MODERN STATISTICAL METHODS Part III
Example Sheet 1 (of 4) RDS/Michaelmas 2025

Questions 3 and 12 will be marked. Some notation: $[m] := \{1, \dots, m\}$; $a \wedge b$ and $a \vee b$ are the the minimum and maximum of a and b respectively; $a_+ := a \vee 0$.

1. Consider minimising the following objective involving response $Y \in \mathbb{R}^n$ and design matrix $X \in \mathbb{R}^{n \times p}$ over $(\mu, \beta) \in \mathbb{R} \times \mathbb{R}^p$:

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

   Here $J : \mathbb{R}^p \to \mathbb{R}$ is an arbitrary penalty function. Suppose $\bar{X}_k = 0$ for $k = 1, \dots, p$. Assuming that a minimiser $(\hat{\mu}, \hat{\beta})$ exists, show that $\hat{\mu} = \bar{Y}$. Now take $J(\beta) = \lambda\|\beta\|_2^2$ so we have the ridge regression objective. Show that

$$\hat{\beta} = (X^\top X + \lambda I)^{-1} X^\top Y.$$

   From here onwards, whenever we refer to ridge regression, we will assume X has had its columns mean-centred.

2. Let $X \in \mathbb{R}^{n \times p}$ $(n > p)$ be a centred data matrix with (thin) SVD $X = UDV^\top$. We saw in lectures that the first principal component was $D_{11}U_1 = XV_1$. V_1 is sometimes known as the first *principal direction*. We may define the kth principal component $u^{(k)}$ and principal direction $v^{(k)}$ for $k > 1$ inductively as follows.

$$v^{(k)} \text{ maximises } \|Xv\|_2 \text{ over } v \in \mathbb{R}^p \text{ with constraints}$$
$$\|v\|_2 = 1 \text{ and } {u^{(j)}}^\top Xv = 0 \text{ for all } j < k;$$
$$u^{(k)} := Xv^{(k)}.$$

   Suppose that $D_{11}, \dots, D_{pp}$ are all distinct and positive. Show that $v^{(k)} = V_k$ and $u^{(k)} = D_{kk}U_k$ (up to an arbitrary sign).

3. Consider performing ridge regression when $Y = \mu\mathbf{1} + X\beta^0 + \varepsilon$, where $X \in \mathbb{R}^{n \times p}$ has full column rank, and $\mathbb{E}\varepsilon = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2 I$. Let the SVD of X be $UDV^\top$ and write $\gamma := U^\top X\beta^0$. Show that

$$\frac{1}{n}\mathbb{E}\|X\beta^0 - X\hat{\beta}_\lambda^R\|_2^2 = \frac{1}{n}\sum_{j=1}^p \left(\frac{\lambda}{\lambda + D_{jj}^2}\right)^2 \gamma_j^2 + \frac{\sigma^2}{n}\sum_{j=1}^p \frac{D_{jj}^4}{(\lambda + D_{jj}^2)^2}.$$

   Now consider varying β^0 but fixing the size of the signal so $\|X\beta^0\|_2^2 = n$ (and keeping X fixed). For what γ is the mean squared prediction error above minimised? For what γ is it maximised?

4. In the following, assume that forming AB where $A \in \mathbb{R}^{a \times b}$, $B \in \mathbb{R}^{b \times c}$ requires $O(abc)$ computational operations, and that if $M \in \mathbb{R}^{d \times d}$ is invertible, then forming M^{-1} requires $O(d^3)$ operations.

   (a) Suppose we wish to apply ridge regression to data $(Y, X) \in \mathbb{R}^n \times \mathbb{R}^{n \times p}$ with $n \gg p$. A complication is that the data is split into m separate datasets of size $n/m \in \mathbb{N}$,

$$Y = \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(m)} \end{pmatrix} \qquad X = \begin{pmatrix} X^{(1)} \\ \vdots \\ X^{(m)} \end{pmatrix},$$

with each dataset located on a different server. Moving large amounts of data between servers is expensive. Explain how one can produce ridge estimates $\hat{\beta}_\lambda$ by communicating only $O(p^2)$ numbers from each server to some central server. What is the total order of the computation time required at each server, and at the central server for your approach?

(b) Now suppose instead that $p \gg n$ and it is instead the variables that are split across m servers, so each server has only a subset of $p/m \in \mathbb{N}$ variables for each observation, and some central server stores Y. Explain how one can obtain the fitted values $X\hat{\beta}_\lambda$ communicating only $O(n^2)$ numbers from each server to the central server. What is the total order of the computation time required at each server, and at the central server for your approach?

5. Consider a high-dimensional regression setting where we have some prior information on the potential relative importance of variables in the matrix of predictors $X \in \mathbb{R}^{n\times p}$, with $p \geq 2$, given by an ordering among the variables, so without loss of generality, X_j is thought of as potentially 'more important' than X_{j+1} for $j \in [p-1]$. (An example of when this may occur is the setting where columns of X are known to have been observed with varying degrees of measurement error.) In such a setting, we may consider performing a sequence of p ridge regressions on those variables indexed by each of the nested sets $[p], [p-1], \ldots, [1]$ in turn (a final variable set could then be chosen via cross-validation, though we do not consider that step here).

(a) Let A be a $n \times n$ non-singular matrix and let $b \in \mathbb{R}^n$. Prove that if $b^\top A^{-1} b \neq 1$, then $A - bb^\top$ is invertible with inverse given by

$$(A - bb^\top)^{-1} = A^{-1} + \frac{A^{-1}bb^\top A^{-1}}{1 - b^\top A^{-1} b}.$$

Explain why $X_j^\top (XX^\top + \lambda I)^{-1} X_j < 1$ for all j whenever $\lambda > 0$.

(b) Assuming the complexity costs of matrix operations given in the previous exercise, show that in the case $p \geq n$, the computational complexity of the algorithm above can be made to be $O(np^2)$.

(c) Now write $X = UDV^\top$ for the SVD of X, with $D \in \mathbb{R}^{p\times p}$, and suppose that we have computed $DU^\top Y \in \mathbb{R}^p$ and $V^\top x \in \mathbb{R}^p$ for some $x \in \mathbb{R}^p$. Given a grid of λ values $\lambda_1 > \lambda_2 > \cdots > \lambda_L$, explain how we may compute all ridge regression predictions $x^\top \hat{\beta}^{\mathrm{R}}_{\lambda_1}, \ldots, x^\top \hat{\beta}^{\mathrm{R}}_{\lambda_L}$ for a given $x \in \mathbb{R}^p$ in $O(Lp)$ operations.

6. Suppose we have a matrix of predictors $X \in \mathbb{R}^{n\times p}$ where $p \gg n$. Explain how to obtain the fitted values of the following ridge regression in $O(n^2 p)$ operations using the kernel trick:

$$\text{Minimise over } \beta \in \mathbb{R}^p, \theta \in \mathbb{R}^{p(p-1)/2}, \gamma \in \mathbb{R}^p,$$
$$\sum_{i=1}^{n} \left(Y_i - \sum_{k=1}^{p} X_{ik}\beta_k - \sum_{k=1}^{p}\sum_{j=1}^{k-1} X_{ik}X_{ij}\theta_{jk} - \sum_{k=1}^{p} X_{ik}^2 \gamma_k \right)^2 + \lambda_1 \|\beta\|_2^2 + \lambda_2 \|\theta\|_2^2 + \lambda_3 \|\gamma\|_2^2.$$

Note we have indexed θ with two numbers for convenience.

7. Let $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 < 1\}$. Show that $k(x, x') = (1 - x^\top x')^{-\alpha}$ defined on $\mathcal{X} \times \mathcal{X}$, where $\alpha > 0$, is a kernel.

8. (a) Let A be a (measurable) subset of $\mathbb{R}$. Suppose that $k_\tau : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a kernel for each $\tau \in A$ and that
$$k(x, y) := \int_A k_\tau(x, y)\, d\tau$$
is finite whenever $x = y$. Show that
$$\int_A |k_\tau(x, y)|\, d\tau < \infty,$$
for all $x, y \in \mathbb{R}^d$, and that k is a kernel.

   (b) Show that the second-order Sobolev kernel $k : [0, 1] \times [0, 1] \to \mathbb{R}$ given by
$$k(x, y) := \int_0^{x \wedge y} (x - z)(y - z), dz$$
is a kernel. Verify further that for $x \in [0, 1]$, the function $k_x : [0, 1] \to \mathbb{R}$ given by
$$k_x(y) := \int_0^{x \wedge y} (x - z)(y - z)\, dz$$
satisfies $k_x''(y) = (x - y)_+$.

   (c) Show that the function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ given by $k(x, y) := (\alpha + \|x - y\|_2^2)^{-1/2}$ is a kernel for each $\alpha > 0$.

9. (a) Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel with associated RKHS $\mathcal{H}$ such that the set of functions $\mathcal{K} := \{k(\cdot, x) : x \in \mathcal{X}\}$ is linearly independent. Show that for any $n \in \mathbb{N}$, any distinct $x_1, \ldots, x_n \in \mathcal{X}$ and any $y_1, \ldots, y_n \in \mathbb{R}$, there exists $f \in \mathcal{H}$ such that $f(x_i) = y_i$ for all $i = 1, \ldots, n$.
   [*Hint: First show that the kernel matrix must be invertible.*]

   (b) Show that when k is a Gaussian kernel on $\mathbb{R}$, the corresponding set $\mathcal{K}$ above is linearly independent.
   [*Hint: Any matrix of the form*
$$A = \begin{pmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^{n-1} \\ 1 & a_2 & a_2^2 & \cdots & a_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & a_n & a_n^2 & \cdots & a_n^{n-1} \end{pmatrix} \in \mathbb{R}^{n \times n}$$
is known as a (square) Vandermonde matrix*, and satisfies* $\det(A) = \prod_{1 \le i < j \le n}(a_j - a_i)$*, so is invertible whenever* $a_1, \ldots, a_n$ *are distinct.*]

10. Let $\hat{\alpha}$ be a minimiser of $\|Y - K\alpha\|_2^2 + \lambda \alpha^\top K \alpha$ over α, with K being a kernel matrix as usual (i.e. symmetric positive semi-definite). Show that $K\hat{\alpha} = K(K + \lambda I)^{-1} Y$.

11. Let $\mathcal{H}$ be a RKHS of functions on $\mathcal{X}$ with reproducing kernel k and suppose $f^0 \in \mathcal{H}$. Let $x_1, \ldots, x_n \in \mathcal{X}$ and let K be the kernel matrix $K_{ij} = k(x_i, x_j)$. Show that
$$\Big(f^0(x_1), \ldots, f^0(x_n)\Big)^\top = K\alpha,$$
for some $\alpha \in \mathbb{R}^n$ and moreover that $\|f^0\|_{\mathcal{H}}^2 \ge \alpha^\top K \alpha$.

12. Consider minimising

$$c(Y, X, f(x_1) + z_1^\top \beta, \ldots, f(x_n) + z_n^\top \beta) + J(\|f\|_{\mathcal{H}}^2)$$

over $f \in \mathcal{H}$ and $\beta \in \mathbb{R}^d$, where $\mathcal{H}$ is an RKHS, $x_1, \ldots, x_n \in \mathcal{X}$ and $z_1, \ldots, z_n \in \mathbb{R}^d$. Here c is an arbitrary loss function and J is strictly increasing. Let k be the reproducing kernel of $\mathcal{H}$. Show that any minimiser $\hat{g}(x, z) = \hat{f}(x) + z^\top \hat{\beta}$ may be written as

$$\hat{g}(x, z) = z^\top \hat{\beta} + \sum_{i=1}^{n} \hat{\alpha}_i k(x, x_i)$$

where $\hat{\alpha}_i \in \mathbb{R}$ for $i = 1, \ldots, n$.

13. This question uses the following facts. Suppose $k = \sum_{j=1}^{p} k_j$ where $k_1, \ldots, k_p$ are kernels with associated RKHS's $\mathcal{H}_1, \ldots, \mathcal{H}_p$ having corresponding norms $\|\cdot\|_{\mathcal{H}_1}, \ldots, \|\cdot\|_{\mathcal{H}_p}$. Then the RKHS $\mathcal{H}$ with reproducing kernel k satisfies

$$\mathcal{H} = \left\{ \sum_{j=1}^{p} f_j : f_j \in \mathcal{H}_j \text{ for all } j = 1, \ldots, p \right\}$$

with squared norm

$$\|f\|_{\mathcal{H}}^2 = \inf \left\{ \sum_{j=1}^{p} \|f_j\|_{\mathcal{H}_j}^2 : \sum_{j=1}^{p} f_j = f,\ f_j \in \mathcal{H}_j \text{ for all } j \right\}.$$

It can be shown that the infimum is achieved uniquely, so given $f \in \mathcal{H}$, there exists a unique $(f_1, \ldots, f_p) \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_p$ such that $\sum_{j=1}^{p} f_j = f$ and $\|f\|_{\mathcal{H}}^2 = \sum_{j=1}^{p} \|f_j\|_{\mathcal{H}_j}^2$.

(a) For $f \in \mathcal{H}$, let

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

Suppose now that $(\hat{f}_1, \ldots, \hat{f}_p)$ minimises

$$Q_2(f_1, \ldots, f_p) = c\left(Y, x_1, \ldots, x_n, \sum_{j=1}^{p} f_j(x_1), \ldots, \sum_{j=1}^{p} f_j(x_n)\right) + \lambda \sum_{j=1}^{p} \|f_j\|_{\mathcal{H}_j}^2$$

over $(f_1, \ldots, f_p) \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_p$. Show that then $Q_2(\hat{f}_1, \ldots, \hat{f}_p) = Q_1(\hat{f})$ where $\hat{f} := \sum_{j=1}^{p} \hat{f}_j$. Show furthermore that $\hat{f}$ minimises $Q_1(f)$ over $f \in \mathcal{H}$.

(b) Finally show that $\hat{f}_j(\cdot) = \sum_{i=1}^{n} \hat{\alpha}_i k_j(\cdot, x_i)$ for all j, where $\hat{\alpha} \in \mathbb{R}^n$ minimises

$$M(\alpha) = c(Y, x_1, \ldots, x_n, K\alpha) + \lambda \alpha^\top K \alpha$$

over $\alpha \in \mathbb{R}^n$.