AIC and relatives

Basic definitions and formulae

K = number of parameters estimated from the data, n = sample size, $\mathcal{L}$ = maximised likelihood.

Akaike's Information Criterion, AIC:

$$AIC = -2\log(\mathcal{L}) + 2K$$

The model with the lowest AIC should give the best predictions when applied to a fresh data set; it gives approximately the same results as cross-validation. That we are actually evaluating our model based on the same data set means that we tend to be over-optimistic about our predictive ability, and the second term tries to correct for this "optimism".

AICc: AIC is an approximation which works fine if n is large, much larger than K. A better approximation is this:

$$AICc = AIC + \frac{2K(K+1)}{n-K-1}$$

The extra term is known as the "small sample correction", but AICc is better than AIC even for large samples. For some models it's not clear what is n: for an occupancy model, should this be the number of sites or the total number of visits to sites?

AICc can also be expressed as:

$$AICc = -2\log(\mathcal{L}) + \frac{2Kn}{n-K-1}$$

QAIC and QAICc: Binomial and Poisson distributions do not have a separate parameter for spread ("dispersion"); this is theoretically determined by other parameters. In practice, the spread is often greater than it should be, and an overdispersion parameter (usually named $\hat{c}$, "c-hat", by biologists) is estimated and used to correct variances and confidence intervals. The correction must also be applied to AIC, resulting in a Quasi-AIC:

$$QAIC = \frac{-2\log(\mathcal{L})}{\hat{c}} + 2K$$

where the number of parameters, K, must include $\hat{c}$. And there's a small-sample-correction version too:

$$QAICc = QAIC + \frac{2K(K+1)}{n-K-1}$$

BIC: The Bayesian or Schwarz Information Criterion is:

$$BIC = -2\log(\mathcal{L}) + \log(n)\,K$$

This implies a greater penalty for extra parameters (unless n < 8), and the BIC-lowest models are simpler than AIC-lowest models. There's nothing Bayesian about the calculation of BIC, but it does tend to give the same results as model selection using Bayes factors.

CAIC: You may also come across something called the "Consistent AIC" (see Bozdogan (1987), Anderson et al (1998)), which is:

$$CAIC = -2\log(\mathcal{L}) + (\log(n) + 1)\,K$$

Values derived from information criteria

In this section we'll refer to AIC, but the exact same calculations can be done based on any of the other information criteria.

Delta AIC, Δ

The values of AIC are not important, it's the differences which matter: $\Delta_i = AIC_i - AIC_{min}$. Two points to bear in mind:

- Adding a totally uninformative parameter to a model (eg, rolling a die), will increase AIC by < 2 units. If an extra parameter is providing useful information, the model will have a *lower* AIC than the model without it. Such models are not supported (B&A p131, Arnold 2010).
- If Δ is small, there is uncertainty about which model is best (if you remember the "which bag" activity, one white stone changes the AIC by 0.2). If the difference in AIC is < 2, it's not clear which model is the best, but with a difference of > 10, the less-good model can be discarded (B&A p70).

Model likelihood

You can think of the model likelihood as an adjusted version of the maximised likelihood, $\mathcal{L}$, which has been adjusted to allow for the number of parameters in the model and standardised so that the best model has likelihood = 1. It is calculated by "undoing" the -2log() operation, so:

$$model\ likelihood = e^{-\frac{1}{2}\Delta_i}$$

You can use ratios of model likelihoods to compare two models; the comparison does not depend on which other models are in the set.

Model weights or Akaike weights

Model weights are in the same proportions as the model likelihoods, but all the weights in the set add up to 1. So: model weight = model likelihood / sum of likelihoods of all the models in the set.

Model weights *do* depend on all the other models in the set. Beware of model redundancy, ie, having 2 models in the set which are equivalent, even if they appear to be parameterised differently (redundant models will have identical values of $\mathcal{L}$).

The model weight can be regarded as the probability that the model is actually the best predictor in the set (B&A pp75-77). This is a subjective (ie, Bayesian) probability, and is the posterior probability of being the best predicting model based on a "savvy" prior which favours models with fewer parameters.

Model averaging

Often there is uncertainty as to which model to use for prediction: Δ is small or model weights are similar. The strategy then is to calculate predictions from each of the plausible models and produce weighted averages of the predictions, using model weights.

Do not try to obtain model-averaged values of the parameters, as they will have different interpretations in different models.

An example

TABLE 1. *AKAIKE Information Criteria corrected for small sample size (AICc) and Akaike weights (W_i) for models ψ* dot *and ψ* habitat *and overall model averaged estimates for site occupancy (ψ) and detection probability (over one sampling period [P] and over entire study [P-hat]) for banteng, gaur, dhole, and leopard at camera traps in Mondulkiri Protected Forest, Cambodia.*

	AICc	W_i	ψ (SE)	P (SE)	P-hat
Psi (.)	255.94	0.7301			
Psi (habitat)	257.93	0.2699			
Banteng			0.51 ± 0.09	0.29 ± 0.04	0.90
Psi (habitat)	85.2	0.96			
Psi (.)	92.1	0.04			
Gaur			0.40 ± 0.08	0.08 ± 0.02	0.42
Psi (habitat)	152.26	0.88			
Psi (.)	156.3	0.12			
Dhole			0.49 ± 0.17	0.13 ± 0.04	0.60
Psi (.)	383.34	0.52			
Psi (habitat)	383.5	0.48			
Leopard			0.70 ± 0.08	0.42 ± 0.03	0.97

The table right is from Gray (2012) and shows the results of occupancy modelling of camera trap data for large mammals in Cambodia.

For each of the four species, he fitted models with and without a habitat covariate for occupancy. He reports AICc and model weights for each, and gives estimates and standard errors of probability of occupancy (ψ) and probability of detection (p) after model averaging.

Bayesian model-selection criteria

In Bayesian analysis, we don't calculate a maximised likelihood, $\mathcal{L}$, when estimating parameters. Instead we can use the "posterior predictive density", based on the probability of observing the data given the posterior distributions of the parameters. To this is added a term representing our "optimism". The calculations are rather involved, but there are 3 options, all of which work like AIC, ie, lowest is best.

- DIC, deviance information criterion, included in the output of many packages based on BUGS or JAGS; not valid for hierarchical models, which include most of our ecological examples (occupancy, mark-recapture, etc).
- WAIC, widely-applicable information criterion or Watanabe-Akaike information criterion; valid for hierarchical models provided observations are independent; cannot be used for models with spatial or temporal autocorrelation.
- posterior predictive loss; valid when observations are not independent.

See Hooten & Hobbs (2015) for examples and a discussion of these criteria.

References

Anderson, D.R., Burnham, K.P., & White, G.C. (1998) Comparison of Akaike information criterion and consistent Akaike information criterion for model selection and statistical inference for capture-recapture studies. *Journal of Applied Statistics*, 25, 263-282.

Arnold, T.W. (2010) Uninformative parameters and model selection using Akaike's Information Criterion. *Journal of Wildlife Management*, 74, 1175-1178.

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

B&A Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference: a practical information-theoretic approach*, 2 edn. Springer-Verlag.

Galipaud, M., Gillingham, M.A.F., David, M., & Dechaume-Moncharmont, F.-X. (2014) Ecologists overestimate the importance of predictor variables in model averaging: a plea for cautious interpretations. *Methods in Ecology and Evolution*, 5, 983-991.

Galipaud, M., Gillingham, M.A.F., & Dechaume-Moncharmont, F.-X. (2017) A farewell to the sum of Akaike weights: the benefits of alternative metrics for variable importance estimations in model selection. *Methods in Ecology and Evolution*, 8, 1558-1678.

Gray, T.N.E. (2012) Studying large mammals with imperfect detection: status and habitat preferences of wild cattle and large carnivores in eastern Cambodia. *Biotropica*, 44, 531-536.

Hooten, M.B. & Hobbs, N.T. (2015) A guide to Bayesian model selection for ecologists. *Ecological Monographs*, 85, 3-28.