THE BAYES INFORMATION CRITERION (BIC)

1. Introduction

Suppose we have a set of models, usually not all of the same dimension, and want to decide which of them fits a data set best. For the Wilks test, recall that we had an m-dimensional model H_0 included in a d-dimensional model H_1, where $m < d$. The maximum of the likelihood over H_1 would always be at least as large, and usually larger, than over H_0 because of the inclusion. But, if the maximum likelihood over H_0 was not too much smaller than over H_1, then in the test, H_0 is not rejected.

2. Model selection and information criteria

In "model selection," there are m models $M_1, ..., M_m$, where usually $m > 2$. The models may be "nested," with inclusions $M_1 \subset M_2 \subset \cdots \subset M_m$, or they may not be. Rather than testing multiple hypotheses on the models two at a time, to see if we reject one or the other, it's convenient to have a criterion for selecting one of the models. Arbitrary levels such as 0.05 may not be appropriate. But, as in the Wilks test, we want to avoid, for example, simply choosing the model with largest (maximum) likelihood, which in the nested case would mean always choosing M_m. That could well be "overfitting." It turns out to be natural to consider maximum log likelihoods rather than likelihoods themselves. Let ML_i be the maximum likelihood over the ith model and $MLL_i = \ln(ML_i)$ the maximum log likelihood over the ith model. Let d_i be the dimension of the ith model M_i. Different "penalties" have been proposed to be subtracted from MLL_i to avoid overfitting. Perhaps the first was the AIC or "Akaike information criterion"

$$AIC_i = MLL_i - d_i$$

(Akaike, 1974). Later, G. Schwarz (1978) proposed a different penalty giving the "Bayes information criterion,"

$$BIC_i = MLL_i - \frac{1}{2} d_i \log n. \tag{1}$$

For either AIC or BIC, one would select the model with the largest value of the criterion.

Date: 18.650, Dec. 4, 2015 .

Schwarz (1978) proved that under some conditions, the BIC is *consistent*, meaning that if one of the models $M_1, ..., M_m$ is correct, so that there is a true θ_0 in that model, then as n becomes large, with probability approaching 1, BIC will select the *best* model, namely the smallest model (model of lowest dimension) containing θ_0. (Of course, if the models are nested, then for θ_0 in one model, it will also be in all the larger models.) Poskitt (1987) and Haughton (1988) extended and improved Schwarz's work, showing that consistency held also under less restrictive conditions. The AIC is not necessarily consistent in this sense, as will be shown. Although that may make the BIC seem preferable, it may be that none of the models $M_1, ..., M_m$ is actually correct, and in such a case it is not so clear which criterion, if either, is best to use.

3. Comparing information criteria with the Wilks test

Suppose we have just two models M_1 and M_2 with $M_1 \subset M_2$, and M_i has dimension d_i with $d_1 < d_2$. To fit with the assumptions of the Wilks test, suppose that there is a true $\theta = \theta_0 \in M_2$. Then M_1 is the best model if $\theta_0 \in M_1$, otherwise M_2 is. For any of three methods, the Wilks test, AIC, and BIC, given a data set, we'd evaluate the maximum log likelihoods MLL_i for $i = 1, 2$. For the Wilks test, with test statistic W defined as $-2log(\Lambda)$ where $\Lambda = ML_1/ML_2$, so $W = 2[MLL_2 - MLL_1]$, for n large enough, and some $\alpha > 0$, we would reject M_1 (and so select M_2) if $W \geq \chi^2_{1-\alpha}(d_2 - d_1)$, otherwise select M_1. If $\theta_0 \notin M_1$, so M_2 is the best model, then ML_1/ML_2 will approach 0 exponentially as $n \to \infty$, and $W \sim cn$ for some $c > 0$, so we will make the correct choice with probability $\to 1$ as $n \to \infty$. A general proof won't be given here, but it will be illustrated later in the special case of binomial probabilities.

If $\theta \in M_1$, the Wilks test will correctly select M_1 with a probability converging to $1 - \alpha$.

The AIC will select M_2 if $W > 2(d_2 - d_1)$, which if $\theta_0 \notin M_1$ will occur and give the correct choice with probability converging to 1 as $n \to \infty$. However, if $\theta_0 \in M_1$, W will converge in distribution to $\chi^2(d_2 - d_1)$ as $n \to \infty$, so the probability of incorrectly rejecting it will again not go to 0 as n becomes large (as in the Wilks test for fixed $\alpha > 0$) because

$$\Pr(W > 2k) \to \Pr(\chi^2(k) > 2k) > 0$$

for $k = d_2 - d_1$.

The BIC will select M_2 if $W > (d_2 - d_1) \log n$. If $\theta_0 \in M_1$, the probability of selecting M_2 will go to 0 as $n \to \infty$, as $(d_d - d_1) \log n$ eventually becomes larger than $\chi^2_{1-\alpha}(d_2 - d_1)$ for any $\alpha > 0$. This illustrates the consistency of BIC, that it will select a lower dimensional model when it is best. If M_2 is the best model, then BIC will select it with probability $\to 1$ as $n \to \infty$, as n becomes larger than $\log n$. So of the three criteria, BIC is the only consistent one.

4. The binomial family

Let M_2 be the binomial model where the success probability $\theta = p$ satisfies $0 < p < 1$, so $d_2 = 1$. Let M_1 be the submodel that p has a specific value p_1, so $d_1 = 0$. Suppose the model holds with a true value p_0. Let's see what happens when $p_0 \neq p_1$. If X successes are observed in n trials, with $0 < X < n$, then the likelihood function is

$$f(X, n, p) := \binom{n}{X} p^X (1-p)^{n-X}.$$

The MLE of p in M_2 is $\hat{p} = X/n$, so $ML_2 = f(X, n, \hat{p})$. We have $ML_1 = f(X, n, p_1)$, so

$$ML_1/ML_2 = (p_1/\hat{p})^X[(1-p_1)/(1-\hat{p})]^{n-X}$$

and $W =$
$2[MLL_2 - MLL_1] = 2(n-X)[\log(1-\hat{p}) - \log(1-p_1)] + 2X[\log(\hat{p}) - \log(p_1)]$. As $n \to \infty$ we will have $X \sim np_0$, $\hat{p} \to p_0$, and $n - X \sim n(1-p_0)$. For p_1 fixed and p varying, the function

$$g(p) = 2(1-p)[\log(1-p) - \log(1-p_1)] + 2p[\log(p) - \log(p_1)]$$

has derivative

$$\begin{aligned} g'(p) &= 2[-1 - \log(1-p) + \log(1-p_1) + 1 + \log(p) - \log(p_1)] \\ &= 2[-\log(1-p) + \log(1-p_1) + \log(p) - \log(p_1) \end{aligned}$$

and second derivative

$$g''(p) = 2\left[\frac{1}{1-p} + \frac{1}{p}\right] > 0,$$

so g' is increasing. We have $g'(p) = 0$ if and only if $p = p_1$, so this is a minimum of g. So $g(p_0) > g(p_1)$ and $W = 2[MLL_2 - MLL_1]$ will indeed approach $+\infty$ as cn for some $c > 0$, namely $c = g(p_0) - g(p_1)$.

5. Multiple regression

For an example, suppose we've observed some (X_j, Y_j), $j = 1, ..., n$, and want to consider models $M_1 \subset M_2 \subset \cdots \subset M_m$ where in M_i,

$$Y_j = P_\theta(X_j) + \varepsilon_j := \theta_0 + \sum_{r=1}^{i} \theta_r f_r(X_j) + \varepsilon_j, \tag{2}$$

ε_j are i.i.d. $N(0, \sigma^2)$, and f_r are some functions.

5.1. **Polynomial regression.** $f_r(x) = x^r$ for each r and x.

Let $f_r(x) := x^r$. Then for a given i, P_θ is a polynomial of degree at most i. For $i = 1$ we'd have ordinary simple y-on-x regression, for $i = 2$ quadratic regression, and so on. In the ith model we'll have $i + 1$ parameters θ_r, with a parameter vector $\theta = (\theta_0, \theta_1, ..., \theta_i)$, where

$$P(x) = P_\theta(x) \equiv \sum_{r=0}^{i} \theta_r x^r. \tag{3}$$

5.1.1. *Interpolation and overfitting.* Suppose all the X_j are distinct. Then there exists a polynomial P of degree $n-1$ such that $P(X_j) = Y_j$ for all $j = 1, ..., n$. To see this, for each $i = 1, ..., n$ the polynomial $P_i(x) = \prod_{j \neq i}(x - X_j)$ is 0 at X_j if and only if $j \neq i$. There is a constant c_i such that $c_i P_i(X_i) = y_i$. Then $P := \sum_{i=1}^{n} c_i P_i$ is of degree $n - 1$ and satisfies $P(X_j) = Y_j$ for all $j = 1, ..., n$ as stated.

For polynomials of degree $n-1$ restricted to $\{X_1, ..., X_n\}$, the P_i are linearly independent. They form a basis, as we have just seen. So the polynomial P just constructed is unique.

In doing polynomial regression of degree i, it will be assumed that $n > i + 1$ to avoid being able to fit the values Y_j exactly. It's actually desirable that n be substantially larger than $i+1$ so as not to "overfit" the data. This is advisable for multiple regression more generally.

5.2. **Residual sums of squares.**

Assuming that X_j are fixed design points, the only random variables are the ε_j, and the likelihood function will be, for P_θ as in (3) or more generally (2),

$$\begin{aligned} f(V, \theta) &= (2\pi\sigma^2)^{-n/2} \prod_{j=1}^{n} \exp\left(-\frac{(Y_j - P_\theta(X_j))^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{j=1}^{n} \frac{(Y_j - P_\theta(X_j))^2}{2\sigma^2}\right) \end{aligned}$$

here $V = \{(X_j, Y_j)\}_{j=1}^n$. To maximize the likelihood with respect to θ for any fixed $\sigma > 0$ is equivalent to minimizing $\sum_{j=1}^n (Y_j - P_\theta(X_j))^2$ (least squares). Let RSS$_i$, the "Residual sum of squares," be the sum so minimized (it's the sum of squares of the regression residuals) for the ith model. Then to find the MLE of σ, we need to maximize $\sigma^{-n} \exp(-RSS_i/(2\sigma^2))$. It's equivalent to maximize the logarithm $-n \log(\sigma) - RSS_i/(2\sigma^2)$ with respect to σ. Because $n > i + 1$ by assumption, $RSS_i > 0$ with probability 1. The expression goes to $-\infty$ as $\sigma \to +\infty$ or as $\sigma \downarrow 0$, using $RSS_i > 0$, as $-n \log(\sigma)$ goes to $+\infty$ but relatively slowly. So to find an interior maximum, we take the derivative with respect to $\sigma > 0$ and set it equal to 0, giving

$$0 = -\frac{n}{\sigma} + \frac{RSS_i}{\sigma^3}, \quad \widehat{\sigma}^2 = \frac{RSS_i}{n}.$$

Then we have $ML_i = (2\pi\hat{\sigma}^2)^{-n/2} \exp(-n/2)$ and

$$MLL_i = -\frac{n}{2} \log(RSS_i) + C_n$$

where C_n is a term depending only on n, not i, and so irrelevant to the comparison of models by either AIC or BIC.

6. Bayesian rationale of the BIC

When we have a set of models including two, neither of which is included in the other, then the Wilks test would no longer apply. Both the AIC and BIC can apply. For the BIC there is a Bayesian rationale. It is asymptotically (as $n \to \infty$) equivalent to choosing the model with highest posterior probability of being the best model, under some not too restrictive conditions. Namely, each model M_i has prior probability $\pi_i > 0$, where $\sum_{i=1}^m \pi_i = 1$, and on each, there is a prior density g_i such that $g_i(\theta) > 0$ and g_i is continuous at each $\theta \in M_i$. The prior density will be with respect to some measure $dA_i(\theta)$, which will be simply $d\theta_1 \cdots d\theta_{d_i}$ if M_i is an open subset of d_i-dimensional Euclidean space, but more often can be viewed as a measure of "area" or "volume" in the possibly curved d_i-dimensional set (manifold) M_i. We will have $\int_{M_i} g_i(\theta) dA_i(\theta) = 1$. The choice of A_i is not too crucial, as for any continuous function $h_i > 0$ defined on M_i one can multiply g_i by h_i while dividing dA_i by it, preserving the ith prior probability for a subset $B \subset M_i$,

$$\pi_i(B) = \int_B g_i(\theta) dA_i(\theta).$$

For each M_i there will also be a likelihood function $f_i(x, \theta)$ defined for $\theta \in M_i$ and each possible observation x. We then have for a vector $X = (X_1, ..., X_n)$ of observations, as usual, $f_i(X, \theta) = \prod_{j=1}^{n} f_i(X_j, \theta)$.

It will be seen that for large n, posterior densities become approximately normal, with mean at the maximum likelihood estimate and a covariance matrix asymptotic to C/n for some matrix C. Let's start with:

Example. Let the binomial parameter p have a $U[0,1]$ prior density. Suppose that the true, unknown value p_0 of p satisfies $0 < p_0 < 1$. In n independent trials, let there be X successes and so $n - X$ failures. The likelihood function is proportional to $p^X(1-p)^{n-X}$ and so that the posterior distribution of p is $\text{Beta}(X+1, n-X+1)$. In the handoxut "Order statistics, quantiles and sample quantiles," Proposition 2, it was shown that if Y_k has a $\text{Beta}(k+1, k+1)$ distribution, then the distribution of $\sqrt{k}(Y_k - \frac{1}{2})$ converges as $k \to \infty$ to $N(0, 1/8)$. Now let's see why we get asymptotic normality also for $X \neq n - X$ if X and $n - X$ are both large, as they will be with high probability for n large since $0 < p_0 < 1$. The $\text{Beta}(X+1, n-X+1)$ density (or equivalently the likelihood) is maximized at $p = \hat{p} = X/n$. Let $\hat{q} = 1 - \hat{p}$. Letting $u = p - \hat{p}$, the likelihood function becomes, omitting an $\binom{n}{X}$ factor not depending on p,

$$(\hat{p}+u)^{n\hat{p}}(\hat{q}-u)^{n\hat{q}} = \hat{p}^{n\hat{p}}\hat{q}^{n\hat{q}}\left(1+\frac{u}{\hat{p}}\right)^{n\hat{p}}\left(1-\frac{u}{\hat{q}}\right)^{n\hat{q}}.$$

Let $ML = \hat{p}^{n\hat{p}}\hat{q}^{n\hat{q}}$ be the maximum of the likelihood and MLL its logarithm (to base e as usual). Then using the Taylor series of the logarithm around 1, the log of the likelihood becomes

$$\begin{aligned}
MLL &+ \log\left[\left(1+\frac{u}{\hat{p}}\right)^{n\hat{p}}\left(1-\frac{u}{\hat{q}}\right)^{n\hat{q}}\right] \\
= MLL &+ n\hat{p}\left(\frac{u}{\hat{p}} - \frac{u^2}{2\hat{p}^2} + \cdots\right) + n\hat{q}\left(-\frac{u}{\hat{q}} - \frac{u^2}{2\hat{q}^2} + \cdots\right) \\
= MLL &- \frac{nu^2}{2\hat{p}} - \frac{nu^2}{2\hat{q}} + O(nu^3) \\
= MLL &- \frac{n^2u^2}{2}\left[\frac{1}{X} + \frac{1}{n-X}\right] + O(nu^3) \\
= MLL &- \frac{u^2}{2}\left(\frac{n^3}{X(n-X)}\right) + O(nu^3).
\end{aligned}$$

This implies that the posterior distribution is asymptotically

$N(\hat{p}, X(n-X)/n^3) = N(\hat{p}, \hat{p}\hat{q}/n)$. Recall that the (non-Bayesian) asymptotic distribution of the MLE $\hat{p}$ is $N(p_0, p_0(1-p_0)/n)$ which is approximately the same, as $\hat{p} \to p_0$ and $\hat{q} \to 1 - p_0$ as $n \to \infty$.

Asymptotic normality of the posterior density in the general case of a parameter $\theta = (\theta_1, ..., \theta_d)$ of dimension d will just be sketched. The log likelihood is

$$LL(X, \theta) = \sum_{j=1}^{n} \log f(X_j, \theta).$$

This is maximized at the MLE $\widehat{\theta} = (\widehat{\theta}_1, ..., \widehat{\theta}_d)$ of θ. Taking a d-dimensional Taylor expansion of $LL(X, \theta)$ around $\widehat{\theta}$, the constant term is MLL, the maximum of the log likelihood. The first order terms are 0 because at a maximum of a smooth function, the gradient is 0. Thus through second order, the Taylor expansion is

$$LL(X, \theta) = MLL + \frac{1}{2} \sum_{i,k=1}^{d} \left[\frac{\partial^2}{\partial\theta_i \partial\theta_k} \sum_{j=1}^{n} \log f(X_j, \widehat{\theta}) \right] (\theta_i - \widehat{\theta}_i)(\theta_k - \widehat{\theta}_k).$$

If there is a true θ_0 in a model being considered, then $\widehat{\theta}$ for that model will be converging to it as n becomes large. By the law of large numbers, as $n \to \infty$, $\frac{1}{n} \sum_{j=1}^{n} \log f(X_j, \theta_0)$ will converge as $n \to \infty$ to $E_{\theta_0} \log f(X_1, \theta_0)$, and likewise for the second partial derivatives of $\log f(x, \theta)$. For the matrix $K(\theta_0)$ of expected second partial derivatives at θ_0, which must be symmetric and negative definite since we are at (or near) a maximum, the positive definite matrix $I(\theta_0) = -K(\theta_0)$ is called the *Fisher information matrix*. To get from the log of the density of a normal distribution to its covariance matrix, we need to take the inverse of a matrix (similarly as in one dimension, the exponent in the density is $-(x-\mu)^2/(2\sigma^2)$ with the variance σ^2 in the denominator), the posterior distribution will be asymptotically $N(\widehat{\theta}, I(\widehat{\theta})^{-1}/n)$. Suppose to simplify that the matrix $I(\theta_0)$ is diagonalized in the given coordinates $(\theta_1, ..., \theta_d)$ with jth diagonal entry $1/\sigma_j^2$ for $j = 1, ..., d$, so that $I(\theta_0)^{-1}$ will also be diagonal, with jth diagonal entry σ_j^2, and $I(\widehat{\theta})^{-1}$ will be approximately the same.

In the case of multiple models M_i for BIC, the posterior densities will not be normalized individually. Rather, the posterior probability $\pi_i(X)$ that M_i is the best model, given X, will be

$$\pi_i(X) = I_i / \sum_{k=1}^{m} I_k \quad \text{where} \quad I_i := \pi_i \int_{M_i} g_i(\theta) f_i(X, \theta) dA_i(\theta)$$

for each $i = 1, ..., m$. (The total posterior probability $\pi_X(M_i)$ would be $\sum\{\pi_j(X) : M_j \subset M_i\}$, which is not what we want.) Finding i to maximize $\pi_i(X)$ is equivalent to finding it to maximize I_i. The integral in I_i is concentrated around $\widehat{\theta}_i$, the MLE of θ in M_i, for large n, and is asymptotic to

$$g_i(\widehat{\theta}_i)ML_i(2\pi)^{d_i/2}n^{-d_i/2}\prod_{j=1}^{d_i}\sigma_j.$$

To maximize this with respect to i, the dominant factor(s) for large n are given by $ML_i n^{-d_i/2}$. To maximize this with respect to i is equivalent to maximizing its log, which is

$$MLL_i - \frac{d_i}{2}\log n,$$

equaling the BIC criterion (1). This is more or less how G. Schwarz arrived at the BIC in his paper.

Notes. The Fisher information matrix and its inverse are well known objects in mathematical statistics. For example, they occur in sections 3.7 and 3.8 of the OCW notes for 18.466, Mathematical Statistics, 2003.

REFERENCES

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Auto. Control* **19**, 716-723.

Haughton, D. (1988). On the choice of a model to fit data from an exponential family, *Ann. Statist.* **16**, 342-355.

Poskitt, D. S. (1987), Complexity and Bayesian model determination, *J. Royal. Statist. Soc. Ser. B* **49**, 199-208.

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Statist.* **6**, 461-464.