(z_i)\}\{g(z_i) - \hat{g}(z_i)\} + \sum_{i=1}^n \varepsilon_i\{g(z_i) - \hat{g}(z_i)\} \\ &\quad + \sum_{i=1}^n \xi_i\{f(z_i) - \hat{f}(z_i)\} + \sum_{i=1}^n \varepsilon_i\xi_i \\ &=: A_1 + A_2 + A_3 + A_4. \end{aligned}$$

By the CLT,

$$\frac{A_4}{\sqrt{n}} \xrightarrow{d} N(0, \mathrm{Var}(\varepsilon_1)\mathrm{Var}(\xi_1)).$$

Now by the triangle inequality and the Cauchy–Schwarz inequality,

$$\begin{aligned} \mathbb{E}|A_1| &\le \sum_{i=1}^n \mathbb{E}\{|f(z_i) - \hat{f}(z_i)||g(z_i) - \hat{g}(z_i)|\} \\ &\le \sum_{i=1}^n [\mathbb{E}\{f(z_i) - \hat{f}(z_i)\}^2]^{1/2}[\mathbb{E}\{g(z_i) - \hat{g}(z_i)\}^2]^{1/2} \\ &\le \left(\sum_{i=1}^n \mathbb{E}\{f(z_i) - \hat{f}(z_i)\}^2\right)^{1/2} \left(\sum_{i=1}^n \mathbb{E}\{g(z_i) - \hat{g}(z_i)\}^2\right)^{1/2}. \end{aligned}$$

Thus, by Markov's inequality, given $\delta > 0$,

$$\mathbb{P}(|A_1|/\sqrt{n} > \delta) \le \frac{\delta^{-1}}{\sqrt{n}}\mathbb{E}|A_1| \le \sqrt{n\mathrm{MSPE}_f\mathrm{MSPE}_g} \to 0,$$

by assumption, so $A_1/\sqrt{n} \xrightarrow{p} 0$. Turning to A_2, observe that

$$\mathbb{E}(\varepsilon_i\varepsilon_j\{g(z_i) - \hat{g}(z_i)\}\{g(z_j) - \hat{g}(z_j)\} \,|\, Y, Z) = \{g(z_i) - \hat{g}(z_i)\}\{g(z_j) - \hat{g}(z_j)\}\mathbb{E}(\varepsilon_i\varepsilon_j \,|\, Y, Z).$$

As $\mathbb{E}(\varepsilon_i\varepsilon_j \,|\, Y, Z) = \mathbb{E}(\varepsilon_i\varepsilon_j \,|\, Z)$, the above display is 0 if $i \neq j$. Therefore given $\delta > 0$, by Markov's inequality and the above,

$$\begin{aligned} \mathbb{P}(|A_2|/\sqrt{n} > \delta) = \mathbb{P}(A_2^2/n > \delta^2) &\le \frac{\delta^{-2}}{n}\mathbb{E}(A_2^2) \\ &= \frac{\delta^{-2}}{n}\sum_{i=1}^n \mathbb{E}[\varepsilon_i^2\{g(z_i) - \hat{g}(z_i)\}^2] \\ &= \delta^{-2}\mathrm{Var}(\varepsilon_1)\mathrm{MSPE}_g \to 0, \end{aligned}$$

by assumption, so $A_2/\sqrt{n} \xrightarrow{p} 0$. Similarly $A_3/\sqrt{n} \xrightarrow{p} 0$. Thus by Slustky's lemma, $\sqrt{n}\tau_N \xrightarrow{d} N(0, \mathrm{Var}(\varepsilon_1)\mathrm{Var}(\xi_1))$.

We now show that $\tau_D \xrightarrow{p} \sqrt{\mathrm{Var}(\varepsilon_1)\mathrm{Var}(\xi_1)}$. Note that by Markov's inequality, given $\delta > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\{f(z_i) - \hat{f}(z_i)\}^2 > \delta\right) \leq \delta^{-1}\mathrm{MSPE}_f \to 0,$$

so

$$\frac{1}{n}\sum_{i=1}^{n}\{f(z_i) - \hat{f}(z_i)\}^2 \xrightarrow{p} 0,$$

and similarly

$$\frac{1}{n}\sum_{i=1}^{n}\{g(z_i) - \hat{g}(z_i)\}^2 \xrightarrow{p} 0.$$

Then applying Lemma 29, we see that

$$\frac{1}{n}\sum_{i=1}^{n}\{x_i - \hat{f}(z_i)\}^2 \xrightarrow{p} \mathrm{Var}(\varepsilon_1) \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n}\{x_i - \hat{f}(z_i)\}^2 \xrightarrow{p} \mathrm{Var}(\xi_1).$$

Thus applying Slutsky's lemma and the CMT shows that $\tau_D \xrightarrow{p} \sqrt{\mathrm{Var}(\varepsilon_1)\mathrm{Var}(\xi_1)}$ and so $T \xrightarrow{d} N(0,1)$ as required. □

Corollary 42. *Consider the setup of Theorem 41 where f and g lie in RKHS $\mathcal{H}$ with reproducing kernel k satisfying an eigendecomposition of the form (1.22), and $\hat{f}$ and $\hat{g}$ are kernel ridge regression estimates using kernel k and optimal choices of tuning parameters. Then $T \xrightarrow{d} N(0,1)$.*

Proof. This follows from Theorem 41 and Theorem 8. □

4.4 Multiple testing

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 $\alpha \in [0,1]$, $i \in I_0$. 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.

4.4.1 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 43. *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:

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

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

- Here H_1 is rejected by 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 44. *The closed testing procedure makes no false rejections with probability at least* $1-\alpha$*. In particular it controls the FWER at level* α*.*

Proof. Assume I_0 is not empty (as otherwise no rejection can be false anyway). The procedure makes a false rejection if and only if $\phi_{I_0} = 1$, 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.

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

It can be shown (see example sheet) that Holm's procedure amounts to ordering 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, and then performing 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.

4.4.2 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

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

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 45. *Suppose that for each $i \in I_0$, p_i is independent of $\{p_j : j \neq i\}$. 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

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

□

Bibliography

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.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

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

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.

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.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2007.

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

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, 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.

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 (Statistical Methodology)*, 68(1):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.

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