Consistency of Model Order Selection Rules

Niklas Wahlström

May 20, 2011

These notes present techniques to be used for analyzing the consistency of model order selection rules. The discussion is focused on the array model and the Bayesian Information Criteria, as it is presented in [4]. However, the analysis can be extended to more general models as well as other model order selection rules, which will be discussed in the lecture.

Furthermore, in the Appendix, the asymptotic distribution of a likelihood ratio is presented and proved, inspired by the treatment in [3].

1 Problem formulation

In array processing, observations are described by the model

$$\mathbf{x}(t) = \sum_{i=1}^{q} \mathbf{a}(\boldsymbol{\theta}_i) s_i(t) + \mathbf{n}(t) \tag{1}$$

where $s_i(t) \in \mathbb{C}$ is jth emitters signal, $\mathbf{a}(\boldsymbol{\theta}_i) \in \mathbb{C}^p$ the sensor response and $\mathbf{n}(t) \in \mathbb{C}^p$ additive noise. In matrix form the following familiar equation is obtained

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \tag{2}$$

where

$$\mathbf{A} = [\mathbf{a}(\boldsymbol{\theta}_1), \cdots \mathbf{a}(\boldsymbol{\theta}_q)]$$

. We assume that

- $q < p$, i.e. that there are more sensors than emitted signals
- $\mathbf{s}(t)$ and $\mathbf{n}(t)$ are stationary, white, zero mean, circular symmetric Gaussian random processes. Especially we assume that $\mathbf{S} = \mathbb{E}[\mathbf{s}(t)\mathbf{s}^*(t)]$ is positive definite and that $\mathbb{E}[\mathbf{n}(t)\mathbf{n}^*(t)] = \sigma^2\mathbf{I}$.
- $\mathbf{A}$ has full rank.

The problem is to detect the number of signals q from the observations $\mathbf{x}(t_1), \cdots, \mathbf{x}(t_N)$.

From the model description we can compute the covariance matrix of $\mathbf{x}(t)$ as

$$\mathbf{R} = \boldsymbol{\Psi} + \sigma^2\mathbf{I} \tag{3}$$

where

$$\mathbf{\Psi} = \mathbf{A}\mathbf{S}\mathbf{A}^* \tag{4}$$

Since $\mathbf{S}$ is a positive definite symmetric matrix we have

$$\begin{aligned} \operatorname{rank}(\mathbf{\Psi}) &= \operatorname{rank}(\mathbf{A}\mathbf{S}\mathbf{A}^*) = \operatorname{rank}\big(\mathbf{A}\sqrt{\mathbf{S}}(\mathbf{A}\sqrt{\mathbf{S}})^*\big) && (5)\\ &= \operatorname{rank}(\mathbf{A}\sqrt{\mathbf{S}}) = \operatorname{rank}(\mathbf{A}) = \min(p,q) = q && (6) \end{aligned}$$

Thus, $\mathbf{\Psi}$ has exactly q non-zero eigenvalues. This gives the following eigenvalues of $\mathbf{R}$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q > \underbrace{\lambda_{q+1} = \cdots = \lambda_p}_{=\sigma^2} \tag{7}$$

The number of signals q can hence be determined from the multiplicity of the smallest eigenvalue of $\mathbf{R}$.

2 Estimation

From previous section we know that the observations $\mathbf{x}(t_1), \cdots, \mathbf{x}(t_N)$ are independent Gaussian random vectors with the covariance matrix (3). For estimation, we therefore consider

$$\mathbf{R}^{(k)} = \mathbf{\Psi}^{(k)} + \sigma^2\mathbf{I} \tag{8}$$

where $\mathbf{\Psi}^{(k)}$ is a positive semi-definite matrix of rank k. With a spectral representation of $\mathbf{R}^{(k)}$ we get

$$\mathbf{R}^{(k)} = \sum_{i=1}^{k}(\lambda_i - \sigma^2)\mathbf{V}_i\mathbf{V}_i^* + \sigma^2\mathbf{I} \tag{9}$$

and $\mathbf{R}^{(k)}$ can thus be parametrized with

$$\mathbf{\Theta}^{(k)} = (\lambda_1, \cdots, \lambda_k, \sigma^2, \mathbf{V}_1^T, \cdots, \mathbf{V}_k^T)^T \tag{10}$$

Now, we have

$$f(\mathbf{x}(t_1), \cdots, \mathbf{x}(t_N)|\mathbf{\Theta}^{(k)}) = \prod_{i=1}^{N}\frac{1}{\pi^p|\mathbf{R}^{(k)}|}\exp\big(-\mathbf{x}(t_i)^*(\mathbf{R}^{(k)})^{-1}\mathbf{x}(t_i)\big) \tag{11}$$

and the log-likelihood function then becomes

$$\begin{aligned} \log l(\mathbf{\Theta}^{(k)}) &= \sum_{i=1}^{N}\big(-\log(|\mathbf{R}^{(k)}|) - \mathbf{x}(t_i)^*(\mathbf{R}^{(k)})^{-1}\mathbf{x}(t_i)\big) + \text{const.}\\ &= -N\log(|\mathbf{R}^{(k)}|) - \sum_{i=1}^{N}\operatorname{tr}\Big(\mathbf{x}(t_i)^*(\mathbf{R}^{(k)})^{-1}\mathbf{x}(t_i)\Big) + \text{const.}\\ &= -N\log(|\mathbf{R}^{(k)}|) - \operatorname{tr}\left((\mathbf{R}^{(k)})^{-1}\sum_{i=1}^{N}\mathbf{x}(t_i)\mathbf{x}(t_i)^*\right) + \text{const.}\\ &= -N\log\big|\mathbf{R}^{(k)}\big| - N\operatorname{tr}\big((\mathbf{R}^{(k)})^{-1}\hat{\mathbf{R}}\big) + \text{const.} \end{aligned} \tag{12}$$

where $\hat{\mathbf{R}}$ is the sample-covariance matrix

$$\hat{\mathbf{R}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}(t_i)\mathbf{x}(t_i)^* \tag{13}$$

It can be shown that the MLE of $\boldsymbol{\Theta}^{(k)}$ is

$$\hat{\lambda}_i = l_i \quad i = 1, \cdots, k \tag{14a}$$

$$\hat{\sigma}^2 = \frac{1}{p-k}\sum_{i=k+1}^{p} l_i \tag{14b}$$

$$\hat{\mathbf{V}}_i = \mathbf{C}_i \quad i = 1, \cdots, k \tag{14c}$$

where $l_1 > l_2 \cdots > l_p$ and $\mathbf{C}_1 . \cdots \mathbf{C}_p$ are eigenvalues and eigenvectors of $\hat{\mathbf{R}}$. Now, from the MLE in (14) we can construct

$$\hat{\mathbf{R}}^{(k)} = \sum_{i=1}^{k}(\hat{\lambda}_i - \hat{\sigma}^2)\hat{\mathbf{V}}_i\hat{\mathbf{V}}_i^* + \hat{\sigma}^2\mathbf{I} \tag{15}$$

By substituting (14) and (15) into (12) then gives

$$\log l(\hat{\boldsymbol{\Theta}}^{(k)}) = -N\log\left|\hat{\mathbf{R}}^{(k)}\right| - N\mathrm{tr}\big((\hat{\mathbf{R}}^{(k)})^{-1}\hat{\mathbf{R}}\big) + \text{const.} \tag{16}$$

Since the eigenvalues of $\hat{\mathbf{R}}^{(k)}$ are $l_1, \cdots, l_k, \hat{\sigma}^2, \cdots \hat{\sigma}^2$, the first term in (16) will be

$$\begin{aligned}
\log\left|\hat{\mathbf{R}}^{(k)}\right| &= \log\left|\hat{\mathbf{R}}^{(k)}\right| \\
&= \log\left(\prod_{i=1}^{k} l_i \prod_{i=k+1}^{p}\hat{\sigma}^2\right) \\
&= \log\left(\prod_{i=1}^{k} l_i\right) + (p-k)\log\left(\hat{\sigma}^2\right) \\
&= -\log\left(\prod_{i=k+1}^{p} l_i\right) + \log\left(\prod_{i=1}^{p} l_i\right) + (p-k)\log\left(\hat{\sigma}^2\right) \\
&= -\log\left(\frac{\sqrt[p-k]{\prod_{i=k+1}^{p} l_i}}{\frac{1}{p-k}\sum_{i=k+1}^{p} l_i}\right)^{(p-k)} + \log\left(\prod_{i=1}^{p} l_i\right)
\end{aligned} \tag{17}$$

For the second term in (16) we have to compute the eigenvalues of $(\hat{\mathbf{R}}^{(k)})^{-1}\hat{\mathbf{R}}$. By construction, the eigenvectors of $\hat{\mathbf{R}}$ are also eigenvectors to $\hat{\mathbf{R}}^{(k)}$ as well as $(\hat{\mathbf{R}}^{(k)})^{-1}$ and the eigenvalues of $(\hat{\mathbf{R}}^{(k)})^{-1}\hat{\mathbf{R}}$ can therefore be computed by element-wise multiplication of the eigenvalues of the two matrices. This gives

$$\begin{aligned}
\mathrm{tr}\big((\hat{\mathbf{R}}^{(k)})^{-1}\hat{\mathbf{R}}\big) &= \left(\sum_{i=1}^{k}\frac{l_i}{l_i} + \sum_{k+1}^{p}\frac{l_i}{\hat{\sigma}^2}\right) \\
&= \left(k + \frac{1}{\frac{1}{p-k}\sum_{i=k+1}^{p} l_i}\sum_{i=k+1}^{p} l_i\right) = p
\end{aligned} \tag{18}$$

which gives

$$\log l(\hat{\mathbf{\Theta}}^{(k)}) = N \log \left(\frac{\sqrt[p-k]{\prod_{i=k+1}^{p} l_i}}{\frac{1}{p-k}\sum_{i=k+1}^{p} l_i} \right)^{(p-k)} - N \log \left(\prod_{i=1}^{p} l_i \right) - Np + \text{const.}$$

$$= N \log \left(\frac{\sqrt[p-k]{\prod_{i=k+1}^{p} l_i}}{\frac{1}{p-k}\sum_{i=k+1}^{p} l_i} \right)^{(p-k)} + \text{const.} \tag{19}$$

Now, the Bayesian Information Criterion Rule is given by minimizing the negative log-likelihood function with an extra penalty coefficient

$$\text{BIC}(k) = -2 \log l(\hat{\mathbf{\Theta}}^{(k)}) + n_k \log(N) + \text{const.} \tag{20}$$

where n_k is the dimension of the state (10).

3 Consistency

For consistency we demand when $N \to \infty$ that

$$P(\arg\min \text{BIC}(k) = q) = 1 \Longrightarrow \tag{21}$$

$$\begin{cases} P(\text{BIC}(k) > \text{BIC}(q) | k < q) = 1 & \text{no underfitting} \\ P(\text{BIC}(k) > \text{BIC}(q) | k > q) = 1 & \text{no overfitting} \end{cases} . \tag{22}$$

Thus, we can prove consistency by proving that the quantity

$$\text{BIC}(q) - \text{BIC}(k) = 2 \log \frac{l(\hat{\mathbf{\Theta}}^{(k)})}{l(\hat{\mathbf{\Theta}}^{(q)})} + (n_q - n_k) \log N \tag{23}$$

is negative w.p.1 both in case of underfitting $k < q$ as well in case of overfitting $k > q$.

3.1 Underfitting

Consider the case $k < q$ and divide the quantity (23) with $2N$

$$\frac{\text{BIC}(q) - \text{BIC}(k)}{2N} = \frac{1}{N} \log \frac{l(\hat{\mathbf{\Theta}}^{(k)})}{l(\hat{\mathbf{\Theta}}^{(q)})} + (n_q - n_k) \frac{\log N}{2N} \tag{24}$$

and consider the first term of (24)

$$\begin{aligned}
\frac{1}{N}\log\frac{l(\hat{\mathbf{\Theta}}^{(k)})}{l(\hat{\mathbf{\Theta}}^{(q)})} &= \log\left(\frac{\prod_{i=k+1}^{p} l_i}{\left(\frac{1}{p-k}\sum_{i=k+1}^{p} l_i\right)^{p-k}} \cdot \frac{\left(\frac{1}{p-q}\sum_{i=q+1}^{p} l_i\right)^{p-q}}{\prod_{i=q+1}^{p} l_i}\right) \\
&= \log\left(\frac{\prod_{i=k+1}^{q} l_i}{1} \cdot \frac{\left(\frac{1}{p-q}\sum_{i=q+1}^{p} l_i\right)^{p-q}}{\left(\frac{1}{p-k}\sum_{i=k+1}^{p} l_i\right)^{p-k}}\right) \\
&= \log\left(\frac{\prod_{i=k+1}^{q} l_i}{\left(\frac{1}{q-k}\sum_{i=k+1}^{q} l_i\right)^{q-k}}\right) \\
&\quad + \log\left(\frac{\left(\frac{1}{p-q}\sum_{i=q+1}^{p} l_i\right)^{p-q}\left(\frac{1}{q-k}\sum_{i=k+1}^{q} l_i\right)^{q-k}}{\left(\frac{1}{q-k}\sum_{i=k+1}^{q} l_i\right)^{q-k}}\right) \\
&= \log\left(\frac{\sqrt[q-k]{\prod_{i=k+1}^{q} l_i}}{\frac{1}{q-k}\sum_{i=k+1}^{q} l_i}\right)^{q-k} \\
&\quad + \log\left(\frac{\left(\frac{1}{p-q}\sum_{i=q+1}^{p} l_i\right)^{\frac{p-q}{p-k}}\left(\frac{1}{q-k}\sum_{i=k+1}^{q} l_i\right)^{\frac{q-k}{p-k}}}{\frac{1}{p-k}\sum_{i=k+1}^{p} l_i}\right)^{p-k} \qquad (25)
\end{aligned}$$

The first term of (25) is non-positive due to the AM-GM inequality

$$\frac{1}{q-k}\sum_{i=k+1}^{q} l_i \geq \sqrt[q-k]{\prod_{i=k+1}^{q} l_i} \qquad (26)$$

Note that, in contrast to what is stated in [4], this inequality *can* be fulfilled with equality. This is the case when all $\{l_i\}_{k+1}^{q}$ are equal, which especially is the case when $k+1=q$.

For the second term, the generalized AM-GM inequality can be applied

$$\sum_{i=1}^{n} w_i A_i \geq \prod A_i^{w_i} \qquad \sum_{i=1}^{n} w_i = 1 \qquad (27)$$

which is fulfilled with equality if and only if all A_i with $w_i > 0$ are equal. By identifying

$$A_1 = \frac{1}{p-q}\sum_{i=q+1}^{p} l_i, \quad A_2 = \frac{1}{q-k}\sum_{i=k+1}^{q} l_i, \quad w_1 = \frac{p-q}{p-k}, \quad w_2 = \frac{q-k}{p-k} \qquad (28)$$

we get

$$\overbrace{\frac{p-q}{p-k}\left(\frac{1}{p-q}\sum_{i=q+1}^{p} l_i\right)+\frac{q-k}{p-k}\left(\frac{1}{q-k}\sum_{i=k+1}^{q} l_i\right)}^{=\frac{1}{p-k}\sum_{i=k+1}^{p} l_i} > \left(\frac{1}{p-q}\sum_{i=q+1}^{p} l_i\right)^{\frac{p-q}{p-k}} \cdot \left(\frac{1}{q-k}\sum_{i=k+1}^{q} l_i\right)^{\frac{q-k}{p-k}} \quad (29)$$

This inequality *can not* be fulfilled with equality since

$$\frac{1}{p-q}\sum_{i=q+1}^{p} l_i - \frac{1}{q-k}\sum_{i=k+1}^{q} l_i \to \frac{1}{p-q}\sum_{i=q+1}^{p} \lambda_i - \frac{1}{q-k}\sum_{i=k+1}^{q} \lambda_i > 0$$

when $N \to \infty$. Thus, we have that the second term in (25) goes to a negative constant as $N \to \infty$. Since the last term of (24) goes to zero as $N \to \infty$, the whole difference (24) will be negative w.p.1 in the large-sample limit.

3.2 Overfitting

Now, we consider the case $k > q$. The first term of the right hand side in (23) can be considered as a *generalized likelihood ratio* (GLR). We can find the asymptotic distribution of this ratio by using Lemma A.2. First, we identify

$$\boldsymbol{\Theta}^{(q)} = \boldsymbol{\theta}_s, \qquad \boldsymbol{\Theta}^{(k)} = \begin{bmatrix} \boldsymbol{\theta}_r \\ \boldsymbol{\theta}_s \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\theta}_{r_0} = \mathbf{0}. \quad (30)$$

Furthermore, since q is the true model order, the correct hypothesis is

$$\mathcal{H}_0 : \boldsymbol{\theta}_r = \boldsymbol{\theta}_{r_0}.$$

According to Lemma A.2 we then have

$$2\log\frac{l(\hat{\boldsymbol{\Theta}}^{(k)})}{l(\hat{\boldsymbol{\Theta}}^{(q)})} \sim \chi^2_{n_k - n_q} \quad (31)$$

This gives

$$P(\mathrm{BIC}(q) - \mathrm{BIC}(k) < 0) = P(\chi^2_{n_k-n_q} < (n_k - n_q)\log N) \to 1 \quad \text{as} \quad N \to \infty \quad (32)$$

By this we have shown that both the probability of underfitting as well as the probability of overfitting goes to zero in the large sample limit and thus, BIC is consistent.

A Appendix

Lemma A.1. *Consider the symmetric matrix*

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \quad (33)$$

Then

$$\min_{\mathbf{y}} \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}^T \begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & \mathbf{C}\end{bmatrix} \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} = \mathbf{x}^T \boldsymbol{\Delta}_{\mathbf{C}} \mathbf{x} \tag{34}$$

where $\boldsymbol{\Delta}_{\mathbf{C}} = \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T$ *is the Schur complement of* $\mathbf{C}$ *in* $\mathbf{M}$

Proof. By using the block triangular factorization (see, e.g. [1])

$$\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & \mathbf{C}\end{bmatrix} = \begin{bmatrix}\mathbf{I} & \mathbf{B}\mathbf{C}^{-1}\\ \mathbf{0} & \mathbf{I}\end{bmatrix} \begin{bmatrix}\boldsymbol{\Delta}_{\mathbf{C}} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}\end{bmatrix} \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{C}^{-1}\mathbf{B}^T & \mathbf{I}\end{bmatrix} \tag{35}$$

we get

$$\min_{\mathbf{y}} \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}^T \begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & \mathbf{C}\end{bmatrix} \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} = \min_{\mathbf{y}} \mathbf{x}^T \boldsymbol{\Delta}_{\mathbf{C}} \mathbf{x} + \underbrace{(\mathbf{C}^{-1}\mathbf{B}^T\mathbf{x} - \mathbf{y})^T \mathbf{C} (\mathbf{C}^{-1}\mathbf{B}^T\mathbf{x} - \mathbf{y})}_{=0 \text{ when } \mathbf{y}=\mathbf{C}^{-1}\mathbf{B}^T\mathbf{x}} = \mathbf{x}^T \boldsymbol{\Delta}_{\mathbf{C}} \mathbf{x} \tag{36}$$

□

Lemma A.2. *Consider a PDF* $p(\mathbf{x};\boldsymbol{\theta})$ *where* $\boldsymbol{\theta}$ *is a vector of unknown parameters. Assume that the MLE of* $\boldsymbol{\theta}$ *is both consistent and efficient. Partition* $\boldsymbol{\theta}$ *as*

$$\boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\theta}_r\\ \boldsymbol{\theta}_s\end{bmatrix} = \begin{bmatrix}r \times 1\\ s \times 1\end{bmatrix} \tag{37}$$

and consider the parameter test

$$\begin{aligned}\mathcal{H}_0 &: \boldsymbol{\theta}_r = \boldsymbol{\theta}_{r_0}\\ \mathcal{H}_1 &: \boldsymbol{\theta}_r \neq \boldsymbol{\theta}_{r_0}\end{aligned}$$

as well as the test statistic

$$L_G(\mathbf{x}) = \frac{p(\mathbf{x}; \hat{\boldsymbol{\theta}}_{r_1}, \hat{\boldsymbol{\theta}}_{s_1})}{p(\mathbf{x}; \boldsymbol{\theta}_{r_0}, \hat{\boldsymbol{\theta}}_{s_0})} = \frac{\max_{\boldsymbol{\theta}_r, \boldsymbol{\theta}_s} p(\mathbf{x}; \boldsymbol{\theta}_r, \boldsymbol{\theta}_s)}{\max_{\boldsymbol{\theta}_s} p(\mathbf{x}; \boldsymbol{\theta}_r = \boldsymbol{\theta}_{r_0}, \boldsymbol{\theta}_s)}. \tag{38}$$

Then, as $N \to \infty$ *the statistic* $\log L_G(\mathbf{x})$ *has the the PDF*

$$2 \log L_G(\mathbf{x}) = \begin{cases}\chi^2_r & \text{under } \mathcal{H}_0\\ \chi^2_r(\lambda) & \text{under } \mathcal{H}_1\end{cases} \tag{39}$$

The noncentrality parameter is

$$\lambda = (\boldsymbol{\theta}_{r_1} - \boldsymbol{\theta}_{r_0})^T \boldsymbol{\Delta}_{\boldsymbol{\mathcal{I}}_{ss}} (\boldsymbol{\theta}_{r_1} - \boldsymbol{\theta}_{r_0}). \tag{40}$$

where $\boldsymbol{\Delta}_{\boldsymbol{\mathcal{I}}_{ss}} = \boldsymbol{\mathcal{I}}_{rr} - \boldsymbol{\mathcal{I}}_{rs}\boldsymbol{\mathcal{I}}_{ss}^{-1}\boldsymbol{\mathcal{I}}_{rs}^T$ *is the Schur complement of* $\boldsymbol{\mathcal{I}}_{ss}$ *in the information matrix*

$$\begin{bmatrix}\boldsymbol{\mathcal{I}}_{rr} & \boldsymbol{\mathcal{I}}_{rs}\\ \boldsymbol{\mathcal{I}}_{rs}^T & \boldsymbol{\mathcal{I}}_{ss}\end{bmatrix} = \boldsymbol{\mathcal{I}} = -\mathbb{E}\left(\frac{\partial^2 \log p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_1} \tag{41}$$

and where $\boldsymbol{\theta}_1 = [\boldsymbol{\theta}_{r_1}^T\ \boldsymbol{\theta}_{s_1}^T]^T$ *is the true value of* $\boldsymbol{\theta}$ *under* $\mathcal{H}_1$.

Proof. Since the MLE of $\boldsymbol{\theta}$ is efficient, we know that [2]

$$\frac{\partial \log p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \tag{42}$$

where $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ is the Fisher information matrix. Since MLE $\boldsymbol{\theta}$ is consistent we have

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}}) + \mathcal{O}(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|) \Longrightarrow \tag{43}$$

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \big(\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}}) + \mathcal{O}(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|)\big)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \tag{44}$$

$$= \boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \mathcal{O}(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2) \to \boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \tag{45}$$

as $N \to \infty$ since $\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\| \to 0$ due to consistency. Integrating and taking the exponential gives

$$p(\mathbf{x};\boldsymbol{\theta}) = K \exp\big[-\frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\big] =: K \exp\big(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})\big) \tag{46}$$

for some constant K. Note, that this can be seen as a Laplace approximation of the likelihood around its maximum $\hat{\boldsymbol{\theta}}$ and where the curvature is given by the information matrix. Furthermore, since $\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}}) \to \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ as $N \to \infty$, we will from here on skip the argument and simply referring to the information matrix as $\boldsymbol{\mathcal{I}}$.

By using the partitioning (37) and (41), we can write

$$p(\mathbf{x};\boldsymbol{\theta}_r,\boldsymbol{\theta}_s) = K \exp\left(\begin{bmatrix}\boldsymbol{\theta}_r \\ \boldsymbol{\theta}_s\end{bmatrix}; \begin{bmatrix}\hat{\boldsymbol{\theta}}_{r_1} \\ \hat{\boldsymbol{\theta}}_{s_1}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\mathcal{I}}_{rr} & \boldsymbol{\mathcal{I}}_{rs} \\ \boldsymbol{\mathcal{I}}_{rs}^T & \boldsymbol{\mathcal{I}}_{ss}\end{bmatrix}\right) \tag{47}$$

By this we evaluate the nominator of (39) as

$$p(\mathbf{x}; \hat{\boldsymbol{\theta}}_{r_1}, \hat{\boldsymbol{\theta}}_{s_1}) = K \tag{48}$$

For the denominator, Lemma A.1, which gives us

$$\begin{aligned} p(\mathbf{x}; \boldsymbol{\theta}_{r_0}, \hat{\boldsymbol{\theta}}_{s_0}) &= \max_{\boldsymbol{\theta}_s} K \exp\left(\begin{bmatrix}\boldsymbol{\theta}_{r_0} \\ \boldsymbol{\theta}_s\end{bmatrix}; \begin{bmatrix}\hat{\boldsymbol{\theta}}_{r_1} \\ \hat{\boldsymbol{\theta}}_{s_1}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\mathcal{I}}_{rr} & \boldsymbol{\mathcal{I}}_{rs} \\ \boldsymbol{\mathcal{I}}_{rs}^T & \boldsymbol{\mathcal{I}}_{ss}\end{bmatrix}\right) \\ &= K \exp(\boldsymbol{\theta}_{r_0}; \hat{\boldsymbol{\theta}}_{r_1}, \boldsymbol{\Delta}_{\boldsymbol{\mathcal{I}}_{ss}}) \end{aligned} \tag{49}$$

where $\boldsymbol{\Delta}_{\boldsymbol{\mathcal{I}}_{ss}} = \boldsymbol{\mathcal{I}}_{rr} - \boldsymbol{\mathcal{I}}_{rs}\boldsymbol{\mathcal{I}}_{ss}^{-1}\boldsymbol{\mathcal{I}}_{rs}^T$ is the Schur complement of $\boldsymbol{\mathcal{I}}_{ss}$ in $\boldsymbol{\mathcal{I}}$.

This leads to

$$\begin{aligned} 2\log L_G(\mathbf{x}) &= 2\log \frac{K}{K \exp(\boldsymbol{\theta}_{r_0}; \hat{\boldsymbol{\theta}}_{r_1}, \boldsymbol{\Delta}_{\boldsymbol{\mathcal{I}}_{ss}})} \\ &= (\hat{\boldsymbol{\theta}}_{r_1} - \boldsymbol{\theta}_{r_0})^T \boldsymbol{\Delta}_{\boldsymbol{\mathcal{I}}_{ss}} (\hat{\boldsymbol{\theta}}_{r_1} - \boldsymbol{\theta}_{r_0}). \end{aligned} \tag{50}$$

Since $\hat{\boldsymbol{\theta}}_{r1}$ is an efficient and unrestricted MLE of $\boldsymbol{\theta}_r$ it will attain the CRLB which can be computed as the corresponding sub-matrix of the inverse of the information matrix $(\boldsymbol{\mathcal{I}}^{-1})_{rr} = (\boldsymbol{\Delta}_{\boldsymbol{\mathcal{I}}_{ss}})^{-1}$. Furthermore, due to consistency, $\hat{\boldsymbol{\theta}}_{r1}$ will attain its true value in the large sample limit. Therefore, under $\mathcal{H}_0$ and as $N \to \infty$ we have

$$\hat{\boldsymbol{\theta}}_{r1} \sim \mathcal{N}(\boldsymbol{\theta}_{r_0}, (\boldsymbol{\Delta}_{\boldsymbol{\mathcal{I}}_{ss}})^{-1}) \tag{51}$$

By setting $\mathbf{z}_0 = \sqrt{\boldsymbol{\Delta}_{\boldsymbol{\mathcal{I}}_{ss}}}(\hat{\boldsymbol{\theta}}_{r_1} - \boldsymbol{\theta}_{r_0}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and we get under $\mathcal{H}_0$

$$2 \log L_G(\mathbf{x}) = \mathbf{z}_0^T \mathbf{z}_0 \sim \chi_r^2 \tag{52}$$

Under $\mathcal{H}_1$ we have

$$\hat{\boldsymbol{\theta}}_{r1} \sim \mathcal{N}(\boldsymbol{\theta}_{r_1}, (\boldsymbol{\Delta}_{\boldsymbol{\mathcal{I}}_{ss}})^{-1}) \tag{53}$$

By setting $\mathbf{z}_1 = \sqrt{\boldsymbol{\Delta}_{\boldsymbol{\mathcal{I}}_{ss}}}(\hat{\boldsymbol{\theta}}_{r_1} - \boldsymbol{\theta}_{r_1}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\mu}_1 = \sqrt{\boldsymbol{\Delta}_{\boldsymbol{\mathcal{I}}_{ss}}}(\boldsymbol{\theta}_{r_1} - \boldsymbol{\theta}_{r_0})$ we get

$$2 \log L_G(\mathbf{x}) = (\mathbf{z}_1 + \boldsymbol{\mu}_1)^T (\mathbf{z}_1 + \boldsymbol{\mu}_1) \sim \chi_r^2(\lambda) \tag{54}$$

where

$$\lambda = \boldsymbol{\mu}_1^T \boldsymbol{\mu}_1 = (\boldsymbol{\theta}_{r_1} - \boldsymbol{\theta}_{r_0})^T \boldsymbol{\Delta}_{\boldsymbol{\mathcal{I}}_{ss}} (\boldsymbol{\theta}_{r_1} - \boldsymbol{\theta}_{r_0}). \tag{55}$$

□

References

[1] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear Estimation*. Prentice-Hall, Inc., Upper Saddle River, NJ, 2000.

[2] S. M. Kay. *Fundamentals of Statistical Signal Processing: Volume 1: Estimation Theory*. Prentice-Hall, PTR, Upper Saddle River, NJ, USA, 2009.

[3] S. M. Kay. *Fundamentals of Statistical Signal Processing: Volume 2: Detection Theory*. Prentice-Hall, PTR, Upper Saddle River, NJ, USA, 2009.

[4] M. Wax and T. Kailath. Detection of signals by information theoretic criteria. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 33(2):387 – 392, apr 1985.