Maximum Likelihood Estimation

Merlise Clyde

STA721 Linear Models

Duke University

August 27, 2015

Outline

Topics

- Likelihood Function
- Projections
- Maximum Likelihood Estimates

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

Models

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?

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

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

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

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

Likelihood Functions

The likelihood function for $\boldsymbol{\mu}, \sigma^2$ is proportional to the sampling distribution of the data

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

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

MLEs

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 $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p \in \mathbb{R}^n$
- The set of all linear combinations of $\mathbf{X}_1, \ldots, \mathbf{X}_p$ is the space spanned by $\mathbf{X}_1, \ldots, \mathbf{X}_p \equiv S(\mathbf{X}_1, \ldots, \mathbf{X}_p)$
- Let $\mathbf{X} = [\mathbf{X}_1 \mathbf{X}_2 \ldots \mathbf{X}_p]$ be a $n \times p$ matrix with columns $\mathbf{X}_j$ then the column space of $\mathbf{X}$, $C(\mathbf{X}) = S(\mathbf{X}_1, \ldots, \mathbf{X}_p)$ space spanned by the (column) vectors of $\mathbf{X}$
- $\boldsymbol{\mu} \in C(\mathbf{X}) : C(\mathbf{X}) = \{\boldsymbol{\mu} \mid \boldsymbol{\mu} \in \mathbb{R}^n$ such that $\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}$ 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

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

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

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

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

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

Projections

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

- idempotent

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

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

Log Likelihood

$\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$ with $\boldsymbol{\mu} = \mathbf{X}\beta$ and $\mathbf{X}$ full column rank
Claim: Maximum Likelihood Estimator (MLE) of $\boldsymbol{\mu}$ is $\mathrm{P}_{\mathbf{X}}\mathbf{Y}$

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

Expand

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

Crossproduct term is zero

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

Likelihood

Substitute decomposition into log likelihood

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

Maximize with respect to $\boldsymbol{\mu}$ for each σ^2
RHS is largest when $\boldsymbol{\mu} = \mathsf{P}_{\mathbf{X}}\mathbf{Y}$ for any choice of σ^2

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

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

MLE of β

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

Similar argument to show that RHS is maximized by minimizing

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

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

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

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

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

MLE of σ^2

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

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

- Set derivative to zero and solve for MLE

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

Estimate of σ^2

Maximum Likelihood Estimate of σ^2

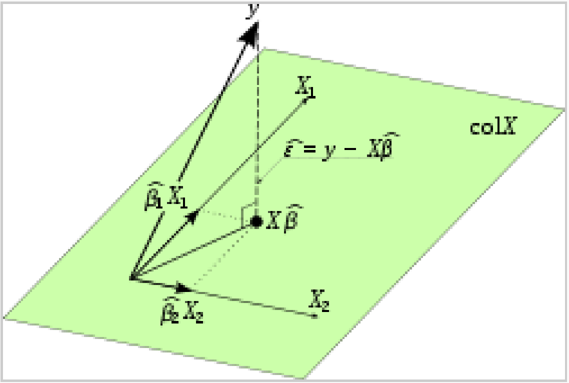
$$\begin{aligned}
\hat{\sigma}^2 &= \frac{\|(\mathbf{I} - \mathrm{P}_{\mathbf{X}})\mathbf{Y}\|^2}{n} \\
&= \frac{\mathbf{Y}^T(\mathbf{I} - \mathrm{P}_{\mathbf{X}})^T(\mathbf{I} - \mathrm{P}_{\mathbf{X}})\mathbf{Y}}{n} \\
&= \frac{\mathbf{Y}^T(\mathbf{I} - \mathrm{P}_{\mathbf{X}})\mathbf{Y}}{n} \\
&=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

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

Geometric View

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

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



Properties

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

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

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

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

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

Estimate of σ^2

MLE of σ^2:

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

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 $\mathbb{R}^n$

Quadratic Forms

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

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

$$\begin{aligned} \frac{\mathbf{Y}^T\mathbf{A}\mathbf{Y} + \mathbf{Y}^T\mathbf{A}^T\mathbf{Y}}{2} &= \mathbf{Y}^T\mathbf{A}\mathbf{Y} \\ \mathbf{Y}^T\frac{(\mathbf{A} + \mathbf{A}^T)}{2}\mathbf{Y} &= \mathbf{Y}^T\mathbf{A}\mathbf{Y} \end{aligned}$$

- 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 $Cov(\mathbf{Y}) = \boldsymbol{\Sigma}$. Then $E[\mathbf{Y}^T\mathbf{A}\mathbf{Y}] = tr\mathbf{A}\Sigma + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$.

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

Proof

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

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

Rearrange

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

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

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

Use the theorem:

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

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 $\mathsf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ then the

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

Spectral Theorem

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 $\boldsymbol{\Lambda}$ with elements λ_i such that $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$

- $\mathbf{U}$ is an orthogonal matrix; $\mathbf{U}^{-1} = \mathbf{U}^T$
- The columns of $\mathbf{U}$ from 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 $\mathbf{A}$)
- $\mathbf{A}^p = \mathbf{U}\boldsymbol{\Lambda}^p\mathbf{U}^T$ (matrix powers)
- a square root of $\mathbf{A} > 0$ is $\mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{U}^T$

Projection Matrix

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

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

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

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

sum of r rank 1 projections.

Next class - distribution theory
Continue Reading Chapter 1-2 and Appendices A & B in Christensen