"AIC-type" Information Criteria

Francis J. DiTraglia

February 26, 2014

> Cite various sources for this material here. Direct them to some references on ML asymptotics or provide a quick reminder.

1 Akaike's Information Criterion

Akaike's Information Criterion (AIC) is often described as summarizing the fit of a model by its maximized Log-likelihood minus a complexity penalty. This is an accurate description of the *formula* for the AIC, but overlooks the deep mathematical ideas that underlie it. The AIC uses an estimate of the Kullback-Leibler divergence to decide which model best approximates the true data generating process.

> Change i subscripts to t subscripts!

1.1 Notation etc.

The ideas presented here apply to arbitrary ML models, but to keep the notation simple, we consider only a scalar random variable Y without conditioning on covariates. We also restrict attention to the iid setting. The notation is as follows:

- Parametric Model: $f(y, \theta)$, sometimes abbreviated $f(y)$

- θ is a p-dimensional vector of parameters

- $Y_1, Y_2, \ldots Y_n$ are iid

- Log-likelihood of sample: $\ell(\theta) = \sum_{i=1}^n \log f(y_i, \theta)$

1.2 Kullback-Leibler Divergence

Consider the following setup:

- Truth: $Y_1, Y_2, \ldots Y_n \sim$ iid $g(y)$

- Model: $f(y)$

- Question: How well does f approximate g?

To find the best approximation to g, we might consider a number of models and choose the one that minimizes some measure of discrepancy. One such measure is the Kullback-Leibler Divergence. Let E_G denote expectation with respect to the true, unknown density $g(y)$.

Definition 1.1 (KL Divergence). *The Kullback-Leibler divergence from $g(y)$ to $f(y)$ is given by*

$$KL(g; f) = E_G \left[\log \left\{ \frac{g(Y)}{f(Y)} \right\} \right] = E_G \left[\log g(Y) \right] - E_G \left[\log f(Y) \right]$$

The quantity $E_G \left[\log f(Y) \right]$ is called the Expected Log-likelihood.

Properties of KL Divergence

1. It is *not* a metric: $KL(g; f) \neq KL(f; g)$

2. $KL(g; f) \geq 0$ with equality iff $f = g$

   Since log is a concave function, $-\log$ is convex. Thus

$$\begin{aligned} KL(g; f) &= E_G \left[\log \left\{ \frac{g(Y)}{f(Y)} \right\} \right] = E_G \left[-\log \left\{ \frac{f(Y)}{g(Y)} \right\} \right] \\ &\geq -\log \left\{ E_G \left[\frac{f(Y)}{g(Y)} \right] \right\} = -\log \left(\int g(y) \frac{f(y)}{g(y)} dy \right) \\ &= -\log \left(\int f(y) dy \right) = -\log(1) = 0 \end{aligned}$$

   by Jensen's Inequality. The inequality is strict only for a non-degenerate random variable and a strictly convex function. Since $-\log$ is strictly convex, the only way to make the inequality strict is for $f(Y)/g(Y)$ to be degenerate. This occurs precisely when $f = g$.

3. *Maximize Expected Log-likelihood $\iff$ Minimize KL Divergence*

   The first term in the KL divergence is a constant: it in no way depends on the model $f(y)$. The expected Log-likelihood enters negatively:

   $$KL(g; f) = E_G\left[\log g(Y)\right] - E_G\left[\log f(Y)\right]$$

   Thus, if we can find a way to estimate the Expected Log-likelihood, we can use the KL divergence for model selection: the larger the Expected Log-likelihood, the smaller the KL divergence, and hence the better the model.

4. The KL divergence equals the negative of **Boltzmann's Entropy** from Statistical Mechanics. Accordingly, it represents the *information lost* when $g(y)$ is encoded by $f(y)$.

1.3 Relationship of MLE to KL

Parameterize f by θ. Then the expected Log-likelihood can be written as:

$$E_G\left[\log f(Y, \theta)\right] = \int \log f(y, \theta) dG(y)$$

Unfortunately, G is unknown. (Remember: G is the true data-generating process.) Replacing G with the empirical distribution $\hat{G}$ yields the estimator:

$$E_{\hat{G}}\left[\log f(Y, \theta)\right] = \frac{1}{n} \sum_{i=1}^{n} \log f(y_i, \theta) = \frac{1}{n}\ell(\theta)$$

By the Weak Law of Large Numbers for iid observations

$$\frac{1}{n}\ell(\theta) \xrightarrow{p} E_G\left[\log f(Y, \theta)\right]$$

Under the standard regularity conditions (see Newey and McFadden) we can strengthen this result to show that

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{n}\ell(\theta) \xrightarrow{p} \arg\max_{\theta \in \Theta} E_G\left[\log f(Y, \theta)\right]$$

Since minimizing the KL divergence is the same as maximizing the expected log-likelihood:

Proposition 1.1. *The ML estimator $\hat{\theta}$ converges in probability to the value of θ that minimizes the KL divergence from unknown true density $g(y)$ to the parametric family $f(y, \theta)$. When $g(y) = f(y, \theta)$ for some value of θ, the divergence is minimized at zero.*

3

1.4   A Naïve Information Criterion

If $g(y)$ were known, we could choose between two parametric models $f(y, \theta)$ and $h(y, \gamma)$ by comparing maximized Log-likelihoods. Define

$$
\begin{aligned}
\theta_0 &= \arg\max_{\theta \in \Theta} E_G \left[\log f(Y, \theta)\right] \\
\gamma_0 &= \arg\max_{\gamma \in \Gamma} E_G \left[\log h(Y, \gamma)\right]
\end{aligned}
$$

If $E_G \left[\log f(Y, \theta_0)\right] > E_G \left[\log h(Y, \gamma_0)\right]$, then the KL divergence from $g(y)$ to the parametric family f_θ is smaller than that from $g(y)$ to h_γ.

From the previous subsection, we know that $\hat{\theta}_{ML} \xrightarrow{p} \theta_0$. Further, $\frac{1}{n}\ell(\theta) \xrightarrow{p} E_G \left[\log f(Y, \theta)\right]$. Of course, n will be constant across models, so why not use the maximized sample likelihood $\ell(\hat{\theta})$ for model comparison? *BECAUSE IT'S BIASED!*

1.5   Intuition for the Bias

The basic problem is that $\ell(\hat{\theta})$ uses the data twice: first to estimate $\hat{\theta}$ and then directly in the sum $\sum_{i=1}^{n} \log f(y_i, \hat{\theta})$. Because $\hat{\theta}$ was chosen to conform to the idiosyncrasies of the data at hand, $\ell(\hat{\theta})$ is overly optimistic.[1]

Since θ_0 is the population minimizer of the KL divergence from g to f_θ,

$$
\begin{aligned}
KL\left[g(y); f(y, \theta)\right] &\geq KL\left[g(y); f(y, \theta_0)\right] \\
E_G\left[\log g(Y)\right] - E_G\left[\log f(Y, \theta)\right] &\geq E_G\left[\log g(Y)\right] - E_G\left[\log f(Y, \theta_0)\right] \\
E_G\left[\log f(Y, \theta)\right] &\leq E_G\left[\log f(Y, \theta_0)\right]
\end{aligned}
$$

for all $\theta \in \Theta$. Recall that $\frac{1}{n}\ell(\theta) = E_{\hat{G}}\left[\log f(Y, \theta)\right]$. By the definition of the maximum likelihood estimate, $\ell(\hat{\theta}) \geq \ell(\theta_0)$. Thus,

$$
E_{\hat{G}}\left[\log f(Y, \hat{\theta})\right] \geq E_{\hat{G}}\left[\log f(Y, \theta_0)\right]
$$

In sample, the estimate $\hat{\theta}$ will show a higher maximized log-likelihood than the value of θ that maximizes the population log-likelihood.

*What we should be evaluating is the likelihood at a **new** set of observations not used to estimate $\hat{\theta}$.*

---

[1]This is similar in spirit is the *post hoc contrast*: measuring two groups on a number of dimensions, choosing the dimension along which they differ the most, and carrying out a *t*-test on this dimension. What a surprise; the result is significant!

1.6 An Asymptotic Approximation to the Bias

First some notation:

- Make the sample size and the dependence of $\hat{\theta}$ on the data explicit by writing

$$\log f\left(\mathbf{y}_n \mid \hat{\theta}(\mathbf{y}_n)\right) = \sum_{i=1}^{n} \log f(y_i \mid \hat{\theta})$$

- Let $E_{G(\mathbf{y}_n)}$ denote expectation with respect to the joint distribution of $\mathbf{Y}^{(n)} = (Y_1, \ldots, Y_n)'$. The realizations of this random vector are $\mathbf{y}_n = y_1, \ldots, y_n$.

- Let $\mathbf{Z}^{(n)} = (Z_1, \ldots, Z_n)'$ be a collection of iid random variables with the same distribution as $\mathbf{Y}^{(n)}$ that were *not* used to estimate $\hat{\theta}$. (Assume that $\mathbf{Z}^{(n)}$ and $\mathbf{Y}^{(n)}$ are mutually independent.) Let $G(\mathbf{z}_n)$ denote expectation with respect to the joint distribution of $\mathbf{Z}^{(n)}$.

- The bias of the maximized Log-likelihood as an estimator of the expected Log-likelihood is given by

$$Bias = E_{G(\mathbf{y}_n)}\left[\log f\left(\mathbf{Y}^{(n)} \mid \hat{\theta}\left(\mathbf{Y}^{(n)}\right)\right) - E_{G(\mathbf{z}_n)}\left\{\log f\left(\mathbf{Z}^{(n)} \mid \hat{\theta}\left(\mathbf{Y}^{(n)}\right)\right)\right\}\right]$$

Calculating the Bias: Since $Z_1, \ldots, Z_n$ are iid, the expectation with respect to $G(\mathbf{z}_n)$ simplifies, and we can write

$$Bias = E_{G(\mathbf{y}_n)}\left[\log f\left(\mathbf{Y}^{(n)} \mid \hat{\theta}\left(\mathbf{Y}^{(n)}\right)\right) - nE_{G(z)}\left\{\log f\left(Z \mid \hat{\theta}\left(\mathbf{Y}^{(n)}\right)\right)\right\}\right]$$

where Z is any of the $Z_1, \ldots, Z_n$ and $G(z)$ is its *marginal* distribution. We expand the bias into $D_1 + D_2 + D_3$ where

$$D_1 = E_{G(\mathbf{y}_n)}\left[\log f\left(\mathbf{Y}^{(n)} \mid \hat{\theta}\left(\mathbf{Y}^{(n)}\right)\right) - \log f\left(\mathbf{Y}^{(n)} \mid \theta_0\right)\right]$$

$$D_2 = E_{G(\mathbf{y}_n)}\left[\log f\left(\mathbf{Y}^{(n)} \mid \theta_0\right) - nE_{G(z)}\left\{\log f\left(Z \mid \theta_0\right)\right\}\right]$$

$$D_3 = E_{G(\mathbf{y}_n)}\left[nE_{G(z)}\left\{\log f\left(Z \mid \theta_0\right)\right\} - nE_{G(z)}\left\{\log f\left(Z \mid \hat{\theta}\left(\mathbf{Y}^{(n)}\right)\right)\right\}\right]$$

Now consider each part separately.

Step 1: Calculation of D_2 This is the easiest part as it involves no estimators. Note first that $E_{G(z)}\{\log f(Z|\theta_0)\}$ is a constant, so that

$$
\begin{aligned}
D_2 &= E_{G(\mathbf{y}_n)}\left[\log f\left(\mathbf{Y}^{(n)}|\theta_0\right)\right] - nE_{G(z)}\left[\log f(Z|\theta_0)\right] \\
&= nE_{G(y)}\left[\log f(Y|\theta_0)\right] - nE_{G(z)}\left[\log f(Z|\theta_0)\right] \\
&= 0
\end{aligned}
$$

where we have used the fact that $Y_1, \ldots, Y_n$ are iid and that Y and Z have the same distribution.

Step 2: Calculation of D_3 Define $\eta(\hat{\theta}) = E_{G(z)}\left[\log f\left(z|\hat{\theta}(\mathbf{Y}^{(n)})\right)\right]$ and Taylor expand around θ_0, the solution to the population moment condition

$$
E_{G(z)}\left[\frac{\partial}{\partial\theta}\log f(z|\theta)\right] = 0
$$

yielding

$$
\eta(\hat{\theta}) = \eta(\theta_0) + \left(\hat{\theta} - \theta_0\right)'\frac{\partial\eta(\theta_0)}{\partial\theta} + \frac{1}{2}\left(\hat{\theta} - \theta_0\right)\frac{\partial^2\eta(\theta_0)}{\partial\theta\partial\theta'}\left(\hat{\theta} - \theta_0\right) + R_\eta
$$

Now, under the appropriate regularity conditions (Lebesgue Dominated Convergence) we can exchange the order of expectation and differentiation, so that

$$
\frac{\partial\eta(\theta_0)}{\partial\theta} = \frac{\partial}{\partial\theta}E_{G(z)}\left[\log f(z|\theta_0)\right] = E_{G(z)}\left[\frac{\partial}{\partial\theta}\log f(z|\theta_0)\right] = 0
$$

by the definition of θ_0 as the solution to the population moment condition. Substituting this into the Taylor Expansion, and assuming further that we may exchange the *second* derivative and expectation with respect to $G(z)$, we have

$$
\begin{aligned}
\eta(\hat{\theta}) &= \eta(\theta_0) + \frac{1}{2}\left(\hat{\theta} - \theta_0\right)'\frac{\partial^2}{\partial\theta\partial\theta'}E_{G(z)}\left[\log f(z|\theta_0)\right]\left(\hat{\theta} - \theta_0\right) + R_\eta \\
&= \eta(\theta_0) + \frac{1}{2}\left(\hat{\theta} - \theta_0\right)'E_{G(z)}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(z|\theta_0)\right]\left(\hat{\theta} - \theta_0\right) + R_\eta
\end{aligned}
$$

Thus, defining

$$
J(\theta_0) = -E_{G(z)}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(z|\theta_0)\right]
$$

we see that

$$
\eta(\hat{\theta}) = E_{G(z)}\left[\log f(z|\theta_0)\right] - \frac{1}{2}\left(\hat{\theta} - \theta_0\right)'J(\theta_0)\left(\hat{\theta} - \theta_0\right) + R_\eta
$$

Substituting this expansion for $E_{G(z)}\left[\log f\left(z|\hat{\theta}(\mathbf{Y}^{(n)})\right)\right]$ into the expression for D_3 cancels the leading first term $nE_{G(z)}\left\{\log f\left(Z|\theta_0\right)\right\}$ in the expectation so that (remember that $\hat{\theta}$ is a function of $\mathbf{y}_n$ and thus, so is R_η)

$$
\begin{aligned}
D_3 &= E_{G(\mathbf{y}_n)}\left[\frac{n}{2}\left(\hat{\theta}-\theta_0\right)' J(\theta_0)\left(\hat{\theta}-\theta_0\right)+R_\eta\right] \\
&= \frac{1}{2}E_{G(\mathbf{y}_n)}\left[\text{trace}\left\{n\left(\hat{\theta}-\theta_0\right)' J(\theta_0)\left(\hat{\theta}-\theta_0\right)\right\}\right]+R \\
&= \frac{1}{2}\text{trace}\left\{J(\theta_0)E_{G(\mathbf{y}_n)}\left[\sqrt{n}\left(\hat{\theta}-\theta_0\right)\sqrt{n}\left(\hat{\theta}-\theta_0\right)'\right]\right\}+R
\end{aligned}
$$

where we have used the facts that: (1) a scalar equals its trace, (2) multiplication within the trace operator is commutative provided that the products remain conformable, (3) trace and expectation can be exchanged because both are linear operators, and (4) $J(\theta_0)$ is a constant with respect to $E_{G(\mathbf{y}_n)}$. We know that

$$
E_{G(\mathbf{y}_n)}\left[\sqrt{n}\left(\hat{\theta}-\theta_0\right)\sqrt{n}\left(\hat{\theta}-\theta_0\right)'\right] \to J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0)
$$

as $n\to\infty$ where

$$
\begin{aligned}
I(\theta_0) &= E_{G(z)}\left[\frac{\partial\log f(Z|\theta)}{\partial\theta}\frac{\partial\log f(Z|\theta)}{\partial\theta'}\right] \\
J(\theta_0) &= -E_{G(z)}\left[\frac{\partial^2\log f(Z|\theta)}{\partial\theta\partial\theta'}\right]
\end{aligned}
$$

Thus, assuming the necessary regularity conditions to ensure that the remainder is indeed of small order, as $n\to\infty$ we have

$$
D_3 \to \frac{1}{2}\text{trace}\left\{J(\theta_0)J(\theta_0)^{-1}I(\theta_0)J(\theta_0)^{-1}\right\} = \frac{1}{2}\text{trace}\left\{I(\theta_0)J(\theta_0)^{-1}\right\}
$$

Step 3: Calculation of D_1 Taylor expand $\ell(\theta_0) = \log f\left(\mathbf{y}_n|\theta_0\right)$ around $\hat{\theta}$, noting that by the definition of the MLE,

$$
\frac{\partial\ell(\hat{\theta})}{\partial\theta} = 0
$$

Thus,

$$
\ell(\theta_0) = \ell(\hat{\theta}) + \frac{1}{2}\left(\theta-\hat{\theta}\right)'\frac{\partial^2\ell(\tilde{\theta})}{\partial\theta\partial\theta'}\left(\theta-\hat{\theta}\right)
$$

7

where $\tilde{\theta}$ is between $\hat{\theta}$ and θ_0, so we have

$$\ell(\hat{\theta}) - \ell(\theta_0) = \frac{1}{2}\sqrt{n}\left(\theta - \hat{\theta}\right)' \left[-\frac{1}{n}\frac{\partial^2 \ell(\tilde{\theta})}{\partial\theta\partial\theta'}\right]\sqrt{n}\left(\theta - \hat{\theta}\right)$$

By the WLLN for iid observations, we know that for fixed θ

$$-\frac{1}{n}\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta'}\log f(y_i|\theta) \xrightarrow{p} J(\theta)$$

Furthermore, by the standard consistency result for MLE $\hat{\theta} \to \theta_0$. Thus, by a Uniform WLLN,

$$-\frac{1}{n}\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'} \xrightarrow{p} J(\theta_0)$$

It follows that

$$
\begin{aligned}
D_1 &= E_{G(\mathbf{y}_n)}\left[\log f\left(\mathbf{Y}^{(n)}\big|\hat{\theta}\left(\mathbf{Y}^{(n)}\right)\right) - \log f\left(\mathbf{Y}^{(n)}|\theta_0\right)\right] \\
&= E_{G(\mathbf{y}_n)}\left[\ell(\hat{\theta}) - \ell(\theta_0)\right] \\
&\to \frac{1}{2}\mathrm{trace}\left\{I(\theta_0)J(\theta_0)^{-1}\right\}
\end{aligned}
$$

using the same arguments as in Step 2.

Finally The Bias: Combining the three steps, we have

$$Bias = D_1 + D_2 + D_3 = \mathrm{trace}\left\{I(\theta_0)J(\theta_0)^{-1}\right\}$$

1.7   Correcting the Bias

Now that we have an asymptotic approximation to the bias of the maximized likelihood as an estimator of the expected log-likelihood, we can correct it. Suppose that we have consistent estimators $\hat{I}$ and $\hat{J}$ of $I(\theta_0)$ and $J(\theta_0)$. Then, by the Continuous Mapping Theorem $\hat{b} = \mathrm{trace}(\hat{I}\hat{J}^{-1})$ is a consistent estimator of the bias term. Recall that the maximized log-likelihood is biased *upwards*, hence to correct it we need to subtract this quantity. This estimate yields **Takeuchi's Information Criterion** (TIC):

$$TIC = 2\left[\sum_{i=1}^{n}\log f\left(y_i|\hat{\theta}\right) - \mathrm{trace}(\hat{I}\hat{J}^{-1})\right]$$

The scaling factor of two is traditional, but has no effect on model comparisons.

Akaike's Information Criterion (AIC) uses a slightly different bias correction. Recall that if there exists a $\theta \in \Theta$ such that the true density $g(y)$ equals the model $f(y|\theta)$, then $I(\theta_0) = J(\theta_0)^{-1}$. This is the famous "Information Matrix Equality." In this case, the bias approximation becomes:

$$Bias = \text{trace}\left\{I(\theta_0)J(\theta_0)^{-1}\right\} = \text{trace}\left\{\mathbb{I}_p\right\} = p$$

that is, the dimension of the parameter vector. Using this as a bias correction yields **Akaike's Information Criterion**:

$$AIC = 2\left[\sum_{i=1}^{n}\log f\left(y_i|\hat{\theta}\right) - p\right]$$

Again, it is traditional to multiply by two. Although the TIC and AIC are similar, there are several subtleties:

- The AIC is derived assuming that the model is *correctly specified*, while the TIC is not. In this sense the AIC is a special case of the TIC.

- Typically, the matrices $I(\theta_0)$ and $J(\theta_0)$ are large, meaning that the estimates will have high variance (we need to estimate $p^2 + p$ elements).

- In contrast, the AIC a has much smaller variance because the bias correction *does not depend on the data.*

- Thus, even if the model is mis-specified, it may be preferable to use AIC rather than TIC unless the sample size is large.

- It has been argued that for models where the Information Matrix is not satisfied, the AIC will still be close to the TIC. (The log-likelihood term should dominate the bias correction in such situations.)

1.8   Friends of AIC

Recall the idea behind the AIC:

1. Use the KL divergence to choose a model.

2. Suffices to compare Expected Log-likelihoods.

3. Sample analogue, maximized log-likelihood, requires a bias correction.

The AIC and TIC use an analytical bias correction. Other possibilities:

- Cross-Validation

- Bootstrap Information Criterion (EIC)

Maximum likelihood estimators can be unstable. If we want to use a different kind of estimator (e.g. maximum penalized likelihood) but are still willing to write down a likelihood, we can use the KL divergence for model selection via the Generalized Information Criterion (GIC) of Konishi and Kitagawa, or its bootstrap equivalent (EGIC).

2 Corrected AIC

3 Cavanaugh & Neath State-Space AIC

4 Schwarz's Bayesian Information Criterion

4.1 Brief Review of Bayesian Statistics

- Treat the parameter θ as a *Random Variable*

- Prior distribution $\pi(\theta)$ encodes information about the parameter θ known *before* observing a sample of data $\mathbf{y} = y_1, \ldots, y_n$.

- Sample information enters through the likelihood $f(\mathbf{y}|\theta)$, the probability model that represents how the distribution of the data depends on the parameter.

- The goal of inference is to calculate the posterior distribution $\pi(\theta|\mathbf{y})$, combining prior and sample information by Bayes' Rule:

Proposition 4.1 (Bayes' Rule)**.**

$$\pi(\theta|\mathbf{y}) = \frac{\pi(\theta)f(\mathbf{y}|\theta)}{\int \pi(\theta)f(\mathbf{y}|\theta)d\theta}$$

Proof. By the definition of a conditional density,

$$\pi(\theta|\mathbf{y}) = \frac{f(\theta, \mathbf{y})}{f(\mathbf{y})}, \text{ and } f(\theta, \mathbf{y}) = \pi(\theta)f(\mathbf{y}|\theta)$$

Thus,

$$f(\mathbf{y}) = \int f(\theta, \mathbf{y})d\theta = \int \pi(\theta)f(\mathbf{y}|\theta)d\theta$$

Combining these,

$$\pi(\theta|\mathbf{y}) = \frac{\pi(\theta)f(\mathbf{y}|\theta)}{\int \pi(\theta)f(\mathbf{y}|\theta)d\theta}$$

$\square$

4.2 Bayesian Model Selection

Bayesian inference means calculating a posterior combining prior and sample information. To choose a model, we put prior probabilities on each candidate and calculate a posterior by Bayes' Rule. Here are the steps:

1. Consider a set of r candidate models $M_1, \ldots, M_r$

2. Prior for ith Model: $p(M_i)$, where $\sum_{i=1}^r p(M_i) = 1$.

3. Posterior: $p(M_i|\mathbf{y})$, satisfying $\sum_{i=1}^r p(M_i|\mathbf{y}) = 1$

4. Bayes' Rule gives:
$$p(M_i|\mathbf{y}) = \frac{p(M_i)f(\mathbf{y}|M_i)}{f(\mathbf{y})}$$

This notation gives a clear analogy with standard Bayesian inference, but hides the dependence on parameters:

- The quantity $f(\mathbf{y}|M_i)$ is called the *Marginal Likelihood* of Model M_i.

- Each model has a likelihood $f_i(\mathbf{y}|\theta_i)$ that depends on a parameter θ_i. Different models may have different parameters.

- Each model also has a prior for its parameter: $\pi(\theta_i)$ is the prior for the parameter θ_i of model M_i.

- The marginal likelihood is obtained by "integrating out" the parameters from the *joint* distribution of $\mathbf{y}$ and θ_i:

$$f(\mathbf{y}|M_i) = \int f_i(\mathbf{y}, \theta_i)d\theta_i = \int \pi(\theta_i)f_i(\mathbf{y}|\theta_i)d\theta_i$$

- By the law of total probability:

$$f(\mathbf{y}) = \sum_{i=1}^r p(M_i)f(\mathbf{y}|M_i)$$

11

This is summarized in the following proposition.

Proposition 4.2 (Model Posteriors)**.**

$$p(M_i|\mathbf{y}) = \frac{p(M_i)f(\mathbf{y}|M_i)}{\sum_{i=1}^{r} p(M_i)f(\mathbf{y}|M_i)}$$

where

$$f(\mathbf{y}|M_i) = \int \pi(\theta_i)f_i(\mathbf{y}|\theta_i)d\theta_i$$

How to Choose a Model? Although there are other possibilities, the usual approach is to choose the model with the highest posterior probability. (In decision-theoretic terms this is equivalent to a 0–1 loss function.) Since the denominator in the above expression is constant across models, it suffices to choose M_i to maximize

$$p(M_i)\int \pi(\theta_i)f_i(\mathbf{y}|\theta_i)d\theta_i$$

If we give equal prior weight to each model by setting $p(M_i) = 1/r$, this reduces to maximizing the Marginal Likelihood alone:

$$\int \pi(\theta_i)f_i(\mathbf{y}|\theta_i)d\theta_i$$

This is called putting a *uniform prior* on the set of candidate models.

4.3 Overview of the BIC

Despite its name, the BIC is *not* a Bayesian procedure. It is a large-sample Frequentist *approximation* to Bayesian model selection:

1. Begin with a uniform prior on the set of candidate models so that it suffices to maximize the Marginal Likelihood.

2. The BIC is a large sample approximation to the Marginal Likelihood:

$$\int \pi(\theta_i)f_i(\mathbf{y}|\theta_i)d\theta_i$$

3. As usual when Bayesian procedures are subjected to Frequentist asymptotics, the priors on parameters vanish in the limit.

4. We proceed by a *Laplace Approximation* to the Marginal Likelihood

4.4 Laplace Approximation

For the moment simplify the notation by suppressing dependence on M_i. We want to approximate:

$$\int \pi(\theta) f(\mathbf{y}|\theta) d\theta$$

This is actually a common problem in applications of Bayesian inference:

- Notice that $\pi(\theta) f(\mathbf{y}|\theta)$ is the *kernel* of some probability density, i.e. the density without its normalizing constant.

- *How do we know this?* By Bayes' Rule

$$\pi(\theta|\mathbf{y}) = \frac{\pi(\theta) f(\mathbf{y}|\theta)}{\int \pi(\theta) f(\mathbf{y}|\theta) d\theta}$$

  is a proper probability density and the denominator is *constant* with respect to θ. (The parameter has been "integrated out.")

- In Bayesian inference, we specify $\pi(\theta)$ and $f(\mathbf{y}|\theta)$, so $\pi(\theta) f(\mathbf{y}|\theta)$ is known. But to calculate the posterior we need to *integrate* to find the normalizing constant.

- Only in special cases (e.g. conjugate families) can we find the exact normalizing constant. Typically some kind of approximation is needed:

    - Importance Sampling
    - Markov-Chain Monte Carlo (MCMC)
    - *Laplace Approximation*

The Laplace Approximation is an *analytical approximation* based on Taylor Expansion arguments. In Bayesian applications, the expansion is carried out around the posterior mode, i.e. the mode of $\pi(\theta) f(\mathbf{y}|\theta)$, but we will expand around the Maximum likelihood estimator.

Proposition 4.3 (Laplace Approximation)**.**

$$\int \pi(\theta) f(\mathbf{y}|\theta) d\theta \approx \frac{\exp\left\{\ell(\hat{\theta})\right\} \pi(\hat{\theta})(2\pi)^{p/2}}{n^{p/2} \left|J(\hat{\theta})\right|^{1/2}}$$

Where $\hat{\theta}$ is the maximum likelihood estimator, *p the dimension of θ and*

$$J(\hat{\theta}) = -\frac{1}{n} \frac{\partial^2 \log f(\mathbf{y}|\hat{\theta})}{\partial\theta\partial\theta'}$$

Proof. A rigorous proof of this result is complicated. The following is a sketch. First write $\ell(\theta)$ for $\log f(\mathbf{y}|\theta)$ so that

$$\pi(\theta)f(\mathbf{y}|\theta) = \pi(\theta)\exp\left\{\log f(\mathbf{y}|\theta)\right\} = \pi(\theta)\exp\left\{\log \ell(\theta)\right\}$$

By a second-order Taylor Expansion around the MLE $\hat{\theta}$

$$\ell(\theta) = \ell(\hat{\theta}) + \frac{1}{2}\left(\theta - \hat{\theta}\right)'\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right) + R_\ell \tag{1}$$

since the derivative of $\ell(\theta)$ is zero at $\hat{\theta}$ by the definition of MLE. A first-order expansion is sufficient for $\pi(\theta)$ because the derivative does not vanish at $\hat{\theta}$

$$\pi(\theta) = \pi(\hat{\theta}) + \frac{\partial\pi(\hat{\theta})}{\partial\theta'}\left(\theta - \hat{\theta}\right) + R_\pi \tag{2}$$

Substituting Equations 1 and 2,

$$
\begin{aligned}
\int \pi(\theta)f(\mathbf{y}|\theta)d\theta &= \int \exp\left\{\ell(\hat{\theta}) + \frac{1}{2}\left(\theta - \hat{\theta}\right)'\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right) + R_\ell\right\} \\
&\qquad \times \left[\pi(\hat{\theta}) + \left(\theta - \hat{\theta}\right)'\frac{\partial\pi(\hat{\theta})}{\partial\theta} + R_\pi\right]d\theta \\
&= \exp\left\{\ell(\hat{\theta})\right\}(I_1 + I_2 + I_3)
\end{aligned}
$$

where

$$
\begin{aligned}
I_1 &= \pi(\hat{\theta})\int \exp\left\{\frac{1}{2}\left(\theta - \hat{\theta}\right)'\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right) + R_\ell\right\}d\theta \\
I_2 &= \frac{\partial\pi(\hat{\theta})}{\partial\theta'}\int \left(\theta - \hat{\theta}\right)\exp\left\{\frac{1}{2}\left(\theta - \hat{\theta}\right)'\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right) + R_\ell\right\}d\theta \\
I_3 &= \int R_\pi \exp\left\{\frac{1}{2}\left(\theta - \hat{\theta}\right)'\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right) + R_\ell\right\}d\theta
\end{aligned}
$$

Under certain regularity conditions (not the standard ones!) we can treat R_ℓ and R_π as approximately equal to zero for large n uniformly in θ, so that

$$
\begin{aligned}
I_1 &\approx \pi(\hat{\theta})\int \exp\left\{\frac{1}{2}\left(\theta - \hat{\theta}\right)'\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right)\right\}d\theta \\
I_2 &\approx \frac{\partial\pi(\hat{\theta})}{\partial\theta'}\int \left(\theta - \hat{\theta}\right)\exp\left\{\frac{1}{2}\left(\theta - \hat{\theta}\right)'\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right)\right\}d\theta \\
I_3 &\approx 0
\end{aligned}
$$

14

Because $\hat{\theta}$ is the MLE,

$$\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}$$

must be negative definite, so

$$-\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}$$

is positive definite. It follows that

$$\exp\left\{\frac{1}{2}\left(\theta-\hat{\theta}\right)'\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta-\hat{\theta}\right)\right\} = \exp\left\{-\frac{1}{2}\left(\theta-\hat{\theta}\right)'\left[\left(-\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1}\right]^{-1}\left(\theta-\hat{\theta}\right)\right\}$$

can be viewed as the kernel of a Normal distribution with mean $\hat{\theta}$ and variance matrix

$$\left(-\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1}$$

Thus,

$$\int \exp\left\{\frac{1}{2}\left(\theta-\hat{\theta}\right)'\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta-\hat{\theta}\right)\right\}d\theta = (2\pi)^{p/2}\left|\left(-\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1}\right|^{1/2}$$

and

$$\int \left(\theta-\hat{\theta}\right)\exp\left\{\frac{1}{2}\left(\theta-\hat{\theta}\right)'\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta-\hat{\theta}\right)\right\}d\theta = 0$$

Therefore,

$$\begin{aligned}
\int \pi(\theta)f(\mathbf{y}|\theta)d\theta &\approx \exp\left\{\ell(\hat{\theta})\right\}\pi(\hat{\theta})\,(2\pi)^{p/2}\left|\left(-\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1}\right|^{1/2}\\
&= \exp\left\{\ell(\hat{\theta})\right\}\pi(\hat{\theta})\,(2\pi)^{p/2}\left|n\left(-\frac{1}{n}\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\right)\right|^{-1/2}\\
&= \frac{\exp\left\{\ell(\hat{\theta})\right\}\pi(\hat{\theta})\,(2\pi)^{p/2}}{n^{p/2}\left|J(\hat{\theta})\right|^{1/2}}
\end{aligned}$$

$\square$

15

4.5   Finally the BIC

Now we re-introduce the dependence on the model M_i. Taking logs of the Laplace Approximation and multiplying by two (again, this is traditional but has no effect on model comparisons)

$$
\begin{aligned}
2 \log f(y|M_i) &= 2 \log \left\{ \int f_i(y|\theta_i)\pi(\theta_i)d\theta_i \right\} \\
&\approx 2\ell(\hat{\theta}_i) - p \log(n) + p \log(2\pi) - \pi(\hat{\theta}_i) - \log \left| J(\hat{\theta}_i) \right|
\end{aligned}
$$

The first two terms are $O_p(n)$ and $O_p(\log n)$, while the last three are $O_p(1)$, hence negligible as $n \to \infty$. This gives us Schwarz's BIC

$$
BIC(M_i) = 2 \log f_i(\mathbf{y}|\hat{\theta}_i) - p \log n
$$

We choose the model M_i for which $BIC(M_i)$ is largest. Notice that the prior on the parameter, $\pi(\theta)$, drops out in the limit, and recall that we began by putting a uniform prior on the *models* under consideration.