Lecture Stat 461-561
Information Criterion

Arnaud Doucet

February 2008

Review of Maximum Likelihood Approach

- We have data $X_i \overset{\text{i.i.d.}}{\sim} g(x)$.
- We model the distribution of these data by a parametric model $\{f(x|\theta); \theta \in \Theta \subseteq \mathbb{R}^p\}$.
- We estimate θ by ML; that is

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} l(\theta) \text{ where } l(\theta) := \sum_{k=1}^{n} \log f(X_k|\theta)$$

- We know that as $n \to \infty$ then $\widehat{\theta}$ converges towards

$$\theta^* = \arg\min_{\theta \in \Theta} KL(g(x); f(x|\theta))$$

where

$$KL(g(x); f(x|\theta)) := \int g(x) \log \frac{g(x)}{f(x|\theta)} dx$$

A Central Limit Theorem

- Assuming the MLE is asymptotically consistent, then we have

$$\mathbf{0} = \left.\frac{\partial l(\theta)}{\partial \theta}\right|_{\widehat{\theta}} = \left.\frac{\partial l(\theta)}{\partial \theta}\right|_{\theta^*} + \left.\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}\right|_{\theta^*} \left(\widehat{\theta} - \theta^*\right) + \ldots$$

- The law of large numbers yields

$$-\frac{1}{n}\left.\frac{\partial l(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}\right|_{\theta^*} \rightarrow \left(J(\theta^*) := -\int g(x) \left.\frac{\partial^2 \log f(x|\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}\right|_{\theta^*} dx\right)$$

- The CLT yields

$$\sqrt{n}\frac{1}{n}\left.\frac{\partial l(\theta)}{\partial \theta}\right|_{\theta^*} \xrightarrow{\mathrm{D}} \mathcal{N}(0, I(\theta^*))$$

where

$$I(\theta^*) = \int g(x) \left.\frac{\partial \log f(x|\theta)}{\partial \theta}\right|_{\theta^*} \left.\frac{\partial \log f(x|\theta)}{\partial \theta^{\mathrm{T}}}\right|_{\theta^*} dx$$

- Using Slutzky's theorem we have

$$\sqrt{n}\left(\widehat{\theta}-\theta^{*}\right) \xrightarrow{\mathrm{D}} \mathcal{N}\left(0, J^{-1}\left(\theta^{*}\right) I\left(\theta^{*}\right) J^{-1}\left(\theta^{*}\right)\right).$$

- This is sometimes known as the sandwitch variance.
- When $g(x)=f\left(x \middle| \theta^{*}\right)$ then we have

$$J\left(\theta^{*}\right)=I\left(\theta^{*}\right)$$

and the 'standard' CLT follows

$$\sqrt{n}\left(\widehat{\theta}-\theta^{*}\right) \xrightarrow{\mathrm{D}} \mathcal{N}\left(0, I^{-1}\left(\theta^{*}\right)\right)$$

Evaluating The Statistical Model

- We want to compare our model $f\left(x|\widehat{\theta}\right)$ to $g(x)$.
- One way consists of evaluating

$$KL\left(g(x);f\left(x|\widehat{\theta}\right)\right)=\int g(x)\log g(x)\,dx-\int g(x)\log f\left(x|\widehat{\theta}\right)dx.$$

  and the larger $\int g(x)\log f\left(x|\widehat{\theta}\right)dx$, the closer the model is to the true one.
- The crucial issue is to obtain an estimate of

$$\mathbb{E}_g\left[\log f\left(X|\widehat{\theta}\right)\right]=\int g(x)\log f\left(x|\widehat{\theta}\right)dx.$$

- Clearly an estimate of $\mathbb{E}_g\left[\log f\left(X|\widehat{\theta}\right)\right]$ is $n^{-1}I\left(\widehat{\theta}\right)$ and an estimate of $n\mathbb{E}_g\left[\log f\left(X|\widehat{\theta}\right)\right]$ is $I\left(\widehat{\theta}\right)$.

Selecting a Model

- In practice, it is difficult to precisely capture the true structure of given phenomena from a limited number of data.
- Often several candidate statistical models are selected and we want to select the model that closely approximates the true distribution of the data.
- Given m candidate models $\left\{f_i\left(x\middle|\theta_i\right); i=1,...,m\right\}$ and the associated ML estimates $\widehat{\theta}_i$, a simple solution would consist of selecting the model j where

$$j = \underset{i\in\{1,...,m\}}{\arg\max}\, l_i\left(\widehat{\theta}_i\right);$$

i.e. picking the model for which $KL\left(g\left(x\right); f_i\left(x\middle|\widehat{\theta}_i\right)\right)$ is the smallest.

Be Careful!

- This approach does not provide a fair comparison of models.
- The quantity $l_i\left(\widehat{\theta}_i\right)$ contains a bias as an estimator of $n\mathbb{E}_g\left[\log f_i\left(X|\widehat{\theta}_i\right)\right]$.
- The reason is that the data $X_1, ..., X_n$ are used twice: to estimate $\widehat{\theta}_i$ and to approximate the expectation with respect to g.
- We will show that the size of the bias depends on the size p of the parameter θ.

Relationship between log-likelihood and expected log-likelihood

- Let us denote $\theta_i^* = \arg\min KL\left(g\left(x\right); f_i\left(x|\,\theta_i\right)\right)$ the 'true' parameter.
- We necessarily have

$$\mathbb{E}_g\left[\log f_i\left(X|\,\widehat{\theta}_i\right)\right] \leq \mathbb{E}_g\left[\log f_i\left(X|\,\theta_i^*\right)\right]$$

whereas

$$l_i\left(\widehat{\theta}_i\right) \geq l_i\left(\theta_i^*\right).$$

- To correct for this bias, we want to estimate

$$b_i\left(g\right) = \mathbb{E}_g\left[l_i\left(\widehat{\theta}_i\right) - n\mathbb{E}_g\left[\log f_i\left(X|\,\widehat{\theta}_i\right)\right]\right]$$

where remember that

$$l_i\left(\widehat{\theta}_i\right) = \sum_{k=1}^{n} \log f_i\left(X_k|\,\widehat{\theta}_i\right)$$

and $\widehat{\theta}_i = \widehat{\theta}_i\left(X_{1:n}\right)$.

Information Criterion

- The information criterion for model i is defined as

$$\mathrm{IC}(i) = -2\sum_{k=1}^{n} \log f_i\left(X_k | \widehat{\theta}_i\right) + 2\left\{\text{estimator for } b_i(g)\right\}$$

- $\mathrm{IC}(i)$ is a biased-corrected estimate of minus the expected log-likelihood.
- So given a collection of models, we will select the model

$$j = \underset{i \in \{1,\ldots,m\}}{\arg\min}\ \mathrm{IC}(i)$$

Derivation of Bias

- We want to estimate the bias $b(g)$ which is given by

$$
\begin{aligned}
& \mathbb{E}_g\left[l\left(\widehat{\theta}\left(X_{1:n}\right)\right)-n\mathbb{E}_g\left[\log f\left(X|\widehat{\theta}\left(X_{1:n}\right)\right)\right]\right] \\
= \; & \underbrace{\mathbb{E}_g\left[l\left(\widehat{\theta}\left(X_{1:n}\right)\right)-l\left(\theta^*\right)\right]}_{D_1} \\
& +\underbrace{\mathbb{E}_g\left[l\left(\theta^*\right)-n\mathbb{E}_g\left[\log f\left(X|\theta^*\right)\right]\right]}_{D_2} \\
& +\underbrace{\mathbb{E}_g\left[n\mathbb{E}_g\left[\log f\left(X|\theta^*\right)\right]-n\mathbb{E}_g\left[\log f\left(X|\widehat{\theta}\left(X_{1:n}\right)\right)\right]\right]}_{D_3}
\end{aligned}
$$

Calculation of second term

- We have

$$\begin{aligned} D_2 &= \mathbb{E}_g \left[l\left(\theta^*\right) - n\mathbb{E}_g \left[\log f\left(X \middle| \theta^*\right)\right]\right] \\ &= \mathbb{E}_g \left[\sum_{k=1}^{n} \log f\left(X_k \middle| \theta^*\right) - n\mathbb{E}_g \left[\log f\left(X \middle| \theta^*\right)\right]\right] \\ &= 0 \end{aligned}$$

- No bias for this term...

Calculation of third term

- Let us denote
$$\eta\left(\widehat{\theta}\right) = \mathbb{E}_g\left[\log f\left(X|\widehat{\theta}\right)\right].$$
- By performing a Taylor expansion around θ^*, we obtain
$$\eta\left(\widehat{\theta}\right) = \eta\left(\theta^*\right) + \left.\frac{\partial \eta\left(\theta\right)}{\partial \theta}\right|_{\theta^*}^{\mathsf{T}}\left(\widehat{\theta}-\theta^*\right) + \frac{1}{2}\left(\widehat{\theta}-\theta^*\right)^{\mathsf{T}}\left.\frac{\partial^2 \eta\left(\theta\right)}{\partial\theta\partial\theta^{\mathsf{T}}}\right|_{\theta^*}\left(\widehat{\theta}-\theta^*\right.$$
- As $\left.\frac{\partial \eta(\theta)}{\partial \theta}\right|_{\theta^*} = \mathbf{0}$ then
$$\eta\left(\widehat{\theta}\right) \approx \eta\left(\theta^*\right) - \frac{1}{2}\left(\widehat{\theta}-\theta^*\right)^{\mathsf{T}} J\left(\theta^*\right)\left(\widehat{\theta}-\theta^*\right)$$
where
$$J\left(\theta^*\right) = -\left.\frac{\partial^2 \eta\left(\theta\right)}{\partial\theta\partial\theta^{\mathsf{T}}}\right|_{\theta^*} = -\int g\left(x\right)\left.\frac{\partial^2 \log f\left(x|\,\theta\right)}{\partial\theta\partial\theta^{\mathsf{T}}}\right|_{\theta^*} dx$$

- It follows that

$$\begin{aligned} D_3 &= \mathbb{E}_g\left[n\left(\eta\left(\theta^*\right)-\eta\left(\widehat{\theta}\right)\right)\right] \\ &\approx \frac{n}{2}\mathbb{E}_g\left[\left(\widehat{\theta}-\theta^*\right)^{\mathsf{T}} J\left(\theta^*\right)\left(\widehat{\theta}-\theta^*\right)\right] \\ &= \frac{n}{2}\mathbb{E}_g\left[\mathrm{tr}\left\{J\left(\theta^*\right)\left(\widehat{\theta}-\theta^*\right)\left(\widehat{\theta}-\theta^*\right)^{\mathsf{T}}\right\}\right] \\ &= \frac{n}{2}\mathrm{tr}\left\{J\left(\theta^*\right)\mathbb{E}_g\left[\left(\widehat{\theta}-\theta^*\right)\left(\widehat{\theta}-\theta^*\right)^{\mathsf{T}}\right]\right\} \end{aligned}$$

- We have from the CLT established previously we have

$$\mathbb{E}_g\left[\left(\widehat{\theta}-\theta^*\right)\left(\widehat{\theta}-\theta^*\right)^{\mathsf{T}}\right] \approx \frac{1}{n} J\left(\theta^*\right)^{-1} I\left(\theta^*\right) J\left(\theta^*\right)^{-1}$$

so

$$D_3 \approx \frac{1}{2}\mathrm{tr}\left\{I\left(\theta^*\right) J\left(\theta^*\right)^{-1}\right\}$$

Calculation of first term

- We have

$$l(\theta) = l\left(\widehat{\theta}\right) + \left.\frac{\partial l(\theta)}{\partial \theta}\right|_{\widehat{\theta}}^{\mathsf{T}} \left(\theta - \widehat{\theta}\right) + \frac{1}{2}\left(\theta - \widehat{\theta}\right)^{\mathsf{T}} \left.\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^{\mathsf{T}}}\right|_{\widehat{\theta}} \left(\theta - \widehat{\theta}\right) + \ldots$$

  so

$$l(\theta^*) - l\left(\widehat{\theta}\right) \approx -\frac{n}{2}\left(\theta^* - \widehat{\theta}\right)^{\mathsf{T}} J(\theta^*)\left(\theta^* - \widehat{\theta}\right)$$

-

$$\begin{aligned} D_1 &= \mathbb{E}_g\left[l\left(\widehat{\theta}(X_{1:n})\right) - l(\theta^*)\right] \\ &\approx \frac{n}{2}\mathbb{E}_g\left[\left(\theta^* - \widehat{\theta}\right)^{\mathsf{T}} J(\theta^*)\left(\theta^* - \widehat{\theta}\right)\right] \\ &= \frac{n}{2}\mathbb{E}_g\left[J(\theta^*)\,\mathrm{tr}\left\{\left(\theta^* - \widehat{\theta}\right)^{\mathsf{T}}\left(\theta^* - \widehat{\theta}\right)\right\}\right] \\ &= \frac{1}{2}\mathrm{tr}\left(I(\theta^*)\,J(\theta^*)^{-1}\right) \end{aligned}$$

Estimate of the Bias

- We have

$$\begin{aligned} b(g) &= D_1 + D_2 + D_3 \\ &\approx \frac{1}{2}\mathrm{tr}\left\{I(\theta^*) J(\theta^*)^{-1}\right\} + 0 + \frac{1}{2}\mathrm{tr}\left\{I(\theta^*) J(\theta^*)^{-1}\right\} \\ &= \mathrm{tr}\left\{I(\theta^*) J(\theta^*)^{-1}\right\} \end{aligned}$$

- Let $\widehat{I}$ and $\widehat{J}$ be consistent estimate of $I(\theta^*)$ and $J(\theta^*)$, say

$$\begin{aligned} \widehat{I} &= \frac{1}{n}\sum_{k=1}^{n} \left.\frac{\partial \log f(X_k|\theta)}{\partial \theta}\right|_{\widehat{\theta}} \left.\frac{\partial \log f(X_k|\theta)}{\partial \theta^{\mathrm{T}}}\right|_{\widehat{\theta}}, \\ \widehat{J} &= -\frac{1}{n}\sum_{k=1}^{n} \left.\frac{\partial^2 \log f(X_k|\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}\right|_{\widehat{\theta}} \end{aligned}$$

  then we can estimate the bias through

$$\widehat{b}(g) = \mathrm{tr}\left(\widehat{I}\widehat{J}^{-1}\right)$$

Information Criterion

- We have for the information criterion for model i

$$IC(i) = -2\sum_{k=1}^{n} \log f_i\left(X_k | \widehat{\theta}_i\right) + 2\mathrm{tr}\left(\widehat{I}_i \widehat{J}_i^{-1}\right)$$

- Assuming $g(x) = f_i(x|\theta_i^*)$ then $I_i(\theta_i^*) = J_i(\theta_i^*)^{-1}$ so $b_i(g) = p_i$ (where p_i is the dimension of θ_i) so in this case

$$AIC(i) = -2\sum_{k=1}^{n} \log f_i\left(X_k | \widehat{\theta}_i\right) + 2p_i.$$

Checking the Equality of Two Discrete Distributions

- Assume we have two sets of data each having k categories and

Category	1	2	$\cdots$	k
Data set 1	n_1	n_2	$\cdots$	n_k
Data set 2	m_1	m_2	$\cdots$	m_k

  where we have $n_1 + \cdots + n_k = n$ observations for the first dataset and $m_1 + \cdots + m_k = m$ for the second.

- We assume these datasets follow the multinomial distributions

$$p\left(n_1, \cdots, n_k \middle| p_1, \cdots, p_k\right) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k},$$

$$p\left(m_1, \cdots, m_k \middle| q_1, \cdots, q_k\right) = \frac{m!}{m_1! \cdots m_k!} q_1^{m_1} \cdots q_k^{m_k}$$

- We want to check whether $p_1, \cdots, p_k \neq q_1, \cdots, q_k$ or $p_1, \cdots, p_k = q_1, \cdots, q_k$

- Assume the two distributions are different then

$$\begin{aligned} l\left(p_{1:k}, q_{1:k}\right) &= \log n! - \sum_{j=1}^{k}\left(\log n_j! + n_j \log p_j\right) \\ &\quad + \log m! - \sum_{j=1}^{k}\left(\log m_j! + m_j \log q_j\right) \end{aligned}$$

- We have the MLE

$$\widehat{p}_j = \frac{n_j}{n},\ \widehat{q}_j = \frac{m_j}{n}$$

- So

$$l\left(\widehat{p}_{1:k}, \widehat{q}_{1:k}\right) = C + \sum_{j=1}^{k}\left(n_j \log \frac{n_j}{n} + m_j \log \frac{m_j}{n}\right)$$

with $C = \log n! + \log m! - \sum_{j=1}^{k}\left(\log n_j! + \log m_j!\right)$ and

$$AIC = -2\left(C + \sum_{j=1}^{k}\left(n_j \log \frac{n_j}{n} + m_j \log \frac{m_j}{n}\right)\right) + 4\left(k-1\right)$$

- If we assume the two distributions are equal then

$$l(r) = C - \sum_{j=1}^{k} (n_j + m_j) \log r_j$$

and

$$\widehat{r}_j = \frac{n_j + m_j}{n + m}$$

- We have

$$l(\widehat{r}_{1:k}) = C + \sum_{j=1}^{k} (n_j + m_j) \log \left(\frac{n_j + m_j}{n + m} \right)$$

and

$$AIC = -2 \left(C + \sum_{j=1}^{k} (n_j + m_j) \log \left(\frac{n_j + m_j}{n + m} \right) \right) + 2(k - 1)$$

Example

- Consider the following data

Category	1	2	3	4	5
Data set 1	304	800	400	57	323
Data set 2	174	509	362	80	214

- From this table we can deduce

Category	1	2	3	4	5
$\widehat{p}_j$	0.16	0.42	0.21	0.03	0.17
$\widehat{q}_j$	0.13	0.38	0.27	0.06	0.16
$\widehat{r}_j$	0.15	0.41	0.24	0.04	0.17

- Ignoring the constant C, the maximum log-likelihood of the models are

$$\text{Model 1} \quad : \quad \sum_{j=1}^{k} \left(n_j \log \frac{n_j}{n} + m_j \log \frac{m_j}{n} \right) = -4567.36,$$

$$\text{Model 2} \quad : \quad \sum_{j=1}^{k} (n_j + m_j) \log \left(\frac{n_j + m_j}{n + m} \right) = -4585.61.$$

- The number of free parameters of the models is $2\,(k-1) = 8$ in model 1 and 4 in model 2. So it follows that AIC for model 1 and model 2 is respectively 9150,73 and 9179,22 so AIC selects Model 1.

Determining the number of bin size of an histogram

- Histograms are use for representing the properties of a set of observations obtained from either a discrete or continuous distribution.
- Assume we have a histogram $\{n_1, n_2, ..., n_k\}$ where k is refer to as the bin size.
- If k is too large or too small, then it is difficult to capture the characteristics of the true distribution.
- We can think of an histogram as a model specified by a multinomial distribution

$$p\left(n_1, \cdots, n_k \middle| p_1, \cdots, p_k\right) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}$$

where $n_1 + \cdots + n_k = n$, $p_1 + \cdots + p_k = 1$.

- In this case, we have seen that the MLE is $\widehat{p}_j = \frac{n_j}{n}$ and

$$l\left(\widehat{p}_{1:k}\right) = C + \sum_{j=1}^{k} n_j \log \frac{n_j}{n}$$

  with $C = \log n! - \sum_{j=1}^{k} \log n_j!$

- We have $k-1$ free parameters so

$$AIC = -2\left\{ C + \sum_{j=1}^{k} n_j \log \frac{n_j}{n} \right\} + 2\left(k-1\right).$$

- We want to compare this model to a model with lower resolutions.

- To compare the histogram model with a simpler one, we may assume $p_{2j-1} = p_{2j}$ for $j = 1, ...m$ and for sake of simplicity, we consider here $k = 2m$. In this case, the new MLE is $\widehat{p}_{2j-1} = \widehat{p}_{2j} = \frac{n_{2j-1}+n_{2j}}{2n}$
- In this case, the AIC is given by

$$AIC = -2\left\{ C + \sum_{j=1}^{m} (n_{2j-1} + n_{2j}) \log \frac{n_{2j-1} + n_{2j}}{2n} \right\} + 2(m-1).$$

- Similarly we can consider the histogram with m bins where $k = 4m$.
- We apply this to galaxy data for which AIC selects 14

bin size	28	14	7
log-like	-189.19	-197.71	-209.52
AIC	432.38	421.43	431.03

Application to Polynomial Regression

- We are given n observations $\{x_i, y_i\}$.
- We want to use the following model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \varepsilon, \ \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

  where k is unknown.
- We have $\theta = \left(\beta_{0:k}, \sigma^2\right)$ and

$$l\left(\theta\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}\left(y_j - \left(\sum_{m=0}^{k}\beta_m x_j^m\right)\right)^2$$

- We can easily establish that

$$l\left(\widehat{\theta}\right) = -\frac{n}{2}\log\left(2\pi\widehat{\sigma}^2\right) - \frac{n}{2}$$

Experimental Results

- In this case for the polynomial model of order k we have $k+2$ unknown parameters so

$$AIC(k) = n(\log 2\pi + 1) + n \log \widehat{\sigma}^2 + 2(k+2)$$

- This yields

k	0	1	2	3	4	5	6	7
$l\left(\widehat{\theta}_k\right)$	22.41	31.19	41.51	42.52	43.75	44.44	45.00	45.45
AIC	-40.81	-56.38	-75.03	-75.04	-75.50	-74.89	-74.00	-72.89

- AIC selects model 4.

Application to Daily Temperature Data

- Daily minimum temperatures y_i in January averaged from 1971 through 2000 for city i
- x_{i1} latitudes, x_{i2} longitudes and x_{i3} altitudes.
- A standard model is

$$y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + a_3 x_{i3} + \varepsilon_i$$

  where $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$.
- But we would also like to compare models where some of the explanatory variables are omitted.

Results

- We have

$$AIC\,(\text{model}) = n\,(\log 2\pi + 1) + n \log \widehat{\sigma}^2 + 2\,(\text{nb. explanatority var.} + 2)$$

- This yields

model	x_1, x_3	x_1, x_2, x_3	x_1, x_2	x_1	x_2, x_3	x_2	x_3
$\widehat{\sigma}^2$	1.49	1.48	5.11	5.54	5.69	7.81	19.96
AIC	88.92	90.81	119.71	119.73	122.43	128.35	151.88

- We select the model

$$y_i = 40.49 - 1.11 x_{i1} - 0.010 x_{i3} + \varepsilon_i$$

with $\varepsilon_i \sim \mathcal{N}\,(0, 1.49)$.

Selection of Order of Autoregressive Model

- We have the following model

$$y_k = \sum_{i=1}^{m} a_i y_{k-i} + \varepsilon_k, \quad \varepsilon_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right).$$

- Assuming we have data $y_1, y_2, ..., y_n$ and that $y_0, y_{-1}, ..., y_{1-m}$ are deterministic then we can easily obtain the MLE estimate and check that

$$l\left(\widehat{\theta}\right) = -\frac{n}{2}\log\left(2\pi\widehat{\sigma}^2\right) - \frac{n}{2}$$

- Since the model with order m has $m+1$ unknown parameters then

$$AIC\left(m\right) = n\left(\log 2\pi + 1\right) + n\log\widehat{\sigma}^2 + 2\left(m+1\right)$$

Application to Canadian Lynx Data

- This corresponds to the logarithms of the annual numbers of lynx trapped from 1821 to 1934; $n = 114$ observations.
- We try 20 candidate models and $m = 11$ is selected using AIC.

Detection of Structural Changes

- We some times encounter the situation in which the stochastic structure of the data changes at a certain time or location.
- Let us start with a simple model where

$$Y_n \sim \mathcal{N}\left(\mu_n, \sigma^2\right)$$

and

$$\mu_n = \begin{cases} \theta_1 & \text{if } n < k \\ \theta_2 & \text{if } n \geq k \end{cases}$$

where k is the unknown so-called changepoint.

- We have N data and, for the model k, the likelihood is

$$L\left(\theta_1, \theta_2, \sigma^2\right) = \prod_{n=1}^{k-1} \mathcal{N}\left(y_n; \theta_1, \sigma^2\right) \prod_{n=k}^{N} \mathcal{N}\left(y_n; \theta_2, \sigma^2\right)$$

- It can be easily established that

$$\begin{aligned}\widehat{\theta}_1 &= \frac{1}{k-1}\sum_{n=1}^{k-1} y_n,\ \widehat{\theta}_2 = \frac{k}{N-k+1}\sum_{n=k}^{N} y_n, \\ \widehat{\sigma}^2 &= \frac{1}{N}\left\{\sum_{n=1}^{k-1}\left(y_n-\widehat{\theta}_1\right)^2 + \sum_{n=k}^{N}\left(y_n-\widehat{\theta}_2\right)^2\right\}\end{aligned}$$

- So we have the maximum log-likelihood given by

$$l\left(\widehat{\theta}_1,\widehat{\theta}_2,\widehat{\sigma}^2\right) = -\frac{N}{2}\log\left(2\pi\widehat{\sigma}^2\right) - \frac{N}{2}$$

  where we emphasize that $\widehat{\sigma}^2$ is a function of k,
- The AIC criterion is given by

$$AIC = N\log\left(2\pi\widehat{\sigma}^2\right) + N + 2\times 3$$

  and the changepoint k can be automatically determined by finding the value of k that gives the smallest AIC.

Application to Factor Analysis

- Suppose we have $x = (x_1, ..., x_p)^{\mathrm{T}}$ a vector of mean μ and variance-covariance Σ.
- The factor analysis model is

$$x = \mu + Lf + \varepsilon$$

where L is a $p \times m$ matrix of factor loadings whereas $f = (f_1, ..., f_m)^{\mathrm{T}}$ and $\varepsilon = (\varepsilon_1, ..., \varepsilon_p)^{\mathrm{T}}$ are unobserved random vectors assumed to satisfy

$$\begin{aligned} \mathbb{E}[f] &= 0,\ Cov[f] = I_m, \\ \mathbb{E}[\varepsilon] &= 0,\ Cov[\varepsilon] = \Psi = \mathrm{diag}(\Psi_1, ..., \Psi_p), \\ Cov[f, \varepsilon] &= 0. \end{aligned}$$

- It then follows that

$$\Sigma = LL^{\mathrm{T}} + \Psi.$$

- Let S denote the sample covariance.
- Under a normal assumption for f and ε, it can be shown that the ML estimates of L and Ψ can be obtained by minimizing

$$\log|\Sigma| - \log|S| + \operatorname{tr}\left(\Sigma^{-1}S\right) - p$$

  subject to $L^{\top}\Psi^{-1}L$ being a diagonal matrix.
- In this case, we have

$$\begin{aligned} AIC(m) &= n\left\{p\log(2\pi) + \log\left|\widehat{\Sigma}\right| + \operatorname{tr}\left(\Sigma^{-1}S\right)\right\} \\ &\quad +2\left\{p(m+1) - \frac{1}{2}m(m-1)\right\} \end{aligned}$$