Maximum Likelihood Estimation and the Bayesian Information Criterion

Donald Richards

Penn State University

The Method of Maximum Likelihood

R. A. Fisher (1912), “On an absolute criterion for fitting frequency curves,” Messenger of Math. **41**, 155–160

Fisher’s first mathematical paper, written while a final-year undergraduate in mathematics and mathematical physics at Gonville and Caius College, Cambridge University

Fisher’s paper started with a criticism of two methods of curve fitting: the method of least-squares and the method of moments

It is not clear what motivated Fisher to study this subject; perhaps it was the influence of his tutor, F. J. M. Stratton, an astronomer

X: a random variable

θ is a parameter

$f(x;\theta)$: A statistical model for X

$X_1, \ldots, X_n$: A random sample from X

We want to construct good estimators for θ

The estimator, obviously, should depend on our choice of f

Protheroe, et al. “Interpretation of cosmic ray composition – The path length distribution,” ApJ., 247 1981

X: Length of paths

Parameter: $\theta > 0$

Model: The exponential distribution,

$$f(x;\theta) = \theta^{-1} \exp(-x/\theta), \quad x > 0$$

Under this model, $E(X) = \theta$

Intuition suggests using $\bar{X}$ to estimate θ

$\bar{X}$ is unbiased and consistent

LF for globular clusters in the Milky Way

X: The luminosity of a randomly chosen cluster

van den Bergh's Gaussian model,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2}\Big)$$

μ: Mean visual absolute magnitude

σ: Standard deviation of visual absolute magnitude

$\bar{X}$ and S^2 are good estimators for μ and σ^2, respectively

We seek a method which produces good estimators automatically: No guessing allowed

Choose a globular cluster at random; what is the "chance" that the LF will be exactly -7.1 mag? exactly -7.2 mag?

For any continuous random variable X, $P(X = x) = 0$

Suppose $X \sim N(\mu = -6.9, \sigma^2 = 1.21)$, i.e., X has probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2}\Big)$$

then $P(X = -7.1) = 0$

However, ...

$$f(-7.1) = \frac{1}{1.1\sqrt{2\pi}} \exp\left(- \frac{(-7.1+6.9)^2}{2(1.1)^2} \right) = 0.37$$

Interpretation: In one simulation of the random variable X, the "likelihood" of observing the number -7.1 is 0.37

$$f(-7.2) = 0.28$$

In one simulation of X, the value $x = -7.1$ is 32% more likely to be observed than the value $x = -7.2$

$x = -6.9$ is the value with highest (or maximum) likelihood; the prob. density function is maximized at that point

Fisher's brilliant idea: The method of maximum likelihood

Return to a general model $f(x;\theta)$

Random sample: $X_1, \ldots, X_n$

Recall that the X_i are independent random variables

The joint probability density function of the sample is

$$f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta)$$

Here the variables are the X's, while θ is fixed

Fisher's ingenious idea: Reverse the roles of the x's and θ

Regard the X's as fixed and θ as the variable

The likelihood function is

$$L(\theta; X_1, \ldots, X_n) = f(X_1;\theta) f(X_2;\theta) \cdots f(X_n;\theta)$$

Simpler notation: $L(\theta)$

$\hat{\theta}$, the maximum likelihood estimator of θ, is the value of θ where L is maximized

$\hat{\theta}$ is a function of the X's

Note: The MLE is not always unique.

Example: “... cosmic ray composition – The path length distribution ...”

X: Length of paths

Parameter: $\theta > 0$

Model: The exponential distribution,

$$f(x;\theta) = \theta^{-1}\exp(-x/\theta), \quad x > 0$$

Random sample: $X_1, \ldots, X_n$

Likelihood function:

$$\begin{aligned} L(\theta) &= f(X_1;\theta) f(X_2;\theta) \cdots f(X_n;\theta) \\ &= \theta^{-n}\exp(-n\bar{X}/\theta) \end{aligned}$$

Maximize L using calculus

It is also equivalent to maximize $\ln L$:

$\ln L(\theta)$ is maximized at $\theta = \bar{X}$

Conclusion: The MLE of θ is $\hat{\theta} = \bar{X}$

LF for globular clusters: $X \sim N(\mu, \sigma^2)$, with both μ, σ unknown

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right)$$

A likelihood function of two variables,

$$\begin{aligned} L(\mu, \sigma^2) &= f(X_1; \mu, \sigma^2) \cdots f(X_n; \mu, \sigma^2) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \right) \end{aligned}$$

Solve for μ and σ^2 the simultaneous equations:

$$\frac{\partial}{\partial \mu} \ln L = 0, \qquad \frac{\partial}{\partial (\sigma^2)} \ln L = 0$$

Check that L is concave at the solution (Hessian matrix)

Conclusion: The MLEs are

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$\hat{\mu}$ is unbiased: $E(\hat{\mu}) = \mu$

$\hat{\sigma}^2$ is not unbiased: $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$

For this reason, we use $S^2 \equiv \frac{n}{n-1}\hat{\sigma}^2$ instead of $\hat{\sigma}^2$

Calculus cannot always be used to find MLEs

Example: "... cosmic ray composition ..."

Parameter: $\theta > 0$

Model: $f(x;\theta) = \begin{cases} \exp(-(x-\theta)), & x \ge \theta \\ 0, & x < \theta \end{cases}$

Random sample: $X_1, \ldots, X_n$

$$L(\theta) = f(X_1;\theta) \cdots f(X_n;\theta)$$

$$= \begin{cases} \exp(-\sum_{i=1}^{n}(X_i - \theta)), & \text{all } X_i \ge \theta \\ 0, & \text{otherwise} \end{cases}$$

$\hat{\theta} = X_{(1)}$, the smallest observation in the sample

General Properties of the MLE

$\hat{\theta}$ may not be unbiased. We often can remove this bias by multiplying $\hat{\theta}$ by a constant.

For many models, $\hat{\theta}$ is consistent.

The Invariance Property: For many nice functions g, if $\hat{\theta}$ is the MLE of θ then $g(\hat{\theta})$ is the MLE of $g(\theta)$.

The Asymptotic Property: For large n, $\hat{\theta}$ has an approximate normal distribution with mean θ and variance $1/B$ where

$$B = nE\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2$$

The asymptotic property can be used to construct large-sample confidence intervals for θ

The method of maximum likelihood works well when intuition fails and no obvious estimator can be found.

When an obvious estimator exists the method of ML often will find it.

The method can be applied to many statistical problems: regression analysis, analysis of variance, discriminant analysis, hypothesis testing, principal components, etc.

The ML Method for Linear Regression Analysis

Scatterplot data: $(x_1, y_1), \ldots, (x_n, y_n)$

Basic assumption: The x_i's are non-random measurements; the y_i are observations on Y, a random variable

Statistical model: $Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, n$

Errors $\epsilon_1, \ldots, \epsilon_n$: A random sample from $N(0, \sigma^2)$

Parameters: α, β, σ^2

$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$: The Y_i's are independent

The Y_i are not identically distributed; they have differing means

The likelihood function is the joint density of the observed data

$$L(\alpha,\beta,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(Y_i - \alpha - \beta x_i)^2}{2\sigma^2} \right)$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left(- \sum_{i=1}^{n} (Y_i - \alpha - \beta x_i)^2 \Big/ 2\sigma^2 \right)$$

Use calculus to maximize $\ln L$ w.r.t. α, β, σ^2

The ML estimators are:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ , \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

The ML Method for Testing Hypotheses

$X \sim N(\mu, \sigma^2)$

Model: $f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

Random sample: $X_1, \ldots, X_n$

We wish to test $H_0 : \mu = 3$ vs. $H_a : \mu \neq 3$

The space of all permissible values of the parameters

$$\Omega = \{(\mu, \sigma) : -\infty < \mu < \infty, \sigma > 0\}$$

H_0 and H_a represent restrictions on the parameters, so we are led to parameter subspaces

$$\omega_0 = \{(\mu, \sigma) : \mu = 3, \sigma > 0\}, \quad \omega_a = \{(\mu, \sigma) : \mu \neq 3, \sigma > 0\}$$

Construct the likelihood function

$$L(\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \right)$$

Maximize $L(\mu, \sigma^2)$ over ω_0 and then over ω_a

The likelihood ratio test statistic is

$$\lambda = \frac{\max_{\omega_0} L(\mu, \sigma^2)}{\max_{\omega_a} L(\mu, \sigma^2)} = \frac{\max_{\sigma > 0} L(3, \sigma^2)}{\max_{\mu \neq 3, \sigma > 0} L(\mu, \sigma^2)}$$

Fact: $0 \leq \lambda \leq 1$

$L(3, \sigma^2)$ is maximized over ω_0 at

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - 3)^2$$

$$\begin{aligned}\max_{\omega_0} L(3, \sigma^2) &= L\Big(3, \tfrac{1}{n}\sum_{i=1}^{n}(X_i - 3)^2\Big) \\ &= \left(\frac{n}{2\pi e \sum_{i=1}^{n}(X_i - 3)^2}\right)^{n/2}\end{aligned}$$

$L(\mu, \sigma^2)$ is maximized over ω_a at

$$\mu = \bar{X}, \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\begin{aligned}\max_{\omega_a} L(\mu, \sigma^2) =& L\left(\bar{X}, \tfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) \\ &= \left(\frac{n}{2\pi e \sum_{i=1}^{n}(X_i - \bar{X})^2}\right)^{n/2}\end{aligned}$$

The likelihood ratio test statistic:

$$\lambda^{2/n} = \frac{n}{2\pi e \sum_{i=1}^{n}(X_i - 3)^2} \div \frac{n}{2\pi e \sum_{i=1}^{n}(X_i - \bar{X})^2}$$
$$= \sum_{i=1}^{n}(X_i - \bar{X})^2 \div \sum_{i=1}^{n}(X_i - 3)^2$$

λ is close to 1 iff $\bar{X}$ is close to 3

λ is close to 0 iff $\bar{X}$ is far from 3

λ is equivalent to a t-statistic

In this case, the ML method discovers the obvious test statistic

The Bayesian Information Criterion

- Suppose that we have two competing statistical models
- We can fit these models using the methods of least squares, moments, maximum likelihood, . . .
- The choice of model cannot be assessed entirely by these methods
- By increasing the number of parameters, we can always reduce the residual sums of squares
- Polynomial regression: By increasing the number of terms, we can reduce the residual sum of squares
- More complicated models generally will have lower residual errors

BIC: Standard approach to model fitting for *large* data sets

The BIC penalizes models with larger numbers of free parameters

Competing models: $f_1(x; \theta_1, \ldots, \theta_{m_1})$ and $f_2(x; \phi_1, \ldots, \phi_{m_2})$

Random sample: $X_1, \ldots, X_n$

Likelihood functions: $L_1(\theta_1, \ldots, \theta_{m_1})$ and $L_2(\phi_1, \ldots, \phi_{m_2})$

$$\text{BIC} = 2 \ln \frac{L_1(\theta_1, \ldots, \theta_{m_1})}{L_2(\phi_1, \ldots, \phi_{m_2})} - (m_1 - m_2) \ln n$$

The BIC balances an increase in the likelihood with the number of parameters used to achieve that increase

Calculate all MLEs $\hat{\theta}_i$ and $\hat{\phi}_i$ and the estimated BIC:

$$\widehat{\text{BIC}} = 2 \ln \frac{L_1(\hat{\theta}_1, \ldots, \hat{\theta}_{m_1})}{L_2(\hat{\phi}_1, \ldots, \hat{\phi}_{m_2})} - (m_1 - m_2) \ln n$$

General Rules:

$\widehat{\text{BIC}} < 2$: Weak evidence that Model 1 is superior to Model 2

$2 \leq \widehat{\text{BIC}} \leq 6$: Moderate evidence that Model 1 is superior

$6 < \widehat{\text{BIC}} \leq 10$: Strong evidence that Model 1 is superior

$\widehat{\text{BIC}} > 10$: Very strong evidence that Model 1 is superior

Competing models for GCLF in the Galaxy

1. A Gaussian model (van den Bergh 1985, ApJ, 297)

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

2. A t-distn. model (Secker 1992, AJ 104)

$$g(x;\mu,\sigma,\delta) = \frac{\Gamma(\frac{\delta+1}{2})}{\sqrt{\pi\delta}\,\sigma\,\Gamma(\frac{\delta}{2})}\left(1+\frac{(x-\mu)^2}{\delta\sigma^2}\right)^{-\frac{\delta+1}{2}}$$

$$-\infty < \mu < \infty, \sigma > 0, \delta > 0$$

In each model, μ is the mean and σ^2 is the variance

In Model 2, δ is a shape parameter

We use the data of Secker (1992), Table 1

We *assume* that the data constitute a *random sample*

ML calculations suggest that Model 1 is inferior to Model 2

Question: Is the increase in likelihood due to larger number of parameters?

This question can be studied using the BIC

Test of hypothesis

H_0: Gaussian model vs. H_a: t- model

TABLE 1. Milky Way sample.

M_V	R_{GC}	E_{B-V}	Name	M_V	R_{GC}	E_{B-V}	Name
-1.75	26.74	0.06	AM4	-7.37	3.50	0.42	N6712
-3.28	26.28	0.02	Pal13	-7.38	5.62	0.10	N6652
-3.86	8.62	0.30	E3	-7.42	4.01	0.08	N6809
-3.88	2.76	0.31	E452SC	-7.43	4.97	1.05	N6553
-4.08	14.93	0.02	Pal12	-7.45	35.39	0.05	N7006
-4.30	2.81	0.19	N6496	-7.48	24.49	0.10	N5694
-4.54	7.19	0.34	Pal11	-7.52	5.35	0.05	N6752
-4.91	16.03	0.03	Pal5	-7.55	13.91	0.27	IC4499
-5.20	6.74	0.26	N6838	-7.57	17.11	0.00	N1261
-5.24	5.14	0.30	Pal8	-7.64	6.97	0.45	N4372
-5.28	20.89	0.11	Arp2	-7.64	7.01	0.02	N7099
-5.52	23.35	0.01	N7492	-7.65	5.26	0.62	N6760
-5.57	35.64	0.40	Pal15	-7.65	20.78	0.05	N5634
-5.87	19.31	0.02	N4147	-7.70	6.27	0.27	N6284
-5.89	2.99	0.37	N6642	-7.77	2.17	0.37	N6293
-5.89	3.89	0.33	N6535	-7.77	9.84	0.03	N4590
-6.04	4.79	0.65	N6366	-7.79	2.65	0.74	N6139
-6.07	2.42	0.80	N6256	-7.85	2.86	0.22	N6093
-6.14	15.12	0.15	N2298	-7.85	18.59	0.01	N1904
-6.16	5.55	0.73	N6544	-7.86	3.20	0.38	N6273
-6.20	3.62	0.25	N6352	-7.86	9.18	0.04	N362
-6.24	2.04	0.62	N6528	-7.88	2.50	0.03	N6723
-6.32	12.19	0.37	N6426	-7.97	28.16	0.01	N6229
-6.41	22.21	0.10	Rp106	-7.98	2.05	0.32	N6333
-6.45	2.11	0.86	N6325	-7.99	4.74	0.56	N5946
-6.48	6.85	0.32	N4833	-8.02	2.57	0.37	N6626
-6.49	11.33	0.03	N288	-8.04	9.40	0.02	N6341
-6.53	5.04	0.08	N6362	-8.14	11.71	0.16	N6864
-6.54	2.01	0.40	N6342	-8.19	4.38	0.21	N6218
-6.66	2.32	0.20	N6717	-8.20	7.27	0.24	N5286
-6.68	4.58	0.43	N5927	-8.23	15.97	0.02	N1851
-6.76	3.50	0.33	N6171	-8.24	4.76	0.25	N5986
-6.80	2.78	0.61	N6453	-8.27	2.27	0.12	N6541
-6.91	16.66	0.00	N5466	-8.29	18.51	0.14	N5824
-6.93	11.40	0.04	N6101	-8.35	5.06	0.35	N6656
-6.95	3.03	0.35	N6144	-8.40	8.22	0.03	N6205
-6.95	12.39	0.03	N6981	-8.59	7.53	0.30	N6356
-6.96	6.61	0.10	N5897	-8.60	3.40	0.87	N6539
-6.97	2.79	0.52	N6304	-8.65	11.68	0.00	N5272
-7.03	2.25	0.38	N6235	-8.70	19.13	0.00	N5024
-7.04	6.05	0.18	N6397	-8.73	6.12	0.03	N5904
-7.08	16.56	0.02	N5053	-8.80	4.32	0.48	N6316
-7.17	8.82	0.21	N3201	-8.82	10.41	0.02	N7089
-7.18	3.14	1.09	N6517	-9.04	10.17	0.10	N7078
-7.19	9.29	0.21	N6779	-9.08	4.18	0.58	N6402
-7.26	11.67	0.11	N6934	-9.24	7.34	0.04	N104
-7.27	6.13	0.36	N6121	-9.25	10.85	0.22	N2808
-7.31	4.00	0.10	N6584	-9.33	13.13	0.14	N6715
-7.34	4.58	0.25	N6254	-9.34	3.38	0.35	N6388
-7.37	3.46	0.05	N6681	-10.28	6.34	0.11	N5139

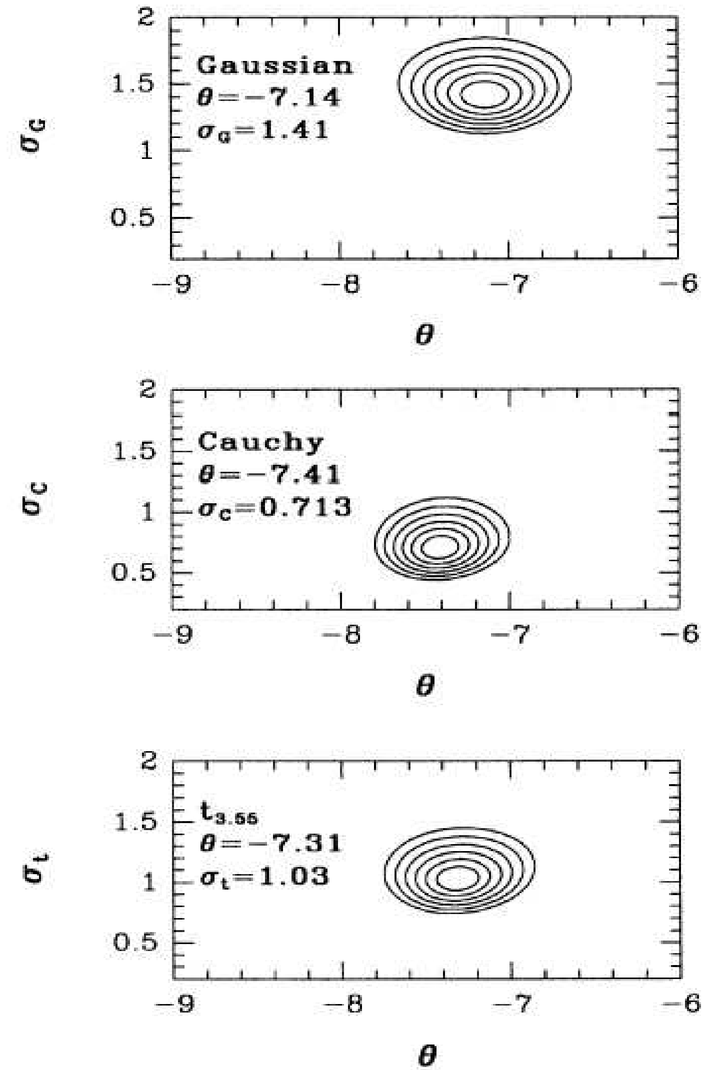

FIG. 1. Maximum-likelihood estimates for the Galactic GCLF expressed as contour plots in a two-dimensional parameter space, for the three distribution functions being considered. The most probable values for the parameters are given in the top left corner of the plot. The contours represent, from inner to outer, the 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0 standard deviation probability limits on the maximum-likelihood parameter estimates.

Model 1: Write down the likelihood function,

$$L_1(\mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \Big)$$

$\hat{\mu} = \bar{X}$, the ML estimator

$\hat{\sigma}^2 = S^2$, a multiple of the ML estimator of σ^2

$$L_1(\bar{X}, S) = (2\pi S^2)^{-n/2} \exp(-(n-1)/2)$$

For the Milky Way data, $\bar{x} = -7.14$ and $s = 1.41$

Secker (1992, p. 1476): $\ln L_1(-7.14, 1.41) = -176.4$

Model 2: Write down the likelihood function

$$L_2(\mu, \sigma, \delta) = \prod_{i=1}^{n} \frac{\Gamma(\frac{\delta+1}{2})}{\sqrt{\pi\delta}\,\sigma\,\Gamma(\frac{\delta}{2})} \left(1 + \frac{(X_i - \mu)^2}{\delta\sigma^2}\right)^{-\frac{\delta+1}{2}}$$

Are the MLEs of μ, σ^2, δ unique?

No explicit formulas for them are known; we evaluate them numerically

Substitute the Milky Way data for the X_i's in the formula for L, and maximize L numerically

Secker (1992): $\hat{\mu} = -7.31$, $\hat{\sigma} = 1.03$, $\hat{\delta} = 3.55$

Secker (1992, p. 1476): $\ln L_2(-7.31, 1.03, 3.55) = -173.0$

Finally, calculate the estimated BIC: With $m_1 = 2$, $m_2 = 3$, $n = 100$

$$\widehat{\text{BIC}} = 2 \ln \frac{L_1(-7.14, 1.41)}{L_2(-7.31, 1.03, 3.55)} - (m_1 - m_2)n$$
$$= -2.2$$

Apply the General Rules on p. 26 to assess the strength of the evidence that Model 1 may be superior to Model 2.

Since $\widehat{\text{BIC}} < 2$, we have weak evidence that the t-distribution model is superior to the Gaussian distribution model.

We fail to reject the null hypothesis that the GCLF follows the Gaussian model over the t-model

Concluding General Remarks on the BIC

The BIC procedure is consistent: If Model 1 is the true model then, as $n \to \infty$, the BIC will determine (with probability 1) that it is.

In typical significance tests, any null hypothesis is rejected if n is sufficiently large. Thus, the factor $\ln n$ gives lower weight to the sample size.

Not all information criteria are consistent; e.g., the AIC is not consistent (Azencott and Dacunha-Castelle, 1986).

The BIC is not a panacea; some authors recommend that it be used in conjunction with other information criteria.

There are also difficulties with the BIC

Findley (1991, Ann. Inst. Statist. Math.) studied the performance of the BIC for comparing two models with different numbers of parameters:

“Suppose that the log-likelihood-ratio sequence of two models with different numbers of estimated parameters is bounded in probability. Then the BIC will, with asymptotic probability 1, select the model having fewer parameters.”