

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.

$$\text{write } \mu = \bar{Y} + c$$

$$\begin{aligned}\|Y - \bar{Y}\mathbb{1} - c\mathbb{1} - X\beta\|_2^2 &= \|Y - \bar{Y}\mathbb{1} - X\beta\|_2^2 + nc^2 - c \sum_{i=1}^n (Y_i - \bar{Y}) - \sum_{j=1}^p X_{ij}\beta_j \\ &= \|Y - \bar{Y}\mathbb{1} - X\beta\|_2^2 + nc^2 - c \sum_{i=1}^n (Y_i - \bar{Y}) + c \sum_{i=1}^n \sum_{j=1}^p X_{ij}\beta_j \\ &\geq \|Y - \bar{Y}\mathbb{1} - X\beta\|_2^2 + nc^2 + c \sum_{j=1}^p \left(\sum_{i=1}^n X_{ij} \right) \beta_j \\ &\quad = 0 \text{ as column mean-centred} \\ &= \|Y - \bar{Y}\mathbb{1} - X\beta\|_2^2 + nc^2 \geq \|Y - \bar{Y}\mathbb{1} - X\beta\|_2^2\end{aligned}$$

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

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

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

differentiate w.r.t. β

$$-\cancel{X^T}(Y - \bar{Y}\mathbb{1}) + \cancel{\lambda} \hat{\beta} = 0$$

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

$$2. \|Xv\|_2^2 = v^T X^T X v = v^T V D^2 V^T v$$

$$a \equiv V^T v, \|a\|_2^2 = 1 \text{ as } VV^T = I$$

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

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

$$V_i^T X^T X v = V_i^T V D^2 V^T v = V_i^T V D^2 a = 0 \quad \forall i = 1, \dots, k-1$$

as $V_i^T V D^2 = \{a_i^2\}_{ii} D^2$ and $a_i^{(i)}$ has only one non-zero component $(a^{(i)})_j = \delta_{ij}$,

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

Thus,

$$\|X\beta^0\|_2^2 = \sum_{j=1}^p a_j^2 D_{jj}^2 \quad \text{is maximized by } a_k = \overset{\pm}{1}, a_{j \neq k} = 0 \text{ as } D_{jj} \geq D_{ij}: j > k \geq 0$$

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

by induction. ~~at k=1~~ \square

$$3. Y = X\beta^0 + \varepsilon \quad \text{Var}(\varepsilon) = \sigma^2 I$$

$$X = UDV^T \quad U^T X \beta^0 = \cancel{UDV^T \beta^0} = \gamma$$

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

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

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

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

$$\frac{1}{n} E \|X\beta^0 - X_{\beta^0}^R\|_2^2 = \frac{1}{n} E \|U^T X \beta^0 - U^T X_{\beta^0}^R\|_2^2 \quad \text{as } UU^T = I$$

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

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

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

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

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

\therefore Minimized for $r_i = \delta_{1i}$ (i.e. $r^* = (1, 0, \dots, 0)^T$), similarly maximized for $r_i = \delta_{p,i}$ (i.e. $r = (0, 0, \dots, 0, 1)^T$).

(Alignment with large / small principal components.)

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

$$\tilde{X} = \begin{array}{c|c} \rho & X \\ \hline & \sqrt{\lambda} I \end{array} \quad n$$

$$\tilde{Y} = \begin{array}{c|c} 1 & y \\ \hline 0 & P \end{array} \quad n$$

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

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

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

$$= \hat{\beta}_\lambda^R$$

$$\tilde{X}^T \tilde{X} = \begin{array}{c|c} X & \sqrt{\lambda} I \end{array}$$

$$\begin{array}{c} X \\ \hline \sqrt{\lambda} I \end{array}$$

The added block only has contributions on the diagonal.

5.(a) Consider P

$$X = \begin{array}{c|c} X_1 & u_1 \\ \hline X_2 & u_2 \end{array} \quad Y = \begin{array}{c|c} 1 & \\ \hline Y_1 & u_1 \\ Y_2 & u_2 \end{array}$$

$$(X^T X)_{jk} = \sum_{j=1}^n X_{ji} X_{jk} = \sum_{j=1}^{u_1} X_{ji} X_{jk} + \sum_{j=u_1+1}^{u_1+u_2} X_{ji} X_{jk} \quad \text{where } n = u_1 + u_2$$

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

$$\text{Similarly } X^T Y = X_1^T Y_1 + X_2^T Y_2.$$

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

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

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

and on central server

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

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

$$X X^T = \sum_{i=1}^p x_i x_i^T$$

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

$$O(n^2 P_m)$$

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

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

6. k_1, k_2, \dots kernels

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

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

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

$$= \alpha_1 \sum_i \beta_i k_1(x_i, x_i) \beta_i + \alpha_2 \sum_i \beta_i k_2(x_i, x_i) \beta_i \geq 0$$

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

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

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

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

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

as k_m are kernels \square

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

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

where $P_i P_i^T$ is a positive semi-definite matrix as $\alpha^T P_i P_i^T \alpha = \|P_i^T \alpha\|^2 \geq 0$
and $D_{ii} \geq 0$.

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

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

as $(P_i^{(1)} \circ P_j^{(2)}) (P_i^{(1)} \circ P_j^{(2)})^T = \tilde{P}_i \tilde{P}_j^T \geq 0$ (pos semi-definite)

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

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

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

$$|x^T x'|^2 \leq \|x\|_2^2 \|x'\|_2^2 < 1 \Rightarrow y = x^T x' \text{ s.t. } |y| < 1$$

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

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

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

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

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

Differentiate over α

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

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

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

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

10. Minimize

$$c(Y, X, f(x_1) + \mu_1, \dots, f(x_n) + \mu_n) + J(\|f\|_X^2)$$

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

Suppose $\hat{g} = \hat{\mu} + u + v$ where $u \in \text{span}\{k(\cdot, x_1), k(\cdot, x_2), \dots, k(\cdot, x_n)\}$ and $v \in V^\perp$.

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

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

$$\begin{aligned} \text{Also } J(\|\hat{f}\|_X^2) &= J(\|u+v\|_X^2) = J(\|u\|_X^2 + \|v\|_X^2) \quad \text{as } \langle u, v \rangle = 0 \\ &\geq J(\|u\|_X^2) \quad \text{equality iff } v=0 \text{ by monotonicity of } J. \end{aligned}$$

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

11. $\Phi = UDV^T$

$$K = \Phi \Phi^T = UDV^T V D U^T = U D^2 U^T = U D (U D)^T$$

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

Eigenvalue decomposition given

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

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

Identify $\tilde{PE} \sim UD$ with $\begin{cases} \text{zero filled columns added if } n > d \\ \text{linearly dependent cols removed if } n < d \end{cases}$

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

From $K = UD^2V^T$ 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 .

$$12. f^* \in \mathcal{X} \quad x_1, \dots, x_n \in \mathcal{X} \quad K_{ij} = k(x_i, x_j) / n$$

$$f^*(x_i) = \langle k(\cdot, x_i), f^* \rangle$$

~~$f^*(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i) + v$~~ $f^*(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i) + v \quad \alpha \in \mathbb{R}^n$

where $v \in V^\perp$, $V = \text{span}\{k(\cdot, x_i), i=1 \dots n\}$

Take inner product with $k(\cdot, x_j)$, $j \in \{1 \dots n\}$

$$f^*(x_j) = \langle k(\cdot, x_j), f^*(\cdot) \rangle = \langle k(\cdot, x_j), \sum_{i=1}^n \alpha_i k(\cdot, x_i) + v \rangle$$

$$= \langle k(\cdot, x_j), \sum_{i=1}^n \alpha_i k(\cdot, x_i) \rangle$$

$$= \sum_{i=1}^n \langle k(\cdot, x_j), k(\cdot, x_i) \rangle \alpha_i = \sum_{i=1}^n \underbrace{k(x_j, x_i)}_{K_{ji}} \alpha_i$$

Thus

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

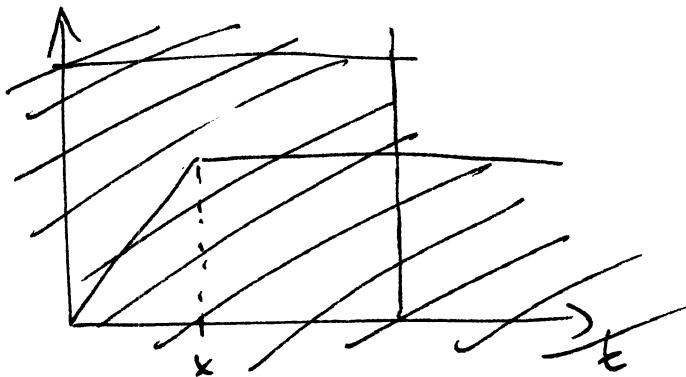
$$\|f^*\|_{\mathcal{X}}^2 = \left\| \sum_{i=1}^n \alpha_i k(\cdot, x_i) \right\|_{\mathcal{X}}^2 + \|v\|_{\mathcal{X}}^2 = \alpha^T K \alpha + \|v\|_{\mathcal{X}}^2 \geq \alpha^T K \alpha$$

13. $f: X \rightarrow \mathbb{R}$, $X = [0, 1]$

$$f_x(t) := \begin{cases} 1 & 0 \leq t \leq x \\ 0 & x < t \leq 1 \end{cases}$$

$$\langle f_x, f_{x'} \rangle := \int_0^1 f_x(t) f_{x'}(t) dt = \min(x, x') = k(x, x')$$

$$\langle k_x, f_{x'} \rangle = \int_0^1 \min(t, x) f_{x'}(t) dt$$



13. $x_1, \dots, x_n \in [0, 1]$, $\alpha \in \mathbb{R}^n$

$$\begin{aligned} \sum_{i,j} \alpha_i \min(x_i, x_j) \alpha_j &= \sum_{i,j} \alpha_i \alpha_j \int_0^1 \mathbb{1}_{[0, x_i]}(t) \mathbb{1}_{[0, x_j]}(t) dt \\ &= \int_0^1 \left(\sum_i \alpha_i \mathbb{1}_{[0, x_i]}(t) \right)^2 dt \geq 0 \end{aligned}$$