ECE 6123
Advanced Signal Processing
Model Order Selection

Peter Willett

Fall 2017

1 Background on Hypothesis Testing

Let us begin with the principle of optimal decision-making. it is simple to show that given a set of simple hypotheses $\mathcal{H}_i$ ($i \in \{1, 2, \ldots, I\}$), the optimal – in the sense of a minimization of the probability of error – decision is to select

$$\mathcal{H}_j \;=\; \arg\max_{\mathcal{H}_i}\{p(\mathbf{u}|\mathcal{H}_i)Pr(\mathcal{H}_i)\} \tag{1}$$

where $\mathbf{u}$ is the observed data and $p(\cdot)$ represents a probability density. A *simple* hypothesis is one in which $p(\mathbf{u}|\mathcal{H}_i)$ has meaning or can be written. To see this, write

$$\Omega_i \;=\; \{\mathbf{u} \text{ such that } \mathbf{u} \in \Omega_i \text{ means decide } \mathcal{H}_i\} \tag{2}$$

Then

$$\begin{aligned} P(error) &= \sum_{i=1}^{I} Pr(\mathbf{u} \notin \Omega_i|\mathcal{H}_i)Pr(\mathcal{H}_i) && (3)\\ &= 1 - \sum_{i=1}^{I} Pr(\mathbf{u} \in \Omega_i|\mathcal{H}_i)Pr(\mathcal{H}_i) && (4)\\ &= 1 - \sum_{i=1}^{I} \int_{\Omega_i} p(\mathbf{u}|\mathcal{H}_i)Pr(\mathcal{H}_i)d\mathbf{u} && (5)\\ &= 1 - \int \sum_{i=1}^{I} \{\mathcal{I}(\text{decide } \mathcal{H}_i)p(\mathbf{u}|\mathcal{H}_i)Pr(\mathcal{H}_i)\}\, d\mathbf{u} && (6) \end{aligned}$$

which is clearly minimized by the rule (1). An example of a simple hypothesis is $\mathcal{H}_i$ that $\{u[n]\}$ is white and Gaussian with mean time series $\{\mu_i[n]\}$.

A *composite*-hypothesis situation, on the other hand, is one in which we have $p(\mathbf{u}|\theta)$ and

$$\mathcal{H}_i \;=\; \{\theta \in \Theta_i\} \tag{7}$$

for some exhaustive set of Θ_i's. Note that if there exists any *prior* probability measure on θ then this is actually a simple hypothesis test, since we can write

$$p(\mathbf{u}|\mathcal{H}_i) \;=\; \int p(\mathbf{u}|\theta)p(\theta|\mathcal{H}_i)d\theta \tag{8}$$

But otherwise the test is *composite*. The most common testing strategy for composite testing is to use the generalized likelihood (GL)

$$\max_{\theta\in\Theta_i}\{p(\mathbf{u}|\theta)\} \tag{9}$$

and in the case of only two hypotheses it would be simpler to express this as a ratio: the GLR.

To be concrete, suppose you have been given a section of time series $\{u[n]\}_{n=0}^{N-1}$. You are asked to fit an AR model to this. What order AR model? If we maximize (9) the answer is: as large as we can make it. This is because a second-order model is a special case of.a third-order model, and hence the maximized likelihood under a third-order assumption can be no smaller than that under a second-order assumption.

Notionally, there comes a point when increasing the order of the model amounts to "fitting the noise" – it is not providing better explanation of the data, it is just able to *wiggle* more to reduce the deviations. However, how to deal with unknown model order is not at all straightforward; the reason is that unless $p(\theta|\mathcal{H}_i)$ and $Pr(\mathcal{H}_i)$ are known, there is no solidly Bayesian means to test. At any rate, there are two ingredients that we must have – a maximized likelihood and an appropriate *penalty* for over-fitting – and we will attack both in subsequent sections.

2 Maximized Likelihood

2.1 The AR Case

According to the AR model

$$u[n] \;=\; \nu[n] - \sum_{k=1}^{M-1} a_k^* u[n-k] \tag{10}$$

the best predictor for $\{u[n]\}$ based on the past is

$$\hat{u}[n] \;=\; \sum_{k=1}^{M-1} a_k^* u[n-k] \tag{11}$$

which leaves prediction error $\{\nu[n]\}$ having power σ_ν^2 – which we usually call $\{f_m[n]\}$ and P_m for the m^{th}-order model – which according to (10) is a white time sequence. It's easy to see that we have

$$\begin{aligned}
\log(p(\mathbf{u})) &= \sum_{n=0}^{N-1} \log(p(u[n]|u[n-1],\ldots,u[0]) && (12)\\
&\longrightarrow \sum_{n=0}^{N-1} \log(p(u[n]|u[n-1],\ldots,u[n-M]) && (13)\\
&= \begin{cases} \left(-\frac{\sum_{n=0}^{N-1} f_m[n]^2}{2P_m} - \frac{N}{2}\log(2\pi P_m)\right) & \in \Re \\ \left(-\frac{\sum_{n=0}^{N-1} |f_m[n]|^2}{P_m} - N\log(\pi P_m)\right) & \notin \Re \end{cases} && (14)
\end{aligned}$$

Presumably this increases with model-order m and decreases with the number of data N.

2.2 The Eigen-method Case

Suppose we have $\{\mathbf{u}_n\}_{n=1}^N$ that are complex Gaussian based on covariance matrix $\mathbf{R}$. We have

$$\begin{aligned}
p(\{\mathbf{u}_n\}_{n=1}^N) &= \frac{1}{|\pi\mathbf{R}|^N} e^{-\sum_{n=1}^N \mathbf{u}_n^H \mathbf{R}^{-1}\mathbf{u}_n} && (15)\\
&= \frac{1}{|\pi\mathbf{R}|^N} e^{-Tr(\sum_{n=1}^N \mathbf{u}_n^H \mathbf{R}^{-1}\mathbf{u}_n)} && (16)\\
&= \frac{1}{|\pi\mathbf{R}|^N} e^{-Tr(\mathbf{R}^{-1}\sum_{n=1}^N \mathbf{u}_n\mathbf{u}_n^H)} && (17)\\
&= \frac{1}{|\pi\mathbf{R}|^N} e^{-NTr(\mathbf{R}^{-1}\hat{\mathbf{R}})} && (18)
\end{aligned}$$

where of course

$$\hat{\mathbf{R}} \equiv \frac{1}{N}\sum_{n=1}^N \mathbf{u}_n\mathbf{u}_n^H \qquad (19)$$

Our goal is to maximize (18) with respect to $\mathbf{R}$. But since this is an eigen-method, we constrain $\hat{\mathbf{R}}$ to be of reduced rank, say $p < M$.

Let us begin by assuming that the eigenvalues of $\mathbf{R}$ (i.e., $\{\lambda_i\}$) are fixed – this means that $|\mathbf{R}|$ is also fixed. We write

$$\hat{\mathbf{R}} = \sum_{i=1}^M \hat{\lambda}_i \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^H \qquad (20)$$

as the eigendecomposition of the empirical covariance matrix. We then have

$$\begin{aligned} Tr(\mathbf{R}^{-1}\hat{\mathbf{R}}) &= \sum_{i=1}^{M} \hat{\lambda}_i Tr(\mathbf{R}^{-1}\hat{\mathbf{v}}_i\hat{\mathbf{v}}_i^H) && (21)\\ &= \sum_{i=1}^{M} \hat{\lambda}_i Tr(\hat{\mathbf{v}}_i^H\mathbf{R}^{-1}\hat{\mathbf{v}}_i) && (22)\\ &\geq \sum_{i=1}^{M} \hat{\lambda}_{(i)}/\lambda_{(i)} && (23) \end{aligned}$$

where $\hat{\lambda}_{(1)} \geq \hat{\lambda}_{(2)} \geq \ldots \geq \hat{\lambda}_{(M)}$ and $\lambda_{(1)} \geq \lambda_{(2)} \geq \ldots \geq \lambda_{(M)}$. Equation (23) follows from the same logic that we applied to minimize the Frobenius norm of a low-rank approximation to a given matrix; the difference is that there we minimized the Frobenius norm and hence maximized the trace-term; here we are *minimizing* the trace term and hence we match the largest $\hat{\lambda}_i$ with the smallest λ_i^{-1} – which means the largest $\hat{\lambda}_i$ is paired to the largest λ_i, second-largest to second-largest, etc. We thus have

$$\log(p(\{\mathbf{u}_n\}_{n=1}^N)) = N\sum_{i=1}^{M} \hat{\lambda}_{(i)}/\lambda_{(i)} - \sum_{i=1}^{M} N\log(\pi\lambda_{(i)}) \quad (24)$$

We take the gradient with respect to $\{\lambda_{(i)}\}_{i=1}^M$ under the constraint that $\lambda_{(i)} = \lambda_0$ for $p < i \leq M$. Setting it to zero we have

$$\begin{aligned} 0 &= -\frac{N\hat{\lambda}_{(i)}}{\lambda_{(i)}^2} + \frac{N}{\lambda_{(i)}} && (25)\\ \implies \lambda_{(i)} &= \hat{\lambda}_{(i)} && (26) \end{aligned}$$

for $i \in \{1, p\}$, and

$$\begin{aligned} 0 &= -N\sum_{i=p+1}^{M} \frac{\hat{\lambda}_{(i)}}{\lambda_0^2} + N\left(\frac{M-p}{\lambda_0}\right) && (27)\\ \implies \lambda_0 &= \frac{1}{M-p}\sum_{i=p+1}^{M} \hat{\lambda}_{(i)} && (28) \end{aligned}$$

for $i \in \{p+1, M\}$. Clearly the maximum

$$\begin{aligned} &\log(p(\{\mathbf{u}_n\}_{n=1}^N))\\ &\leq N\left(\sum_{i=1}^{p}\frac{\hat{\lambda}_{(i)}}{\hat{\lambda}_{(i)}} + \frac{\sum_{i=p+1}^{M}\hat{\lambda}_{(i)}}{\frac{1}{M-p}\sum_{i=p+1}^{M}\hat{\lambda}_{(i)}}\right) \end{aligned}$$

$$- N \left(\sum_{i=1}^{p} \log(\pi \hat{\lambda}_{(i)}) + (M-p) \log \left(\pi \frac{1}{M-p} \sum_{i=p+1}^{M} \hat{\lambda}_{(i)} \right) \right) \tag{29}$$

$$= -N \left(p + (M-p) + \sum_{i=1}^{M} \log(\hat{\lambda}_{(i)}) - (M-p) \sum_{i=p+1}^{M} \log \left(\hat{\lambda}_{(i)}^{\frac{1}{M-p}} \right) \right)$$

$$- N \left(M \log(\pi) + (M-p) \log \left(\frac{1}{M-p} \sum_{i=p+1}^{M} \hat{\lambda}_{(i)} \right) \right) \tag{30}$$

$$= N \left((M-p) \log \left(\frac{\left(\prod_{i=p+1}^{M} \hat{\lambda}_{(i)} \right)^{\frac{1}{M-p}}}{\frac{1}{M-p} \sum_{i=p+1}^{M} \hat{\lambda}_{(i)}} \right) - M \log(\pi e) - \log(|\hat{\mathbf{R}}|) \right) \tag{31}$$

So, in words: the hard work of the test statistic is done by the *ratio of the geometric to arithmetic means of the eigenvalues in the (empirical) noise subspace.*

2.3 A Little Bit of Random Matrix Theory

RMT is an emerging field for statisticians, with much activity. The results are not simple to prove, and no effort will be given here to offer proofs. There are applications in testing and especially in communications. Signal processors are interested, but are struggling to find applications.

First, please be aware that we are interested (here) in square Hermitian matrices. There are two such classes. The first is the **Wigner** class that involves an $M \times M$ matrix $\mathbf{A} = \mathbf{A}^H$ that is composed of zero-mean complex Gaussian random variables with $1/M$ as their variance[1]. The second is the **Wishart** class of random matrices where

$$\hat{\mathbf{R}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{u}_n \mathbf{u}_n^H \tag{32}$$

where $\mathcal{E}\{\mathbf{u}_n \mathbf{u}_n^H\} = \mathbf{S}$ which is of dimension $M \times M$. In the Wishart class we sometimes are interested in asymptotics where

$$\lim_{N \to \infty} \left\{ \frac{M}{N} \right\} = \gamma \tag{33}$$

shows that there is a scaling between matrix size and estimation accuracy – it does not apply to a situation of near-convergence to a good estimate of the covariance matrix.

[1] Obviously this can be scaled; but all entries must be *iid*.

Wishart Density. It can be shown that the probability density function (pdf) of $\hat{\mathbf{R}}$ is

$$p(\hat{\mathbf{R}}) \;=\; \frac{\left|\hat{\mathbf{R}}\right|^{N-M-1} e^{-\frac{1}{2}Tr(\mathbf{S}^{-1}\hat{\mathbf{R}})}}{2^{\frac{MN}{2}}|\mathbf{S}|^{\frac{N}{2}}\Gamma_M(\frac{N}{2})} \tag{34}$$

where

$$\Gamma_M\left(\frac{N}{2}\right) \;\equiv\; \pi^{\frac{M(M-1)}{4}} \prod_{i=1}^{M} \Gamma\left(\frac{N}{2} - \frac{i-1}{2}\right) \tag{35}$$

is the "multi-variate Gamma function" and in which Γ denotes the usual Gamma function. The pdf (34) is usually written as $\hat{\mathbf{R}} \sim W_M(\mathbf{S}, N)$. It is not asymptotic, and applies for any N and M. The pdf (34) looks fascinating, but I'll admit that I've never seen an application of the Wishart pdf.

Semi-Circle Law. This applies to the Wigner case. It says that the *marginal* pdf of any eigenvalue has pdf

$$p(\lambda) \;=\; \frac{\sqrt{4-\lambda^2}}{2\pi} \tag{36}$$

This (36) is not precisely the pdf for any finite-size matrix, but can be shown to be the asymptotic pdf as $M \to \infty$.

Marcenko-Pastur Law. This is the analog of (36) for Wishart matrices, which is probably more useful for us. In this case the result is asymptotic: the scaled situation of (33). For the case $\gamma < 1$ we have

$$p(\lambda) \;=\; \begin{cases} \frac{\sqrt{(b_+-\lambda)(\lambda-b_-)}}{2\pi\gamma\lambda} & b_- \le \lambda \le b_+ \\ 0 & \text{else} \end{cases} \tag{37}$$

in which

$$b_- \;\equiv\; (1-\sqrt{\gamma})^2 \tag{38}$$

$$b_+ \;\equiv\; (1+\sqrt{\gamma})^2 \tag{39}$$

For $\gamma > 1$ we have

$$p(\lambda) \;=\; \frac{1}{1-\gamma}\delta(\lambda) \;+\; \frac{1}{\gamma}\begin{cases} \frac{\sqrt{(b_+-\lambda)(\lambda-b_-)}}{2\pi\gamma\lambda} & b_- \le \lambda \le b_+ \\ 0 & \text{else} \end{cases} \tag{40}$$

in which

$$b_- \;\equiv\; 0 \tag{41}$$

$$b_+ \;\equiv\; (1+\sqrt{\gamma})^2 \tag{42}$$

This difference – that there are zero eigenvalues – is not so surprising, in that if $\gamma < 1$ it is necessarily that $\hat{\mathbf{R}}$ be singular, since there are fewer snapshots than dimensions.

There are (many) other interesting RMT results. One example is the Tracy-Widom theory for the pdf of the largest eigenvalue. Obviously this would be quite useful when testing for a nontrivial *signal subspace* from data. It is not presented since it is quite complex.

2.4 Asymptotic Distribution of the MLE

Under mild but non-trivial regularity conditions the MLE $\hat{\theta}$ can be converges, as the number of samples upon which is computed goes to infinity, to Gaussian, with mean θ (the true parameter) and covariance $\mathbf{J}_\theta^{-1}$; that is, we have

$$p(\hat{\theta}) \;\approx\; \sqrt{\left|\frac{\mathbf{J}_\theta}{2\pi}\right|} e^{-\frac{1}{2}(\hat{\theta}-\theta)^T \mathbf{J}_\theta (\hat{\theta}-\theta)} \tag{43}$$

The latter quantity $\mathbf{J}_\theta$ is the *Fisher information matrix* (FIM). Generally one does not know the true θ so one is content to use $\mathbf{J}_{\hat{\theta}}$ – this is called the *observed information* (OI), which has little theoretical backing but it often useful in situations where $\mathbf{J}_\theta$ is not independent[2] of θ. There is nothing to be embarrassed about in using the OI instead of the FIM; just be aware that it is an approximation.

3 Penalty Criteria

If we knew $Pr(\mathcal{H}_i)$ and $p(\theta|\mathcal{H}_i)$ then we would have (1) as

$$\mathcal{H}_j \;=\; \arg\max_{\mathcal{H}_i} \left\{ \int p(\mathbf{u}|\mathcal{H}_i, \theta) p(\theta|_i) d\theta Pr(\mathcal{H}_i) \right\} \tag{44}$$

and we would be done. We know neither. But we would like some means to *penalize* more-complex models, such that we could select

$$\mathcal{H}_j \;=\; \arg\max_{\mathcal{H}_i} \left\{ \int p(\mathbf{u}|\mathcal{H}_i, \theta) p(\theta|_i) d\theta Pr(\mathcal{H}_i) - \kappa_p \right\} \tag{45}$$

[2] An example of such lack of dependence is the estimation of the mean of Gaussian data; but such nice behavior is the exception rather than the rule.

as the penalty that applies to a model with p free parameters (such[3] as AR(p)). But in fact we need more than this, since some p^{th}-order models are more attractive than others. It is bests to let the data decide.

There are several "penalty terms" for model order that have some appeal: the Akaike information criterion (AIC), Rissanen's minimum descriptor length (MDL) and the Bayesian information criterion (BIC) come to mind. There are others, and it is a field of continual developments. No penalty term has a really rigorous development; but that is forgivable since the problem of model order selection (without prior information) is not well-posed.

3.1 AIC

First, please recall (or be introduced to) the Kullback-Leibler (KL) divergence between to probability measures (densities)

$$d_{kl}(p,q) \equiv \int p \log\left(\frac{p}{q}\right) \tag{46}$$

We have $d_{kl} = 0$ if and only if $p = q$; otherwise $d_{kl} > 0$. The KL divergence has a great deal of importance in information theory, and is of paramount importance in *large deviations* theory where it describes convergence exponents. And, indeed, if $p(x,y)$ is a joint distribution and $q(x,y)$ has the same marginals but is the special case that the two are independent, then $d_{kl}(p,q)$ is the same as Shannon's Information. But for our purposes, just be aware that d_{kl} is a measure of the difference between p and q.

Akaike assumed:

θ_0 is the true parameter for the true model, which has dimension (number of parameters to be estimated) p_0.

θ is the expected value of the parameter, of order p, for the model being tested.

$\hat{\theta}$ is the maximum-likelihood estimate (MLE) of the parameter, of order p, for the model being tested.

Akaike in 1975 wanted to choose the best model in the sense of minimizing

$$d_{kl}(p_{\theta_0}, p_\theta) \equiv \int p_{\theta_0} \log\left(\frac{p_{\theta_0}}{p_\theta}\right) \tag{47}$$

[3] In the eigenmethod case, the number of free parameters, in the notation just used, is pM, corresponding to the requisite eigenvalues and eigenvectors in the signal-subspace. It is noted that each eigenvector only requires $M - 1$ parameters due to its unit-length requirement.

which amounts to maximizing

$$\int p_{\theta_0}(\mathbf{u}) \log\left(p_\theta(\mathbf{u})\right) d\mathbf{u} \tag{48}$$

where we have defined $\mathbf{u} \equiv \{\mathbf{u}_n\}_{n=1}^N$. Under the the assumption that $\hat{\theta}$ is sufficient for θ we have both

$$p_\theta(\mathbf{u}) \;=\; p_{\hat{\theta}}(\mathbf{u}) p_\theta(\hat{\theta}) \tag{49}$$

which follows from the *factorization theorem* for sufficient statistics; and the asymptotic MLE distribution expression (43). Substituting (49) and (43) into (48) we propose to maximize

$$\int p_{\theta_0}(\mathbf{u}) \left(\log\left(p_{\hat{\theta}}(\mathbf{u})\right) - \frac{1}{2}(\hat{\theta}-\theta)^T \mathbf{J}_\theta (\hat{\theta}-\theta) + \frac{1}{2}\log\left(\left|\frac{\mathbf{J}_\theta}{2\pi}\right|\right)\right) d\mathbf{u} \tag{50}$$

over the model type and order.

The AIC development says that the first term in (50) is the maximized likelihood. The second term assumes that the covariance is indeed $\mathbf{J}^{-1}$, so the expectation results in p, the dimension of $\hat{\theta}$. The third term is ignored. Hence in its raw form the AIC maximizes

$$\arg\max_p \left\{ \max_{\theta \in \Theta_p} \{\log(p(\{\mathbf{u}_n\}_{n=1}^N))\} - p \right\} \tag{51}$$

As can be seen, however, (at least) these problems can be identified:

- The integration in the first term of (50) is ignored.
- It is not clear why $\mathbf{J}^{-1}$ should be the covariance in the second term of (50) when θ_0 is true.
- It is unclear why the third term in (50) can be ignored.
- It is unexplained why the integration in (48) should be over $\mathbf{u}$ when in fact $\mathbf{u}$ is known.

There is a "corrected" form of the AIC for finite data sizes – that is, finite N. It is

$$\arg\max_p \left\{ \max_{\theta \in \Theta_p} \{\log(p(\{\mathbf{u}_n\}_{n=1}^N))\} - \frac{Np}{N-p} \right\} \tag{52}$$

The AIC is probably the first attempt to address the issue of model-order selection, and should be complimented for that; and in fact it works reasonably well for small N. But its development is a Swiss cheese.

3.2 MDL

Rissanen originally developed the MDL with an idea from information theory. A nice intuition is from a notional example. Consider we have an alphabet of 2 letters (OK: here "letters" means bits), N data from this alphabet, and two coding strategies:

1. Treat all symbols are equally likely. N data can be represented by N bits.

2. Randomly[4] generate 2^{10} symbol-probability choices $\{\{p_{i,n}\}_{i=1}^{32}\}_{n=1}^{1024}$, in which $p_{i,n}$ is the probability of symbol i under model n, and of course we must have $\sum_{i=1}^{32} p_{i,n} = 1$. Then for the N data perform a Huffman coding procedure for each $\{p_{i,n}\}_{i=1}^{32}$. Use the shortest coded symbol stream, which should be less than N. Since you must also encode the identity of the code used, the number of coded bits is $\min_n\{N\bar{L}_n\}+10$.

Clearly there is more "overhead" needed in the second strategy[5]; but if the data fits it better (shorter coded length) by enough compared to the overhead, then it might be a better strategy. Suppose we used 2^{20} $\{p_{i,n}\}$'s – presumably the best $\bar{L}_n$ should be lower than for 2^{10}, but is it worth the extra 10 bits needed to tell the decoder which codebook we used?

As I indicated, RIssanen originally was motivated by the ideas above – find the best encoding of the data – which is reminiscent both of Kolmogorov complexity theory and of "universal" source coding. But I find Djuric's 1998 paper the most appealing way to develop MDL. Djuric starts with (1) and takes $Pr(\mathcal{H}_i)$ uniform (and hence ignorable). He then writes

$$p(\mathbf{u}|\mathcal{H}_i) \;=\; \int p(\mathbf{u}|\theta,\mathcal{H}_i)p(\theta|\mathcal{H}_i)d\theta \tag{53}$$

and takes $p(\theta|\mathcal{H}_i)$ uniform as well. Let us put this into a form that we can use:

$$p(\mathbf{u}|\mathcal{H}_i) \;=\; \int p(\theta|\mathcal{H}_i)e^{N\left[\frac{1}{N}\sum_{n=1}^{N}\log(p(\mathbf{u}_|\theta,\mathcal{H}_i))\right]}d\theta \tag{54}$$

We have to discuss Laplace's method of integral approximation now. Consider

$$I(t) \;=\; \int_V f(\mathbf{y})e^{-tg(\mathbf{y})}d\mathbf{y} \tag{55}$$

[4]For uniformity this would be according to the Dirichlet density and model.

[5]We are not interested in the overhead to compute the codes, although this may be considerable; we are only interested in the encoded length.

where $g(y)$ attains its minimum at $\mathbf{y} = \mathbf{c}$ which is an interior point[6] of V. Then since we know $\nabla g(\mathbf{y})|_{\mathbf{y}=\mathbf{c}} = 0$ we can approximate

$$I(t) \;\longrightarrow\; \int_{\mathcal{B}(\mathbf{c})} f(\mathbf{c}) e^{-t\left[g(c) - \frac{1}{2}(\mathbf{y}-\mathbf{c})^T \mathbf{G}(\mathbf{y}-\mathbf{c})\right]} d\mathbf{y} \tag{56}$$

as $t \to \infty$, where $\mathcal{B}(\mathbf{c})$ is a small ball surrounding $\mathbf{c}$ and

$$\mathbf{G} \;\equiv\; \nabla^2 g(\mathbf{y})|_{\mathbf{y}=\mathbf{c}} \tag{57}$$

is the Hessian. We get

$$I(t) \;\longrightarrow\; f(\mathbf{c}) e^{-tg(\mathbf{c})} \sqrt{\left|\frac{2\pi}{t\mathbf{G}}\right|} \tag{58}$$

after integrating and recognizing the multivariate Gaussian form of the integral.

For us doing the MDL derivation we have the correspondences from our problem to the Laplace integral and solution in (55)-(58) given by

$$\begin{aligned}
f(\cdot) &\leftarrow p(\theta|\mathcal{H}_i) \text{ (uniform)} && (59)\\
t &\leftarrow N \text{ (the number of samples)} && (60)\\
\mathbf{c} &\leftarrow \hat{\theta} \text{ (the MLE)} && (61)\\
\mathbf{y} &\leftarrow \theta && (62)\\
g(\cdot) &\leftarrow -\frac{1}{N}\sum_{n=1}^{N} \log(p(\mathbf{u}_n|\theta, \mathcal{H}_i)) && (63)\\
\mathbf{G} &\leftarrow +\mathbf{J}_1 && (64)
\end{aligned}$$

where $\mathbf{J}_1$ is the FIM for one snapshot of data, and recall the negative sign in the definition of the FIM when the second-derivative is used. Consequently we can write

$$\begin{aligned}
\log(p(\mathbf{u}|\mathcal{H}_i)Pr(\mathcal{H}_i)) \;\rightarrow\; & \log(p(\hat{\theta}|\mathcal{H}_i)) \;+\; \log(p(\mathbf{u}|\hat{\theta}, \mathcal{H}_i)) \;-\; \frac{p}{2}\log(2\pi) \\
& -\; \frac{1}{2}\log(|N\mathbf{J}_1|) \;+\; \log(Pr(\mathcal{H}_i)) \qquad (65)
\end{aligned}$$

Ignoring the terms that don't scale with N – meaning the first, third and fifth terms – we have at last the task to look for

$$\arg\max_p \left\{ \max_{\theta \in \Theta_p} \{\log(p(\{\mathbf{u}_n\}_{n=1}^N))\} \;-\; \frac{1}{2}\log(|\mathbf{J}|) \right\} \tag{66}$$

[6]The situation that $\mathbf{c}$ is on the boundary of V is also treatable by Laplace's method, but is not at issue here.

where $\mathbf{J} = N\mathbf{J}_1$ is the FIM of the full data.

One interpretation of (66) is that the penalty term is the maximized logarithm of (43) – with a zero exponent. That is, it is perhaps a fair point of comparison of the maximized likelihood against what it should be.

I am very fond of Djuric's development, and of the "full" result (66). Nonetheless it is worth mentioning that one might consider setting $\mathbf{J} = N\sigma^2\mathbf{I}$. In that case we have

$$\log(|\mathbf{J}|) \;=\; p\log(N) \;+\; p\log(\sigma^2) \tag{67}$$

Again ignoring the terms not increasing with N, we have the *original* MDL

$$\arg\max_p \left\{ \max_{\theta\in\Theta_p} \{\log(p(\{\mathbf{u}_n\}_{n=1}^N))\} \;-\; \frac{p}{2}\log(N) \right\} \tag{68}$$

which is certainly very simple but gives no visibility into models of the same order. It is worth mentioning that Rissanen, in later papers, enhanced his development to incorporate the FIM.

3.3 BIC

The BIC is actually equivalent to the form (68) of the MDL. It is "derived" by assuming the model is from the exponential family.