SI151

Discussion 1

2020.3.9

Review

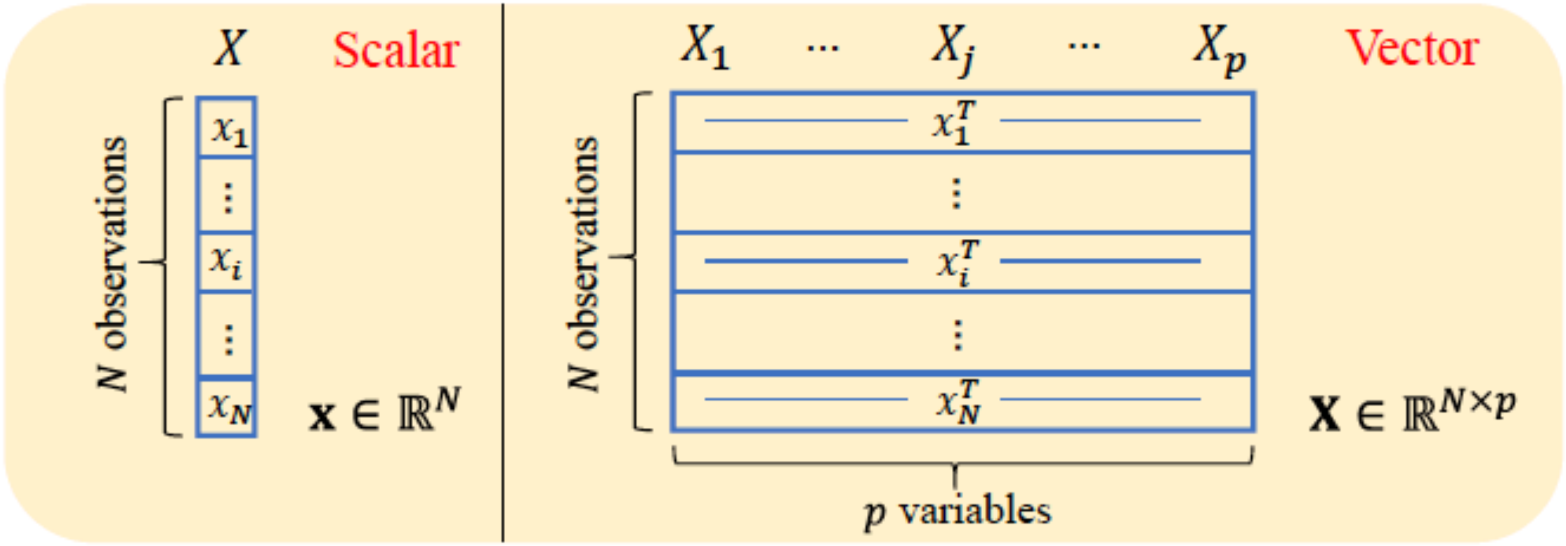
Variable Types and Terminology

Input: a variable X. If X is a vector, its j-th element is X_j

an observation x_i
(scalar or vector)

Typically, we use i to denote the index of observations, while use j to denote the index of variables.

Model



Simple Approach 1: Least Squares

- Training procedure:
  Method of *least-squares*
- N = #observations
- Minimize the *residual sum of squares*

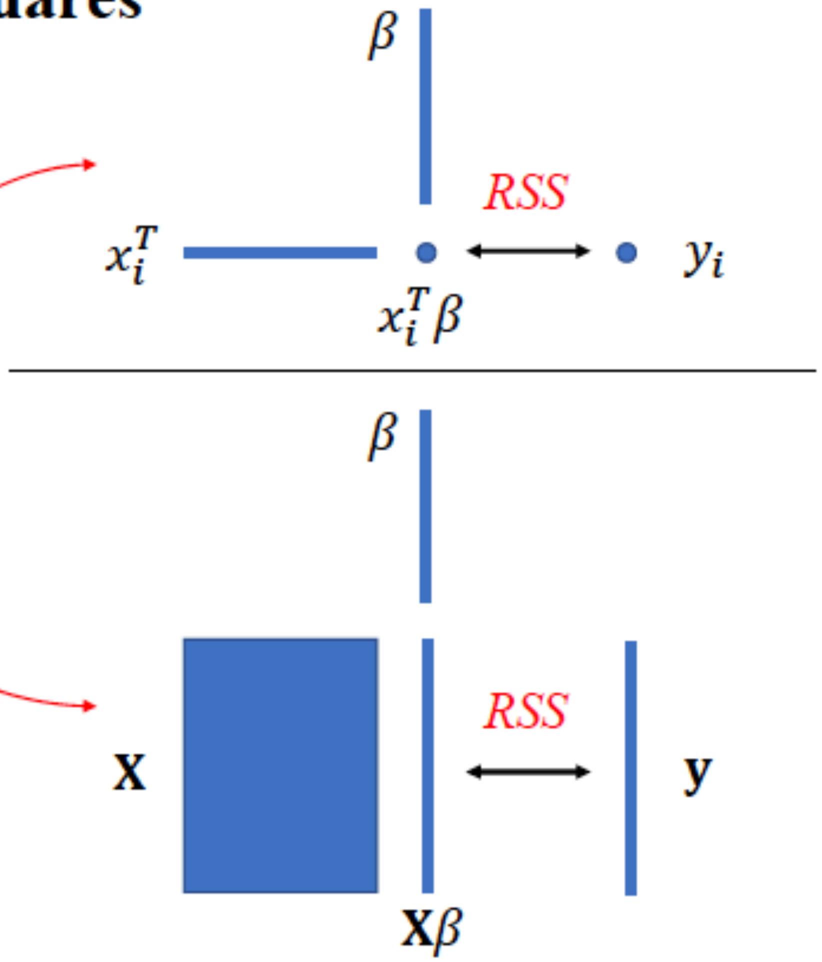
$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} \boxed{(y_i - x_i^T \beta)^2}$$

Or equivalently,

$$\begin{aligned}\mathrm{RSS}(\beta) &= \boxed{(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)} \\ &= \|\mathbf{y} - \mathbf{X}\beta\|_2^2\end{aligned}$$

- This quadratic function always has a global minimum, but it may not be unique.

Q: What is the difference among x_i, x_i^T, $\mathbf{x}$, X and $\mathbf{X}$?



Simple Approach 1: Least Squares

- Training procedure:
  Method of *least-squares*
- N = #observations
- Minimize the *residual sum of squares*

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2$$

Or equivalently,

$$\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$
$$= \|\mathbf{y} - \mathbf{X}\beta\|_2^2$$

- This quadratic function always has a global minimum, but it may not be unique.
- Differentiating w.r.t. β yields the *normal equations*

$$\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) = 0$$

- If $\mathbf{X}^T\mathbf{X}$ is nonsingular, then the unique solution is

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- The fitted value at an arbitrary input x_0 is

$$\hat{y}(x_0) = x_0^T \hat{\beta}$$

- The entire fitted surface is characterized by $\hat{\beta}$.

Differential of Vector(Matrix)

- *(scalar to scalar)* $\mathrm{d}f = f'(x)dx$
- *(scalar to vector)* $df = \sum_i \frac{\partial f}{\partial x_i} dx_i = \frac{\partial f}{\partial \boldsymbol{x}}^T d\boldsymbol{x}$

- $RSS(\beta) = (X\beta - y)^T (X\beta - y)$
- $X\beta \in R^{n\times 1}, y \in R^{n\times 1}, RSS \in R$
- $dRSS(\beta) = (Xd\beta)^T (X\beta - y) + (X\beta - y)^T (Xd\beta) = 2(X\beta - y)^T X d\beta$
- $dRSS = \frac{\partial RSS}{\partial \beta}^T d\beta$
- $\frac{\partial RSS}{\partial \beta} = (2(X\beta - y)^T X)^T = 2X^T(X\beta - y) = 0 \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y$

- *Question: What about vector to vector?*

- $\frac{\partial Ax}{\partial x} = A^T, \frac{\partial x^T A}{\partial x} = A$

$$RSS(\beta) = (y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta)$$

$$= y^T y - 2\beta^T \mathbf{X}^T y + \beta^T \mathbf{X}^T \mathbf{X}\beta$$

Now, to minimize the function, set the derivative to zero

$$\frac{\partial RSS(\beta)}{\partial \beta} = -2\mathbf{X}^T y + 2\mathbf{X}^T \mathbf{X}\beta = 0$$

$$\Rightarrow \quad \mathbf{X}^T \mathbf{X}\beta = \mathbf{X}^T y$$

$$\Rightarrow \quad \hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^T y$$

Statistical Decision Theory

- Given:
  - random input vector $X \in \mathbb{R}^p$,
  - random output variable $Y \in \mathbb{R}$,
  - joint distribution $\Pr(X, Y)$,
- Goal: we seek a function $f(X)$ for predicting Y given values of X.
- To penalize prediction errors, we introduce the *loss function* $L\big(Y, f(X)\big)$.
- *Squared error loss*:

$$L\big(Y, f(X)\big) = (Y - f(X))^2.$$

- *Expected prediction error (EPE)*:

$$\begin{aligned}\mathrm{EPE}(f) &= \mathrm{E}(Y - f(X))^2 \\ &= \int (y - f(x))^2 \Pr(dx, dy).\end{aligned}$$

- Since $\Pr(X, Y) = \Pr(Y|X)\Pr(X)$, EPE can also be written as

$$\mathrm{EPE}(f) = \mathrm{E}_X \mathrm{E}_{Y|X}([Y - f(X)]^2 | X).$$

- Thus, it suffices to minimize EPE *pointwise*:

$$f(x) = \mathrm{argmin}_c \, \mathrm{E}_{Y|X}\,([Y - c]^2 | X = x)$$

Regression function: $f(x) = E(Y|X = x)$.

- $EPE(f) = E\left[\left(Y - f(X)\right)^2\right] = \int_{x,y}\left(y - f(x)\right)^2 \mathrm{P}(dx, dy)$
- $Bayes's\ rule: P(x, y) = P(y|x)P(x), P(dx, dy) = P(dy|dx)P(dx)$
- $EPE(f) = \int_x \int_{y|x} (y - f(x))^2 |x) P(dy|dx)\, P(dx) = E_X E_{Y|X}[\left(Y - f(X)\right)^2 |X]$
- Hint: $E_{XY}(g(X, Y)) = E_X E_{Y|X}(g(X, Y))$

$$f(x) = \mathrm{argmin}_c\, \mathrm{E}_{Y|X}\,([Y - c]^2 | X = x)$$

$$E[(Y - c)^2 | X = x] = E[(Y - E(Y|X = x) + E(Y|X = x) - c)^2]$$
$$= var(Y|X = x) + 2E[(Y - E(Y|X = x))(E(Y|X = x) - c)] + E[(E(Y|X = x) - c)^2]$$

f(x)=c=E(Y|X=x)[our result: Regression function]
min E=var(Y|X=x)

$$f(x) = \arg\min_{f} E_{Y|X}([Y - f]^2 | X = x)$$

$$\Rightarrow \frac{\partial}{\partial f} \int [Y - f]^2 \Pr(y|x)\, dy = 0$$

$$= \int \frac{\partial}{\partial f} [y - f]^2 \Pr(y|X)\, dy = 0$$

$$\Rightarrow 2 \int y Pr(y|x) dy = 2f \int \quad Pr(y|x)\, dy$$

$$\Rightarrow 2E[Y|X] = 2f$$

$$\Rightarrow f = E[Y|X = x].$$

Statistical Decision Theory

- Linear regression assumes that the regression function is approximately linear
$$f(x) \approx x^T \beta.$$
- This is a model-based approach.
- Plugging this $f(x)$ into EPE,

$$\mathrm{EPE}(f) = E(Y - f(X))^2$$
$$= \mathrm{E}((Y - X^T\beta)^T(Y - X^T\beta))$$

- Differentiating w.r.t. β, leads to
$$\beta = [\mathrm{E}(XX^T)]^{-1}\mathrm{E}(XY)$$

$$E\left(X(X^T\beta - Y)\right) = 0$$
$$E(XX^T)\beta = E(XY)$$
$$\beta = [E(XX^T)]^{-1}E(XY)$$

- Again, linear regression replaces the theoretical expectation by averaging over the observed data
$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N}(y_i - x_i^T\beta)^2$$
$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$
- Summary – approximation of $f(X)$
  - Least squares:
    globally linear function
  - Nearest neighbors:
    locally constant function.

Regression function: $f(x) = E(Y|X = x)$.

Statistical Decision Theory

- Additional methods in our course are often model-based but more flexible than the linear model.
- For example, additive models

$$f(X) = \sum_{j=1}^{p} f_j(X_j)$$

  - Coordinate function f_j is arbitrary.
  - Approximate *univariate* conditional expectations *simultaneously* for each f_j.
  - Model assumption: additivity.

- What happens if we use another loss function?

$$L_1(Y, f(X)) = E|Y - f(X)|$$

- In this case,

$$\hat{f}(x) = \text{median}(Y|X = x)$$

- More robust than the conditional mean.
- Summary:
  - L_1 criterion not differentiable.
  - Squared error is the most popular.

$$f(x) = \arg\min_f E_{Y|X}(|Y - f||X = x)$$

$$= \frac{\partial}{\partial f}\int |Y - f|\Pr(y|X)\,dy = 0$$

By LLM, we have $\int |Y - f|\Pr(y|X)dy = Lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}|Y_i - f| \approx \frac{1}{n}\sum_{i=1}^{n}|Y_i - f|$

$$\approx \frac{\partial}{\partial f}\frac{1}{n}\sum_{i=1}^{n}|Y_i - f| = 0$$

$$\frac{\partial}{\partial f}|Y_i - f| = \begin{cases} -1, & Y_i - f > 0 \\ 1, & Y_i - f < 0 \\ 0, & Y_i = f \end{cases}.$$

$$= \frac{1}{n}\sum_{i=1}^{n} -sign(Y_i - f) = 0$$

$$= \sum_{i=1}^{n} sign(Y_i - f) = 0.$$

$$\boxed{\hat{f}(x) = \text{median}(Y|X = x).}$$

Statistical Decision Theory

- Procedure for categorical output variable G with values from $\mathcal{G}$.
- Loss function is $K \times K$ matrix $\mathbf{L}$, where $K = \mathrm{card}(\mathcal{G})$
- $\mathbf{L}(k, l)$ is the price paid for misclassifying an observation belonging to class $\mathcal{G}_k$ as class $\mathcal{G}_l$
- $\mathbf{L}$ is zero on the diagonal
- We often use the zero-one loss function

$$\mathbf{L}(k, l) = 1 - \delta_{kl}$$

where $\delta_{kl} = 1$ if $k = l$, otherwise $\delta_{kl} = 0$

- Expected prediction error (EPE)

$$\mathrm{EPE} = \mathrm{E}[L(G, \hat{G}(X))]$$

where expectation taken w.r.t. $\Pr(G, X)$

- Conditioning on X yields

$$\mathrm{EPE} = \mathrm{E}_X \sum_{k=1}^{K} L[\mathcal{G}_k, \hat{G}(X)] \Pr(\mathcal{G}_k | X)$$

- Again, it suffices to pointwise minimization

$$\hat{G}(x) = \mathrm{argmin}_{g \in \mathrm{G}} \sum_{k=1}^{K} L(\mathcal{G}_k, g) \Pr(\mathcal{G}_k | X = x)$$

- Or simply

Bayes classifier

$$\hat{G}(x) = \max_{g \in \mathrm{G}} \Pr(g | X = x)$$

Local Models in High Dimensions

$$E_{\mathcal{T}}\big(\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0)\big)\big(E_{\mathcal{T}}(\hat{y}_0) - f(x_0)\big) = 0$$

$$\begin{aligned}
\mathrm{MSE}(x_0) &= E_{\mathcal{T}}[f(x_0) - \hat{y}_0]^2 \\
&= E_{\mathcal{T}}[\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0) + E_{\mathcal{T}}(\hat{y}_0) - f(x_0)]^2 \\
&= E_{\mathcal{T}}\left[\big(\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0)\big)^2 + 2\underline{\big(\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0)\big)\big(E_{\mathcal{T}}(\hat{y}_0) - f(x_0)\big)} + \big(E_{\mathcal{T}}(\hat{y}_0) - \underbrace{f(x_0)}_{Constant}\big)^2\right] \\
&= E_{\mathcal{T}}\left[\big(\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0)\big)^2\right] + \big(E_{\mathcal{T}}(\hat{y}_0) - f(x_0)\big)^2 \\
&= \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) + \mathrm{Bias}^2(\hat{y}_0)
\end{aligned}$$

This is known as the bias-variance decomposition.

Exercise

ESL EX2.2

- Show how to compute the Bayes decision boundary for the simulation example in Figure 2.5.

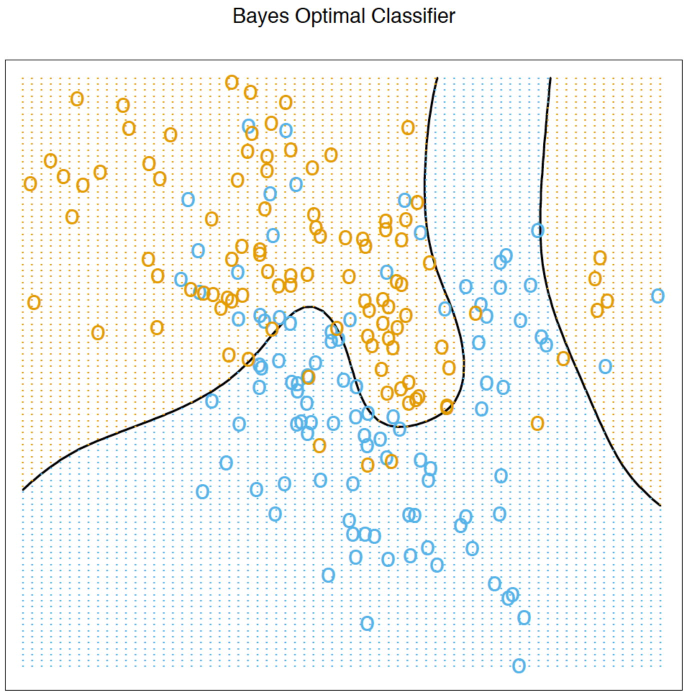


FIGURE 2.5. *The optimal Bayes decision boundary for the simulation example of Figures 2.1, 2.2 and 2.3. Since the generating density is known for each class, this boundary can be calculated exactly (Exercise 2.2).*

Class 1	Class 2
1. 10 means generated from a bivariate Gaussian $N((0,1)^T, I)$. 2. 100 Samples selected as follows a. For each observation, m_k was selected with probability $\frac{1}{10}$. b. Then a sample was generated from the bivariate Gaussian $N(m_k, \frac{I}{5})$.	1. 10 means generated from a bivariate Gaussian $N((1,0)^T, I)$. 2. 100 Samples selected as follows a. For each observation, n_i was selected with probability $\frac{1}{10}$. b. Then a sample was generated from the bivariate Gaussian $N(n_i, \frac{I}{5})$.

$$Boundary = \left\{x: \max_{g \in G} \Pr(g|X = x) = \max_{k \in G} \Pr(k|X = x)\right\}.$$

$$Boundary = \{x: \Pr(g|X = x) = \Pr(k|X = x)\}$$

$$= \left\{x: \frac{\Pr(g|X = x)}{\Pr(k|X = x)} = 1\right\}.$$

$$\frac{\Pr(g|X=x)}{\Pr(k|X=x)} = \frac{\Pr(X=x|g)\Pr(g)\,/\Pr(X=x)}{\Pr(X=x|k)\Pr(k)\,/\Pr(X=x)} = \frac{\Pr(X=x|g)\Pr(g)}{\Pr(X=x|k)\Pr(k)} = 1$$

$$\Pr(X=x|g) = \prod_{k=1}^{10} \frac{1}{5\sqrt{2\pi}} \exp\left(-\frac{(x-m_k)^2}{2\cdot 25}\right)$$

$$\log(\Pr(X=x|g)) = \sum_{k=1}^{10} \log\left(\frac{1}{5\sqrt{2\pi}}\right) - \frac{(x-m_k)^2}{2\cdot 25}.$$

$$Boundary = \left\{x: \sum_{k=1}^{10} \log\left(\frac{1}{5\sqrt{2\pi}}\right) - \frac{(x-m_k)^2}{2\cdot 25} = \sum_{i=1}^{10} \log\left(\frac{1}{5\sqrt{2\pi}}\right) - \frac{(x-n_i)^2}{2\cdot 25}\right\}$$

$$= \left\{x: \sum_{k=1}^{10} (x-m_k)^2 = \sum_{i=1}^{10} (x-n_i)^2\right\}$$

ESL EX2.5

- (a) Derive equation (2.27). The last line makes use of (3.8) through a conditioning argument.

$$\begin{aligned}
\mathrm{EPE}(x_0) &= \mathrm{E}_{y_0|x_0}\mathrm{E}_{\mathcal{T}}(y_0-\hat{y}_0)^2 \\
&= \mathrm{Var}(y_0|x_0) + \mathrm{E}_{\mathcal{T}}[\hat{y}_0 - \mathrm{E}_{\mathcal{T}}\hat{y}_0]^2 + [\mathrm{E}_{\mathcal{T}}\hat{y}_0 - x_0^T\beta]^2 \\
&= \mathrm{Var}(y_0|x_0) + \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) + \mathrm{Bias}^2(\hat{y}_0) \\
&= \sigma^2 + \mathrm{E}_{\mathcal{T}}x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\sigma^2 + 0^2. \qquad (2.27)
\end{aligned}$$

$$Y = X^T\beta + \epsilon$$

$$Var_T(\hat{y}_0) = Var_T(x_0^T\hat{\beta}) = x_0^T Var_T(\hat{\beta})x_0.$$

Hint: $Cov(\mathbf{A}x+a)=\mathbf{A}Cov(x)\mathbf{A}^T\, for\ x \in R_p$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta+\epsilon)$$

$$Var_T(\hat{\beta}) = Var((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Var_T(\epsilon)X(\mathbf{X}^T\mathbf{X})^{-1} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

$$Var_T(\hat{y}_0) = x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\,\sigma^2 = E_T x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\,\sigma^2$$

- (b) Derive equation (2.28), making use of the cyclic property of the trace operator [trace(AB) = trace(BA)], and its linearity (which allows us to interchange the order of trace and expectation).

$$\begin{aligned} \mathrm{E}_{x_0}\mathrm{EPE}(x_0) &\sim \mathrm{E}_{x_0} x_0^T \mathrm{Cov}(X)^{-1} x_0 \sigma^2/N + \sigma^2 \\ &= \mathrm{trace}[\mathrm{Cov}(X)^{-1}\mathrm{Cov}(x_0)]\sigma^2/N + \sigma^2 \\ &= \sigma^2(p/N) + \sigma^2. \end{aligned}$$

By (a), we have $EPE(x_0) = \sigma^2 + \mathrm{E}_{\mathcal{T}} x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\sigma^2 + 0^2.$

$\boxed{x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\sigma^2 = E_{\mathcal{T}} x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\sigma^2}$

$$E_{x_0}\left[x_0^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}x_0\sigma^2\right] = trace\left(E_{x_0}\left[x_0^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}x_0\sigma^2\right]\right)$$

$$= E_{x_0}\left[trace\left(x_0^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}x_0\sigma^2\right)\right] = \sigma^2 E_{x_0}\left[trace\left(x_0x_0^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)\right]$$

$$= \sigma^2 trace\left(E_{x_0}[x_0x_0^T]\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right) = \sigma^2 trace\left(\frac{Cov(X)Cov(X)^{-1}}{N}\right) = \frac{\sigma^2 trace(I_p)}{N}$$

$$= \frac{\sigma^2 p}{N}$$

$$Cov(X) = E[(X-\mu)(X-\mu)^T] = E[XX^T] - \mu\mu^T$$

Our assumption is the matrix X has mean 0 along the columns, so $\mu = 0$

$$\frac{\mathbf{X}^T\mathbf{X}}{\mathrm{N}} = \begin{bmatrix} - & \mathbf{x}_1^T & - \\ - & \mathbf{x}_2^T & - \\ & \vdots & \\ - & \mathbf{x}_n^T & - \end{bmatrix} \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \\ | & | & & | \end{bmatrix} / N$$

$$= \begin{bmatrix} \frac{\mathbf{x}_1^T\mathbf{x}_1}{N} & \frac{\mathbf{x}_1^T\mathbf{x}_2}{N} & \cdots & \frac{\mathbf{x}_1^T\mathbf{x}_p}{N} \\ \frac{\mathbf{x}_2^T\mathbf{x}_1}{N} & \frac{\mathbf{x}_2^T\mathbf{x}_2}{N} & \cdots & \frac{\mathbf{x}_2^T\mathbf{x}_p}{N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathbf{x}_p^T\mathbf{x}_1}{N} & \frac{\mathbf{x}_p^T\mathbf{x}_1}{N} & \cdots & \frac{\mathbf{x}_p^T\mathbf{x}_p}{N} \end{bmatrix}$$

$$= \begin{bmatrix} \widehat{Cov}(X_1, X_1) & \widehat{Cov}(X_1, X_2) & \cdots & \widehat{Cov}(X_1, X_p) \\ \widehat{Cov}(X_2, X_1) & \widehat{Cov}(X_2, X_2) & \cdots & \widehat{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{Cov}(X_p, X_1) & \widehat{Cov}(X_p, X_2) & \cdots & \widehat{Cov}(X_p, X_p) \end{bmatrix}$$

So that in the first line, if N is large and assuming $E(X) = 0$, then $\mathbf{X}^T\mathbf{X} \to NCov(X)$.