171:290 Model Selection
Lecture IV: The Takeuchi Information Criterion, TIC

Joseph E. Cavanaugh

Department of Biostatistics
Department of Statistics and Actuarial Science
The University of Iowa

September 18, 2012

Introduction

- The Akaike information criterion, AIC, is derived as an estimator of the expected Kullback discrepancy between the true model and a fitted candidate model.
- The asymptotic justification of the criterion requires two strong assumptions:
  - (i) that the true model is contained in the candidate class under consideration,
  - (ii) that the vector of maximum likelihood estimators satisfies the conventional large-sample properties of MLE's.
- The Takeuchi (1976) information criterion, TIC, is based on a development where the true model assumption (i) is relaxed.

Introduction

Outline:

- Review of AIC
- Framework for TIC
- Derivation of TIC
- Discussion
- Simulation Study to Compare TIC, AIC, MAIC, AICc

Review of AIC

Constructs:

- **True** or **generating model**: $g(y)$.
- **Candidate** or **approximating model**: $f(y|\,\theta_k)$.
- **Candidate class:**

$$\mathcal{F}(k) = \{f(y|\,\theta_k) \mid \theta_k \in \Theta(k)\}\,.$$

- **Fitted model:** $f(y|\,\hat{\theta}_k)$.

Review of AIC

- **Kullback discrepancy** between $g(y)$ and $f(y|\,\hat{\theta}_k)$ with respect to $g(y)$:

$$d(\hat{\theta}_k) = \mathrm{E}\{-2\ln f(y|\,\theta_k)\}|_{\theta_k=\hat{\theta}_k}.$$

- **Expected Kullback discrepancy:**

$$\Delta(k) = \mathrm{E}\{d(\hat{\theta}_k)\}.$$

Review of AIC

- **Expected Fisher Information**:

$$\mathrm{I}(\theta_k) = \mathrm{E}\left[-\frac{\partial^2 \ln f(y|\,\theta_k)}{\partial\theta_k \partial\theta_k'}\right].$$

- **Observed Fisher Information**:

$$\mathcal{I}(\theta_k, y) = \left[-\frac{\partial^2 \ln f(y|\,\theta_k)}{\partial\theta_k \partial\theta_k'}\right].$$

Review of AIC

- In the asymptotic justification of AIC, we impose the assumption that $g(y) \in \mathcal{F}(k)$.
- Thus, the true model or density is a member of the parametric class $\mathcal{F}(k)$, and can therefore be written as $f(y|\,\theta_o)$, where $\theta_o \in \Theta(k)$.
- We also assume that the vector of maximum likelihood estimators satisfies the conventional large-sample properties of MLE's.

Review of AIC

Representation for $\Delta(k)$ leading to AIC:

$$
\begin{aligned}
&\Delta(k) \\
&= \mathrm{E}\{d(\hat{\theta}_k)\} \\
&= \mathrm{E}\{-2 \ln f(y|\,\hat{\theta}_k)\} \\
&\quad + \Big[\mathrm{E}\{-2 \ln f(y|\,\theta_o)\} - \mathrm{E}\{-2 \ln f(y|\,\hat{\theta}_k)\}\Big] \\
&\quad + \Big[\mathrm{E}\{d(\hat{\theta}_k)\} - \mathrm{E}\{-2 \ln f(y|\,\theta_o)\}\Big] \\
&= \mathrm{E}\{-2 \ln f(y|\,\hat{\theta}_k)\} \\
&\quad + \Big[\mathrm{E}\{(\hat{\theta}_k - \theta_o)'\{\mathcal{I}(\hat{\theta}_k, y)\}(\hat{\theta}_k - \theta_o)\}\Big] \\
&\quad + \Big[\mathrm{E}\{(\hat{\theta}_k - \theta_o)'\{\mathrm{I}(\theta_o)\}(\hat{\theta}_k - \theta_o)\}\Big] \\
&\quad + o(1)
\end{aligned}
$$

Review of AIC

$$
\begin{aligned}
&= \mathrm{E}\{-2\ln f(y|\,\hat{\theta}_k)\} \\
&\quad +2\left[\mathrm{E}\{(\hat{\theta}_k-\theta_o)'\{\mathrm{I}(\theta_o)\}(\hat{\theta}_k-\theta_o)\}\right] + o(1) \\
&= \mathrm{E}\{-2\ln f(y|\,\hat{\theta}_k)\} + 2k + o(1)
\end{aligned}
$$

TIC Framework

- In the asymptotic justification of TIC, we do not impose the assumption that $g(y) \in \mathcal{F}(k)$.
- We must therefore consider a framework where the candidate model is improperly specified, perhaps due to an omission of necessary explanatory variables, an inappropriate probabilistic formulation, or an inadequate functional characterization of the deterministic component.
- We will utilize the large-sample properties of maximum likelihood estimators under misspecification.

TIC Framework

- Recall the **Kullback-Leibler Information** between $g(y)$ and $f(y|\,\theta_k)$ with respect to $g(y)$:

$$I(\theta_k) = \mathrm{E}\left\{ \ln \frac{g(y)}{f(y|\,\theta_k)} \right\},$$

where E denotes the expectation under $g(y)$.

- The **pseudo true parameter** is defined as

$$\bar{\theta}_k = \mathrm{argmin}_{\theta_k \in \Theta(k)} I(\theta_k).$$

- When $g(y) \in \mathcal{F}(k)$, we can write $g(y)$ as $f(y|\,\theta_o)$, where $\theta_o \in \Theta(k)$. In this case, $\bar{\theta}_k = \theta_o$.
- When $g(y) \notin \mathcal{F}(k)$, $\bar{\theta}_k$ can be regarded as follows: of all models in $\mathcal{F}(k)$, the model $f(y|\,\bar{\theta}_k)$ provides the best approximation to $g(y)$ (in the sense of Kullback-Leibler information).

TIC Framework

- When $g(y) \notin \mathcal{F}(k)$, the MLE $\hat{\theta}_k$ will converge to the pseudo true parameter $\bar{\theta}_k$.
- When $g(y) \notin \mathcal{F}(k)$, the large-sample variance covariance matrix of $\hat{\theta}_k$ is given by

$$\Sigma(\bar{\theta}_k) = [\mathrm{I}(\bar{\theta}_k)]^{-1}\mathrm{J}(\bar{\theta}_k)[\mathrm{I}(\bar{\theta}_k)]^{-1},$$

  where

$$\mathrm{J}(\theta_k) = \mathrm{E}\left[\left\{\frac{\partial \ln f(y|\,\theta_k)}{\partial\theta_k}\right\}\left\{\frac{\partial \ln f(y|\,\theta_k)}{\partial\theta_k}\right\}'\right].$$

- When $g(y) \notin \mathcal{F}(k)$, under appropriate regularity conditions, the large-sample distribution of $\hat{\theta}_k$ is $N_k(\bar{\theta}_k, \Sigma(\bar{\theta}_k))$.

TIC Framework

- When $g(y) \in \mathcal{F}(k)$, we can write $g(y)$ as $f(y|\,\theta_o)$, with $\theta_o \in \Theta(k)$. We then have $\bar{\theta}_k = \theta_o$, and $\mathrm{I}(\theta_o) = \mathrm{J}(\theta_o)$. It follows that

$$\Sigma(\bar{\theta}_k) = \Sigma(\theta_o) = [\mathrm{I}(\theta_o)]^{-1}\mathrm{J}(\theta_o)[\mathrm{I}(\theta_o)]^{-1} = [\mathrm{I}(\theta_o)]^{-1}.$$

- When $g(y) \in \mathcal{F}(k)$, under appropriate regularity conditions, the large-sample distribution of $\hat{\theta}_k$ is $N_k(\theta_o, [\mathrm{I}(\theta_o)]^{-1})$.

TIC Derivation

- The derivation of TIC is based on the following representation for $\Delta(k)$:

$$
\begin{aligned}
&\Delta(k) \\
&= \mathrm{E}\{d(\hat{\theta}_k)\} \\
&= \mathrm{E}\{-2\ln f(y|\,\hat{\theta}_k)\} \\
&\quad + \left[\mathrm{E}\{-2\ln f(y|\,\bar{\theta}_k)\} - \mathrm{E}\{-2\ln f(y|\,\hat{\theta}_k)\}\right] \qquad (1) \\
&\quad + \left[\mathrm{E}\{d(\hat{\theta}_k)\} - \mathrm{E}\{-2\ln f(y|\,\bar{\theta}_k)\}\right]. \qquad (2)
\end{aligned}
$$

- The following lemma asserts that (1) and (2) are both within $o(1)$ of $\mathrm{tr}\left\{\mathrm{J}(\bar{\theta}_k)[\mathrm{I}(\bar{\theta}_k)]^{-1}\right\}$.

TIC Derivation

Lemma

$$E\{-2\ln f(y|\,\bar{\theta}_k)\} - E\{-2\ln f(y|\,\hat{\theta}_k)\}$$
$$= tr\left\{J(\bar{\theta}_k)[I(\bar{\theta}_k)]^{-1}\right\} + o(1), \tag{1}$$
$$E\{d(\hat{\theta}_k)\} - E\{-2\ln f(y|\,\bar{\theta}_k)\}$$
$$= tr\left\{J(\bar{\theta}_k)[I(\bar{\theta}_k)]^{-1}\right\} + o(1). \tag{2}$$

Proof of Lemma

Proof:

- First, consider taking a second-order expansion of $-2 \ln f(y|\,\bar{\theta}_k)$ about $\hat{\theta}_k$, and evaluating the expectation of the result.
- We obtain

$$\begin{aligned} \mathrm{E}\{-2 \ln f(y|\,\bar{\theta}_k)\} &= \mathrm{E}\{-2 \ln f(y|\,\hat{\theta}_k)\} \\ &\quad + \mathrm{E}\left\{(\hat{\theta}_k - \bar{\theta}_k)' \{\mathcal{I}(\hat{\theta}_k, y)\}(\hat{\theta}_k - \bar{\theta}_k)\right\} \\ &\quad + o(1). \end{aligned}$$

- Thus,

$$\begin{aligned} &\mathrm{E}\{-2 \ln f(y|\,\bar{\theta}_k)\} - \mathrm{E}\{-2 \ln f(y|\,\hat{\theta}_k)\} \\ &= \mathrm{E}\left\{(\hat{\theta}_k - \bar{\theta}_k)' \{\mathcal{I}(\hat{\theta}_k, y)\}(\hat{\theta}_k - \bar{\theta}_k)\right\} + o(1). \qquad (3) \end{aligned}$$

Proof of Lemma

- Next, consider taking a second-order expansion of $d(\hat{\theta}_k)$ about $\bar{\theta}_k$, again evaluating the expectation of the result.
- We obtain

$$
\begin{aligned}
\mathrm{E}\{d(\hat{\theta}_k)\} &= \mathrm{E}\{-2 \ln f(y|\,\bar{\theta}_k)\} \\
&\quad + \mathrm{E}\left\{(\hat{\theta}_k - \bar{\theta}_k)'\{\mathrm{I}(\bar{\theta}_k)\}(\hat{\theta}_k - \bar{\theta}_k)\right\} \\
&\quad + o(1).
\end{aligned}
$$

- Thus,

$$
\begin{aligned}
&\mathrm{E}\{d(\hat{\theta}_k)\} - \mathrm{E}\{-2 \ln f(y|\,\bar{\theta}_k)\} \\
&= \mathrm{E}\left\{(\hat{\theta}_k - \bar{\theta}_k)'\{\mathrm{I}(\bar{\theta}_k)\}(\hat{\theta}_k - \bar{\theta}_k)\right\} + o(1). \qquad (4)
\end{aligned}
$$

Proof of Lemma

- Since $\hat{\theta}_k$ is consistent for $\bar{\theta}_k$, the expectations of the quadratic forms

$$(\hat{\theta}_k - \bar{\theta}_k)' \{\mathcal{I}(\hat{\theta}_k, y)\} (\hat{\theta}_k - \bar{\theta}_k) \text{ and } (\hat{\theta}_k - \bar{\theta}_k)' \{\mathrm{I}(\bar{\theta}_k)\} (\hat{\theta}_k - \bar{\theta}_k)$$

are within $o(1)$ of one another.

Proof of Lemma

- Consider the expectation of the latter quadratic form. We have

$$
\begin{aligned}
&\mathrm{E}\{(\hat{\theta}_k - \bar{\theta}_k)'\{\mathrm{I}(\bar{\theta}_k)\}(\hat{\theta}_k - \bar{\theta}_k)\} \\
&= \mathrm{tr}\left\{\mathrm{I}(\bar{\theta}_k)\mathrm{E}\{(\hat{\theta}_k - \bar{\theta}_k)(\hat{\theta}_k - \bar{\theta}_k)'\}\right\} \\
&= \mathrm{tr}\left\{\mathrm{I}(\bar{\theta}_k)\Sigma(\bar{\theta}_k)\right\} + o(1) \\
&= \mathrm{tr}\left\{\mathrm{I}(\bar{\theta}_k)[\mathrm{I}(\bar{\theta}_k)]^{-1}\mathrm{J}(\bar{\theta}_k)[\mathrm{I}(\bar{\theta}_k)]^{-1}\right\} + o(1) \\
&= \mathrm{tr}\left\{\mathrm{J}(\bar{\theta}_k)[\mathrm{I}(\bar{\theta}_k)]^{-1}\right\} + o(1)
\end{aligned}
$$

- The preceding along with (3) and (4) establishes (1) and (2). $\square$

TIC Derivation

Representation for $\Delta(k)$ leading to TIC:

$$
\begin{aligned}
&\Delta(k) \\
&= \mathrm{E}\{d(\hat{\theta}_k)\} \\
&= \mathrm{E}\{-2\ln f(y|\,\hat{\theta}_k)\} \\
&\quad + \Big[\mathrm{E}\{-2\ln f(y|\,\bar{\theta}_k)\} - \mathrm{E}\{-2\ln f(y|\,\hat{\theta}_k)\}\Big] \\
&\quad + \Big[\mathrm{E}\{d(\hat{\theta}_k)\} - \mathrm{E}\{-2\ln f(y|\,\bar{\theta}_k)\}\Big] \\
&= \mathrm{E}\{-2\ln f(y|\,\hat{\theta}_k)\} \\
&\quad + \Big[\mathrm{E}\{(\hat{\theta}_k - \bar{\theta}_k)'\{\mathcal{I}(\hat{\theta}_k, y)\}(\hat{\theta}_k - \bar{\theta}_k)\}\Big] \\
&\quad + \Big[\mathrm{E}\{(\hat{\theta}_k - \bar{\theta}_k)'\{\mathrm{I}(\bar{\theta}_k)\}(\hat{\theta}_k - \bar{\theta}_k)\}\Big] \\
&\quad + o(1)
\end{aligned}
$$

TIC Derivation

$$
\begin{aligned}
&= \mathrm{E}\{-2 \ln f(y|\,\hat{\theta}_k)\} \\
&\quad +2\left[\mathrm{E}\{(\hat{\theta}_k - \bar{\theta}_k)'\{\mathrm{I}(\bar{\theta}_k)\}(\hat{\theta}_k - \bar{\theta}_k)\}\right] + o(1) \\
&= \mathrm{E}\{-2 \ln f(y|\,\hat{\theta}_k)\} \\
&\quad +2\left[\mathrm{tr}\left\{\mathrm{J}(\bar{\theta}_k)[\mathrm{I}(\bar{\theta}_k)]^{-1}\right\}\right] + o(1)
\end{aligned}
$$

- Note: When $g(y) \in \mathcal{F}(k)$, $\mathrm{I}(\bar{\theta}_k) = \mathrm{J}(\bar{\theta}_k)$. We have $\mathrm{tr}\left\{\mathrm{J}(\bar{\theta}_k)[\mathrm{I}(\bar{\theta}_k)]^{-1}\right\} = k$.

TIC Derivation

TIC is defined by Takeuchi (1976) as follows:

$$\text{TIC} = -2 \ln f(y|\,\hat{\theta}_k) + 2\left[\widehat{\text{tr}}\left\{\text{J}(\hat{\theta}_k)[\text{I}(\hat{\theta}_k)]^{-1}\right\}\right],$$

where

$$\left[\widehat{\text{tr}}\left\{\text{J}(\hat{\theta}_k)[\text{I}(\hat{\theta}_k)]^{-1}\right\}\right]$$

is an estimator of

$$\left[\text{tr}\left\{\text{J}(\bar{\theta}_k)[\text{I}(\bar{\theta}_k)]^{-1}\right\}\right].$$

Penalty Term of TIC

- Recall the definitions for $\mathrm{J}(\hat{\theta}_k)$ and $\mathrm{I}(\hat{\theta}_k)$:

$$\begin{aligned}
\mathrm{I}(\hat{\theta}_k) &= \mathrm{E}\left[-\frac{\partial^2 \ln f(y|\,\theta_k)}{\partial\theta_k\partial\theta_k'}\right]\Big|_{\theta_k=\hat{\theta}_k}, \\
\mathrm{J}(\hat{\theta}_k) &= \mathrm{E}\left[\left\{\frac{\partial \ln f(y|\,\theta_k)}{\partial\theta_k}\right\}\left\{\frac{\partial \ln f(y|\,\theta_k)}{\partial\theta_k}\right\}'\right]\Big|_{\theta_k=\hat{\theta}_k}.
\end{aligned}$$

- In general, $\mathrm{J}(\hat{\theta}_k)$ and $\mathrm{I}(\hat{\theta}_k)$ will depend on $g(y)$, and will not be directly accessible. $\mathrm{J}(\hat{\theta}_k)$ and $\mathrm{I}(\hat{\theta}_k)$ must be estimated to evaluate the penalty term of TIC.

Penalty Term of TIC

- The observed Fisher information evaluated at $\hat{\theta}_k$, $\mathcal{I}(\hat{\theta}_k, y)$, is often used to estimate $\mathrm{I}(\hat{\theta}_k)$.
- The estimation of $\mathrm{J}(\hat{\theta}_k)$ is less straightforward.
- Suppose that the collection of data y is comprised of n independent elements: $y = \{y_1, y_2, \ldots, y_n\}$. In this setting, the score vector

$$\mathrm{u}(\theta_k | y) = \frac{\partial \ln f(y | \theta_k)}{\partial \theta_k}$$

is represented by the sum of n contributions:

$$\mathrm{u}(\theta_k | y) = \sum_{i=1}^{n} \mathrm{u}_i(\theta_k | y_i), \text{ where } \mathrm{u}_i(\theta_k | y_i) = \frac{\partial \ln f(y_i | \theta_k)}{\partial \theta_k}.$$

Penalty Term of TIC

- Note that we can write $\mathrm{J}(\hat{\theta}_k)$ as

$$\mathrm{J}(\hat{\theta}_k) = \mathrm{E}\left[\mathrm{u}(\theta_k|\, y)\mathrm{u}(\theta_k|\, y)'\right]\Big|_{\theta_k=\hat{\theta}_k}.$$

- The matrix

$$\sum_{i=1}^{n} \mathrm{u}_i(\hat{\theta}_k|\, y_i)\mathrm{u}_i(\hat{\theta}_k|\, y_i)'$$

  is often used to estimate $\mathrm{J}(\hat{\theta}_k)$.

- Question: Why not use

$$\mathrm{u}(\hat{\theta}_k|\, y)\mathrm{u}(\hat{\theta}_k|\, y)' = \left\{\sum_{i=1}^{n} \mathrm{u}_i(\hat{\theta}_k|\, y_i)\right\}\left\{\sum_{i=1}^{n} \mathrm{u}_i(\hat{\theta}_k|\, y_i)'\right\}?$$

Penalty Term of TIC

- Bootstrap estimators of $\text{tr}\left\{ \text{J}(\bar{\theta}_k)[\text{I}(\bar{\theta}_k)]^{-1} \right\}$ have also been proposed. For example, see Cavanaugh and Shumway, 1997; Ishiguro, Sakamoto, and Kitagawa, 1997; Shibata, 1997; Burnham and Anderson, 2002.

Penalty Term of TIC

- An alternative approach is built upon the same assumption used in the development of MAIC.
- Let $\mathcal{F} = \{\mathcal{F}(k_1), \mathcal{F}(k_2), \ldots, \mathcal{F}(k_L)\}$ represent the candidate family.
- Assume that the largest candidate class in $\mathcal{F}$ is $\mathcal{F}(K)$ (i.e., $K = \max\{k_1, k_2, \ldots, k_L\}$).
- Assume that $g(y) \in \mathcal{F}(K)$.
- $g(y)$ can therefore be written as $f(y|\,\theta_o)$, where $\theta_o \in \Theta(K)$.
- Under the preceding assumption, the largest candidate model, $f(y|\,\theta_K)$, is either correctly specified or overspecified.
- Let $\hat{\theta}_K$ denote the MLE of θ_K.
- θ_o can be consistently estimated with $\hat{\theta}_K$.

Penalty Term of TIC

- Let E_{θ_o} denote the expectation under $f(y|\,\theta_o)$. Denote the previous information matrices as follows:

$$
\begin{aligned}
\mathrm{I}(\theta_o, \theta_k) &= \mathrm{E}_{\theta_o}\left[-\frac{\partial^2 \ln f(y|\,\theta_k)}{\partial\theta_k \partial\theta_k'}\right], \\
\mathrm{J}(\theta_o, \theta_k) &= \mathrm{E}_{\theta_o}\left[\left\{\frac{\partial \ln f(y|\,\theta_k)}{\partial\theta_k}\right\}\left\{\frac{\partial \ln f(y|\,\theta_k)}{\partial\theta_k}\right\}'\right].
\end{aligned}
$$

- The estimator of $\left[\mathrm{tr}\left\{\mathrm{J}(\theta_o, \bar{\theta}_k)[\mathrm{I}(\theta_o, \bar{\theta}_k)]^{-1}\right\}\right]$ is found by computing

$$
\left[\mathrm{tr}\left\{\mathrm{J}(\hat{\theta}_K, \hat{\theta}_k)[\mathrm{I}(\hat{\theta}_K, \hat{\theta}_k)]^{-1}\right\}\right].
$$

Penalty Term of TIC

- In the normal linear regression setting, when the penalty term of TIC is evaluated via $2\left[\text{tr}\left\{\text{J}(\hat{\theta}_K, \hat{\theta}_k)[\text{I}(\hat{\theta}_K, \hat{\theta}_k)]^{-1}\right\}\right]$, we obtain

$$\text{TIC} = -2\ln f(y|\,\hat{\theta}_k) + 2\left[\frac{\hat{\sigma}_*^2}{\hat{\sigma}^2}\left\{(p+2) - \frac{\hat{\sigma}_*^2}{\hat{\sigma}^2}\right\}\right].$$

- Here, p denotes the rank of the design matrix for the candidate model of interest.
- $\hat{\sigma}_*^2$ denotes the maximum likelihood estimator of the error variance associated with the largest candidate model, $f(y|\,\theta_K)$.

Penalty Term of TIC

- Compare to MAIC:

$$\begin{aligned}\text{MAIC} &= -2\ln f(y|\,\hat{\theta}_k) + \frac{2n(p+1)}{(n-p-2)} \\ &+ \left[2p\left\{\frac{(n-p)\hat{\sigma}_*^2}{(n-P)\hat{\sigma}^2} - 1\right\} - 2\left\{\frac{(n-p)\hat{\sigma}_*^2}{(n-P)\hat{\sigma}^2} - 1\right\}^2\right].\end{aligned}$$

- Here, P denote the rank of the design matrix for the largest candidate model.

Properties of AIC, AICc, MAIC, and TIC

The following table illustrates the relationships among AIC, AICc, MAIC, and TIC as estimators of $\Delta(k)$.

	Assumes $g(y) \in \mathcal{F}(k)$	Relaxes $g(y) \in \mathcal{F}(k)$ assumption
Requires large samples	AIC	TIC
Designed for smaller samples	AICc	MAIC

TIC versus AIC

- Under the assumption that $g(y) \in \mathcal{F}(k)$, we have $g(y) = f(y|\,\bar{\theta}_k)$, and $\text{tr}\left\{\text{J}(\bar{\theta}_k)[\text{I}(\bar{\theta}_k)]^{-1}\right\} = k$.
- If $g(y)$ is well approximated by $f(y|\,\bar{\theta}_k)$, then $\text{tr}\left\{\text{J}(\bar{\theta}_k)[\text{I}(\bar{\theta}_k)]^{-1}\right\} \approx k$.
- If $g(y)$ is not well approximated by $f(y|\,\bar{\theta}_k)$, then $\text{tr}\left\{\text{J}(\bar{\theta}_k)[\text{I}(\bar{\theta}_k)]^{-1}\right\}$ could be quite different than k.
    - A data-dependent estimator of $\text{tr}\left\{\text{J}(\bar{\theta}_k)[\text{I}(\bar{\theta}_k)]^{-1}\right\}$ might be substantially less biased than k.
    - However, a data-dependent estimator might also be highly variable.

TIC versus AIC

- Suppose the candidate collection of models contains some models that serve as good approximations of $g(y)$ and some that are badly misspecified.
- For the badly misspecified models, the goodness-of-fit term of the criterion will be inflated, leading to large values of the criterion.
- The penalty term of the criterion is pertinent in delineating among the models that are close to $g(y)$.
- Among models that are close to $g(y)$, the penalty term of AIC should provide a good asymptotic approximation to the true bias adjustment.

Simulation Study

- In simulation studies conducted in the linear regression setting, we often define the **signal-to-noise ratio** as the variance of the linear form in the regressor variables over the variance of the error component.
- In traditional regression applications, the linear form in the regressors is regarded as deterministic and thereby has a variance of zero.
- However, in simulation studies, this SNR definition is sensible since the regressors are randomly generated.
- The SNR definition is amenable to a familiar interpretation: if a correctly specified model is fit to data generated under a true model with a signal-to-noise ratio of SNR, the coefficient of determination for the fit will be approximately SNR/(1+SNR).

Simulation Study

- In simulation studies to evaluate the performance of model selection criteria, two common settings tend to promote a high frequency of underfitted selections: (1) small sample sizes, (2) low signal-to-noise ratios.
- Critics of model selection criteria with data-dependent penalty terms sometimes argue that such estimators of the bias adjustment are highly inaccurate in settings conducive to underfitting.
- Do the stochastic penalty terms of TIC/MAIC yield a practical improvement over the nonstochastic penalty terms of AIC/AICc?

Simulation Study

Study Outline:

- In each of three simulation sets, one thousand samples of size $n = 40$ are generated from a true regression model which has an $n = 40$ by $p_o = 5$ design matrix, and a parameter vector of the form $\beta_o = (1, 1, 1, 1, 1)'$.
- For every sample, candidate models with nested design matrices of ranks $p = 2, 3, \ldots, P = 8$ are fit to the data.
  - The first column of every design matrix is a vector of ones.
  - The design matrix of rank $p_o = 5$ is correctly specified.
- The covariates are generated as *iid* replicates from a uniform $(0, 10)$ distribution.

Simulation Study

- In the three simulation sets, the error variance σ_o^2 is chosen to produce signal-to-noise ratios (SNR's) of 0.5, 1.0, and 1.5.
- We examine the effectiveness of AIC, AICc, TIC, and MAIC at selecting p, the order of the model.

Simulation Study

Set I: Order selections with SNR = 0.5.

p	AIC	AICc	TIC	MAIC
2	35	72	36	83
3	47	91	62	103
4	125	175	150	200
5	495	536	568	531
6	111	71	87	47
7	95	35	60	26
8	92	20	37	10

Simulation Study

Set II: Order selections with SNR = 1.0.

p	AIC	AICc	TIC	MAIC
2	2	10	3	12
3	8	11	10	18
4	45	69	57	85
5	606	736	697	777
6	149	102	121	68
7	112	51	72	28
8	78	21	40	12

Simulation Study

Set III: Order selections with SNR = 1.5.

p	AIC	AICc	TIC	MAIC
2	0	0	0	0
3	0	0	0	0
4	2	5	4	6
5	696	846	792	896
6	147	93	115	71
7	82	34	50	19
8	73	22	39	8

Simulation Study

Conclusions:

- In settings where SNR is low or the sample size is small, TIC and MAIC tend to choose underfitted models more frequently than AIC and AICc (respectively).
- For a fixed sample size, as SNR grows, the underfitting propensities of the criteria are attenuated. TIC and MAIC tend to outperform AIC and AICc (respectively).
- (Not illustrated in study.) For a fixed SNR, as the sample size grows, the probability of the criteria choosing an underfitted model converges to zero, and TIC, MAIC, AICc, and AIC all exhibit the same selection properties.