

1 Kernel Machines

Consider a linear model

$$Y_i = x_i^T \beta^0 + \varepsilon_i, \quad i = 1, \dots, n, \quad x_i \in \mathbb{R}^p \text{ fixed}$$

where $\mathbb{E}\varepsilon = 0$, $\text{Var}(\varepsilon) = \sigma^2 I_n$. We have

$$\begin{aligned} \hat{\beta}^{\text{ols}} &= \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \sum_{i=1}^n (Y_i - x_i^T \beta)^2 \\ &= \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \|Y - X\beta\|^2 \\ &= (X^T X)^{-1} X^T Y. \end{aligned}$$

Classical theory:

- $\hat{\beta}^{\text{ols}}$ unbiased,

$$\text{Var}(\hat{\beta}^{\text{ols}}) = \sigma^2 (X^T X)^{-1} = i^{-1}(\beta^0)$$

Where i is the Fisher information.

- Cramér-Rao lower bound: if an estimator $\tilde{\beta}$ is unbiased then

$$\text{Var}(\tilde{\beta}) - i^{-1}(\beta^0) \underset{\text{positive semi-definite}}{\geq} 0.$$

- If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, then $\hat{\beta}^{\text{ols}}$ is the MLE of β^0 . Furthermore $\sqrt{n}(\hat{\beta}^{\text{ols}} - \beta^0) \sim \mathcal{N}(0, n\sigma^2(X^T X)^{-1})$. From this we can derive confidence intervals, hypothesis test, etc.

In a general model with parameter $\theta \in \mathbb{R}^p$, n independent observations, under regularity, we have asymptotic normality, i.e $\sqrt{n}(\hat{\theta}^{\text{MLE}} - \theta^0) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta^0))$ (with p fixed).

Question: what happens when p is large relative to n ?

- If $p > n$, $\hat{\beta}^{\text{ols}}$ is not even defined.
- If $p \approx n$, $\text{Var}(\hat{\beta}^{\text{ols}})$ explodes since $X^T X$ is near singular.
- More generally, if $p, n \rightarrow \infty$ then asymptotic normality can break down.

Recall the bias-variance decomposition:

$$\begin{aligned} \text{mse}(\tilde{\beta}) &= \mathbb{E}_{\beta^0, \sigma^2} [(\tilde{\beta} - \beta^0)(\tilde{\beta} - \beta^0)] \\ &= \mathbb{E}_{\beta^0, \sigma^2} \left\| \tilde{\beta} - \mathbb{E}\tilde{\beta} + \mathbb{E}\tilde{\beta} - \beta^0 \right\|^2 \\ &= \text{Var}(\tilde{\beta}) + \left\| \mathbb{E}(\tilde{\beta}) - \beta^0 \right\|^2. \end{aligned}$$

We introduce bias to reduce the variance.

1.1 Ridge regression

Define

$$(\hat{\mu}_\lambda^R, \hat{\beta}_\lambda^R) = \operatorname{argmin}_{(\mu, \beta) \in \mathbb{R} \times \mathbb{R}^p} \left[\|Y - \mu \mathbf{1} - X\beta\|^2 + \underbrace{\lambda \|\beta\|^2}_{\text{penalty for large } \beta} \right].$$

λ is called a *regularisation* or *tuning* parameter. We shall assume the columns of X have been standardised (mean 0, variance 1).

After standardisation, we can show that

$$\hat{\mu}_\lambda^R = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y}.$$

Hence, if we replace Y with $Y - \mathbf{1}\bar{Y}$ we can write

$$\begin{aligned} \hat{\beta}_\lambda^R &= \operatorname{argmin}_{\beta \in \mathbb{R}^p} [\|Y - X\beta\|^2 + \lambda \|\beta\|^2] \\ &= \underbrace{(X^T X + \lambda I_p)^{-1}}_{\text{always invertible}} X^T Y. \end{aligned}$$

Theorem 1. For $\lambda > 0$ sufficiently small,

$$\operatorname{mse}(\hat{\beta}^{ols}) - \operatorname{mse}(\hat{\beta}_\lambda^R) = \mathbb{E}\|\hat{\beta}^{ols} - \beta^0\|^2 - \mathbb{E}\|\hat{\beta}_\lambda^R - \beta^0\|^2 > 0. \quad (*)$$

Proof. We have

$$Y = X\beta^0 + \varepsilon.$$

The bias of $\hat{\beta}_\lambda^R$ is

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

While we have variance

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

Then (*) becomes

$$\begin{aligned} &\mathbb{E}\|\hat{\beta}^{ols} - \beta^0\|^2 - \mathbb{E}\|\hat{\beta}_\lambda^R - \beta^0\|^2 \\ &= \sigma^2 (X^T X)^{-1} - \sigma^2 (X^T X + \lambda I) X^T X (X^T X + \lambda I)^{-1} \\ &\quad - \lambda^2 (X^T X + \lambda I)^{-1} \beta^0 (\beta^0)^T (X^T X + \lambda I)^{-1} \\ &= \vdots \quad \quad \quad (\text{use SVD } X = UDV^T) \\ &= \lambda (X^T X + \lambda I)^{-1} [\sigma^2 \{2I_p + \lambda (X^T X)^{-1}\} - \lambda \beta^0 (\beta^0)^T] (X^T X + \lambda I)^{-1}. \end{aligned}$$

We want to show this is positive definite. This is equivalent to

$$\begin{aligned}\sigma^2 [2I + \lambda(X^T X)^{-1}] - \lambda\beta^0(\beta^0)^T &> 0 \\ \iff 2\sigma^2 I - \lambda\beta^0(\beta^0)^T &> 0 \\ \iff 2\sigma^2 \|z\|^2 - \lambda(z^T \beta^0)^2 &> 0 \quad \forall z \in \mathbb{R}^p.\end{aligned}\tag{†}$$

We also have $(z^T \beta^0)^2 \leq \|z\|^2 \|\beta^0\|^2$ by Cauchy-Schwarz. Hence (†) holds for all $\lambda < \frac{2\sigma^2}{\|\beta^0\|^2}$. \square

Singular value decomposition

Suppose $n \geq p$, so we can always write $X \in \mathbb{R}^{n \times p}$ as

$$X = UDV^T \quad (\text{“thin SVD”})$$

where $U \in \mathbb{R}^{n \times p}$, $V \in \mathbb{R}^{p \times p}$, with orthonormal columns, $D \in \mathbb{R}^{p \times p}$ diagonal with $D_{11} \geq D_{22} \geq \dots \geq D_{pp} \geq 0$.

The fitted values in ridge regression are

$$\begin{aligned} \hat{Y}_\lambda^R &= X\hat{\beta}_\lambda^R = X(X^T X + \lambda I)^{-1} X^T Y \\ &= UDV^T (VD^2 V^T + \lambda I)^{-1} V D U^T Y \quad (\text{using } VV^T = V^T V = I) \\ &= UD(D^2 + \lambda I)^{-1} D U^T Y \\ &= \sum_{j=1}^p U_j \frac{D_{jj}^2}{D_{jj}^2 + \lambda} U_j^T Y \end{aligned}$$

where U_j denotes the j th column of U . For reference, in OLS regression

$$\hat{Y}^{ols} = X\hat{\beta}^{ols} = X(X^T X)^{-1} X^T Y = \sum_{j=1}^p U_j U_j^T Y.$$

So ridge “projects” onto columns of U , but it shrinks j th component by a factor

$$\frac{D_{jj}^2}{D_{jj}^2 + \lambda}.$$

Hence it shrinks small singular values to 0 rapidly.

Note. The matrix $X(X^T X)^{-1} X^T Y$ is known as the “hat matrix” and it represents an orthogonal projection onto the column space of X .

The SVD of X is related to principal component analysis.

Definition. The k th *principal component* $U^{(k)}$ of X and *principal direction* $v^{(k)}$ of X are defined recursively by

$$v^{(k)} = \operatorname{argmax}_{v \in \mathbb{R}^p} \|Xv\|^2 \text{ subject to } \|v\| = 1, (v^{(j)})^T X^T X v = 0 \quad \forall j < k$$

and

$$u^{(k)} = Xv^{(k)}.$$

Lemma 2. If $D_{jj} > 0$ for all $j \in \{1, \dots, p\}$ then $v^{(k)} = V_k$, $u^{(k)} = D_{kk} U_k$.

Message: ridge is good when the signal (β^0) is large for the top principal components of X .

Computation: we can compute \hat{Y}_λ^R for any value of λ quickly after doing an SVD, which has cost $\mathcal{O}(np^2)$.

1.2 v -fold cross-validation

We assume that (x_i, Y_i) , $i = 1, \dots, n$ is iid from some distribution (random design matrix). Let (x^*, Y^*) be another independent observation from this distribution. We may wish to pick λ minimising the mean-squared prediction error (MSPE) conditional on (X, Y) :

$$\mathbb{E}\{(Y^* - (x^*)^T \hat{\beta}_\lambda^R(X, Y))^2 | (X, Y)\}.$$

A less ambitious goal is to minimise the MSPE

$$\mathbb{E}\{(Y^* - (x^*)^T \hat{\beta}_\lambda^R(X, Y))^2\} = \mathbb{E}\left[\mathbb{E}\{(Y^* - (x^*)^T \hat{\beta}_\lambda^R(X, Y))^2 | (X, Y)\}\right]. \quad (\ddagger)$$

We can try to estimate this quantity for different values of λ , using data splitting.

- Let $(X^{(1)}, Y^{(1)}), \dots, (X^{(v)}, Y^{(v)})$ be groups of data points of roughly equal size. These are called *folds*.
- Let $(X^{(-k)}, Y^{(-k)})$ denote all the folds except the k th.
- Let $\kappa(i)$ be the fold to which sample i (i.e. (X_i, Y_i)) belongs.

Our estimator of (\ddagger) is

$$\text{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ Y_i - x_i^T \underbrace{\hat{\beta}_\lambda^R(X^{(-\kappa(i))}, Y^{(-\kappa(i))})}_{\text{using all folds except the ones containing } (x_i, Y_i)} \right\}^2.$$

Then define

$$\lambda_{\text{CV}} = \operatorname{argmin}_{\lambda \in \{\lambda_1, \dots, \lambda_m\}} \text{CV}(\lambda).$$

We use the estimator

$$\hat{\beta}_{\lambda_{\text{CV}}}^R(X, Y).$$

How to choose v ?

Note.

- The expectation of each summand in $\text{CV}(\lambda)$ is almost the same as \ddagger , which is what we want to estimate. The only difference is the size of the training set. Hence the bias of $\text{CV}(\lambda)$ is small when v is large [the extreme of this is $v = n$, called “leave one out” cross-validation].
- When v is large, the estimator $\hat{\beta}_\lambda^R(X^{(-k)}, Y^{(-k)})$ is similar for different values of k , which leads to positively correlated summands in $\text{CV}(\lambda)$, leading to high variance.
- A common choice is $v = 5$ or $v = 10$.

1.3 Kernel trick

We have

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

Note that

$$\begin{aligned} X^T (X X^T + \lambda I) &= (X^T X + \lambda I) X^T \\ \implies (X^T X + \lambda I)^{-1} X^T &= X^T (X X^T + \lambda I)^{-1} \\ \implies X \underbrace{(X^T X + \lambda I)^{-1}}_{p \times p} X^T Y &= X X^T \underbrace{(X X^T + \lambda I)^{-1}}_{n \times n} Y. \end{aligned}$$

The computation cost of the LHS is $\mathcal{O}(np^2 + p^3)$ while the RHS is $\mathcal{O}(pn^2 + n^3)$.

- When $p \gg n$, the 2nd expression is cheaper to compute;
- The fitted values in ridge regression only depend on X through the “Gram matrix” $K = X X^T$, with entries $K_{ij} = \langle x_i, x_j \rangle$.

Suppose we wish to fit a quadratic model:

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

This can be done with a linear model where we replace the predictors $x_i \in \mathbb{R}^p$ with a new “feature” vector:

$$\phi(x_i) = (x_{i1}, x_{i2}, \dots, x_{ip}, x_{i1}x_{i1}, x_{i1}x_{i2}, \dots, x_{ip}x_{ip}) \in \mathbb{R}^{p+p^2}.$$

We call ϕ a “feature map”. Now we have $\mathcal{O}(p^2)$ predictors. If $p^2 \gg n$, to compute ridge fitted values, we want to use the 2nd expression, with cost $\mathcal{O}(p^2 n^2 + n^3)$.

However, the part that scales as $\mathcal{O}(p^2 n^2)$ is just the computation of the Gram matrix with entries $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$.

The *kernel trick* offers a shortcut for computing K .

Idea:

$$\begin{aligned} \left(\frac{1}{2} + x_i^T x_j \right)^2 - \frac{1}{4} &= \left(\frac{1}{2} + \sum_k x_{ik} x_{jk} \right)^2 - \frac{1}{4} \\ &= \sum_k x_{ik} x_{jk} + \sum_{k,l} x_{ik} x_{il} x_{jk} x_{jl} \\ &= \langle \phi(x_i), \phi(x_j) \rangle = K_{ij}. \end{aligned}$$

The LHS can be computed in $\mathcal{O}(p)$ iterations, so we can obtain K in $\mathcal{O}(n^2 p)$ iterations, and we can compute the fitted values in ridge regression in $\mathcal{O}(n^2 p + n^3)$, which is not worse than the linear model!

Notes.

- For many feature maps ϕ , there are similar shortcuts.
- Instead of focusing on ϕ , we can directly think of the function $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ as a measure of “similarity” between inputs x_i, x_j .

Question: for which similarities k is there a feature map ϕ such that

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle?$$

1.4 Kernels

Definition. An *inner product space* is a real vector space \mathcal{H} endowed with a map $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ satisfying:

- (i) Symmetry: for all $u, v \in \mathcal{H}$ we have $\langle u, v \rangle = \langle v, u \rangle$;
- (ii) Bilinearity: for all $a, b \in \mathbb{R}$ and all $u, v, w \in \mathcal{H}$ we have

$$\langle au + bv, w \rangle = a\langle u, w \rangle + b\langle v, w \rangle.$$

- (iii) Positive-definiteness: we have $\langle u, u \rangle \geq 0$ for all $u \in \mathcal{H}$, with equality if and only if $u = 0$.

Suppose that regression inputs x_1, \dots, x_n take values in an abstract set \mathcal{X} (so far we’ve had $\mathcal{X} = \mathbb{R}^p$, but the x_i ’s could be functions; images; graphs; etc.).

Goal: characterise similarity functions $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ for which there is an inner product space \mathcal{H} and a feature map $\phi : \mathcal{X} \rightarrow \mathcal{H}$ such that

$$k(x, x') = \langle \phi(x), \phi(x') \rangle \quad \forall x, x' \in \mathcal{X}.$$

Definition. A (*positive-definite*) *kernel* k is a symmetric map $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ such that for all $n \in \mathbb{N}$ and all $x_1, \dots, x_n \in \mathcal{X}$, the matrix K with entries $K_{ij} = k(x_i, x_j)$ is positive semi-definite.

Remark. A kernel is not an inner product on \mathcal{X} in general. Indeed, \mathcal{X} does not even need to be a vector space, and k need not be bilinear. However, we do have a version of the Cauchy-Schwarz inequality for kernels.

Proposition 3. *Let k be a kernel on \mathcal{X} . Then*

$$k(x, x')^2 \leq k(x, x)k(x', x') \quad \forall x, x' \in \mathcal{X}.$$

Proof. Since k is a kernel,

$$\begin{pmatrix} k(x, x) & k(x, x') \\ k(x', x) & k(x', x') \end{pmatrix} \geq 0.$$

Hence this has non-negative determinant and $k(x, x)k(x', x') - k(x, x')^2 \geq 0$. \square

Proposition 4. *Any similarity k defined by*

$$k(x, x') = \langle \phi(x), \phi(x') \rangle \quad \forall x, x' \in \mathcal{X}$$

is a kernel.

Proof. Symmetry of k is clear. Let $x_1, \dots, x_n \in \mathcal{X}$ be arbitrary and take any vector $\alpha \in \mathbb{R}^n$. We need to show $\alpha^T K \alpha \geq 0$. Indeed

$$\begin{aligned} \alpha^T K \alpha &= \sum_{i,j} \alpha_i K_{ij} \alpha_j \\ &= \sum_{i,j} \alpha_i \langle \phi(x_i), \phi(x_j) \rangle \alpha_j \\ &= \left\langle \sum_{i=1}^n \alpha_i \phi(x_i), \sum_{j=1}^n \alpha_j \phi(x_j) \right\rangle && \text{(linearity of } \langle \cdot, \cdot \rangle \text{)} \\ &\geq 0. && \text{(positive-definiteness of } \langle \cdot, \cdot \rangle \text{)} \end{aligned}$$

□

Examples of kernels

Proposition 5 (Closure property). *Suppose k_1, k_2, \dots are kernels on \mathcal{X} . Then*

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

Proof. Example Sheet 1. □

Some examples of kernels are:

- Linear kernel: $k(x, x') = x^T x'$ (for $\mathcal{X} = \mathbb{R}^p$);
- Polynomial kernel: $k(x, x') = (1 + x^T x')^d$, $d \in \mathbb{N}$ ($\mathcal{X} = \mathbb{R}^p$). Note $(x, x') \mapsto 1$ is a kernel so this is a kernel by the previous proposition;
- Gaussian kernel: $k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$, $\sigma^2 > 0$ the *bandwidth* of the kernel. Indeed note

$$\exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) = \underbrace{\exp\left(-\frac{\|x\|^2}{2\sigma^2}\right) \exp\left(-\frac{\|x'\|^2}{2\sigma^2}\right)}_{:=k_1(x, x')} \underbrace{\exp\left(\frac{x^T x'}{\sigma^2}\right)}_{:=k_2(x, x')}.$$

It suffices to show k_1, k_2 are kernels. For k_1 we have $k_1(x, x') = \langle \phi(x), \phi(x') \rangle$ where $\phi : \mathbb{R}^p \rightarrow \mathbb{R}$ is defined by

$$\phi(x) = \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right).$$

For k_2 we have that $(x, x') \mapsto x^T x'$ is a kernel and k_2 can be Taylor expanded so is the limit of kernels;

- Sobolev kernel: let $\mathcal{X} = [0, 1]$ and set $k(x, x') = \min(x, x') = \text{Cov}(Wx, Wx')$ where $(W_t)_{t \geq 0}$ a Brownian motion (positive definite as a covariance);
- Jaccard similarity kernel: let $\mathcal{X} = \mathcal{P}(\{1, \dots, p\})$ and set

$$k(x, x') = \begin{cases} \frac{|x \cap x'|}{|x \cup x'|} & \text{if } x \cup x' \neq \emptyset \\ 0 & \text{otherwise} \end{cases}.$$

(For proof this is a kernel see Example Sheet 1.)

Remark. There is no finite-dimensional feature map $\phi : \mathbb{R}^p \rightarrow \mathbb{R}^m$ representing the Gaussian kernel.

Theorem 6 (Moore-Aronzajn Theorem). *For every kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, there exists a feature map ϕ taking values in some inner product space \mathcal{H} such that $k(x, x') = \langle \phi(x), \phi(x') \rangle$ for all $x, x' \in \mathcal{X}$.*

Proof. Take \mathcal{H} to be the vector space of functions from \mathcal{X} to \mathbb{R} of the form

$$f(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i) \quad n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X}.$$

In other words, \mathcal{H} is the linear span of functions of the form $f(\cdot, x)$ for $x \in \mathcal{X}$. Our feature map $\phi : \mathcal{X} \rightarrow \mathcal{H}$ will be $\phi(x) = k(\cdot, x)$. We now define the inner product $\langle \cdot, \cdot \rangle$ on \mathcal{H} . Let

$$f(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i) \quad n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X}$$

and

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

Then define

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

In particular, the final two expressions show $\langle \cdot, \cdot \rangle$ is well-defined (it doesn't matter how we represent f, g as these linear combinations).

We observe directly from the definition that $\langle \phi(x), \phi(x') \rangle = \langle k(\cdot, x), k(\cdot, x') \rangle = k(x, x')$ as required. We must show $\langle \cdot, \cdot \rangle$ is indeed an inner product. It is certainly bilinear and symmetric. So we show it is positive-definite. Note that

$$\langle f, f \rangle = \sum_{i=1}^n \sum_{j=1}^n \alpha_i k(x_i, x_j) \alpha_j \geq 0 \quad (\ddagger)$$

since k is a kernel. It remains to show $\langle f, f \rangle$ implies $f(x) = 0$ for all $x \in \mathcal{X}$.

Note that $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ is a kernel. Indeed, given functions f_1, \dots, f_m and $\gamma_1, \dots, \gamma_n \in \mathbb{R}$ we have

$$\sum_{i=1}^n \sum_{j=1}^n \gamma_i \langle f_i, f_j \rangle \gamma_j = \left\langle \sum_{i=1}^n \gamma_i f_i, \sum_{j=1}^n \gamma_j f_j \right\rangle \geq 0$$

by (\ddagger) .

Now note that

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

by the Cauchy-Schwarz property for kernels. Hence $\langle f, f \rangle = 0$ implies $f(x) = 0$ for all $x \in \mathcal{X}$. \square

Remark. The space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ constructed in the proof has the property that

$$f(x) = \langle f, \underbrace{k(\cdot, x)}_{\phi(x)} \rangle.$$

As a consequence

$$|f(x) - g(x)| = |\langle f - g, k(\cdot, x) \rangle| \leq \|f - g\|_{\mathcal{H}} k(x, x)^{1/2}.$$

Hence convergence in $(\mathcal{H}, \|\cdot\|)$ implies pointwise convergence.

Lemma 7. Let \mathcal{H} be a Hilbert space and $\mathcal{V} \subseteq \mathcal{H}$ a closed subspace. Then $\mathcal{H} = \mathcal{V} \oplus \mathcal{V}^\perp$, i.e for any $f \in \mathcal{H}$ we have $f = u + v$ where $u \in \mathcal{V}$ and $v \in \mathcal{V}^\perp$ and u, v are unique.

Proof. See Part II Linear Analysis. \square

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

The function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ defined by $(x, x') \mapsto \langle k_x, k_{x'} \rangle = k_x(x')$ is known as the reproducing kernel of \mathcal{H} .

Remark. By the Riesz Representation Theorem, it is equivalent to define an RKHS as a Hilbert space where the evaluation operator $E_x : f \mapsto f(x)$ is a continuous linear operator.

The Moore-Aronzajn Theorem says that whenever k is a kernel, there is an inner product space \mathcal{H} of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ where $f(x) = \langle f, k(\cdot, x) \rangle$ and thus $k(x, x') = \langle k(\cdot, x), k(\cdot, x') \rangle$.

This implies that if $(f_n)_{n \geq 1}$ is Cauchy in \mathcal{H} ,

$$|f_n(x) - f_m(x)| \leq \sqrt{k(x, x)} \|f_n - f_m\|_{\mathcal{H}} \rightarrow 0.$$

Hence $(f_n)_{n \geq 1}$ has a pointwise limit $f^* : \mathcal{X} \rightarrow \mathbb{R}$ by completeness of \mathbb{R} . So we can complete \mathcal{H} by including all limits of Cauchy sequences (Hausdorff completion) to obtain a Hilbert space $\overline{\mathcal{H}}$. By construction, $\overline{\mathcal{H}}$ is a RKHS with reproducing kernel k .

Proposition 8. If \mathcal{G} is a RKHS of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ such that $\mathcal{G} \supseteq \mathcal{H}$, then $\overline{\mathcal{H}} = \mathcal{G}$.

Proof. Example Sheet 1. \square

Notation: from now on the RKHS is \mathcal{H} (i.e $\mathcal{H} = \overline{\mathcal{H}}$).

Examples.

- **Linear kernel:** $k(x, x') = x^T x'$. Then $\mathcal{H} = \{f : f(x) = x^T \beta, \beta \in \mathbb{R}^p\}$. If $f(x) = x^T \beta$ then $\|f\|_{\mathcal{H}}^2 = \|\beta\|^2$.
- **Sobolev kernel:** $k(x, x') = \min(x, x')$ with $\mathcal{X} = [0, 1]$. Then \mathcal{H} is the space of continuous functions $f : [0, 1] \rightarrow \mathbb{R}$ with $f(0) = 0$, for which

$$\int_0^1 |f'(x)|^2 dx < \infty$$

where f' is the weak derivative.

The Representer Theorem

If \mathcal{H} is the RKHS of the linear kernel, we can express ridge regression as

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{H}} \left\{ \sum_{i=1}^n (Y_i - \underbrace{f(x_i)}_{x_i^T \beta})^2 + \lambda \underbrace{\|f\|_{\mathcal{H}}^2}_{\|\beta\|^2} \right\}.$$

In *kernel ridge regression*, we solve this problem in a more general RKHS with kernel k , e.g the Gaussian kernel.

Theorem 9 (Representer Theorem). *Let:*

- $c : \mathbb{R}^n \times \mathcal{X}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ be an arbitrary loss;
- $J : [0, \infty) \rightarrow \mathbb{R}$ be strictly increasing;
- $x_1, \dots, x_n \in \mathcal{X}$, $Y \in \mathbb{R}^n$;
- \mathcal{H} an RKHS with representing kernel k ;
- $K_{ij} = k(x_i, x_j)$, $i, j \in [n]$.

Then \hat{f} minimises

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

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

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

Example. In kernel ridge regression we just need to solve the quadratic program

$$\hat{\alpha} = \operatorname{argmin}_{\alpha \in \mathbb{R}^n} \|Y - K\alpha\|^2 + \lambda \alpha^T K \alpha = (K + I\lambda)^{-1}.$$

Then the fitted values are given by $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$.

Intuition: to make a prediction at “test” point x^* the terms in $\hat{f}(x^*)$ that contribute the most are those for training points x_i with similarity $k(x^*, x_i)$ large.

Proof of the Representer Theorem. Note $V = \operatorname{span}\{k(\cdot, x_1), \dots, k(\cdot, x_n)\}$ is a closed (as its finite dimensional) subspace of \mathcal{H} . Hence any $f \in \mathcal{H}$ can be written as $f = u + v$ for $u \in V$ and $v \in V^\perp$.

We have $f(x_i) = \langle k(\cdot, x_i), u + v \rangle = \langle k(\cdot, x_i), u \rangle = u(x_i)$. Then

$$\|f\|_{\mathcal{H}}^2 = \|v\|_{\mathcal{H}}^2 + \|u\|_{\mathcal{H}}^2.$$

In the expression for Q_1 , the first term only depends on u , and the second term is $J(\|f\|_{\mathcal{H}}^2) \geq J(\|u\|_{\mathcal{H}}^2)$ with equality if and only if $v = 0$. Hence any minimiser

of Q_1 is contained in \mathcal{V} .

So write $f(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ for the minimiser. Now note

$$(f(x_1), \dots, f(x_n)) = K\alpha$$

$$\|f\|_{\mathcal{H}}^2 = \sum_{i,j=1}^n \alpha_i k(x_i, x_j) \alpha_j = \alpha^T K \alpha$$

so therefore for any $f \in \mathcal{V}$, $Q_1(f) = Q_2(\alpha)$. Hence $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$ minimises Q_1 if and only if $\hat{\alpha}$ minimises Q_2 . \square

Now we will assume that

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

where $\|f^0\|_{\mathcal{H}} \leq 1$.

Note. This is equivalent to $cY_i = cf^0(x_i) + c\varepsilon_i$ so $\|cf^0\|_{\mathcal{H}} = c\|f^0\|_{\mathcal{H}}$, $\text{Var}(c\varepsilon_i) = \sigma^2 c^2$. So the “signal-to-noise ratio” is

$$\frac{\text{Var}(c\varepsilon_i)}{\|cf^0\|_{\mathcal{H}}^2} = \frac{\text{Var}(\varepsilon_i)}{\|f^0\|_{\mathcal{H}}^2} \geq \sigma^2.$$

Theorem 10. Let K have eigenvalues $d_1 \geq d_2 \geq \dots \geq d_n \geq 0$. Then

$$\begin{aligned} \text{MSPE}(\hat{f}_n) &= \frac{1}{n} \mathbb{E} \left\{ \sum_{i=1}^n (f^0(x_i) - \hat{f}_n(x_i))^2 \right\} \\ &\leq \frac{\sigma^2}{n} \sum_{i=1}^n \frac{d_i^2}{(d_i + \lambda)^2} + \frac{\lambda}{4n} \\ &\leq \frac{\sigma^2}{n} \frac{1}{\lambda} \sum_{i=1}^n \min\left(\frac{d_i}{4}, \lambda\right) + \frac{\lambda}{4n}. \end{aligned}$$

Proof. From the Representer Theorem $(\hat{f}_n(x_1), \dots, \hat{f}_n(x_n))^T = K(K + \lambda I)^{-1}Y$. As $f^0 \in \mathcal{H}$ we have $(f^0(x_1), \dots, f^0(x_n))^T = K\alpha$ for some $\alpha \in \mathbb{R}^n$ (see Example Sheet). Moreover, $\|f^0\|_{\mathcal{H}}^2 \geq \alpha^T K \alpha$. Let the UDU^T be the eigen-decomposition of K , with $D_{ii} = d_i$. Define $\Theta = U^T K \alpha$. Then

$$\begin{aligned} n\text{MSPE}(\hat{f}_n) &= \mathbb{E} \left\| K(K + \lambda I)^{-1} \underbrace{(U\Theta + \varepsilon)}_Y - \underbrace{U\Theta}_{(f^0(x_1), \dots, f^0(x_n))^T} \right\|^2 \\ &= \mathbb{E} \|UDU^T(UDU^T + \lambda I)^{-1}(U\Theta + \varepsilon) - U\Theta\|^2 \\ &= \mathbb{E} \|D(D + \lambda I)^{-1}(\Theta + U^T \varepsilon) - \Theta\|^2 \quad (U^T U = I) \\ &= \underbrace{\mathbb{E} \|\{D(D + \lambda I)^{-1} - I\}\varepsilon\|^2}_{:= (1)} + \underbrace{\mathbb{E} \|D(D + \lambda I)^{-1}U^T \varepsilon\|^2}_{:= (2)}. \quad (\mathbb{E}\varepsilon = 0) \end{aligned}$$

So

$$\begin{aligned} (2) &= \mathbb{E} [\{D(D + \lambda I)^{-1}U^T \varepsilon\}^T \{D(D + \lambda I)^{-1}U^T \varepsilon\}] \\ &= \mathbb{E} [\text{tr}(\{D(D + \lambda I)^{-1}U^T \varepsilon\}^T \{D(D + \lambda I)^{-1}U^T \varepsilon\})] \\ &= \mathbb{E} [\text{tr}(D(D + \lambda I)^{-1} \varepsilon \varepsilon^T D(D + \lambda I)^{-1})] \quad (\text{circular property of tr}) \\ &= \text{tr}(D(D + \lambda I)^{-1} \sigma^2 I D(D + \lambda I)^{-1}) \\ &= \sigma^2 \sum_{i=1}^n \frac{d_i^2}{(d_i + \lambda)^2}. \end{aligned}$$

Also

$$(1) = \sum_{i=1}^n \frac{\lambda^2 \Theta_i^2}{(d_i + \lambda)^2}.$$

Since $\Theta = DU^T \alpha$, so if $d_i = 0$ then $\Theta_i = 0$. So let D^+ be a diagonal matrix with $D_{ii}^+ = \begin{cases} d_i^{-1} & \text{if } d_i \neq 0 \\ 0 & \text{otherwise} \end{cases}$.

Then,

$$\begin{aligned} \sum_{i:d_i > 0} \frac{\Theta_i^2}{d_i} &= \|\sqrt{D^+} \Theta\|^2 = \alpha^T K U D^+ U^T K \alpha \\ &= \alpha^T U D D^+ D U^T \alpha \\ &= \alpha^T U D U^T \alpha \quad (D D^+ D = D) \\ &= \alpha^T K \alpha \leq 1. \end{aligned}$$

Then

$$\begin{aligned} (1) &= \sum_{i:d_i > 0} \frac{\Theta_i^2}{d_i} \frac{d_i \lambda^2}{(d_i + \lambda)^2} \leq \max_{1 \leq i \leq n} \frac{d_i \lambda^2}{(d_i + \lambda)^2} \sum_{i:d_i > 0} \frac{\Theta_i^2}{d_i} \\ &\leq \max_{1 \leq i \leq n} \frac{d_i \lambda^2}{(d_i + \lambda)^2} \\ &\leq \frac{\lambda}{4}. \quad ((a+b)^2 \geq 4ab) \end{aligned}$$

Combining the bounds for (1) and (2) gives the first inequality. Finally, for the final inequality we note that

$$\frac{d_i^2}{(d_i + \lambda)^2} \leq \min \left\{ 1, \frac{d_i^2}{4d_i \lambda} \right\} = \frac{1}{\lambda} \min \left\{ \lambda, \frac{d_i}{4} \right\}.$$

□

Question: when is the upper bound good?

Random design

Let $(\mathcal{X}, \mathcal{B}, \mathbb{P})$ be a probability space, where \mathcal{X} is a metric space, \mathcal{B} is the Borel σ -algebra on \mathcal{X} . Assume that $x_1, \dots, x_n \sim^{\text{iid}} \mathbb{P}$.

Theorem 11 (Mercer's Theorem). *Under mild assumptions on k, \mathbb{P} , there is an orthonormal basis (e_i) of $\mathcal{L}^2(\mathbb{P})$, i.e*

$$\int_{\mathcal{X}} e_l(x) e_j(x) d\mathbb{P}(x) = \mathbb{1}\{l = j\}$$

and eigenvalues (μ_i) with $\sum_{i=1}^n \mu_i < \infty$ such that

$$\mu_j e_j(x') = \int_{\mathcal{X}} k(x, x') e_j(x) d\mathbb{P}(x).$$

Furthermore

$$k(x, x') = \sum_{l=1}^{\infty} \mu_l e_l(x) e_l(x')$$

and this series is absolutely convergent.

Proof. Not given. □

Let $\hat{\mu}_1, \dots, \hat{\mu}_n$ be (random) eigenvalues of K/n . As it turns out, when n is large $\hat{\mu}_i \approx \mu_i$. Let $\gamma = \lambda/n$, then a previous theorem gives

$$\text{MSPE}(\hat{f}_{\gamma n}) \leq \frac{\sigma^2}{\gamma} \frac{1}{n} \sum_{i=1}^n \min\left(\frac{\hat{\mu}_i}{4}, \gamma\right) + \frac{\gamma}{4}.$$

Then the MSPE is a random variable depending on x_1, \dots, x_n .

Lemma 12.

$$\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n \min\left(\frac{\hat{\mu}_i}{4}, \gamma\right)\right) \leq \frac{1}{n} \sum_{i=1}^n \min\left(\frac{\mu_i}{4}, \gamma\right).$$

Proof. Not given. □

This lemma means we can bound

$$\underbrace{\mathbb{E}[\text{MSPE}(\hat{f}_{n\gamma})]}_{\text{over } Y \text{ and } x_1, \dots, x_n} \leq \frac{\sigma^2}{\gamma} \frac{1}{n} \sum_{i=1}^n \min\left(\frac{\mu_i}{4}, \gamma\right) + \frac{\gamma}{4}. \quad (*)$$

Theorem 13. *Under the assumptions of Mercer's Theorem, there is a sequence $(\gamma_n)_{n \geq 1}$ such that for fixed $\sigma^2 > 0$,*

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

Proof. Let $\phi : [0, \infty) \rightarrow [0, \infty)$ be defined by

$$\phi(\gamma) = \sum_{j=1}^{\infty} \min \left(\frac{\mu_j}{4}, \gamma \right).$$

Note ϕ is increasing, and as $\sum_{j=1}^{\infty} \mu_j < \infty$, $\lim_{\gamma \downarrow 0} \phi(\gamma) = 0$. Define $\gamma_n = n^{-1/2} \sqrt{\phi(n^{-1/2})}$, so $\gamma_n = o(n^{-1/2})$. Thus for n large enough, $\phi(\gamma_n) \leq \phi(n^{-1/2})$ and the upper bound in (*) is

$$\sigma^2 \frac{\phi(\gamma_n)}{n\gamma_n} + \frac{\gamma_n}{4} \leq \frac{\sigma^2 \phi(n^{-1/2})}{n^{1/2} \sqrt{\phi(n^{-1/2})}} + o(n^{-1/2}) = o(n^{-1/2}).$$

□

When we know (μ_j) , in some cases we can get a better bound on the MSPE.

Example. If k is the Sobolev kernel and \mathbb{P} is the Lebesgue measure on $[0, 1]$, one can show that

$$\frac{\mu_i}{4} = \frac{1}{\pi^2(2i-1)^2}.$$

Then for any integer j ,

$$\sum_{i=1}^{\infty} \min \left(\frac{\mu_i}{4}, \gamma_n \right) \leq \gamma_n j + \sum_{i=j+1}^{\infty} \frac{1}{\pi^2(2i-1)^2}.$$

So if we take $j = \frac{(\pi^2 \gamma_n)^{-1/2} + 1}{2}$ we get upper bound

$$\begin{aligned} & \frac{\gamma_n}{2} \left(\frac{1}{\sqrt{\pi^2 \gamma_n}} + 1 \right) + \frac{1}{\pi^2} \int_{(\pi^2 \gamma_n)^{-1/2} + 1}^{\infty} \frac{1}{(2x-1)^2} dx \\ &= \mathcal{O}(\gamma_n^{1/2}) + \mathcal{O}(\gamma_n) = \mathcal{O}(\sqrt{\gamma_n}). \end{aligned}$$

By (*) we have

$$\mathbb{E}(\text{MSPE}(\hat{f}_{\gamma_n, n})) \leq \mathcal{O} \left(\frac{\sigma^2}{n\gamma_n} \sqrt{\gamma_n} + \gamma_n \right).$$

Picking $\gamma_n \sim \left(\frac{\sigma^2}{n} \right)^{2/3}$ gives an error of at most $\mathcal{O} \left(\left(\frac{\sigma^2}{n} \right)^{2/3} \right)$.

Support Vector Machines

Suppose we have data $(x_i, Y_i)_{i \in [n]}$ where $x_i \in \mathbb{R}^p$, $Y_i \in \{-1, 1\}$. Suppose the two response classes can be separated by a hyperplane through the origin. Let β be a unit vector which is normal to the hyperplane.

There could be many separating hyperplanes. One way of choosing a single one of these is to maximise an empty margin, i.e

$$\max_{\substack{M > 0 \\ \beta \in \mathbb{S}^{p-1}}} M \text{ subject to } Y_i x_i^T \beta \geq M \text{ for all } i \in [n].$$

Reparameterising by $\beta \rightarrow \beta/M$, this problem becomes

$$\max_{\beta \in \mathbb{R}^p} \frac{1}{\|\beta\|} \text{ subject to } Y_i x_i^T \beta \geq 1 \text{ for all } i \in [n]$$

or equivalently

$$\min_{\beta \in \mathbb{R}^p} \|\beta\|^2 \text{ subject to } Y_i x_i^T \beta \geq 1 \text{ for all } i \in [n].$$

Instead, what if just a few samples fall on the wrong side of the margin? A different estimator, known as a *support vector classifier* replaces the constraint $Y_i x_i^T \beta \geq 1$ with a penalty $[(1 - Y_i)x_i^T \beta]_+$.

Remark. This works even if there is no separating hyperplane.

So our problem is

$$\min_{\beta \in \mathbb{R}^p} \left[\lambda \|\beta\|^2 + \sum_{i=1}^n (1 - Y_i x_i^T \beta)_+ \right].$$

λ is a tuning parameter which balances “maximum margin” objective and penalty.

In general, we may want to estimate a hyperplane which does not pass through the origin; $x^T \beta + \mu = 0$. We can define a similar optimisation:

$$\min_{\substack{\beta \in \mathbb{R}^p \\ \mu \in \mathbb{R}}} \left[\lambda \|\beta\|^2 + \sum_{i=1}^n (1 - Y_i(x_i^T \beta + \mu))_+ \right].$$

If \mathcal{H} is the RKHS for the linear kernel, this problem can be written as

$$(\hat{\mu}, \hat{f}) = \operatorname{argmin}_{(\mu, f) \in \mathbb{R} \times \mathcal{H}} \left[\sum_{i=1}^n (1 - Y_i(f(x_i) + \mu))_+ + \lambda \|f\|_{\mathcal{H}}^2 \right]$$

where $\hat{f}(x) = x^T \hat{\beta}$.

A *support vector machine* is defined by this optimisation with a generic RKHS \mathcal{H} with reproducing kernel k .

Prediction: given $(\hat{\mu}, \hat{f})$ and a new input x^* we predict $\hat{Y}^* = \operatorname{sgn}(\hat{f}(x^*) + \hat{\mu})$.

Note. In \mathcal{X} the separating ‘hyperplane’ is not necessarily linear, but upon mapping (via ϕ) to \mathcal{H} (i.e $x \mapsto \phi(x) = k(\cdot, x)$) it becomes a hyperplane since the class boundary $\{x \in \mathcal{X} : f(x) + \mu = 0\}$ is mapped to $\{k(\cdot, x) : \langle k(\cdot, x), f \rangle_{\mathcal{H}} + \mu = 0\}$.

Using a slight generalisation of the Representer Theorem (see Example Sheet 1), we can show that

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

where

$$(\hat{\alpha}, \hat{\mu}) = \operatorname{argmin}_{(\alpha, \mu) \in \mathbb{R}^n \times \mathbb{R}} \sum_{i=1}^n (1 - Y_i(K_i^T \alpha + \mu))_+ + \lambda \alpha^T K \alpha$$

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

Remark. We can have $\hat{\alpha}_i = 0$ for some i , so we do not use the corresponding x_i at all in the estimator.

Kernel Logistic Regression

We have standard logistic regression

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

Maximising the likelihood with (x_i, Y_i) , $i \in [n]$ is equivalent to solving

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

As in ridge regression, we may wish to penalise $\|\beta\|^2$:

$$\min_{\beta \in \mathbb{R}^p} \left[\sum_{i=1}^n \log (1 + \exp(-Y_i x_i^T \beta)) + \lambda \|\beta\|_2^2 \right].$$

This is the same as

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

where \mathcal{H} is the linear RKHS.

In kernel logistic regression we build the class boundary $\hat{f}(\cdot)$ by solving this problem with an arbitrary RKHS.

Question: how does this compare with the Support Vector Machine?

In each case, the objective is

$$\sum_{i=1}^n l(Y_i f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2$$

with $l(z) = (1 - z)_+$ and $l(z) = \log(1 + e^{-z})$ for the SVM and logistic regression respectively.

1.5 Large-scale Kernel Machines

Suppose for a kernel k , there is a feature map $\phi : \mathcal{X} \rightarrow \mathbb{R}^q$. Let $K_{ij} = k(x_i, x_j)$

and $\Phi = \begin{pmatrix} \phi(x_1) \\ \vdots \\ \phi(x_n) \end{pmatrix} \in \mathbb{R}^{n \times q}$ so that $K = \Phi \Phi^T$.

Consider kernel ridge regression. There are two ways of computing the fitted values:

$$K(\underbrace{K + I\lambda}_{n \times n})^{-1} Y \text{ or;}$$

$$\Phi(\underbrace{\Phi^T \Phi + \lambda I}_{q \times q})^{-1} \Phi^T Y.$$

These have costs $\mathcal{O}(n^3)$ and $\mathcal{O}(q^3 + nq^2)$ respectively. So when n is much larger than q , we want to use the latter expression.

In other kernel machines, it is helpful to have a low rank kernel matrix $K = \Phi \Phi^T$.

Example. Consider the optimisation problem resulting from the representer theorem

$$\min_{\alpha \in \mathbb{R}^n} [c(Y, x_1, \dots, x_n, K\alpha) + \lambda \alpha^T K \alpha].$$

The gradient of the penalty term is $2\lambda K\alpha$. Computing this has cost $\mathcal{O}(n^2)$ (since K is $n \times n$), but if $K = \Phi \Phi^T$ ($q < n$) we can compute $2\lambda \Phi \Phi^T \alpha$ in $\mathcal{O}(nq)$ iterations.

Problem: what if there is no feature map ϕ onto \mathbb{R}^q with $q \ll n$? For example, the Gaussian kernel.

Idea: find an approximation $\hat{\Phi}$ such that $K \approx \hat{\Phi} \hat{\Phi}^T$. Our approach will be to develop a random feature map $\hat{\Psi} : \mathcal{X} \rightarrow \mathbb{R}^b$ satisfying

$$\mathbb{E}[\hat{\Psi}(x)^T \hat{\Psi}(x')] = k(x, x') \text{ for all } x, x' \in \mathcal{X}.$$

Then, we can let $\hat{\Psi}_i, i \in [L]$ be iid copies of $\hat{\Psi}$; define the approximate feature map

$$\hat{\phi} : x \mapsto \frac{1}{\sqrt{L}} (\hat{\Psi}_1(x), \dots, \hat{\Psi}_L(x)) \in \mathbb{R}^{b \times L}.$$

Then $\hat{\phi}(x)^T \hat{\phi}(x') = \frac{1}{L} \sum_{i=1}^L \hat{\Psi}_i(x)^T \hat{\Psi}_i(x')$. In particular $\mathbb{E}[\hat{\phi}(x)^T \hat{\phi}(x)] = k(x, x')$ and $\text{Var}[\hat{\phi}(x)^T \hat{\phi}(x)] = \mathcal{O}(L^{-1})$.

Then approximate $K \approx \hat{\Phi} \hat{\Phi}^T$ where $\hat{\Phi} = \begin{pmatrix} \hat{\phi}(x_1) \\ \vdots \\ \hat{\phi}(x_n) \end{pmatrix}$. In some cases the error

$$\|K - \hat{\Phi}^T \hat{\Phi}\|$$

is small with $Lb \ll n$.

Random Fourier Feature

Theorem 14 (Bochner). *Let $k : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ be a continuous kernel. Then k is shift-invariant (there exists h such that $k(x, x') = h(x - x')$) if and only if there exists $c > 0$ and some distribution F in \mathbb{R}^p such that if $W \sim F$, then*

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

Proof. Not given. □

Example. If $k(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$ is the Gaussian kernel, then we have the representation in the theorem with $W \sim \mathcal{N}(0, \sigma^{-2}I)$.

We can use the theorem to construct a random feature map

$$\hat{\Psi}(x) = \sqrt{2c} \cos(W^T x + U) \in \mathbb{R}$$

where $W \sim F$, $U \sim \text{Unif}(-\pi, \pi)$ are independent.

Lemma 15.

$$\mathbb{E}(\hat{\Psi}(x)\hat{\Psi}(y)) = k(x, y) \text{ for all } x, y \in \mathbb{R}^p.$$

Proof. The LHS is

$$\begin{aligned} & 2c\mathbb{E}[\cos(W^T x + U) \cos(W^T y + U)] \\ &= 2c\mathbb{E}[\cos(W^T x) \cos U - \sin(W^T x) \sin U] \times [\cos(W^T y) \cos U - \sin(W^T y) \sin U] \\ &= c\mathbb{E}[\cos(W^T x) \cos(W^T y) + \sin(W^T x) \sin(W^T y)] \quad (\text{since } \mathbb{E}[\cos u \sin u] = 0) \\ &= c\mathbb{E}[\cos(W^T(x - y))] \\ &= k(x, y). \end{aligned} \quad (\text{Bochner's Theorem})$$

□

2 The Lasso & Beyond

Consider the standard linear model

$$Y = X\beta^0 + \varepsilon, \quad \mathbb{E}\varepsilon = 0, \text{Var } \varepsilon = \sigma^2 I.$$

Then

$$\begin{aligned} \text{MSPE}(\hat{\beta}^{\text{ols}}) &= \frac{1}{n} \mathbb{E} \|X\beta^0 - X\hat{\beta}^{\text{ols}}\|^2 \\ &= \frac{1}{n} \mathbb{E} [\text{tr}((\beta^0 - \hat{\beta}^{\text{ols}})(\beta^0 - \hat{\beta}^{\text{ols}})^T X^T X)] \\ &= \frac{1}{n} \text{tr} \left(\underbrace{\mathbb{E}[(\beta^0 - \hat{\beta}^{\text{ols}})(\beta^0 - \hat{\beta}^{\text{ols}})^T]}_{\text{Var}(\hat{\beta}^{\text{ols}})} X^T X \right) \\ &= \frac{1}{n} \text{tr}(\sigma^2 (X^T X)^{-1} X^T X) \\ &= \frac{1}{n} \text{tr}(\sigma^2 I_p) = \frac{\sigma^2 p}{n}. \end{aligned}$$

Let $S = \{k : \beta_k^0 \neq 0\}$ be the “relevant” predictors.

Question: what if $s := |S| \ll p$?

The model $Y = X_s \beta_s^0 + \varepsilon$, where X_s is the matrix with columns which are columns of X with index in S , and β_s^0 are the coefficients for predictors in S . So if we fit a model with design matrix X_s instead of X , we get $\text{MSPE} = \frac{\sigma^2 s}{n} \ll \frac{\sigma^2 p}{n}$.

In practice, we don’t know S , but we can try to estimate it (variable selection).

Best subset regression

Fit every model with a subset $M \subseteq \{1, \dots, p\}$ of the predictors. Then choose the best M by cross-validation.

Problem: there are 2^p possibilities, which is too large even for relatively small p .

Forward selection

This is a greedy way of approximating best subset regression.

1. Start by fitting intercept-only model;
2. Add to the model the predictor that decreases the sum-of-squares residuals the most;
3. Repeat step 2 until we have m predictors.

We treat m as a tuning parameter, chosen by cross-validation.

2.1 The Lasso

$$(\hat{\mu}_\lambda^L, \hat{\beta}_\lambda^L) \in \operatorname{argmin}_{(\mu, \beta) \in \mathbb{R} \times \mathbb{R}^p} \left[\frac{1}{2n} \|Y - \mu \mathbf{1} - X\beta\|^2 + \lambda \|\beta\|_1 \right]$$

where $\|\beta\|_1 = \sum_{k=1}^p |\beta_k|$. As we did for ridge regression, we can remove μ by standardising the columns of X and centering the response Y :

$$\hat{\beta}_\lambda^L \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \left[\frac{1}{2n} \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right].$$

Note that $\hat{\beta}_\lambda^L$ minimises

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

Similarly, $\hat{\beta}_\lambda^R$ minimises

$$\|Y - X\beta\|^2 \text{ subject to } \|\beta\| \leq \|\hat{\beta}_\lambda^R\|.$$

Fact: in general, $\hat{\beta}_\lambda^R$ has all non-zero entries, whereas $\hat{\beta}_\lambda^L$ can have many entries equal to zero.

Prediction error of the Lasso (slow rate)

Assume the columns of X are standardised, and Y is the centred response

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

Further assume $\varepsilon \in \mathcal{N}(0, \sigma^2 I)$.

Theorem 16. Let $\hat{\beta}$ be any Lasso solution with $\lambda = A\sigma\sqrt{\log(p)/n}$. Then with probability $\geq 1 - 2p^{-(A^2/2-1)}$,

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

Remarks.

- Instead of bounding the MSPE (the expectation of the LHS) we bound the SPE with high-probability;
- This is called the “slow rate” with respect to n , since we know the MSPE usually decreases as $\mathcal{O}(n^{-1})$ (e.g MSPE of $\hat{\beta}^{\text{ols}}$).
- However, we trade the factor of p in the numerator with $\sqrt{\log p} \|\beta^0\|_1$, which can be much smaller than in OLS in general.
- We make no assumptions about X !

Lemma 17. Let $\|X^T \varepsilon\|_\infty = \max_k |(\varepsilon^T X)_k|$ and let $\Omega = \left\{ \frac{\|X^T \varepsilon\|_\infty}{n} \leq \lambda \right\}$ then

$$\mathbb{P}(\Omega) \geq 1 - 2p^{-(A^2/2-1)}.$$

Proof of slow rate. By the definition of $\hat{\beta}$

$$\frac{1}{2n} \left\| \underbrace{Y - X\hat{\beta}}_{X(\beta^0 - \hat{\beta}) + \varepsilon - \mathbf{1}\bar{\varepsilon}} \right\|^2 + \lambda \|\hat{\beta}\|_1 \leq \frac{1}{2n} \left\| \underbrace{Y - X\beta^0}_{\varepsilon - \mathbf{1}\bar{\varepsilon}} \right\|^2 + \lambda \|\beta^0\|_1.$$

Since $X^T \mathbf{1} = 0$, rearranging terms (and using the previous lemma) gives

$$\begin{aligned} \frac{1}{2n} \|X(\beta^0 - \hat{\beta})\|^2 &\leq \frac{1}{n} \varepsilon^T X(\beta^0 - \hat{\beta}) + \lambda \|\beta^0\|_1 - \lambda \|\hat{\beta}\|_1 \\ &\leq \|\varepsilon^T X\|_\infty \|\beta^0 - \hat{\beta}\|_1 + \lambda \|\beta^0\|_1 - \lambda \|\hat{\beta}\|_1 \\ &\leq \lambda \left[\|\beta^0 - \hat{\beta}\|_1 + \|\beta^0\|_1 - \|\hat{\beta}\|_1 \right]. \quad (\text{On } \Omega) \end{aligned}$$

Thus by the triangle inequality

$$\frac{1}{n} \|X(\beta^0 - \hat{\beta})\|^2 \leq 4\lambda \|\beta^0\|_1.$$

□

Concentration Inequalities

Definition. We say a random variable W is σ -sub-Gaussian for some parameter $\sigma > 0$ if

$$\mathbb{E}[e^{\alpha(W - \mathbb{E}W)}] \leq e^{\frac{\alpha^2 \sigma^2}{2}}.$$

Proposition 18. If W is σ -sub-Gaussian, then

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

Proof. Apply Markov's inequality to $\mathbb{P}(W - \mathbb{E}W \geq t) = \mathbb{P}(\exp(\alpha(W - \mathbb{E}W)) \geq \exp(\alpha t))$ and minimise over α (Chernoff bound). \square

All bounded random variables are sub-Gaussian.

Lemma 19. If W is a random variable taking values in $[a, b]$, then it is $(\frac{b-a}{2})$ -sub-Gaussian.

Proof. See Part III Topics in Statistical Theory. \square

Proposition 20. Let W_1, \dots, W_n be independent random variables where W_i is σ_i -sub-Gaussian. Let $\gamma \in \mathbb{R}^n$. Then $\gamma^T W = \sum_{i=1}^n \gamma_i W_i$ is sub-Gaussian with parameter $\sqrt{\sum_{i=1}^n \gamma_i^2 \sigma_i^2}$.

Proof. Without loss of generality, assume $\mathbb{E}W_i = 0$ for all $i \in [n]$. Then

$$\mathbb{E} \left[\exp \left(\alpha \sum_{i=1}^n \gamma_i W_i \right) \right] = \prod_{i=1}^n \mathbb{E} [\exp(\alpha \gamma_i W_i)] \leq \exp \left(\alpha^2 \sum_{i=1}^n \frac{\gamma_i^2 \sigma_i^2}{2} \right).$$

\square

Recall:

Lemma 21. Let $\|X^T \varepsilon\|_\infty = \max_k |(\varepsilon^T X)_k|$ and let $\Omega = \left\{ \frac{\|X^T \varepsilon\|_\infty}{n} \leq \lambda \right\}$ then

$$\mathbb{P}(\Omega) \geq 1 - 2p^{-(A^2/2-1)}.$$

We will prove a stronger result:

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

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

Proof. We have

$$\begin{aligned}\mathbb{P}\left(\frac{\|X^T \varepsilon\|_\infty}{n} > \lambda\right) &\leq \sum_{j=1}^p \mathbb{P}\left(\frac{|X_j^T \varepsilon|}{n} > \lambda\right) \\ &= \sum_{j=1}^p \left[\mathbb{P}\left(\frac{X_j^T}{n} > \lambda\right) + \mathbb{P}\left(-\frac{X_j^T}{n} > \lambda\right) \right].\end{aligned}$$

By the previous proposition, $\pm \frac{X_j^T \varepsilon}{n}$ is mean zero sub-Gaussian with parameter $\left(\frac{\sigma^2 \|X_j\|^2}{n}\right)^{1/2} = \frac{\sigma}{\sqrt{n}}$. Hence, the above expression is bounded above by

$$2p \exp\left(-\frac{\lambda^2}{\left(2\frac{\sigma^2}{n}\right)}\right) = 2p \exp\left(-A^2 \frac{\log p}{2}\right) = 2p^{1-A^2/2}.$$

□

Now we recall some facts from complex analysis.

Proposition 22. *Let $C \subseteq \mathbb{R}^d$ be convex.*

- (i) *Let $f_1, \dots, f_m : C \rightarrow \mathbb{R}$ be convex and $c_1, \dots, c_m \geq 0$. Then $c_1 f_1 + \dots + c_m f_m$ is convex.*
- (ii) *Let $A : \mathbb{R}^m \rightarrow \mathbb{R}^d$ be affine ($A(x) = Mx + b$). Let $D = A^{-1}(C) = \{x : A(x) \in C\}$ and let $f : C \rightarrow \mathbb{R}$ be convex. Then D is convex and the composition $f \circ A : D \rightarrow \mathbb{R}$, $x \mapsto f(A(x))$ is a convex function.*
- (iii) *If $f : C \rightarrow \mathbb{R}$ is twice continuously differentiable with C open, then*
 - (a) *f is convex if and only if its Hessian $H(x)$ is positive semi-definite for all $x \in C$;*
 - (b) *f is strictly convex if its Hessian $H(x)$ is positive definite for all $x \in C$.*

Lagrangian method

Consider a problem of the form

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

where $g : C \rightarrow \mathbb{R}^b$. Let c^* be the optimal value. The *Lagrangian* of this problem is defined as

$$L(x, \theta) = f(x) + \theta^T g(x), \quad \theta \in \mathbb{R}^b.$$

Note that for all θ ,

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

The Lagrangian method involves finding θ^* such that the minimiser x^* of the LHS in the above has $g(x^*) = 0$, in which case this is a minimiser of the original problem.

Subgradient

Definition. Given a convex $C \subseteq \mathbb{R}^d$, convex $f : C \rightarrow \mathbb{R}$, define the *subdifferential* of f at x $\partial f(x) \subseteq \mathbb{R}^d$ defined by

$$\partial f(x) = \{v \in \mathbb{R}^d : f(y) \geq f(x) + v^T(y - x) \quad \forall y \in C\}.$$

An element $v \in \partial f(x)$ is called a *subgradient* of f at x .

Proposition 23. If $f : C \rightarrow \mathbb{R}$ is convex and differentiable at $x \in \text{int}(C)$, then

$$\partial f(x) = \{\nabla f(x)\}.$$

Proposition 24. If $f, g : C \rightarrow \mathbb{R}$ are convex with $\text{int}(C) \neq \emptyset$, then

$$\partial(\alpha f)(x) = \alpha \partial f(x) = \{\alpha v : v \in \partial f(x)\}$$

for all $\alpha \in \mathbb{R}$. Also

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

Karush-Kuhn-Tucker (KKT) conditions

Proposition 25. *Given $f : C \rightarrow \mathbb{R}$ convex, $x^* \in \operatorname{argmin}_{x \in C} f(x)$ if and only if $0 \in \partial f(x^*)$.*

Proof. Trivial. □

Subdifferential of $\|\cdot\|_1$

By the triangle inequality,

$$\|tx + (1-t)y\|_1 \leq t\|x\|_1 + (1-t)\|y\|_1 \quad \forall t \in (0, 1)$$

so $\|\cdot\|_1$ is convex. Let $A = \{k_1, \dots, k_m\} \subseteq \{1, \dots, d\}$. For $x \in \mathbb{R}^d$ we write x_A for the vector $(x_{k_1}, \dots, x_{k_m}) \in \mathbb{R}^m$. For $X \in \mathbb{R}^{n \times d}$ we write X_A for the matrix with columns $(X_{k_1}, \dots, X_{k_m})$. We also write x_{-j} and x_{-jk} for $x_{\{j\}^c}$ and $x_{\{j,k\}^c}$ respectively. Also X_A^T and X_A^{-1} denote $(X_A)^T$ and $(X_A)^{-1}$ respectively.

We also define the sgn function by

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

for $x_i \in \mathbb{R}$. For $x \in \mathbb{R}^d$ we define $\operatorname{sgn}(x) = (\operatorname{sgn}(x_1), \dots, \operatorname{sgn}(x_d))$.

Proposition 26. *For $x \in \mathbb{R}^d$, let $A = \{j : x_j \neq 0\}$. Then*

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

Proof. Let $g_j : \mathbb{R}^d \rightarrow \mathbb{R}$ be defined by $x \mapsto |x_j|$. Define $g = \sum_{j=1}^d g_j$ so $g(x) = \|x\|_1$ for all $x \in \mathbb{R}^d$. Then by a previous proposition, $\partial g(x) = \sum_{j=1}^d \partial g_j(x)$. When $x_j \neq 0$, g_j is differentiable at x so $\partial g_j(x) = \{\partial g_j(x)\} = \{\operatorname{sgn}(x_j)e_j\}$, where $e_j \in \mathbb{R}^d$ has all entries 0 except j th entry 1.

When $x_j = 0$,

$$\begin{aligned} v \in \partial g(x_j) &\iff g_j(y) \geq g_j(x) + v^T(y - x) \quad \forall y \in \mathbb{R}^d \\ &\iff |y_j| \geq |x_j| + v^T(y - x) \quad \forall y \in \mathbb{R}^d \\ &\iff |y_j| \geq v^T(y - x) \quad \forall y \in \mathbb{R}^d. \end{aligned}$$

We claim this holds if and only if $v_j \in [-1, 1]$ and $v_{-j} = 0$. Indeed if $v_{-j} = 0$ the above becomes $|y_j| \geq v_j y_j$ which holds as long as $v_j \in [-1, 1]$. Conversely, taking $y \in \mathbb{R}^d$ with $y_j = 0$ and $y_{-j} = x_{-j} + v_{-j}$ gives $0 = |y_j| \geq |v_{-j}|^2$, implying $v_{-j} = 0$. Taking $y \in \mathbb{R}^d$ with $y_{-j} = 0$ and $y_j = \operatorname{sgn}(v_j)$ gives $1 \geq |\operatorname{sgn}(v_j)| = |y_j| \geq v_j \operatorname{sgn}(v_j) = |v_j|$.

The proposition now follows from $\partial g(x) = \sum_{j=1}^d \partial g_j(x)$. □

Lasso solutions

We have

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

KKT says that $0 \in \partial Q_\lambda(\hat{\beta}_\lambda^L)$ for any solution $\hat{\beta}_\lambda^L$. This is equivalent to

$$-\frac{1}{n} X^T (Y - X\hat{\beta}_\lambda^L) + \lambda \hat{v} = 0$$

for $\hat{v} \in \partial \|\hat{\beta}_\lambda^L\|_1$, i.e. $\|\hat{v}\|_\infty \leq 1$ and $\hat{v}_{\hat{S}_\lambda} = \text{sgn}(\hat{\beta}_{\lambda, \hat{S}_\lambda}^L)$, where $\hat{S}_\lambda = \{k : \hat{\beta}_{\lambda, k}^L \neq 0\}$.

It turns out the Lasso fitted values are unique!

Proposition 27. Fix $\lambda > 0$, suppose $\beta^{(1)}, \beta^{(2)}$ are two lasso solutions. Then $X\beta^{(1)} = X\beta^{(2)}$.

Proof. We have $Q_\lambda(\beta^{(1)}) = Q_\lambda(\beta^{(2)}) = c^*$, the optimal value of Q_λ . By strict convexity of $\|\cdot\|_2^2$,

$$\left\| Y - \frac{X\beta^{(1)}}{2} - \frac{X\beta^{(2)}}{2} \right\|_2^2 \leq \frac{\|Y - X\beta^{(1)}\|_2^2}{2} + \frac{\|Y - X\beta^{(2)}\|_2^2}{2}$$

with equality if and only if $X\beta^{(1)} = X\beta^{(2)}$. We'll show this is an equality. We construct a chain of inequalities

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

Hence all of these inequalities are in fact equalities. The equality at $(*)$ implies $X\beta^{(1)} = X\beta^{(2)}$. \square

Define the *equicorrelation set*

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

which is well defined, since $X\hat{\beta}_\lambda^L$ does not depend on the choice of $\hat{\beta}_\lambda^L$. The KKT conditions tell us that \hat{E}_λ contains all non-zero entries in $\hat{\beta}_\lambda^L$, i.e. \hat{S}_λ .

Proposition 28. If $\text{rank}(X_{\hat{E}_\lambda}) = |\hat{E}_\lambda|$, then $\hat{\beta}_\lambda^L$ is unique.

Proof. Say $\beta^{(1)}, \beta^{(2)}$ are lasso solutions. We know $\beta_{-\hat{E}_\lambda}^{(1)} = \beta_{-\hat{E}_\lambda}^{(2)} = 0$ so

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

by uniqueness of the fitted values. If $X_{\hat{E}_\lambda}$ has full column rank, the only solution is $\beta_{\hat{E}_\lambda}^{(1)} - \beta_{\hat{E}_\lambda}^{(2)} = 0$. \square

Variable selection

How good is the lasso at recovering the non-zero entries of β^0 ? When do we have $\text{sgn}(\hat{\beta}_\lambda^L) = \text{sgn}(\beta^0)$?

In the lectures we'll deal with the noiseless case $Y = X\beta^0$ (see Example Sheet for more general case).

Define $S = \{k : \beta_k^0 \neq 0\}$ and $N = S^c$. Assume without loss of generality that $S = \{1, \dots, s\}$. Assume $\text{rank}(X_S) = s$. Define $\Delta = X_N^T X_S (X_S^T X_S)^{-1} \text{sgn}(\beta_S^0)$ so $\Delta_k = ((X_S^T X_S)^{-1} X_S^T X_k) \text{sgn}(\beta_S^0)$, for $k > s$. Note that $(X_S^T X_S)^{-1} X_S^T X_k$ represents the coefficients in regression of non-significant variables X_k onto significant variables X_S .

Fix $\lambda > 0$ and define conditions

- (A) $\|\Delta\|_\infty \leq 1$ (irrepresentable condition);
- (B) $|\beta_k^0| > \lambda |\text{sgn}(\beta_S^0)^T (X_S^T X_S)^{-1}|$ for all $k \in S$ (strong signal);
- (C) There exists a lasso solution $\hat{\beta}_\lambda^L$ with $\text{sgn}(\hat{\beta}_\lambda^L) = \text{sgn}(\beta^0)$.

Theorem 29. *A & B imply C, and C implies A.*

Proof. Write $\hat{\beta} = \hat{\beta}_\lambda^L$, $\hat{S} = \{k : \hat{\beta}_k \neq 0\}$. The KKT conditions for the lasso say

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

where $\|\hat{v}\|_\infty \leq 1$ and $v_{\hat{S}} = \text{sgn}(\hat{\beta}_{\hat{S}})$. Blockwise, this is

$$\frac{1}{n} \begin{pmatrix} X_S^T X_S & X_S^T X_N \\ X_N^T X_S & X_N^T X_N \end{pmatrix} \begin{pmatrix} \beta_S^0 - \hat{\beta}_S \\ -\hat{\beta}_N \end{pmatrix} = \lambda \begin{pmatrix} \hat{v}_S \\ \hat{v}_N \end{pmatrix}. \quad ((I))$$

First we show C implies A: if $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^0)$ then $\hat{v}_S = \text{sgn}(\beta_S^0)$ and $\hat{\beta}_N = 0$. The top block of (I) gives

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

Since X_S has full rank, $X_S^T X_S$ is invertible and $\beta_S^0 - \hat{\beta}_S = n\lambda (X_S^T X_S)^{-1} \text{sgn}(\beta_S^0)$. Plugging this into the bottom block of (I) gives

$$\lambda \frac{1}{n} X_N^T X_S (n(X_S^T X_S)^{-1} \text{sgn}(\beta_S^0)) = \lambda \hat{v}_N$$

so

$$\Delta = X_N^T X_S (X_S^T X_S)^{-1} \text{sgn}(\beta_S^0) = \hat{v}_N.$$

Since $\|\hat{v}\|_\infty \leq 1$ we have $\|\Delta\|_\infty \leq 1$, i.e (A).

Now we show A & B imply C: it is enough to exhibit $(\hat{\beta}, \hat{v})$ satisfying (I) with $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^0)$. Take

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

It is easy to verify this solves (I). By condition (A), $\|\Delta\|_\infty \leq 1$ so $\|\hat{v}\|_\infty \leq 1$. Now we just check $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^0)$. Condition (B) gives that $\text{sgn}(\beta_S^0) = \text{sgn}(\hat{\beta}_S)$, and $\hat{\beta}_N = 0 = \beta_N^0$. Finally, we need $\hat{v}_{\hat{S}} = \text{sgn}(\hat{\beta}_{\hat{S}})$. But by construction, $\hat{S} = S$ and we showed that $\hat{v}_S := \text{sgn}(\beta_S^0) = \text{sgn}(\hat{\beta}_S)$. \square

Prediction & Estimation (fast rates)

Return to a noisy linear model

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

where ε_i are independent and σ -sub-Gaussian random variables.

Definition. Given $X \in \mathbb{R}^{n \times p}$ and $S \subseteq \{1, \dots, p\}$ non-empty, define the *compatibility factor*

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

for $s := |S|$.

Remark. This is similar to a variational characterisation of the smallest eigenvalue c_{\min} of $\frac{1}{n}X^T X$:

$$c_{\min} = \inf_{\substack{\delta \in \mathbb{R}^p \\ \delta \neq 0}} \frac{\delta^T (\frac{1}{n}X^T X) \delta}{\delta^T \delta} = \inf_{\substack{\delta \in \mathbb{R}^p \\ \delta \neq 0}} \frac{\frac{1}{n} \|X\delta\|_2^2}{\|\delta\|_2^2}.$$

ϕ^2 is sometimes called a “restricted eigenvalue”. Note that

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

Hence $\phi^2 \geq c_{\min}$.

The compatibility condition says $\phi^2 > 0$. In high-dimensions ($p > n$) we have $c_{\min} = 0$, but we can have $\phi^2 > 0$ for some S .

Theorem 30. Suppose $\phi^2 > 0$, let $\lambda^* = A\sigma\sqrt{\frac{\log p}{n}}$ with $A > 2\sqrt{2}$. Then with probability $\geq 1 - 2p^{-(A^2/8-1)}$ we have for all $\lambda \geq \lambda^*$ that

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

Corollary 31. If $\lambda = \lambda^*$,

$$\frac{1}{n} \|X(\beta^0 - \hat{\beta}_{\lambda^*}^L)\|_2^2 \leq \frac{16A^2 \log p}{\phi^2} \frac{\sigma^2 s}{n}$$

and

$$\|\beta^0 - \hat{\beta}_{\lambda^*}^L\|_1 \leq \frac{16A\sigma s}{\phi^2} \sqrt{\frac{\log p}{n}}.$$

Proof of theorem. Let $\hat{\beta} = \hat{\beta}_\lambda^L$. We have from a previous result that $\Omega = \{\frac{2\|X^T \varepsilon\|_\infty}{n} \leq \lambda\}$ has $\mathbb{P}(\Omega) \geq 1 - 2p^{-(A^2/8-1)}$ and on Ω

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

Let $a = \|X(\hat{\beta} - \beta^0)\|_2^2/(n\lambda)$. Dividing (II) through by λ gives

$$a + 2(\|\hat{\beta}_S\|_1 + \|\hat{\beta}_N\|_1) \leq \|\hat{\beta}_S - \beta_S^0\|_1 + \|\hat{\beta}_N\|_1 + 2\|\beta_S^0\|_1.$$

Hence

$$\begin{aligned} a + \|\hat{\beta}_N\|_1 &\leq \|\hat{\beta}_S - \beta_S^0\|_1 + 2\|\beta_S^0\|_1 - 2\|\hat{\beta}_S\|_1 \\ &\leq \|\hat{\beta}_S - \beta_S^0\|_1 + 2\|\hat{\beta}_S - \beta_S^0\|_1 \\ &= 3\|\hat{\beta}_S - \beta_S^0\|_1 \end{aligned}$$

so adding $\|\hat{\beta}_S - \beta_S^0\|_1$ to both sides gives

$$a + \|\hat{\beta} - \beta^0\|_1 \leq 4\|\hat{\beta}_S - \beta_S^0\|_1. \quad ((III))$$

Let $\delta = \hat{\beta} - \beta^0$, so we have $\|\delta_N\|_1 \leq 3\|\delta_S\|_1$ and hence

$$\phi^2 \leq \frac{\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2}{\frac{1}{s}\|\hat{\beta}_S - \beta_S^0\|_1} \implies \|\hat{\beta}_S - \beta_S^0\|_1 \leq \sqrt{\frac{s}{n}} \frac{1}{\phi} \|X(\beta^0 - \hat{\beta})\|_2.$$

Plugging this into (III):

$$\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 + \lambda\|\beta^0 - \hat{\beta}\|_1 \leq \frac{4\lambda}{\phi} \sqrt{\frac{s}{n}} \|X(\hat{\beta} - \beta^0)\|_2. \quad ((IV))$$

This implies

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

so plugging into (IV) gives the desired inequality. \square

The compatibility condition

- $\phi^2 > 0$ is a weaker assumption than the smallest eigenvalue of $\frac{1}{n}X^T X$, c_{\min} , is positive;
- If $p > n$, $c_{\min} = 0$ but we can have $\phi^2 > 0$;
- We want to prove that if the rows x_i of X are centered iid random variables, under certain conditions on covariance $\Sigma^0 = \mathbb{E}[x_1 x_1^T] = \mathbb{E}[\frac{1}{n}X X^T]$, and certain assumptions on the tails of x_1 , we can have $\phi^2 > 0$ for *all* subsets S of a given size, with high probability.

Define

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

Lemma 32. Suppose $\phi_{\Theta}^2(S) > 0$ and Σ is such that $\max_{j,k} |\Theta_{jk} - \Sigma_{jk}| \leq \frac{\phi_{\Theta}^2(S)}{32|S|}$. Then $\phi_{\Sigma}^2(S) \geq \frac{\phi_{\Theta}^2(S)}{2}$.

Proof. For simplicity, we will neglect dependence on S . Write $s = |S|$ and define $B = \{\delta \in \mathbb{R}^p : \|\delta_S\|_1 = 1, \|\delta_N\|_1 \leq 3\}$. Note

$$\phi_{\Theta}^2(S) = |S| \inf_{\delta \in B} \delta^T \Theta \delta \quad \forall \Theta.$$

Now for $\delta \in B$

$$\begin{aligned} s\delta^T \Sigma \delta &= s\delta^T \Theta \delta - s\delta^T (\Theta - \Sigma) \delta \\ &\geq \phi_{\Theta}^2 - s|\delta^T (\Theta - \Sigma) \delta|. \end{aligned}$$

Note

$$\begin{aligned} |\delta^T (\Theta - \Sigma) \delta| &\leq \|\delta\|_1 \|(\Theta - \Sigma) \delta\|_{\infty} && \text{(Hölder)} \\ &\leq \|\delta\|_1^2 \max_{j,k} |\Theta_{jk} - \Sigma_{jk}| && \text{(Hölder)} \\ &\leq \|\delta\|_1^2 \frac{\phi_{\Theta}^2}{32s}. \end{aligned}$$

As $\delta \in B$ we have $\|\delta\|_1 = \|\delta_S\|_1 + \|\delta_N\|_1 \leq 4$. Hence,

$$s\delta^T \Sigma \delta \geq \phi_{\Theta}^2 - s \frac{4^2 \phi_{\Theta}^2}{32s} = \frac{\phi_{\Theta}^2}{2}$$

for any $\delta \in B$, so taking the infimum over $\delta \in B$ gives the result. \square

Plan:

- If the rows of X are iid with covariance Σ^0 then $\hat{\Sigma} = \frac{1}{n}X^T X$ is an estimate of Σ^0 ;

- Use concentration inequalities to find high probability bounds for $\max_{j,k} \left| \Sigma_{jk}^0 - \hat{\Sigma}_{jk} \right|$;
- Apply the previous lemma to argue $\phi_{\hat{\Sigma}}^2 \geq \frac{1}{2} \phi_{\Sigma^0}^2$.

Question: when can we assume $\phi_{\Sigma^0}^2 > 0$?

Example. Take $\Sigma^0 = I$ (predictors uncorrelated). Then $\phi_{\Sigma^0}^2(S) > c_{\min}(\Sigma^0) = 1$ for any subset $S \subseteq \{1, \dots, p\}$.

Concentration Inequalities Continued

Goal: obtain tail bounds on products of sub-Gaussian random variables.

Definition. We say a random variable W satisfies the *Bernstein condition* with parameters (σ, b) , where $\sigma, b > 0$ if

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

Proposition 33 (Bernstein's inequality). *Let W_1, \dots, W_n be independent random variables with mean μ . Suppose that W_i is Bernstein(σ, b) for all $i \in [n]$. Then*

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

and

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

Proof. Fix i , let $W = W_i$. Then for $|\alpha| < 1/b$

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

For the tail bound we have

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

so by Markov's inequality

$$\begin{aligned}\mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n W_i - \mu \geq t \right) &= \mathbb{P} \left(\exp \left(\frac{\alpha}{n} \sum_{i=1}^n W_i - \mu \right) \geq e^{\alpha t} \right) \\ &\leq \exp \left(n \frac{(\alpha/n)^2 \sigma^2 / 2}{1 - b|\alpha/n|} - \alpha t \right) \quad \forall |\alpha/n| < 1/b.\end{aligned}$$

Taking $\alpha/n = \frac{t}{bt + \sigma^2} \in (0, 1/b)$ gives the result. \square

Lemma 34. *Let W, Z be sub-Gaussian with parameters σ_W, σ_Z respectively. Then WZ is Bernstein($8\sigma_W\sigma_Z, 4\sigma_W\sigma_Z$) [W, Z need not be independent].*

Proof. We have

$$\begin{aligned}\mathbb{E}(W^{2k}) &= \mathbb{E} \left[\int_0^\infty \mathbb{1}\{y < w^{2k}\} dy \right] = \int_0^\infty \mathbb{P}(y < W^{2k}) dy \\ &= 2k \int_0^\infty t^{2k-1} \mathbb{P}(|W| > t) dt \\ &\leq 4k \int_0^\infty t^{2k-1} \exp\left(-\frac{t^2}{2\sigma_W^2}\right) dt \\ &= 2^{k+1} \sigma_W^{2k} k!\end{aligned}$$

by Fubini's Theorem. Also, for any random variable Y we have $\mathbb{E}|Y - \mathbb{E}Y|^k \leq 2^k \mathbb{E}[|Y|^k]$ by the binomial theorem and Jensen. Hence

$$\begin{aligned}\mathbb{E}[|WZ - \mathbb{E}[WZ]|^k] &\leq 2^k \mathbb{E}[|WZ|^k] \leq 2^k \sqrt{\mathbb{E}[W^{2k}] \mathbb{E}[Z^{2k}]} \\ &\leq k! 2^{2k+1} (\sigma_W \sigma_Z)^k \\ &= \frac{k!}{2} (8\sigma_W \sigma_Z)^2 (4\sigma_W \sigma_Z)^{k-2}.\end{aligned}$$

□

Random Design

Theorem 35. *Suppose the rows of X are iid, and each entry of X is v -sub-Gaussian. Let $\hat{\Sigma} = \frac{1}{n} X^T X$ and $\Sigma^0 = \mathbb{E}\hat{\Sigma}$. Define “worst case” compatibility factors:*

$$\phi_{\hat{\Sigma},s}^2 = \min_{S:|S|=s} \phi_{\hat{\Sigma}}^2(S), \quad \phi_{\Sigma^0,s}^2 = \min_{S:|S|=s} \phi_{\Sigma^0}^2(S).$$

Suppose $\phi_{\Sigma^0,s}^2 > c > 0$. Then

$$\mathbb{P}(\sigma_{\hat{\Sigma},s}^2 \geq \frac{\phi_{\Sigma^0,s}^2}{2}) \geq 1 - 2p \left(2^{-M \frac{n}{s^2 \log p}} \right)$$

for some constant $M > 0$ independent of n, s and p .

Remarks.

- This theorem says the compatibility condition holds uniformly over all S of size s with high probability if $\frac{n}{s^2 \log p}$ is large, i.e $s \ll \sqrt{n/\log p}$.
- The condition $\phi_{\Sigma^0,s}^2 > 0$ holds for c being the smallest eigenvalue of Σ^0 .
- The theorem does not require X_{ij}, X_{ik} for $k \neq j$ to be independent or even uncorrelated.

Proof. By a previous lemma,

$$\begin{aligned} \mathbb{P}\left(\phi_{\hat{\Sigma},s}^2 \geq \frac{\phi_{\Sigma^0,s}^2}{2}\right) &\geq \mathbb{P}\left(\max_{j,k} |\hat{\Sigma}_{jk} - \Sigma_{jk}^0| \leq \frac{\phi_{\Sigma^0,s}^2}{32s}\right) \\ &\geq 1 - p^2 \min_{j,k} \mathbb{P}\left(|\hat{\Sigma}_{jk} - \Sigma_{jk}^0| \geq \frac{\phi_{\Sigma^0,s}^2}{32s}\right) \end{aligned}$$

so it suffices to show

$$\min_{j,k} \mathbb{P}\left(|\hat{\Sigma}_{jk} - \Sigma_{jk}^0| \geq \frac{\phi_{\Sigma^0,s}^2}{32s}\right) \leq 2p^{-M \frac{n}{s^2 \log p}}$$

for some $M > 0$. Indeed note

$$\mathbb{P}\left(|\hat{\Sigma}_{jk} - \Sigma_{jk}^0| \geq \frac{\phi_{\Sigma^0,s}^2}{32s}\right) = \mathbb{P}\left(\left|\sum_{i=1}^n \frac{X_{ij}X_{ik}}{n} - \Sigma_{jn}^0\right| \geq \frac{\phi_{\Sigma^0,s}^2}{32s}\right).$$

Since $X_{ij}X_{ik}$ is a product of independent v -sub-Gaussian random variables, it is Bernstein($8v^2, 4v^2$). By Bernstein's inequality we then have

$$\mathbb{P}\left(\left|\sum_{i=1}^n \frac{X_{ij}X_{ik}}{n} - \Sigma_{jn}^0\right| \geq \frac{\phi_{\Sigma^0,s}^2}{32s}\right) \leq 2 \exp\left(-\frac{n \left(\frac{\phi_{\Sigma^0,s}^2}{32s}\right)^2}{2 \left(64v^4 + 4v^2 \left(\frac{\phi_{\Sigma^0,s}^2}{32s}\right)\right)}\right)$$

and

$$-\frac{n \left(\frac{\phi_{\Sigma^0,s}^2}{32s}\right)^2}{2 \left(64v^4 + 4v^2 \left(\frac{\phi_{\Sigma^0,s}^2}{32s}\right)\right)} = -\frac{n}{s^2} C' \frac{(\phi_{\Sigma^0,s}^2)^2}{C'' + \frac{\phi_{\Sigma^0,s}^2}{s}}$$

and

$$\begin{aligned} C' \frac{(\phi_{\Sigma^0,s}^2)^2}{C'' + \frac{\phi_{\Sigma^0,s}^2}{s}} &\geq \frac{(\phi_{\Sigma^0,s}^2)^2}{C'' + \phi_{\Sigma^0,s}^2} \\ &\geq \frac{C^2}{C'' + C} \end{aligned}$$

where the last inequality follows since the penultimate term is increasing in $\phi_{\Sigma^0,s}^2$ and $\phi_{\Sigma^0,s}^2 > 0$. Hence the desired inequality holds with $M = C' \frac{C^2}{C'' + C}$. \square

Computing $\hat{\beta}_\lambda^L$

The lasso objective is of the form

$$f(x) = g(x) + \sum_{j=1}^d h_j(x_j), \quad x \in \mathbb{R}^d$$

where g is convex and differentiable, h_j is convex.

Coordinate descent

Initialise at $x^{(0)} \in \mathbb{R}^d$. For each $m = 1, \dots, M$ set

$$\begin{aligned} x_1^{(m)} &= \operatorname{argmin}_{x_1 \in \mathbb{R}} f(x_1, x_2^{(m-1)}, x_3^{(m-1)}, \dots, x_d^{(m-1)}) \\ x_2^{(m)} &= \operatorname{argmin}_{x_2 \in \mathbb{R}} f(x_1^{(m)}, x_2, x_3^{(m-1)}, \dots, x_d^{(m-1)}) \\ &\vdots \\ x_d^{(m)} &= \operatorname{argmin}_{x_d \in \mathbb{R}} f(x_1^{(m)}, x_2^{(m)}, x_3^{(m)}, \dots, x_d). \end{aligned}$$

Lemma 36. Suppose $A_0 = \{x \in \mathbb{R}^d : f(x) \leq f(x^{(0)})\}$ is compact. Then

- (i) $f(x^{(m)}) \rightarrow f(x^*)$ as $m \rightarrow \infty$, where $x^* \in \operatorname{argmin}_{x \in A_0} f(x)$;
- (ii) If x^* is the unique minimiser, then $x^{(m)} \rightarrow x^*$.

Proof. Not given. □

For the lasso, each coordinate descent step is

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

where

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

This objective is strictly convex, hence it has a unique solution, which satisfies the following KKT condition:

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

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

The solution is given by

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

where

$$S_\lambda(u) := \operatorname{sgn}(u)(|u| - \lambda)_+$$

is the “soft-thresholding” function.

For fixed λ , we can find $\hat{\beta}_\lambda^L$ by coordinate descent. Suppose we want to compute $\hat{\beta}_\lambda^L$ for λ in $\lambda_0 > \dots > \lambda_L$.

Warm starts:

- Find $\hat{\beta}_{\lambda_0}^L$
- For $\ell = 1, \dots, L$, find $\hat{\beta}_{\lambda_\ell}^L$ by CD initialised at $\hat{\beta}_{\lambda_{\ell-1}}$.

Active set strategy:

1. Initialise $A_\ell = \{k : \hat{\beta}_{\lambda_{\ell-1},k}^L \neq 0\}$
2. Perform coordinate descent only on coordinates in A_ℓ to obtain $\hat{\beta}$ (with $\hat{\beta}_j = 0$ for all $j \notin A_\ell$)
3. Let $V = \{k : |X_k^T(Y - X\hat{\beta})| > \lambda_\ell\}$ be the coordinates where the KKT conditions are violated.
4. If V is empty, set $\hat{\beta}_{\lambda_\ell} = \hat{\beta}$. Otherwise, update $A_\ell \rightarrow A_\ell \cup V$ and go back to step 2.

Extensions of the Lasso

The square-root Lasso

Prediction and estimation error bounds for the Lasso required setting $\lambda = A\sigma\sqrt{\log p/n}$. Although to do this, we need to know σ ! Instead we can try to estimate σ . Define

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

If $\hat{\beta}_\lambda^L$ is accurate, then $\hat{\sigma}_\lambda^L$ should be accurate.

Idea: find optimal λ by setting $\lambda = \lambda_0$, then iterating

- Find $\hat{\beta}$ by solving $\text{Lasso}(\lambda)$
- Set $\hat{\sigma} = \frac{1}{\sqrt{n}} \|Y - X\hat{\beta}\|$
- Set $\lambda = A\hat{\sigma}\sqrt{\log p/n}$.

This is equivalent to alternating minimising with respect to β and σ , the objective

$$Q_\gamma^{\text{sq}}(\beta, \sigma) = \frac{1}{2n\sigma} \|Y - X\beta\|^2 + \frac{\sigma}{2} + \gamma \|\beta\|_1$$

where $\gamma = A\sqrt{\log n/p}$, $\lambda = \gamma\sigma$.

A more direct route to minimising $Q_\gamma^{\text{sq}}(\beta, \sigma)$ is to minimise

$$\min_{\sigma > 0} Q_\gamma^{\text{sq}}(\beta, \sigma) = \frac{1}{\sqrt{n}} \|Y - X\beta\| + \gamma \|\beta\|_1 := Q_\gamma^{\text{sq}}(\beta)$$

with respect to β , provided $Y \neq X\beta$. A minimiser of $Q_\gamma^{\text{sq}}(\beta)$ is called the *square-root Lasso estimator*, written $\hat{\beta}_\gamma^{\text{sq}}$. Contrast the loss with

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

Define $\hat{\sigma}_\gamma^{\text{sq}} = \frac{1}{\sqrt{n}} \|Y - X\hat{\beta}_\gamma^{\text{sq}}\|$. From this definition, we can see that if $\lambda = \hat{\sigma}_\gamma^{\text{sq}}\gamma$, then

$$\hat{\beta} \in \operatorname{argmin}_\beta Q_\gamma^{\text{sq}}(\beta) \iff \hat{\beta} \in \operatorname{argmin}_\beta Q_\lambda^L(\beta).$$

Conclusion: square-root Lasso is simply a reparameterisation of the regular path of the Lasso.

Theorem 37. Consider a model $Y = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Let $\hat{\beta}$ be a square-root Lasso estimator with $\gamma = B\sqrt{\log p/n}$, $B > 2\sqrt{2}$. Consider an asymptotic regime: $n, p \rightarrow \infty$ with $\frac{s \log p}{n} \rightarrow 0$ and compatibility factor $\phi^2 > c > 0$. Then with probability tending to 1,

$$\begin{aligned} \frac{1}{n} \|X(\beta^0 - \hat{\beta})\|^2 &\leq \frac{17B^2 \log p}{\phi^2} \frac{s\sigma^2}{n} \text{ and} \\ \|\beta^0 - \hat{\beta}\|_1 &\leq \frac{17Bs}{\phi^2} \sqrt{\log p/n}. \end{aligned}$$

Proof. Let $\hat{\sigma} := \hat{\sigma}_\gamma^{\text{sq}} = \frac{1}{\sqrt{n}} \|Y - X\hat{\beta}\|$. So $\hat{\beta}$ is a Lasso estimator with parameter $\lambda = \hat{\sigma}\gamma$. Take $\lambda_j = \sigma A_j \sqrt{\log p/n}$ for $j = 1, 2, \dots$, where $2\sqrt{2} < A_1 < B < A_2$ with $16A^2 \leq 17B^2$ (this implies $16A_2 < 17B$).

Strategy:

- Show that $\lambda_1 \leq \hat{\sigma}\gamma \leq \lambda_2$ with high probability.
- By a previous theorem, $\hat{\beta}_{\lambda_1}^L, \hat{\beta}_{\lambda_2}^L$ are “accurate”.
- Deduce from this that $\hat{\beta} = \hat{\beta}_{\hat{\sigma}\gamma}^L$ is “accurate”.

In the Example Sheet we’ll show there exists a sequence $a_n \rightarrow 0$ such that on a sequence of events $\Omega_n^{(1)}$ with $\mathbb{P}(\Omega_n^{(1)}) \rightarrow 1$, we have

$$1 - a_n \leq \frac{\sigma}{\hat{\sigma}_{\lambda_j}^L} \leq 1 + a_n \text{ for } j = 1, 2.$$

Hence on $\Omega_n^{(1)}$, for n large enough,

$$\begin{aligned}\gamma_1 &:= \frac{\lambda_1}{\hat{\sigma}_{\lambda_1}^L} = \frac{\sigma}{\hat{\sigma}_{\lambda_1}^L} A_1 \sqrt{\frac{\log p}{n}} \leq (1 + a_n) A_1 \sqrt{\frac{\log p}{n}} \\ &\leq B \sqrt{\frac{\log p}{n}} \\ &= \gamma.\end{aligned}$$

And similarly,

$$\gamma \leq \gamma_2 := \frac{\lambda_2}{\hat{\sigma}_{\lambda_2}^L}. \quad (1)$$

Note $\hat{\sigma}_{\lambda_j}^L = \hat{\sigma}_{\gamma_j}^{\text{sq}}$ for $j = 1, 2$. We'll show $\gamma_1 \leq \gamma$ implies $\hat{\sigma}_{\gamma_1}^{\text{sq}} \leq \hat{\sigma}_{\gamma}^{\text{sq}}$. By optimality,

$$\begin{aligned}\frac{1}{\sqrt{n}} \|Y - X \hat{\beta}_{\gamma_1}^{\text{sq}}\| + \gamma_1 \|\hat{\beta}_{\gamma_1}^{\text{sq}}\| &\leq \frac{1}{\sqrt{n}} \|Y - X \hat{\beta}_{\gamma}^{\text{sq}}\| + \gamma_1 \|\hat{\beta}_{\gamma}^{\text{sq}}\|_1 \\ \frac{1}{\sqrt{n}} \|Y - X \hat{\beta}_{\gamma}^{\text{sq}}\| + \gamma_1 \|\hat{\beta}_{\gamma}^{\text{sq}}\| &\leq \frac{1}{\sqrt{n}} \|Y - X \hat{\beta}_{\gamma_1}^{\text{sq}}\| + \gamma_1 \|\hat{\beta}_{\gamma_1}^{\text{sq}}\|_1.\end{aligned} \quad (I)$$

Adding these and rearranging,

$$\begin{aligned}(\underbrace{\gamma - \gamma_1}_{\geq 0}) (\|\hat{\beta}_{\gamma_1}^{\text{sq}}\|_1 - \|\hat{\beta}_{\gamma}^{\text{sq}}\|_1) &\geq 0 \\ \implies \|\hat{\beta}_{\gamma_1}^{\text{sq}}\| &\geq \|\hat{\beta}_{\gamma}^{\text{sq}}\|_1.\end{aligned}$$

Plugging this into (I) gives

$$\hat{\sigma}_{\gamma_1}^{\text{sq}} = \frac{1}{\sqrt{n}} \|Y - X \hat{\beta}_{\gamma_1}^{\text{sq}}\| \leq \frac{1}{\sqrt{n}} \|Y - X \hat{\beta}_{\gamma}^{\text{sq}}\| = \hat{\sigma}_{\gamma}^{\text{sq}}.$$

And similarly

$$\hat{\sigma}_{\gamma}^{\text{sq}} \leq \hat{\sigma}_{\gamma_2}^{\text{sq}}. \quad (2)$$

Combining (1) and (2) we get

$$\lambda_1 \leq \gamma \hat{\sigma}_{\gamma_1}^{\text{sq}} \leq \gamma \hat{\sigma} \leq \gamma \hat{\sigma}_{\gamma_2}^{\text{sq}} \leq \lambda_2$$

note that the middle term is the λ parameter corresponding to $\hat{\beta}$. Now, apply a previous theorem with $\lambda^* = \lambda_1$, which says that on a sequence of events $\Omega_n^{(2)}$ with $\mathbb{P}(\Omega_n^{(2)}) \rightarrow 1$, for all $\lambda \geq \lambda_1$ we have

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

Then with $\Omega_n = \Omega_n^{(1)} \cap \Omega_n^{(2)}$ we have $\mathbb{P}(\Omega_n) \rightarrow 1$ and on Ω_n we have $\hat{\sigma}_{\gamma} \geq \lambda_1$ hence

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

Finally, we note that as we are working on a subset of $\Omega_n^{(1)}$, we have

$$\hat{\sigma}\gamma \leq \lambda_2 \leq 17\sigma B\sqrt{\log p/n} \implies \hat{\sigma}^2\gamma^2 \leq \lambda_2^2 \leq 17\sigma^2 B^2 \frac{\log p}{n}$$

which imply the desired inequalities. \square

Other loss functions

We can apply an ℓ_1 penalty to other log-likelihoods

$$Q_\lambda(\beta) = \sum_{i=1}^n \ell(Y_i, x_i^T \beta) + \lambda \|\beta\|_1.$$

The usual Lasso corresponds to $\ell(Y_i, x_i^T \beta) = (Y_i - x_i^T \beta)^2$. Logistic regression corresponds to $\ell(Y_i, x_i^T \beta) = \log(1 + e^{-Y_i x_i^T \beta})$.

The ℓ_1 penalty encourages sparsity in $\hat{\beta} = \operatorname{argmin}_\beta Q_\lambda(\beta)$. β can be computed via coordinate descent, and has similar statistical guarantees as for Lasso.

Group Lasso

Let G_1, \dots, G_q be a partition of $\{1, \dots, p\}$. The group Lasso penalty is

$$\sum_{j=1}^q m_j \|\beta_{G_j}\|_2, \quad m_j > 0.$$

- When $|G_j| = 1$ for all j , $m_j = \lambda$, this is the normal Lasso.
- Encourages whole groups of coefficients to be shrunk to 0.
- Typically $m_j = \sqrt{|G_j|}$.

Example.

- We code a categorical predictor with M categories using $M - 1$ indicator variables X_G .
- Want to model non-linear relationship between Y_i, x_{ij} using basis functions $h_1(x_{ij}), \dots, h_M(x_{ij})$.
-

Fused Lasso

If coefficients $\beta_1^0, \dots, \beta_p^0$ have some natural order, and we believe the sequence is piecewise constant, we can use penalty

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

The first term encourages sparsity in $(\hat{\beta}_j - \hat{\beta}_{j+1})_{j=1}^{p-1}$, i.e. $\hat{\beta}$ is piecewise constant.

Example. Suppose $Y_i = \mu_i^0 + \varepsilon_i$ where (μ_i^0) is piecewise constant. Use

$$\hat{\mu} = \operatorname{argmin}_{\mu \in \mathbb{R}^n} \left[\frac{1}{2n} \|Y - \mu\|_2^2 + \lambda_1 + \sum_{j=1}^{p-1} |\mu_j - \mu_{j+1}| \right].$$

This is known as “total variation denoising”.

Bias reduction

The Lasso tends to shrink both non-significant and significant coefficients, leading to bias.

Adaptive Lasso

Let $\hat{\beta}^{\text{init}}$ be a Lasso estimator, $\hat{S}_{\text{init}} = \{k : \hat{\beta}_k^{\text{init}} \neq 0\}$. Use the estimator

$$\hat{\beta}^{\text{adapt}} = \underset{\substack{\beta \in \mathbb{R}^p \\ \hat{\beta}_{\hat{S}_{\text{init}}^c} = 0}}{\text{argmin}} \left\{ \frac{1}{2n} \|Y - X\beta\|^2 + \lambda \sum_{k \in \hat{S}_{\text{init}}} \frac{|\beta_k|}{|\hat{\beta}_k^{\text{init}}|} \right\}.$$

This reduces the penalty on large coefficients.

Minimax convex penalty

Have penalty

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

where

$$p'_{\lambda, \gamma}(u) = \left(\lambda - \frac{u}{\gamma} \right)_+, \quad p'_{\lambda, u}(0) = 0.$$

Graphical Models

We have iid observations of some random variable in \mathbb{R}^d . We wish to understand the relationship of the coordinates.

Definition. If X, Y, Z are random vectors with joint density $f_{X,Y,Z}$ (with respect to a product measure μ), we say X is *conditionally independent of Y given Z* , or $X \perp\!\!\!\perp Y|Z$ if

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

Equivalently $X \perp\!\!\!\perp Y|Z \iff f_{X|Y,Z}(x|y, z) = m(x, z)$ for some function m (where m is $f_{X|Z}$). We write $X \not\perp\!\!\!\perp Y|Z$ for the negation of $X \perp\!\!\!\perp Y|Z$. We can represent conditional independence relationships between Z_1, \dots, Z_p through an undirected graph $\mathcal{G} = (V, E)$ with vertices $V = \{1, \dots, p\}$ and edges $E \subseteq \{\{i, j\} : i \neq j, i, j \in V\}$.

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

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

“Structure learning” refers to learning the CIG from data.

Directed Acyclic Graphs & Causality

This subsection is **non-examinable**.

Gaussian Graphical Models

Let Z be a random variable in \mathbb{R}^p . Let x_1, \dots, x_n be iid samples with $x_z =^d Z$. We wish to estimate the CIG of (the coordinates of) Z .

This task is easier when Z is normal. Let $Z \sim \mathcal{N}_p(\mu, \Sigma)$ and Σ is positive definite. For any partition of $\{1, \dots, p\}$ into two subsets A, B we can write

$$Z = \begin{pmatrix} Z_A \\ Z_B \end{pmatrix}, \mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}.$$

Note that Σ being positive definite implies both Σ_{AA} and Σ_{BB} are too. Let $\Omega = \Sigma^{-1}$ be the *precision matrix* of Z . We'll show $\Omega_{ij} = 0$ iff $Z_i \perp\!\!\!\perp Z_j | Z_{-ij}$.

Message: to estimate the CIG we just need to identify zeros in Ω .

Proposition 38 (Blockwise matrix inversion). *Let*

$$M = \begin{pmatrix} P & Q^T \\ Q & R \end{pmatrix} > 0$$

be a block matrix. Then

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

where $S = P - Q^TR^{-1}Q$ is the Schur complement of R .

Proof. Plug in and verify. □

Applying this proposition to $M = \Sigma$, $P = \Sigma_{AA}$ yields 2 useful identities:

1. $\Omega_{AA}^{-1} := (\Omega_{AA})^{-1} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$;
2. $\Sigma_{BB}^{-1}\Sigma_{BA} = -\Omega_{BA}\Omega_{AA}^{-1}$.

Proposition 39. *If $Z \sim \mathcal{N}_p(\mu, \Sigma)$ then*

$$Z_A | Z_B = z_B \sim \mathcal{N}_{|A|}(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(z_B - \mu_B), \underbrace{\Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}}_{\Omega_{AA}^{-1}}).$$

Note. The parameter z_B only impacts the mean, not the variance.

Proof. Let $M = \Sigma_{AB}\Sigma_{BB}^{-1}$. We claim $Z_A - MZ_B \perp\!\!\!\perp Z_B$. Since $(Z_A - MZ_B, Z_B)$ is normal, it is enough to show $\text{Cov}(Z_A - MZ_B, Z_B) = 0$.

Note

$$\begin{aligned} \text{Cov}(Z_B, Z_A - MZ_B) &= \Sigma_{BA} - \Sigma_{BB}M^T \\ &= \Sigma_{BA} - (\Sigma_{AB})^T \\ &= 0. \end{aligned}$$

We have $Z_A - MZ_B \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$ where $\bar{\mu} = \mathbb{E}[Z_A - MZ_B] = \mu_A - \Sigma_{AB}\Sigma_{BB}^{-1}\mu_B$ and

$$\begin{aligned}\bar{\Sigma} &= \text{Var}(Z_A - MZ_B) = \Sigma_{AA} + \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BB}\Sigma_{BB}^{-1}\Sigma_{BA} - 2\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA} \\ &= \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}.\end{aligned}$$

Finally, note

$$Z_A = (Z_A - MZ_B) + MZ_B$$

so conditional on $Z_B = z_B$, Z_A is distributed as a $\mathcal{N}(\bar{\mu} + Mz_B, \bar{\Sigma})$. \square

Corollary 40. *We have $Z_i \perp\!\!\!\perp Z_j | Z_{-ij}$ if and only if $\Omega_{ij} = 0$.*

Proof. Take $A = \{i, j\}$. By the previous proposition, Ω_{AA}^{-1} is the conditional covariance of (Z_i, Z_j) given Z_{-ij} . As the conditional distribution is normal,

$$\begin{aligned}Z_i \perp\!\!\!\perp Z_j | Z_{-ij} &\iff \Omega_{AA}^{-1} \text{ is diagonal} \\ &\iff \Omega_{AA} \text{ is diagonal} \\ &\iff \Omega_{ij} = 0.\end{aligned}$$

\square

Graphical Lasso

When Σ is positive definite, $\mathcal{N}_p(\mu, \Sigma)$ has density

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

Given observations $x_1, \dots, x_n \sim^{\text{iid}} \mathcal{N}_p(\mu, \Sigma)$ the log-likelihood of (μ, Ω) 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

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

Then

$$\begin{aligned}\sum_{i=1}^n (x_i - \mu)^T \Omega (x_i - \mu) &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu)^T \Omega (x_i - \bar{x} + \bar{x} - \mu) \\ &= \sum_{i=1}^n (x_i - \bar{x})^T \Omega (x_i - \bar{x}) + \sum_{i=1}^n (\bar{x} - \mu)^T \Omega (\bar{x} - \mu) \\ &= \text{ntr}(S\Omega) + n(\bar{x} - \mu)^T \Omega (\bar{x} - \mu).\end{aligned}$$

Thus

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

and since $(\bar{x} - \mu)^T \Omega (\bar{x} - \mu) \geq 0$ with equality iff $\bar{x} - \mu = 0$, i.e $\mu = \bar{x}$. Hence

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

The MLE of Ω is

$$\hat{\Omega}^{\text{MLE}} = \text{argmin}_{\Omega > 0} \{ \text{tr}(S\Omega) - \log \det(\Omega) \}.$$

The objective is convex, and $\{\Omega > 0\}$ is convex. Also

$$\begin{aligned} \frac{\partial}{\partial \Omega_{jk}} \log \det(\Omega) &= (\Omega^{-1})_{jk} \\ \frac{\partial}{\partial \Omega_{jk}} \text{tr}(S\Omega) &= (S)_{jk}. \end{aligned}$$

Hence if $S > 0$ the stationary point has $S = \Omega^{-1}$, so $\hat{\Omega}^{\text{MLE}} = S^{-1}$.

Note. If $n < p$, S has rank $< p$ and the MLE doesn't exist.

The graphical lasso estimator for Ω is

$$\hat{\Omega}_\lambda^L = \text{argmin}_{\Omega > 0} \{ \text{tr}(S\Omega) - \log \det(\Omega) + \lambda \|\Omega\|_1 \}$$

where $\|\Omega\|_1 = \sum_{jk} |\Omega_{jk}|$. So $\hat{\Omega}_\lambda^L$ will tend to have entries equal to 0 and we can use the zeros to estimate the CIG.

Guassan Graphical Models

$$Z \sim \mathcal{N}_p(\mu, \Sigma), \quad \Sigma > 0.$$

Apply proposition 39 where $A = \{k\}$, we see that the distribution of $Z_k | Z_{-k} = z_{-k}$ can be written as

$$Z_k = m_k + z_{-k}^T \Sigma_{-k, -k}^{-1} \Sigma_{-k, -k} + \varepsilon_k$$

$$m_k = \mu_k - \Sigma_{k, -k} \Sigma_{-k, -k}^{-1} \mu_{-k}$$

where $\varepsilon_k | Z_{-k} = z_{-k} \sim \mathcal{N}(0, \Omega_{kk}^{-1})$. This is a normal linear model with coefficients $\Sigma_{-k, -k}^{-1} \Sigma_{-k, k}$.

Note. $Z_k \perp\!\!\!\perp Z_j | Z_{-jk}$ if and only if the conditional distribution of $Z_k | Z_{-k} = z_{-k}$ does not depend on z_j , which happens if and only if the entry in $\Sigma_{-k, -k}^{-1} \Sigma_{-k, k}$ corresponding to Z_j is 0.

By our blockwise inversion lemma we have

$$\Sigma_{-k, -k}^{-1} \Sigma_{-k, k} = - \underbrace{\Omega_{-k, k}}_{\text{vector}} \underbrace{\Omega_{kk}^{-1}}_{\text{scalar}}.$$

So $(\Sigma_{-k, -k}^{-1} \Sigma_{-k, k})_j = 0$ iff $(\Omega_{-k, k})_j = 0$.

Nodewise regresison

Have data x_1, \dots, x_n iid with distribution Z . Let $X = (x_1, \dots, x_n)^T$. For each $k = 1, \dots, p$

1. Regress X_k onto X_{-k} using e.g square root Lasso;
2. Let \hat{S}_k be the selected variables.

Construct the CIG by letting $\{j, k\} \in E$ if $j \in \hat{S}_k$ OR (AND) $k \in \hat{S}_j$.

3 High-dimensional inference

Debiased Lasso

Low dimensions: $n > p$. We can find confidence sets for β^0 using the fact

$$\hat{\beta}^{\text{OLS}} - \beta^0 \sim \mathcal{N}(0, \sigma^2 (X^T X)^{-1}).$$

Can also test $H_0 : \beta_j^0 = 0$ for example.

In high-dimensions ($n \ll p$; $\|\beta^0\|_0 = s \ll p$) the error $\hat{\beta}_\lambda^L - \beta^0$ is **not** normal. The bias is large and dependent on β^0 .

Idea: suppose $n > p$ and X has full rank. Let R_j be the residual in regressing X_j onto X_{-j} by OLS. Note $R_j \perp \text{col}(X_{-j})$ and $R_j \neq 0$.

If $\tilde{\beta}$ is any estimator of β^0 , we can eliminate the bias by defining a new estimator with j th coefficient

$$\begin{aligned} \tilde{\beta}_j + \frac{R_j^T (Y - X\tilde{\beta})}{R_j^T X_j} &= \tilde{\beta}_j + \frac{R_j^T X (\beta^0 - \tilde{\beta})}{R_j^T X_j} + \frac{R_j^T \varepsilon}{R_j^T X_j} \\ &= \tilde{\beta}_j + \beta_j^0 - \tilde{\beta}_j + \frac{R_j^T \varepsilon}{R_j^T X_j} \\ &= \beta_j^0 + \underbrace{\frac{R_j^T \varepsilon}{R_j^T X_j}}_{\mathcal{N}\left(0, \sigma^2 \frac{\|R_j\|^2}{(R_j^T X_j)^2}\right)}. \end{aligned}$$

We apply this idea in high-dimensions ($n \ll p$). R_j can be the residual of regressing X_j onto X_{-j} with the square root Lasso. Now R_j is not necessarily orthogonal to $\text{col}(X_{-j})$, hence the bias is not completely eliminated. In particular, we can apply this to $\tilde{\beta}$ a square-root Lasso.

Define

$$\begin{aligned} \hat{\beta}^{\text{sq}} &= \underset{\beta}{\text{argmin}} \left\{ \frac{1}{\sqrt{n}} \|Y - X\beta\| + \gamma \|\beta\|_1 \right\}, \\ \hat{\theta}_j^{\text{sq}} &= \underset{\theta}{\text{argmin}} \left\{ \frac{1}{\sqrt{n}} \|X_j - X_{-j}\theta\| + \gamma \|\theta\|_1 \right\}, \\ R_j &= X_j - X_{-j} \hat{\theta}_j^{\text{sq}} \text{ for } j = 1, \dots, p. \end{aligned}$$

The *debiased Lasso* \hat{b} is given by

$$\hat{b}_j = \hat{\beta}_j^{\text{sq}} + \frac{R_j^T (Y - X\hat{\beta}^{\text{sq}})}{R_j^T X_j}.$$

Theorem 41. Let $\gamma = A\sqrt{\frac{\log p}{n}}$. Suppose $R_j \neq 0$ for all $j = 1, \dots, p$. Then, under the model $Y = X\beta^0 + \varepsilon$, $\varepsilon \in \mathcal{N}_n(0, \sigma^2 I)$ we have

$$\frac{R_j^T X_j}{\sigma \|R_j\|} (\hat{b}_j - \beta_j^0) = \delta_j + \xi_j \text{ for } j = 1, \dots, p \quad (*)$$

where $\xi = (\xi_1, \dots, \xi_p)^T$ is normal with $\xi_j \sim \mathcal{N}(0, 1)$ for all j and

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

Note. By a previous theorem we have $\|\hat{\beta}^{\text{sq}} - \beta^0\|_1 \leq s\sqrt{\frac{\log p}{n}}\sigma$, so the RHS tends to 0 if $s \log p / \sqrt{n} \rightarrow 0$.

Proof. We have

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

also

$$\frac{R_j^T X_j (\beta_j^0 - \hat{\beta}_j^{\text{sq}})}{R_j^T X_j} = \beta_j^0 - \hat{\beta}_j^{\text{sq}}$$

so putting this together

$$(*) = \underbrace{\frac{R_j^T X_{-j}}{\sigma \|R_j\|} (\beta_{-j}^0 - \hat{\beta}_{-j}^{\text{sq}})}_{:=\delta_j} + \underbrace{\frac{R_j^T \varepsilon}{\sigma \|R_j\|}}_{:=\xi_j}.$$

We see that $\xi = (\xi_1, \dots, \xi_p)$ is multivariate normal and ξ_j has mean 0, variance 1. Also

$$\begin{aligned} \sigma |\delta_j| &= \left| (\beta_{-j}^0 - \hat{\beta}_{-j}^{\text{sq}})^T \frac{X_{-j}^T R_j}{\|R_j\|} \right| \\ &\leq \underbrace{\|\beta_{-j}^0 - \hat{\beta}_{-j}^{\text{sq}}\|_1}_{\leq \|\beta^0 - \hat{\beta}^{\text{sq}}\|_1} \frac{\|X_{-j}^T R_j\|_\infty}{\|R_j\|}. \end{aligned} \quad (\text{H\"older})$$

We have

$$\frac{\|X_{-j}^T R_j\|_\infty}{\|R_j\|} = \frac{1}{\sqrt{n}} \frac{\|X_{-j}^T (X_j - X_{-j} \hat{\theta}_j^{\text{sq}})\|_\infty}{\frac{1}{\sqrt{n}} \|X_j - X_{-j} \hat{\theta}_j^{\text{sq}}\|_1}.$$

The KKT conditions for the square root Lasso implies this is at most $n\gamma = A\sqrt{n \log p}$ and so

$$|\delta_j| \leq \frac{A\sqrt{\log p}}{\sigma} \|\hat{\beta}^{\text{sq}} - \beta^0\|_1.$$

□

Question: how do we use this result for inference? Under the assumption X has $\phi^2 > c > 0$ from before, we have

$$\|\beta^0 - \hat{\beta}^{\text{sq}}\|_1 \leq \text{constant} \times \sigma s \sqrt{\frac{\log p}{n}}.$$

Hence

$$|\delta_j| \leq \text{constant} \times \frac{s \log p}{\sqrt{n}}.$$

Consider an asymptotic regime where $\frac{s \log p}{\sqrt{n}} \rightarrow 0$. Then $|\delta_j| \rightarrow 0$. So we can define an approximate $(1 - \alpha)$ -confidence interval for β_j^0 by

$$C = \left[\hat{b}_j - z_{\alpha/2} \frac{\hat{\sigma} \|R_j\|}{R_j^T X_j}, \hat{b}_j + z_{\alpha/2} \frac{\hat{\sigma} \|R_j\|}{R_j^T X_j} \right]$$

where

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

Question: is it okay to plug in $\hat{\sigma}$ for σ in C ?

Lemma 42. Suppose $T = f + \varepsilon$ where $\varepsilon \in \mathbb{R}^n$ and $\varepsilon_1, \dots, \varepsilon_n$ are iid, $\mathbb{E}\varepsilon_1 = 0$, $\text{Var}(\varepsilon_1) = \sigma^2 > 0$. Consider an asymptotic regime with increasing n . Let \hat{f} be an estimator of f such that there exists $a_n \rightarrow 0$ with

$$\mathbb{P} \left(\frac{\|\hat{f} - f\|^2}{n} \leq a_n \right) \rightarrow 1.$$

Then if $\hat{\sigma} = \frac{1}{\sqrt{n}} \|Y - \hat{f}\|$ then there is a sequence $b_n \rightarrow 0$ with

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

Remark. It is easy to show $W_n \xrightarrow{\mathbb{P}} W$ if and only if there exists a sequence $(a_n)_{n \geq 1}$ with $a_n \rightarrow 0$ such that $\mathbb{P}(|W_n - W| < a_n) \rightarrow 1$.

As a corollary of this and the previous theorem, we have $\mathbb{P}(\beta_j^c \in C) \xrightarrow{n \rightarrow \infty} 1 - \alpha$.

Question: how wide is the confidence interval for β_j^0 ? It depends on σ as well as the factor

$$\begin{aligned} \frac{X_j^T R_j}{\|R_j\|} &= \frac{X_j^T (X_j - X_{-j} \hat{\theta}_j^{\text{sq}})}{\|X_j - X_{-j} \hat{\theta}_j^{\text{sq}}\|} \\ &= \|X_j - X_{-j} \hat{\theta}_j^{\text{sq}}\| + \underbrace{\frac{\hat{\theta}_j^{\text{sq}} X_{-j}^T (X_j - X_{-j} \hat{\theta}_j^{\text{sq}})}{\|X_j - X_{-j} \hat{\theta}_j^{\text{sq}}\|}}_{(*)}. \end{aligned}$$

The KKT condition for $\hat{\theta}_j^{\text{sq}}$ is

$$\frac{X_{-j}^T (X_j - X_{-j} \hat{\theta}_j^{\text{sq}})}{\|X_j - X_{-j} \hat{\theta}_j^{\text{sq}}\|} = \gamma \sqrt{n} \hat{\nu}$$

where $\|\hat{\nu}\|_\infty \leq 1$, and $(\hat{\theta}_j)_k \neq 0 \Rightarrow \hat{\nu}_k = \text{sgn}(\hat{\theta}_j^{\text{sq}})_k$. Hence (*) is

$$\gamma \sqrt{n} \left(\hat{\theta}_j^{\text{sq}} \right)^T \hat{\nu} = \gamma \sqrt{n} \|\hat{\theta}_j^{\text{sq}}\|_1.$$

Therefore

$$\frac{1}{\sqrt{n}} \frac{X_j^T R_j}{\|R_j\|} = \underbrace{\frac{\|X_j - X_{-j} \hat{\theta}_j^{\text{sq}}\|}{\sqrt{n}}}_{(I)} + \underbrace{\gamma \|\hat{\theta}_j^{\text{sq}}\|_1}_{(II)}.$$

To understand this quantity, we can model the design matrix such that conditional on X_{-j} we have

$$X_j = X_{-j} \theta_j^0 + \xi$$

with θ_j^0 sparse, $\xi \sim \mathcal{N}_n(0, \nu^2 I)$. If X_{-j} has $\phi^2 > c > 0$ then by the previous lemma,

$$(I) \xrightarrow{\mathbb{P}} \nu$$

and by the triangle inequality,

$$(II) \leq \underbrace{\gamma \|\theta_j^0\|_1}_{\rightarrow 0 \text{ for } \theta_j^0 \text{ sparse enough}} + \underbrace{\gamma \|\theta_j^0 - \hat{\theta}_j^{\text{sq}}\|_1}_{\xrightarrow{\mathbb{P}} 0 \text{ by a previous thm}}.$$

Hence we expect $\frac{1}{\sqrt{n}} \frac{X_j^T R_j}{\|R_j\|} \rightarrow \nu$ with high probability.

So we conclude $\sqrt{n}(\hat{b}_j - \beta_j^0)$ to be approximately $\mathcal{N}(0, \sigma^2/\nu^2)$.

Proof of Lemma 42. The hypothesis says $\|f - \hat{f}\|/n \xrightarrow{\mathbb{P}} 0$ and we want to prove $\sigma/\hat{\sigma} \xrightarrow{\mathbb{P}} 1$. By the continuous mapping theorem it is enough to show $\hat{\sigma}^2 \xrightarrow{\mathbb{P}} \sigma^2$. We have

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n} \|f + \varepsilon - \hat{f}\|^2 \\ &= \underbrace{\frac{1}{n} \|f - \hat{f}\|^2}_{\xrightarrow{\mathbb{P}} 0} + \underbrace{\frac{1}{n} \|\varepsilon\|^2}_{\xrightarrow{P} \sigma^2} + \frac{2}{n} \varepsilon^T (f - \hat{f}). \end{aligned}$$

Furthermore

$$\frac{2}{n} \varepsilon^T (f - \hat{f}) \leq 2 \frac{\|\varepsilon\|}{\sqrt{n}} \frac{\|f - \hat{f}\|}{\sqrt{n}} \xrightarrow{\mathbb{P}} 0. \quad (\text{Slutsky})$$

Hence $\hat{\sigma}^2 \xrightarrow{\mathbb{P}} \sigma^2$.

□

Conditional Independence Testing

Let $(x_i, y_i, z_i) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^p$ be iid random variables for $i = 1, \dots, n$. We wish to test $H_0 : x_1 \perp\!\!\!\perp y_1 | z_1$. Previously we showed that if (x_1, y_1, z_1) is $\mathcal{N}(\mu, \Omega^{-1})$, then

$$x_1 \perp\!\!\!\perp y_1 | z_1 \iff \Omega_{x,y} = 0.$$

The graphical Lasso uses an estimator $\hat{\Omega}_\lambda^L$ for Ω . When n is large the zeros in $\hat{\Omega}_\lambda^L$ recover zeros in Ω . This falls short of a hypothesis test.

Main idea: assume $H_0 : x_1 \perp\!\!\!\perp y_1 | z_1$ holds, (x_1, y_1, z_1) not necessarily Gaussian. Then

$$\begin{aligned} & \mathbb{E}[\{x_1 - \mathbb{E}(x_1 | z_1)\}\{y_1 - \mathbb{E}(y_1 | z_1)\} | z_1] \\ &= \mathbb{E}[x_1 - \mathbb{E}(x_1 | z_1) | z_1] \mathbb{E}[y_1 - \mathbb{E}(y_1 | z_1) | z_1] = 0 \end{aligned}$$

almost-surely. By the tower property

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

We will approximate $\mathbb{E}(x_1 | z_1)$ and $\mathbb{E}(y_1 | z_1)$ with our favourite high-dimensional regression method, then use the product of residuals to define a test statistic.

First we define a slightly null hypothesis

$$\tilde{H}_0 : \begin{cases} x_1 = f(z_1) + \varepsilon_1 \\ y_1 = g(z_1) + \xi_1 \\ \mathbb{E}\xi_1 = \mathbb{E}\varepsilon_1 = 0 \\ \text{Var}(\xi_1) = \text{Var}(\varepsilon_1) < \infty \\ \varepsilon_1, \xi_1, z_1 \text{ independent} \end{cases}.$$

Note. \tilde{H}_0 implies $x_1 \perp\!\!\!\perp y_1 | z_1$ but it adds a ‘homoskedasticity’ assumption to H_0 .

Under \tilde{H}_0 we have

$$\text{Var}(\varepsilon_1 \xi_1) = \mathbb{E}(\varepsilon_1^2 \xi_1^2) = \text{Var}(\varepsilon_1) \text{Var}(\xi_1).$$

Given regressors \hat{f}, \hat{g} built from $(x_i, z_i)_{1 \leq i \leq n}$ and $(y_i, z_i)_{1 \leq i \leq n}$ respectively, define

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

Since $\mathbb{E}[\varepsilon_i \xi_i] = 0$, the CLT suggests $\sqrt{n} \tau_N$ is approximately $\mathcal{N}(0, \text{Var}(\varepsilon_1) \text{Var}(\xi_1))$. To estimate the variance, we use

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

The test statistic

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

should be approximately $\mathcal{N}(0, 1)$. A test which rejects \tilde{H}_0 when

$$|T| > \Phi^{-1}(1 - \alpha/2)$$

should have size approximately α (Φ denotes the CDF of a $\mathcal{N}(0, 1)$).

To make this formal, let

$$\begin{aligned} \text{MSPE}_f &= \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n \{f(z_i) - \hat{f}(z_i)\}^2 \right) \\ \text{MSPE}_g &= \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n \{g(z_i) - \hat{g}(z_i)\}^2 \right). \end{aligned}$$

Theorem 43. Suppose $\text{MSPE}_f, \text{MSPE}_g \rightarrow 0$ and $\text{MSPE}_f \text{MSPE}_g = o(n^{-1})$, then under \tilde{H}_0 , we have $T \xrightarrow{d} \mathcal{N}(0, 1)$.

Corollary 44. Previous hypothesis test has size $\rightarrow \alpha$ as $n \rightarrow \infty$.

Proof of Theorem. We have

$$\begin{aligned} n\tau_N &= \sum_{i=1}^n \{f(z_i) + \varepsilon_i - \hat{f}(z_i)\} \{g(z_i) + \xi_i - \hat{g}(z_i)\} \\ &= \sum_{i=1}^n \{f(z_i) - \hat{f}(z_i)\} \{g(z_i) - \hat{g}(z_i)\} \end{aligned} \tag{A1}$$

$$+ \sum_{i=1}^n \varepsilon_i \{g(z_i) - \hat{g}(z_i)\} \tag{A2}$$

$$+ \sum_{i=1}^n \xi_i \{f(z_i) - \hat{f}(z_i)\} \tag{A3}$$

$$+ \sum_{i=1}^n \varepsilon_i \xi_i. \tag{A4}$$

By the CLT, $\frac{A_4}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \text{Var}(\varepsilon_1) \text{Var}(\xi_1))$. Then

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

By Markov's inequality, for all $\delta > 0$ we have

$$\begin{aligned} \mathbb{P}\left(\frac{|A_1|}{\sqrt{n}} > \delta\right) &\leq \frac{\delta^{-1}}{\sqrt{n}} \mathbb{E}|A_1| \\ &\leq \delta^{-1} \sqrt{n \text{MSPE}_f \text{MSPE}_g} \\ &\xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

Hence $|A_1|/\sqrt{n} \xrightarrow{\mathbb{P}} 0$. Now for A_2 ,

$$\begin{aligned} \mathbb{P}\left(\frac{|A_2|}{\sqrt{n}} > \delta\right) &= \mathbb{P}\left(\frac{A_2^2}{n} > \delta^2\right) \\ &\leq \frac{\delta^{-2}}{n} \mathbb{E}A_2^2 \\ &= \frac{\delta^{-2}}{n} \sum_{i=1}^n \mathbb{E}[\varepsilon_i^2 (g(z_i) - \hat{g}(z_i))^2] \quad (\varepsilon_i \perp\!\!\!\perp z_i, y_i \text{ and } \mathbb{E}\varepsilon_i = 0) \\ &= \delta^{-2} \text{Var}(\varepsilon_i) \text{MSPE}_g \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

Therefore $|A_2|/\sqrt{n} \xrightarrow{\mathbb{P}} 0$. By symmetry $|A_3|/\sqrt{n} \xrightarrow{\mathbb{P}} 0$.

By Slutsky, $\sqrt{n}\tau_N \xrightarrow{d} \mathcal{N}(0, \text{Var}(\varepsilon_1) \text{Var}(\xi_1))$. Now we show $\tau_D \xrightarrow{\mathbb{P}} \sqrt{\text{Var}(\varepsilon_1) \text{Var}(\xi_1)}$. Indeed by Markov's inequality

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

and so $\frac{1}{n} \sum_{i=1}^n \{f(z_i) - \hat{f}(z_i)\}^2 \xrightarrow{\mathbb{P}} 0$. Hence by a previous lemma, $\frac{1}{n} \sum_{i=1}^n \{x_i - \hat{f}(z_i)\}^2 \xrightarrow{\mathbb{P}} \text{Var}(\varepsilon_1)$ and by symmetry $\frac{1}{n} \sum_{i=1}^n \{y_i - \hat{g}(z_i)\}^2 \xrightarrow{\mathbb{P}} \text{Var}(\xi_1)$.

Slutsky's lemma and the CMT imply $\tau_D \xrightarrow{\mathbb{P}} \text{Var}(\varepsilon_1) \text{Var}(\xi_1)$ and $T = \frac{\sqrt{n}\tau_N}{\tau_D} \xrightarrow{d} \mathcal{N}(0, 1)$. \square

Example. Suppose $f, g : \mathbb{R}^p \rightarrow \mathbb{R}$ are in an RKHS \mathcal{H} whose reproducing kernel k is the Bochner kernel

$$k(z, z') = h(z - z') = \sum_{j=1}^{\infty} \mu_j e_j(z) e_j(z')$$

where (e_j) is an orthonormal basis for $\mathcal{L}^2(\mathbb{P})$. By Theorem 13, if $\|f\|_{\mathcal{H}} \leq 1$ and $z_i \sim^{\text{iid}} \mathbb{P}$ then there exists a tuning parameter λ_n such that if \hat{f} are the λ_n -kernel ridge regression fitted values, then $\text{MSPE}_f = o(n^{-1/2})$.

Corollary 45. If $z_i \sim^{\text{iid}} \mathbb{P}$, $\|f\|_{\mathcal{H}}, \|g\|_{\mathcal{H}} \leq 1$ and \hat{f}, \hat{g} are λ_n -kernel ridge regressors then under \tilde{H}_0 , $T \xrightarrow{d} \mathcal{N}(0, 1)$.

Multiple Testing

We wish to test many hypotheses H_1, \dots, H_m simultaneously. Suppose that a subset $I_0 \subseteq \{1, \dots, m\}$ are true and let $m_0 = |I_0|$. We have p -values p_1, \dots, p_m satisfying for all $\alpha \in [0, 1]$

$$\mathbb{P}_{H_i}(p_i \leq \alpha) \leq \alpha.$$

A *multiple testing procedure* takes as input p_1, \dots, p_m (or any means of testing H_1, \dots, H_m individually) and outputs a set of rejections $\mathcal{B} \subseteq \{1, \dots, m\}$.

Let $N = |\mathcal{B} \cap I_0|$ be the number of false rejections, and $R = |\mathcal{B}|$ the number of rejections. The *familywise error rate* (FWER) is $\mathbb{P}(N \geq 1)$.

The *Bonferroni procedure* rejects H_i if and only if $p_i \leq \frac{\alpha}{m}$.

Theorem 46. For any $I_0 \subseteq \{1, \dots, m\}$, Bonferroni has $\mathbb{P}(N \geq 1) \leq \mathbb{E}N \leq \frac{m_0}{m} \alpha \leq \alpha$.

Proof. First inequality follows from Markov's inequality. Then

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

□

Closed testing

For any $I \subseteq \{1, \dots, m\}$ let $H_I = \bigcap_{i \in I} H_i = \{H_i \text{ is true } \forall i \in I\}$ be an *intersection hypothesis*. Suppose we have a test ϕ_I taking values in $\{0, 1\}$ such that

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

for each $I \subseteq \{1, \dots, m\}$. We call ϕ_I a *local test*. A *closed testing procedure* (CTP) rejects H_I if and only if, for all $J \supseteq I$, the local test rejects H_J (i.e. $\phi_J = 1$).

Theorem 47. *The CTP has FWER $\leq \alpha$.*

Proof. Wlog take I_0 non-empty. The CTP makes a false rejection only if $\phi_{I_0} = 1$, and this has probability $\leq \alpha$. \square

A CTP defines a multiple testing procedure if we consider $H_{\{1\}}, H_{\{2\}}, \dots, H_{\{m\}}$.

Holm's procedure

A CTP where the local test is defined by

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

In other words, local test rejects H_I ($\phi_I = 1$) if the Bonferroni procedure applied to $\{H_i : i \in I\}$ rejects at least one hypothesis. By the previous result, $\mathbb{P}_{H_I}(\phi_I = 1) \leq \alpha$. Can see (Example Sheet) Holm is equivalent to

1. Sorting p -values $p_{(1)} \leq \dots p_{(n)}$. Write hypothesis corresponding to $p_{(i)}$ as $H_{(i)}$;
2. For each $i = 1, \dots, m$:
 - If $p_{(i)} \leq \frac{\alpha}{m-i+1}$, reject $H_{(i)}$, continue;
 - Otherwise, accept $H_{(i)}, \dots, H_{(m)}$, stop.

Definition (False Discovery Rate (FDR)). The *False Discovery Rate* (FDR) is defined by

$$\text{FDR} := \mathbb{E}_{H_{I_0}} \left[\frac{N}{\max(1, R)} \right]$$

i.e it is the expected proportion of rejections which are false.

The Benjamini-Hochberg procedure works as follows. Let

$$\hat{k}_{\text{B-H}} = \max\{i \in \{1, \dots, m\} : p_{(i)} \leq \frac{i\alpha}{m}\}$$

and reject $H_{(1)}, \dots, H_{(\hat{k}_{\text{B-H}})}$.

Theorem 48. Suppose that for all $i \in I_0$, p_i is independent of $\{p_j : j \neq i\}$. Then the Benjamini-Hochberg procedure has

$$\text{FDR} \leq \alpha \frac{m_0}{m} \leq \alpha.$$

Problem: what if the p -values are not independent? For example $H_i : \beta_i^0 = 0$ in a sparse regression model.

There is a fix:

Benjamini-Yekutieli

Reject $H_{(1)}, \dots, H_{(\hat{k}_{\text{B-Y}})}$ where

$$\hat{k}_{\text{B-Y}} = \max\{i : p_{(i)} \leq \frac{i\alpha}{m\gamma_m}\}$$

where $\gamma_m = 1 + \frac{1}{2} + \dots + \frac{1}{m} \sim \log m$.

On Example Sheet 4 it is shown that Benjamini-Yekutieli has $\text{FDR} \leq \alpha$ even if the p -values have arbitrary dependence.

Proof of Theorem. For each $i \in I_0$, let $p^{-i} = \{p_1, \dots, p_{i-1}, p_{i+1}, \dots, p_m\}$ and if $p_{(1)}^{-i} \leq \dots \leq p_{(m-1)}^{-i}$ are the order statistics, define

$$R_i = \max\{j : p_{(j)}^{(-i)} \leq \frac{\alpha(j+1)}{m}\}.$$

Fix $r \in \{1, \dots, m\}$, $i \in I_0$. Then

$$\begin{aligned} \{p_i \leq \frac{\alpha r}{m}, R = r\} &= \{p_i \leq \frac{\alpha r}{m}, p_{(r)} \leq \frac{\alpha r}{m}, p_{(s)} > \frac{\alpha s}{m} \forall s > r\} \\ &= \{p_i \leq \frac{\alpha r}{m}, p_{(r-1)}^{-i} \leq \frac{\alpha r}{m}, p_{(s-1)}^{-i} > \frac{\alpha s}{m} \forall s > r\} \\ &= \{p_i \leq \frac{\alpha r}{m}, R_i = r - 1\}. \end{aligned}$$

Thus

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

□

The rest of these notes are **non-examinable**.

Wang & Randas (2021)

Definition. An e -value e_k is a statistic which has

$$\mathbb{E}_{H_k} e_k \leq 1, \quad e_i \geq 0.$$

If we have e -values e_1, \dots, e_m with order statistics $e_{(1)} \geq \dots \geq e_{(m)}$, the e -BH procedure rejects $H_{(1)}, \dots, H_{(\hat{k}_e)}$ where

$$\hat{k}_e = \max \left\{ i : \frac{i e_{(i)}}{m} \geq \frac{1}{\alpha} \right\}.$$

Note. $p_i = 1/e_i$ is a p -value since

$$\mathbb{P} \left(\frac{1}{e_i} \leq \alpha \right) = \mathbb{P} \left(e_i \geq \frac{1}{\alpha} \right) \leq \frac{\mathbb{E} e_i}{1/\alpha} \leq \alpha.$$

Theorem 49. The e -BH procedure controls FDR at level α even if (e_i) have arbitrary dependence.

Defining e -values can be much easier than defining p -values.

Examples.

- Sequential hypotheses;
 - Use super-martingales
- Bayes factors with point null hypotheses are p -values;
- Likelihood ratio statistic is an e -value even for finite samples.

Neural Networks

Multi-layer perceptron (MLP)

Have data (z_i, y_i) with $z_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$. We have a model

$$y_i = f(z_i; \theta) + \varepsilon_i$$

with $\mathbb{E}\varepsilon_i = 0$, $\text{Var}(\varepsilon_i) = \sigma^2$.

Non-linearity: $\sigma : \mathbb{R}^d \rightarrow \mathbb{R}$ with e.g. $\sigma(x)_i = \max(x_i, 0)$ (Relu).

Let $h_i^0 = z_i$. For $j \in [D]$ let $h_i^j = \sigma(W^{j-1}h_i^{j-1} + b^{j-1}) \in \mathbb{R}^{p_j}$ with $p_D = 1$. Let $f(z_i; \theta) = h_i^D$.

Double descent phenomenon

$\theta(W^0, h^0, \dots, W^{D-1}, h^{D-1}) \in \mathbb{R}^p$. The training error falls as $\gamma := p/n$ grows, until $\gamma = 1$ where it flattens out. On the other hand, the test error is like a parabola for $\gamma \in [0, 1]$, overfitting with $\gamma \sim 1$. But as γ grows above 1, the test error actually decreases again. So there's no disadvantage to adding more and more parameters.

Lazy training

Let θ_0 be a random initialisation of θ , such that $f(z_i, \theta_0) \approx 0$ for all i .

We will train the parameters θ in some small neighbourhood of the initialisation θ_0 . In this regime,

$$f(z_i, \theta) \approx \underbrace{\nabla_{\theta} f(z_i, \theta_0)^T}_{:= x_i^T} \underbrace{(\theta - \theta_0)}_{:= \beta}.$$

We'll apply gradient descent to the objective

$$\begin{aligned} \sum_{i=1}^n (y_i - f(z_i; \theta))^2 &\approx \sum_{i=1}^n (y_i - \nabla f(z_i; \theta_0)^T \beta)^2 \\ &= \|Y - X\beta\|^2 \end{aligned}$$

Proposition 50. *Gradient descent on this objective with $\beta^{(0)} = 0$, we converge to the “ridgeless estimator”*

$$\hat{\beta} = \operatorname{argmin}_{b \in \mathbb{R}^p} \{\|b\|_2 : b \text{ is a minimiser of } \|Y - Xb\|^2\}.$$

Remark. If $\hat{\beta}_{\lambda}^R = \operatorname{argmin}_{\beta} \{\|Y - X\beta\|^2 + \lambda\|\beta\|^2\}$ then $\hat{\beta} = \lim_{\lambda \downarrow 0} \hat{\beta}_{\lambda}^R$.

Note. Ridge regression with features $x_i = \nabla_{\theta} f(z_i; \theta_0)$ is equivalent to doing kernel ridge regression on original inputs (z_i) with kernel

$$k(z, z') = \nabla f(z; \theta_0)^T \nabla f(z'; \theta_0)$$

called the neural tangent kernel.

Summary: under the lazy training regime, the Neural Network is simply doing kernel ridge regression with a fixed kernel (the NTK).

This is not the full story however. Many useful Neural Networks do “representation learning”, i.e they learn a kernel.

Question: can we understand double descent in a linear model

$$y_i = x_i^T \beta + \varepsilon_i, \quad \mathbb{E}\varepsilon_i = 0, \text{Var}(\varepsilon_i) = \sigma^2$$

where $(x_i) \sim^{\text{iid}} \mathbb{P}$?

Example.

1. $x_i \sim \mathcal{N}(0, I)$;
2. “Latent variable model” $x_i \sim \mathcal{N}(0, WW^T + I)$ for WW^T a low-rank component. In other words, $x_i = Wz_i + \xi_i$ with $z_i, \xi_i \sim \mathcal{N}(0, I)$ and $W \in \mathbb{R}^{p \times d}$, $d \ll p$. β is in the column space of W , i.e β is aligned with the top eigenvectors of $\text{Cov}(x_i)$.
3. Misspecified model:

$$y_i = x_i^T \beta + \underbrace{w_i^T}_{\text{unobserved}} \eta + \varepsilon_i$$

simulate from this model; then fit ridgeless estimator using only (x_i, y_i) .

4. $x_i = f(z_i; \theta)$ where $f(\cdot; \theta)$ is a two-layer MLP with random parameters, $z_i \sim^{\text{iid}} \mathcal{N}(0, I)$.