MODERN STATISTICAL METHODS Part III
Example Sheet 1 (of 4) Michaelmas 2019

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 $k = 1, \ldots, p$. 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
$$\hat{\beta} = (X^TX + \lambda I)^{-1}X^TY.$$
From here onwards, whenever we refer to ridge regression, we will assume X has had its columns mean-centred.
Solution: Differentiating w.r.t. μ, we have that the minimising μ, $\hat{\mu}$ is defined by the following equation
$$\mathbf{1}^T(Y - \hat{\mu}\mathbf{1} - X\beta) = \mathbf{1}^T(Y - \hat{\mu}\mathbf{1}) = 0,$$
giving $\hat{\mu} = \bar{Y}$. Thus the minimising β in fact minimises
$$\|Y - \bar{Y}\mathbf{1} - X\beta\|_2^2 + J(\beta).$$
Letting $\tilde{Y} = Y - \bar{Y}\mathbf{1}$ and specialising to the ridge objective, our optimisation problem is to minimise
$$\|\tilde{Y} - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$
over $\beta \in \mathbb{R}^p$. Differentiating w.r.t. β we see the minimiser $\hat{\beta}$ satisfies
$$X^T(\tilde{Y} - X\hat{\beta}) = \lambda\hat{\beta}$$
so
$$\begin{aligned} X^T\tilde{Y} &= (X^TX + \lambda I)\hat{\beta} \\ \hat{\beta} &= (X^TX + \lambda I)^{-1}X^T\tilde{Y}. \end{aligned}$$
Finally note that $X^TY = X^T\tilde{Y}$.

2. Consider performing ridge regression when $Y = X\beta^0 + \varepsilon$, where $X \in \mathbb{R}^{n\times p}$ has full column rank, and $\mathrm{Var}(\varepsilon) = \sigma^2 I$. Let the SVD of X be UDV^T and write $U^TX\beta^0 = \gamma$. 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 suppose the size of the signal is n, so $\|X\beta^0\|_2^2 = n$. For what γ is the mean squared prediction error above minimised? For what γ is it maximised?
Solution: From lectures, we know that
$$X\hat{\beta}_\lambda^R = UD^2(D^2 + \lambda I)^{-1}U^T(X\beta^0 + \varepsilon) = UD^2(D^2 + \lambda I)^{-1}(\gamma + U^T\varepsilon).$$
Also, $X\beta^0 = UU^TX\beta^0 = U\gamma$. Thus $\mathbb{E}\|X\beta^0 - X\hat{\beta}_\lambda^R\|_2^2$ equals
$$\mathbb{E}\|U\{I - D^2(D^2 + \lambda I)^{-1}\}\gamma + UD^2(D^2 + \lambda I)^{-1}U^T\varepsilon\|_2^2 = \sum_{j=1}^p \left(\frac{\lambda}{\lambda + D_{jj}^2}\right)^2 \gamma_j^2 + \mathbb{E}\|UD^2(D^2 + \lambda I)^{-1}U^T\varepsilon\|_2^2.$$
Applying the 'trace trick' to the second term gives
$$\begin{aligned} \mathbb{E}\|UD^2(D^2 + \lambda I)^{-1}U^T\varepsilon\|_2^2 &= \mathrm{tr}\mathbb{E}\{UD^2(D^2 + \lambda I)^{-1}U^T\varepsilon\varepsilon^TUD^2(D^2 + \lambda I)^{-1}U^T\} \\ &= \sigma^2\mathrm{tr}\{UD^4(D^2 + \lambda I)^{-2}U^T\} \\ &= \sigma^2\mathrm{tr}\{U^TUD^4(D^2 + \lambda I)^{-2}\} \\ &= \sigma^2\sum_{j=1}^p \frac{D_{jj}^4}{(\lambda + D_{jj}^2)^2}. \end{aligned}$$

For the last part, note that $\|\gamma\|_2^2 = \|X\beta^0\|_2^2$ as $X\beta^0 = U\gamma$ and only the first term in the expression for MSPE depends on γ. As

$$\left(\frac{\lambda}{\lambda + D_{jj}^2}\right)^2$$

is increasing in j, the MSPE is minimised when $\gamma_1^2 = n$ (and all other entries are zero), so all the signal is in the direction of the first principal component. It is maximised when $\gamma_p^2 = n$ (and all other entries are zero).

3. Show that the ridge regression estimates can be obtained by ordinary least squares regression on an augmented data set with $\sqrt{\lambda}I$ added to the bottom of X (where I here is $p \times p$), and p zeroes added to the end of the response Y.
Solution: The least squares objective is

$$\|Y - X\beta\|_2^2 + \|0 - \sqrt{\lambda}I\beta\|_2^2 = \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2.$$

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?
   Solution: On server j we compute $\hat{\Sigma}^{(j)} := {X^{(j)}}^T X^{(j)} \in \mathbb{R}^{p\times p}$ and $\hat{\rho}^{(j)} := {X^{(j)}}^T Y \in \mathbb{R}^p$. These are sent to the central server, which computes

   $$\hat{\Sigma} := X^T X = \sum_{j=1}^m \hat{\Sigma}^{(j)} \qquad \hat{\rho} := X^T Y = \sum_{j=1}^m \hat{\rho}^{(j)}.$$

   The ridge regression estimates can then be calculated as $\hat{\beta}_\lambda = (\hat{\Sigma} + \lambda I)^{-1}\hat{\rho}$. Thus the computation at each server is $O(p^2 n/m)$, whilst the cost at the central server is $O(p^2 m + p^3)$: $p^2 m$ for adding the $\hat{\Sigma}^{(j)}$ and p^3 for inverting $\hat{\Sigma} + \lambda I$.

   (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?
   Solution: Let the data on server j be $X^{(j)} \in \mathbb{R}^{n\times p/m}$. Form $K^{(j)} = X^{(j)}{X^{(j)}}^T$ at each server, and send this $n \times n$ matrix to the central server. At the central server, form $K = \sum_{j=1}^m K^{(j)}$ and compute $K(K + \lambda I)^{-1}Y$. The computation at each server is $O(n^2 p/m)$ and the cost at the central server is $O(n^3 + n^2 m)$.

5. Prove Proposition 4 in our notes. *Hint: For part (ii) it may help to consider the eigendecompositions of positive semi-definite matrices $K^{(1)}$ and $K^{(2)}$ derived from kernels k_1 and k_2 in the form $K^{(1)} = PDP^T = \sum_{i=1}^n P_i P_i^T D_{ii}$ for example.*
Solution: For (i), given observations $x_1, \dots, x_n$, consider the derived kernel matrices $K_1, K_2, \dots \in \mathbb{R}^{n\times n}$ (here we go against the convention of the course and do not mean the first column of K by K_1). We have

$$a^T(\alpha_1 K_1 + \alpha_2 K_2)a = \alpha_1 a^T K_1 a + \alpha_2 a^T K_2 a \geq 0.$$

Also
$$a^T(\lim_{m\to\infty} K_m)a = \lim_{m\to\infty} a^T K_m a \geq 0.$$
Turning to (ii), write $K_1 = \sum_{i=1}^n P_i P_i^T D_{ii}$, $K_2 = \sum_{i=1}^n Q_i Q_i^T \Lambda_{ii}$. Note $D_{ii}, \Lambda_{mm} \geq 0$ as K_1 and K_2 are positive semi-definite. Thus the entrywise or Hadamard product $K_1 \circ K_2$ has jkth entry
$$\sum_{i,m} P_{ji} P_{ki} D_{ii} Q_{jm} Q_{km} \Lambda_{mm} = \sum_{i,m} (P_i \circ Q_m)_j D_{ii} \Lambda_{mm} (P_i \circ Q_m)_k.$$
This is the jkth entry of
$$\sum_{i,m} (P_i \circ Q_m) D_{ii} \Lambda_{mm} (P_i \circ Q_m)^T$$
which is a linear combination of positive semi-definite matrices with non-negative coefficients $D_{ii}\Lambda_{mm} \geq 0$.

6. Let $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 < 1\}$. Show that $k(x, x') = (1 - x^T x')^{-\alpha}$ defined on $\mathcal{X} \times \mathcal{X}$, where $\alpha > 0$, is a kernel.
Solution: Note that $|x^T x'| \leq \|x\|_2 \|x'\|_2 < 1$ by Cauchy–Schwarz so, Taylor's theorem tells us that
$$k(x, x') = 1 + \alpha x^T x' + \frac{1}{2!}\alpha(\alpha+1)(x^T x')^2 + \cdots .$$
The ratio test shows that the series converges whenever $|x^T x'| < 1$. This is an infinite sum of products of kernels, and so is a kernel by Proposition 4 in our notes.

7. 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 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.
Solution: Form matrices $K^{(1)}, K^{(2)}, K^{(3)} \in \mathbb{R}^{n\times n}$:
$$\begin{aligned} K^{(1)} &= XX^T \\ K^{(2)}_{ij} &= (x_i^T x_j)^2 \\ K^{(3)}_{ij} &= \sum_{k=1}^p X_{ik}^2 X_{jk}^2. \end{aligned}$$
Finally calculate $K = \lambda_1^{-1} K^{(1)} + (2\lambda_2)^{-1} K^{(2)} + \{\lambda_3^{-1} - (2\lambda_2)^{-1}\} K^{(3)}$. We may then see that the fitted values are
$$K(K+I)^{-1}Y.$$
Note that computation of K requires $O(n^2 p)$ operations.

8. Let $\hat{\alpha}$ be a minimiser of $\|Y - K\alpha\|_2^2 + \lambda \alpha^T K \alpha$ over α, with K being a kernel matrix as usual (i.e. symmetric positive semi-definite). Show that $K\hat{\alpha} = K(K+\lambda)^{-1}Y$.
Solution: Differentiating, we obtain
$$K(Y - K\hat{\alpha}) = \lambda K \hat{\alpha}$$
so
$$\begin{aligned} KY &= (K + \lambda I)K\hat{\alpha} \\ (K+\lambda I)^{-1} KY &= K\hat{\alpha}. \end{aligned}$$
Finally note that $K(K+\lambda I) = (K+\lambda I)K$, so $(K+\lambda I)^{-1}K = K(K+\lambda I)^{-1}$.

9. Consider minimising

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

over $f \in \mathcal{H}$ and $\mu \in \mathbb{R}$ where $\mathcal{H}$ is an RKHS. 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 $\hat{g}(\cdot) = \hat{f}(\cdot) + \hat{\mu}$ may be written as

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

where $\hat{\alpha}_i \in \mathbb{R}$ for $i = 1, \ldots, n$.
Solution: Write $\hat{g} = \hat{f} + \hat{\mu}$. Note we may decompose $\hat{f} = u + v$ where $u \in V := \text{span}\{k(\cdot, x_1), \ldots, k(\cdot, x_n)\}$ and $v \in V^{\perp}$. Then

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

Meanwhile, by Pythagoras' theorem we have

$$J(\|\hat{f}\|_{\mathcal{H}}^2) = J(\|u\|_{\mathcal{H}}^2 + \|v\|_{\mathcal{H}}^2) \geq J(\|u\|_{\mathcal{H}}^2)$$

with equality iff. $v = 0$. Thus by optimality of $\hat{g}$, $v = 0$.

10. This question proves a result needed for Theorem 7 in our notes. 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

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

for some $\alpha \in \mathbb{R}^n$ and moreover that $\|f^0\|_{\mathcal{H}}^2 \geq \alpha^T K \alpha$.
Solution: Let $V = \text{span}\{k(\cdot, x_1), \ldots, k(\cdot, x_n)\}$ and write $f^0 = u + v$ where $u \in V$ and $v \in V^{\perp}$. Then

$$f^0(x_i) = \langle f^0, k(\cdot, x_i)\rangle = \langle u, k(\cdot, x_i)\rangle.$$

Write $u = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$. Then

$$f^0(x_i) = \sum_{j=1}^{n} \alpha_j \langle k(\cdot, x_j), k(\cdot, x_i)\rangle = \sum_{j=1}^{n} \alpha_j k(x_j, x_i) = K_i^T \alpha,$$

where K_i is the ith column (or row) of K. Thus $K\alpha = \Big(f^0(x_1), \ldots, f^0(x_n)\Big)^T$. By Pythagoras' theorem

$$\|f^0\|_{\mathcal{H}}^2 = \|u\|_{\mathcal{H}}^2 + \|v\|_{\mathcal{H}}^2 \geq \|u\|_{\mathcal{H}}^2 = \alpha^T K \alpha.$$

11. Show from first principles that the Sobolev kernel is indeed a (positive definite) kernel.
Solution: Let $x_1, \ldots, x_n \in [0, 1]$ and assume without loss of generality that $x_1 \geq x_2 \geq \cdots \geq x_n$. Let $\delta_j = x_j - x_{j+1} \geq 0$ for $j = 1, \ldots, n-1$ and set $\delta_n = x_n$. Also let $J^{(j)}$ be the matrix with $J_{ik}^{(j)} = 1$ for all $i, k \leq j$ and all other entries equal to zero. Then if $i < j$,

$$\min(x_i, x_j) = x_j = \sum_{k=j}^{n} \delta_k = \left(\sum_k \delta_k J^{(k)}\right)_{ij}.$$

Thus $K = \sum_k \delta_k J^{(k)}$. Each $J^{(k)}$ is positive semi-definite as $a^T J^{(k)} a = (\sum_{j=1}^k a_j)^2 \geq 0$, whence K is also.

12. Let $\mathcal{H}$ be an RKHS with reproducing kernel k. Show that if $h_x \in \mathcal{H}$ has the property that $\langle h_x, f\rangle = f(x)$ for all $f \in \mathcal{H}$, then $h_x(\cdot) = k(\cdot, x)$.
Solution: We know that $\langle k(\cdot, x), f\rangle = f(x)$ for all $f \in \mathcal{H}$. Thus

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

for all $x' \in \mathcal{X}$.

13. Prove that if k is a reproducing kernel for RKHS's $\mathcal{H}_1$ and $\mathcal{H}_2$, then $\mathcal{H}_1 = \mathcal{H}_2$, so the RKHS is uniquely determined by k. *Hint: First argue that it is enough to show the result for* $\mathcal{H}_1 \subseteq \mathcal{H}_2$. *Next consider decomposing each* $f \in \mathcal{H}_2$ *as* $f = u + v$ *with* $u \in \mathcal{H}_1$ *and* $v \in \mathcal{H}_1^\perp$ *and argue that* $v = 0$.

    Solution: First note that $\mathcal{H}_0 := \mathcal{H}_1 \cap \mathcal{H}_2$ is an RKHS with reproducing kernel k. It is enough to show that $\mathcal{H}_0 = \mathcal{H}_1$. Take $f \in \mathcal{H}_1$. As $\mathcal{H}_0$ is a closed subspace of $\mathcal{H}_1$, we may decompose $f = u + v$ where $u \in \mathcal{H}_0$ and $v \in \mathcal{H}_0^\perp$. But

$$f(x) = \langle k(\cdot, x), f \rangle = \langle k(\cdot, x), u \rangle = u(x)$$

    as $\langle k(\cdot, x), v \rangle = 0$ owing to $k(\cdot, x) \in \mathcal{H}_0$. Thus $f \in \mathcal{H}_0$, so $\mathcal{H}_1 = \mathcal{H}_0$.