Behav Ecol Sociobiol (2011) 65:13–21
DOI 10.1007/s00265-010-1037-6

REVIEW

A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion

Matthew R. E. Symonds · Adnan Moussalli

Received: 19 April 2010 / Revised: 19 July 2010 / Accepted: 29 July 2010 / Published online: 25 August 2010

Abstract Akaike's information criterion (AIC) is increasingly being used in analyses in the field of ecology. This measure allows one to compare and rank multiple competing models and to estimate which of them best approximates the "true" process underlying the biological phenomenon under study. Behavioural ecologists have been slow to adopt this statistical tool, perhaps because of unfounded fears regarding the complexity of the technique. Here, we provide, using recent examples from the behavioural ecology literature, a simple introductory guide to AIC: what it is, how and when to apply it and what it achieves. We discuss multimodel inference using AIC—a procedure which should be used where no one model is strongly supported. Finally, we highlight a few of the pitfalls and problems that can be encountered by novice practitioners.

Keywords Akaike's information criterion · Information theory · Model averaging · Model selection · Multiple regression · Statistical methods

Communicated by L. Garamszegi

This contribution is part of the Special Issue "Model selection, multimodel inference and information-theoretic approaches in behavioural ecology" (see Garamszegi 2010).

M. R. E. Symonds (✉)
Department of Zoology, University of Melbourne,
Melbourne, Victoria 3010, Australia
e-mail: symondsm@unimelb.edu.au

A. Moussalli
Sciences Department, Museum Victoria,
GPO Box 666E, Melbourne, Victoria 3001, Australia

Introduction

Increasingly, ecologists are applying novel model selection methods to the analysis of their data. Of these novel methods, information theory (IT) and in particular the use of Akaike's information criterion (AIC) is becoming widespread (Akaike 1973; Burnham and Anderson 2002; Garamszegi 2010). Unfortunately, the literature describing AIC can be intimidating to those who are not fluent with statistical phraseology. This short introduction is intended for those behavioural ecologists who are unfamiliar with the practicalities of AIC.

Considerable literature exists discussing the origin, philosophy and application of AIC (e.g. Burnham and Anderson 2001, 2004; Burnham et al. 2010), and criticism of AIC is likewise prevalent (e.g. Guthery et al. 2005; Richards 2005; Stephens et al. 2005; Link and Barker 2006). It is not our aim here to provide in-depth discussion of the philosophical background to AIC, nor to advocate or discourage use of the method. Such discussion can be found in the other contributions to this special issue. Our intention is simply to provide, for those intending to use the method, a basic user's guide to model selection and multimodel inference using AIC. For a full background to AIC, readers are referred to the key text by Burnham and Anderson (2002). Additionally, Hilborn and Mangel (1997), Johnson and Omland (2004), Mazerolle (2006), Towner and Luttbeg (2007) and Stephens et al. (2007) have provided some excellent overviews of techniques of model selection for ecologists more generally.

Using AIC in behavioural ecology

Most behavioural ecologists use traditional statistics involving null hypothesis significance testing (NHST), with

assessment of significance through associated p values. Mundry (2010) and Burnham et al. (2010) provide discussion of the philosophical and inferential differences between the NHST approach and IT-AIC approach. In behavioural ecology, the scenario where using AIC is likely to be proposed is when an analysis explores a range of variables that may be associated with a particular trait or behaviour. Such studies are often fundamentally explorative, seeking to identify strong associations worthy of further investigation and experimentation. For example, Thorup et al. (2006) sought to identify the environmental covariates associated with migration behaviour decisions in ospreys on a particular day—covariates that included wind speed and direction, precipitation, time of year and previous migration history. Thus, the model involved is inherently multivariate (i.e. has more than one possible predictor), and accordingly, multiple models will need to be considered. A traditional approach might seek to identify a 'best model' through forward or backward stepwise variate selection, a procedure whose shortcomings are well documented (Whittingham et al. 2006; but see Hegyi and Garamszegi 2010). Not least of these problems is that parameters can appear as significant or non-significant, depending on what other parameters are present in the model.

AIC compares multiple competing models (or working hypotheses) all at once, asking "how certain are we that any given model is the best approximating model?" In doing so, model selection uncertainty can be quantified and accounted for, and inference can be based on a set of models in cases where no single model stands out as being the best model. AIC therefore enables the user to make biological inferences that are unconditional on a specific model (as do other information criteria, such as the Bayesian Information Criterion—see Johnson and Omland 2004). If there is any uncertainty over the model, then you are implicitly in a multiple-model framework, whether you admit it or not. AIC provides a means of expressing and evaluating this explicitly.

Any model that we produce is, at best, only going to be an approximation of the biological phenomenon being studied. We can never actually know exactly what determines every aspect of, say, migration behaviour in ospreys. So many factors are involved that the truth would be irreducibly complex. Any model we test will thus only be an approximation of the truth. Burnham et al. (2010) describe the fundamentals underlying the formulation of AIC under information theory, but to briefly encapsulate, AIC is a numerical value by which to rank competing models in terms of information loss in approximating the unknowable truth. Accordingly, AIC as a value by itself is meaningless. It derives meaning from comparison with the AIC values of other models with the model having the lowest AIC value representing the 'best approximating model'. As we shall see, however, there is often uncertainty regarding the identity of the best approximating model.

Calculating AIC

Calculation of AIC is not difficult. Recent versions of most statistical software packages provide AIC values for general linear models (Table 1). AIC is calculated using the number of fitted parameters, including the intercept, in the model (k), and either the maximum likelihood estimate for the model (L) or the residual sum of squares of the model (RSS), two measures that are also easily derived from the output of any statistics package. In the case of least-squares regression analyses, the value of k must be increased by 1 to reflect the variance estimate as an extra model parameter.

AIC is calculated as

$$\mathrm{AIC} = -2\ln(L) + 2k$$

if using likelihood or

$$\mathrm{AIC} = n\left[\ln\left(\frac{\mathrm{RSS}}{n}\right)\right] + 2k$$

if using residual sum of squares, where n is the sample size.

For small sample sizes (roughly approximated as being when n/k is less than 40 and k is the number of fitted parameters in the most complex model), a modified version of AIC ($\mathrm{AIC_c}$) is recommended:

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1}$$

In practice, because $\mathrm{AIC_c}$ approximates AIC at large sample sizes, it is often advised that $\mathrm{AIC_c}$ is used as default (but see "Problems and pitfalls" later).

AIC is affected by overdispersion in the data, that is when there is more variability in the data than would be expected from the fitted model (i.e. the model is a poor fit). Overdispersion is very common with count data, which are typically modelled using Poisson regression. The causes of overdispersion are numerous, and there are several ways of dealing with it (see Richards 2008 for a recent primer). In terms of AIC analyses, it is usually recommended that QAIC is used:

$$\mathrm{QAIC} = \frac{-2\ln(L)}{\hat{c}} + 2k$$

where $\hat{c}$ is the variance inflation factor or overdispersion coefficient that is sometimes generated by statistics packages (calculated as the χ^2 goodness of fit of the most complex model in the candidate set divided by its degrees

Table 1 Major statistical packages and how they implement Akaike's information criterion

Package	Website	Versions of AIC calculated	Additional Notes
Genstat 12	www.vsni.co.uk/software/genstat/	AIC only—automatically calculated for generalised linear models and for restricted maximum likelihood (REML) for linear mixed models	For generalised linear models, the model variance is not taken into account in the count of fitted parameters; for REML, the variance parameters in the random model are included in the parameter count
JMP 8	www.jmp.com	AIC_c only—automatically calculated when analysing in the stepwise regression menu	Stepwise regression menu also allows one to compare AIC_c across all possible models, calculate Akaike weights and perform model averaging
Minitab 15	www.minitab.com	None	
SPSS 18	www.spss.com/statistics/18/	AIC and AIC_c—calculated in several procedures (e.g. generalised linear models, mixed models, time series analysis)	For generalised linear models, the model variance is included in the parameter count
R	http://www.r-project.org/	AIC only, in two commands: extractAIC {stats} &AIC {stats}	For full details, see http://stat.ethz.ch/R-manual/R-patched/library/stats/html/AIC.html
SAS 9.2	www.sas.com	AIC and AIC_c supplied as part of the 'Fit Statistics' table in numerous procedures	For simple and generalised linear models, the model variance is not taken into account in the count of fitted parameters, but it is included for the generalised linear mixed model procedure. AICc is only provided under the generalised linear and mixed model applications. Number of parameters is explicitly stated in the output
Statistica 9	www.statsoft.com	AIC only—calculated as part of regression output	Provides an option to report AIC for all possible models. Output shows how many parameters are used
Systat 12	www.systat.com	AIC and AIC_c supplied as part of output of numerous procedures	

of freedom). QAIC should be employed if $\hat{c}$ is greater than 1, and, since the overdispersion coefficient is a parameter, k should be increased by 1 (Burnham and Anderson 2001). As with AIC, a version of QAIC for small sample sizes, $QAIC_c$, can be employed:

$$QAIC_c = QAIC + \frac{2k(k+1)}{n-k-1}$$

Whichever version of AIC is employed, it must be consistent across models (i.e. do not mix AIC, AIC_c, QAIC and $QAIC_c$). Hereafter, we shall use only the term AIC, but our discussion applies to any of these iterations.

As mentioned above, the models are ranked by AIC, with the best approximating model being the one with the lowest (most negative) AIC value. AIC thus takes into account how well the model fits the data (by using likelihood or RSS), but models with greater numbers of fitted parameters (k) will have higher AIC values, all other things being equal. In other words, models with fewer parameters will be favoured.

To illustrate with an example, we shall use an analysis of song structures in dark-eyed juncos (*Junco hyemalis*) originally analysed by Cardoso et al. (2007) with standard NHST and stepwise approaches. Specifically, we focus on syllable length and whether there are tradeoffs between this and other aspects of song structure and complexity. For example, longer syllables may require longer gaps (for recovery) between utterances, or songs that might be more complex in other regards (greater range of frequency, more trills, etc.) might have shorter syllables. The data consist of 188 different syllable types (n=188), with eight song complexity variables being examined (see Table 2 for details), plus intercept and variance as additional fitted parameters in the model (k=10). As with any statistical analysis, problems can arise using highly correlated traits as independent predictor variables. In this case, an analysis of the full model (i.e. the model with all eight variables included) indicated substantial tolerance between variables, suggesting that all eight variables are sufficiently independent (see also Cardoso et al. 2007). Nor was there any indication of overdispersion in the data set. Since n/k is less than 40, calculation of AIC_c is the most appropriate. We calculated AIC_c values for every possible combination of variables and intercept (what is known as an all-subset approach). The results of this are shown in Table 2. This all-subset approach to model selection is one that is likely to be used by behavioural ecologists with observational data and several putative predictor variables; however, it is important to realise that this approach is fundamentally explorative (i.e. more about hypothesis generating than testing, providing a basis for subsequent work and data collection where one more explicitly tests a specific number of hypotheses). Such an approach can be seen as a form of 'fishing expedition', and Burnham and Anderson (2002,

Table 2 95% confidence set of best-ranked regression models (the 24 models whose cumulative Akaike weight, acc w_i, ≤0.95) examining effect of song complexity variables on syllable length in the song of dark-eyed juncos

	Candidate models	*k*	RSS	AIC_c	Δi	w_i	acc w_i	ER
1	FB + NFI + NE + LR + LG	7	0.05155	−1527.294	0	0.295	0.295	
2	FB + PF + NFI + NE + LR + LG	8	0.05143	−1525.557	1.737	0.124	0.418	2.38
3	FB + NFI + NE + L2V + LR + LG	8	0.05145	−1525.462	1.832	0.118	0.536	2.50
4	FB + NFI + NE + LH + LR + LG	8	0.05155	−1525.115	2.179	0.099	0.636	2.97
5	FB + PF + NFI + NE + L2V + LR + LG	9	0.05130	−1523.830	3.465	0.052	0.688	5.65
6	FB + PF + NFI + NE + LH + LR + LG	9	0.05143	−1523.354	3.940	0.041	0.729	7.17
7	FB + NFI + NE + LH + L2V + LR + LG	9	0.05145	−1523.255	4.039	0.039	0.768	7.53
8	FB + NFI + NE + LR	6	0.05344	−1522.681	4.613	0.029	0.797	10.04
9	FB + NE + LR + LG	6	0.05373	−1521.668	5.627	0.018	0.814	16.66
10	FB + PF + NFI + NE + LH + L2V + LR + LG	10	0.05130	−1521.598	5.696	0.017	0.832	17.25
…								
23	FB + NFI + NE + LH + L2V + LR	8	0.05323	−1521.869	8.230	0.005	0.946	59.00
24	FB + NFI + NE + LH + L2V + LG	8	0.05329	−1521.670	8.429	0.004	0.950	73.75

Complete descriptions are provided in Cardoso et al. (2007). For definition of other terms, see text

FB frequency bandwidth, *PF* peak frequency, *NFI* number of frequency inflections, *NE* number of elements, *LH* length of harmonics, *L2V* length of two voices, *LR* length of rattles, *LG* length of gaps

p. 147) call it "poor strategy". However, we suspect that, in behavioural ecology, there often exists insufficient knowledge of the system under study such that explorative methods like the all-subset approach present above are the only way forward. Further, how one defines a reduced candidate set of models is also a complex matter (see Dochtermann and Jenkins 2010; Burnham et al. 2010). Finally, if the aim is to proceed beyond model selection and produce a predictive model through multimodel inference and model averaging, as we shall do here, then the all-subset approach with full-model averaging needs to be employed.

In the example, the best AIC model contains the same five variables identified as significant predictors by Cardoso et al. (2007) using backward stepwise selection: frequency bandwidth, number of frequency inflections, number of elements, length of rattles and length of gaps. However, this should not be taken as tacit confirmation that AIC automatically produces the same result as stepwise selection. Additionally, making inference based on the best approximating model alone may not be desirable. One of the main purposes of calculating AIC is to present a range of models and their relative AIC scores. By comparing the different models, we can measure *how much* better the best approximating model is compared to the next best models. The simplest way of doing this is to calculate the difference (Δ_i or ΔAIC_i) between the AIC value of the best model and the AIC value for each of the other models.

Δ_i is used to calculate two additional measures used to assess the relative strengths of each candidate model (Burnham and Anderson 2002, pp. 75–79). The first of these is the evidence ratio (ER):

$$\mathrm{ER} = \frac{\exp\left(-\frac{1}{2}\Delta_{best}\right)}{\exp\left(-\frac{1}{2}\Delta_i\right)}$$

which provides a measure of how much more likely the best model is than model *i* (Δ_{best} is the Δ value for the best model=0). In Table 2, the first model is approximately 2.4 and 2.5 times more likely to be the best approximating model than the second and third models, respectively The evidence ratio can also be used to compare any two models individually (simply replace Δ_{best} in the equation with Δ_j, the Δ value for model *j*).

The second, more commonly seen, measure is the Akaike weight, w_i. The Akaike weight for a given model, *i*, is calculated from Δ_i values as:

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

The Akaike weight is a value between 0 and 1, with the sum of Akaike weights of all models in the candidate set being 1, and can be considered as analogous to the probability that a given model is the best approximating model (although there are some who disagree with this, e.g. Link and Barker 2006; Bolker 2008; Richards 2005). Thus, in Table 2, the best model has a w_i of 0.295—which can be interpreted as meaning that there is 29.5% chance that it really is the best approximating model describing the data given the

candidate set of models considered. With this relatively low weight, we cannot be certain that this model is the best. That is to say, there exists model selection uncertainty.

Model weights can also be used to estimate the relative importance of variables under consideration. This is done by summing the Akaike weights for each model in which that variable appears. In our example, the variable 'peak frequency' (PF) has an Akaike weight=0.124+0.052+0.041+0.017+…. and so on down the complete list of models. If a particular predictor appears in all of the top models, then its summed Akaike weight will tend towards 1. If that predictor only appears in the very unlikely models, its weight will tend towards 0. As with the Akaike model weight (w_i), the predictor weight can be interpreted as equivalent to the probability that that predictor is a component of the best model. Similarly, these summed weights can be used to rank the various predictors in terms of importance (Burnham and Anderson 2002, p. 168). The Akaike weights for the predictors in the junco example are shown in Table 3. Reassuringly, the five variables identified by Cardoso et al. (2007) and that feature in the best AIC model all have high (>0.9) parameter weights, considerably higher than the remaining three variables.

Rejecting models

In any set of models being compared, some will obviously be 'better' (i.e. have lower AIC scores) than others, but can we discount some models altogether? Clearly, the candidate models, in the first place, should make biological sense, and due to the relative nature of AIC, you should not be comparing models that are all poor fits for the data (see "Problems and pitfalls"). Thus, qualifying the utility or worth of either the global model or the best AIC model in terms of 'goodness of fit' is essential. There is still debate about when a model can be considered uninformative (see Richards et al. 2010; Burnham et al. 2010), but as a coarse guide, models with Δ_i values less than 2 are considered to be essentially as good as the best model, and models with Δ_i up to 6 should probably not be discounted (Richards 2005). Above this, model rejection might be considered, and certainly models with Δ_i values greater than 10 are sufficiently poorer than the best AIC model as to be considered implausible (Burnham and Anderson 2002).

A rather less procrustean approach is to produce a 'confidence set' or 'credibility set' of models that are the most realistically likely to be the best approximating model (Burnham and Anderson 2002). This is done by ranking all the models from the best downwards and

Table 3 Model-averaged estimates for eight aspects of song complexity predicting syllable length in juncos

	Int	FB	PF	NFI	NE	LH	L2V	LR	LG
1	0.011 (0.0056)	0.005 (0.0015)		0.006 (0.0022)	0.023 (0.0026)			0.205 (0.071)	1.066 (0.406)
2	0.021 (0.0156)	0.005 (0.0015)	-1.8×10^{-6} (2.84×10^{-6})	0.006 (0.0022)	0.023 (0.0026)			0.204 (0.071)	1.064 (0.406)
3	0.011 (0.0056)	0.005 (0.0016)		0.006 (0.0022)	0.023 (0.0027)		−0.098 (0.166)	0.193 (0.074)	1.041 (0.408)
4	0.011 (0.0056)	0.005 (0.0016)		0.006 (0.0022)	0.023 (0.0027)	0.010 (0.167)		0.205 (0.072)	1.067 (0.406)
5	0.023 (0.0158)	0.005 (0.0016)	-2.1×10^{-6} (2.86×10^{-6})	0.007 (0.0023)	0.022 (0.0027)		−0.116 (0.167)	0.189 (0.074)	1.034 (0.408)
6	0.021 (0.0156)	0.005 (0.0016)	-1.9×10^{-6} (2.84×10^{-6})	0.006 (0.0023)	0.023 (0.0027)	0.010 (0.167)		0.203 (0.072)	1.064 (0.406)
7	0.011 (0.0057)	0.005 (0.0017)		0.006 (0.0023)	0.023 (0.0028)	−0.001 (0.168)	−0.098 (0.167)	0.194 (0.075)	1.041 (0.408)
8	0.010 (0.0057)	0.004 (0.0015)		0.007 (0.0022)	0.025 (0.0025)			0.243 (0.071)	
9	0.013 (0.0056)	0.006 (0.0015)			0.024 (0.0027)			0.173 (0.072)	1.160 (0.413)
10	0.023 (0.0158)	0.005 (0.0017)	-2.2×10^{-6} (2.86×10^{-6})	0.007 (0.0023)	0.022 (0.0028)	−0.002 (0.167)	−0.116 (0.168)	0.189 (0.075)	1.034 (0.408)
w		0.982	0.298	0.945	1	0.256	0.3	0.934	0.906
$\widehat{\bar{\beta}}$	0.014	0.005	-2.0×10^{-6}	0.006	0.023	0.020	−0.113	0.204	1.075
$\bar{\beta}$	0.014	0.005	-5.9×10^{-7}	0.006	0.023	0.005	−0.034	0.190	0.974
$\widehat{se}\left(\widehat{\bar{\beta}}\right)$	0.0090	0.0015	2.5×10^{-6}	0.0021	0.0025	0.142	0.155	0.070	0.389
$\widehat{se}\left(\bar{\beta}\right)$	0.0090	0.0015	7.8×10^{-7}	0.0020	0.0025	0.033	0.047	0.064	0.348

Typically, the standard model-averaging method would only consider those variates in the best AIC model. For comparison, however, estimates ($\widehat{\bar{\beta}}$ and $\bar{\beta}$, see text) from both model-averaging methods are shown for all covariates. Given that this example is based on an all-subset candidate set and the best AIC model is not strongly weighted, full-model averaging would be the preferred approach. As in Table 2, parameter estimates (±SE) from the top 10 models are shown, and weighted averages for estimates and error are displayed at the bottom of the table

FB frequency bandwidth, *PF* peak frequency, *NFI* number of frequency inflections, *NE* number of elements, *LH* length of harmonics, *L2V* length of two voices, *LR* length of rattles, *LG* length of gaps, *Int* intercept, *w* variate weight

proceeding down the list until the cumulative Akaike weight exceeds 0.95, or whatever value you choose, and then rejecting the rest. This produces a 95% confidence set of models—in other words, we are 95% confident that one of the models within this credibility set is the best approximating model. In our example, there were 24 models constituting the 95% confidence set (Table 2). The 95% value is arbitrary and derives from the frequentist approach but is the one most people are familiar with when dealing with confidence.

One hazard of considering a lot of models, as typically occurs in an all-subset context, is that the interpretation of Δ_i can become problematic because models strongly competing with the best AIC model (i.e. $\Delta_i<2$) may differ very little structurally, having merely one additional parameter for instance. Little has been gained by making the model more complex, with the likelihood of the 'nested' best AIC model and competing models being essentially equivalent (Burnham and Anderson 2002, p. 131). Thus, models 2 to 7 in our example (Table 2) are simply slightly more complex versions of model 1. In the interests of parsimony, Richards (2008, see also Richards et al. 2010 for further discussion) recommends post hoc elimination from the candidate set models that are more complicated versions of any model with a lower AIC value. However, Richards' technique has not been widely employed, and as yet it is uncertain whether inference is consistently improved by this approach.

Model selection uncertainty and multimodel inference

Occasionally, IT-AIC analyses present clear-cut results where the Akaike weight of the best model is considerably higher than the next best model. In situations where one can clearly identify a best model (say it has an Akaike weight of >0.9—Burnham and Anderson 2002), it is appropriate to make inference based on that model alone. For example, in an analysis of likelihood of clutch desertion by mallards with experimentally manipulated broods, Ackerman and Eadie's (2003) best model, which included proportion of clutch remaining and clutch age as predictors, had an Akaike weight of 0.91, with the next best model having a relatively paltry weight of 0.09 (the only other model considered had negligible weight). Often, however, AIC analyses can identify several, perhaps dozens, of 'almost as good' models. If, for example, there are several models with Δ_i less than 2, then they strongly compete for the position of being the best approximating model. It is clear that in this case only presenting the best model would be disingenuous. Presenting all the models, or at least the confidence set of best models, is essential. Likewise, presentation of the predictor weights (see earlier) helps to measure the relative likelihood that each predictor is part of the best model. Moreover, where model selection uncertainty is evident, inference needs to be multimodel-based, that is, model averaging should be employed using the full set of models.

Model averaging

Model averaging produces parameter and error estimates that are not conditional on any one model but instead derive from weighted averages of these values across multiple models. Model averaging in an AIC framework is still an area that raises some thorny issues concerning both methodology and indeed whether it improves inference at all (see Richards 2005; Richards et al. 2010). When it comes to actually doing the model averaging, there are two subtly different approaches that can produce quite different results. The first approach derives from what Burnham and Anderson (2002) refer to as 'natural averaging', as it keeps the averaged parameter estimate in the original scale. This approach to model averaging is applied in cases where there is strong, but not unequivocal (e.g. w>0.90), support for the best AIC model, and parameters are averaged only for those variates in the best AIC model (Buckland et al 1997). The averaged parameter estimate is calculated as follows:

$$\widehat{\overline{\beta}} = \frac{\sum_{i=1}^{R} w_i \widehat{\beta}_i}{\sum_{i=1}^{R} w_i}$$

where $\widehat{\beta}_i$ is the estimate for the predictor in a given model *i,* and w_i is the Akaike weight of that model. In this instance, $\widehat{\beta}_i$ is averaged only over the models for which the variate of interest appears. For instance, in the junco example in Table 2, $\widehat{\beta}_i$ and w_i values for PF would only be taken from models 2, 5, 6, 10 and so on. Unconditional variance (an estimate of variance not conditional on a single model) can then be calculated using the following equation:

$$\widehat{var}\left(\widehat{\overline{\beta}}\right) = \left[\sum w_i \sqrt{\widehat{var}\left(\widehat{\beta}_i\right) + \left(\widehat{\beta}_i - \widehat{\overline{\beta}}\right)^2}\right]^2$$

where $\widehat{var}\left(\widehat{\beta}_i\right)$ is the variance of the parameter estimate in model *i*, and $\widehat{\beta}_i$ and $\widehat{\overline{\beta}}$ are as defined above (Buckland et al. 1997). Note that AIC weights for the subset of models for which the variate of interest appears need to be renormalized (i.e. summed up to 1) before calculating the unconditional variance. For calculation of standard error, $\widehat{se}\left(\widehat{\overline{\beta}}\right)$, one simply omits the overall squared term. Notice that there are two components to this unconditional error:

error in parameter estimation $\widehat{var}\left(\widehat{\beta}_i\right)$, and the error due to model selection uncertainty $\left(\widehat{\beta}_i - \widetilde{\widehat{\beta}}\right)^2$. Consequently, unconditional error estimates are typically greater than when inference is based on a single model.

The second approach to model averaging, known as full-model averaging (see Lukacs et al. 2009), should be employed in cases where high model selection uncertainty exists (i.e. the best AIC model is not strongly weighted); thus, inference needs to be based on all models in the candidate set. Such a situation typically arises in all-subset modelling, as in our example. Here, the estimator is denoted as $\widetilde{\widehat{\beta}}$; parameter estimates for all variates of the global model are averaged, and *all* models are considered. Models not containing the variates of interest simply contribute zero to the calculation of the average. Hence, the above formula simply reduces to:

$$\widetilde{\widehat{\beta}} = \sum_{i=1}^{R} w_i \widehat{\beta}_i$$

Consequently, $\widehat{\beta}$ shrinks towards zero by the amount representing the degree by which the variate is uninformative (i.e. $\widetilde{\widehat{\beta}} = \widehat{\beta}(1 - \text{variate weight})$). The second approach essentially produces a predictive formula for the global model (in our example, all eight variables) with each averaged parameter estimate being weighted so that variates with low Akaike parameter weights will have little influence on prediction. Note, in terms of predictions, this method in parameter averaging yields identical results to weighted averaging of model predictions. A simple analytical estimator of the unconditional variance under full-model averaging unfortunately remains elusive. Burnham and Anderson (2002) do recommend an estimate worthy of further investigation (see also Lukacs et al. 2009), namely:

$$\widehat{var}\left(\widetilde{\widehat{\beta}}\right) = \sum w_i \left[\widehat{var}\left(\widehat{\beta}_i\right) + \left(\beta_i - \widetilde{\widehat{\beta}}\right)^2\right]$$

For illustrative purpose, Table 3 illustrates both approaches for the junco data. The averaged parameter estimates indicate that the five variates with high Akaike weights and that feature in the best approximating model are all positively related to syllable length. This indicates that longer syllables require longer gaps between utterances (suggesting an energetic tradeoff) but that there are no tradeoffs between other aspects of song complexity—rather syllable length reflects the frequency bandwidth, number of frequency inflections, number of elements and length of rattles produced by juncos. Note, in contrast to the five important variables, model-averaged parameter estimates for peak frequency, length of harmonics and length of two voices are smaller than they are using the first model-averaging approach, reflecting their low weight (see Table 3). If the aim is to formulate how a particular predictor relates to the response variable, then these shrunken estimates may not be biologically realistic or helpful. On the other hand, if the aim is prediction, then you want poorly weighted variates to contribute less to the prediction than strongly weighted variates.

Problems and pitfalls

Any statistical analysis in behavioural ecology is only worthwhile if the variables chosen are biologically meaningful (i.e. might reasonably be thought to be linked to the trait of interest). Because AIC is a relative measure of how good a model is among a candidate set of models given the data, it is particularly prone to poor choices of model formulation. You can have a set of essentially meaningless variables and yet the analysis will still produce a best model. It is therefore important to assess the goodness of fit (χ^2, R^2) of the model that includes all the predictors under study. If this global model is a good fit, then you can rest assured that the best approximating model will be a good fit also. In the junco example, the R^2 of the global model is 0.49. In behavioural ecology, with noisy data sets of observational data such as this, this is an agreeably higher than average value and can be interpreted as a 'large' effect (Møller and Jennions 2002).

Ideally, however, you should not be in a situation where you are uncertain if any of your variables are informative or not. Burnham and Anderson (2001, 2002, 2004) emphasise the importance of justifying the inclusion of each parameter, and each model within the candidate set. Justification for selection of parameters, or models, is a separate topic (see Burnham et al. 2010; Dochtermann and Jenkins 2010) but should derive from prior biological knowledge or analysis (e.g. experiments, field observations, previously published analyses). For example, Symonds and Johnson (2008), in predicting community composition in Australian birds, used variables that had been identified as significant environmental predictors of species richness in these communities by Hawkins et al. (2005).

We earlier mentioned the importance of not using correlated traits as independent predictor variables. Likewise, if the data do not conform to the assumptions of the statistical procedures (e.g. using non-normally distributed data in least-squares regressions), then calculation of AIC_c and Akaike weights can become unreliable (Richards 2005). Additionally, AIC has not been well tested in relation to more complex model formulations involving random effects and non-linear or polynomial terms, so it is still an open question as to whether AIC performs well when comparing such models. For simplicity's sake, our discussion has focused on simple linear regression models, primarily of observational data; however, AIC can also be

applied to analyses of mechanistic models of behaviour and controlled experiments (see, e.g. Lau and Glimcher 2005; Luttbeg et al. 2009; Richards et al. 2010), where the problems of data dredging are mitigated. A couple of good examples of the use of AIC in an experimental framework in behavioural ecology are provided in the review by Garamszegi et al. (2009).

Although it is not our intention here to cover in detail the criticisms of the IT-AIC approach, one aspect that has been identified as a weakness is that, despite explicitly taking into account the number of predictors, AIC still tends to favour overly complex models (Kass and Raftery 1995; Link and Barker 2006). This maybe particularly true if none of the models are good fits for the data. In terms of selecting a best approximating model, this can lead to over-fitting. Recent simulation studies also suggest that full-model averaging can help to reduce the problems caused by the model selection bias towards over-complex (and indeed under-complex) models (Lukacs et al. 2009).

Researchers using AIC sometimes report *p* values for their models in addition to AIC values and use this as a basis for further inference. Reviewers have different levels of tolerance for this practice, but ultimately there is inconsistency in doing this because of the philosophical difference (see Mundry 2010; Burnham et al. 2010) between the two approaches. Nevertheless, we believe that in certain cases there is a benefit in reporting results from both the NHST- and AIC-based approaches in that it provides comparison. However, the onus is very much on the authors to argue which approach inference should be based on. Where model selection uncertainty is clearly apparent, we believe that it would become evident that multimodel inference using AIC provides the greater capacity for sound inference.

Although we have suggested taking AIC values from the output of statistic packages, it pays to check exactly how the package has calculated them (in Table 1, we provide some indications). The appropriate version of AIC for an analysis may not necessarily be the one given by the package. For example, it may provide only the raw AIC values, and not AIC_c, which is recommended for analyses with smaller sample sizes.

On a final practical note, another common mistake is to use different data sets for the different models that are being compared. This may sound obvious but can easily and absent-mindedly occur if there are missing data for some parameters (M. Symonds, *personal experience*). In such cases, AIC cannot be compared between models. The safest approach to address this issue is to delete incomplete records. Unfortunately, this can result in reduced power of the analysis (and loss of a lot of hard work). There are ways around this problem, recently outlined by Nakagawa and Freckleton (2008, 2010), including model-based augmentation and multiple imputation methods.

Acknowledgments This paper originates from a presentation by the first author at the 12th International Behavioral Ecology Congress at Cornell University in the post-conference symposium 'Advances in statistical philosophy and experimental design in behavioral ecology' organised by László Garamszegi and Shinichi Nakagawa. We thank László Garamszegi for inviting us to contribute to this issue. We are indebted to Gonçalo Cardoso for allowing us to use and reanalyse his junco data for the example in this paper. We thank six anonymous reviewers for their time and efforts. The members of the Animal Behaviour and Evolution group at the University of Melbourne also provided useful feedback on the manuscript. MRES was financially supported by the Australian Research Council.

References

Ackerman JT, Eadie JM (2003) Current versus future reproduction: an experimental test of parental decisions using nest desertion by mallards (*Anas platyrhynchos*). Behav Ecol Sociobiol 54:264–273

Akaike H (1973) Information theory as an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) Second international symposium on information theory. Budapest, Akaemiai Kiado, pp 267–281

Bolker BM (2008) Ecological models and data in R. Princeton University Press, Princeton

Buckland ST, Burnham KP, Augustin NH (1997) Model selection: an integral part of inference. Biometrics 53:603–618

Burnham KP, Anderson DR (2001) Kullback-Leibler information as a basis for strong inference in ecological studies. Wildlife Res 28:111–119

Burnham KP, Anderson DR (2002) Model selection and multimodel inference, 2nd edn. Springer, New York

Burnham KP, Anderson DR (2004) Multimodel inference: understanding AIC and BIC in model selection. Sociol Methods Res 33:261–304

Burnham KP, Anderson DR, Huyvaert KP (2010) AICc model selection in the ecological and behavioral sciences: some background, observations and comparisons. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1029-6

Cardoso GC, Atwell JW, Ketterson ED, Price TD (2007) Inferring performance in the songs of dark-eyed juncos (*Junco hyemalis*). Behav Ecol 18:1051–1057

Dochtermann NA, Jenkins SH (2010) Developing multiple hypotheses in behavioral ecology. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1039-4

Garamszegi LZ (2010) Information-theoretic approaches to statistical analysis in behavioural ecology: an introduction. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1028-7

Garamszegi LZ, Calhim S, Dochtermann N, Hegyi G, Hurd PL, Jørgensen C, Katsukake N, Lajeunesse MJ, Pollard KA, Schielzeth H, Symonds MRE, Nakagawa S (2009) Changing philosophies and tools for statistical inferences in behavioral ecology. Behav Ecol 20:1363–1375

Guthery FS, Brennan LA, Peterson MJ, Lusk JJ (2005) Information theory in wildlife science: critique and viewpoint. J Wildl Manag 69:457–465

Hawkins BA, Diniz-Filho JAF, Soeller SA (2005) Water links the historical and contemporary components of the Australian bird diversity gradient. J Biogeog 32:1035–1042

Hegyi G, Garamszegi LZ (2010) Using information theory as a substitute for stepwise regression in ecology and behavior. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1036-7

Hilborn R, Mangel M (1997) The ecological detective: confronting models with data. Princeton University Press, Princeton

Johnson JB, Omland KS (2004) Model selection in ecology and evolution. Trends Ecol Evol 19:101–108

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773–795

Lau B, Glimcher PW (2005) Dynamic response-by-response models of matching behavior in rhesus monkeys. J Exp Anal Behav 84:555–579

Link WA, Barker RJ (2006) Model weights and the foundations of multimodel inference. Ecology 87:2626–2635

Lukacs PM, Burnham KP, Anderson DR (2009) Model selection bias and Freedman's paradox. Ann Inst Stat Math 62:117–125

Luttbeg B, Hammond JI, Sih A (2009) Dragonfly larvae and tadpole frog space use games in varied light conditions. Behav Ecol 20:13–21

Mazerolle MJ (2006) Improving data analysis in herpetology: using Akaike's Information Criterion (AIC) to assess the strength of biological hypotheses. Amphibia-Reptilia 27:169–180

Møller AP, Jennions MD (2002) How much variance can be explained by ecologists and evolutionary biologists? Oecologia 132:492–500

Mundry R (2010) Issues in information theory based statistical inference—a commentary from a frequentist's perspective. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1040-y

Nakagawa S, Freckleton RP (2008) Missing inaction: the dangers of ignoring missing data. Trends Ecol Evol 23:592–596

Nakagawa S, Freckleton RP (2010) Model averaging, missing data and multiple imputation: a case study for behavioural ecology. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1044-7

Richards SA (2005) Testing ecological theory using the information-theoretic approach: examples and cautionary results. Ecology 86:2805–2814

Richards SA (2008) Dealing with overdispersed count data in applied ecology. J Appl Ecol 45:218–227

Richards SA, Whittingham MJ, Stephens PA (2010) Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1035-8

Stephens PA, Buskirk SW, Hayward GD, Martínez del Rio C (2005) Information theory and hypothesis testing: a call for pluralism. J Appl Ecol 42:4–12

Stephens PA, Buskirk SW, Martínez del Rio C (2007) Inference in ecology and evolution. Trends Ecol Evol 22:192–197

Symonds MRE, Johnson CN (2008) Species richness and evenness in Australian birds. Am Nat 171:480–490

Thorup K, Alerstam T, Hake M, Kjellén N (2006) Traveling or stopping of migrating birds in relation to wind: an illustration for the osprey. Behav Ecol 17:497–502

Towner MC, Luttbeg B (2007) Alternative statistical approaches to the use of data as evidence for hypotheses in human behavioral ecology. Evol Anthropol 16:107–118

Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP (2006) Why do we still use stepwise modeling in ecology and behaviour? J Anim Ecol 75:1182–1189