Choosing models with AIC & BIC (Zwick, 7 March 7, 2013)

Models are selected from the one of the measures that OCCAM outputs for different models applied to the training set data, namely the Bayesian Information Criterion (BIC) also known as the Schwartz Criterion (Schwartz 1978). BIC is a way of linearly integrating the error of a model and its complexity (DF) which differs from the Akaike Information Criterion (AIC) (Akaike 1994) by its inclusion of a factor which depends on the sample size, N:

$$AIC = -2N \sum p \ln q + 2\,DF.$$
$$BIC = -2N \sum p \ln q + \ln(N)\,DF$$

These measure are unaffected by adding the constant $N \sum p \ln p$, which gives

$$AIC' = 2N \sum p \ln (p/q) + 2\,DF.$$
$$BIC' = 2N \sum p \ln (p/q) + \ln(N)\,DF$$

The first term of AIC and BIC is now the familiar likelihood-ratio (LR, sometimes written equivalently as L^2) Chi-square measure of a model. **Good** models have **low** values of these measures, since LR, the model **error**, is ideally small and so is DF, the model **complexity**. In OCCAM, however, AIC and BIC are given *relative to a reference model*, usually taken to be the bottom (independence) model:

$$\Delta AIC = AIC'(ref) - AIC'(model) = AIC(ref) - AIC(model) = \Delta LR + 2 * \Delta DF$$
$$\Delta BIC = BIC'(ref) - BIC'(model) = BIC(ref) - BIC(model) = \Delta LR + \ln(N) * \Delta DF$$

In this case (for reference=<u>bottom</u>), ΔAIC and ΔBIC have *high* values for good models, since ΔLR is the information *captured* in the model, and since ΔDF = df(ref) – df(model), being negative, diminishes the measure the more complex the model is. Including the ln(N) factor in ΔBIC penalizes more complex models (for large enough N). BIC is more conservative than AIC in recommending departures from the reference independence model. In my experience, models picked by ΔBIC do better on generalization (test data) than the more complex models picked by ΔAIC.

If reference = <u>top</u>, it is *also* the case that good models have *high* ΔAIC and ΔBIC, since LR(reference) = 0, so ΔLR = - LR(model), and thus negative. We want LR(model), the model error, to be as small as possible. We also want ΔDF = df(reference) – df(model) to be as big as possible, which means a simple model. Having the smallest LR(model) and the biggest ΔDF will maximize ΔAIC or ΔBIC.

Akaike, H. (1994). "Implications of Informational Point of View on the Development of Statistical Science." In *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, H. Bozdogan, ed., pp. 27-38, Kluwer Academic Publishers, the Netherlands.

Schwartz, G. (1978). *Ann. Stat.* 6, pp. 461-464.