Information Criteria: AIC and BIC

Statistics · Module 10 — Model Selection Learning

Node 10.2

Position in Knowledge Graph

System: Statistics. Structural Domain: Model Selection and Learning (Module 10). Node 10.2.

Information criteria convert the bias–variance tradeoff (Node 10.1) into explicit model-ranking formulas by penalizing the maximized log-likelihood for the number of free parameters. AIC targets predictive accuracy via KL divergence minimization; BIC approximates the Bayesian marginal likelihood. Together they define the two canonical analytic approaches to model selection, feeding into model averaging (Node 10.6).

Formal Definition

Let $\mathcal{M}_1, \ldots, \mathcal{M}_K$ be a collection of candidate parametric models. Model $\mathcal{M}_j$ has k_j free parameters and MLE $\hat{\theta}^{(j)}$ with maximized log-likelihood $\ell_j = \ell(\hat{\theta}^{(j)}; x)$.

Akaike Information Criterion:

$$\mathrm{AIC}_j = -2\ell_j + 2k_j.$$

Bayesian Information Criterion:

$$\mathrm{BIC}_j = -2\ell_j + k_j \log n.$$

The preferred model minimizes the criterion. Both penalize the log-likelihood for model complexity, differing only in the penalty weight per parameter.

Intuition

The maximized log-likelihood ℓ_j always improves as parameters are added, but this improvement partly reflects fitting noise rather than signal. Information criteria correct for this optimism by adding a penalty that grows with the number of parameters. AIC adds a fixed penalty of 2 per parameter, derived from the expected optimism of in-sample log-likelihood. BIC adds a penalty of $\log n$ per parameter, derived from Bayesian model comparison. For $n \geq 8$, BIC penalizes more heavily and selects simpler models.

The choice between them reflects the inferential goal: AIC for prediction, BIC for identifying the true model.

Structural Role

Dependencies: Builds on the bias–variance tradeoff (Node 10.1), the KL divergence (Node 2.3), and asymptotic likelihood theory (Node 7.6). **Enables:** Model averaging (Node 10.6), where BIC approximations to marginal likelihoods provide the posterior model weights.

AIC and BIC operationalize two philosophically distinct goals. AIC is the frequentist criterion: it targets the model that best predicts new data from the same distribution. BIC is the Bayesian criterion: it targets the model that most likely generated the observed data. Neither dominates the other; the appropriate choice depends on context.

Mathematical Development

Derivation of AIC from KL Divergence

Let p_0 be the true density and p_θ the model density. The KL divergence from p_0 to p_θ is

$$D_{\mathrm{KL}}(p_0 \| p_\theta) = \mathbb{E}_{p_0}\left[\log \frac{p_0(X)}{p_\theta(X)}\right] = \text{const} - \mathbb{E}_{p_0}[\log p_\theta(X)].$$

Minimizing KL divergence over θ is equivalent to maximizing $\mathbb{E}_{p_0}[\log p_\theta(X)]$, the expected log-likelihood under new data. Define

$$\Delta_j = \mathbb{E}_{X^{\text{new}}}[\log p_{\hat{\theta}^{(j)}}(X^{\text{new}})] - \frac{1}{n}\ell_j.$$

This is the **optimism**: how much the in-sample log-likelihood overestimates the out-of-sample log-likelihood. Akaike's key result:

Theorem (Akaike, 1973). Under regularity conditions (model $\mathcal{M}_j$ correctly specified, $\hat{\theta}^{(j)}$ consistent and asymptotically normal),

$$\mathbb{E}[\Delta_j] = -\frac{k_j}{n} + o(1/n).$$

Proof sketch. Expand $\ell(\hat{\theta}^{(j)}; X^{\text{new}})$ around $\theta_0^{(j)}$ to second order. The cross term between the estimation error $\hat{\theta}^{(j)} - \theta_0^{(j)}$ and the score at X^{new} has mean zero. The quadratic term contributes $-\frac{1}{2}(\hat{\theta}^{(j)} - \theta_0^{(j)})^\top I(\theta_0^{(j)})(\hat{\theta}^{(j)} - \theta_0^{(j)})$, which has expectation $-k_j/(2n)$ by the asymptotic distribution of the MLE. Meanwhile, the in-sample quadratic term contributes $+k_j/(2n)$. The total optimism is $-k_j/n$, so

$$\mathbb{E}_{X^{\text{new}}}\left[\frac{1}{n}\ell(\hat{\theta}^{(j)}; X^{\text{new}})\right] \approx \frac{1}{n}\ell_j - \frac{k_j}{n}.$$

Multiplying by $-2n$ gives

$$\mathbb{E}[-2\ell(\hat{\theta}^{(j)}; X^{\text{new}})] \approx -2\ell_j + 2k_j = \mathrm{AIC}_j.$$

Derivation of BIC from Laplace Approximation

The marginal likelihood of model $\mathcal{M}_j$ under prior $\pi(\theta^{(j)})$ is

$$m_j(x) = \int \exp(\ell(\theta^{(j)}; x))\, \pi(\theta^{(j)})\, d\theta^{(j)}.$$

Apply the Laplace approximation: expand $\ell(\theta; x)$ around $\hat{\theta}^{(j)}$ to second order,

$$\ell(\theta; x) \approx \ell_j - \frac{1}{2}(\theta - \hat{\theta}^{(j)})^\top \hat{I}_n (\theta - \hat{\theta}^{(j)}),$$

where $\hat{I}_n = -\nabla^2 \ell(\hat{\theta}^{(j)}; x)$. The Gaussian integral gives

$$m_j(x) \approx \exp(\ell_j)\,(2\pi)^{k_j/2}\,|\hat{I}_n|^{-1/2}\,\pi(\hat{\theta}^{(j)}).$$

Taking $-2\log$:

$$-2\log m_j(x) \approx -2\ell_j + \log|\hat{I}_n| - k_j \log(2\pi) - 2\log \pi(\hat{\theta}^{(j)}).$$

Under standard asymptotics, $\hat{I}_n = n \cdot \hat{I}_1 + O(\sqrt{n})$, so $\log|\hat{I}_n| = k_j \log n + O(1)$. All remaining terms are $O(1)$ as $n \to \infty$. Therefore

$$-2\log m_j(x) = -2\ell_j + k_j \log n + O(1) = \mathrm{BIC}_j + O(1).$$

Properties: Consistency and Efficiency

BIC Consistency. If the true model $\mathcal{M}^*$ is among the candidates, then for any overfitting model $\mathcal{M}_j \supsetneq \mathcal{M}^*$,

$$\mathrm{BIC}_j - \mathrm{BIC}^* = \underbrace{(-2\ell_j + 2\ell^*)}_{O_p(1)} + \underbrace{(k_j - k^*)\log n}_{\to\infty} \to +\infty.$$

So BIC selects the true model with probability $\to 1$. AIC does not: its penalty $2(k_j - k^*)$ is finite, so it selects the overfitting model with non-vanishing probability.

AIC Efficiency. When no candidate model is true (all are approximations), AIC asymptotically selects the model minimizing KL divergence to p_0. BIC may select a model that is too simple.

Corrected AIC

For small n relative to k_j, the $O(1/n)$ approximation in Akaike's theorem is inaccurate. The corrected AIC is

$$\mathrm{AICc}_j = -2\ell_j + 2k_j \cdot \frac{n}{n - k_j - 1}.$$

This is exact for Gaussian linear models; for other models it is an improved approximation. As $n/k_j \to \infty$, $\mathrm{AICc}_j \to \mathrm{AIC}_j$.

Worked Examples

Example 1: Polynomial Regression Model Selection

Data: $n = 30$ observations from $Y = 2 + 3X - X^2 + \varepsilon$, $\varepsilon \sim \mathcal{N}(0,1)$, $X \sim \mathrm{Uniform}(-2,2)$. Fit polynomials of degree 1, 2, 3, 5.

Each model has $k_j = j + 2$ parameters ($j+1$ coefficients plus σ^2). Suppose the maximized log-likelihoods are $\ell_1 = -48.2$, $\ell_2 = -39.1$, $\ell_3 = -38.9$, $\ell_5 = -38.4$.

AIC: $\mathrm{AIC}_1 = 96.4 + 6 = 102.4$; $\mathrm{AIC}_2 = 78.2 + 8 = 86.2$; $\mathrm{AIC}_3 = 77.8 + 10 = 87.8$; $\mathrm{AIC}_5 = 76.8 + 14 = 90.8$. AIC selects degree 2.

BIC ($\log 30 = 3.40$): $\text{BIC}_1 = 96.4 + 10.2 = 106.6$; $\text{BIC}_2 = 78.2 + 13.6 = 91.8$; $\text{BIC}_3 = 77.8 + 17.0 = 94.8$; $\text{BIC}_5 = 76.8 + 23.8 = 100.6$. BIC also selects degree 2.

Both criteria correctly identify the true model order. The degree-3 model gains only $\Delta\ell = 0.2$ over degree 2, insufficient to offset the additional parameter penalty.

Example 2: AIC vs. BIC Disagreement

Data: $n = 500$, true model has 5 predictors. Model A: 5 predictors, $\ell_A = -700.0$, $k_A = 6$. Model B: 8 predictors, $\ell_B = -696.5$, $k_B = 9$.

AIC: $\text{AIC}_A = 1400 + 12 = 1412$; $\text{AIC}_B = 1393 + 18 = 1411$. AIC selects Model B (the larger model) by 1 unit.

BIC ($\log 500 = 6.21$): $\text{BIC}_A = 1400 + 37.3 = 1437.3$; $\text{BIC}_B = 1393 + 55.9 = 1448.9$. BIC selects Model A (the true model).

The three extra parameters in Model B improve the likelihood by 3.5, which exceeds AIC's penalty of 6 but not BIC's penalty of 18.6. This illustrates: AIC favors prediction (Model B is slightly better out-of-sample), BIC favors parsimony (Model A is the true model).

Cross-Links

AI. Neural architecture search and hyperparameter tuning replace analytic criteria with validation-set evaluation, but AIC-like principles (penalizing effective parameters) appear in pruning criteria and the minimum description length principle for model compression.

Economics. AIC and BIC are standard in VAR lag-length selection, ARMA order selection, and structural model comparison. Economists often report both, noting that disagreement signals the prediction-vs-identification tension.

Linear Algebra. The effective degrees of freedom in penalized regression, $\text{df}(\lambda) = \text{Tr}(H_\lambda)$, generalizes the parameter count k and can be substituted into AIC for regularized models. The trace operation connects model complexity to the spectrum of the hat matrix.

Quantum Mechanics. In quantum state tomography, model selection between density matrix parameterizations of different rank uses likelihood ratio tests and information criteria adapted to the positive semidefinite constraint, with the BIC penalty reflecting the dimensionality of the state manifold.

Structural Compression

Invariant

Information criteria decompose model quality into fit ($-2\ell_j$) and a complexity penalty ($2k_j$ or $k_j \log n$); model selection minimizes the sum.

Mechanism

AIC: Akaike's theorem shows the in-sample log-likelihood overestimates the expected out-of-sample log-likelihood by k_j/n in expectation; multiplying by $-2n$ yields the penalty $2k_j$. BIC: the Laplace approximation to the marginal likelihood introduces a $(k_j/2)\log n$ penalty for parameter uncertainty integrated over the prior.

Cross-System Echo

AIC is asymptotically equivalent to LOOCV (Stone, 1977) under squared-error loss for linear models. BIC is the Schwarz approximation to the Bayes factor (Node 6.6) for model comparison. In information theory, both connect to the minimum description length (MDL) principle: the best model compresses the data most efficiently.

Exercises

1. Starting from the KL divergence $D_{\mathrm{KL}}(p_0 \| p_{\hat{\theta}})$, carry out the full Taylor expansion to derive the AIC penalty $2k$. Identify where the asymptotic normality of the MLE is used.
2. Derive the BIC via the Laplace approximation. Show explicitly that the prior density $\pi(\hat{\theta})$ and the term $k \log(2\pi)$ are absorbed into the $O(1)$ remainder.
3. Show that BIC is consistent: if the true model $\mathcal{M}^*$ is among the candidates and has the fewest parameters, then $\mathbb{P}(\hat{\jmath}_{\mathrm{BIC}} = *) \to 1$ as $n \to \infty$.
4. Show that for Gaussian linear regression with known σ^2, AIC reduces to $\mathrm{RSS}/\sigma^2 + 2k$ and BIC reduces to $\mathrm{RSS}/\sigma^2 + k \log n$, both up to additive constants independent of the model.
5. For a dataset with $n = 20$ and two models ($k_1 = 3, \ell_1 = -25.0$; $k_2 = 6, \ell_2 = -21.5$), compute AIC, AICc, and BIC for both models. Which does each criterion prefer?
6. Derive the AICc correction for Gaussian linear regression by computing the exact expected optimism (using the non-central chi-squared distribution of the in-sample RSS) rather than the asymptotic approximation.
7. Argue, structurally, why no single information criterion can simultaneously be consistent (selecting the true model) and efficient (minimizing predictive risk when no model is true), and relate this impossibility to the bias–variance tradeoff.

Statistics · Module 10 — Model Selection Learning · Node 10.2
Concepts: model-selection, likelihood, shannon-entropy, bias-variance-tradeoff
Prerequisites: 10.1, 2.3, 7.6
Enables: 10.6

hbar.university — own.your.cognition