

Maximum Likelihood Estimation

Merlise Clyde

STA721 Linear Models

Duke University

August 27, 2015

Topics

- Likelihood Function
- Projections
- Maximum Likelihood Estimates

Readings: Christensen Chapter 1-2, Appendix A, and Appendix B

Take an random vector $\mathbf{Y} \in \mathbb{R}^n$ which is observable and decompose

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

into $\boldsymbol{\mu} \in \mathbb{R}^n$ (unknown, fixed) and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ unobservable error vector (random)

Usual assumptions?

- $E[\epsilon_i] = 0 \ \forall i \Leftrightarrow E[\boldsymbol{\epsilon}] = \mathbf{0} \quad \Rightarrow E[\mathbf{Y}] = \boldsymbol{\mu}$ (mean vector)
- ϵ_i independent with $\text{Var}(\epsilon_i) = \sigma^2$ and $\text{Cov}(\epsilon_i, \epsilon_j) = 0$
- Matrix version

$$\text{Cov}[\boldsymbol{\epsilon}] \equiv [(E[\epsilon_i - E[\epsilon_i]])(E[\epsilon_j - E[\epsilon_j]])]_{ij} = \sigma^2 \mathbf{I}_n$$

$\Rightarrow \text{Cov}[\mathbf{Y}] = \sigma^2 \mathbf{I}_n$ (errors are uncorrelated)

- $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ implies that $Y_i \stackrel{\text{ind}}{\sim} N(\mu_i, \sigma^2)$

Likelihood Functions

The likelihood function for μ, σ^2 is proportional to the sampling distribution of the data

$$\begin{aligned}\mathcal{L}(\mu, \sigma^2) &\propto \prod_{i=1}^n \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp -\frac{1}{2} \left\{ \frac{(y_i - \mu_i)^2}{\sigma^2} \right\} \\ &\propto (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2} \frac{\sum_i (Y_i - \mu_i)^2}{\sigma^2} \right\} \\ &\propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2} \frac{(\mathbf{Y} - \mu)^T (\mathbf{Y} - \mu)}{\sigma^2} \right\} \\ &\propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2} \frac{\|\mathbf{Y} - \mu\|^2}{\sigma^2} \right\} \\ &\propto (2\pi)^{-n/2} |\mathbf{I}_n \sigma^2|^{-1/2} \exp \left\{ -\frac{1}{2} \frac{\|\mathbf{Y} - \mu\|^2}{\sigma^2} \right\}\end{aligned}$$

Last line is the density of $\mathbf{Y} \sim N_n(\mu, \sigma^2 \mathbf{I}_n)$

Find values of $\hat{\boldsymbol{\mu}}$ and $\hat{\sigma}^2$ that maximize the likelihood $\mathcal{L}(\boldsymbol{\mu}, \sigma^2)$ for $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\sigma^2 \in \mathbb{R}^+$

$$\begin{aligned}\mathcal{L}(\boldsymbol{\mu}, \sigma^2) &\propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2} \frac{\|\mathbf{Y} - \boldsymbol{\mu}\|^2}{\sigma^2} \right\} \\ \mathcal{L}(\boldsymbol{\mu}, \sigma^2) \log(\mathcal{L}) &\propto -\frac{n}{2} \log(\sigma^2) - \frac{1}{2} \frac{\|\mathbf{Y} - \boldsymbol{\mu}\|^2}{\sigma^2}\end{aligned}$$

or equivalently the log likelihood

Clearly, $\hat{\boldsymbol{\mu}} = \mathbf{Y}$ but $\hat{\sigma}^2 = 0$ is outside the parameter space

Need restrictions on $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$

Column Space

- Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p \in \mathbb{R}^n$
- The set of all linear combinations of $\mathbf{X}_1, \dots, \mathbf{X}_p$ is the space spanned by $\mathbf{X}_1, \dots, \mathbf{X}_p \equiv S(\mathbf{X}_1, \dots, \mathbf{X}_p)$
- Let $\mathbf{X} = [\mathbf{X}_1 \mathbf{X}_2 \dots \mathbf{X}_p]$ be a $n \times p$ matrix with columns \mathbf{X}_j ; then the column space of \mathbf{X} , $C(\mathbf{X}) = S(\mathbf{X}_1, \dots, \mathbf{X}_p)$ space spanned by the (column) vectors of \mathbf{X}
- $\boldsymbol{\mu} \in C(\mathbf{X}) : C(\mathbf{X}) = \{\boldsymbol{\mu} \mid \boldsymbol{\mu} \in \mathbb{R}^n \text{ such that } \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu} \text{ for some } \boldsymbol{\beta} \in \mathbb{R}^p\}$ (also called the Range of \mathbf{X} , $R(\mathbf{X})$)
- $\boldsymbol{\beta}$ are the “coordinates” of $\boldsymbol{\mu}$ in this space
- $C(\mathbf{X})$ is a subspace of \mathbb{R}^n

Many equivalent ways to represent the same mean vector – inference should be independent of the coordinate system used

Projections

- $\mu = \mathbf{X}\beta$ with \mathbf{X} full rank $\mu \in C(\mathbf{X})$
- $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
- $\mathbf{P}_\mathbf{X}$ is the orthogonal projection operator on the column space of \mathbf{X} ; e.g.
- $\mathbf{P} = \mathbf{P}^2$ idempotent (projection)

$$\begin{aligned}\mathbf{P}_\mathbf{X}^2 = \mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{P}_\mathbf{X}\end{aligned}$$

- $\mathbf{P} = \mathbf{P}^T$ symmetry (orthogonal)

$$\begin{aligned}\mathbf{P}_\mathbf{X}^T &= (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\ &= (\mathbf{X}^T)^T((\mathbf{X}^T\mathbf{X})^{-1})^T(\mathbf{X})^T \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{P}_\mathbf{X}\end{aligned}$$

- $\mathbf{P}_\mathbf{X}\mu = \mathbf{P}_\mathbf{X}\mathbf{X}\beta = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}\beta = \mu$

Claim: $\mathbf{I} - \mathbf{P}_\mathbf{X}$ is an orthogonal projection onto $C(\mathbf{X})^\perp$

- idempotent

$$\begin{aligned}(\mathbf{I} - \mathbf{P}_\mathbf{X})^2 &= (\mathbf{I} - \mathbf{P}_\mathbf{X})(\mathbf{I} - \mathbf{P}_\mathbf{X}) \\&= \mathbf{I} - \mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{X} + \mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X} \\&= \mathbf{I} - \mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{X} + \mathbf{P}_\mathbf{X} \\&= \mathbf{I} - \mathbf{P}_\mathbf{X}\end{aligned}$$

- Symmetry $\mathbf{I} - \mathbf{P}_\mathbf{X} = (\mathbf{I} - \mathbf{P}_\mathbf{X})^T$
- $\mathbf{u} \in C(\mathbf{X})^\perp \Rightarrow \mathbf{u} \perp C(\mathbf{X})$ that is $u \in C(\mathbf{X})^\perp$ and $v \in C(\mathbf{X})$
then $\mathbf{u}^T \mathbf{v} = 0$
- $(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{u} = \mathbf{u}$ (projection)
- if $\mathbf{v} \in C(\mathbf{X})$, $(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{v} = \mathbf{v} - \mathbf{v} = \mathbf{0}$

$\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$ with $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and \mathbf{X} full column rank

Claim: Maximum Likelihood Estimator (MLE) of $\boldsymbol{\mu}$ is $\mathbf{P}_\mathbf{X}\mathbf{Y}$

- Log Likelihood:

$$\log \mathcal{L}(\boldsymbol{\mu}, \sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2} \frac{\|\mathbf{Y} - \boldsymbol{\mu}\|^2}{\sigma^2}$$

- Decompose $\mathbf{Y} = \mathbf{P}_\mathbf{X}\mathbf{Y} + (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}$
- Use $\mathbf{P}_\mathbf{X}\boldsymbol{\mu} = \boldsymbol{\mu}$
- Simplify $\|\mathbf{Y} - \boldsymbol{\mu}\|^2$

Expand

$$\begin{aligned}\|\mathbf{Y} - \boldsymbol{\mu}\|^2 &= \|(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y} + \mathbf{P}_\mathbf{X}\mathbf{Y} - \boldsymbol{\mu}\|^2 \\&= \|(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y} + \mathbf{P}_\mathbf{X}\mathbf{Y} - \mathbf{P}_\mathbf{X}\boldsymbol{\mu}\|^2 \\&= \|(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y} + \mathbf{P}_\mathbf{X}(\mathbf{Y} - \boldsymbol{\mu})\|^2 \\&= \|(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}\|^2 + \|\mathbf{P}_\mathbf{X}(\mathbf{Y} - \boldsymbol{\mu})\|^2 + 2(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{P}_\mathbf{X}^T (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y} \\&= \|(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}\|^2 + \|\mathbf{P}_\mathbf{X}(\mathbf{Y} - \boldsymbol{\mu})\|^2 + 0 \\&= \|(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}\|^2 + \|\mathbf{P}_\mathbf{X}\mathbf{Y} - \boldsymbol{\mu}\|^2\end{aligned}$$

Crossproduct term is zero

$$\begin{aligned}\mathbf{P}_\mathbf{X}^T (\mathbf{I} - \mathbf{P}_\mathbf{X}) &= \mathbf{P}_\mathbf{X} (\mathbf{I} - \mathbf{P}_\mathbf{X}) \\&= \mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{X} \mathbf{P}_\mathbf{X} \\&= \mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{X} \\&= 0\end{aligned}$$

Substitute decomposition into log likelihood

$$\begin{aligned}\log \mathcal{L}(\boldsymbol{\mu}, \sigma^2) &= -\frac{n}{2} \log(\sigma^2) - \frac{1}{2} \frac{\|\mathbf{Y} - \boldsymbol{\mu}\|^2}{\sigma^2} \\&= -\frac{n}{2} \log(\sigma^2) - \frac{1}{2} \left(\frac{\|(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}\|^2}{\sigma^2} + \frac{\|\mathbf{P}_\mathbf{X}\mathbf{Y} - \boldsymbol{\mu}\|^2}{\sigma^2} \right) \\&= \underbrace{-\frac{n}{2} \log(\sigma^2) - \frac{1}{2} \frac{\|(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}\|^2}{\sigma^2}}_{\text{constant with respect to } \boldsymbol{\mu}} + \underbrace{-\frac{1}{2} \frac{\|\mathbf{P}_\mathbf{X}\mathbf{Y} - \boldsymbol{\mu}\|^2}{\sigma^2}}_{\leq 0} \\&= \text{constant with respect to } \boldsymbol{\mu} \leq 0\end{aligned}$$

Maximize with respect to $\boldsymbol{\mu}$ for each σ^2

RHS is largest when $\boldsymbol{\mu} = \mathbf{P}_\mathbf{X}\mathbf{Y}$ for any choice of σ^2

$$\therefore \hat{\boldsymbol{\mu}} = \mathbf{P}_\mathbf{X}\mathbf{Y}$$

is the MLE of $\boldsymbol{\mu}$ (yields fitted values $\hat{\mathbf{Y}} = \mathbf{P}_\mathbf{X}\mathbf{Y}$)

$$\begin{aligned}\mathcal{L}(\mu, \sigma^2) &= -\frac{n}{2} \log(\sigma^2) - \frac{1}{2} \left(\frac{\|(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}\|^2}{\sigma^2} + \frac{\|\mathbf{P}_\mathbf{X}\mathbf{Y} - \mu\|^2}{\sigma^2} \right) \\ \mathcal{L}(\beta, \sigma^2) &= -\frac{n}{2} \log(\sigma^2) - \frac{1}{2} \left(\frac{\|(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}\|^2}{\sigma^2} + \frac{\|\mathbf{P}_\mathbf{X}\mathbf{Y} - \mathbf{X}\beta\|^2}{\sigma^2} \right)\end{aligned}$$

Similar argument to show that RHS is maximized by minimizing

$$\|\mathbf{P}_\mathbf{X}\mathbf{Y} - \mathbf{X}\beta\|^2$$

Therefore $\hat{\beta}$ is a MLE of β if and only if satisfies

$$\mathbf{P}_\mathbf{X}\mathbf{Y} = \mathbf{X}\hat{\beta}$$

If $\mathbf{X}^T\mathbf{X}$ is full rank, the MLE of β is

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

- Plug-in MLE of $\hat{\boldsymbol{\mu}}$ for $\boldsymbol{\mu}$ and differentiate with respect to σ^2

$$\begin{aligned}\log \mathcal{L}(\hat{\boldsymbol{\mu}}, \sigma^2) &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2} \frac{\|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}\|^2}{\sigma^2} \\ \frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\mu}}, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2} \|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}\|^2 \left(\frac{1}{\sigma^2} \right)^2\end{aligned}$$

- Set derivative to zero and solve for MLE

$$\begin{aligned}0 &= -\frac{n}{2} \frac{1}{\hat{\sigma}^2} + \frac{1}{2} \|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}\|^2 \left(\frac{1}{\hat{\sigma}^2} \right)^2 \\ \frac{n}{2} \hat{\sigma}^2 &= \frac{1}{2} \|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}\|^2 \\ \hat{\sigma}^2 &= \frac{\|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}\|^2}{n}\end{aligned}$$

Maximum Likelihood Estimate of σ^2

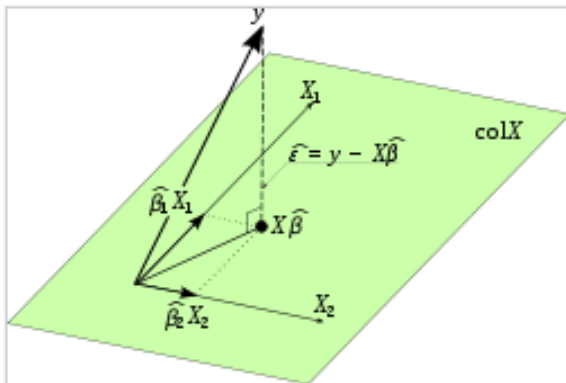
$$\begin{aligned}\hat{\sigma}^2 &= \frac{\|(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}\|^2}{n} \\&= \frac{\mathbf{Y}^T(\mathbf{I} - \mathbf{P}_X)^T(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}}{n} \\&= \frac{\mathbf{Y}^T(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}}{n} \\&= \frac{\mathbf{e}^T\mathbf{e}}{n}\end{aligned}$$

where $\mathbf{e} = (\mathbf{I} - \mathbf{P}_X)\mathbf{Y}$ **residuals** from the regression of \mathbf{Y} on \mathbf{X}

Geometric View

- Fitted Values $\hat{\mathbf{Y}} = \mathbf{P}_X \mathbf{Y} = \mathbf{X} \hat{\boldsymbol{\beta}}$
- Residuals $\mathbf{e} = (\mathbf{I} - \mathbf{P}_X) \mathbf{Y}$
- $\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e}$

$$\|\mathbf{Y}\|^2 = \|\mathbf{P}_X \mathbf{Y}\|^2 + \|(\mathbf{I} - \mathbf{P}_X) \mathbf{Y}\|^2$$



Properties

$\hat{\mathbf{Y}} = \hat{\boldsymbol{\mu}}$ is an unbiased estimate of $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$

$$\begin{aligned} E[\hat{\mathbf{Y}}] &= E[\mathbf{P}_\mathbf{X}\mathbf{Y}] \\ &= \mathbf{P}_\mathbf{X}E[\mathbf{Y}] \\ &= \mathbf{P}_\mathbf{X}\boldsymbol{\mu} \\ &= \boldsymbol{\mu} \end{aligned}$$

$E[\mathbf{e}] = \mathbf{0}$ if $\boldsymbol{\mu} \in C(\mathbf{X})$

$$\begin{aligned} E[\mathbf{e}] &= E[(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}] \\ &= (\mathbf{I} - \mathbf{P}_\mathbf{X})E[\mathbf{Y}] \\ &= (\mathbf{I} - \mathbf{P}_\mathbf{X})\boldsymbol{\mu} \\ &= \mathbf{0} \end{aligned}$$

Will not be $\mathbf{0}$ if $\boldsymbol{\mu} \notin C(\mathbf{X})$ (useful for model checking)

Estimate of σ^2

MLE of σ^2 :

$$\hat{\sigma}^2 = \frac{\mathbf{e}^T \mathbf{e}}{n} = \frac{\mathbf{Y}^T (\mathbf{I} - \mathbf{P}_X) \mathbf{Y}}{n}$$

Is this an unbiased estimate of σ^2 ?

Need expectations of quadratic forms $\mathbf{Y}^T \mathbf{A} \mathbf{Y}$ for \mathbf{A} an $n \times n$ matrix
 \mathbf{Y} a random vector in \mathbb{R}^n

Without loss of generality we can assume that $\mathbf{A} = \mathbf{A}^T$

- $\mathbf{Y}^T \mathbf{A} \mathbf{Y}$ is a scalar
- $\mathbf{Y}^T \mathbf{A} \mathbf{Y} = (\mathbf{Y}^T \mathbf{A} \mathbf{Y})^T = \mathbf{Y}^T \mathbf{A}^T \mathbf{Y}$

$$\frac{\mathbf{Y}^T \mathbf{A} \mathbf{Y} + \mathbf{Y}^T \mathbf{A}^T \mathbf{Y}}{2} = \mathbf{Y}^T \mathbf{A} \mathbf{Y}$$

$$\mathbf{Y}^T \frac{(\mathbf{A} + \mathbf{A}^T)}{2} \mathbf{Y} = \mathbf{Y}^T \mathbf{A} \mathbf{Y}$$

- may take $\mathbf{A} = \mathbf{A}^T$

Expectations of Quadratic Forms

Theorem

Let \mathbf{Y} be a random vector in \mathbb{R}^n with $E[\mathbf{Y}] = \boldsymbol{\mu}$ and $\text{Cov}(\mathbf{Y}) = \boldsymbol{\Sigma}$.
Then $E[\mathbf{Y}^T \mathbf{A} \mathbf{Y}] = \text{tr} \mathbf{A} \boldsymbol{\Sigma} + \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu}$.

Result useful for finding expected values of Mean Squares; no normality required!

Start with $(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{A}(\mathbf{Y} - \boldsymbol{\mu})$, expand and take expectations

$$\begin{aligned} E[(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{A}(\mathbf{Y} - \boldsymbol{\mu})] &= E[\mathbf{Y}^T \mathbf{A} \mathbf{Y} + \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} - \boldsymbol{\mu}^T \mathbf{A} \mathbf{Y} - \mathbf{Y}^T \mathbf{A} \boldsymbol{\mu}] \\ &= E[\mathbf{Y}^T \mathbf{A} \mathbf{Y}] + \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} - \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} - \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} \\ &= E[\mathbf{Y}^T \mathbf{A} \mathbf{Y}] - \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} \end{aligned}$$

Rearrange

$$\begin{aligned} E[\mathbf{Y}^T \mathbf{A} \mathbf{Y}] &= E[(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{A}(\mathbf{Y} - \boldsymbol{\mu})] + \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} \\ &= E[\text{tr}(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{A}(\mathbf{Y} - \boldsymbol{\mu})] + \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} \\ &= E[\text{tr} \mathbf{A}(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T] + \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} \\ &= \text{tr} E[\mathbf{A}(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T] + \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} \\ &= \text{tr} \mathbf{A} E[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T] + \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} \\ &= \text{tr} \mathbf{A} \boldsymbol{\Sigma} + \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} \end{aligned}$$

$$\text{tr} \mathbf{A} \equiv \sum_{i=1}^n a_{ii}$$

Expectation of $\hat{\sigma}^2$

Use the theorem:

$$\begin{aligned}E[\mathbf{Y}^T(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}] &= \text{tr}(\mathbf{I} - \mathbf{P}_\mathbf{X})\sigma^2\mathbf{I} + \boldsymbol{\mu}^T(\mathbf{I} - \mathbf{P}_\mathbf{X})\boldsymbol{\mu} \\&= \sigma^2\text{tr}(\mathbf{I} - \mathbf{P}_\mathbf{X}) \\&= \sigma^2r(\mathbf{I} - \mathbf{P}_\mathbf{X}) \\&= \sigma^2(n - r(\mathbf{X}))\end{aligned}$$

Therefore an unbiased estimate of σ^2 is

$$\frac{\mathbf{e}^T\mathbf{e}}{n - r(\mathbf{X})}$$

If \mathbf{X} is full rank ($r(\mathbf{X}) = p$) and $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ then the

$$\begin{aligned}\text{tr}(\mathbf{P}_\mathbf{X}) &= \text{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) \\&= \text{tr}(\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}) \\&= \text{tr}(\mathbf{I}_p) = p\end{aligned}$$

Theorem

If \mathbf{A} ($n \times n$) is a symmetric real matrix then there exists a \mathbf{U} ($n \times n$) such that $\mathbf{U}^T \mathbf{U} = \mathbf{U} \mathbf{U}^T = \mathbf{I}_n$ and a diagonal matrix $\mathbf{\Lambda}$ with elements λ_i such that $\mathbf{A} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$

- \mathbf{U} is an orthogonal matrix; $\mathbf{U}^{-1} = \mathbf{U}^T$
- The columns of \mathbf{U} form an Orthonormal Basis for \mathbb{R}^n
- rank of \mathbf{A} equals the number of non-zero eigenvalues λ_i
- Columns of \mathbf{U} associated with non-zero eigenvalues form an ONB for $C(\mathbf{A})$ (eigenvectors of \mathbf{A})
- $\mathbf{A}^p = \mathbf{U} \mathbf{\Lambda}^p \mathbf{U}^T$ (matrix powers)
- a square root of $\mathbf{A} \geq 0$ is $\mathbf{U} \mathbf{\Lambda}^{1/2} \mathbf{U}^T$

Projections

Projection Matrix

If P is an orthogonal projection matrix, then its eigenvalues λ_i are either zero or one with $\text{tr}(P) = \sum_i (\lambda_i) = r(P)$

- $P = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$
- $P = P^2 \Rightarrow \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^T$
- $\mathbf{\Lambda} = \mathbf{\Lambda}^2$ is true only for $\lambda_i = 1$ or $\lambda_i = 0$
- Since $r(P)$ is the number of non-zero eigenvalues,
 $r(P) = \sum \lambda_i = \text{tr}(P)$

$$P = [\mathbf{U}_P \mathbf{U}_{P^\perp}] \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{n-r} \end{bmatrix} \begin{bmatrix} \mathbf{U}_P^T \\ \mathbf{U}_{P^\perp}^T \end{bmatrix} = \mathbf{U}_P \mathbf{U}_P^T$$

$$P = \sum_{i=1}^r \mathbf{u}_i \mathbf{u}_i^T$$

sum of r rank 1 projections.

Next Class

distribution theory

Continue Reading Chapter 1-2 and Appendices A & B in
Christensen