CS 5233: FDS - DATA MODELING

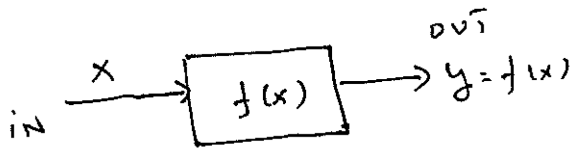
STATISTICAL LEAST SQUARES

S. Lakshmivarahan

April 30, 2022

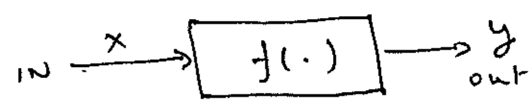
1 A classification of problems:

Forward



- f(x) is known
- x is given
- compute y

Inverse



- x and y are given
- Identify f(·)
- x - is x-ray, y - x-ray picture
- Identify f(·)

a) simple pendulum: $T = 2\pi\sqrt{l/g}$

- If l and g are given, Compute T - Forward
- If (l_i, T_i), $1 \le i \le m$ given, estimate g - Inverse

b) Economics:

$$\underbrace{\text{change in unemployment}}_{\text{observe it.}} = \underbrace{w_0 + w_1}_{\text{want to discover the relation}} * \underbrace{\text{change in GDP}}_{=\text{output}} + \text{noise}$$

~~want~~ observe it.

want to discover the relation

= output ↑ we can observe it

(c) Satellite problem: Thermal energy radiated is proportional to the 4^{th} power of the temperature of the radiating surface: $E = \alpha T^4$

- knowing E as measured by the satellite and α, estimate T.

(2) Statistical/ Least squares method is one of the methods for modeling that is used in solving inverse problems.

- $x \in \mathbb{R}^n \longrightarrow \boxed{f(\cdot)} \longrightarrow y \in \mathbb{R}$

$f(\cdot)$: Linear / Non-linear

W-parameters (a) Example 1: $y = x^T w + v$ $\quad v \sim$ Noise $\sim N(0, \sigma^2)$ — Known

- Here y and x and properties of noise v are known. Find / Estimate w.
- Here $f(x) = x^T w$ is linear in x and w.

(b) Example 2: $y = a e^{bx} + v$

$x \in \mathbb{R} \longrightarrow \boxed{f(\cdot)} \longrightarrow y \in \mathbb{R}$

- Here $f(x) = a e^{bx}$. It is nonlinear in the unknown (a, b) and also nonlinear in x

(c) Example 3: $y = a_0 + a_1 x + a_2 x^2 + v$

$x \longrightarrow \boxed{f(\cdot)} \longrightarrow y$

- $f(x) = a_0 + a_1 x + a_2 x^2$. Nonlinear in x and linear in parameters (a_0, a_1, a_2).

(3) <u>linear least squares problem</u> :- (statistical)

- let
$$y_i = w_0 + x_{i1} w_1 + x_{i2} w_2 \cdots + x_{i,n-1} w_{n-1} + v_i \longrightarrow (1)$$

for $1 \le i \le m$, where assume $m > n$.

- Define $y = (y_1, y_2, \cdots y_m)^T \in R^m$

$$w = (w_0, w_1 \cdots w_{n-1})^T \in R^n$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1\,n-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2\,n-1} \\ \vdots & & & & \\ 1 & x_{m1}, & x_{m2}, & \cdots & x_{m,n-1} \end{bmatrix} \in R^{m \times n}$$

$$v = [v_1, v_2, \cdots v_m]^T \in R^m$$

- (1) becomes in <u>matrix-vector notation</u>:

$$y = X w + v \longrightarrow (2)$$

where $v \sim N(0, R)$, $\boxed{\sigma^2 \text{ is known}}$, $R = \sigma^2 I_m$

- <u>Define residual</u>

$$r(w) = y - X w \longrightarrow (2)$$

(4) - <u>Cost function</u>: $J : R^n \to R$

$$J(w) = \frac{1}{2} (y - X w)^T R^{-1} (y - X w)$$

is the <u>weighted sum of squared residuals.</u>

- Goal is to <u>minimize</u> $J(w)$ w. r. to w.

- This is a <u>multivariate minimization problem</u>

- <u>Rewrite $J(w)$ by multiplying out:</u>

$$J(w) = \frac{1}{2}\left[y^T R^{-1} y - y^T R^{-1} X w - (Xw)^T R^{-1} y + (Xw)^T R^{-1} (Xw) \right] \rightarrow (3)$$

- Recall: $(Xw)^T = w^T X^T$

$$\therefore (Xw)^T R^{-1} y = w^T X^T R^{-1} y \rightarrow (4)$$

- Recall, if a is a scalar: $a^T = a$.

- Since $y^T R^{-1} X w$ is scalar:

$$y^T R^{-1} X w = (y^T R^{-1} X w)^T = w^T X^T R^{-1} y \rightarrow (5)$$

where since R is symmetric, so is R^{-1} and $(R^{-1})^T = R^{-1}$.

- using (4) and (5) in (3):

$$J(w) = \frac{1}{2}\Big[\underbrace{y^T R^{-1} y}_{\text{Independent of } w} - \underbrace{2\, y^T R^{-1} X w}_{\text{Linear in } w} + \underbrace{w^T (X^T R^{-1} X) w}_{\text{Quadratic in } w} \Big] \rightarrow (6)$$

- <u>Recall:</u> $\boxed{A = A^T}$

$$f(w) = a^T w \Rightarrow \nabla f(w) = a \quad (a \in \mathbb{R}^n)$$
$$f(w) = w^T A w \Rightarrow \nabla f(w) = 2Aw \quad (7)$$

- <u>Consider the linear term:</u> Set $a^T = y^T R^{-1} X$

$$\underbrace{(y^T R^{-1} X)}\, w = a^T w, \quad \nabla[\;\;] = a \Rightarrow \nabla\left[(y^T R^{-1} w) X\right] = X^T R^{-1} y \rightarrow (8)$$

- <u>Consider the quadratic term:</u> $A = X^T R^{-1} X$

Verify $A^T = A$

$$\nabla [x^T A x] = 2Ax = 2(x^T R^{-1} x) w \rightarrow (9)$$

- <u>Combining</u> (7) to (9) with (6) and simplify:

$$\nabla J(w) = \frac{1}{2}\left[-2x^T R^{-1} y + 2x^T R^{-1} x\, w\right]$$

$$= -x^T R^{-1}[y - xw] \longrightarrow (10)$$

- <u>Hessian of $J(w)$</u> :-

$$\nabla^2 J(w) = (x^T R^{-1} x) \in \mathbb{R}^{n \times n} \text{ - SPD}$$

- <u>Equate the gradient to zero</u>:

$$\nabla J(w) = 0 \implies (x^T R^{-1} x) w = x^T R^{-1} y \rightarrow (11)$$

- If $R = \sigma^2 I_m$, $R^{-1} = \frac{1}{\sigma^2} I_m$ and (11)

becomes $\boxed{(x^T x) w = x^T y} \longrightarrow (12)$

- (12) is called <u>Normal equation</u>. Assume $X \in \mathbb{R}^{m \times n}$ is a full rank matrix: $\text{Rank}(x) = n$. This guarantees $x^T x$ is SPD:

$a^T (x^T x) x = (xa)^T (xa) = \|xa\|_2^2 \geq 0$ with equality only if $a = 0$ since the columns of x are linearly independent.

- The least square solution: $\hat{W}_{LS}$:

Solving (12):

$$\boxed{\hat{W}_{LS} = (X^T X)^{-1} X^T y} \longrightarrow (13)$$

$\left.\begin{array}{l}\text{use} \cdot \text{cholesky} \\ \cdot \text{QR decomp.}\end{array}\right\}$ To Solve (13)

4 Unbiasedness: $y = XW + V$

$$\therefore \hat{W}_{LS} = (X^T X)^{-1} X^T [XW + V]$$

$$= (X^T X)^{-1} (X^T X) W + (X^T X)^{-1} X^T V$$

$$= W + (X^T X)^{-1} X^T V \longrightarrow (14)$$

$$\therefore E[\hat{W}_{LS}] = W + (X^T X)^{-1} X^T E(V) = W \longrightarrow (15)$$

$\therefore \hat{W}_{LS}$ is unbiased.

5 Covariance of $\hat{W}_{LS}$:- $\boxed{\text{cov}(\hat{W}_{LS}) = \sigma^2 (X^T X)^{-1}}$

$$\text{cov}(\hat{W}_{LS}) = E\left(\left[\hat{W}_{LS} - E(\hat{W}_{LS})\right]\left[\hat{W}_{LS} - E(\hat{W}_{LS})\right]^T\right) \longrightarrow (16)$$

from (14): $\hat{W}_{LS} - E(\hat{W}_{LS})$

$$= W + (X^T X)^{-1} X^T V - W = (X^T X)^{-1} X^T V \longrightarrow (17)$$

Substituting (17) in (16)

$$\text{cov}(\hat{W}_{LS}) = E\left[\left((X^T X)^{-1} X^T V\right)\left((X^T X)^{-1} X^T V\right)^T\right]$$

$$= E\left[(X^T X)^{-1} X^T (V V^T) X (X^T X)^{-1}\right]$$

$$= (X^T X)^{-1} X^T \underbrace{\left[E(V V^T)\right]}_{= R = \sigma^2 I_m} X (X^T X)^{-1}$$

$$= \sigma^2 (X^T X)^{-1} (X^T X) (X^T X)^{-1} = \underline{\sigma^2 (X^T X)^{-1}} \longrightarrow (18)$$

- Recall $X \in \mathbb{R}^{m \times n}$. Do an SVD of X

$$X = U \Lambda^{1/2} V^T, \qquad \left.\begin{array}{l} (X^T X) V = V \Lambda \\ (X X^T) U = U \Lambda \end{array}\right\} \longrightarrow (19)$$

$$\boxed{(X^T X) = V \Lambda V^T}$$

- What is the total sum of the variance of the components of $\hat{W}_{LS}$? This is the $\mathrm{tr}[\mathrm{cov}(\hat{W}_{LS})]$

$$\begin{aligned}
\mathrm{tr}[\mathrm{cov}(\hat{W}_{LS})] &= \mathrm{tr}[\sigma^2 (X^T X)^{-1}] \\
&= \sigma^2 \, \mathrm{tr}[(X^T X)^{-1}] \\
&= \sigma^2 \, \mathrm{tr}([V \Lambda V^T]^{-1}) \\
&= \sigma^2 \, \mathrm{tr}[V \Lambda^{-1} V^T] \\
&= \sigma^2 \, \mathrm{tr}[V^T V \Lambda^{-1}] \\
&= \sigma^2 \, \mathrm{tr}[\Lambda^{-1}] \\
&= \sigma^2 \sum_{i=1}^{n} \frac{1}{\lambda_i} \qquad \longrightarrow (20)
\end{aligned}$$

$$\boxed{\begin{aligned} &\mathrm{tr}(ABC) \\ &= \mathrm{tr}(CAB) \\ &= \mathrm{tr}[BCA] \end{aligned}} \qquad \boxed{V^T = V^{-1}} \qquad \boxed{V^T V = I}$$

$$\boxed{\underset{\max}{\lambda_1} \geq \lambda_2 \geq \cdots \geq \underset{\min}{\lambda_n} > 0}$$

- Consequences:- The smallest eigenvalue of $(X^T X)$ determine the behavior of the total variance of the components of $\hat{W}_{LS}$. Thus, if the columns of X inherit collinearity property, then λ_n may be positive but very small and $\frac{1}{\lambda_n}$ can be very large.

Question: How tame the estimate $\hat{W}_{LS}$ in this case?

6) Estimation of σ^2:

- We assumed σ^2 is known. If not, can we estimate σ^2?

- Recall $\hat{w}_{LS} = (X^T X)^{-1} X^T y = X^+ y \quad \rightarrow (21)$

  where $\boxed{X^+ = (X^T X)^{-1} X^T}$ is called generalized inverse of X.

- Define a new matrix

$$P = X X^+ = X (X^T X)^{-1} X^T$$

- Verify $P^T = P$ and $P^2 = P. \quad \longrightarrow (22)$

~~Consider the residual at $\hat{w}_{LS}$~~

~~Recall: $\hat{y} = X \hat{w}_{LS}$, $y = X w$~~

- Define fitted value of y given by $\hat{y}$:

$$\hat{y} = X \hat{W}_{LS} = \underline{X (X^T X)^{-1} X^T} y = P y \quad \rightarrow (23)$$

- Define error e:

$$e = y - \hat{y} = y - P y = (I - P) y \quad \rightarrow (24)$$

- Verify $(I - P) X = X - P X = X - X (X^T X)^{-1} (X^T X) = X - X = 0$

$$\cancel{E(e) = E[(I-P)y] = E[(I-P)(Xw + v)]}$$

$$\therefore \quad e = (I-P)y = (I-P)(\cancel{X}w + v)$$

$$= [\underbrace{(I-P)X}_{=0}]w + (I-P)v$$

$$= (I-P)v.$$

$$\therefore \quad E(e) = (I-P)E(v) = 0 \implies \boxed{\text{Mean of } e \text{ is zero}} \longrightarrow (25)$$

$\therefore$ Total variance in the components of e:

$$E(e^T e) = E\left[\{(I-P)v\}^T \{(I-P)v\}\right]$$

$(I-P)^T = I-P$

$(I-P)^2 = I-P$

$$= E\left[v^T (I-P)^T (I-P) v\right]$$

$$= E[\underbrace{v^T (I-P) v}_{\text{scalar}}]$$

$$= E\left[\operatorname{tr}[v^T (I-P) v]\right]$$

$$\boxed{E(vv^T) = \sigma^2 I_m}$$

$$= E\left[\operatorname{tr}[v v^T (I-P)]\right]$$

$$= \operatorname{tr}\left[E(vv^T)(I-P)\right]$$

$$= \sigma^2 \operatorname{tr}[I_m - P]$$

$$= \sigma^2 [\underbrace{\operatorname{tr}(I_m)}_{=m} - \underbrace{\operatorname{tr}(P)}_{=n}]$$

$$= \sigma^2 (m-n) \longrightarrow (26)$$

$$\operatorname{tr}(I_m) = m$$

$$P = X(X^TX)^{-1}X^T$$

$$\operatorname{tr}(P) = \operatorname{tr}[X(X^TX)^{-1}X^T]$$

$$= \operatorname{tr}[(X^TX)(X^TX)^{-1}]$$

$$= \operatorname{tr}[I_n]$$

$$= n$$

- This suggests an estimator for σ^2:

$$\hat{\sigma}^2 = \frac{e^T e}{(m-n)} \longrightarrow (27)$$

- This is an unbiased estimator of σ^2.

---

7. - Now to the case when x is such that $(x^T x)$ is nearly singular. or multicollinearity in x

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0 \longrightarrow (28)$

be the eigenvalues of $(X^T X) \in R^{n \times n}$ - Smaller Gramian of x. Then

$$\lambda_1^{1/2} \geq \lambda_2^{1/2} \geq \cdots \geq \lambda_n^{1/2} > 0 \longrightarrow (29)$$

are the singular values of x.

- $\kappa(x) = \dfrac{\lambda_1^{1/2}}{\lambda_n^{1/2}}$ is a measure of the collinearity: larger $\kappa(x)$ more likely $(x^T x)$ is near singular.

- In this case: $(x^T x) V = \cancel{V}\Lambda$, $\boxed{V V^T = V^T V = I \quad V^{-1} = V^T}$

$$\Rightarrow (x^T x) = V \Lambda V^T$$

$$(x^T x)^{-1} = (V \Lambda V^T)^{-1} = (V^T)^{-1} \Lambda^{-1} V^{-1}$$

$$= V \Lambda^{-1} V^T = \sum_{i=1}^{n} \frac{1}{\lambda_i} V_i V_i^T \rightarrow (30)$$

$\therefore$ If λ_n is very small, small changes in λ_n will reflect as big changes in $(x^T x)^{-1}$ and hence $\hat{W}_{LS} = (x^T x)^{-1} x^T y \rightarrow (31)$

(8) One solution to reduce this sensitivity is

Ridge regression:

- Reformulate by changing the cost function:

$$J_R(w) = \frac{1}{2\sigma^2}(y - Xw)^T(y - Xw) + \frac{1}{2}\alpha \|w\|_2^2 \rightarrow (32)$$

where $\cancel{\alpha}\ \lambda > 0$ is the penalty parameter.

- This penalty term is called a regularization term.

$$\nabla J_R(w) = \frac{1}{2\sigma^2}\left[-2X^Ty + 2(X^TX)w + \alpha 2w\right]$$

$$= 0$$

$$\Rightarrow (X^TX + \alpha I) = X^Ty$$

$$\therefore \hat{w}_{LS}^R = (X^TX + \alpha I)^{-1}X^Ty \quad = X_\lambda^+ y \rightarrow (33)$$

$$\boxed{\therefore X_\lambda^+ = (X^TX + \alpha I)^{-1}X^T}$$

(9) To make sense of (33) express X in SVD.

$$(X^TX)V = V\Lambda \qquad (XX^T)U = U\Lambda, \quad X = U\Lambda^{1/2}V^T$$

$$\downarrow$$

- From (33): $(X^TX) = V\Lambda V^T, \quad V^TV = VV^T = I \rightarrow (34)$

$$X_\lambda^+ = [V\Lambda V^T + \alpha VV^T]^{-1}[U\Lambda^{1/2}V^T]^T$$

$$= [V(\Lambda + \alpha I)V^T]^{-1}[U\Lambda^{1/2}V^T]^T$$

$$= V(\Lambda + \alpha I)^{-1}\underbrace{V^TV}_{=I_m}\Lambda^{1/2}U^T$$

$$\boxed{\begin{gathered} V^{-1} = V^T \\ (AB)^{-1} = B^{-1}A^{-1} \end{gathered}}$$

$$= V \underbrace{(\Lambda + \alpha I)^{-1} \Lambda^{1/2}} U^T = \underline{\text{SVD of } X^+_\lambda} \longrightarrow (35)$$

$$= V D U^T \quad \text{where } D = \operatorname{diag}(d_{11}\, d_{22} \cdots d_{nn})$$

$$d_{ii} = \frac{\lambda_i^{1/2}}{\lambda_i + \alpha} \qquad (1 \le i \le n)$$

$$\therefore \quad X^+_\lambda = \sum_{i=1}^{n} \left(\frac{\lambda_i^{1/2}}{\lambda_i + \alpha} \right) v_i u_i^T \longrightarrow (36)$$

· The addition of $\alpha > 0$ to λ_n stabilizes the estimate $\hat{w}_{LS}$

(10) <u>For Comparison</u>: Consider a special case when columns of X are orthogonal:

that is: $X^T X = n I$. Then

$$\hat{W}^R_{LS} = (X^T X + \alpha I)^{-1} X^T [\overbrace{X W + N}^{= y}] = (X^T X + \alpha I)^{-1} [X^T X W + X^T N]$$

$$= (nI + \alpha I)^{-1} n I W + (X^T X + \alpha I)^{-1} X^T N$$

$$= \left(\frac{n}{\alpha + n} \right) W + \frac{X^T N}{\alpha + n}$$

$$\therefore \quad E[\hat{X}^R_{LS}] = \left(\frac{n}{n + \alpha} \right) W + \left(\frac{X^T}{n + \alpha} \right) \underbrace{E(N)}_{= 0}$$

$$= \left(\frac{n}{n + \alpha} \right) W \ne W \Rightarrow \underline{\text{biased}}$$

$$\mathrm{cov}(\hat{w}_{LS}^{R}) = \mathrm{cov}\left(\frac{n}{n+\alpha} w + \frac{x^T N}{n+\alpha}\right)$$

$$= \mathrm{cov}\left(\frac{x^T N}{n+\alpha}\right)$$

$$= \frac{1}{(n+\alpha)^2} E\left[(x^T N)(x^T N)^T\right]$$

$$= \frac{1}{(n+\alpha)^2} x^T \underbrace{E(NN^T)}_{=\sigma^2 I_m} x$$

$$= \frac{\sigma^2 I_m \, x^T x}{(n+\alpha)^2} = \frac{n\sigma^2}{(n+\alpha)^2} I_m$$

$$(x^T x) = n I_m$$
$$\therefore (x^T x)^{-1} = \frac{1}{n} I_m$$
$$\therefore I_m = n(x^T x)^{-1}$$

$$= \frac{n}{(n+\alpha)^2} \sigma^2 \boxed{n(x^T x)^{-1}}$$

$$= \frac{n^2 \sigma^2}{(n+\alpha)^2} (x^T x)^{-1}$$

Recall $\mathrm{cov}(\hat{w}_{LS}) = \sigma^2 (x^T x)^{-1}$

$$\therefore \mathrm{cov}(\hat{w}_{LS}^{R}) < \mathrm{cov}(\hat{x}_{LS}$$

since $\dfrac{n^2}{(n+\alpha)^2} \sigma^2 (x^T x)^{-1} < \sigma^2 (x^T x)^{-1}$

$$\therefore \mathrm{cov}(\hat{w}_{LS}) - \mathrm{cov}(\hat{w}_{LS}^{R})$$

$$= \sigma^2 (x^T x)^{-1} - \frac{n^2}{(n+\alpha)^2} \sigma^2 (x^T x)^{-1}$$

$$= \sigma^2 (X^T X)^{-1} \left[1 - \frac{n^2}{(n+\alpha)^2} \right]$$

$$= \sigma^2 (X^T X)^{-1} \left[\frac{\alpha (2n + \alpha)}{(n+\alpha)^2} \right] > 0$$

(i) Bias in $\hat{w}^R_{LS} = \frac{n}{n+\alpha}$

Difference in Variance $= \left[1 - \frac{n^2}{(n+\alpha)^2} \right] > 0$

(ii) $\hat{w}^R_{LS}$ has lesser variance but biased.

This is called bias - Covariance trade-off