Paper 099-2018

Agreeing to Disagree: Using SAS® to Make Reasoned Decisions When Information Criteria Select Different Models

Wendy Christensen, University of California, Los Angeles

ABSTRACT

Information criteria are a useful and flexible tool for model selection when the null hypothesis testing framework is unsuitable or undesirable. Over 30 procedures in SAS/STAT® 14.3 provide information criteria for fitted models automatically. Typically, the model with the lowest information criterion value is considered the "best" model, the model with the second-lowest value is considered the "second-best" model, and so on. In practice, many statistical modelers use multiple information criteria (e.g. AIC and BIC) to select among a set of candidate models. It is possible, however, for different information criteria to rank models differently, potentially leading to seemingly conflicting results. The focus of this breakout session is twofold. First, I provide a conceptual overview of why differences in model ranking occur and describe practical approaches to resolving disagreements. Second, I demonstrate how to use Base SAS® functions to identify if ranking differences have occurred among a set of fitted candidate models and to compute useful summary output to assist with making a well-reasoned decision about which model to select.

INTRODUCTION

The scientific method is generally conceptualized as an iterative process of collecting observations, positing hypotheses, and evaluating those hypotheses by comparing them to new observations. As part of the cyclic nature of this process, it is common for researchers to develop more than one theory about the underlying mechanisms of a given phenomenon. Because researchers often express theories as statistical models, model selection methods have become a vital part of the scientific method as it is currently practiced. Introductory statistics texts routinely describe methods of model selection that are based in the null hypothesis testing framework (e.g. R^2 change tests), but compelling alternatives exist for situations where the restrictions of null hypothesis testing are undesirable.

One such alternative is the use of information criteria, which are derived from an information-theoretic framework of model selection. Model selection using information criteria offers much greater flexibility compared to null hypothesis testing methods in three key ways (Hamaker, van Huttum, Kuiper, & Hoijtink, 2010). First, the models being compared do not have to be nested; that is, the only thing the models need to have in common is the outcome variable. Second, because models do not have to be compared in a pairwise fashion, any number of models can be compared simultaneously. Finally, there is no need to designate any model as a null model (or fit a model solely for that purpose) as part of the model selection process.

Model selection using information criteria, as described in Burnham and Anderson's (2002) widely-cited text, may be conceptualized as a four-step process. First, the researcher defines a set of candidate models, each of which is a statistical representation one of the theories under consideration. Second, using the same outcome variable *Y* across all models, each of the candidate models is fit to the same data. Third, one or more information criteria are computed for each model. Finally, models are compared by ranking them based on the computed information criteria values, such that the model with the smallest information criterion value is the "best" model, the model with the second smallest value is the "second-best" model, and so on. If multiple information criteria are computed for each model, then the ranking of models is done within-criterion. For example, if both AIC and BIC are being used, then there will be a set of AIC-based ranks and a set of BIC-based ranks.

Given that the use of multiple information criteria will produce multiple sets of model rankings, it is then possible for models to have differences in between-criterion ranks? In other words, is it possible for information criteria to disagree? The answer is a definitive yes! In fact, sharp-eyed readers of research literature will almost certainly find examples of this occurring in models fit to real data (e.g. Bauer, Gottfredson, Dean, & Zucker, 2013; Salom, Betts, Williams, Najman, & Alati, 2015). This paper is divided

into three parts. The first is a conceptual overview of why information criteria sometimes select different candidate models. The second part is a brief discussion of a few potential approaches to resolving disagreement among information criteria. The third part is a demonstration of how to use Base SAS functionality to help with making reasoned decisions when this occurs.

SOURCES OF (POTENTIAL) DISAGREEMENT

Many information criteria have been developed since Akaike (1974) introduced what is now known as the Akaike Information Criterion (AIC). Table 1 shows the four most frequently produced information criteria in SAS/STAT 14.3 procedures (for more detail about how this was determined, see Christensen, 2018). Schwarz (1978) proposed another information criterion, BIC, which was derived based on the asymptotic behavior of Bayes estimators. AICC (Hurvich & Tsai, 1989) and CAIC (Bozdogan, 1987) are both modifications to AIC, the former of which was developed for smaller sample sizes and the latter of which was developed to reduce overparameterization (i.e. overfitting).

Efficient information criteria		Consistent information criteria	
AIC	$-2LL + 2K$	BIC	$-2LL + \ln(N)K$
AICC	$-2LL + 2K\left(\frac{N}{N-K-1}\right)$	CAIC	$-2LL + [\ln(N) + 1]K$

Table 1. Formulae for two efficient (AIC, AICC) and two consistent (BIC, CAIC) information criteria

There are a few notable similarities among the formulae shown in Table 1. One similarity is the model-specific deviance, denoted *-2LL*. Another similarity is the model-specific number of parameters, denoted *K*. When computing different information criteria for the same candidate model, each information criterion will have the deviance and the number of parameters in common. Finally, three of the information criteria – AICC, BIC, and CAIC – include the sample size in the data set, denoted *N*. Because each candidate model is fit to the same data, this will be the same across candidate models. The key difference between different information criteria is in how the different "inputs" from each candidate model are used as part of the computation.

Multiple information criteria exist because each one was developed to optimize something different than the others. A major point of comparison among information criteria is whether a criterion is asymptotically *efficient* or *consistent*. The difference between these relates to the assumptions made about the *dimensionality* (i.e. the number of parameters) of the "true" or "generating" model and the set of candidate models. Efficient information criteria were optimized for situations where the generating model is of infinite dimension, while consistent information criteria were optimized for situations where the generating model is of finite dimension. Stated differently, the former presumes that the generating model from which data emerge is too complex to correctly specify, while the latter presumes that the generating model can be correctly specified and could be included in the set of candidate models (Aho, Derryberry, & Peterson, 2014).

AIC and AICC are examples of efficient information criteria, while BIC and CAIC are examples of consistent information criteria. A common argument for the use of efficient information criteria is that the assumption of a generating model of infinite dimension is more realistic in practice; that is, reality is often too complex to correctly specify in a statistical model (Burnham & Anderson, 2002; Hurvich & Tsai, 1989). In contrast, a common argument for the use of consistent criteria is that, as the sample size approaches infinity, they have a 100% probability of selecting the generating model as long as it is in the set of candidate models (Bozdogan, 1987). Ultimately, whether efficient or consistent information criteria are more appropriate for model selection depends on the context of the research being conducted (Aho, Derryberry, & Peterson, 2014).

To demonstrate how the computational differences resulting from these differences in optimization could potentially lead to different model rankings, the top of Table 2 shows different "inputs" for three hypothetical candidate models. The candidate models are ordered by model complexity, such that Model 1 has the lowest deviance and the highest number of parameters, Model 3 has the highest deviance and

lowest number of parameters, and Model 2 is roughly in-between. Because candidate models are always fit to the same data, the sample size is the same for the information criteria that incorporate it. The bottom of Table 2 shows the resulting information criteria values for the three hypothetical models. Model 1 has the lowest AIC (340) and AICC (~341) values of the three models. Model 2, however, has the lowest BIC (~412) and CAIC (~422) values. Disagreement has occurred, and a researcher in this situation may wonder what to do next.

	Model 1	Model 2	Model 3
Deviance	300	350	400
K	20	10	5
N	500	500	500
AIC	340	370	410
AICC	341.753653	370.449898	410.121457
BIC	424.292162	412.146081	431.07304
CAIC	444.292162	422.146081	436.07304

Table 2. Hypothetical models that result in selection disagreement, selected models shaded

APPROACHES TO MODEL SELECTION DISAGREEMENT

To date, there do not appear to be any universally-applicable recommendations for what to do when information criteria select different models. In fact, it is likely that a universally-applicable solution does not exist. Instead, what follows in this section is a *descriptive* overview of potential approaches available if the information criteria being used for model selection disagree. It is important to note that this is not intended to be an exhaustive list of all potential approaches. Given this, the presence or absence of a method from this section should not be taken as a judgement about the utility of the approach.

One potential approach is to choose one information criterion to select among a set of candidate models. A major advantage of this method is that it entirely eliminates the problem of selection disagreement; that is, it takes two (or more) to disagree. If this approach is taken, then the choice of which information criterion to use for model selection must be considered carefully. This choice could be made based on which information criterion one believes is most appropriate based on its mathematical properties and for what it optimizes (Aho, Derryberry, & Peterson, 2014). Another source of guidance could be simulation studies of the performance of different information criteria for the types of models being fit, the specification of those models, and the available sample size.

It may also be possible to resolve disagreement by ascertaining if the differences in ranks are due to small differences among the within-criterion values. These differences can be assessed in two ways: by examining the simple differences or by computing model weights (Burnham & Anderson, 2002). To find the simple differences, denoted Δ_i, each information criterion value is subtracted by that of the model with the lowest value. For example, if the lowest criterion value is 100, the second lowest is 105, and the third lowest is 110, then the simple differences of these models are 0, 5, and 10, respectively. The model weights (denoted w_i) are computed for each model (denoted *r*) in a set of candidate models (denoted *R*) as follows:

$$w_i = \frac{e^{-\frac{1}{2}\Delta_i}}{\sum_{r=1}^{R} e^{-\frac{1}{2}\Delta_r}}$$ (p. 75).

Small simple differences and small model weights suggest weaker support of a given model over other models in the set, while larger differences and weights suggest stronger support. The advantage of evaluating model differences is that the disagreement among information criteria may be resolvable if the

differences in the ranks were due to small differences in the within-criterion values. The disadvantage is that this approach is only tenable in the specific case of small differences in information criteria values.

Another potential approach is to compute multiple information criteria and select the model that a simple majority of the computed criteria indicates as the best model. This approach is akin to a popularity contest or a first-past-the-post voting system, such that the model that gets the most "votes" is considered the best model. An advantage of this method is that a model selected by multiple information criteria may be a more robust model across different kinds of optimization, especially if both efficient and consistent criteria are used (Kuha, 2004). A disadvantage of this method is that different "voting" patterns may undermine this (apparent) robustness. Take, for example, a researcher using five information criteria to evaluate a set of candidate models. If four of the five criteria select one model, then using the most-selected model seems reasonable. On the other hand, if only two of the five criteria select one model and the other three each select different models, then the robustness of the two-vote model may be less convincing. Another disadvantage, as demonstrated in Table 2, is that ties are possible if an even number of information criteria are used.

Because model selection using information criteria is just one of many model selection methods, yet another potential approach is to use information criteria alongside other indices of model fit. In fact, this is the norm in the applied literature for some types of statistical techniques, such as structural equation models (e.g. West, Taylor, & Wu, 2012). Some SAS/STAT procedures, such as the CALIS procedure, will produce multiple measures of model fit automatically. Presumably, a model that is selected by multiple selection methods would, all else being equal, be more convincing than a model selected by just one method. However, the availability of alternative indices of model fit depends greatly on the type of models being fit and the available data.

Finally, while the selection of a single model among a set of models is often what motivates model selection, this is not an inherent requirement. It is possible that, after careful consideration, the researcher could decide to proceed with more than one model. In addition, if models in a set are closely related and robust parameter estimates are of interest, then a model averaging approach may be appropriate (Burnham & Anderson, 2002).

MAKING REASONED DECISIONS USING SAS

Regardless of the choice of approach, Base SAS offers many options for creating useful descriptive output to help with decision-making. Before using any of these options, the user must first obtain the information criteria values for each candidate model. For a detailed explanation of how to use the Output Delivery System to obtain information criteria from the output of SAS/STAT procedures, see Christensen (2018).

EXAMPLE 1: CREATING AND VISUALLY INSPECTING RANKS

While it is sometimes feasible for the user to manually inspect a list of information criteria to determine the best model, an automated method is often preferable. The RANK procedure can be used to quickly compute the within-criterion ranks for any number of candidate models. Take, for example, a case where three candidate models are evaluated using four information criteria, the values of which are read into a new DATA step:

```
data model_onebest;
input Name $ AIC AICC BIC CAIC;
cards;
M1 300 400 500 600
M2 100 200 300 400
M3 200 300 400 500;
run;
```

The RANK procedure can compute ranks for multiple information criteria simultaneously. To do this, list the appropriate variable names (matching the information criteria of interest) in the VAR statement. The RANKS statement creates new variables that contain the numeric rank assigned to each model by the RANK procedure. By default, the RANK procedure assigns a rank of 1 to the lowest column value, a 2 to

the second-lowest column value, etc. For further convenience, the SORT procedure can be used to place the lowest-ranked models at the top of the printed output:

```
proc rank data=model_onebest out=model_onebest_rank;
var AIC AICC BIC CAIC;
ranks AIC_r AICC_r BIC_r CAIC_r;
run;

proc sort data=model_onebest_rank;
by AIC AICC BIC CAIC;
run;
```

Sorted ranks, all criteria select M2

Obs	Name	AIC	AICC	BIC	CAIC	AIC_r	AICC_r	BIC_r	CAIC_r
1	M2	100	200	300	400	1	1	1	1
2	M3	200	300	400	500	2	2	2	2
3	M1	300	400	500	600	3	3	3	3

Figure 1. A set of models with no disagreement among criteria, with values and ranks

The results of this are displayed in Figure 1, which shows the model names, the information criteria values, and the ranks associated with each information criterion. These have been sorted by rank, so Model 2 (M2) is at the top. In this example, all of the information criteria select M2, so there is no disagreement.

EXAMPLE 2: DETERMINING IF DISAGREEMENT HAS OCCURED

Another DATA step can be used to create a new variable to determine if all of the information criteria select the same model or if there is disagreement among the criteria. Take, for example, a data set similar to the previous one, except that there is disagreement. Specifically, Model 1 (M1) will be selected by AIC, AICC, and BIC, but Model 3 (M3) will be selected by CAIC. First, the RANK procedure is used to create the numeric ranks within each information criterion:

```
data model_twobest;
input name $ AIC AICC BIC CAIC;
cards;
M1 100 200 300 600
M2 200 300 400 500
M3 300 400 500 400
;
run;

proc rank data=model_twobest out=model_twobest_rank;
var AIC AICC BIC CAIC;
ranks AIC_ranks AICC_ranks BIC_ranks CAIC_ranks;
run;
```

Next, the information criteria values and the numeric ranks are inserted into a new DATA step using the SET statement. Within the same DATA step, three new variables are created. The first variable (total) sums across the numeric ranks of the information criteria within each model. The second variable (IC_num) is a user-inputted constant, equaling the number of information criteria being used for model selection. These two variables are then used as part of an IF statement to create a third variable

(all_select) that indicates if a model was unanimously selected or not. If a model is unanimously selected, then all_select will equal 1 for a single model and 0 for all other models; otherwise, it will equal 0 for all models.

```
data twobest_select;
set model_onebest_rank;
total=AIC_ranks+AICC_ranks+BIC_ranks+CAIC_ranks;
IC_num=4;
if total=IC_num then all_select=1; else all_select=0;
run;
```

All four criteria select M2

Obs	Name	all_select
1	M2	1
2	M3	0
3	M1	0

Disagreement exists

Obs	Name	all_select
1	M1	0
2	M2	0
3	M3	0

Figure 2. The all_select variable when there is no disagreement (left) and when there is (right)

Figure 2 shows how the all_select variable changes between the first example (with no disagreement) and the current example (with disagreement). In the first case, shown in the left portion of Figure 2, the all_select variable is equal to 1 for M2 and 0 for the other models. Even without looking at the original information criteria values, the user can see not only that there was an unanimously-selected model, but also that the selected model was M2. In the second case, shown in the right portion of Figure 2, the all_select variable is always equal to 0. From this, the user knows that there is disagreement among the four information criteria and that further consideration will be necessary.

If minimizing the examination of individual models is of interest, the FREQ procedure can be used to quickly determine if there is disagreement:

```
proc freq data=onebest_select;
table all_select;
run;

proc freq data=twobest_select;
table all_select;
run;
```

The results of the FREQ output will match those shown in Figure 2. If there is no disagreement among information criteria, then the frequency table will show a single value of 1. The number of zeros will be equal to the number of candidate models minus one. If there is disagreement, then the frequency table will contain only zeros, equal to the total number of candidate models.

EXAMPLE 3: GENERATING OUTPUT TO HELP MAKE REASONED DECISIONS

Several potential approaches to a disagreement in model selection using information criteria were discussed in the previous section. Three of those solutions, which are presented in this section in order of the amount of code needed beyond what was used in Examples 1 and 2, can be readily implemented in Base SAS. The first one, using a single information criterion for model selection, requires no additional coding if the criterion was among those computed initially. The second one, the “popularity contest” method, requires the creation of few extra variables within a DATA step to properly count the “votes”. The

third one, computing the simple differences between information criteria and model weights, uses the SQL procedure to streamline what would otherwise be a tedious cycle of obtaining output from the MEANS procedure and manipulating it through multiple DATA steps.

Example 3.1: Examining the ranks for a single information criterion

Because the RANKS statement of the RANK procedure retains the ranks, the PRINT procedure can be used to selectively display just one set of ranks. For example, if the user decides to select a model based only on the ranks produced by AIC, both the original AIC values and the associated ranks can be selectively printed using the VAR statement:

```
proc print data=twobest_select;
var AIC AIC_ranks;
run;
```

Per the AIC-specific ranks shown in Figure 2 as part of Example 2, M1 would be selected as the “best” model. The disagreement is resolved because the conflicting ranks produced by CAIC (as well as the agreement of AICC and BIC) would be ignored.

Example 3.2: Model selection by “popularity contest”

In this method, the ranks produced by the RANK procedure for each information criterion are converted into votes. To do this, the within-criterion ranks are used to create new variables – Vote_AIC, Vote_AICC, Vote_BIC, Vote_CAIC – where a rank of 1 (i.e. the “best” model) remains equal to one and all other ranks are recoded as zero. These new variables are then summed within-model, creating a new variable called Votes.

```
data pop;
set twobest_select;
if AIC_ranks = 1 then Vote_AIC=1; else Vote_AIC=0;
if AICC_ranks = 1 then Vote_AICC=1; else Vote_AICC=0;
if BIC_ranks = 1 then Vote_BIC=1; else Vote_BIC=0;
if CAIC_ranks = 1 then Vote_CAIC=1; else Vote_CAIC=0;
Votes = Vote_AIC + Vote_AICC + Vote_BIC + Vote_CAIC;
run;
```

Voting results

Obs	Name	Votes
1	M1	3
2	M2	0
3	M3	1

Figure 3. A count of the model selected as “best” by four information criteria

Figure 3 shows the result of this method. Recall from Example 2 that AIC, AICC, and BIC selected M1, while CAIC selected M3. This is reflected in the Votes variable. M1 has three votes, M2 has zero votes, and M3 has one vote. If this approach was chosen to resolve the disagreement among information criteria, M1 would be selected.

Example 3.3: Computing information criteria simple differences and model weights

Burnham and Anderson (2002, p. 70-79) provided insightful commentary about how to interpret what they referred to as Akaike differences and Akaike weights. To better map onto their commentary, the code presented here is focused on AIC and closely mimics their example values (p. 72). It is important to note that this code can be used for other information criteria; the choice to focus on AIC was made to converse space in this manuscript.

Say, for example, that a researcher obtained AIC values for seven candidate models fit to the same data set:

```
data ex;
input Model AIC;
cards;
M1  100.0
M2  101.2
M3  101.9
M4  103.5
M5  104.1
M6  105.8
M7  107.3
;
run;
```

The first step for both the differences and the model weights is to obtain the AIC value of the model with the lowest value (the “best” model). In this case, the model with the lowest AIC value is M1. The SQL procedure efficiently isolates M1’s AIC value (100.0) with a SELECT statement and stores it as a numeric macro variable (&AIC_min).

```
proc sql;
select min(AIC) into :AIC_min
from ex;
quit;

%put &AIC_min;
```

Because the minimum AIC value is stored as a macro variable, it can be freely reference in later DATA steps. To compute the simple differences, a new variable (AIC_diff) is created by subtracting the &min_AIC value from the AIC value of each model.

```
data ex1;
set ex;
AIC_diff = AIC - &AIC_min;
run;
```

If just the simple differences are of interest, then no more code is needed. To compute the weights, as shown in the equation in the second section, the relative likelihoods for each model must first be computed. These relative likelihoods are what is in the numerator of the equation presented earlier. This can be done within the same DATA step that creates the AIC_diff variable by adding a new variable (num):

```
data ex1;
set ex;
AIC_diff = AIC - &AIC_min;
num = exp(-0.5*AIC_diff);
run;
```

Next, the denominator of the equation must be computed. This is done by summing the relative likelihoods of each model. Once again, the SQL procedure offers an efficient way to obtain this and use it in future DATA steps by creating a new macro variable (&denom):

```
proc sql;
select sum(num) into :denom
from ex1;
quit;

%put &denom;
```

To create the model weights, each model's relative likelihood (num) is divided by the sum of the model relative likelihoods (&denom). The results of these steps are shown in Figure 4, which matches (within rounding) what is shown in Burnham and Anderson (p. 77).

```
data ex2;
set ex1;
weights=num/&denom;
run;
```

AIC differences and weights

Obs	Model	AIC	AIC_diff	num	weights
1	M1	100.0	0.0	1.00000	0.43121
2	M2	101.2	1.2	0.54881	0.23665
3	M3	101.9	1.9	0.38674	0.16677
4	M4	103.5	3.5	0.17377	0.07493
5	M5	104.1	4.1	0.12873	0.05551
6	M6	105.8	5.8	0.05502	0.02373
7	M7	107.3	7.3	0.02599	0.01121

Figure 4. The AIC differences and weights, matching Burnham and Anderson (2002, p. 77)

CONCLUSION

Information criteria offer a flexible framework for model selection, but that flexibility is occasionally accompanied by disagreements among information criteria. Because different information criteria were developed to optimize different things, it is not surprising that candidate models can be ranked differently when fit to the same data. While there is no one-size-fits-all solution to resolving disagreements, there are multiple approaches that may be considered. Regardless of which solution is used, Base SAS functionality can help the user make a reasoned decision about how to proceed with model selection.

REFERENCES

Aho, K., Derryberry, D., & Peterron, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, *93*(3), 631-636.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.

Bauer, D. J., Gottfredson, N. C., Dean, D., & Zucker, R. A. (2013). Analyzing repeated measures data on individuals nested within groups: Accounting for dynamic group effects. *Psychological Methods*, *18*(1), 1-14.

Burnham, K. P., & Anderson, D. R. (2003). Model selection and multimodel inference: A practical information-theoretic approach (2nd ed.). New York, NY: Springer.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345-370.

Christensen, W. (2018). *Model selection using information criteria (made easy in SAS®)*. Paper presented at SAS Global Forum, Denver, CO.

Hamaker, E. L., van Hattum, P., Kuiper, R. M., & Hoijtink, H. (2011). Model selection based on information criteria in multilevel modeling. In Hox, J. J. & Roberts, J. K (Eds.), *Handbook of advanced multilevel analysis* (pp. 231-255). New York, NY: Routledge.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, 297-307.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods and Research*, *33*(1), 188-229.

Salom, C. L., Betts, K. S., Williams, G. M., Najman, J. M, & Alati, R. (2015) Predictors of comorbid polysubstance use and mental health disorders in young adults – a latent class analysis. *Addiction*, *111*(1), 156-164.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464.

West, W. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In Hoyle, R. H. (Eds), *Handbook of structural equation modeling* (pp. 209-231). Guilford Press; New York, NY.

RECOMMENDED READING

- *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*
- *Model Selection for Ecologists: The Worldviews of AIC and BIC*
- *Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Wendy Christensen
wchristensen@ucla.edu
https://github.com/wendychristensen

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.