Summary of Basic Curve Fitting and Definitions

Least squares optimization

Task: to fit a set of data given by a set of N pairs of x-values and y-values, (x_i, y_i).
Select a function, *f*, with a number of unknown parameters, *p*.
Given any set of values for the parameters, calculate the function evaluated at the data points, $\hat{y}_i = f(x_i)$.
Calculate the errors, also called residuals *r*, as the difference between fitted and actual values: $r_i = y_i - \hat{y}_i$.

The sum of squared errors, *SSE*, is given by: $SSE = \sum_{i=1}^{N} r_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$

A basic least squares fit adjusts the parameters of the function to generate different estimated until *SSE* cannot be reduced any further. However, often the procedure produces a local optimum (changing parameters a little bit does not improve things) that is a poor fit and a long way from a global optimum (the overall best set of parameters to fit the function). Good choice of initial parameters helps avoid this. In general as functions get complicated, try to start with optimal parameters from related simpler functions.

A weighted least squares fit, allows different points to contribute by different amounts to the overall error and can be used to prevent dominance of data points that are very far from the rest of the data. In this case, this sum is: $SSE' = \sum_{i=1}^{N} w_i r_i^2$, where w_i could be calculated iteratively and decrease with the previous calculation of r_i or depend on x_i or y_i.

R-square: *SSE* (usually without weighting so $w_i = 1$) can be combined with *SST*, the total sum of variance: $SST = \sum_{i=1}^{N} w_i (y_i - \bar{y})^2$ where $\bar{y}$ is the mean, to yield an R-square value, R^2, representing the fraction of Explained Variance (variance in y-values of data that is explained by the fit to the range of x-values): $R^2 = 1 - \frac{SSE}{SST}$. Clearly, better fits yield *SSE* nearer zero, thus R^2 closer to 1.

Comparing models with different numbers of parameters

Since more and more parameters can always make the fit better, one wants to stop before “fitting noise”. Thus, when comparing different functions, *f*, with different numbers *p* of parameters, one should pick the model with the largest “adjusted-R-square” equal to $1 - \frac{SSE}{SST} \cdot \frac{N-1}{N-p-1}$. This essentially requires that to keep an extra parameter the fitting function must do better than precisely fit one extra point.

Aikake Information Criterion (AIC)

An alternative method for model comparison, with justification based on information theory, is the use of AIC (this penalizes extra parameters less severely than a similar method, Bayesian Information Criterion = BIC, and in general appears to provide a better comparison).

$$AIC = 2p - 2\ln(L)$$

where AIC is lower for a better model, p is the number of parameters and L is the likelihood of a model—namely the probability of obtaining the dataset given the model.

The likelihood, L, can be either determined by analytic calculation if the model and distribution of errors is well defined, or estimated by Monte-Carlo simulation by adding an appropriate level of noise. If the level of noise—the estimated error in each measurement combined with the expected variability given the model—is not known, then AIC is less useful, since L depends on this with extreme sensitivity. Examples where the variability/noise is precisely known, given a model, are case where observations are discrete (eg coin-tossing, dice rolling, counting numbers, DNA sequencing so there is essentially no measurement error and all variability lies in the definition of a stochastic model, say via the Binomial distribution).

In cases where models can be compared by AIC we can attribute confidence or significance of our decision to choose one model (the one with lowest AIC) over another, since the relative probability of data arising from a less optimal model, i, is:

$$\exp\left[\frac{AIC_0 - AIC_i}{2}\right]$$

where AIC_i is the AIC of the less optimal model and AIC_0 is that of the optimal model.

Corrected Aikake Information Criterion (AICc)

When the number of parameters is large or number of data points, N, is small, even AIC can over-fit the data, so a correction term should be added:

$$AIC_c = AIC + \frac{2p(p+1)}{N-p-1} = 2p\left(\frac{N}{N-p-1}\right) - 2\ln(L).$$

Bayesian Information Criterion (BIC)

The correction term to prevent overfitting using BIC is stronger than that for AIC. The BIC reads:

$$BIC = p\ln(N) - 2\ln(L).$$

Notes

1) If statistical error gives rise to measurements that are distributed as a Gaussian about some true value, then the log-likelihood is simply proportional to the sum of the squared errors and selection of the model with either largest AIC or BIC is identical to least square error minimization if the models being compared have the same number of parameters.

2) The difference between AIC and BIC arises in the way that they are derived. The goal of AIC is to select the model closest to the data. BIC assumes one of the models being compared is correct and aims to select the model that is most likely to be the correct model given the data. However, while BIC is shown to achieve this with probability 1 given infinite data, in practical cases, BIC can get it wrong more often than AIC even when one of the models is the "correct" model that produced the data.

3) My preference is to produce dummy data sets that follow the models being compared with different numbers of parameters. Then calculate AIC, BIC (also adjusted-R-square if you like) and see which analysis most often identifies the correct model and use that best analysis for the real data.