嘉南藥理科技大學專題研究計畫成果報告

Discriminating among nonnested normal linear models

計畫類別：☑個別型計畫　　□整合型計畫

計畫編號：90-HA-10

執行期間：90 年 1 月 1 日至 90 年 12 月 31 日

計畫主持人：陳青浩

共同主持人：

計畫參與人員：

執行單位：嘉南藥理科技大學 醫務管理系

中華民國 90 年 12 月 31 日

嘉南藥理科技大學九十年度教師專題研究計畫成果報告

Discriminating among nonnested normal linear models

計畫編號：90-HA-10
執行期限：90 年 1 月 1 日至 90 年 12 月 31 日
主持人：陳青浩　嘉南藥理科技大學 醫務管理系

Abstract

Empirical researchers are often confronted with the problem of making choice among alternative statistical models, the choice of a model selection criterion will depend on the amount of *a priori* information and plans for future of the models. The problem of choice among nonnested models was originally discussed by Hotelling(1940), he derived a test for the selection, later, Cox(1962) discussed the selection from the classical point of view, a considerable amount of work in the area of choice among alternative models has been done recently.

The purpose of this study will be to list the available material related to the choice among the nonnested normal linear models. The criteria will be classified and their selection rules will be described and introduced. Section 1 covers specialized methods for comparing linear models, Fequentist and Bayesian methods for general models are treated in Section 2 and 3, respectively, the use of a predictive density for model selection is covered in Section 4.

Keywords： nonnested, linear model, fequentist, Bayesian, predictive density.

Results

Section 1. Methods for linear models

The Coefficient of multiple determination R^2 is the proportion of the total variation about the mean in the data explained by the model

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SS(model)}{SS(total)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1.1)$$

where $\bar{y} = \sum \frac{y_i}{n}$. Furthermore $R^2 = 1 - \frac{RSS}{SS(total)}$, where RSS=SS(total)-SS(model) is the residual sum of squares. Thus selecting the model with the largest R^2 is equivalent to selecting the model with the smallest residual sum of squares RSS.

The numerator of R^2 tends to increase when the dimension is increases causing a higher dimensional model to have a larger R^2 than a lower dimension model, in particular for the nested models. To lessen this effect, the R^2 statistic is often adjusted to explicitly incorporate the dimension. The adjusted coefficient of multiple

determination $\bar{R}^2$ is given by

$$\bar{R}^2 = 1 - \frac{RSS/(n-p)}{SS(total)/(n-1)} = 1-(1-R^2)\frac{(n-1)}{(n-p)} \quad \text{..... (1.2)}$$

It is clear that if SS(total) and n are fixed, then $\bar{R}^2$ increase if and only if the mean squared error $MSE = \frac{RSS}{(n-p)}$ decreases. Thus selecting the model with the largest $\bar{R}^2$ is equivalent to selecting the model with the smallest mean squared error MSE.

Mallows(1973) proposed a model selection criterion based on an estimate of the total mean squared error of the n fitted values scaled by the true error variance. In a plot of C_p against p, the C_p values for those models with small bias should be near the line $C_p = p$. The C_p value below the line $C_p = p$ may suggest a violation of the model assumptions.

Allen(1976) proposed a model selection method based on a cross-validation assessment of the predictive ability of the model. For a given component y_i of $\mathbf{y}$, let $\hat{y}_{i(i)}$ denote the predicted value of y_i when the model is fitted with y_i omitted. The squared error of prediction $(y_i - \hat{y}_{i(i)})^2$ provides a measure of the predictive ability of the model with respect to y_i. Combining these squared errors of prediction gives Allen's prediction sum of squares statistic

$$PRESS = \sum_{i=1}^{n}(y_i - \hat{y}_{i(i)})^2 \quad \text{..... (1.3)}$$

A small PRESS value indicates that the model fits well in the sense of having small prediction errors.

Section 2. Frequentist methods for general models

Several frequentist methods for choosing a model from a finite collection of candidate models $M_1, \cdots, M_k$ are described in the following. From a frequentist point of view, a model M_j consists of a specification of the sampling distribution of the observation vector $\mathbf{y}$ given a k_j dimensional model specific parameter vector θ_j. Let $f(\mathbf{y}|\theta_j, M_j)$ denote the conditional density of $\mathbf{y}$ given θ_j under model M_j. The maximum likelihood principle suggests selecting the model with largest maximized likelihood $\hat{L}_j(\mathbf{y}) = f(\mathbf{y}|\hat{\theta}_j, M_j)$, where $\hat{\theta}_j$ is the maximum likelihood estimate of θ_j under

model M_j. This approach is often expressed in the equivalent form of minimizing

$$-2\log\hat{L}_j(\mathbf{y}) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.1)$$

Choosing a model based on the maximized likelihood tends to favor higher dimensional models especially when the models are nested. Several modifications of maximum likelihood criterion of the form $-2\log\hat{L}_j(\mathbf{y})+h(n,k_j)$ have been proposed. The quantity $h(n,k_j)$ serves as a penalty to lessen the tendency of the maximum likelihood criterion to favor higher dimensional models. Two commonly used criteria of this form are Akaike's(1973) information criterion AIC

$$\mathrm{AIC}_j = -2\log\hat{L}_j(\mathbf{y})+2k_j \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.2)$$

and the Bayes information criterion BIC of Schwarz(1978) and Leonard(1982)

$$\mathrm{BIC}_j = -2\log\hat{L}_j(\mathbf{y})+k_j\log(n) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.3)$$

In addition to the BIC procedure, two other commonly used criteria are Shibata(1980) information criterion

$$\mathrm{SIC}_j = -2\log\hat{L}_j(\mathbf{y})+2n\log(n+2k_j) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.4)$$

and

$$\mathrm{HQIC}_j = -2\log\hat{L}_j(\mathbf{y})+2k_jc\log(\log(n)) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.5)$$

of Hannan and Quinn(1979), where $c>1$ is a constant.

Assuming the true model is among the competing models, Hurvich and Tsai(1989) and (1991) derived a bias corrected version of AIC for regression models. Their computations yield the corrected AIC statistic

$$\mathrm{CAIC}_j = -2\log\hat{L}_j(\mathbf{y})+\frac{2n(p_j+1)}{n-p_j-2} = \mathrm{AIC}_j+\frac{2(p_j+1)(p_j+2)}{n-p_j-2} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.6)$$

Besides the above information criteria, Cox(1961), Atkinson(1970) and Quandt(1972) provided methods of separate families of densities. Two families of densities are said to be separate when an arbitrary density in one family cannot be obtained as a limit of a sequence of densities in the other family. Cox suggested the exponential mixture model with the density proportional to

$$f_1(\mathbf{y}\mid\theta_1,M_1)^{\lambda}f_2(\mathbf{y}\mid\theta_2,M_2)^{1-\lambda} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.7)$$

as special cases of a comprehensive model. Letting λ vary between 0 and 1

produces a continuous range of competing models. Atkinson treated this approach and developed a test for departures from one model in the direction of the other in the sense of a test that $\lambda = 0$ or $\lambda = 1$. He argued that the hypothesis $\lambda = 1/2$ indicates that both models fit the data equally well or equally badly. Several of discussants of Atkinson(1970) indicates that this interpretation of $\lambda = 1/2$ may not be appropriate. Atkinson also considered this method with the λ and $1-\lambda$ replaced by arbitrary positive constant λ_1 and λ_2.

Atkinson also suggested that a convex combination model with density

$$\lambda f_1(\mathbf{y} \mid \theta_1, M_1) + (1-\lambda) f_2(\mathbf{y} \mid \theta_2, M_2) \quad \ldots\ldots\ldots\ldots\ldots\ldots (2.8)$$

might serve as a useful comprehensive model. For testing nonnested hypotheses, Quandt(1972) presented an approach based on this comprehensive model. Here λ may be interpreted as the probability that nature has chosen $f_1(\mathbf{y} \mid \theta_1, M_1)$ for generating $\mathbf{y}$. Compared to (2.7), the comprehensive model in (2.8) is easier to interpret since the comprehensive model is the mixture of two distributions. Quandt suggested forming a confidence interval for λ, and he stated that if the interval does not overlap 1, then $f_1(\mathbf{y} \mid \theta_1, M_1)$ is not appropriate. Similarly, if the confidence interval for λ does not overlap 0, then $f_2(\mathbf{y} \mid \theta_2, M_2)$ is not appropriate. If the confidence interval overlaps neither 1 nor 0, then the information is inadequate for discriminating between the two models. If the interval overlaps both 1 and 0, this indicates that both models are appropriate.

Section 3. Bayesian methods for general models

The Bayesian approach and certain approximations to the Bayesian approach are discussed in the following section. Let $\pi(M_j)$ denote the prior probability of model M_j, with $\sum \pi(M_j)=1$. Under model M_j, let $\pi(\theta_j \mid M_j)$ denote the prior density of θ_j and let $f(\mathbf{y} \mid \theta_j, M_j)$ denote the conditional density of the observation vector $\mathbf{y}$ given θ_j and M_j. The posterior density of θ_j given $\mathbf{y}$ and M_j will be denoted by $\pi(\theta_j \mid \mathbf{y}, M_j)$. The marginal or predictive density of $\mathbf{y}$ under model M_j is

$$m(\mathbf{y} \mid M_j) = \int f(\mathbf{y} \mid \theta_j, M_j) \pi(\theta_j \mid M_j) d\theta_j \quad \ldots\ldots\ldots\ldots\ldots\ldots (3.1)$$

This marginal density can be viewed as a weighted likelihood function since it is the

average of the conditional sampling densities with respect to the prior $\pi(\theta_j | M_j)$. The posterior probability of model M_j is

$$P(M_j | \mathbf{y}) = \frac{\pi(M_j) m(\mathbf{y} | M_j)}{\sum \pi(M_j) m(\mathbf{y} | M_j)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3.2)$$

the posterior model probability $P(M_j | \mathbf{y})$ can be viewed as a revision of the prior model probability $\pi(M_j)$ based on the observed data.

Bayesian model selection methods are usually based on Bayes factors or on the posterior model probabilities. For simplicity consider the comparison of two models M_1 and M_2, the Bayes factor $B(\mathbf{y})$ for model M_1 in favor of model M_2 is defined by

$$B(\mathbf{y}) = \frac{P(M_1 | \mathbf{y}) / P(M_2 | \mathbf{y})}{\pi(M_1) / \pi(M_2)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3.3)$$

The Bayes factor is a data based multiplier for adjusting the prior odds ratio to give the posterior odds ratio. A value of $B(\mathbf{y})$ greater than one indicates that the data increase the *a prio* evidence in favor of model M_1 over model M_2. The Bayes factor can also be expressed as the ratio of the predictive densities under M_1 and M_2, that is

$$B(\mathbf{y}) = \frac{m(\mathbf{y} | M_1)}{m(\mathbf{y} | M_2)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3.4)$$

Thus a value of $B(\mathbf{y})$ greater than one indicates that model M_1 fits the data better than M_2 in the sense that the predictive density evaluated at the observed data is larger under M_1 than it is under M_2. Proper prior distributions for the model specification parameters θ_j must be specified if the Bayes factor is to be used for model comparison.

If the improper prior $\pi(\theta_j | M_j)$ is deemed suitable, then $c_j \pi(\theta_j | M_j)$, where c_j is an arbitrary positive constant, should be just as suitable. Thus the Bayes factor

$$B(\mathbf{y}) = \frac{c_1 m(\mathbf{y} | M_1)}{c_2 m(\mathbf{y} | M_2)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3.5)$$

is undetermined under improper priors since the ratio c_1/c_2 is undefined.

Two common fixes for this problem are described below. The first solution is to avoid the specification of these prior distributions altogether by considering an asymptotic approximation to the Bayes factor. The second solution is to use part of

the data to update the improper priors yielding proper "prior" distributions.

The description of the large sample approximation to the Bayes factor given below is adapted from the presentation in Kass and Raftery(1995). This approach is based on the Laplace approximation which is treated in more detail in Tierney and Kadane(1986), Kass and Kadane(1989), and Barndorff-Nielson and Cox(1989).

Assume that the posterior density, which is proportional to $f(\mathbf{y} \mid \theta_j, M_j)\pi(\theta_j \mid M_j)$, is highly peaked about the posterior mode $\tilde{\theta}_j$. Kass and Rafetery argued that this assumption is usually reasonable when $f(\mathbf{y} \mid \theta_j, M_j)$ is highly peaked near its maximum likelihood estimate $\hat{\theta}_j$. Application to Laplace approximation of the Bayes factor $B(\mathbf{y})$ gives the approximation

$$\tilde{B}(\mathbf{y}) = (2\pi)^{\frac{1}{2}(k_1-k_2)} \frac{f(\mathbf{y} \mid \tilde{\theta}_1, M_1)\pi(\tilde{\theta}_1 \mid M_1)\left|\tilde{\Sigma}_1\right|^{\frac{1}{2}}}{f(\mathbf{y} \mid \tilde{\theta}_2, M_2)\pi(\tilde{\theta}_2 \mid M_2)\left|\tilde{\Sigma}_2\right|^{\frac{1}{2}}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.6)$$

For large value of n, $\tilde{\theta}_j$ will be approximately equal to the maximum likelihood estimate $\hat{\theta}_j$, and $\tilde{\Sigma}_j^{-1}$, which is the negative of the Hessian matrix of second derivatives of $\log\left[f(\mathbf{y} \mid \theta_j, M_j)\pi(\theta_j \mid M_j)\right.$ evaluated at $\tilde{\theta}_j$, will be approximately equal to the observed information matrix $\hat{\Sigma}_j^{-1}$. Substituting these values into the above expression yields the alternative approximation to $B(\mathbf{y})$ given by

$$\hat{B}(\mathbf{y}) = (2\pi)^{\frac{1}{2}(k_1-k_2)} \frac{f(\mathbf{y} \mid \hat{\theta}_1, M_1)\pi(\hat{\theta}_1 \mid M_1)\left|\hat{\Sigma}_1\right|^{\frac{1}{2}}}{f(\mathbf{y} \mid \hat{\theta}_2, M_2)\pi(\hat{\theta}_2 \mid M_2)\left|\hat{\Sigma}_2\right|^{\frac{1}{2}}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.7)$$

When the observations $\mathbf{y}$ form a random sample, the quantity $\log\left|\hat{\Sigma}_j\right|$ will have the same asymptotic behavior as $-k_j \log n$. Leonard(1982) suggested dropping terms that are asymptotically negligible and approximating $\log\hat{B}(\mathbf{y})$ by

$$S(\mathbf{y}) = \log\frac{\hat{L}_1(\mathbf{y})}{\hat{L}_2(\mathbf{y})} - \frac{1}{2}(k_1 - k_2)\log n \quad \ldots\ldots\ldots\ldots\ldots\ldots (3.8)$$

where $\hat{L}_j(\mathbf{y})$ denotes the maximized log likelihood under model M_j. For nested linear models, Kass and Wasserman (1995) showed that the statistic $S(\mathbf{y})$ roughly approximates the logarithm of the Bayes factor $B(\mathbf{y})$ in the sense that

$$\lim_{n\to\infty} \frac{S(\mathbf{y}) - \log B(\mathbf{y})}{\log B(\mathbf{y})} = 0 \ldots\ldots\ldots\ldots\ldots\ldots (3.9)$$

Notice that $S(\mathbf{y})$ does not depend on the priors $\pi(\hat{\theta}_j \mid M_j)$, and

$$-2S(\mathbf{y}) = BIC_1 - BIC_2 \ldots\ldots\ldots\ldots\ldots\ldots (3.10)$$

where BIC_j is the Bayes information criterion for model M_j. Hence this large sample approximation leads to the model selection procedure advocated in Schwarz(1978) and Leonard(1982). That is, to select the model M_j to minimize BIC_j.

An obvious solution to the nonexistence of the Bayes factor under improper priors is to use part of the data to update the improper priors giving proper "priors" from which a Bayes factor can be computed. Let $\mathbf{y}^t = (\mathbf{y}_1^t, \mathbf{y}_2^t)$ denote a decomposition of the n-dimensional observation vector into an m-dimensional training sample $\mathbf{y}_1$ and an (n-m)-dimensional model comparison sample. Berger and Perichhi(1996) defined a training sample $\mathbf{y}_1$ to be proper if the(3Uposteriors $\pi(\theta_j \mid \mathbf{y}_1, M_j)$ are proper for all of the models M_j under considerations. A proper training sample $\mathbf{y}_1$ is said to be minimal if no proper subset of $\mathbf{y}_1$ is proper. Once a proper training sample $\mathbf{y}_1$ is chosen, it is used to update the improper prior $\pi(\theta_j \mid M_j)$ giving

$$\pi^*(\theta_j \mid M_j) = \pi(\theta_j \mid \mathbf{y}_1, M_j) = \frac{f(\mathbf{y}_1 \mid \theta_j, M_j)\pi(\theta_j \mid M_j)}{\int f(\mathbf{y}_1 \mid \theta_j, M_j)\pi(\theta_j \mid M_j)d\theta_j} \ldots\ldots\ldots\ldots\ldots\ldots (3.11)$$

Notice that $c_j\pi(\theta_j \mid M_j)$ will yield the same value, since the constant c_j appears in the numerator and the denominator. Assuming that the integral in the denominator of

the above expression converges, this updated prior density is well defined despite c_j being unspecified. The remaining data $\mathbf{y}_2$ can then be used to compute the marginal densities

$$m(\mathbf{y}_2 \mid M_j) = \int f(\mathbf{y}_2 \mid \theta_j, M_j)\pi^*(\theta_j \mid M_j)d\theta_j \quad \text{..... (3.12)}$$

from which the Bayes factor can be computed. Berger and Pericchi defined the intrinsic Bayes factor as the average of the Bayes factors computed using all possible minimal training samples. They considered intrinsic Bayes factors based on both arithmetic and geometric averages.

When the sample size is small or the data correspond to a highly structured model, this training sample method may not be practical. Spiegelhalter and Smith (1982) proposed a sort of thought experiment to determine the ratio c_1/c_2 for comparing two nested normal linear models. They assumed that the improper prior $\pi(\beta_j, \sigma_j \mid M_j)$ has the limiting conjugate form

$$\pi(\beta_j, \sigma_j \mid M_j)\pi(\sigma_j \mid M_j) = c_j(2\pi\sigma_j^2)^{-\frac{p_j}{2}}\sigma_j^{-1} \quad \text{..... (3.13)}$$

where the c_j are undefined constants. The Bayes factor for the reduced model M_1 in favor of the full model M_2 is

$$B_{12}(\mathbf{y}) = \frac{c_1}{c_2}\left(\frac{|\mathbf{X}_2^t\mathbf{X}_2|}{|\mathbf{X}_1^t\mathbf{X}_1|}\right)^{\frac{1}{2}}\left(1+\frac{p_2-p_1}{n-p_2}F\right)^{-\frac{n}{2}} \quad \text{..... (3.14)}$$

where F is the usual F test statistic, and $\mathbf{X}_j$ denotes the full rank p_j design matrix for the model M_j. Spiegelhalter and Smith introduced the idea of an imaginary training sample of minimal size which would provide the maximum possible support to the simpler model. This sample would give a Bayes factor $B_{12}(\mathbf{y})$ just barely greater than 1. From equation (3.14), Spiegelhalter and Smith suggested that a reasonable choice for c_1/c_2 will be $\left(|\mathbf{X}_2^t\mathbf{X}_2|/|\mathbf{X}_1^t\mathbf{X}_1|\right)^{-\frac{1}{2}}$.

Aitkin(1991) suggested using the entire sample as the training sample and using the same data for model comparison. This leads to the posterior Bayes factor for model M_1 versus M_2

$$POB(\mathbf{y}) = \frac{\int f(\mathbf{y} \mid \theta_1, M_1)\pi(\theta_1 \mid \mathbf{y}, M_1)d\theta_1}{\int f(\mathbf{y} \mid \theta_2, M_2)\pi(\theta_2 \mid \mathbf{y}, M_2)d\theta_2} \quad \text{..... (3.15)}$$

O'Hagan(1995) suggested using a fractional power of the likelihood to normalize the likelihood and obtain a fractional Bayes factor. Let $b=m/n$ denote the proportion of the entire sample $\mathbf{y}$ used in the training sample $\mathbf{y}_1$. If both m and n are large, the likelihood $f(\mathbf{y}_1 \mid \theta_j, M_j)$ will approximate the full likelihood $f(\mathbf{y} \mid \theta_j, M_j)$ raised to the power b. Thus O'Hagan suggested the fractional Bayes factor

$$FB(\mathbf{y}) = q_1(b,\mathbf{y}) \Big/ q_2(b,\mathbf{y}) \quad \ldots\ldots\ldots\ldots\ldots\ldots (3.16)$$

where

$$q_j(b,\mathbf{y}) = \frac{\int f(\mathbf{y} \mid \theta_j, M_j)\pi(\theta_j \mid M_j)d\theta_j}{\int f^b(\mathbf{y} \mid \theta_j, M_j)\pi(\theta_j \mid M_j)d\theta_j} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3.17)$$

Section 4. The Predictive methods

Note that the Bayes factor for model M_1 in favor of model M_2 can be viewed as the ratio of the predictive densities of these models evaluated at the observed data. Thus the argument that leads to the Bayes factor as an appropriate quantity for model selection also suggests the use of the predictive density $m(\mathbf{y} \mid M_j)$ as a basis for the model selection. To use $m(\mathbf{y} \mid M_j)$ for model selection, the values of these predictive densities at the observed data value are compared and the model for which $m(\mathbf{y} \mid M_j)$ is largest is selected.

Laud and Ibrahim(1995) suggested a modification of this idea as described below, once the data $\mathbf{y}$ are obtained, define a predictive density for a replicate experiment PDRE as the density of a hypothetical new observation vector $\mathbf{z}$ given $\mathbf{y}$ and the model. Laud and Ibrahim define the PDRE of $\mathbf{z}$ given $\mathbf{y}$ and M_j as

$$\text{PDRE}(\mathbf{z} \mid \mathbf{y}, M_j) = \int f(\mathbf{z} \mid \theta_j, M_j)\pi(\theta_j \mid \mathbf{y}, M_j)d\theta_j \quad \ldots\ldots\ldots\ldots\ldots\ldots (4.1)$$

where $\pi(\theta_j \mid \mathbf{y}, M_j)$ is the posterior density for θ_j given $\mathbf{y}$ under model M_j. They suggested two model selection criteria based on the PDRE. The first criterion is

$$\text{PDRE}(\mathbf{y} \mid \mathbf{y}, M_j) \quad \ldots\ldots\ldots\ldots\ldots\ldots (4.2)$$

which is the PDRE for model M_j evaluated at the observed data $\mathbf{y}$. The ratio of

the $\mathrm{PDRE}(\mathbf{y} \mid \mathbf{y}, M_j)$ values for two different models is the posterior Bayes factor proposed by Aitkin (1991). The second criterion is

$$\mathrm{L_m} = E(\mathbf{z}-\mathbf{y})^t(\mathbf{z}-\mathbf{y}) = \sum_{i=1}^{n}\left[(E(Z_i)-y_i)^2 + \mathrm{var}(Z_i)\right] \quad \ldots\ldots (4.3)$$

where the expectation is with respect to $\mathbf{Z} \sim \mathrm{PDRE}(\mathbf{z} \mid \mathbf{y}, M_j)$. As indicated above $\mathrm{L_m}$ can be viewed as a combination of the squared distance from $E(\mathbf{Z})$ to $\mathbf{y}$ and the total variability $\mathrm{tr}(\mathrm{cov}(\mathbf{Z}))$. To use these criteria for model selection, the model is chosen to maximize equation(4.2) or to minimize equation(4.3).

Stone(1977) and Geisser and Eddy(1979) considered methods based on cross validation estimates of certain predictive densities. Geisser and Eddy(1979) proposed two model selection methods for the situation when $\mathbf{y} = (y_1, y_2, \cdots, y_n)^t$ represents a random sample. In this context, the density of $\mathbf{y}$ under model M_j is given by $f(\mathbf{y} \mid \theta_j, Mj) = \prod_{i=1}^{n} f(y_i \mid \theta_j, M_j)$. In stead of using the joint marginal density as the predictive density, Geisser and Eddy proposed densities based on products of certain predictive densities for the individual components of $\mathbf{y}$. Their predictive sample reuse quasi-likelihood method selects the model to maximize

$$\mathrm{PSRQL}_j(\mathbf{y}) = \prod_{i=1}^{n} f(y_i \mid \hat{\theta}_{j(i)}, M_j) \quad \ldots\ldots (4.4)$$

where $\hat{\theta}_{j(i)}$ denotes the maximum likelihood estimator of θ_j under model M_j based on the $(n-1)\times$ vector $\mathbf{y}_{(i)} = (y_1, \cdots y_{i-1}, y_{i+1}, \cdots, y_n)^t$ obtaining by removing the i^{th} observation y_i from $\mathbf{y}$.

The second method they proposed is the predictive sample reuse quasi-Bayes method which selects the model to maximize

$$\mathrm{PSRQB}_j(\mathbf{y}) = \prod_{i=1}^{n} f_p(y_i \mid \mathbf{y}_{(i)}, M_j) = \prod_{i=1}^{n} \int f(y_i \mid \hat{\theta}_{j(i)}, M_j)\pi(\theta_j \mid \mathbf{y}_{(i)}, M_j)d\theta_j \quad \ldots(4.5)$$

where $\pi(\theta_j \mid \mathbf{y}_{(i)}, M_j)$ is the posterior density of θ_j given $\mathbf{y}_{(i)}$ under model M_j. They suggested using a diffuse prior to compute the posterior densities used in expression (4.5).

Stone(1977) showed that the predictive sample reuse quasi-likelihood predicting

density $\mathrm{PSRQL}_j(\mathbf{y})$ is asymptotically equivalent to the AIC criterion. That is, as the sample size n tends to infinity, $-2\log \mathrm{PSRQL}_j(\mathbf{y})$ converges to AIC_j of equation (2.2).

Bibliography

Aitkin , M. (1991). Posterior Bayes factors. *J. Roy. Statist. Soc. Ser. B,* **53**, 111-142.

Akaike H. (1973). Information theory and an extension of the maximum likelihood principle. In. Proc. *Of the Second International Symposium on Information Theory,* B.N. Petrov and F. Czaki, eds. Akademiai Kiado, Budapest, 267-281.

Atkinson, A.C. (1970). A method of discriminating between models (with discussion). *J. Amer. Statist. Assoc.,* **32**, 323-345.

Barndorff-Nielson, O.E. and D.R. Cox(1989) .*Asympototic Techniques For Use in Statistics.* Chapman and Hall, London.

Berger, J.O. and L.R. Pericchi (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.,* **91**, 109-122.

Cox, D.R. (1962). Further results on tests of separate families of hypotheses . *J. Roy. Statist. Soc. Ser. B,* **24**, 406-424.

Cox, D.R. (1990). Role of models in statistical analysis. *Statist. Sci.* **5**, 169-174.

Drer, A.R. (1974). Hypothesis testing procedures for separate families of hypotheses. *J. Amer. Statist. Assoc.,* 69, 140-145.

Efron,B. (1984). Comparing nonested linear models. *J. Amer. Statist. Assoc.,* 79, 791-803.

Fraser D.A.S. and R.J. Gebotys (1987). Nonnested linear models: a conditional confidence approach. *Canad. J. Statist.,* **15**, 375-386.

Gelfand, A.E., D.K. Dey and H.Chang (1992) Model determination using predictive distributions with implementation via sampling-based methods. *In Bayesian Statistics 4,* J. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds. Oxford University Press, Oxford, 501-514.

Gelfand, A.E. and D.K. Dey (1994). Bayesian model choice : asymptotics and exact caculations. *J. Roy. Statist. Soc. Ser. B,* **36**, 111-147.

Geisser, S. and J. Cornfield.(1963). Posterior distributions for multivariate normal parameters. *J. Roy. Statist. Soc. Ser. B,* **25**, 368-376.

Geisser, S. (1964). Posterior odds for multivariate normal classification. *J. Roy. Statist. Soc. Ser. B,* **26**, 69-76.

Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.,* **70**,

Geisser, S. and W.F. Eddy (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.,* **74**,

153-160. corr, **75**, 765.

Hannan, E.J. and B.G. Quinn. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B,* **41**, 190-195

Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the general problem on nuisance parameters. *Ann. Math. Statist.,* **11**, 271-283.

Hurvich, C.M. and C.L.Tsai (1989). Regression and time series model selection in small samples. *Biometrika,* **76**, 297-307.

Hurvich, C.M. and C.L.Tsai (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika,* **78**, 499-509.

Kass, R.E. and A.E. Raftery (1995). Bayes factors. *J. Amer. Statist. Assoc.,* **90**, 773-795.

Kass, R.E. and L. Wasserman (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion.. *J. Amer. Statist. Assoc.,* **90**, 928-934.

Laud, P.W. and J.G. Ibrahim (1995). Predictive model selection. *J. Roy. Statist. Soc. Ser. B,* **57**, 247-262.

Lehmann, E.L. (1990). Model specification : the views of Fisher and Neyman and later developments. *Statist. Sci.,* **5**, 160-168.

Leonard, T. (1982). Comment on the paper by M. Lejeune and G.D. Faulkenberry. *J. Amer. Statist. Assoc.,* 77, 657-658.

Laud, P.W. and J.G.Ibrahim (1995). Predictive model selection. *J. Roy. Statist. Soc. Ser. B,* **57**, 247-262.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics,* **15**, 661-675.

Pauler D.K. (1995). The Schwarz criterion and related methods for model selection in linear regression. Technical Report, **634.** Carnegie Mellon University, Pittsburg.

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

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.,* **8**, 147-164.

Spiegelhalter, D.J. and A.F.M. Smith (1982). Bayes factors and choice criteria for linear models. *J. Roy. Statist. Soc. Ser. B,* **42**, 213-220.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. Roy. Statist. Soc. Ser. B, 57*, 99-138.

Stone, M.(1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc. Ser. B,* **36**, 111-147.

Stone, M.(1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser. B,* **39**, 44-47.