Here $Y = \begin{bmatrix} Y_{11} & \cdots & Y_{1q} \\ & & \\ Y_{n1} & \cdots & Y_{nq} \end{bmatrix} = \begin{bmatrix} - & y_1^T & - \\ & \vdots & \\ - & y_n^T & - \end{bmatrix}_{n \times q}$ $\quad y_i \in \mathbb{R}^q$

$$X = \begin{bmatrix} - & x_1^T & - \\ & \vdots & \\ - & x_n^T & - \end{bmatrix}_{n \times K}, \quad x_i \in \mathbb{R}^K$$

$$B = \begin{bmatrix} | & & | \\ \beta_1 & \cdots & \beta_q \\ | & & | \end{bmatrix}_{K \times q}. \quad \beta_i \in \mathbb{R}^K$$

$$\varepsilon = \begin{bmatrix} - & \varepsilon_1^T & - \\ & \vdots & \\ - & \varepsilon_n^T & - \end{bmatrix}_{n \times q} \quad \varepsilon_i \in \mathbb{R}^q$$

$\varepsilon_i \overset{iid}{\sim} N(0, \underset{q \times q}{\Sigma})$ and $y_i \sim N(\underset{q \times K}{B^T} \underset{K \times 1}{x_i}, \Sigma)$

$\forall i = 1, \ldots, n$, y_i are idpt.

4 $AIC = -2\log(\text{likelihood}) + 2(\#\text{ parameters})$

$$= nq\log 2\pi + n\log|\hat{\Sigma}_k| + nq + 2kq + q(q+1)$$

we can ignore the constant w.r.t "k".

$$\Rightarrow AIC = n\log|\hat{\Sigma}_k| + 2kq$$

OR $$AIC_k = \log|\hat{\Sigma}_k| + \frac{2kq}{n}$$

$BIC = -2(\log\text{likelihood}) + \log(n) \times (\#\text{ parameters})$

$$= \underbrace{nq\log 2\pi}_{\text{No } k} + n\log|\hat{\Sigma}_k| + \underbrace{nq}_{\text{No } k} + \log(n)kq + \underbrace{\frac{q(q+1)\log(n)}{2}}_{\text{No } k}$$

$$\Rightarrow BIC_k = \log|\hat{\Sigma}_k| + \frac{\log(n)kq}{n}$$

5:

BIC tends to select small order model and therefore are less prone to overfitting.

$$X = \begin{bmatrix} - x_1^T - \\ \vdots \\ - x_n^T - \end{bmatrix}_{n \times p} \to i^{th} \text{ row: } x_i^T.$$

Question 2

$$AIC = -2 \times \underline{\log(\text{likelihood})} + 2 \times \underline{(\# \text{ parameters})}$$

1):

$$-2 \log(\text{likelihood}) = n \log(2\pi) + n \log \sigma^2 + \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

$$-2 \log(\text{likelihood}) \Big|_{\theta = \hat{\theta}_{MLE}} = n \log(2\pi) + n \log \hat{\sigma}^2_{MLE} + n$$

where $\hat{\sigma}^2_{MLE} = \frac{RSS}{n} = \frac{(y - \hat{y})^T (y - \hat{y})}{n}$

$$= \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n}$$

$$= \frac{\sum_{i=1}^{n} (y_i - x_i^T \hat{\beta})^2}{n}$$

Since the constant $n \log 2\pi + n$ play no role in model selection and can be ignored.

$$AIC = n \log \hat{\sigma}^2_K + 2(\# \text{ parameters})$$

$$= n \log \hat{\sigma}^2_K + 2(K+1)$$

σ^2

β

$$\Rightarrow AIC_K = \frac{\log \hat{\sigma}^2_K}{n} + \frac{2(K+1)}{n}$$

2): Suppose the # of parameters in the true model is k_0, but we fit a candidate model of # parameter is $k_0 + L$ $(L>0)$.

overfitting measure by the AIC means: $\boxed{AIC_{k_0+L}} < \boxed{AIC_{k_0}}$

$$Pr\{AIC_{k_0+L} < AIC_{k_0}\}$$

$$= Pr\left\{\log(\hat{\sigma}^2_{k_0+L}) + \frac{2(k_0+L+1)}{n} < \log(\hat{\sigma}^2_{k_0}) + \frac{2(k_0+1)}{n}\right\}$$

where $\hat{\sigma}^2_{k_0+L} = \frac{RSS_{k_0+L}}{n} \Rightarrow \log\hat{\sigma}^2_{k_0+L} = \log(RSS_{k_0+L}) - \log n$

$\hat{\sigma}^2_{k_0} = \frac{RSS_{k_0}}{n} \Rightarrow \log\hat{\sigma}^2_{k_0} = \log(RSS_{k_0}) - \log n$

$$Pr\left\{\log RSS_{k_0+L} + \frac{2(k_0+L+1)}{n} < \log RSS_{k_0} + \frac{2(k_0+1)}{n}\right\}$$

$$= Pr\left\{\log\frac{RSS_{k_0+L}}{RSS_{k_0}} < \frac{-2L}{n}\right\}$$

$$\log\frac{RSS_{k_0+L}}{RSS_{k_0}} = -\log\frac{RSS_{k_0}}{RSS_{k_0+L}}$$

$$= Pr\left\{\log\frac{RSS_{k_0}}{RSS_{k_0+L}} > \frac{2L}{n}\right\}$$

$$= Pr\left\{\frac{RSS_{k_0}}{RSS_{k_0+L}} > e^{\frac{2L}{n}}\right\}$$

$$= \Pr\left\{ \frac{RSS_{k_0} - RSS_{k_0+L}}{RSS_{k_0+L}} > e^{\frac{2L}{n}} - 1 \right\}$$

where $\dfrac{RSS_{k_0} - RSS_{k_0+L}}{\sigma^2} \sim \chi^2_L$

idpt

$\dfrac{RSS_{k_0+L}}{\sigma^2} \sim \chi^2_{n-k_0-L}$

$$\Rightarrow \Pr\left\{ \frac{\chi^2_L / L}{\chi^2_{n-k_0-L} / (n-k_0-L)} > \frac{(n-k_0-L)}{L}\left(e^{\frac{2L}{n}} - 1\right) \right\}$$

$$= \Pr\left\{ F_{L,\, n-k_0-L} > \frac{(n-k_0-L)}{L}\left(e^{\frac{2L}{n}} - 1\right) \right\}$$

why $RSS_{K_0} - RSS_{K_0+L} \perp RSS_{K_0+L}$?

Let $X^*_{K_0} = [\, X_{K_0} \mid \underbrace{0}_{n\times L} \,]_{n\times(K_0+L)}$ $\quad X_{K_0} = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_{K_0} \\ | & & | \end{bmatrix}$

$X_{K_0+L} = [\, X_{K_0} \mid X_L \,]$ $\quad X_L = \begin{bmatrix} | & & | \\ x_{K_0+1} & \cdots & x_{K_0+L} \\ | & & | \end{bmatrix}$

$\underbrace{H_{K_0}}_{\in \mathbb{R}^{K_0+L} \times \mathbb{R}^{K_0+L}} = X^*_{K_0} (X^{*T}_{K_0} X^*_{K_0})^{-1} X^{*T}_{K_0}$ $\quad H_{K_0+L} = X_{K_0+L} (X^T_{K_0+L} X_{K_0+L})^{-1} X_{K_0+L}$

Fact: $\text{colSpace}(X^*_{K_0}) \subseteq \text{colSpace}(X_{K_0+L})$

thus $H_{K_0} H_{K_0+L} = H_{K_0+L} H_{K_0} = H_{K_0}$.

$RSS_{K_0} = y^T (I - H_{K_0}) y \quad RSS_{K_0+L} = y^T (I - H_{K_0+L}) y$.

$RSS_{K_0} - RSS_{K_0+L} = y^T (H_{K_0+L} - H_{K_0}) y$.

$\text{cov}\{ (H_{K_0+L} - H_{K_0}) y,\ (I - H_{K_0+L}) y \}$

$= \sigma^2 (H_{K_0+L} - H_{K_0})(I - H_{K_0+L}) = \sigma^2 (H_{K_0+L} - H_{K_0+L} - H_{K_0} + H_{K_0}) = 0$

3): $\dfrac{n-K_0-L}{L} \left(e^{\frac{2L}{n}} - 1 \right)$

$= \dfrac{n-K_0-L}{L} \left(\dfrac{2L}{n} + O(\tfrac{1}{n^2}) \right)$

$= 2\,\dfrac{n-K_0-L}{n} + O(\tfrac{1}{n}) \xrightarrow[\text{as } n\to\infty]{} 2$

and $F_{L,\, n-K_0-L} \longrightarrow \dfrac{\chi^2_L}{L}$ (why? since $\dfrac{\chi^2_{(n)}}{n} = \dfrac{1}{n}\sum_{i=1}^{n} z_i^2 \xrightarrow{LNN} \mathbb{E}\{\chi^2_{(1)}\} = 1$

$z_i \overset{iid}{\sim} N(0,1)$

$z_i^2 \overset{iid}{\sim} \chi^2(1)$

Thus, as $n \to \infty$,

$P_r\{ AIC_{K_0+L} < AIC_{K_0} \} = P_r\{ \chi^2_L > 2L \}$.