ESL Exercises*

Chapter 2

Ex. 2.1

This result holds for standard Euclidean norm. For example we could consider a metric space $(\mathbb{R}^3, \|\cdot\|)$ where $\|a_0e_0 + a_1e_1 + a_2e_2\| = 3a_0^2 + a_1^2 + a_2^2$ and $\hat{y} = (0.4, 0.3, 0.3)$.

- Task 1: largest element of $\hat{y}$ is the first element.
- Task 2: $\|\hat{y} - t_0\| = 1.26$, $\|\hat{y} - t_{1,2}\| = 1.06$. $\arg\min \|\hat{y} - t_i\| = 1, 2 \neq 0$.

For standard Euclidean norm, we have,

$$\|t_k - \hat{y}\| = \sum_{i \neq k} \hat{y}_i^2 + (1 - \hat{y}_k)^2 = 1 + TSS(\hat{y}) - 2y_k.$$

Thus we have,

$$\begin{aligned} & \arg\min \|t_k - \hat{y}\| \\ = \; & \arg\min(1 + TSS(\hat{y}) - 2y_k) \\ = \; & \arg\min(-2y_k) \\ = \; & \arg\max(\hat{y}). \end{aligned}$$

Ex. 2.2

By the caption of Figure 2.5, we know $\mathbb{P}(x|C_k)$ and $\mathbb{P}(C_k)$. The Bayes decision boundary of the 0 - 1 loss is determined by, $\mathbb{P}(C_0|x) = \mathbb{P}(C_1|x)$ or equivalently

$$\mathbb{P}(x|C_0)\mathbb{P}(C_0) = \mathbb{P}(C_1|x)\mathbb{P}(C_1).$$

Ex. 2.3

$X = \min_{dist}\{X_1, ..., X_N\} \in \mathbb{R}^1$, $1 - F_X(x) = \mathbb{P}(X > x) = (1 - x^p)^N$. Median of X is simply the x_0 where $F_X(x_0) = \frac{1}{2}$.

Ex. 2.4

For samples follows multivariate Gaussian, $r^2 = \sum_{i=1}^p x_i^2 \sim \chi_p^2$. Fix a direction a, projection $x \cdot a = \sum_{i=1}^p x_i a_i \sim \mathcal{N}(0, \sum_{i=1}^p a_i^2) = \mathcal{N}(0, 1)$. (one can easily prove this using characteristic functions) Thus after projection, the mean is closer to zero in the $\mathbb{L}^2$ sense.

*https://github.com/xincui-math

Ex. 2.5

$$
\begin{aligned}
EPE(x_0) &= \mathbb{E}_{\mathcal{T},x_0}(y_0 - \hat{y_0})^2 \\
&= \mathbb{E}_{\mathcal{T},x_0}(y_0 - \mathbb{E}_{\mathcal{T},x_0} y_0)^2 + \mathbb{E}_{\mathcal{T},x_0}(\mathbb{E}_{\mathcal{T},x_0} y_0 - \hat{y_0})^2 \\
&= \mathbb{E}_{\mathcal{T},x_0}\epsilon^2 + \mathbb{E}_{\mathcal{T},x_0}(x_0\beta - x_0\hat{\beta})^2 \\
&= \mathbb{E}\epsilon^2 + x_0\mathbb{E}_{\mathcal{T}}(\beta - \hat{\beta})^2 \\
&= \mathbb{E}\epsilon^2 + x_0\mathbb{E}_{\mathcal{T}}\left(\beta - (X^TX)^{-1}X^T(X\beta+\epsilon)\right)^2 \\
&= \mathbb{E}\epsilon^2 + x_0\mathrm{Var}\left((X^TX)^{-1}X^T\epsilon\right) \\
&= \sigma^2 + x_0(X^TX)^{-1}.
\end{aligned}
$$

Ex. 2.6

Decompose sample space $\Omega = \bigoplus_{x_i} \Omega_i = \bigoplus_i \{(x_i, y_{ij})\}$.

$$
\begin{aligned}
SSR(\Omega_i) &= \sum_j [y_{ij} - f_\theta(x_i)]^2 \\
&= \sum_j y_{ij}^2 - 2\sum_j y_{ij} f_\theta(x_i) + n_i f_\theta(x_i)^2 \\
&= n_i \left(f_\theta(x_i) - \frac{1}{n_i}\sum_j y_{ij} \right)^2 + \phi(y_{ij}).
\end{aligned}
$$

Hence the problem reduces to weighted least square weights n_i.

Ex. 2.7

(a) Representations

Linear regression:

$$
\begin{aligned}
\hat{f}(x_0) &= x_0\hat{\beta} \\
&= x_0(X^TX)^{-1}X^Ty \\
&= \sum_i [x_0(X^TX)^{-1}X^T]_i y_i.
\end{aligned}
$$

K-nearest neighbourhood:

$$
\hat{f}(x_0) = \sum_i \frac{1}{k} I_{i \in argmin_k \overrightarrow{d}(x_0, \mathcal{X})} y_i.
$$

(b) $\mathbb{E}_{\mathcal{Y}|\mathcal{X}}(MSE)$

$$\begin{aligned}
& \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[f(x_0)-\hat{f}(x_0)\right]^2 \\
= \; & \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[f(x_0)-\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0)+\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0)-\hat{f}(x_0)\right]^2 \\
= \; & \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[f(x_0)-\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0)\right]^2+\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)-\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0)\right]^2 \\
& -2\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[f(x_0)-\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0)\right]\left[\hat{f}(x_0)-\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0)\right] \\
= \; & \left[f(x_0)-\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0)\right]^2+\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)-\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0)\right]^2 \\
= \; & \left[f(x_0)-\sum_{i=1}^{N} l_i(x_0,\mathcal{X})f(x_i)\right]^2+\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\sum_{i=1}^{N} l_i(x_0,\mathcal{X})\epsilon_i\right]^2 \\
= \; & \left[f(x_0)-\sum_{i=1}^{N} l_i(x_0,\mathcal{X})f(x_i)\right]^2+\sum_{i=1}^{N} l_i^2(x_0,\mathcal{X})\sigma^2.
\end{aligned}$$

(c) $\mathbb{E}_{\mathcal{Y},\mathcal{X}}(MSE)$

Notice that $\hat{f}(x_0)$ is $(\mathcal{Y},\mathcal{X})$ measurable, $\mathbb{E}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0)=\hat{f}(x_0)$.

$$\begin{aligned}
& \mathbb{E}_{\mathcal{Y},\mathcal{X}}\left[f(x_0)-\hat{f}(x_0)\right]^2 \\
= \; & \left[f(x_0)-\mathbb{E}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0)\right]^2+\mathbb{E}_{\mathcal{Y},\mathcal{X}}\left[\hat{f}(x_0)-\mathbb{E}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0)\right]^2 \\
= \; & \left[f(x_0)-\hat{f}(x_0)\right]^2.
\end{aligned}$$

Ex. 2.8

TODO: code up.

Ex. 2.9

Write train set as (X_0, y_0), test set as (X_1, y_1), projection matrix as P_i. Rewrite $\hat{\beta}=(X_0^T X_0)^{-1}X_0^T\epsilon_0+\beta$. On train set, we have the following,

$$\begin{aligned}
& \mathbb{E}_0\left[(y_0-X_0\hat{\beta})^T(y_0-X_0\hat{\beta})\right] \\
= \; & \mathbb{E}_0\left[(\epsilon_0-P_0\epsilon)^T(\epsilon_0-P_0\epsilon_0)\right] \\
= \; & (N-k)\sigma^2.
\end{aligned}$$

On test set, we have,

$$
\begin{aligned}
& \mathbb{E}_1\left[(y_1 - X_1\hat{\beta})^T(y_1 - X_1\hat{\beta})\right] \\
= \; & \mathbb{E}_1\left[(\epsilon_1 - X_1(X_0^TX_0)^{-1}X_0^T\epsilon_0)^T(\epsilon_1 - X_1(X_0^TX_0)^{-1}X_0^T\epsilon_0)\right] \\
= \; & N\sigma^2 + trace\left[X_0(X_0^TX_0)^{-1}X_1^TX_1(X_0^TX_0)^{-1}X_0^T\right]\sigma^2 \\
= \; & N\sigma^2 + trace\left[X_1(X_0^TX_0)^{-1}X_1^T\right]\sigma^2 \\
\geq \; & N\sigma^2.
\end{aligned}
$$

Chapter 3

Ex. 3.1

For simplicity, let's denote variable with tilde as skipping skipping column i. Here we explore bit more in this problem to clarify how to compute F-score.

- RSS_0: $(y - X\beta)^T(y - X\beta)$
- RSS_1: $(y - \tilde{X}\tilde{\beta})^T(y - \tilde{X}\tilde{\beta})$
- rss_1: $(y - X\beta + X_i\beta_i)^T(y - X\beta + X_i\beta_i)$

or say, RSS_1 uses refit $\tilde{\beta}$, rss_1 uses original β.

$$rss_1 - RSS_0 = X_i^TX_i\beta_i^2 = (X^TX)_{ii}\beta_i^2.$$

Denote

- $span\tilde{X} = span\{X_k | k \neq i\}$
- $spanX = span\{X_k\}$
- P_A, projection matrix to $spanA$.
- $\tilde{X}^{\perp}$ is the orthogonal complement of $span\{\tilde{X}\}$ inside $span\{X\}$.

$$RSS_1 - RSS_0 = y^TP_{\tilde{X}}y - y^TP_Xy = y^TP_{\tilde{X}^{\perp}}y.$$

$$F_i = \frac{(RSS_1 - RSS_0)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)} = \frac{y^TP_{\tilde{X}^{\perp}}y}{\hat{\sigma}^2} = \frac{x_i^TP_{\tilde{X}^{\perp}}x_i}{\hat{\sigma}^2}\hat{\beta}i^2 = \frac{\|X_i^{\perp}\|^2}{\hat{\sigma}^2}\hat{\beta}i^2.$$

$$z_i^2 = \frac{\hat{\beta}_i^2}{\hat{\sigma}^2(X^TX)_{ii}^{-1}}.$$

Now let's compute $(X^TX)^{-1}_{ij}$. For arbitrary $\epsilon \sim \mathcal{N}(0, I_N)$,

$$\begin{aligned} & (X^TX)^{-1}_{ij} \\ = & \; COV\left((X^TX)^{-1}X^T\epsilon, (X^TX)^{-1}X^T\epsilon\right) \\ = & \; \mathbb{E}(\beta_{\epsilon,X,i}\beta_{\epsilon,X,j}) \\ = & \; \frac{X_i^{\perp} \cdot X_j^{\perp}}{\|X_i^{\perp}\|^2\|X_j^{\perp}\|^2} \end{aligned}$$

Thus we have $z_i^2 = \frac{\hat{\beta_i}^2}{\hat{\sigma}^2(X^TX)^{-1}_{ii}} = \frac{\hat{\beta_i}^2\|X_i^{\perp}\|^2}{\hat{\sigma}^2} = F_i$. And $f_i = \frac{\hat{\beta_i}^2(X^tX)_{ii}}{\hat{\sigma}^2}$

Ex. 3.2

- $\mathrm{Var}(a^T\beta) = a^TCOV_\beta a = (X^TX)^{-1}\sigma^2$.
- $C_\beta = \{(\hat{\beta} - \beta)^T(X^TX)(\hat{\beta} - \beta) \leq \hat{\sigma}^2 {\chi^2_{p+1}}^{(1-\alpha)}\}$

The first estimation needs σ, the second one doesn't. (code TBD)

Ex. 3.3

For quantity $a^T\beta$, we have unbias estimator $\hat{\theta} = a^T\hat{\beta} = a^T(X^TX)^{-1}X^Ty$. Giving another unbiased estimation c^Ty, write $c^T = a^T(X^TX)^{-1}X^T + D$.

$$\mathbb{E}(\left[a^T(X^TX)^{-1}X^T + D\right]y) = a^T\beta + DX\beta.$$

Hence we have $DX = 0$.

$$\begin{aligned} & \mathrm{Var}(c^Ty) \\ = & \; \left[a^T(X^TX)^{-1}X^T + D\right]\left[a^T(X^TX)^{-1}X^T + D\right]^T\sigma^2 \\ = & \; [a^T(X^TX)^{-1}a + DD^T]\sigma^2 \\ \succeq & \; a^T(X^TX)^{-1}a\sigma^2 \\ = & \; \mathrm{Var}(a^T\hat{\beta}). \end{aligned}$$

Ex. 3.4

Give $X = (X_0, X_1, ..., X_{n-1})$ and y.

- $\tilde{\beta}_0 = \frac{\mathrm{cov}(X_0,y)}{\mathrm{cov}(X_0,X_0)}$
- $\tilde{\beta}_i = \frac{\mathrm{cov}(z_i,y)}{\mathrm{cov}(z_i,z_i)}$ with $z_i = x_i - \sum_{j=0}^{i-1}\gamma_{ij}x_j$, $\gamma_{ij} = \frac{\mathrm{cov}(x_i,x_j)}{\mathrm{cov}(x_j,x_j)}$.

$$y = \sum_{i=0}^{n}\tilde{\beta}_iz_i = \sum_{i=1}^{n}\tilde{\beta}_iz_i = \sum_{i=1}^{n}\tilde{\beta}_i(x_i - \sum_{j=1}^{i-1}\gamma_{ij}x_j) = \sum_{i=1}^{n}(\tilde{\beta}_i - \sum_{j=i+1}^{n}\Gamma_{ji})x_i.$$

Ex. 3.5

- $\beta_0^c = \beta_0 + \sum_{j=0}^{p} \bar{x_j}\beta_j$
- $\beta_i^c = \beta_i$

Ex. 3.6

Assume prior $\beta \sim N(0, \tau I)$, data samples from $y \sim N(X\beta, \sigma^2 I)$. Posterior distribution has PDF proportional to:

$$p(\beta|D) \sim exp\left[-\frac{(y-X\beta)^T(y-X\beta)}{2\sigma^2} - \frac{\beta^T\beta}{2\tau^2}\right]$$

Q1: Equivalent to mode: $\lambda = \frac{\sigma^2}{\tau^2}$.

Q2: Equivalent to posterior mean:

$$-\frac{1}{2}(\beta - m_1)^T m_2^{-1}(\beta - m_1) = -\frac{(y-X\beta)^T(y-X\beta)}{2\sigma^2} - \frac{\beta^T\beta}{2\tau^2}$$

Solves the system, we have,

$$\begin{cases} m_1 = (\frac{X^TX}{\sigma^2} + \frac{I}{\tau^2}) \\ m_2 = (X^TX + \frac{\sigma^2}{\tau^2}X^Ty). \end{cases}$$

Ex. 3.7

Direct consequence of $p(\beta|D) \sim exp\left[-\frac{(y-X\beta)^T(y-X\beta)}{2\sigma^2} - \frac{\beta^T\beta}{2\tau^2}\right]$.

Ex. 3.8

For special matrix, $X = (e, x_1, x_2, x_3, ..., x_p)$, centered matrix is given by

$$\widetilde{X} = (x_1 - \overline{x_1}, ..., x_p - \overline{x_p}).$$

Gram-Schmidt processes gives the following. $x_i = \sum_{d=1}^{i-1} q_d r_{di} + \frac{\langle e, x_i \rangle}{|e|}\frac{e}{|e|}$.

$$Q_{lower}R_{lower} = \widetilde{X} = U\Sigma V^*.$$

Hence column span of Q_{lower} is the same as column span of U.

Denote $\tilde{R} = \Sigma^{-1}R = V^*$. Consider r_i be i-th column vector of $\tilde{R}$. Then use induction not hard to see $\tilde{R}$ is diagonal. This implies ΣV is diagonal of 1/-1.

Ex. 3.9

Denote $z_i = x_i - \sum_{k=1}^{r} q_k(x_i, q_k)$, variance explained increment has norm,

$$\|\hat{\beta}_i z_i\| = |\langle y, x_i - \sum_{k=1}^{r} q_k(x_i, q_k) \rangle|.$$

For new set of feature it is equivalent (and faster) to pick the following,

$$\text{argmax}_i \|(y^TX - y^TQQ^TX)_i\|.$$

Ex. 3.10

Exercise 3.1 shows F statistics for dropping i-th variable corresponds to z_i^2. Hence we just need to drop the variable with lowest $|z|$.

Ex. 3.11

$$
\begin{aligned}
& \left(\frac{dtr[(Y-XB)^T(Y-XB)]}{dB}\right)_{ij} \\
= \quad & \frac{d}{db_{ij}} \sum_{p,q,v} (y_{pq} - X_{pv}B_{vq})^2 \\
= \quad & -\sum_{p,q,v} 2(y_{pq} - X_{pv}B_{vq}) X_{ps} \delta_{sq}^{ij} \\
= \quad & -\sum_{p,v} 2(y_{pj} - X_{pv}B_{vj}) X_{ps} \delta_s^i \\
= \quad & -\sum_{p,v} 2(y_{pj} - X_{pv}B_{vj}) X_{pi} \\
= \quad & -2(X^TY - X^TXB)_{ij}.
\end{aligned}
$$

Consider symmetric square root $\Sigma^{-\frac{1}{2}}$, the solution is

$$X^TY\Sigma^{-\frac{1}{2}} - X^TXB\Sigma^{-\frac{1}{2}} = 0.$$

Ex. 3.12

$$\begin{pmatrix} y \\ 0 \end{pmatrix} = \begin{pmatrix} \widetilde{X} \\ \sqrt{\lambda} I \end{pmatrix} \beta + \epsilon.$$

$RSS(\beta) = (y - \tilde{X}\beta)^T(y - \tilde{X}\beta) + \lambda\|\beta\|^2$, same as Ridge regression.

Ex. 3.13

Now consider $z_i = Xv_i = \lambda_i u_i$.

- $\langle z_i, y\rangle = \lambda_i \langle u_i, y\rangle$.
- $\langle z_i, z_i\rangle = \lambda_i^2$.

$$\hat{\beta}^{pcr}(p) = \sum_{i=1}^{p} \frac{\langle z_i, y\rangle}{\langle z_i, z_i\rangle} v_i = VD^{-1}U^Ty = \hat{\beta}^{ls}.$$

Ex. 3.14

- $z_1 = \sum_{j=1}^p \langle x_j, y\rangle x_j$.
- $x_j^1 = x_j^0 - \widehat{\phi_{1j}} \frac{z_1}{\langle z_1, z_1\rangle}$

- $\langle x_j^1, y\rangle = \langle x_j^0 - \widehat{\phi_{1j}} \frac{z_1}{\langle z_1, z_1\rangle}, y\rangle = \widehat{\phi_{1j}} - \widehat{\phi_{1j}} \frac{\langle z_1, y\rangle}{\langle z_1, z_1\rangle}$.

Using x_i are orthogonal, we have,

$$\langle z_1, y\rangle = \langle z_1, z_1\rangle = \sum_{j=1}^{p} \widehat{\phi_{1j}}^2.$$

This implies $\widehat{\phi_{2j}} = \langle x_j^1, y\rangle = 0$.

Ex. 3.15

(PLS)

$$\max_{\alpha} \operatorname{corr}^2(y, X\alpha) \operatorname{Var}(X\alpha)$$
$$\text{subject to } \|\alpha\| = 1, \alpha^T S\phi_l = 0.$$

The problem is equivalent to the following.

$$\max_{\alpha} \operatorname{cov}(y, X\alpha)$$
$$\text{subject to } \|\alpha\| = 1, \alpha^T S\phi_l = 0.$$

Decompose $\operatorname{span}(X) = X_{proj} \oplus X_{ortho}$ using inner product. Restrict α to $(X\alpha)_{proj} = 0$, Lagrange multiplier gives,

$$\begin{aligned} & L(x, \lambda) \\ = \ & \operatorname{cov}(y, X\alpha) - \lambda\alpha^T\alpha \\ = \ & \sum_i \left[\langle y, x_{iortho}\rangle\alpha_i - \lambda\alpha_i^2\right] + \langle y, (X\alpha)_{proj}\rangle. \end{aligned}$$

This implies $\alpha_i \sim \langle y, x_{iortho}\rangle$.

Ex. 3.16

Consider $y = \sum_i \alpha_i x_i + \epsilon$, where $\langle x_i, x_j\rangle = \delta_{ij}$.

Notice for any S, we always have estimated $\widehat{\beta_i^{(S)}} = \widehat{\beta}_i = \langle y, x_i\rangle$.

(1) Best M-subset

$$\begin{aligned} SSR(S) &= \left\|\sum_{i\in S} \widehat{\beta}_i x_i\right\|^2 \\ &= \sum_{i\in S} \widehat{\beta}_i^2. \end{aligned}$$

It is equivalent to pick the largest M $\widehat{\beta}_i$ in full regression.

(2) Ridge.

$$\begin{aligned} & \min\{\|y - X\beta\|^2 + \lambda\|\beta\|^2\} \\ = \ & \min_{\beta} \sum_i [(\lambda+1)\beta_i^2 - 2\langle x_i, y\rangle\beta_i] \end{aligned}$$

Hence $\beta_i^{Ridge} = \frac{\widehat{\beta}_i}{\lambda+1}$.
(3) Lasso.

$$\begin{aligned} & \min\{\|y - X\beta\|^2 + 2\lambda\|\beta\|\} \\ = \quad & \min_{\beta} \sum_i \left[\beta_i^2 - 2[\hat{\beta}_i - \lambda \text{sign}(\beta_i)]\beta_i\right] \end{aligned}$$

(I) $\beta_i > 0$

- $\hat{\beta}_i \geq \lambda$: $\beta_i = \hat{\beta}_i - \lambda$
- $\hat{\beta}_i < \lambda$: $\beta_i = 0$

(II) $\beta_i < 0$

- $\hat{\beta}_i \geq -\lambda$: $\beta_i = 0$
- $\hat{\beta}_i < \lambda$: $\beta_i = \hat{\beta}_i + \lambda$

Rephrase the above analysis gives $\beta_j^{Lasso} = \text{sign}(\hat{\beta}_j)(\hat{\beta}_j - \lambda)_+$.

Ex. 3.17

Code TBD.

Ex. 3.18

Solving β is equivalent represent y_{proj} in coordinate X. The PLS solves a set of orthonormal basis iteratively for space span$\{X\}$, namely z_m. Under z_m,

$$\beta_{z_m} = \frac{\langle z_m, y \rangle}{\langle z_m, z_m \rangle}_I .$$

This formula matches the conjugate gradients algorithm.

Ex. 3.19

(1) L^2 norm as a decreasing function of λ in Ridge regression.

$$\begin{aligned} & \frac{d}{d\lambda}\|\beta^{ridge}\|^2 \\ = \quad & \frac{d}{d\lambda}[y^T X(X^T X + \lambda I)^{-1}(X^T X + \lambda I)^{-1} X^T y] \\ = \quad & -2\lambda[y^T X(X^T X + \lambda I)^{-1}(X^T X + \lambda I)^{-1}(X^T X + \lambda I)^{-1} X^T y] \\ = \quad & -2\lambda\beta_\lambda^T (X^T X + \lambda I)^{-1}\beta_\lambda \\ \leq \quad & 0 \end{aligned}$$

(2) L^1 norm may not be a decreasing function of λ in Lasso regression.

Ex. 3.20

(CCR problem) Follow the exact same approach in [3]. $c = \Sigma_{YY}^{1/2} u$, $d = \Sigma_{XX}^{1/2} v$. Cauchy Schwarz gives,

$$u^T Y^T X v \leq \frac{\|\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}c\|}{\|c\|}.$$

Equality holds when $d = \lambda \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}c$. Perform SVD decomposition on $\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2} = UDV^T$, and consider $\tilde{c} = V^T c$.

$$\frac{\|\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}c\|}{\|c\|} = \frac{\|\tilde{c}^T \Sigma \tilde{c}\|}{\|\tilde{c}\|}.$$

Optimization result $\tilde{c}$ gives identity matrix. Hence c gives V, right singular vectors, equality condition gives d are left singular vectors of $\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}$. This gives same conclusion (after transpose).

$$\begin{aligned} u_1 &= \Sigma_{YY}^{-1/2} u_1^* \\ v_1 &= \Sigma_{XX}^{-1/2} v_1^* \end{aligned}$$

Ex. 3.21

$$\begin{aligned} & tr\left[(y - XB)\Sigma^{-1}(y - XB)^T\right] \\ = \quad & tr\left[(y^* - XB\Sigma^{-1/2})^T(y^* - XB\Sigma^{-1/2})\right] \\ = \quad & tr\left[(Z - A)(Z - A)^T\right] + \text{const}(B) \end{aligned}$$

where $Z = \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}$, $A = \Sigma_{YY}^{-1/2} B^T \Sigma_{XX}^{1/2}$. Apply Theorem 3.7.4 in [4],

$$A_{opt} = \sum_{j=1}^{m} d_j U_j V_j^T.$$

where U, V are singular vectors of Z. Hence $B = \Sigma_{XX}^{-1}\Sigma_{YX}\sum_i u_{ccr,i} u_{ccr,i}^T$, where $u_{ccr,i} = \Sigma_{YY}^{-1/2} u_i$. U is orthgonal gives $U_m^T \Sigma_{YY}^{1/2}$ is generalized inverse of $\Sigma_{YY}^{-1/2} U_m$.

Ex. 3.22

Replace Σ_{YY} to $\Sigma_{residual}$.

Ex. 3.23

(a)

$$\begin{aligned} & \|\langle x_j, y - u(\alpha)\rangle\| \\ = \quad & \|\langle x_j, y - \alpha X(X^T X)^{-1} X^T y\rangle\| \\ = \quad & \|(1 - \alpha)(X^T y)_j\| \\ = \quad & N\lambda|1 - \alpha|. \end{aligned}$$

(b) $(y - \alpha X\hat{\beta})^T(y - \alpha X\hat{\beta}) = N + \alpha(\alpha - 2)y^T X\hat{\beta}$. Let $\alpha = 1$, we have, $y^T X\hat{\beta} = N - RSS$. Hence $(y - \alpha X\hat{\beta})^T(y - \alpha X\hat{\beta}) = N(1-\alpha)^2 + \alpha(2-\alpha)RSS$.

$$\mathrm{corr}(x_i, y - u(\alpha)) = \frac{\langle x_i, y - u(\alpha)\rangle}{\|x_i\|\|y - u(\alpha)\|} = \frac{N\lambda|1-\alpha|}{\sqrt{N}\sqrt{N(1-\alpha)^2 + \alpha(2-\alpha)RSS}}.$$

(c) $(X_{\mathcal{A}_k}^T X_{\mathcal{A}_k})^{-1} X_{\mathcal{A}_k}^T r_k$ is OLS of r_k with $X_{\mathcal{A}_k}$. Notice that X has mean 0, std 1, $X_{\mathcal{A}_k}$ has same correlations with r_k. Hence (2) directly gives the result.

Ex. 3.24

$$\cos(x_i, X\hat{\beta}) = \frac{(X^T X\hat{\beta})_i}{\|x_i\|\|X\hat{\beta}\|} = \frac{(X_{\mathcal{A}_k}^T r_k)_i}{\sqrt{N}\|X\hat{\beta}\|}.$$

$(X_{\mathcal{A}_k}^T r_k)_i$ is constant among i from LAR algorithm.

References

[1] Hastie, T., Tibshirani, R.,, Friedman, J. (2001). The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc..

[2] John L. Weatherwax, David Epstein A Solution Manual and Notes for: The Elements of Statistical Learning by Jerome Friedman, Trevor Hastie, and Robert Tibshirani, https://waxworksmath.com/Authors/G_M/Hastie/WriteUp/Weatherwax_Epstein_Hastie_Solution_Manual.pdf.

[3] Canonical correlation. Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 30 Nov 2020. Web. 15 Feb. 2021, https://en.wikipedia.org/wiki/Canonical_correlation.

[4] Brillinger, David R. Time series: data analysis and theory. Society for Industrial and Applied Mathematics, 2001.