Modern Statistical Methods
Example sheet 1

1. Sufficient to minimize $\|Y - \mu\mathbb{1} - X\beta\|_2^2$ w.r.t. μ as $J(\beta)$ only.

Write $\mu = \bar{Y} + c$

$$\|Y - \bar{Y}\mathbb{1} - c\mathbb{1} - X\beta\|_2^2 = \|Y - \bar{Y}\mathbb{1} - X\beta\|_2^2 + nc^2 - c\sum_{i=1}^{n}\left(Y_i - \bar{Y} - \sum_{j=1}^{p} X_{ij}\beta_j\right)$$

$$= \|Y - \bar{Y}\mathbb{1} - X\beta\|_2^2 + nc^2 - c\sum_{i=1}^{n}(Y_i - \bar{Y}) + c\sum_{i=1}^{n}\sum_{j=1}^{p} X_{ij}\beta_j$$

$$= \|Y - \bar{Y}\mathbb{1} - X\beta\|_2^2 + nc^2 + c\sum_{j=1}^{p}\underbrace{\left(\sum_{i=1}^{n} X_{ij}\right)}_{=0 \text{ as columns mean-centred}}\beta_j$$

$$= \|Y - \bar{Y}\mathbb{1} - X\beta\|_2^2 + nc^2 \geq \|Y - \bar{Y}\mathbb{1} - X\beta\|_2^2$$

$\therefore \hat{\mu} = \bar{Y}$

Substituting $Y \to Y - \bar{Y}\mathbb{1}$,

$$\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

differentiate w.r.t. β

$$-X^T(Y - X\hat{\beta}) + \lambda\hat{\beta} = 0$$

$$\Rightarrow \hat{\beta} = (X^TX + \lambda I)^{-1}X^TY$$

2. $\|Xv\|_2^2 = v^TX^TXv = v^TVD^2V^Tv$

$a \equiv V^Tv$, $\|a\|_2^2 = 1$ as $VV^T = I$

$$\|Xv\|_2^2 = a^TD^2a = \sum_{i=1}^{k-1} a_i^2 D_{ii}^2 + \sum_{j=k}^{p} a_j^2 D_{jj}^2$$

For $i = 1, \ldots, k-1$, know that $v^{(i)} = V_i$, thus require

$$V_i^TX^TXv = V_i^TVD^2V^Tv = V_i^TVD^2a = 0 \quad \forall\, i = 1, \ldots, k-1$$

as $V_i^TVD^2 = [a_1^{(i)T}D^2$ and $a_1^{(i)}$ has only one non-zero component $(a^{(i)})_j = \delta_{ij}$,

$a_i = 0 \quad \forall\, i = 1, \ldots, k-1$

Thus,

$\|Xv\|_2^2 = \sum_{j=k}^{p} a_j^2 D_{jj}^2$ is maximized by $a_k = \pm 1$, $a_{j\neq k} = 0$ as $D_{kk} \geq D_{jj} : j > k \geq 0$.

$\Rightarrow (a)_j = \delta_{kj}$ and $v^{(k)} = V_k$, $u^{(k)} = UDV^T V_k = D_{kk} U_k$

by induction. □

3. $Y = X\beta^0 + \varepsilon \qquad Var(\varepsilon) = \sigma^2 I$

$X = UDV^T \qquad U^T X\beta^0 = \gamma$

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

$X\hat{\beta}_\lambda^R = UDV^T(VD^2V^T + \lambda I)^{-1} VDU^T(X\beta^0 + \varepsilon)$

$= UD(D^2 + \lambda I)^{-1} DU^T(X\beta^0 + \varepsilon)$

$= UD(D^2 + \lambda I)^{-1} D\gamma + UD(D^2 + \lambda I)^{-1} DU^T\varepsilon$

$\frac{1}{n}\mathbb{E}\|X\beta^0 - X\hat{\beta}_\lambda^R\|_2^2 = \frac{1}{n}\mathbb{E}\|U^TX\beta^0 - U^TX\hat{\beta}_\lambda^R\|_2^2$ as $UU^T = I$

$= \frac{1}{n}\mathbb{E}\|\gamma - D(D^2+\lambda I)^{-1}D\gamma - D(D^2+\lambda I)^{-1}DU^T\varepsilon\|_2^2$

$= \frac{1}{n}\|\gamma - D(D^2+\lambda I)^{-1}D\gamma\|_2^2 + \frac{1}{n}\mathbb{E}\|D(D^2+\lambda I)^{-1}DU^T\varepsilon\|_2^2$ as $\mathbb{E}\varepsilon = 0$.

$= \frac{1}{n}\sum_{j=1}^{p}\left(\frac{\lambda}{\lambda + D_{jj}^2}\right)^2 \gamma_j^2 + \frac{\sigma^2}{n}\sum_{j=1}^{p}\frac{D_{jj}^4}{(\lambda + D_{jj}^2)^2}$.

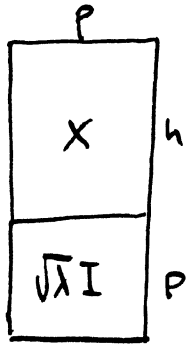
$\|X\beta^0\|_2^2 = \|U^TX\beta^0\|_2^2 = \|\gamma\|_2^2 = n$

$\sum_{j=1}^{p}\left(\frac{\lambda}{\lambda + D_{jj}^2}\right)^2 \gamma_j^2 \geq \sum_{j=1}^{p}\left(\frac{\lambda}{\lambda + D_{11}^2}\right)^2 \gamma_j^2$

$\therefore$ Minimized for $\gamma_i = \delta_{1i}$ (i.e. $\gamma^* = (1, 0, \ldots, 0)^T$), similarly maximized for $\gamma_i = \delta_{pi}$ (i.e. $\gamma = (0, 0, \ldots, 0, 1)^T$).

(Alignment with large / small principal components.)

4. $\hat{\beta}^{OLS} = (X^T X)^{-1} X^T Y$

$\tilde{X} =$ 

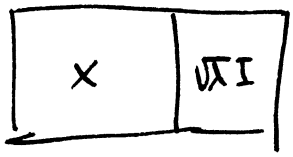
$\tilde{Y} =$ 

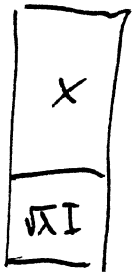
$\hat{\beta}_\lambda = (\tilde{X}^T \tilde{X})^{-1} \underbrace{\tilde{X}^T \tilde{Y}}_{= X^T Y}$

$= (\tilde{X}^T \tilde{X})^{-1} X^T Y$

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

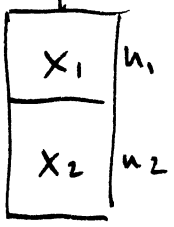
$= \hat{\beta}^R_\lambda$

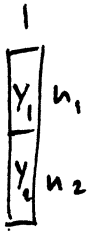
$\tilde{X}^T \tilde{X} =$ 



The added block only has contributions on the diagonal.

5. (a) Consider

$X =$ 

$Y =$ 

$$(X^T X)_{jk} = \sum_{i=1}^{n} X_{ij} X_{ik} = \sum_{i=1}^{n_1} X_{ij} X_{ik} + \sum_{i=n_1+1}^{n_1+n_2} X_{ij} X_{ik} \quad \text{where } n = n_1 + n_2$$

$$\Rightarrow X^T X = X_1^T X_1 + X_2^T X_2$$

Similarly $X^T Y = X_1^T Y_1 + X_2^T Y_2$.

Therefore, on each server compute $X^{(i)T} X^{(i)}$ and $X^{(i)T} y^{(i)}$,

$O(p^2 \frac{n}{m} + p \frac{n}{m}) = O(p^2 \frac{n}{m})$ time;

$p^2 + p = p(p+1)$ numbers communicated,

and on central server

$O(\underbrace{mp^2 + mp}_{\text{additions}} + \underbrace{p^3}_{\text{inversion}} + \underbrace{p^2}_{\text{multiplication}}) = O(p^3)$ for const m.

(6) $X\hat{\beta}_\lambda = XX^T(XX^T + \lambda I)^{-1} y$

$$XX^T = \sum_{i=1}^{P} x_i x_i^T$$

Hence calculate $X^{(i)} X^{(i)T}$ at each server

$$O(n^2 \tfrac{P}{m})$$

and communicate n^2 numbers, calculate $X\hat{\beta}_\lambda$ at central server

$$O(\underbrace{n^2 m}_{\text{additions}} + \underbrace{n^3}_{\text{inversion}} + \underbrace{n^2}_{\text{multiplication}}) = O(n^3) .$$

6. $k_1, k_2, \ldots$ kernels

(i) If $\alpha_1, \alpha_2 \geq 0$ then $\alpha_1 k_1 + \alpha_2 k_2$ is a kernel.

Let $x_1, \ldots, x_n \in \mathcal{X}$, $\beta_1, \ldots, \beta_n \in \mathbb{R}$

$$\beta^T(\alpha_1 K_1 + \alpha_2 K_2)\beta = \sum_{i,j} \beta_i (\alpha_1 k_1(x_i, x_j) + \alpha_2 k_2(x_i, x_j)) \beta_j$$

$$= \alpha_1 \sum \beta_i k_1(x_i, x_j) \beta_j + \alpha_2 \sum \beta_i k_2(x_i, x_j) \beta_j \geq 0$$

as k_1, k_2 kernels and $\alpha_1, \alpha_2 \geq 0$ □

If $\lim_{m\to\infty} k_m(x, x') =: k(x, x')$ exists for all $x, x' \in \mathcal{X}$, then k is a kernel.

Let $x_1, \ldots, x_n \in \mathcal{X}$, $\beta_1, \ldots, \beta_n \in \mathbb{R}$

$$\beta^T K \beta = \sum_{i,j} \beta_i k(x_i, x_j) \beta_j = \sum_{i,j} \beta_i \left(\lim_{m\to\infty} k_m(x_i, x_j)\right) \beta_j$$

$$= \sum_{i,j} \lim_{m\to\infty} \left(\beta_i k_m(x_i, x_j) \beta_j\right) = \lim_{m\to\infty} \sum_{i,j} \beta_i k_m(x_i, x_j) \beta_j \geq 0$$

as k_m are kernels □

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

$$K_{ij}^{(1)} = P^{(1)} D^{(1)} P^{(1)T} = \sum_{i=1}^{n} P_i^{(1)} P_i^{(1)T} D_{ii}^{(1)}$$

where $P_i P_i^T$ is a positive semi-definite matrix as $\alpha^T P_i P_i^T \alpha = \|P_i^T \alpha\|^2 \geq 0$ $\forall \alpha$

and $D_{ii} \geq 0$.

$$K = K^{(1)} \circ K^{(2)} = \sum_{i,j} D_{ii}^{(1)} D_{jj}^{(2)} \left(P_i^{(1)} P_i^{(1)T}\right) \circ \left(P_j^{(2)} P_j^{(2)T}\right)$$

$$= \sum_{i,j} D_{ii}^{(1)} D_{jj}^{(2)} \left(P_i^{(1)} \circ P_j^{(2)}\right)\left(P_i^{(1)} \circ P_j^{(2)}\right)^T \geq 0$$

as $\left(P_i^{(1)} \circ P_j^{(2)}\right)\left(P_i^{(1)} \circ P_j^{(2)}\right)^T = \tilde{P}_{ij} \tilde{P}^T \geq 0$ (pos semi-definite)

and $D_{ii}^{(1)}, D_{jj}^{(2)} \geq 0 \quad \forall i, j$, finally the sum of pos. semi-definite matrices is positive semi definite as shown above. □

7. $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 < 1\}$ $\qquad k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$k(x, x') = (1 - x^T x')^{-\alpha}$, $\alpha > 0$

$|x^T x'|^2 \leq \|x\|_2^2 \|x'\|_2^2 < 1 \quad \Rightarrow y \equiv x^T x', |y| < 1$

$f(y) = (1-y)^{-\alpha}$ is analytic for $y \in (-1, 1)$

$k_m(x, x') := \sum_{n=0}^{m} \frac{\alpha(\alpha+1)\ldots(\alpha+n-1)}{n!} (x^T x')^n$

then $k(x, x') = \lim_{m \to \infty} k_m(x, x')$

$(x^T x') \xrightarrow{4.ii} (x^T x')^n \xrightarrow{4.i} \sum^{m} \xrightarrow[\lim]{4.i} k(x, x')$ is a kernel

(is a kernel)

□

9. $\|Y - K\alpha\|_2^2 + \lambda \alpha^T K \alpha$, $K \geq 0$, $K^T = K$?

Differentiate over α

$$-K^T(Y - K\hat{\alpha}) + \lambda K \hat{\alpha} = 0$$

~~$\Rightarrow K(KK\hat{\alpha} + \lambda K\hat{\alpha} - KY) = 0$~~

$$-KY + KK\hat{\alpha} + \lambda K\hat{\alpha} = 0$$

$$\Rightarrow K\hat{\alpha} = (K+\lambda)^{-1} KY = K(K+\lambda)^{-1} Y$$

10. Minimise

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

over $f \in \mathcal{H}$, $\mu \in \mathbb{R}$, c is arbitrary, J is strictly increasing

Suppose $\hat{g} = \hat{\mu} + u + v$ where $u \in \text{span}\{k(\cdot, x_1), \dots, k(\cdot, x_n)\}$ and $v \in V^\perp$.

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

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

~~Also~~ i.e. $c(Y, X, \hat{\mu} + u + v) = c(Y, X, \hat{\mu} + u)$

Also $J(\|\hat{f}\|_{\mathcal{H}}^2) = J(\|u+v\|_{\mathcal{H}}^2) = J(\|u\|_{\mathcal{H}}^2 + \|v\|_{\mathcal{H}}^2)$ as $\langle u, v \rangle = 0$

$$\geq J(\|u\|_{\mathcal{H}}^2)$$ equality iff $v = 0$ by monotonicity of J.

$$\therefore \hat{g}(\cdot) = \hat{\mu} + u(\cdot) = \hat{\mu} + \sum_{i=1}^{n} \hat{\alpha}_i k(\cdot, x_i) \quad , \hat{\alpha}_i \in \mathbb{R}^n \quad \square$$

11. $\Phi = UDV^T$

$$K = \Phi \Phi^T = UDV^T VDU^T = UD^2U^T = UD(UD)^T$$

$$(K)_{ij} = \sum_{k=1}^{d} (UD)_{ik} (UD)_{jk} = \sum_{k=1}^{d} u_i^{(k)} u_j^{(k)}$$

Eigenvalue decomposition gives

$$K = PEP^T \quad \text{as } K^T = K$$

$$= PE(PE)^T \quad \text{where } \tilde{E}\tilde{E} = E \text{ (sqrt on diagonal)}$$

Identify $P\tilde{E} \sim UD$ with { zero filled columns added if $n > d$
~~linearly independent cols discarded if $n < d$~~

There are the principal components as $K = UD(UD)^T$ provides $n^2 > nd$ ~~linear~~ equations for the components of $u^{(k)}$ $k = 1, \ldots, d$ (not all eq independent).

From $K = UD^2U^T$ can see that this provides an eigenvalue decomposition of K (provided $U(D)$ augmented with extra columns). Eigenvalue decompositions are unique up to ordering.

If $d > n$, fill K with extra zeros.

$\therefore$ Can deduce principal components from eigenvalue decomposition of K.