Model Selection and the Cult of AIC

Mark J Brewer Mark.Brewer@bioss.ac.uk, @bulboussquidge

Adam Butler Adam.Butler@bioss.ac.uk

Biomathematics and Statistics Scotland
http://www.bioss.ac.uk

UAB, Barcelona, 16th June 2017

Being a statistician

Stuart Hurlbert (San Diego State University):

> *Statistics, as practiced and taught, is a strange discipline. Most courses in statistics are not taught by professional statisticians, the most widely used texts in statistics are not written by them, and most of the literature of statistical criticism (like that in this editorial*) is not written by them.*

* More later

A strange discipline

Frequently, ecologists tell me I know nothing about statistics:

A strange discipline

Frequently, ecologists tell me I know nothing about statistics:

- Using SAS to fit mixed models (and not R)

A strange discipline

Frequently, ecologists tell me I know nothing about statistics:

- Using SAS to fit mixed models (and not R)
- Not making a 5-level factor a random effect

A strange discipline

Frequently, ecologists tell me I know nothing about statistics:

- Using SAS to fit mixed models (and not R)
- Not making a 5-level factor a random effect
- Estimating variance components as zero

A strange discipline

Frequently, ecologists tell me I know nothing about statistics:

- Using SAS to fit mixed models (and not R)
- Not making a 5-level factor a random effect
- Estimating variance components as zero
- Not using GAMs for binary explanatory variables, or mixed models with no factors

A strange discipline

Frequently, ecologists tell me I know nothing about statistics:

- Using SAS to fit mixed models (and not R)
- Not making a 5-level factor a random effect
- Estimating variance components as zero
- Not using GAMs for binary explanatory variables, or mixed models with no factors
- Not using AIC for model selection

Two key questions

"**Always** use AIC for model selection or model comparison":

Two key questions

"**Always** use AIC for model selection or model comparison":

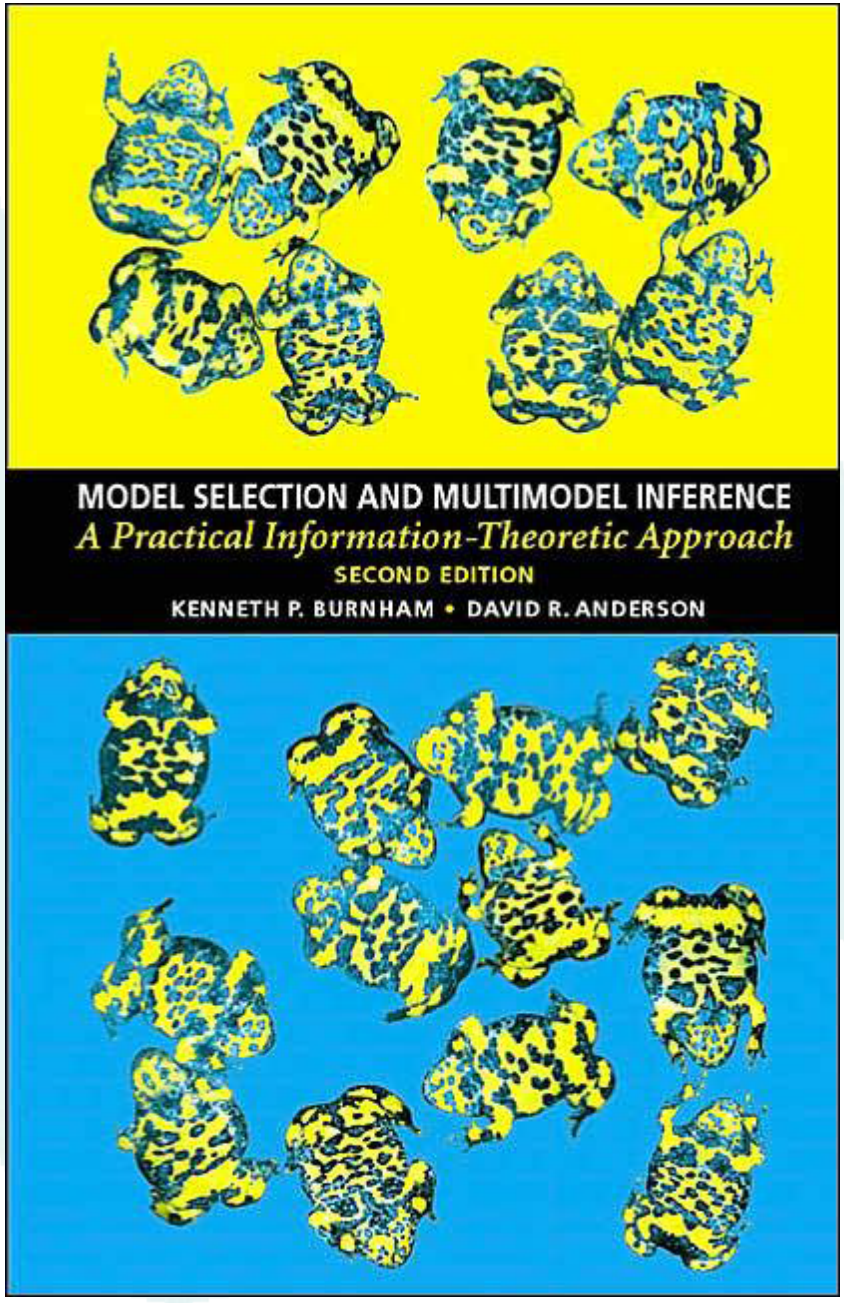
- Is this instruction correct?

Two key questions

"**Always** use AIC for model selection or model comparison":

- Is this instruction correct?
- **Why** is it wrong?

The Cult of AIC



Burnham & Anderson (2002)

Outline

- Model selection—what, why and how?
- Contrast AIC with hypothesis testing / *p*-values
  - Forum articles in *Ecology*, March 2014
- AIC, BIC — or something else?
  - Theoretical properties — useful?
  - Simulation study
- Practical model selection
- Miscellanea

Model selection — What?

- One response variable (y)
- Multiple explanatory variables (x's)
- Will fit some kind of regression model
- Response equal to some function of the x's
- **Linear regression**: easy to count number of parameters (GLMs)

Model selection — Why?

- Understand which explanatory variables are important
- Find evidence of relationship(s) between variable(s) and response
- Quantify the effects of explanatory variables on the response
- Use model to predict response values
- Predict response at another location or time

Inference, Estimation, Prediction

- Understand which explanatory variables are important
- Find evidence of relationship(s) between variable(s) and response
- Quantify the effects of explanatory variables on the response
- Use model to predict response values
- Predict response at another location or time

Model selection — Why?

- Q: Why not just include all explanatory variables?

Model selection — Why?

- Q: Why not just include all explanatory variables?
- A: They will (almost certainly) be correlated.

Model selection — Why?

- Q: Why not just include all explanatory variables?
- A: They will (almost certainly) be correlated.

Correlation between explanatory variables. . .

Model selection — Why?

- Q: Why not just include all explanatory variables?
- A: They will (almost certainly) be correlated.

Correlation between explanatory variables. . .

- makes it hard to identify important effects;

Model selection — Why?

- Q: Why not just include all explanatory variables?
- A: They will (almost certainly) be correlated.

Correlation between explanatory variables. . .

- makes it hard to identify important effects;
- causes bias in effect sizes/parameter estimates;

Model selection — Why?

- Q: Why not just include all explanatory variables?
- A: They will (almost certainly) be correlated.

Correlation between explanatory variables. . .

- makes it hard to identify important effects;
- causes bias in effect sizes/parameter estimates;
- leads to overfitting, and poor out-of-sample prediction.

Overfitting

FINAL FINAL

POLICYFORUM

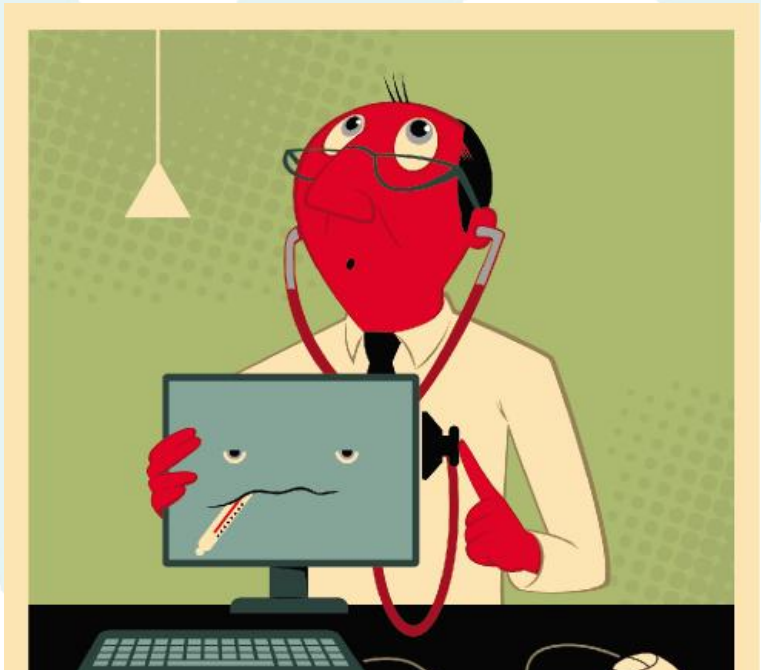
BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2]* Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?



the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These

Significance (June 2017)

SIGNIFICANCE

Explore this journal >

In Detail

Bursting the big data bubble

Wai Mun Fong

First published: 9 June 2017 Full publication history

DOI: 10.1111/j.1740-9713.2017.01035.x View/save citation

Cited by (CrossRef): 0 articles Check for updates Citation tools

Am score 10

View issue TOC
Volume 14, Issue 3
June 2017
Pages 20–23

Abstract

In the financial world, big data is hailed as a potential game changer for predicting stock market performance. But without adequate safeguards, big data analyses may result in spurious correlations, misguided predictions and disappointing returns. By **Wai Mun Fong**

Model selection — How?

How can we compare models?

Model selection — How?

How can we compare models?

- Using statistical hypothesis tests; *p*-values
  - *only two models at a time*
  - *likelihood ratio tests (LRTs)*
  - *t-tests (coefficients); F-tests (ANOVA)*

Model selection — How?

How can we compare models?

- Using statistical hypothesis tests; *p*-values
  - *only two models at a time*
  - *likelihood ratio tests (LRTs)*
  - *t-tests (coefficients); F-tests (ANOVA)*
- Calculate measure of (relative) model fit
  - *as many models as you like*
  - *R^2, R^2_{adj}, Mallows' C_p*
  - *information criteria: AIC, BIC, DIC, WAIC*

Model selection — How?

How can we compare models?

- Using statistical hypothesis tests; *p*-values
  - *only two models at a time*
  - *likelihood ratio tests (LRTs)*
  - *t-tests (coefficients); F-tests (ANOVA)*
- Calculate measure of (relative) model fit
  - *as many models as you like*
  - *R^2, R^2_{adj}, Mallows' C_p*
  - *information criteria: AIC, BIC, DIC, WAIC*
- Simultaneous selection and estimation (LASSO, NNET, CART)

Fit?

- Compare “fit” of different models
- Define “fit” — (maximum) **likelihood**?
- Increases when adding more variables
- Need to **penalise** fit somehow (to avoid overfitting)
- Choice between different approaches reflects choice of penalty
  - sometimes just **strength** of penalty
- Three ways of comparing likelihoods. . .

Definitions: AIC, BIC, LRT

- AIC (*Akaike's Information Criterion*):

$$\text{AIC} = -2 \log \ell \left(\hat{\theta} \right) + 2p$$

- BIC (*Bayesian Information Criterion*):

$$\text{BIC} = -2 \log \ell \left(\hat{\theta} \right) + p \log n$$

- LRT (*Likelihood Ratio Test*); vs $\chi^2_{p_2 - p_1}$:

$$-2 \log \ell \left(\hat{\theta}_1 \right) + 2 \log \ell \left(\hat{\theta}_2 \right)$$

AIC_C for small samples

- AIC:

$$AIC = -2 \log \ell \left(\hat{\theta} \right) + 2p$$

- AIC_C:

$$AIC_C = AIC + \frac{2p(p+1)}{n-p-1}$$

ΔAIC vs. LRT

- ΔAIC:

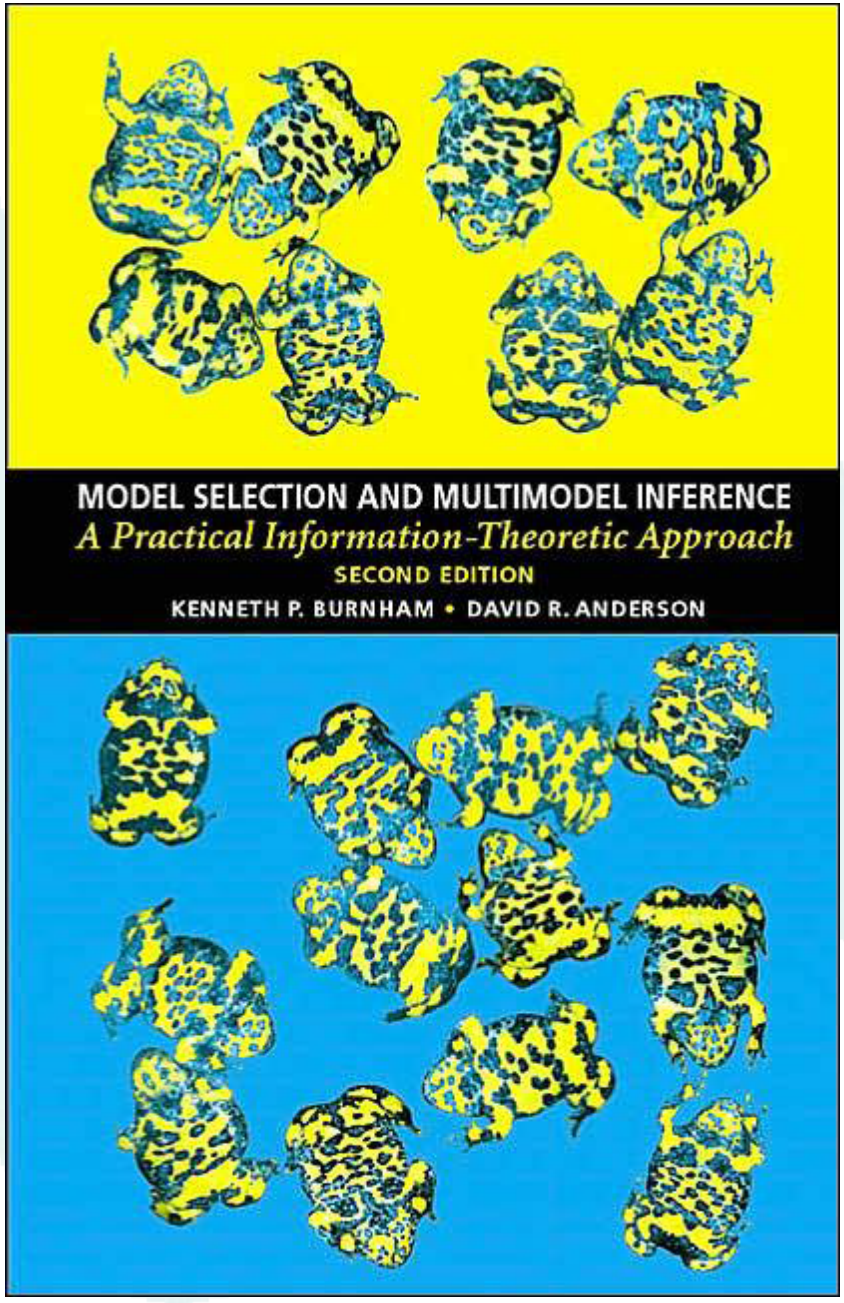
$$\Delta\text{AIC} = -2\log\ell\left(\hat{\theta}_1\right) + 2\log\ell\left(\hat{\theta}_2\right) - 2(p_2 - p_1)$$

- LRT; compare following with $\chi^2_{p_2-p_1}$:

$$-2\log\ell\left(\hat{\theta}_1\right) + 2\log\ell\left(\hat{\theta}_2\right)$$

- If $p_1 = p_2$, no *need* for LRT—comparing likelihoods equivalent to comparing AIC

The Cult of AIC



B & A
(2002)

The Cult of AIC

The good points:

- Emphasis on model uncertainty in addition to parameter uncertainty
- Corresponding move away from thinking we have found a single “best” or “right” model
- Exhortation to present effect sizes
- Negative view of post hoc power analyses
- Valid criticisms of *p*-values and hypothesis testing, BIC, data dredging

B & A recipe for modelling

Procedure for model selection:

- Decide on scientific questions to answer
- Define **small** candidate set of models
- Fit models, calculate AIC, find “best” model
- Use **threshold** on ΔAIC to determine set of “plausible” models
- Calculate model weights via ΔAIC
- (Assess variable importance using weights)

The Cult of AIC

Model selection then becomes a simple function minimization, where AIC . . . is the criterion to be minimized. AIC selection is ***objective*** *and represents a very different paradigm to that of null hypothesis testing and is free from the* ***arbitrary*** α ***levels****, the* ***multiple-testing*** *problem, and the fact that some candidate* ***models might not be nested****.*

"objective" — $>$ 50 times;
"arbitrary" — 41 times

AIC in practice

Minimum AIC selects “best” model, but what about others?

Support for for suboptimal model by ΔAIC:

- 0 – 2: substantial
- 4 – 7: considerably less
- $>$10: essentially none

Burnham and Anderson (*Springer*, 2002) §2.6
BUT: “is free from the arbitrary α levels” (p.89)

AIC in practice

Minimum AIC selects “best” model, but what about others?

Support for for suboptimal model by ΔAIC:

- 0 – 7: plausible
- 7 - 14: equivocal
- $>$14: implausible

Burnham, Anderson and Huyvaert (2011)
Behavioral Ecology and Sociobiology

Theory: AIC, BIC, LRT

Properties of different likelihood comparisons:

- AIC finds “best” model for making predictions given new data (MSE sense)
- BIC finds “true” model if it exists (consistent) or “quasi-true” model otherwise
- LRT compares two nested models *asymmetrically*, assesses “improvement” of one model over another

Which to use should depend on purpose of modelling. . .

Purpose of modelling

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

Galit Shmueli

Abstract. Statistical modeling is a powerful tool for developing and testing theories by way of causal explanation, prediction, and description. In many disciplines there is near-exclusive use of statistical modeling for causal explanation and the assumption that models with high explanatory power are inherently of high predictive power. Conflation between explanation and prediction is common, yet the distinction must be understood for progressing scientific knowledge. While this distinction has been recognized in the philosophy of science, the statistical literature lacks a thorough discussion of the many differences that arise in the process of modeling for an explanatory versus a predictive goal. The purpose of this article is to clarify the distinction between explanatory and predictive modeling, to discuss its sources, and to reveal the practical implications of the distinction to each step in the modeling process.

Key words and phrases: Explanatory modeling, causality, predictive modeling, predictive power, statistical strategy, data mining, scientific research.

1. INTRODUCTION

Looking at how statistical models are used in different scientific disciplines for the purpose of theory building and testing, one finds a range of perceptions regarding the relationship between causal explanation

focus on the use of statistical modeling for causal explanation and for prediction. My main premise is that the two are often conflated, yet the causal versus predictive distinction has a large impact on each step of the statistical modeling process and on its consequences.

Purpose of modelling

Theory suggests:

AIC optimal for **prediction**

BIC useful for **explanation**

LRT for assessing whether a given effect or set of effects is “important”

How useful is this distinction in practice?

AIC and *p*-values

Ecology, 95(3), 2014, pp. 611–617

In defense of *P* values

PAUL A. MURTAUGH[1]

Department of Statistics, Oregon State University, Corvallis, Oregon 97331 USA

Abstract. Statistical hypothesis testing has been widely criticized by ecologists in recent years. I review some of the more persistent criticisms of *P* values and argue that most stem from misunderstandings or incorrect interpretations, rather than from intrinsic shortcomings of the *P* value. I show that *P* values are intimately linked to confidence intervals and to differences in Akaike's information criterion (ΔAIC), two metrics that have been advocated as replacements for the *P* value. The choice of a threshold value of ΔAIC that breaks ties among competing models is as arbitrary as the choice of the probability of a Type I error in hypothesis testing, and several other criticisms of the *P* value apply equally to ΔAIC. Since *P* values, confidence intervals, and ΔAIC are based on the same statistical information, all have their places in modern statistical practice. The choice of which to use should be stylistic, dictated by details of the application rather than by dogmatic, a priori considerations.

Key words: *AIC; confidence interval; null hypothesis;* P *value*; *significance testing*.

In the 1970s, a number of authors argued for the systematic use of null and alternative hypotheses when

stem from misuse of these procedures by practitioners. Hurlbert and Lombardi (2009) also consider criticisms

AIC and *p*-values

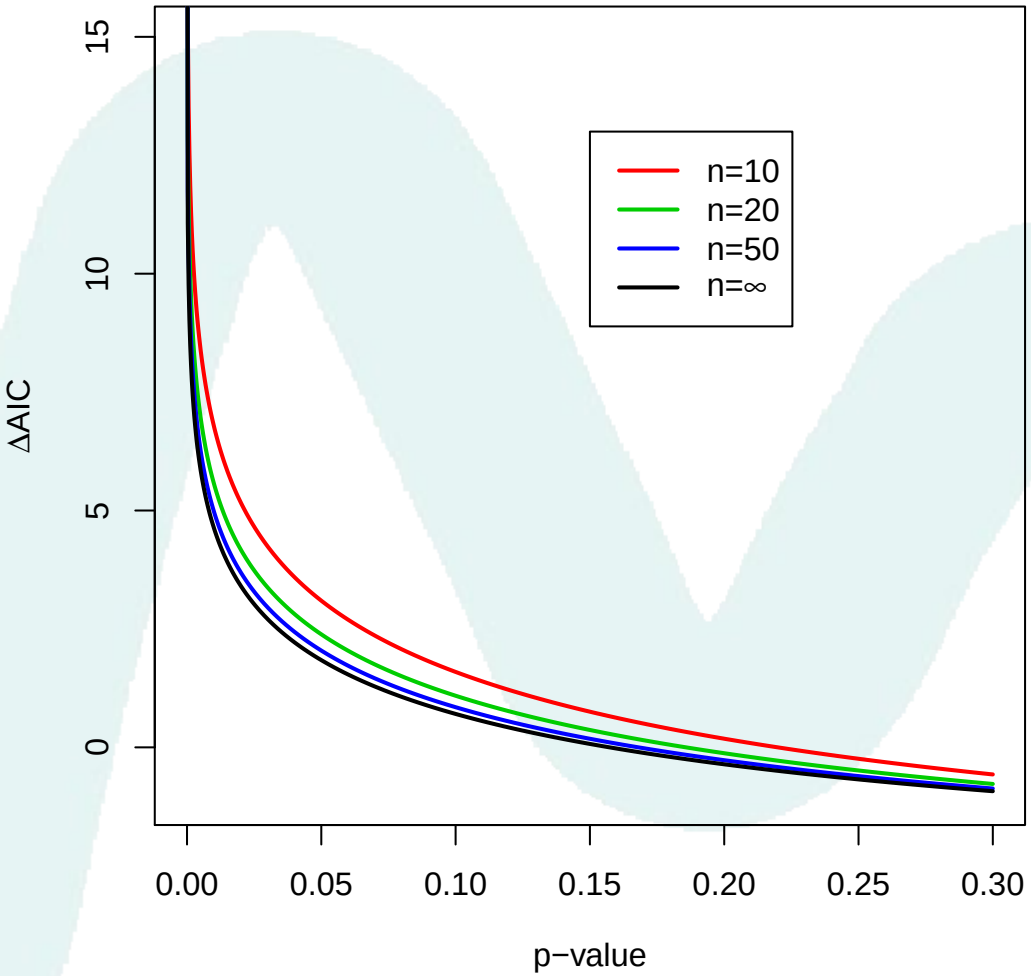
- ΔAIC and (LRT) *p*-values linked
- Criticisms of *p*-values more on **misuse**
- Report *p*-values and effect size
- AIC, BIC and *p*-values methods *all* have their uses

AIC and *p*-values

From Murtaugh *Ecology* (2014), Fig. 2:

Direct comparison

Two models, differ by one parameter



AIC and *p*-values

- Two models, differ by one parameter
- Larger model must have larger log likelihood by 1 to have lower AIC
- Corresponds to *p*-value of 0.157
  - why AIC likes “bigger models”
- For $\Delta\text{AIC} = 2$, $p \approx 0.046$
- ΔAIC and *p* are **different points on the same curve(s)**

AIC in practice. . .

Var 1	Var 2	Var 3	Var 4	Var 5	AIC	ΔAIC	Weight
	-0.539 (± 0.244)		**-0.602** (± 0.190)		1789.73	0.00	0.26
	-0.674 (± 0.336)		**-0.609** (± 0.192)	0.173 (± 0.295)	1791.38	1.65	0.12
	-0.544 (± 0.245)	0.003 (± 0.008)	**-0.566** (± 0.214)		1791.60	1.86	0.10
-0.090 (± 0.333)	**-0.541** (± 0.244)		**-0.574** (± 0.217)		1791.66	1.93	0.10
			-0.641 (± 0.201)		1792.23	2.50	0.08
-0.070 (± 0.335)	**-0.670** (± 0.336)		**-0.586** (± 0.220)	0.167 (± 0.296)	1793.34	3.61	0.04
			-0.622 (± 0.198)	-0.212 (± 0.222)	1793.34	3.61	0.04
	-0.662 (± 0.344)	0.001 (± 0.008)	**-0.591** (± -0.591)	0.155 (± 0.316)	1793.35	3.62	0.04

AIC in practice. . .

Models Within Two Units of the Best Model

Models having Δ_i within about 0–2 units of the best model should be examined to see whether they differ from the best model by 1 parameter *and* have essentially the same values of the maximized log-likelihood as the best model.

In this case, the larger model is not really supported or competitive, but rather is "close" only because it adds 1 parameter and therefore will be within 2 Δ_i units, even though the fit, as measured by the log-likelihood value, is not improved.

Previous slide

That sounds like a rationale for doing a LRT

AIC vs. *p*-values

Consider (from Burnham and Anderson (*Springer*, 2002) §6.9.3) a set of models with equal AIC, differing by *j* parameters from a null model.

LRT statistics and *p*-values shown in the table.

"The solution to this dilemma is a matter of which model selection approach has a sound theoretical basis: AIC does, LRT does not."

j	χ^2	P
1	2	0.157
2	4	0.135
3	6	0.112
4	8	0.092
5	10	0.075
6	12	0.062
7	14	0.051
8	16	0.042
9	18	0.035
10	20	0.029
15	30	0.012
20	40	0.005
25	50	0.005
30	60	0.001

AIC vs. *p*-values

Is this valid?

No!!

Conclusions differ because there are two fundamentally different questions being answered:

AIC: is it worth having **all** the extra parameters?

LRT: is it worth having **any** of the extra parameters?

j	χ^2	P
1	2	0.157
2	4	0.135
3	6	0.112
4	8	0.092
5	10	0.075
6	12	0.062
7	14	0.051
8	16	0.042
9	18	0.035
10	20	0.029
15	30	0.012
20	40	0.005
25	50	0.005
30	60	0.001

"Uninformed *pronunciamentos*"

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Editorial

David Trafimow and Michael Marks
New Mexico State University

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

With the banning of the NHSTP from BASP, what are the implications for authors? The following are anticipated questions and their corresponding answers.

Question 1. *Will manuscripts with p-values be desk rejected automatically?*

Answer to Question 1. No. If manuscripts pass the preliminary inspection, they will be sent out for review. But prior to publication, authors will have to remove all vestiges of the NHSTP (*p*-values, *t*-values, *F*-values, a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that when in a state of ignorance, the researcher should assign an equal probability to each possibility. The problems are well documented (Chihara, 1994; Fisher, 1973; Glymour, 1980; Popper, 1983; Suppes, 1994; Trafimow, 2003, 2005, 2006). However, there have been Bayesian proposals that at least somewhat circumvent the Laplacian assumption, and there might even be cases where there are strong grounds for assuming that the numbers really are there (see Fisher, 1973, for an

AIC, BIC — or something else?

Ecology, 95(3), 2014, pp. 631–636

Model selection for ecologists: the worldviews of AIC and BIC

Ken Aho,[1,4] DeWayne Derryberry,[2] and Teri Peterson[3]

[1]*Department of Biological Sciences, Idaho State University, Pocatello, Idaho 83209 USA*
[2]*Department of Mathematics, Idaho State University, Pocatello, Idaho 83209 USA*
[3]*Division of Health Sciences, Idaho State University, Pocatello, Idaho 83209 USA*

Introduction

Ecologists frequently ask questions that are best addressed with a model comparison approach. Under this system, the merit of several models is considered without necessarily requiring that (1) models are nested, (2) one of the models is true, and (3) only current data be used. This is in marked contrast to the pragmatic blend of Neyman-Pearson and Fisherian significance testing conventionally emphasized in biometric texts (Christensen 2005), in which (1) just two hypotheses are under consideration, representing a pairwise comparison of models, (2) one of the models, H_0, is assumed to be true, fication or confirmation of scientific claims because it does not explicitly consider prior information. Scientists often do not consider a single data set to be adequate for research hypothesis rejection (Quinn and Keough 2002:35), particularly for complex hypotheses with a low degree of falsifiability (i.e., Popper 1959:266). Similarly, the support of hypotheses in the generation of scientific theories requires repeated corroboration (Ayala et al. 2008).

Third, ecologists and other scientists are frequently concerned with the plausibility of existing or default models, what statistician would consider null hypotheses

Argument for AIC above BIC

Aho *et al.* argue that in biology, any "true" model is too complex to model exactly ($p >> n$)

Very many, very small effects.

> *The correct model is not one of the candidate models, so consistency is irrelevant.*

While consistency *per se* seems not a useful concept here, BIC **might** still be useful

Shibata, *Biometrika*, 1981

. . . paper which shows "optimality" of AIC (and hence not BIC), with respect to:

> *[t]he expectation, with respect to future observations, of the sum of squared errors of prediction*

BUT: future observations are from the **same population as the current data**.

Shibata, *Biometrika*, 1981

. . . paper which shows "optimality" of AIC (and hence not BIC), with respect to:

> *[t]he expectation, with respect to future observations, of the sum of squared errors of prediction*

BUT: future observations are from the **same population as the current data**.

Ecologists routinely apply model for one time/place to a *different* time/place!

AIC vs. BIC - Simulation Study

Don't want to fall into a trap:

> *the pitfall of many simulation studies: (1) there is a simple (low dimension), true model, which (2) is included as one of the contending models and (3) the goal is to select that true model. Studies of this type favor (at least for large sample sizes) BIC over AIC because BIC is designed for this circumstance.*

Buckland, Burnham, and Augustin, *Biometrics* (1997)

AIC vs. BIC - Simulation Study

Burnham and Anderson (*Springer*, 2002), §6.4:

> *Simulation studies . . . have been done, but the results when comparing AIC and BIC performance depend on the nature of the data generating model* ***(such as having many tapering effects or not)****, on whether the model set contains the generating model, on the sample sizes considered, and on the objective: select the true model or select the K-L best approximating model.*

AIC vs. BIC - Comparisons

Murtaugh, *Ecology Letters* (2009):
"Performance of several variable-selection methods applied to real ecological data"

> *I argue that there is no 'best' method of variable selection and that any of the regression-based approaches discussed here is capable of yielding useful predictive models.*

Validation: AIC, BIC, *F*; stepwise(!)/all subsets

AIC vs. BIC - Comparisons

Raffalovich *et al*., *Journal of Applied Statistics* (2008): “Model selection procedures in social research: Monte-Carlo simulation results”

> *Our results recommend BIC as the criterion of choice, but stepwise regression, with strict entry and removal criteria, is a feasible second choice.*

True model in data set — naturally favours BIC

AIC vs. BIC - Comparisons

Burnham and Anderson, *Sociological Methods & Research* (2004): "Multimodel Inference: Understanding AIC and BIC in Model Selection"

> *. . . it must be realized that these two criteria for computing model weights have their optimal performance under different conditions:* ***AIC for tapering effects*** *and* ***BIC for when there are no effects at all or a few big effects and all others are zero effects (no intermediate effects, no tapering effects)****.*

A simulation framework

Given true, generating model:

$$y_i = \alpha + \boldsymbol{\beta}' \boldsymbol{x}_i + r_i$$

Define simulation framework:

- Genuine out-of-sample fit assessment
- **Heterogeneity of data sets**
- Assess predictive fit quality using RMSE
- Compare AIC with BIC *and* other penalties?

A simulation framework

BioSS

Paper and R package:

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7**, 679–692 doi: 10.1111/2041-210X.12541

SPECIAL FEATURE: 5TH ANNIVERSARY OF *METHODS IN ECOLOGY AND EVOLUTION*

The relative performance of AIC, AIC_C and BIC in the presence of unobserved heterogeneity

Mark J. Brewer[1,*], Adam Butler[2] and Susan L. Cooksley[3]

[1]*Biomathematics and Statistics Scotland, Craigiebuckler, Aberdeen, AB15 8QH,UK;* [2]*Biomathematics and Statistics Scotland, JCMB, The King's Buildings, Edinburgh, EH9 3JZ,UK; and* [3]*The James Hutton Institute, Craigiebuckler, Aberdeen, AB15 8QH, UK*

Summary

1. Model selection is difficult. Even in the apparently straightforward case of choosing between standard linear regression models, there does not yet appear to be consensus in the statistical ecology literature as to the right approach.

A simulation framework

Important — allow for data set heterogeneity:

- *Different sites, times, species, etc*
- Changes in covariate means and variances
- Changes in between covariate correlations
- Changes in model parameters, in a "random effect" sense

Generate data sets from **known model**, use **out-of-sample assessment of prediction**.

A simulation framework

Process:

1. Generate data set (and response y)
2. Fit all possible models (`dredge` in R)
3. Select best models by AIC_c, AIC, BIC etc
4. Generate **second set** of data
5. Predictions of y (models at 3; data at 4)

Repeat 1000 times; calculate mean RMSE.

A simulation framework

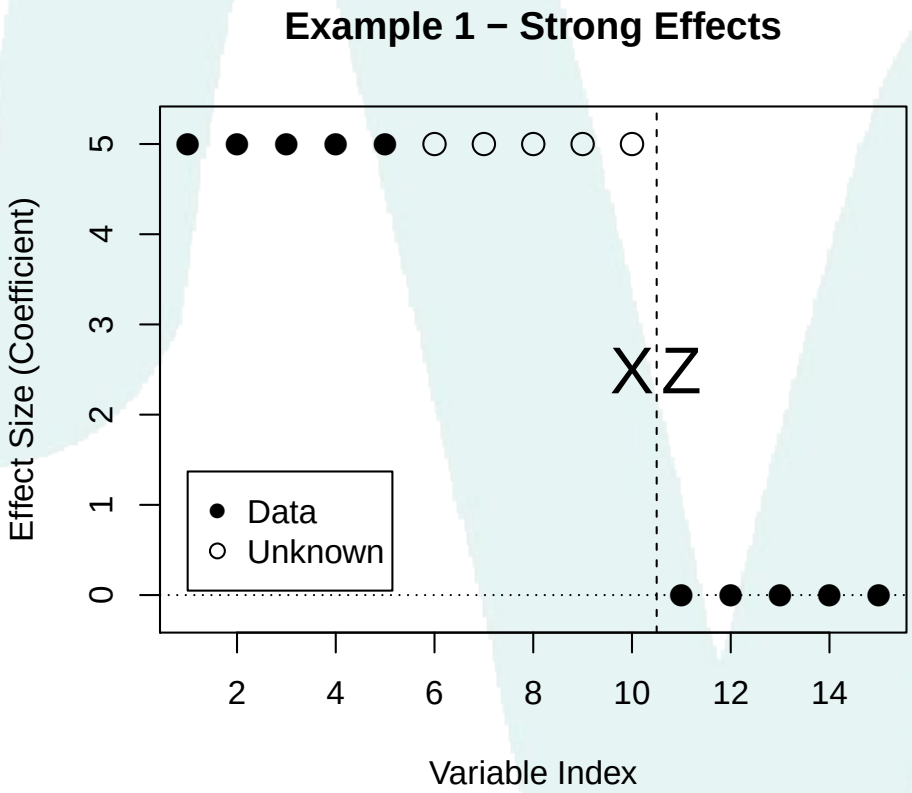
- Generate covariate data
  - ***x*** which enter the generating model (**real**)
  - ***z*** which don't (**spurious**)

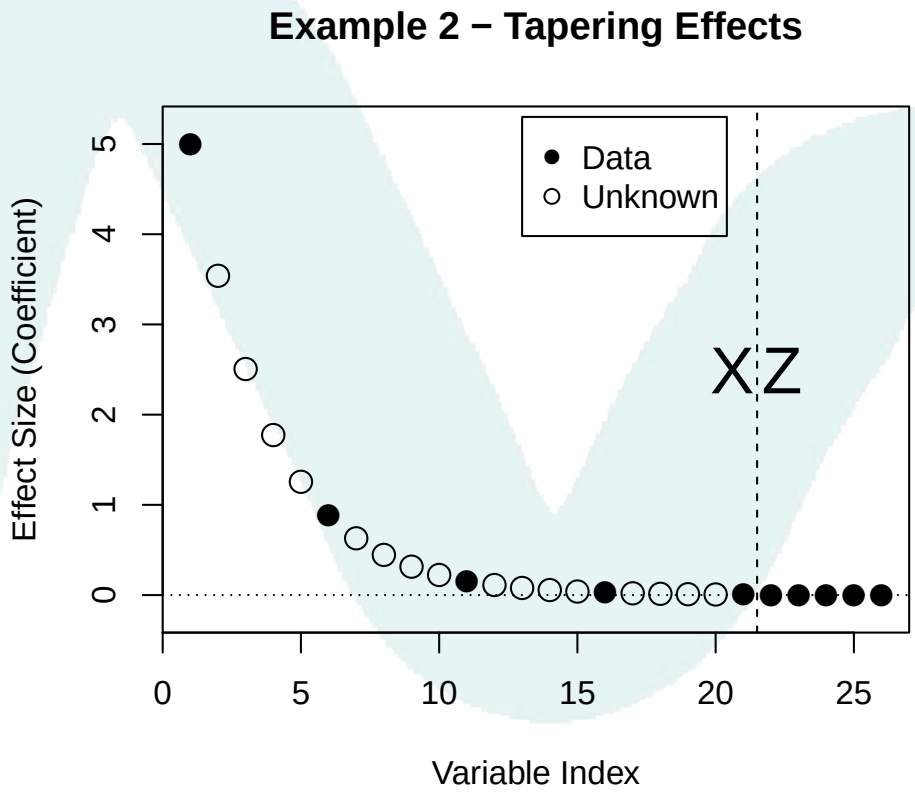
  from *multivariate Normal*
- For model fitting, **remove** some ***x***'s from covariate set (omitted)
- R package on GitHub:

  `MarkJBrewer/ICsims`

Simulation study

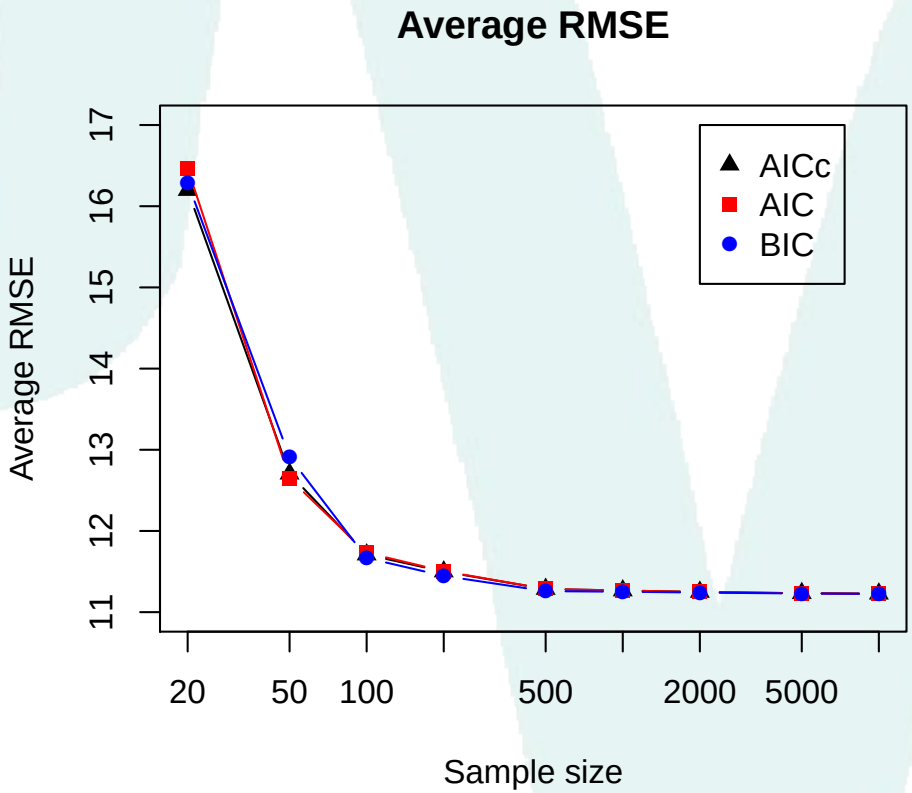
Two examples: Strong vs. tapering effects. . .

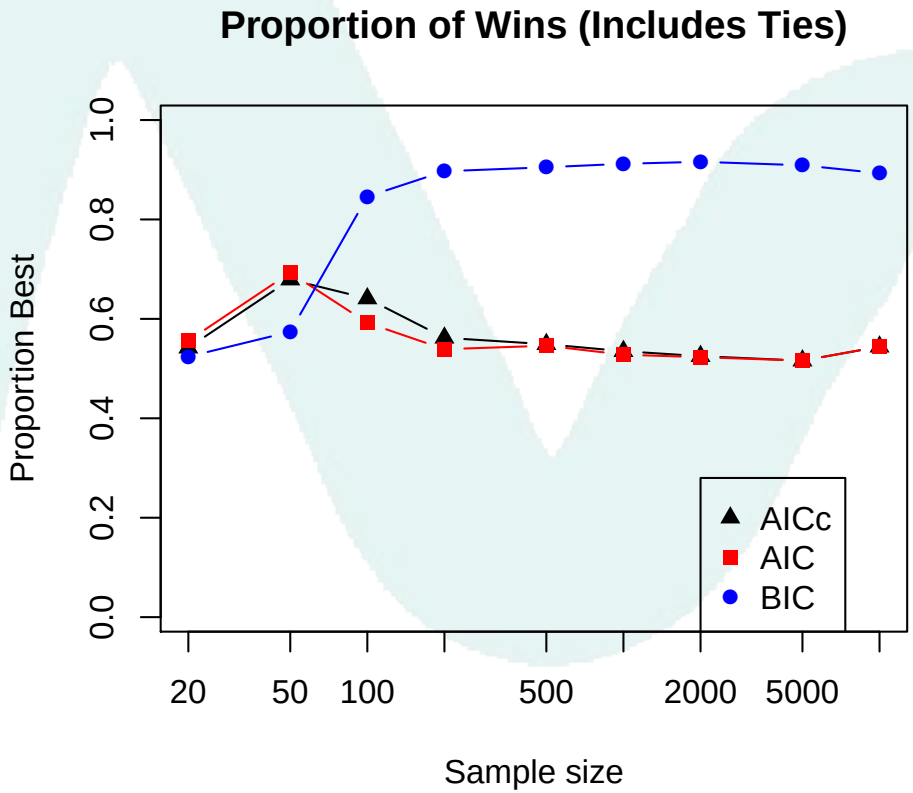




Simulation study

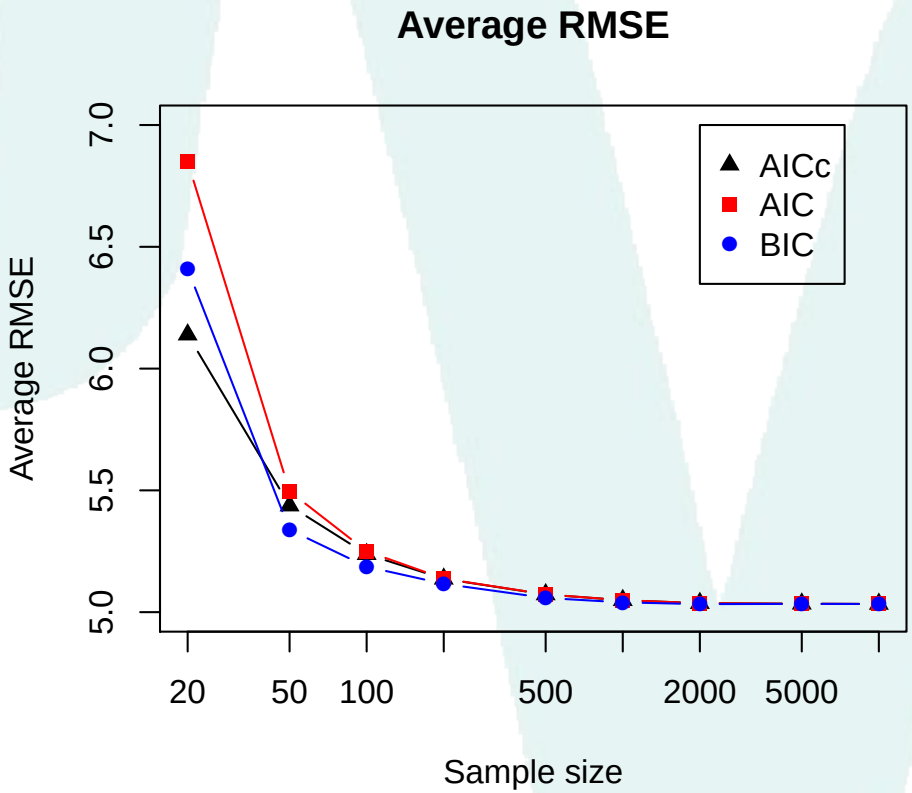
Example 1 (Strong effects):
sample new (uncorrelated) ***x*** and ***z*** only

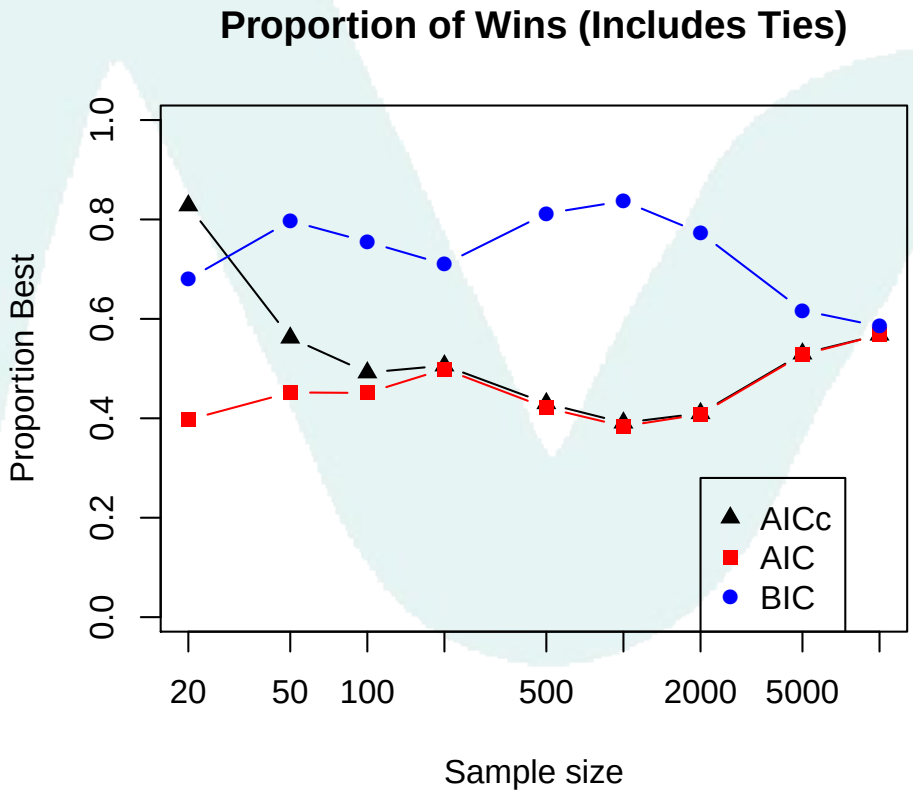




Simulation study

Example 2 (Tapering effects):
sample new (uncorrelated) ***x*** and ***z*** only





Simulation study

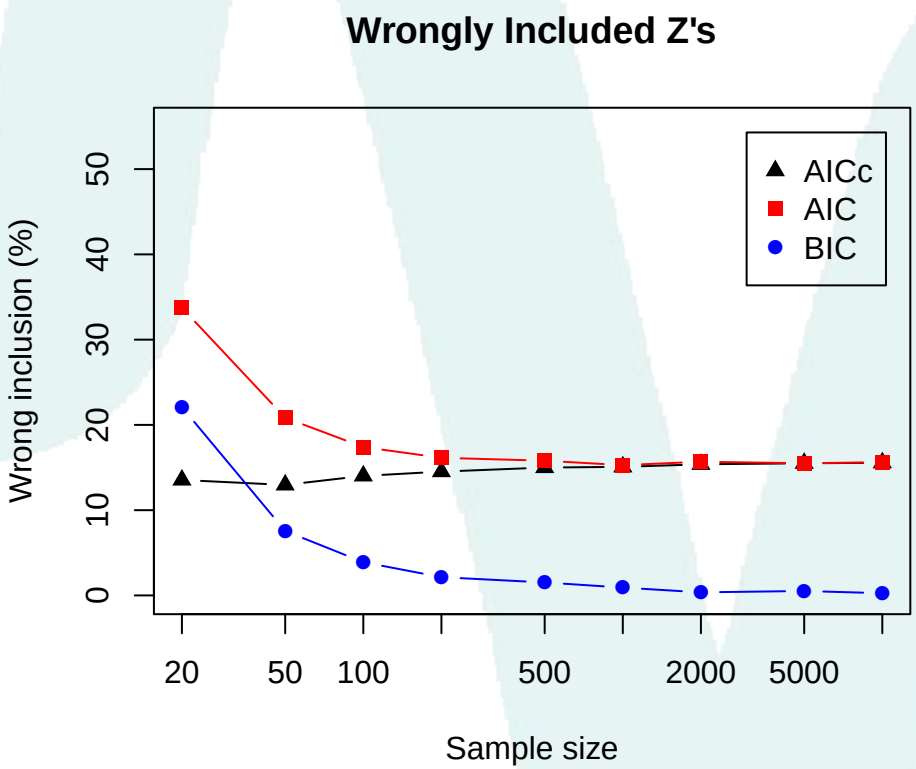
But we should "use AIC for prediction"!?

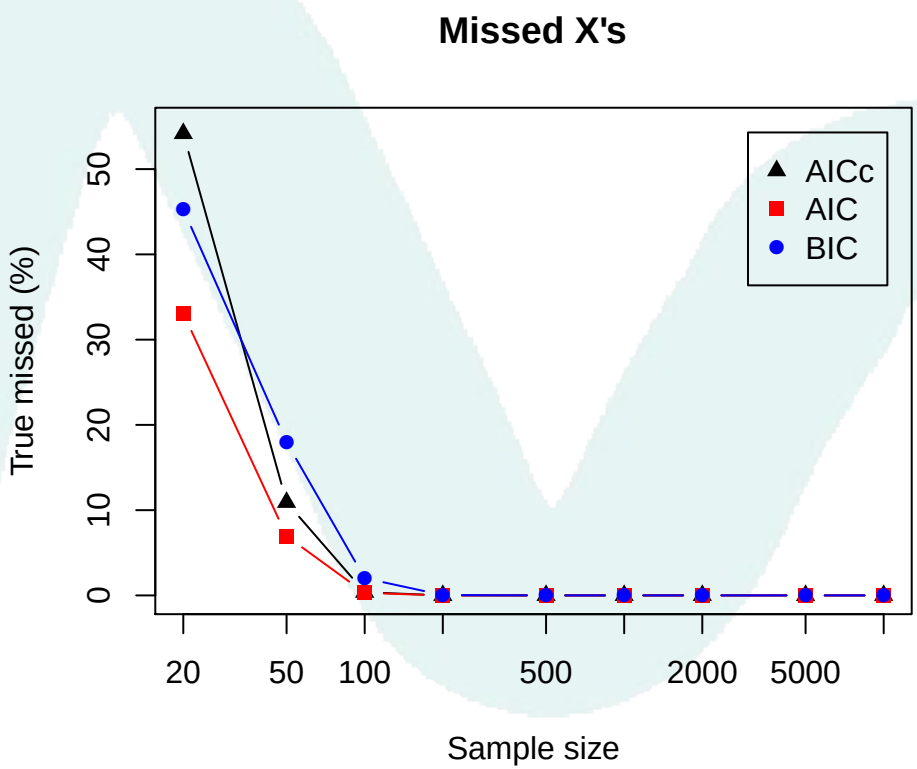
Why is BIC doing better than AIC here?

- BIC more often prefers smaller models, with (real) tiny effects set to zero
- Zero can be a better estimate than the actual one (can be wrong sign!)

Simulation Study

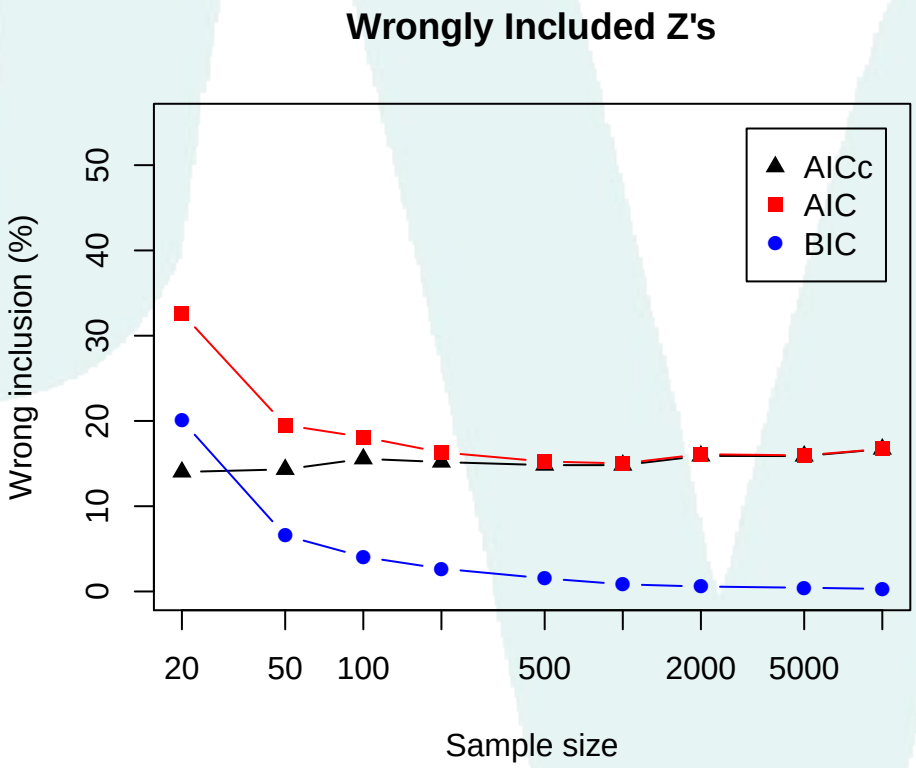
Example 1 — Strong effects (within sample)...

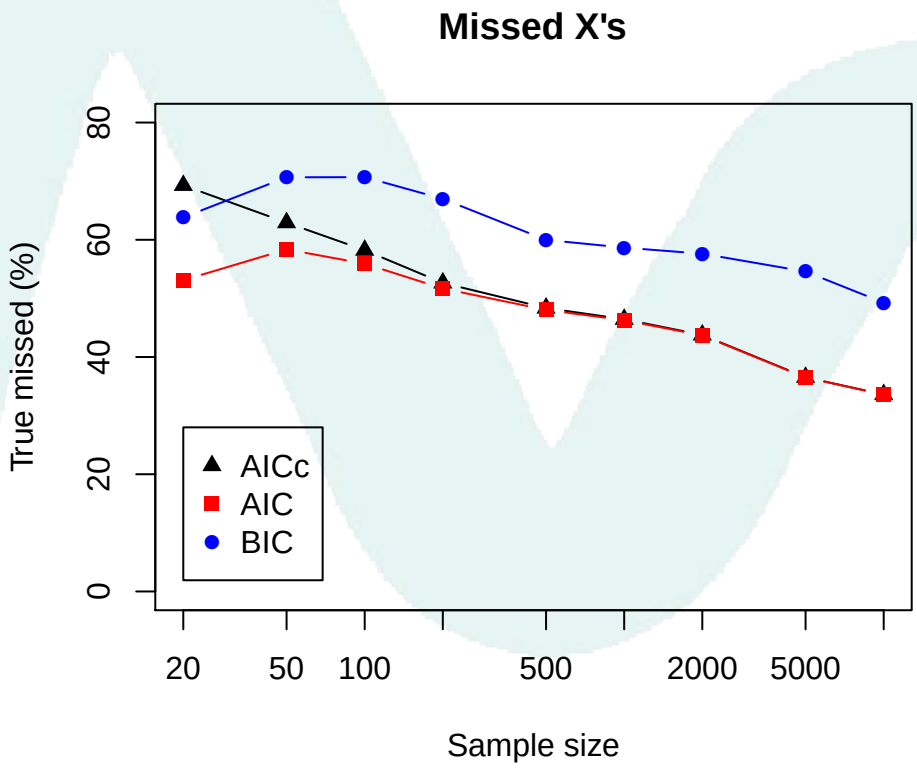




Simulation Study

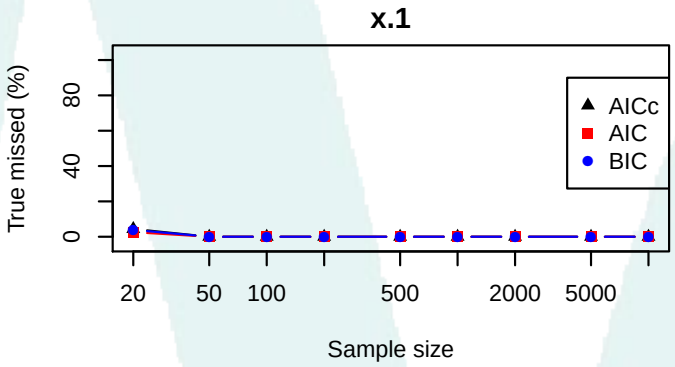
Example 2 — Tapering effects (within sample)...

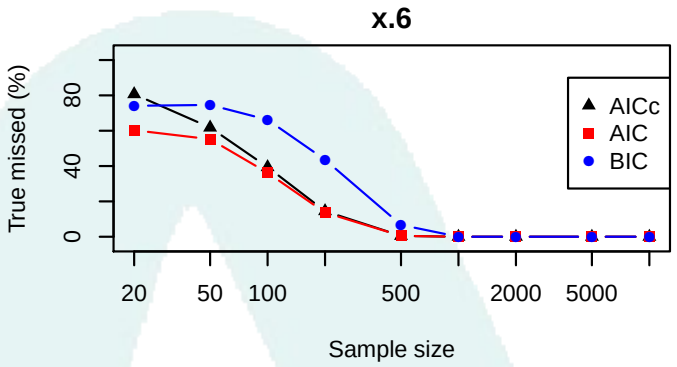


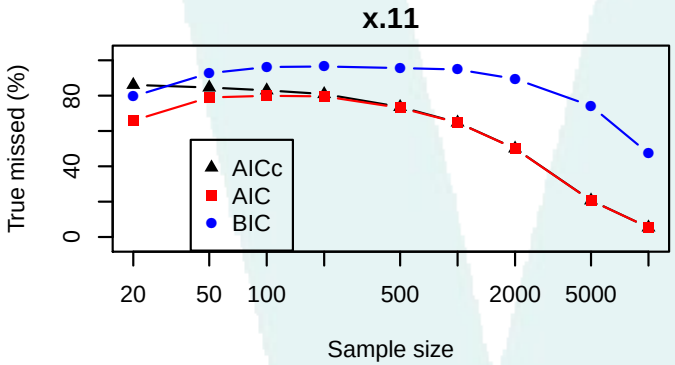


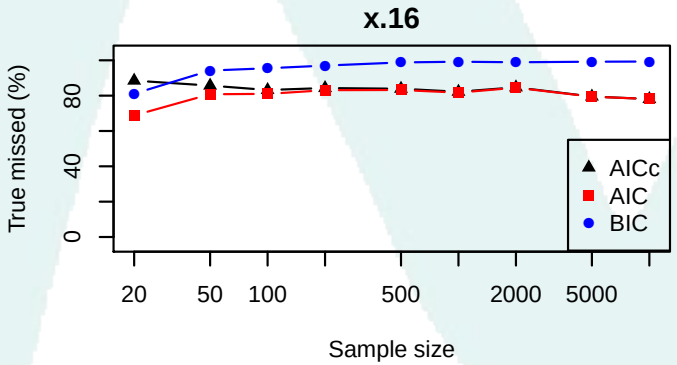
Simulation Study

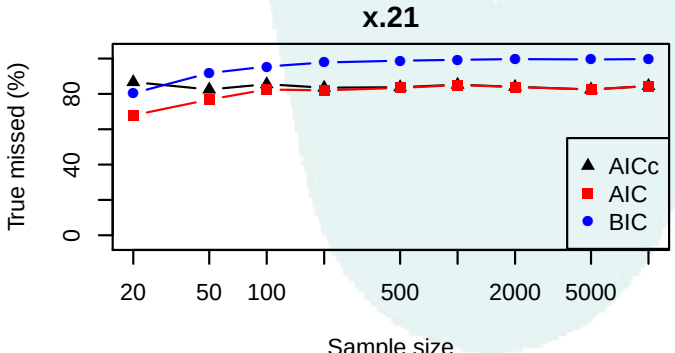
Example 2 — Tapering effects, individual variables





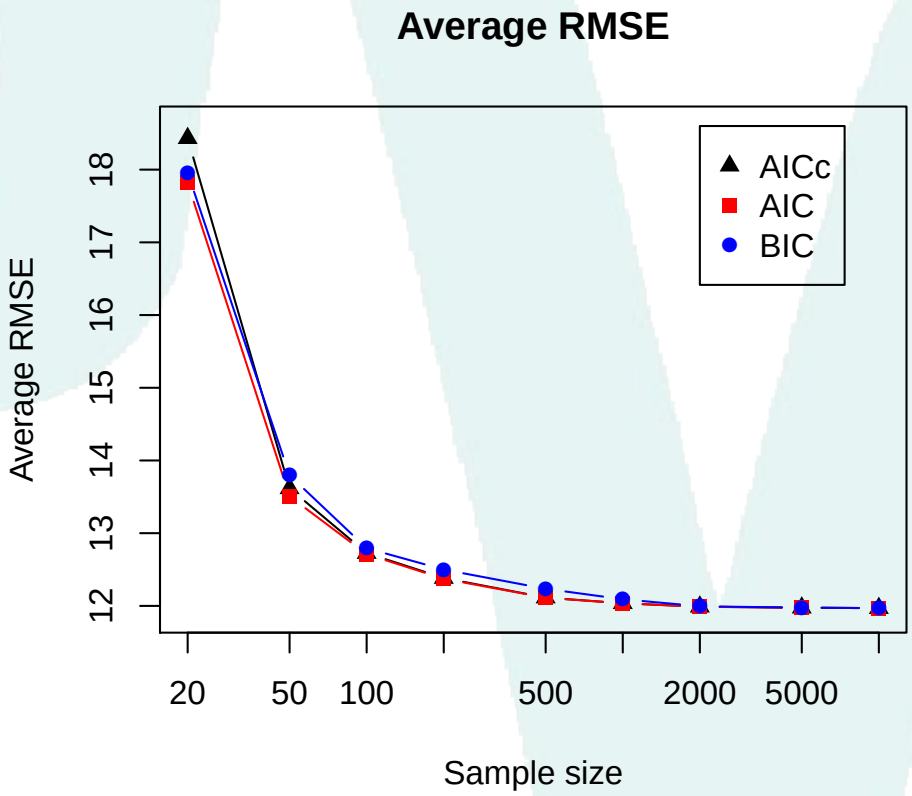


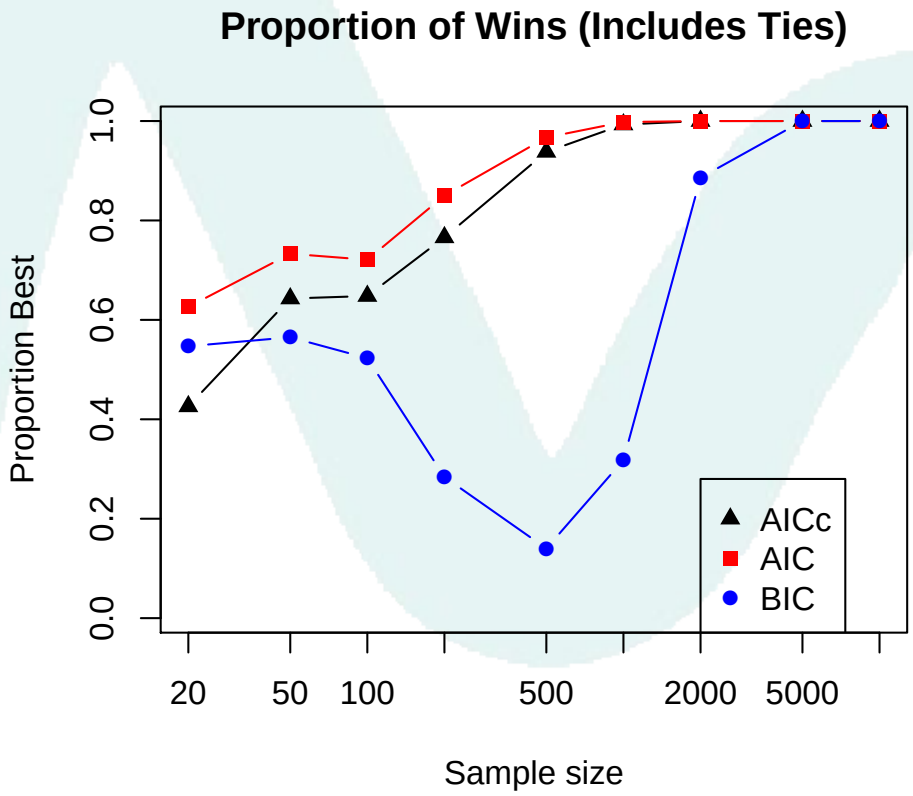




Simulation study

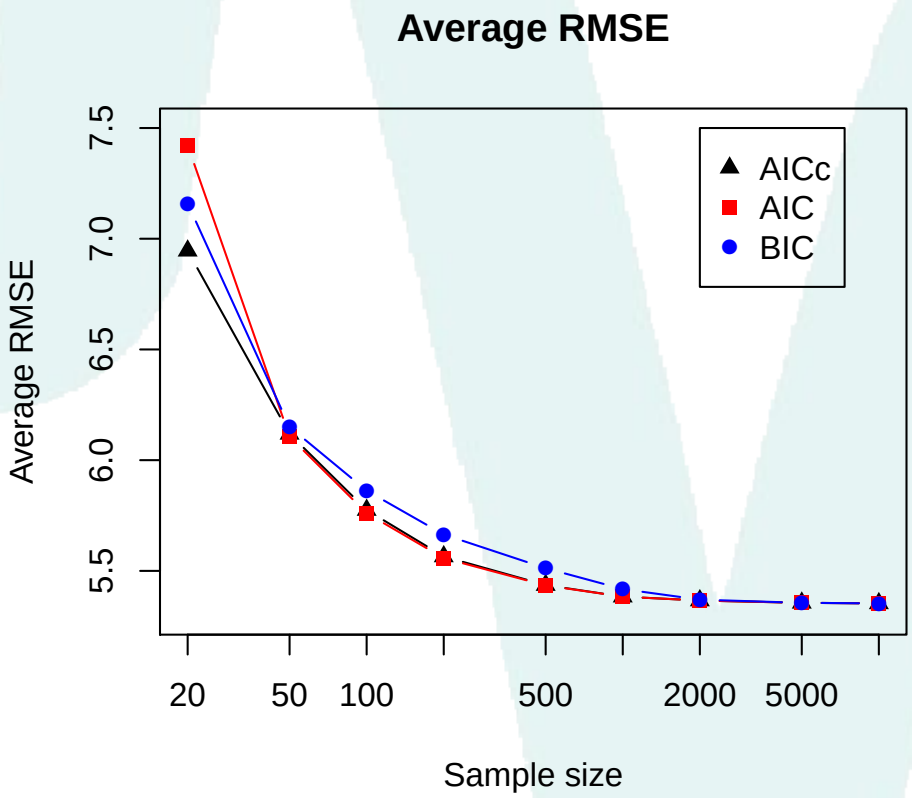
Example 1 (Strong effects):
sample new (correlation 0.1) ***x*** and ***z*** only

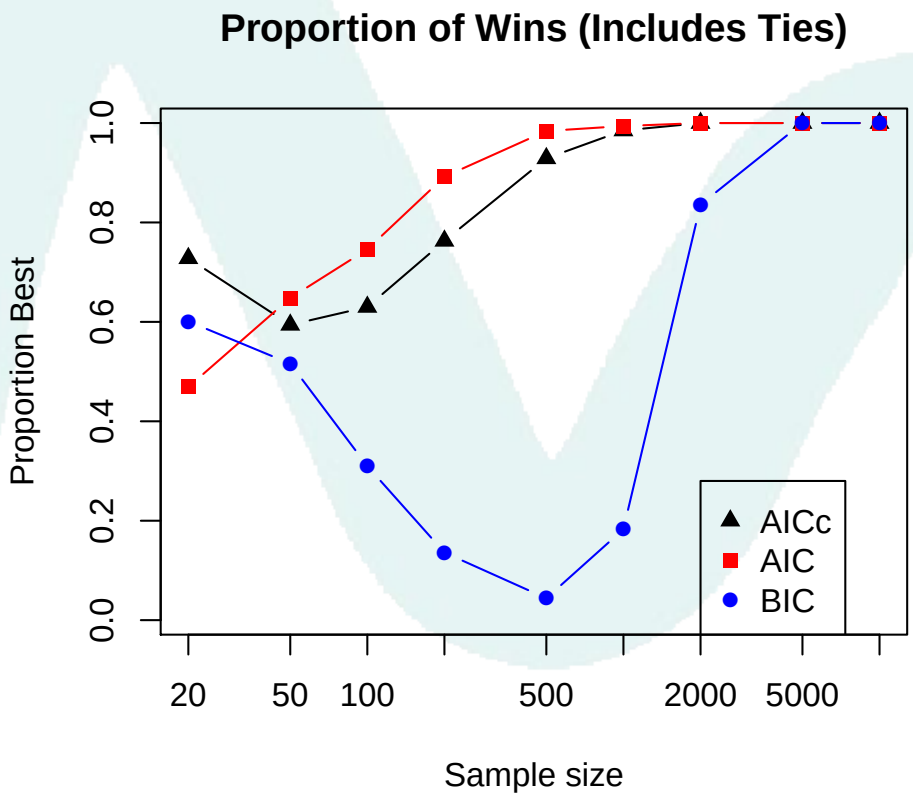




Simulation study

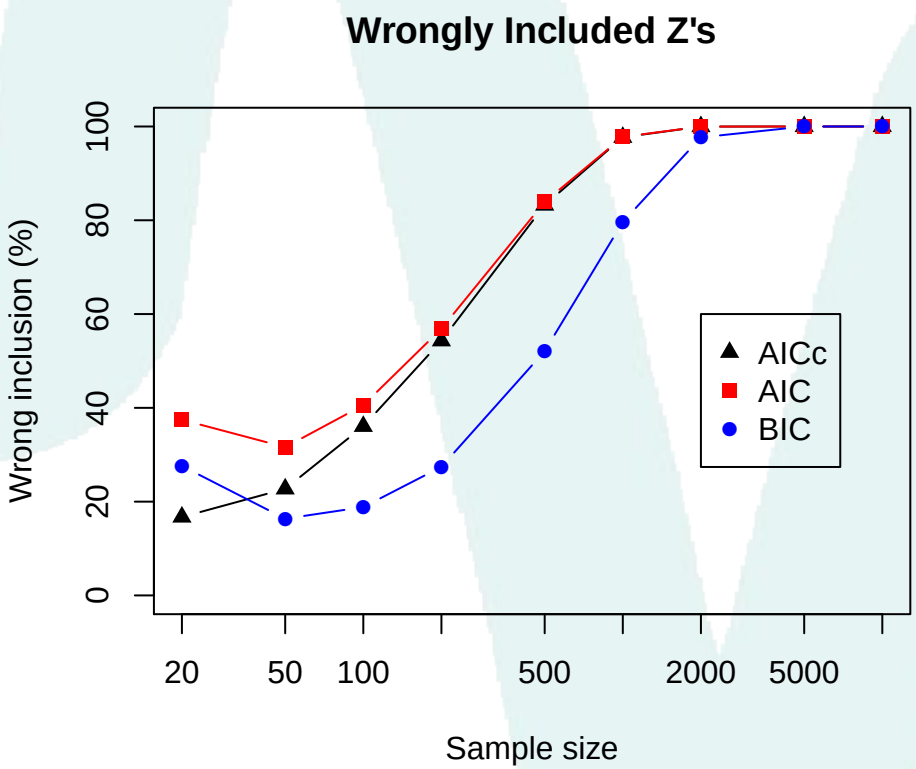
Example 2 (Tapering effects):
sample new (correlation 0.1) ***x*** and ***z*** only

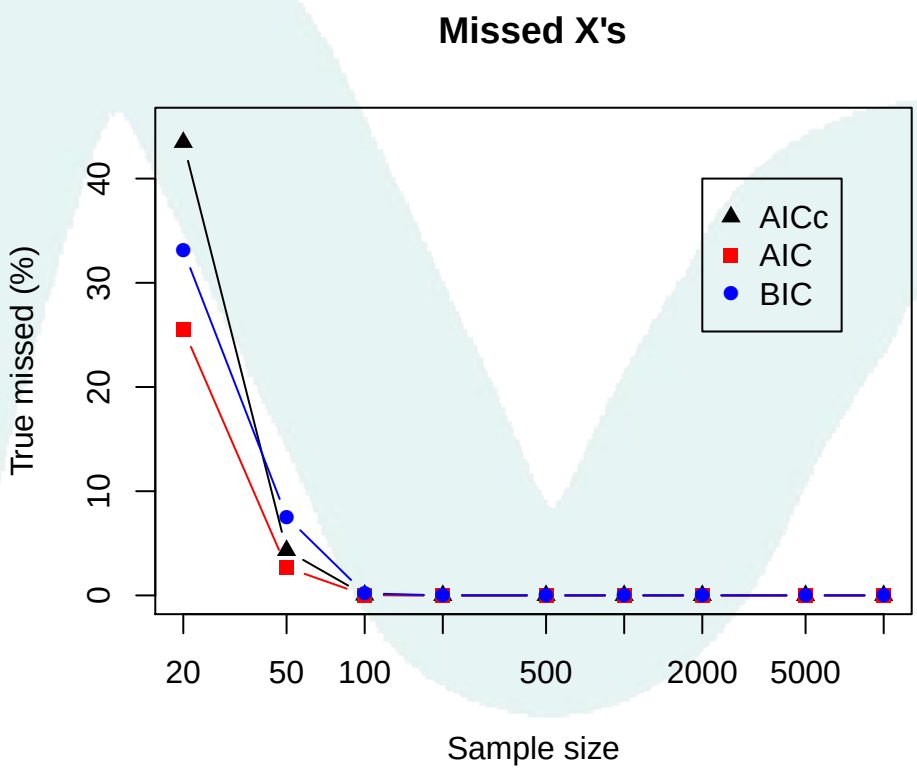




Simulation Study

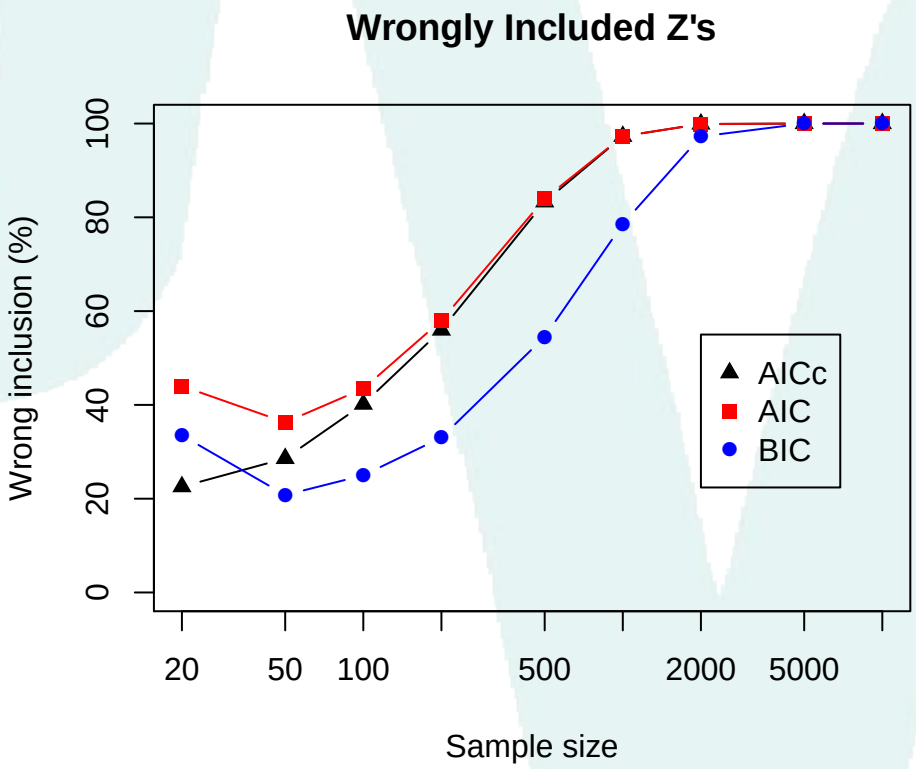
Example 1 — Strong effects (within sample). . .

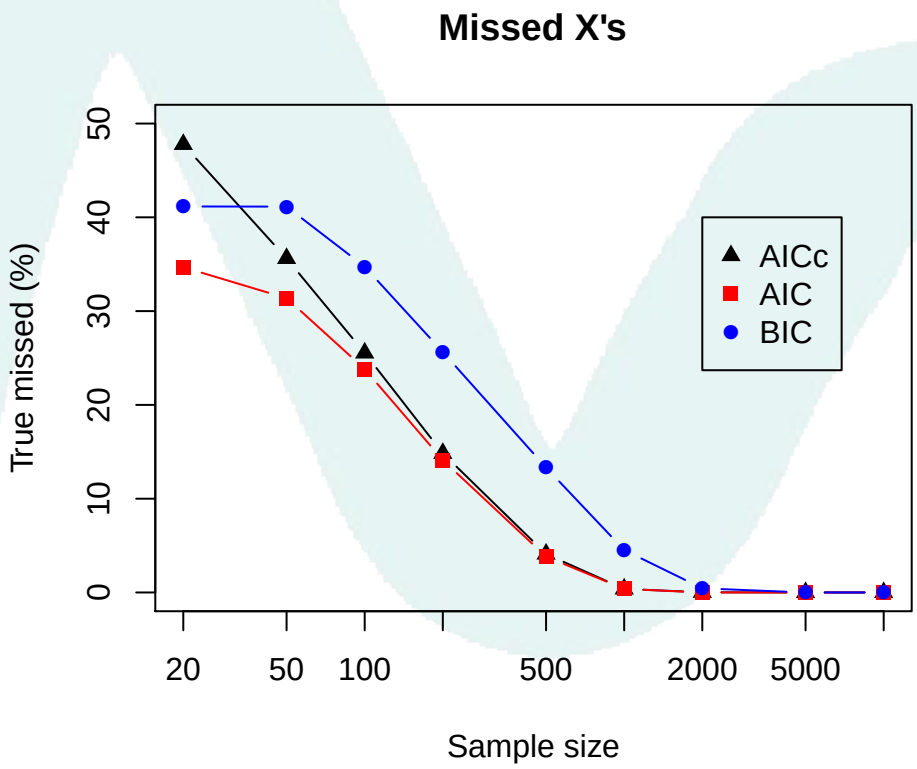




Simulation Study

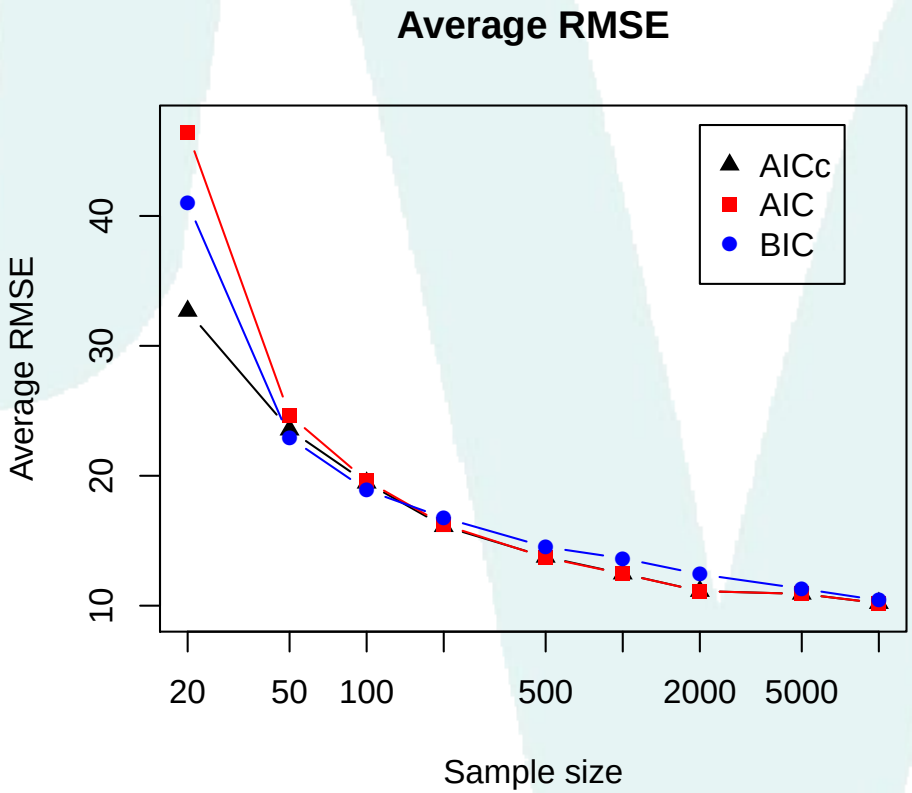
Example 2 — Tapering effects (within sample)…

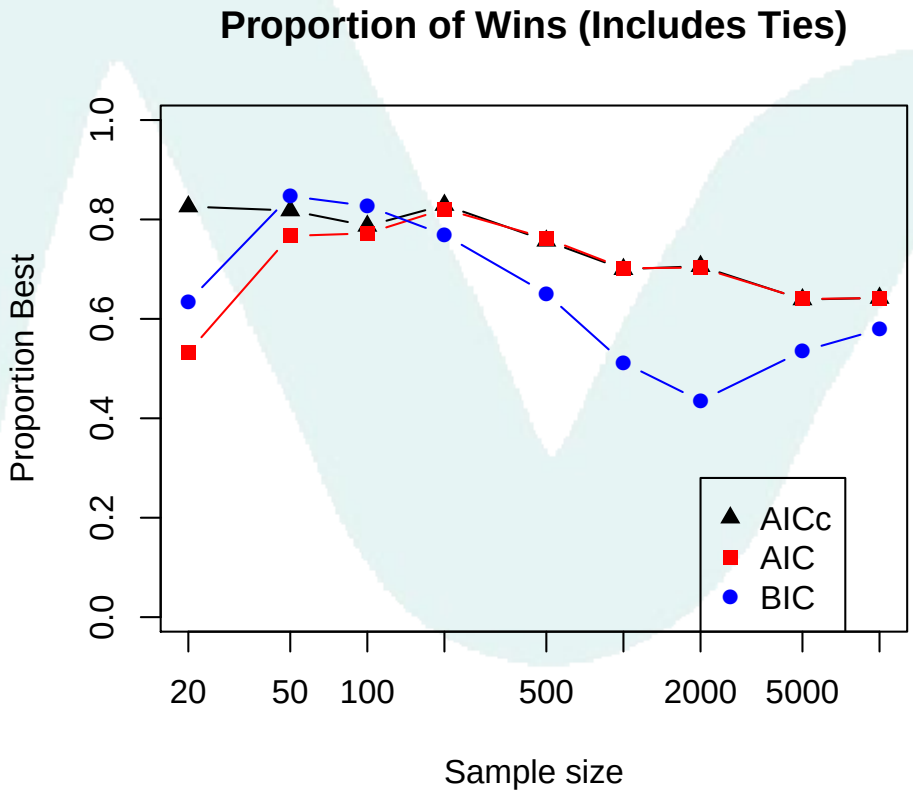




Simulation study

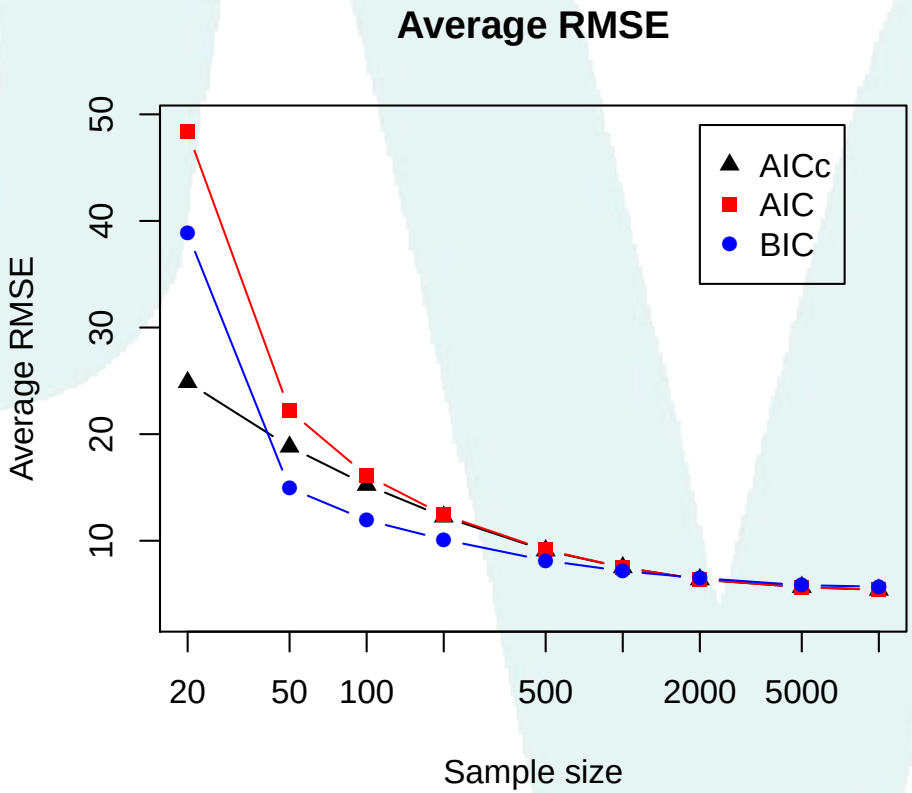
Example 1 (Strong effects):
high correlation variability (**heterogeneity**)

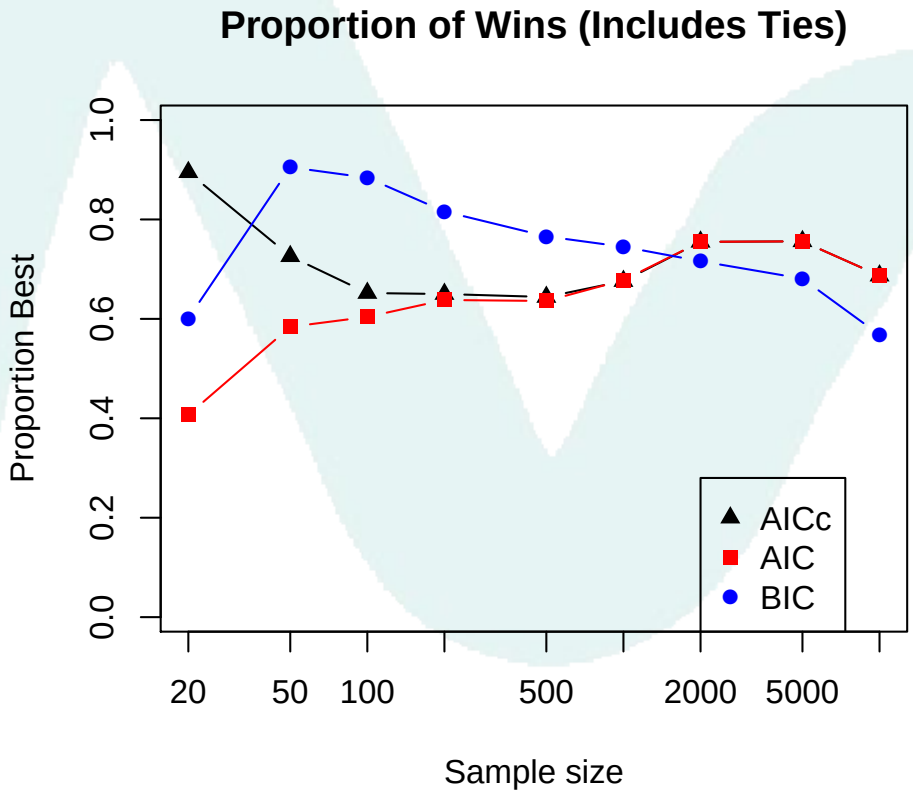




Simulation study

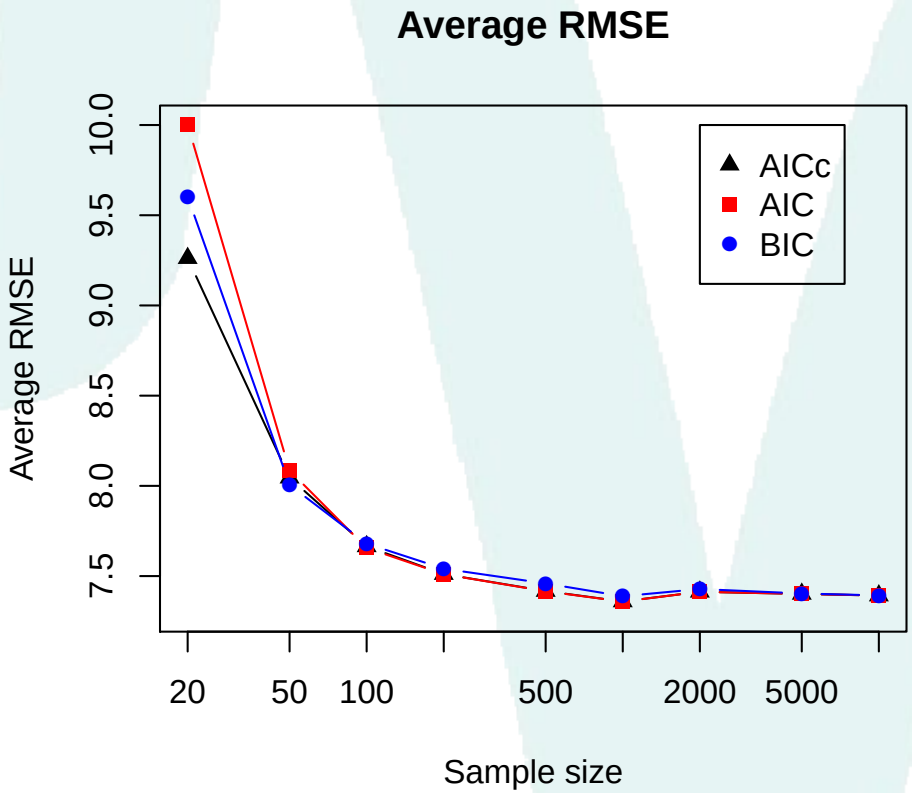
Example 2 (Tapering effects):
high correlation variability (**heterogeneity**)

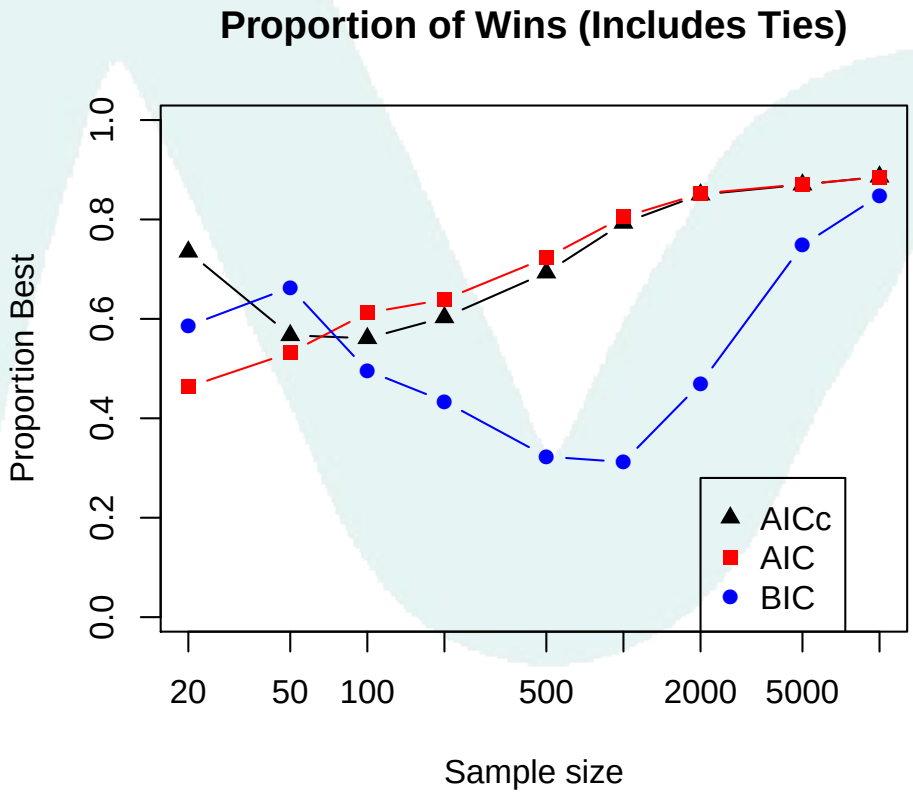




Simulation study

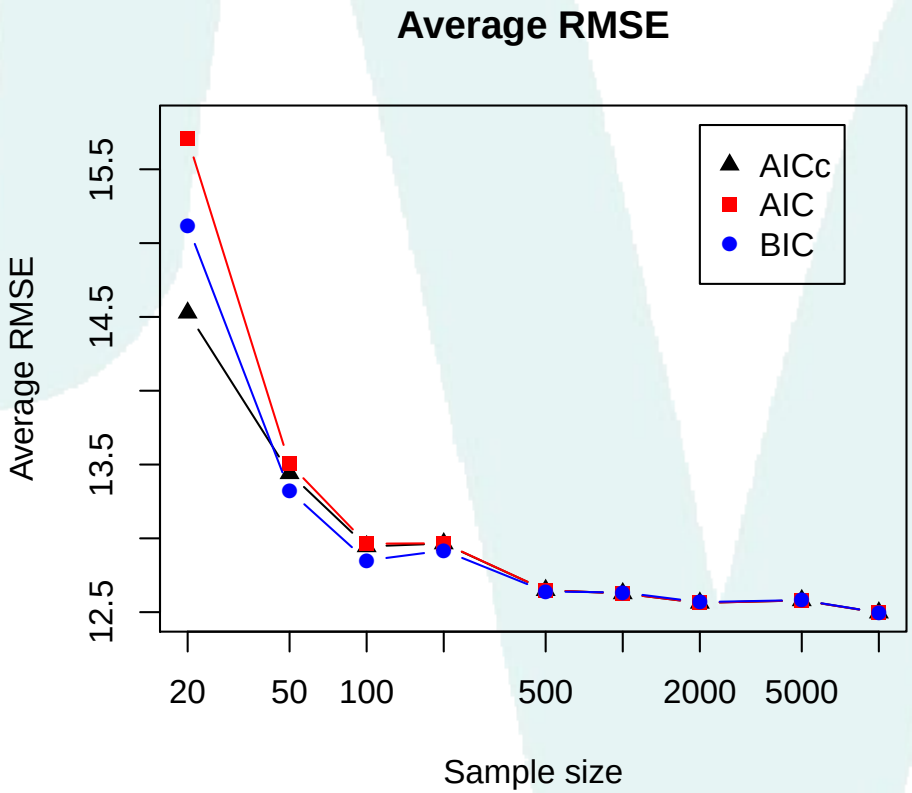
Example 2 (Tapering effects):
low parameter variability (**heterogeneity**)

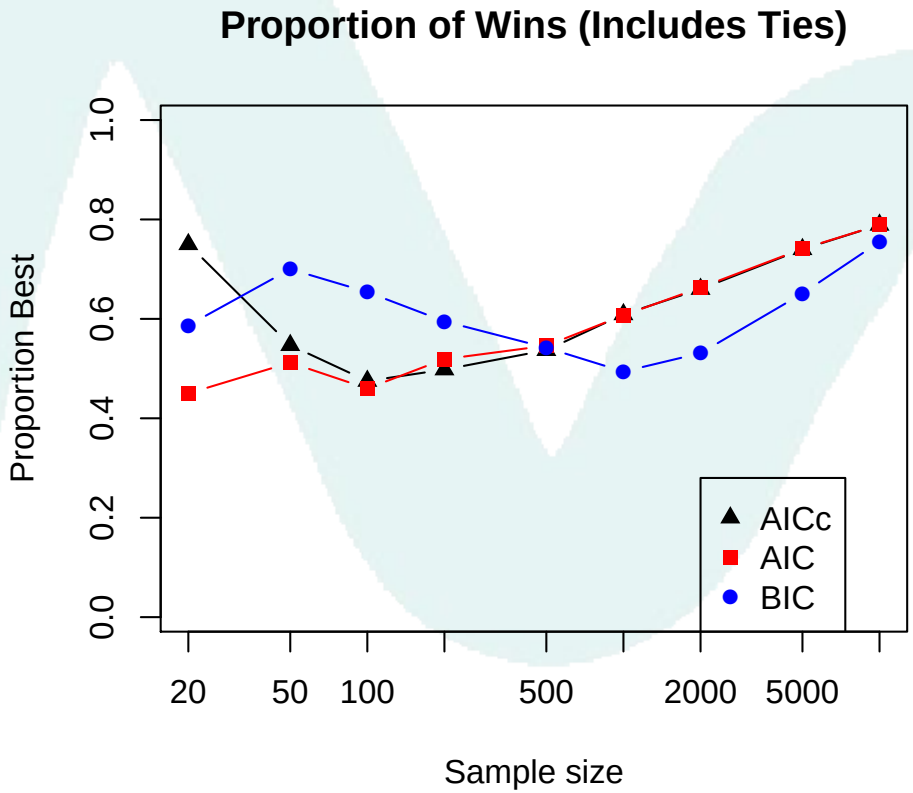




Simulation study

Example 2 (Tapering effects):
high parameter variability (**heterogeneity**)





Simulation study — summary

- **Strength** of between-data set heterogeneity important
- BIC better than AIC for more heterogeneity
- Be more cautious for predictive model — use a **stronger penalty**
- **BUT** RMSE differences often small — does it always matter?

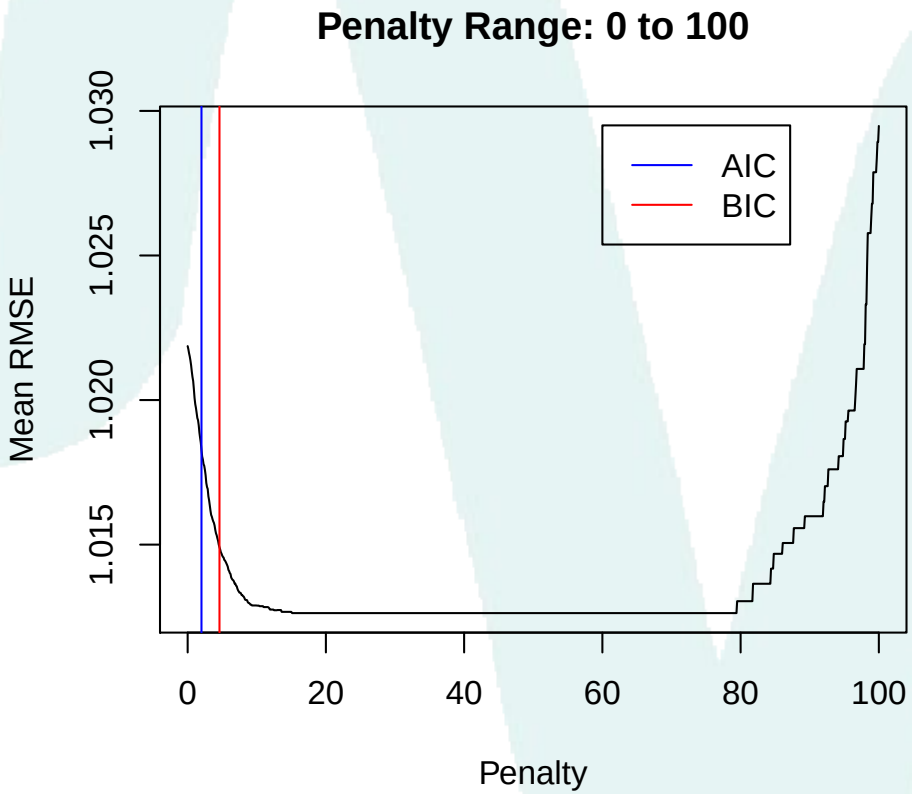
So what is “optimal”?

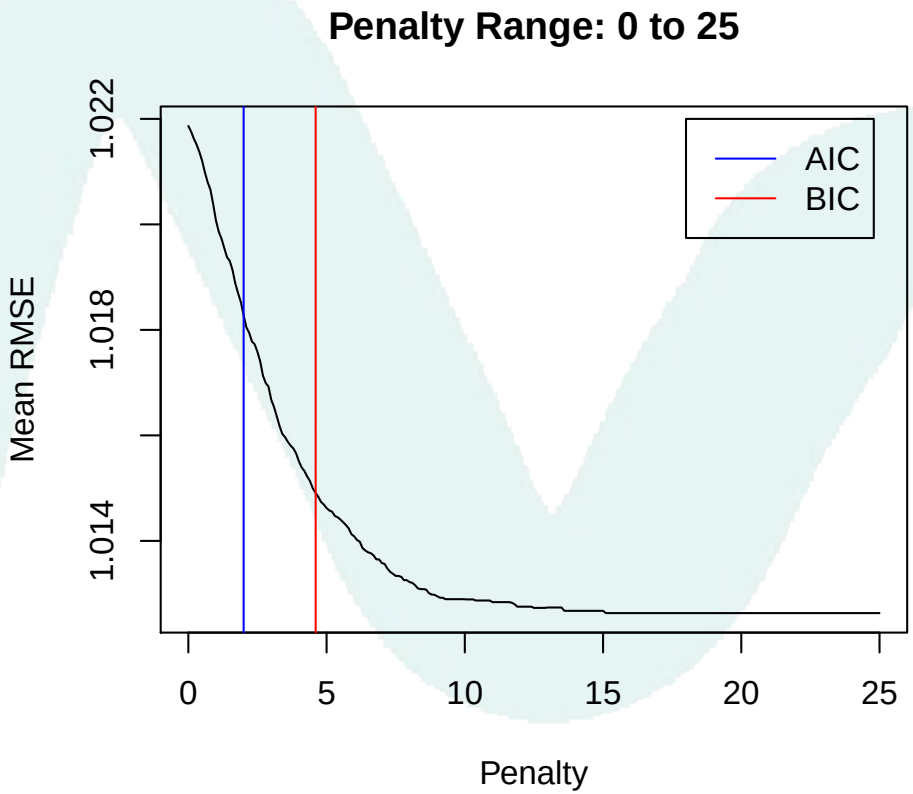
Simulation study, prediction vs. penalty:

- Consider penalty as a **multiplier** on p
- AIC multiplier is 2
- BIC multiplier is $log(n)$
- **How is mean RMSE for predictions affected by choice of penalty?**
- Two examples, $n = 100$, 2000 sims. . .

So what is “optimal”?

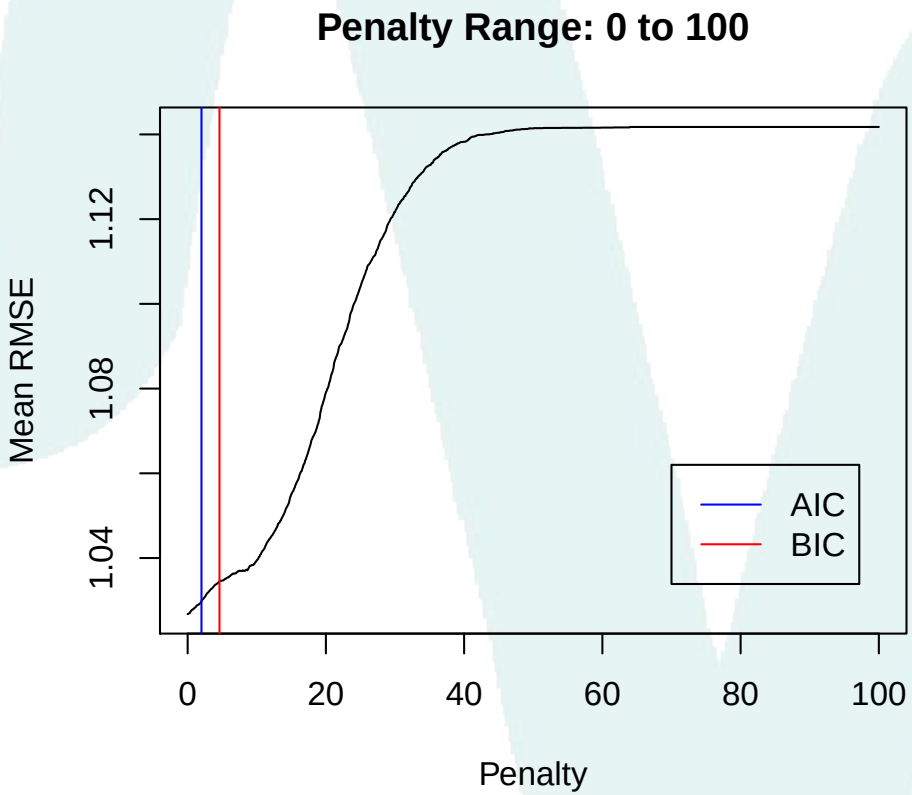
Example 3 — two strong effects, one very weak

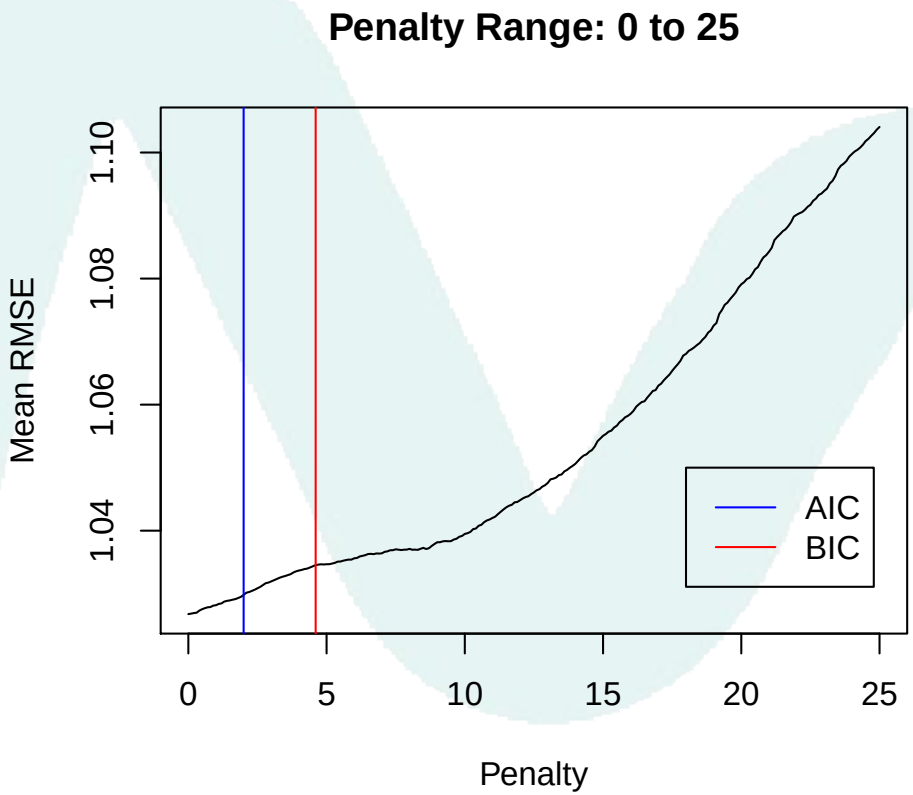




So what is “optimal”?

Example 4 — two weak effects, one very weak





So what is “optimal”?

- Optimal penalty dependent on **relative effect sizes**
- Neither AIC nor BIC is “best” ($n = 100$)
- Suggestion: weakest effect “contaminates” strongest effect in Example 3 through the correlation
- These were **simple** examples. . .
- “Small” samples $\approx$ not infinite??

Discussion

- Some people **really love** AIC
- Between-sample heterogeneity means AIC not always optimal
- Can use `ICsims` to explore penalties with real data
- AIC-chosen models may not be **transferable**(†) from one place/time to another

(†)Wenger and Olden (2012) *Methods in Ecology and Evolution*

Miscellanea

Model importance

Akaike weights for R models as:

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{r=1}^{R} \exp\left(-\frac{1}{2}\Delta_r\right)}$$

where $\Delta_i = \text{AIC}_i - \text{AIC}_{\text{min}}$.

Idea is to quantify relative "importance" of models.

Model importance

Var 1	Var 2	Var 3	Var 4	Var 5	AIC	ΔAIC	Weight
	-0.539 (± 0.244)		**-0.602** (± 0.190)		1789.73	0.00	0.26
	-0.674 (± 0.336)		**-0.609** (± 0.192)	0.173 (± 0.295)	1791.38	1.65	0.12
	-0.544 (± 0.245)	0.003 (± 0.008)	**-0.566** (± 0.214)		1791.60	1.86	0.10
-0.090 (± 0.333)	**-0.541** (± 0.244)		**-0.574** (± 0.217)		1791.66	1.93	0.10
			-0.641 (± 0.201)		1792.23	2.50	0.08
-0.070 (± 0.335)	**-0.670** (± 0.336)		**-0.586** (± 0.220)	0.167 (± 0.296)	1793.34	3.61	0.04
			-0.622 (± 0.198)	-0.212 (± 0.222)	1793.34	3.61	0.04
	-0.662 (± 0.344)	0.001 (± 0.008)	**-0.591** (± -0.591)	0.155 (± 0.316)	1793.35	3.62	0.04

Model importance

Akaike weights for R models as:

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{r=1}^{R} \exp\left(-\frac{1}{2}\Delta_r\right)}$$

where $\Delta_i = \text{AIC}_i - \text{AIC}_{\text{min}}$.

Terminology often loose here, using “weight”, “probability” and even “likelihood” interchangeably.

Probability ok. . . from a Bayesian perspective?

Model importance

model	predictors
1	x_1
2	x_1, x_2
3	x_1, x_2, x_3
4	x_1, x_2, x_3, x_4
5	x_1, x_2, x_3, x_4, x_5
6	$x_1, x_2, x_3, x_4, x_5, x_6$

- $y = 100 + 15x_1 + 10x_2 + 5x_3 + 3x_7 + \epsilon$
- Mean model weights ($n =$ sims $= 1000$): 0.000, 0.000, 0.449, 0.262, 0.168, 0.121
- So any model which is not too many parameters different from the "best" model will have reasonable weight

Variable importance

- Again use Akaike weights
- Relative importance for variable x_i is sum of weights for models containing x_i
- Dependent on model set, and **balance** of variables in models
  - B & A suggest using “balanced” model set
- **However** this then restricts choice of “candidate set”…

Variable importance

- Again: AIC says nothing about variables *per se*
- Variables gain “importance” in “+1” models
- Note, weights essentially arbitrary for “spurious” variables
- Murray and Conner (2009) *Ecology*
  - Plain correlations to find spurious variables
  - Hierarchical partitioning on remainder
  - “Akaike weights performed poorly in all respects”

Spin

Burnham and Anderson, *Ecology* (2014)

HISTORICAL STATISTICS

Murtaugh (2014) reviews several of the historical methods for data analysis in simple situations; these methods focus on "testing" a null hypothesis by computing a "test statistic," assuming its asymptotic distribution, setting an arbitrary α level, and computing a P value. The P value usually leads to an arbitrary simplistic binary decision as to whether the result is "statistically significant" or not. In other cases, the P value is stated and interpreted as if it were evidential. The P value is defined as the pre-data probability: Prob{*a test statistic as large as, or larger, than that observed, given the null*}. That is, the anticipated data are being thought of as random variables.

Spin

Burnham and Anderson, *Ecology* (2014)

21st-Century Statistical Science

Methods based on Bayes theorem or Kullback-Leibler information (Kullback and Leibler 1951) theory allow advanced, modern approaches and, in this context, science is best served by moving forward from the historical methods (progress should not have to ride in a hearse). We will focus on the information-theoretic methods in the material to follow. Bayesian methods, and the many data resampling methods, are also useful and other approaches might also become important in the years ahead (e.g., machine learning, network theory). We will focus on the IT approaches as they are so compelling and easy to both compute and understand. We must assume the reader has a basic familiarity with IT methods (see Burnham and Anderson 2001, 2002, 2004, Anderson 2008).

Acknowledgements

Rural and Environment Science and Analytical Services (RESAS) Division of the Scottish Government