Alternative Model Comparison Methods

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 12, Hilary Term 2009

February 25, 2009

Consider two alternative models $M_1 = \{f(x;\theta), \theta \in \Theta_1\}$ and $M_2 = \{f(x;\theta), \theta \in \Theta_2\}$ for a sample $X = x$.

Without having a prior distribution we could in principle address the question of which of these are more adequate by considering the *maximized likelihood ratio*

$$\Lambda = \frac{\sup_{\Theta_1} L(\theta)}{\sup_{\Theta_2} L(\theta)} = \frac{L(\hat{\theta}_1)}{L(\hat{\theta}_2)}.$$

Note that the quantities $L(\hat{\theta}_j)$ can be considered as the *profile likelihoods* $\hat{L}_j$ of the 'model label' j, considering θ as a nuisance parameter.

Thus, this ratio is analogous to the Bayes factor, which is the ratio of the *integrated likelihoods*.

If the models are *nested* in the sense that

$$\Theta_1 \subseteq \Theta_2$$

the likelihood ratio

$$\Lambda = \frac{\sup_{\Theta_1} L(\theta)}{\sup_{\Theta_2} L(\theta)} = \frac{L(\hat{\theta}_1)}{L(\hat{\theta}_2)}$$

will always be less than or equal to 1, so will always prefer the larger model as a description for the data.

There are many reasons this is not adequate, hence Λ as above is rarely used as a measure of relative accuracy of two models and some penalty for complexity is applied, for example as in the BIC.

If the models are nested, one may in principle consider the *p-value*

$$p = P\{-2 \log \Lambda \geq -2 \log \lambda_{\text{obs}}; M_1\} \tag{1}$$

i.e. the probability that the ratio Λ is less that the observed value, assuming the simpler model is true.

If the p-value is very small, corresponding to Λ_1 being unusually small, this will be taken as evidence against M_1, and so M_2 is favoured.

In contrast, if p is moderate, M_1 would be favoured over M_2 as the simpler explanation of the data.

This approach has several problems, including:

- it does not make clear sense unless M_2 has been established as adequate

This approach has several problems, including:

- it does not make clear sense unless M_2 has been established as adequate
- it does not make sense if the models M_i are not nested

This approach has several problems, including:

- it does not make clear sense unless M_2 has been established as adequate
- it does not make sense if the models M_i are not nested
- when many models M_i are considered, it is hard to control the probability of favouring an incorrect model by chance

This approach has several problems, including:

- it does not make clear sense unless M_2 has been established as adequate
- it does not make sense if the models M_i are not nested
- when many models M_i are considered, it is hard to control the probability of favouring an incorrect model by chance
- Arbitrary thresholds for the significance level must be set and it is hard to give precise guidelines for this.

This approach has several problems, including:

- it does not make clear sense unless M_2 has been established as adequate
- it does not make sense if the models M_i are not nested
- when many models M_i are considered, it is hard to control the probability of favouring an incorrect model by chance
- Arbitrary thresholds for the significance level must be set and it is hard to give precise guidelines for this.

Nevertheless, the methodology is often used and it appears often to behave a lot better than what theory immediately suggests. Recent systematic studies of principles associated with this way seem to confirm that.

Consider the problem of choosing between including different subsets of variables in linear regression.

Consider the problem of predicting an n-dimensional vector Y with expectation μ from explanatory variables X. The total mean square prediction error would be

$$\mathbf{E}(||Y - \hat{Y}||^2) = \mathbf{E}\{||\mu - \hat{\mu}||^2\} + \mathbf{E}\{||Y - \mathbf{E}(Y)||^2\},$$

where $||v||^2 = \sum_i v_i^2$ is the squared error norm.

The second term in this expression is the intrinsic random error and we can do nothing about it. The first term is the *squared prediction risk*

$$R = \mathbf{E}\{||\mu - \hat{\mu}||^2\}$$

and we would wish to choose a model for $\mu(X)$ which makes this risk small.

If it holds that $\mu = X\beta$ and we use a linear model of the form

$$\mu_S(X) = X(S)\beta_S$$

where S is a subset of d elements of the covariates so

$$x_i(S) = (x_{ij}, j \in S)$$

we thus have the prediction risk

$$R = \mathbf{E}\{||X\beta - X(S)\hat{\beta}_S||^2\} = d\sigma^2 + B(S)$$

where $B(S)$ is a bias term

$$B(S) = ||\mu - \mu_S(X)||^2 = ||X\beta - X(S)\beta_S||^2$$

with $B(S) = 0$ if the true distribution satisfies $\beta_j = 0$ for $j \notin S$.

The corresponding residual sum of squares has expectation

$$\mathbf{E}(\mathrm{RSS}) = \mathbf{E}\{||Y - X(S)\hat{\beta}||^2\} = (n-d)\sigma^2 + B(S).$$

Thus, if we add $(2d-n)\sigma^2$ to both sides this equation, we get an unbiased estimate of the prediction risk from the residual sum of squares

$$\hat{R}(S) = \mathrm{RSS} + (2d-n)\sigma^2.$$

Mallows C_p uses now an unbiased estimate of σ^2, typically based on the residual sum of squares for the model with all the variables included, to estimate the risk so that

$$C_p = \frac{\mathrm{RSS}}{\hat{\sigma}^2} + 2d - n.$$

Choosing a model S can now be based on this criterion. Note that this also penalizes models with many parameters.

Akaike's Information Criterion (AIC) is based on exactly the same idea as C_p, but it is more general and is not restricted to regression models.

Akaike suggests assessing the prediction error by the *Kullback-Leibler distance* to the true distribution g:

$$D(g,\theta) = \int g(x)\log f(x,\theta)\,dx - \int g(x)\log g(x)\,dx = S(g,\theta) + H(g).$$

The AIC is an approximately unbiased estimate of $-2nS(g,\hat{\theta})$ which can be shown to reduce to

$$\mathrm{AIC}_i = l(\hat{\theta}_i) - d_i$$

so

$$2\Delta\mathrm{AIC} = D + 2(d_1 - d_2).$$

AIC is equivalent to Mallows C_p for linear regression models.

AIC gives typically lower penalty for complexity than BIC, in particular as $n \to \infty$, since the penalty for complexity is $2(d_2 - d_1)$ rather than $\log n(d_2 - d_1)$.

In particular the *AIC does not share the consistency property of the BIC*.