Sensitivity and Specificity of Information Criteria for Model Selection in Prevention and Psychology Datasets

John J. Dziak, Donna L. Coffman, Stephanie T. Lanza, and Runze Li
The Methodology Center, The Pennsylvania State University

Abstract

We simulated case studies in factor analysis and in latent class analysis (LCA), based on prevention and psychological literature, to show how the performance of commonly used information criteria depends on (1) sample size, (2) the complexity of structure to be detected in the true population, and (3) the kind of performance considered most important (e.g., avoidance of false positives, avoidance of false negatives, low estimation error). For large samples, BIC is preferable to AIC because AIC can overfit. However, for small samples, AIC and adjusted BIC do better than BIC or CAIC at describing the true population, due to a tendency for BIC and especially CAIC to underfit.

Background

Several commonly used model selection criteria are equivalent to the penalized fit function $2k + A_n p$, where $k$ is the log-likelihood, $p$ is the number of free parameters in the model, and $A_n$ is a penalty chosen by the theory motivating the criterion. These criteria include the Akaike Information Criterion (AIC; Akaike, 1974), Schwarz’s Bayesian Information Criterion (BIC; 1978), the sample-size-adjusted BIC (see Rissanen, 1978, 2001, and the Consistent AIC (CAIC; Bozdogan, 1986).

These information criteria are sometimes used as heuristics for choosing models. However, they disagree as to which model is best. In deciding how to evaluate them as sources of information, it is useful to think of them as differing approaches to balancing good fit with small model size. The value of $A_n$ controls the degree of emphasis on keeping the model small (see Lin & Dayton, 1997). For comparing a given pair of models of different sizes, they are very similar to likelihood ratio tests, with an unconditional version derived by A. (Pitman, 1991). Thus, when they disagree, the IC’s can be thought of as suggesting a range of model sizes, and the choice among these involves a value judgment comparing the likely costs of a model which is too small or too large in one’s particular situation.

We illustrate the different performances of the IC’s by a set of simulations in the context of choosing the number of factors in factor analysis and the number of latent classes in latent class analysis. Some especially relevant past simulation work was done by Nylund, Asparouhov and Muthén (2007) and Yang & Yang (2007).

Factor Analysis Simulation Example

Settings

We simulated two scenarios: where the true model has 3 factors or 4 factors. In each case we used an analysis of the 24-item psychological test data of Harman (1976) to choose parameters for the population model. We compared the performance of the IC’s for 1000 samples each of size $n$, at each value of $n$ in 50, 60, 70, ..., 600. For each sample, we fit 1, 2, 3, 4, and 5 factor models and observed which model size was selected by each IC.

General form (lower bound) $-2\ell + A_n p$

$\text{AIC} = \hat{A} + 2$

$\text{AIC}^3 = \hat{A} + 3$

$\text{BIC} = \hat{A} + \ln (n)$

$\text{BIC} (\text{ABIC}) = \hat{A} + \ln \left( \frac{n}{2} \cdot \frac{4}{4} \right)$

$\text{CAIC} = \hat{A} + \ln (n + 1)$

Results for Three-Parameter True Model

1. Probability of overfitting (choosing more than the correct number of factors).
2. Probability of underfitting (choosing less than the correct number of factors).
3. Probability of choosing the correct number of factors.
4. Average root mean squared estimation error for the parameters of the population covariance matrix.

Results for Four-Parameter True Model

- Underfitting occurred frequently when the sample size was too small, but rarely otherwise. The required sample size to avoid underfitting depended on the criterion (i.e., higher for BIC than for AIC) and on how complex the data generating model was (i.e., it was harder to detect all of the factors when there was a true but somewhat subtle fourth factor than when there were only three).
- Overfitting was not a problem except in the case of AIC.
- Because AIC was less likely to underfit, it had a slightly better average MSE for modest sample sizes.

Latent Class Analysis Simulation Example

Settings

We simulated two scenarios: where the true model has 3 classes or 4 classes. The parameters for the true models were adapted from estimates presented in Collins & Lanza (2010, p. 12) of 6 dichotomous items on adolescent delinquency from the Add Health public-use data wave, one. 1000 random datasets were generated for each size $n = 50, 100, 150, 200, 300, 400, ..., 3000$ for the four-class true model but only up to 1000 for the three-class true model.

Results for Three-Class True Model

1. Probability of overfitting (choosing more than the correct number of classes).
2. Probability of underfitting (choosing less than the correct number of classes).
3. Probability of choosing the correct number of classes.
4. Average root mean squared estimation error for the probabilities of the cells of the underlying contingency table of possible class types.

Results for Four-Class True Model

- Underfitting rates were high when sample sizes were too small (e.g., in the 2-class case, this corresponded to $n < 200$ for AIC, ABC or AIC$^3$ and to $n < 400$ for BIC and CAIC). The needed sample size to avoid underfitting also depended on the complexity of the true model (the needed sample sizes went up to about 1000 and 2000 for the 4-class case instead of 200 and 400, possibly because the fourth class was rather small).$
-$ Overfitting rates were small except when $n > 300$ or for AIC.
- The accuracy of probability estimates tended to be more heavily impaired by underfitting than by overfitting.

Conclusions

There is no best information criterion for model selection, much as there is no best alpha level for testing. However, some may be better or worse for a given situation and given goals. The criteria each have a different theoretical derivation but algebraically they are just different values of $A_n$, i.e., different tradeoffs between simplicity and fit, or, between specificity and sensitivity.

Although BIC has the theoretical property of "consistency" for large $n$, in several scenarios described above it performs poorly even for modest $n$. It is less likely than AIC to choose an excessively large model, but will sometimes choose a misleadingly simple model.

Limitations

The simulated case studies shown here are only intended as illustrations. It would be impossible to survey every possible scenario, so we do not make general recommendations about which IC to use, or whether to use IC’s at all instead of another approach. One unrealistic feature of all of the simulations is that they assume a literally true data-generating model among the range of models considered. This makes it simple to define overfitting, underfitting, or correct fitting, by comparison to the correct model. In empirical studies with real data, it is harder to describe what these intuitive ideas mean. The squared error measures do not require a parametric true model but do not take interpretability into account. There is no one consistent measure of model performance, which is why there is no one best method for selecting good models.