Statistica Sinica **7**(1997), 265-284

INFORMATION CRITERIA FOR MULTIPLE DATA SETS AND RESTRICTED PARAMETERS

Richard M. Dudley and Dominique Haughton

Massachusetts Institute of Technology and Bentley College

Abstract: In this paper, we extend information criteria for model selection to the case of K independent data sets corresponding to different true parameters $\theta_1, \ldots, \theta_K$ and to situations where some of the models may have the same dimension and may include boundaries. New criteria are introduced: SBICR, which combines criteria from different data sets, and IBICR, which treats one data set at a time. We apply the criteria to a set of 2×2 contingency tables (mosquito data) and to some data on baseball players' performance. Consistency results are given for the criteria under some assumptions. The best model will be the smallest one containing all the θ_i. A model m_j is called competitive if the vector $\theta^{(K)}$ of true parameters is in the closure of the set $m_j^{(K)}$ of $\phi^{(K)}$'s where m_j is the best model. We find that, under reasonable assumptions, for submodels of an exponential family, if for all competitive $m_j, m_j^{(K)}$ is not too thin close to $\theta^{(K)}$, the SBICR procedure is asymptotically close to Bayes procedures. This article extends results in Haughton (Ann. Statist. 1988, Sankhyā 1989) and Poskitt (J. Roy. Statist. Soc. Ser. B 1987).

Key words and phrases: BIC, Jeffreys' prior, model selection.

1. Introduction

For choosing between models m_j of different dimensions k_j, some penalties have been proposed to be subtracted from the maximum log likelihood, yielding what are called information criteria. Akaike (1974) defined a criterion AIC in which the penalty is the dimension k_j. Schwarz (1978) gave a criterion BIC for exponential families with penalty function $(k_j/2)\log n$ where n is the sample size. BIC was based on the leading terms in an asymptotic expansion of posterior probabilities. Haughton (1984), Propositions 3.4 and 3.5.1, and (1988), Proposition 2.2 and Theorem 2.3, carried the expansion to more terms, given also in Theorem 4.1(B) below, and extended the validity of the expression to smoothly curved submodels of exponential families. Poskitt (1987), Corollary 2.2, independently obtained a similar expansion for models which are open subsets of a Euclidean space, under some regularity conditions (which hold only locally for some exponential families), for data whose law may be in none of the models compared, and allowing for a general utility function.

Shibata (1976) and Hannan (1980) for ARMA processes, and Woodroofe (1982) under some general regularity conditions, showed that the AIC criterion is not consistent: in other words, as the size n of the data set increases, the probability of choosing the wrong model doesn't necessarily approach 0. The Schwarz criterion BIC is consistent, as proved for curved submodels of exponential families in Haughton (1988), Proposition 1.2 and Remark 1.2, (1989), Proposition 1; (see also Woodroofe (1982) section 8).

We now propose to extend BIC in two ways.

• First, we consider a situation in which a single model must be selected in the presence of several independent data sets, for which the true parameters may be different. In this case, applying BIC separately to each data set could lead to conflicting choices of a model. Let θ_k be the true parameter for the kth data set, $k = 1, \ldots, K$. We aim to define criteria to help us find the best model, namely the smallest model (in the sense of inclusion) containing all of $\theta_1, \ldots, \theta_K$.

• Our second extension applies to models in which some parameters are restricted. Consider for example a model m_1 with a real parameter θ, $-\infty < \theta < \infty$, two submodels of m_1, say m_2 with $\theta \geq 0$ and m_3 with $\theta \leq 0$, and a "null hypothesis" model m_4 with $\theta = 0$. Application of AIC or BIC in such a case can lead to unfortunate results. Suppose the estimate of one of the parameters is not in m_2, only by a small margin and only for one of several data sets. Then AIC and BIC will favor m_1 over m_2 since the penalty functions are the same. Yet the model m_2 is in a sense more parsimonious than m_1. We propose that there should be some penalty for removing restrictions as well as for raising the dimension. Section 2.2 treats "quartets" of models such as these and gives an application.

If there is only one data set, with one true parameter θ, then in this example it must be either in m_2 or m_3, so m_1 cannot be the best model. But suppose there are two independent data sets with parameters θ_1, θ_2. Then in the plane of possible parameters, the set where m_2 is the best model is the first quadrant, where m_3 is the best model is the third quadrant, and where m_1 is the best model is the union of the second and fourth quadrants, where θ_1 and θ_2 have different signs (Figure 1).

The extension of BIC to situations with multiple data sets and possibly restricted parameters is fairly complex. Yet such situations arise quite naturally as illustrated in our examples in Section 2. We propose two new model selection criteria, SBICR and IBICR. Let us first define SBICR:

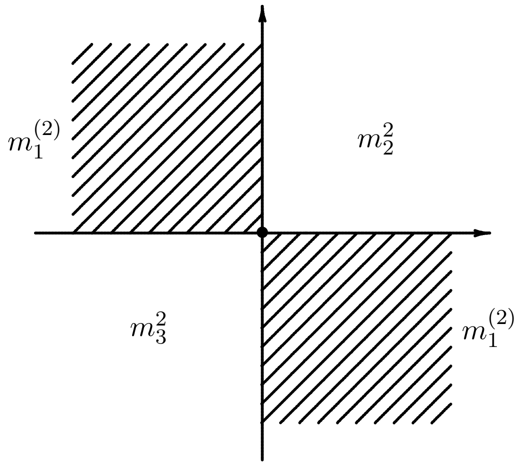

Figure 1. Projections of $m_1^{(2)}$, m_2^2, m_3^2 onto the space of means for two data sets.
Note that $m_1^{(2)}$ is not too thin near $(0,0)$. $\bullet = m_4^2$.

Let $m_1, \ldots, m_J$ be a set of models (subsets of a parameter space). Let $m_j^K := m_j \times \cdots \times m_j$ (to K factors). Let N_j be the union of all m_r^K such that $m_r \subset m_j$ strictly, and $m_j^{(K)} := m_j^K \setminus N_j$. Then m_j^K is the set where all the parameters are in the model m_j, and $m_j^{(K)}$ is the set where m_j is the best model (assumptions (3.0) and (3.1) below imply that there is a unique best model). Also, let b_j, $j = 1, \ldots, J$, be non negative numbers called bonuses. Suppose we are given K independent data sets, consisting of $n_1, \ldots, n_K$ observations. For each j, find the supremum of the overall log likelihood (for all observations) over $m_j^{(K)}$, subtract the *summed* BIC penalty $(k_j/2)\sum_{k=1}^K \log(n_k)$, and add b_j. Choose the model for which the resulting quantity is largest, or make any choice among those tied for largest. We will call this procedure SBICR or SBICR$(\{b_j\}_{j=1}^J)$. We will define bonuses based on Jeffreys priors, if finite, in (1.3) below and other bonuses in Section 2.2. Let SBIC* be SBICR with bonuses $b_j = (K/2)k_j \log(2\pi)$. For one data set, SBIC* will be called BIC* (see Haughton, Haughton, and Izenman (1990)).

We now define our second criterion IBICR. Consider as previously a set $\{b_j\}_{j=1}^J$ of bonuses. First suppose there is just one data set. For each j, after finding the supremum of the log likelihood over m_j and subtracting the BIC penalty function, the bonus b_j will be added to form a new criterion function BICR $=$ BICR$(\{b_j\}_{j=1}^J)$ for model choice. Bonuses will be called *consistent* if $b_i > b_j$ whenever $m_i \subset m_j$ strictly and the dimensions of m_i and m_j are the same. Then the criterion will favor smaller models of a given dimension, as long as the unknown parameters θ_k are estimated to be in or near the smaller model. We will always take bonuses to be used in IBICR to be consistent. Now assume we have K data sets. For the kth data set separately, apply BICR to choose a model $m_{j(k)}$. Then take the smallest model including all the $m_{j(k)}$, assuming such a model exists. We call this the *individual* BICR or IBICR procedure.

When all models are smoothly curved submodels of the natural parameter space Θ of an exponential family, with no boundary except at the boundary of Θ, and have distinct dimensions, under the assumptions of Haughton (1988), Proposition 2.2 and Theorem 2.3, and when Jeffreys priors are proper on models, we propose to specify the bonuses b_j as follows (we give an application in Section 2.1):

Suppose we are given a prior $\sum_j \alpha_j \mu_j$ where μ_j is concentrated on m_j and has a density f_j. Let $post_x(m_j)$ be the posterior probability that m_j is the best model (the smallest model containing the true parameter θ). If x is a vector of observations i.i.d. (P_θ) for some $\theta \in \Theta$, and if the true parameter θ is in m_j, we have:

$$\log post_x(m_j) = T(n,j) + A_x + O_p(n^{-1/2}) \tag{1.1}$$

(Haughton (1984, 1988), Poskitt (1987)) where A_x does not depend on j and

$$T(n,j) = \log \alpha_j + \hat{L}_j - \frac{1}{2} k_j \log(n/(2\pi)) + \log f_j(\overline{\theta}_n^j) - \frac{1}{2} \log \det\{i_{rs}(\overline{\theta}_n^j)\}, \tag{1.2}$$

$\hat{L}_j$ is the maximum log likelihood and $\overline{\theta}_n^j$ the maximum likelihood estimator on m_j, i_{rs} is the Fisher information matrix of the model m_j in a given parameterization $\eta_1, \ldots, \eta_{k_j}$ for m_j near $\overline{\theta}_n^j$, and f_j is the density of μ_j with respect to $d\eta_1 \cdots d\eta_{k_j}$. Let, now, the observations be i.i.d. P, where P need not be in the exponential family P_ϕ, $\phi \in \Theta$. Poskitt (1987) pointed out that (1.1) and (1.2) can still hold if there is a *pseudo-true value* θ_0 in the interior of m_j, where the Kullback-Leibler information $I(P_\theta, P)$ has a unique maximum over m_j at $\theta = \theta_0$ (see Sawa (1978)).

The (non-normalized) *Jeffreys measure* $M := M_j$ has density $(\det i_{rs})^{1/2}$. In any region of m_j where two parameterizations apply, M_j does not depend on the choice of parameterization (Jeffreys (1946), Kass (1989), pp. 199-200). If $0 < M(m_j) < \infty$ then $M_j/M(m_j)$ is known as the *Jeffreys prior* on m_j. Taking it as μ_j, we define a *bonus*

$$b_j := K c_j \quad \text{where} \quad c_j := c(m_j) := \frac{1}{2} k_j \log(2\pi) - \log M_j(m_j), \tag{1.3}$$

which does not depend on the observations or sample size(s). We call the resulting criterion SBICJ (J for Jeffreys); for one data set we call it BICJ.

If the α_j are all equal, $K = 1$, and the Jeffreys prior is finite and has density f_j, then $\mathrm{BICJ}(m_j) - \mathrm{BICJ}(m_i) \equiv \mathrm{BIC}(m_j) - \mathrm{BIC}(m_i) + c_j - c_i \equiv T(n, j) - T(n, i)$ for any i, j, just as desired, and likewise for multiple data sets and SBICJ (since model dimensions are distinct, SBICJ is the sum of the BICJ's for each data set). In fact, SBICR with any constant bonuses corresponds to the use of Jeffreys priors for some α_j. In this sense Jeffreys priors fit well with SBICR, despite arguments made against Jeffreys and other fixed priors on other grounds e.g. by Good (1965), pp. 45-46. Here (1.1) would apply, under the given conditions on models, for $\theta \in m_i \cap m_j$ but also more generally: in Table 1, Columns 5-6 indicate that the "$O_p(n^{-1/2})$" error in (1.1) is small in practice.

When some of the models may have the same dimension, and/or Jeffreys priors are improper, strategies for choosing bonuses will depend on the specific problem. We propose, in Section 2.2, bonuses for positive and negative halves of the real line, and give an application to baseball data.

A large part of the paper is devoted to SBICR. In Section 3 we give a list of assumptions, then state a consistency fact for SBICR under these assumptions. Then in Section 4, Theorem 4.1 shows that under further assumptions, SBICR acts approximately as a Bayes procedure. Part (E) of the theorem covers, notably, cases where a true parameter is on the boundary of a model and/or at least two models have the same dimension.

It follows from Theorem 3.2 that the IBICR is consistent even when sample sizes are so different that for some k and r, $n_k \ll \log(n_r)$, when SBICR and Bayes procedures may not be consistent as shown in Example 4.2 at the end of the paper. On the other hand, as stated in Theorem 4.1, the SBICR is closer to Bayes procedures for priors of a kind to be described in Section 4.

Theorem 4.1 relies on the new concepts of *competitive*, *fully competitive*, and *well competitive* models, defined in Section 4. A model m_j is *competitive* if the true vector of parameter vectors $\theta^{(K)} = (\theta_1, \ldots, \theta_K)$ is in the closure of the set $m_j^{(K)}$. In Theorem 4.1(A), we show that non competitive models are selected by SBICR only with exponentially decreasing probabilities as the sample sizes get large. Roughly speaking, a model is *well competitive* if $m_j^{(K)}$ is not too thin near $\theta^{(K)}$ (see Figure 1). For well competitive models, we find that the difference between SBICR criteria for two models equals the difference between the logs of their posterior probabilities (for suitable priors) plus $O_p(1)$ (see Theorem 4.1 (E)). *Fully competitive* models (defined in Section 4) are competitive models where, near $\theta^{(K)}$, m_j^K is a manifold and N_j is a union of lower dimensional sets. For fully competitive models, we get a closer agreement between SBICR and the Bayes procedure.

We assume in Section 4 that models are (possibly curved) submodels of an exponential family. This assumption is needed in the proofs of parts (A), (B) and (E) of the main Theorem 4.1.

No proofs are given in the paper, except for a brief sketch of the proof of Theorem 4.1. A longer version of the paper with complete proofs is available from either author, on-line or in hard copy.

2. Two Applications

2.1. Contingency tables: mosquito data

A study of the effectiveness of an insect electrocuting device against mosquitoes gives data (Nasci, Harris and Porter (1983), Table 4; Rasmussen (1992), pp. 161, 382) that can be arranged into five 2×2 contingency tables:

$$\begin{pmatrix} 31 & 49 \\ 94 & 90 \end{pmatrix}, \begin{pmatrix} 44 & 151 \\ 146 & 172 \end{pmatrix}, \begin{pmatrix} 129 & 30 \\ 194 & 219 \end{pmatrix}, \begin{pmatrix} 15 & 12 \\ 54 & 60 \end{pmatrix}, \begin{pmatrix} 11 & 17 \\ 39 & 21 \end{pmatrix}.$$

Each table is for a different experimental time period. The first row gives numbers of female mosquitoes electrocuted. The second gives numbers of female mosquitoes approaching human bait. The columns are for different back yards, called “Site 1” and “Site 2”, which in fact varied by design among 6 adjoining yards, so that no systematic effects are to be expected based on the labels “1” and “2”. We compare the independence (between rows and columns) model $m_2 = \mathcal{I}_2$ and the three-dimensional full multinomial model $m_1 = \mathcal{M}_4$, where each table gives a data set. Here for $\mathcal{M}_4$, $\{n_{ij}\}_{i,j=1}^2$ are multinomial $(N; p_{11}, p_{12}, p_{21}, p_{22})$, for any $p_{ij} > 0$ whose sum is 1. Let $p_{i\cdot} = p_{i1} + p_{i2}$ and $p_{\cdot i} = p_{1i} + p_{2i}$, $i = 1, 2$, and similarly for $n_{i\cdot}, n_{\cdot i}$. The independence submodel $\mathcal{I}_2$ is the 2-dimensional surface where $p_{ij} = p_{i\cdot}p_{\cdot j}$ for each i, j. Here the maximized log likelihood (for all five data sets together) on $m_j^{(5)}$ is the same as on m_j^5, so the SBICR is the sum of five BIC criteria (one for each data set) plus a bonus b_j.

Let $\mathcal{M}_m$ be the multinomial family of all laws $\{p_i\}_{i=1}^m$ on a finite set of m points with $p_i > 0$ for all i. The Jeffreys prior probability on $\mathcal{M}_m$ is (cf. Kass (1989)) a particular Dirichlet distribution $d\mathcal{J}_m := \pi^{-m/2}\Gamma(m/2)(\Pi_{i=1}^m p_i)^{-1/2} dp_1\, dp_2 \cdots dp_{m-1}$ for $p_i > 0$, $p_1 + p_2 + \cdots + p_{m-1} < 1$ where $p_m \equiv 1 - p_1 - p_2 - \cdots - p_{m-1}$ (see e.g. Johnson and Kotz (1972), Chapter 40, Sec. 5). The Jeffreys measure of $\mathcal{M}_m$ is $M(\mathcal{M}_m) = \pi^{m/2}/\Gamma(m/2)$ and so, since $\mathcal{M}_m$ has dimension $m - 1$, $c(\mathcal{M}_m) = \log \Gamma(m/2) + (m - 1)(\log 2)/2 - (\log \pi)/2$. Thus $c(\mathcal{M}_4) = (\log(8/\pi))/2$ and $M(\mathcal{I}_2) = \pi^2$ so $c(\mathcal{I}_2) = \log(2/\pi)$. Since the MLEs are simple to find for $\mathcal{M}_4$ and $\mathcal{I}_2$, the SBICJ applies easily to choosing between these models for any number of data sets. Note that in this case SBIC* and SBICJ coincide since $M(\mathcal{M}_4) = \pi^2 = M(\mathcal{I}_2)$. By the way, Jeffreys (1961), pp. 259 ff. selects between $\mathcal{M}_4$ and $\mathcal{I}_2$ for one 2×2 contingency table based on uniform, not “Jeffreys” (1946) priors.

The posterior probability of $\mathcal{I}_2$ vs. $\mathcal{M}_4$ for Jeffreys (1946) priors is $r/(r + 1)$ where $r = (N + 1)[\Pi_{i=1}^2 \Gamma(n_{i\cdot} + \frac{1}{2})\Gamma(n_{\cdot i} + \frac{1}{2})]/\{[\Pi_{i=1}^2 \Pi_{j=1}^2 \Gamma(n_{ij} + \frac{1}{2})]N!\}$, from normalization of other Dirichlet distributions (cf. Good (1976), (5.1) and (2.7) for $\phi(k)dk = d\delta_{1/2}(k)$ $(k \equiv 1/2)$). So we can illustrate how BICJ and SBICJ approximate Bayes procedures (with Jeffreys priors). We omit the details of the calculations but give results in Table 1:

Table 1. Results for mosquito data

Col.	2	3	4	5	6	7
#	X^2	p-val(χ^2)	p-val(hyp.)	$post_{\rm J}(\mathcal{I}_2)$	$post_{\rm BICJ}(\mathcal{I}_2)$	$post_{\rm BIC}(\mathcal{I}_2)$
1	3.404	0.06503	0.08643	0.5398	0.5385	0.7453
2	28.26	$1.063 \cdot 10^{-7}$	$1.072 \cdot 10^{-7}$	$3.867 \cdot 10^{-6}$	$3.856 \cdot 10^{-6}$	$9.664 \cdot 10^{-6}$
3	54.49	$1.560 \cdot 10^{-13}$	$5.097 \cdot 10^{-14}$	$2.068 \cdot 10^{-12}$	$2.061 \cdot 10^{-12}$	$5.166 \cdot 10^{-12}$
4	0.5856	0.4441	0.5817	0.7813	0.7794	0.8986
5	5.145	0.02331	0.04185	0.2255	0.2228	0.4181

Legend: # = identifying number for each 2×2 table; X^2 = chi-squared statistic; p-val(χ^2) is its p-value by the χ^2_1 distribution; p-val(hyp.) is the 2-sided p-value $2H$, H = the hypergeometric tail probability $< 1/2$; $post_{\rm J}(\mathcal{I}_2)$ = Jeffreys posterior probability of $\mathcal{I}_2$ (vs. $\mathcal{M}_4$); $post_{\rm BICJ}(\mathcal{I}_2)$ = its approximation via BICJ, calculated as $r/(r+1)$, where $r = \exp(\text{BICJ}(\mathcal{I}_2) - \text{BICJ}(\mathcal{M}_4))$; $post_{\rm BIC}(\mathcal{I}_2)$ = its approximation via BIC.

We note that: BICJ = BIC* gives a very satisfactory approximation, in terms of relative as well as absolute error, to the Jeffreys posterior probabilities (Columns 5, 6). For the second and third tables, which are far from independence, we see (1.1) working for "pseudo-true" values (Poskitt (1987)). The approximation via BIC (column 7) is too large by a factor about $(2\pi)^{1/2}$ in the odds ratio r, due to lacking the BIC* correction constant. The chi-squared p-value is not such a good approximation to the hypergeometric one (columns 3,4). The p-values of the "null hypothesis" $\mathcal{I}_2$ are substantially smaller than its posterior probabilities. This is a known phenomenon: Berger and Mortera (1991) show that often p-values are much smaller than posteriors for the null hypothesis under all priors in a large class. The BIC* correction does make the discrepancy smaller than for BIC itself. Independence is clearly rejected, for the second and third tables, by any method shown, and for the five tables, SBICJ$(\mathcal{M}_4) = -26.959$, SBICJ$(\mathcal{I}_2) = -66.165$.

The results show that $p_{11} < p_{1\cdot}p_{\cdot 1}$ for the second table and $p_{11} > p_{1\cdot}p_{\cdot 1}$ for the third. So, the labeling of varying sites as 1 or 2 (columns of the matrices) does not "explain" (and was not intended to) the directions of observed effects. (They may be explained by large differences between species of mosquitoes: Nasci et al. (1983), Table 3.)

2.2. Quartets of models: baseball data

Consider a family of distributions depending on a parameter vector (θ_1, θ_2), a model m_1 where θ_1 is unrestricted, and models m_2, m_3, and m_4 where $\theta_1 \geq 0$, $\theta_1 \leq 0$, and $\theta_1 = 0$ respectively. Given K data sets, we propose bonuses $b_j := b_{jK}$ for the quartet m_1, m_2, m_3, m_4 of models as follows: $b_{1K} = 0$, $b_{2K} = b_{3K} = K \log 2 + \log(1 + 2^{-K})$, $b_{4K} = K \log \pi$.

A rationale for the bonuses is as follows. Suppose the prior is $\frac{1}{4} \sum_{j=1}^{4} \mu_j^K$ where μ_j is a probability measure on m_j having a density f_j, $d\mu_j = f_j(\theta_1, \theta_2) d\theta_1 d\theta_2$ for $j = 1, 2, 3$, $d\mu_4 = f_4(\theta_2) d\theta_2$. Suppose that for $j = 2, 3$, $f_j(\theta_1, \theta_2) = 2f_1(\theta_1, \theta_2)$ for $\theta_1 \geq 0$ and $\theta_1 \leq 0$ respectively, where f_j is 0 for other θ_1 in each case. The prior probabilities γ_j of $m_j^{(K)}$ are then $\gamma_2 = \gamma_3 = \frac{1}{4}(1 + 2^{-K})$, $\gamma_1 = \frac{1}{4}(1 - 2^{1-K})$, $\gamma_4 = 1/4$. The prior probability densities g_j on $m_j^{(K)}$, normalized to integrate to one on $m_j^{(K)}$, are the product of f_1 over the coordinates times 2^K for $j = 2$ or 3, or times $(1 - 2^{1-K})^{-1}$ for $j = 1$. Bonuses appear to be most important in distinguishing between models on the boundary between $m_1^{(K)}$ and $m_j^{(K)}$, for $j = 2$ or 3, so that the maximum likelihood estimators for $m_1^{(K)}$ and $m_2^{(K)}$ or $m_3^{(K)}$ coincide. In that case,

$$\log \gamma_j + \log g_j - \log \gamma_1 - \log g_1 = K \log 2 + \log(1 + 2^{-K}) = b_{jK} - b_{1K}$$

as desired. It should however be noted that the higher order asymptotics of posterior probabilities need special treatment on the boundaries of models such as m_2 and m_3, but we do not deal with that problem in the present paper.

To define a suitable bonus for m_4, we take as prior on m_4 the conditional $\mu_4 = \mu_1 | \{\theta_1 = 0\}$, yielding $f_4 = f_1(0, \cdot) / \int_{-\infty}^{\infty} f_1(0, \theta_2) d\theta_2$. Then the normalized prior density g_4 on $m_4^{(K)}$ $(= m_4^K)$ equals the product of K copies of f_4. Defining as above $b_{4K} - b_{3K} := \log \gamma_4 + \log g_4 - \log \gamma_3 - \log g_3$, we get $b_{4K} - b_{3K} = -K \log \int f_1(0, \theta_2) d\theta_2 - \log(1 + 2^{-K}) - K \log 2$. Let us now take $f_1(\theta_1, \theta_2)$ to be of the form $f_1(\theta_1, \theta_2) = h_1(\theta_1) h_2(\theta_2)$ where h_1, h_2 are any proper prior densities for θ_1, θ_2. Then $b_{4K} - b_{3K} = -K \log(2h_1(0)) - \log(1 + 2^{-K})$. If we choose h_1 to be a Cauchy density, then $b_{4K} - b_{3K} = K \log(\pi/2) - \log(1 + 2^{-K})$, so $b_{4K} = K \log \pi$.

An example of such a quartet of models is as follows. Chatterjee et al. (1995) analyze data (given on a diskette) on a set of some 162 cases of major league baseball players who became free agents at the end of a season with one team, say in year "fy" (free agent year), and were hired by and played for another team the following year ("ny" or "next year"). Let S be a statistic measuring a player's performance, S_{fy} and S_{ny} its values in the given years, and $X_i := S_{ny} - S_{fy}$ for the ith case. Chatterjee et al. (1995) then model the X_i as $N(\mu, \sigma^2)$ and give alternative theories based on players' psychology favoring models m_2 $(\mu \geq 0)$ or

m_3 ($\mu \leq 0$). Let m_4 ($\mu = 0$) be the null hypothesis that the change of teams via free agency makes on average no difference in S. For multiple data sets we can also consider the model m_1 (where μ is unrestricted), which will be the best model if m_2 holds for some data sets and m_3 for others, as for the mosquito data.

In the 162 cases on the diskette, only several batting statistics are given and (thus) no pitchers are included. We re-examined the data and adjoined a new statistic "TPR" from Thorn and Palmer (1995), a large compendium of baseball data. TPR (Total Player Rating) includes offensive contributions besides hitting (getting on base by bases on balls, stealing bases) and defense (fielding). TPR ranges from -2.0 to 2.0 for average players with extremes of perhaps -5 to +8. For three cases where players had changed teams twice consecutively, so that one "ny" became the next "fy", we kept the first but not the second overlapping pair of years. We also dropped 9 cases where, according to Thorn and Palmer (1995), the player had not in fact changed teams between the years with data on the diskette, and/or the data years were not consecutive and the player had played in the intervening year. Chatterjee et al. (1995) omitted from their analysis cases where fy or $ny = 1981$, a strike-shortened season, but both batting average (BA) and TPR compare players to others in the same season, so we included the 1981 cases.

Pietrusza (1995) says that for $fy = 1985$, 1986 and 1987, most free agents wanted by their current teams received no offers from competing teams. We thus separated the data into two sets, one for these years and another for $1976 \leq fy \leq 1984$, and omitted data for the year $fy = 1988$. The number of times at bat in a season ranged from 4 and 6 for two players (who had no hits, giving outlier batting averages .000) to over 600 for another. Thus players' skills were observed with very different variances for different players. We adjusted for this as follows. For the ith case in the kth data set, $k = 1, 2$, let abf_{ki} (resp. abn_{ki}) be the number of times at bat in fy (resp. ny). Let $\tau_{ki} = (abf_{ki}^{-1} + abn_{ki}^{-1})^{1/2}$. Then we assume that S_{ki}, which equals $S_{ny} - S_{fy}$ for $S =$ BA or TPR, for the (k, i) case, are independent $N(\mu_k, \sigma_k^2 \tau_{ki}^2)$ for some μ_k and σ_k^2 (which also depend of course on which statistic, BA or TPR, we consider). Let $\kappa_k := (n_k / \sum_{i=1}^{n_k} \tau_{ki}^{-2})^{1/2}$. The least-squares and maximum likelihood estimate of μ_k is $\hat{\mu}_k = \kappa_k^2 \sum_{i=1}^{n_k} S_{ki}/(n_k \tau_{ki}^2)$. So for each k, $X_{ki} := \hat{\mu}_k + \kappa_k(S_{ki} - \hat{\mu}_k)/\tau_{ki}$ are approximately i.i.d. $N(\mu_k, \kappa_k^2 \sigma_k^2)$. We applied SBICR, IBICR, and a 1-sample t-test for each k, to the variables X_{ki}.

Chatterjee et al. (1995), p. 103 found that (without normalization) batting average was significantly lower, by .011, in ny than in fy. Table 2 gives our results, for the normalized X_{ki}. $\bar{Y}$, the sample mean of the X_{ki} for a fixed k, was used in forming the t-statistic. The differences $|\bar{Y} - \hat{\mu}_k|$ were small: less than 0.0028 for the first three data sets and 0.016 for TPR, $85 \leq fy \leq 87$. For batting average, $85 \leq fy \leq 87$, BICR and the t-test chose m_3: free agents on average did worse in ny than in fy, an effect apparently due more to team management decisions than to players' psychology; in the other three cases BICR chose m_4, saying that players did equally well before and after becoming free agents. Thus IBICR chooses m_3 for batting average and m_4 for TPR. SBICR selects m_4 for both statistics.

Table 2. Change in batting average or TPR with free agency

	n	$\bar{Y}$	t	p-value	(S)BICR(m_j) − (S)BICR(m_4)			IBICR
					$j=1$	$j=2$	$j=3$	
BA: $fy < 85$	110	0.001	0.153	0.879		-2.385	-2.396	
BA: $fy = 85, 86, 87$	28	-0.021	-3.186	0.004		-1.712	2.756	
SBICR, IBICR					-1.826	-4.685	-0.229	m_3
TPR: $fy < 85$	110	-0.082	-0.534	0.595		-2.396	-2.253	
TPR: $fy = 85, 86, 87$	28	-0.479	-1.530	0.138		-1.712	-0.549	
SBICR, IBICR					-5.142	-4.696	-3.389	m_4

Legend: TPR = total player rating (Thorn and Palmer (1995)), BA = batting average, fy = last two digits of free agency year, $\bar{Y}$ = sample mean of normalized (see text) change X_{ki} in statistic (BA or TPR) from before to after free agency. All other tabulated statistics are also based on X_{ki}. (S)BICR = BICR in BA and TPR rows, =SBICR in "SBICR, IBICR" rows.

The following is in response to a referee's comment. Suppose that for each player and year we treat the times at bat as a data set, consisting of a "1" for each hit and a "0" for each out. We would then have $2 \cdot (110+28) = 276$ data sets, of sample sizes abf_{ki}, abn_{ki}, $k = 1$, $i = 1, \ldots, 110$; $k = 2$, $i = 1, \ldots, 28$, with an i.i.d. assumption only within each of the 276. The theory would need an extension (which is possible) to cover hypotheses relating parameters for different data sets (with the same k, i: same player in two different years). Let $\overline{abf}_k$ (resp. $\overline{abn}_k$) for $k = 1, 2$ be the sample means of the numbers abf_{ki} (resp. abn_{ki}). These four numbers, and the corresponding sample medians, are all between 250 and 350. On average, then, the sample sizes are large enough, *but* two individual sample sizes $abn_{1i} = 6$, $abn_{2j} = 4$ for some i and j would be too small for asymptotics. So, we have preferred the above variance stabilization approach.

3. Consistency

In this section, we introduce assumptions for consistency of SBICR and IBICR. We then state Theorem 3.1 (on consistency of SBICR) and Theorem 3.2 (on consistency of IBICR). Let $(X, \mathcal{S})$ be a measurable space and Θ a parameter space, so that for each $\phi \in \Theta$, a probability measure P_ϕ is defined on $(X, \mathcal{S})$. Let m_j, $j = 1, \ldots, J$, be subsets of Θ, to be called models, with $m_i \neq m_j$

for $i \neq j$. We assume that on each m_j a measure μ_j is defined (e.g., a prior), finite on compact sets and strictly positive on non-empty open sets. Suppose that for each $k = 1, \ldots, K$, we have a parameter $\theta_k \in \Theta$ and i.i.d. observations $X_1^k, \ldots, X_{n_k}^k$ with distribution P_{θ_k}. We assume that

(3.0) For some (unknown) j, all the θ_k are in m_j;
(3.1) For any i, j, $m_i \cap m_j$ is empty or $m_i \cap m_j = m_r$ for some $r = 1, \ldots, J$.

Since there are finitely many models, (3.0) and (3.1) imply that there is a best model: a unique smallest model containing all of $\theta_1, \ldots, \theta_K$.

Schwarz (1978) writes about the "true" model, which apparently means the best model in our sense, the smallest one containing the true parameter. Likewise, the "model that is *a posteriori* most probable" (Schwarz (1978)) or has highest posterior probability (Haughton (1988)) really means the model having the highest posterior probability of being the best or true model, as opposed to the posterior probability of the model as a subset of the parameter space, which can include other models. In past work, these other models had lower dimensions; here, they may have the same dimension.

In this paper, we consider only the simple *loss function* which is 0 when the best model is chosen and 1 otherwise. We assume that each P_ϕ has a density $f(y, \phi)$ with respect to some σ-finite measure ν on $(X, \mathcal{S})$. Let A be a subset of Θ. Let $\overline{f}(A, n) := \overline{f}(x, A, n) := \sup\{\prod_{i=1}^n f(x_i, \phi) : \phi \in A\}$, for $x = (x_1, \ldots, x_n)$, the maximum likelihood over A for n observations $x_1, \ldots, x_n$.

In what follows, when we say θ is the true parameter, we mean that the probabilities are taken under P_θ^n, in other words for $X_1, \ldots, X_n$ i.i.d. P_θ. Below, as noted, there may be different true parameters for different data sets. We next have an assumption in probability, followed by a strong (almost sure) form:

(3.2) If θ is the true parameter, and $\theta \in m_j$ for a model m_j, then for any neighborhood U of θ, if A is the complement of U, then for some $c > 0$, $\frac{1}{n}\log(\overline{f}(A, n)) < \text{ess.sup}_{v \in U \cap m_j} E \log(f(\cdot, v)) - c$, with probability converging to 1 as $n \to \infty$, where the essential supremum is with respect to μ_j.
(3.2$'$) Assumption (3.2) still holds if "with probability converging to 1 as $n \to \infty$" is replaced by "almost surely for n large enough."

Assumption (3.2$'$) holds for exponential families (e.g. Haughton (1988), proof of Proposition 1.2, and Haughton (1989), proof of Proposition 1, Case 1).

For any v, as is well known, we have $E_\theta(\log(f(\cdot, v)/f(\cdot, \theta))) \leq 0$ since $\log x \leq x - 1$, $x \geq 0$, and so $E_\theta \log f(\cdot, v) \leq E_\theta \log f(\cdot, \theta)$. Thus by the law of large numbers, if (3.2) holds, $\log \overline{f}(\{\theta\}, n) > nc + \log \overline{f}(A, n)$ with probability converging to 1 as $n \to \infty$, or eventually a.s. if (3.2$'$) holds. It follows that under (3.2), if θ is in a model m_1 and not in the closure of a model m_2, then as $n \to \infty$, $\log(\overline{f}(m_2, n)/\overline{f}(m_1, n)) < -cn$ with probability converging to 1.

Two alternate, related assumptions are:
(3.3) If θ is the true parameter then as $n \to \infty$, $\log(\overline{f}(\Theta, n)/\overline{f}(\{\theta\}, n)) = O_p(1)$.
(3.3$'$) For the true θ, as $n \to \infty$, $\log(\overline{f}(\Theta, n)/\overline{f}(\{\theta\}, n)) = o(\log n)$ almost surely.

Assumptions (3.3) and (3.3$'$) also hold for exponential families (e.g. Haughton (1988), proof of Proposition 1.2, and Haughton (1989), proof of Proposition 1, Case 2 respectively, with $O(\log\log n)$ in place of $o(\log n)$). Under (3.3), if θ belongs to models m_1 and m_2, then as $n \to \infty$, $\log(\overline{f}(m_2, n)/\overline{f}(m_1, n)) = O_p(1)$. Our next assumptions are:
(3.4) For each $i \neq j$, if a point in the closure of m_i is in m_j, then it is in m_i.
(3.5) If $m_j \subset m_i$ of the same dimension and m_j is the best model, no parameter θ_r is in the closure of $m_i \backslash m_j$.
(3.5$'$) If m_j is the best model, no parameter θ_r is on the boundary of m_j.
(3.6) For all r, k, we have $n_r, n_k \to \infty$ in such a way that $\log(n_r) = o(n_k)$.
(3.7) There is a largest model, including all the others.

Theorem 3.1. *For K independent data sets with true parameters $\theta_1, \ldots, \theta_K$, suppose* (3.2) *and* (3.3) *hold for $\theta = \theta_k$ for all $k = 1, \ldots, K$. Suppose that the finite set of models satisfies* (3.1) *and* (3.4)*, and that* (3.0)*,* (3.5) *and* (3.6) *hold. Then a best model m_j exists, and SBICR chooses m_j with probability converging to* 1*, for any $\{b_i\}_{i=1}^{J}$. Or if instead of* (3.2) *and* (3.3) *we assume* (3.2$'$) *and* (3.3$'$)*, then almost surely SBICR chooses the best model for all $n_1, \ldots, n_K$ large enough and such that* $\log(n_r)/n_k$ *is small enough for all r, k.*

Note. In the quartet of models $\mathbf{R}$, $\mathbf{R}^+$, $\mathbf{R}^-$, $\{0\}$ mentioned above, under assumption (3.5), if $\mathbf{R}^+$ or $\mathbf{R}^-$ is the best model, then no true parameter is 0.

The IBICR procedure, as defined in the introduction, is consistent, as follows:

Theorem 3.2. *Under the conditions of Theorem* 3.1*, without* (3.6)*, for any consistent choice of bonuses, IBICR eventually chooses the best model m_j with probability converging to* 1*. If* (3.2$'$) *and* (3.3$'$) *hold for $\theta = \theta_k$ for all $k = 1, \ldots, K$, then almost surely IBICR chooses the best model for all $n_1, \ldots, n_K$ large enough.*

4. Bayes Model Choice and Asymptotic Expansions

In this section, we introduce the assumptions and definitions to be used in Theorem 4.1, where we show that SBICR and Bayes procedures are close under reasonable conditions.

Suppose we are given a prior probability distribution μ on Θ. We assume that
(4.1) $\mu(\bigcup_{j=1}^{J} m_j) = 1$, $\mu(m_j) > 0$ for all $j = 1, \ldots, J$, and $\mu(m_i) < \mu(m_j)$ whenever $m_i \subset m_j$ with $i \neq j$.

We also assume we are given probabilities $\alpha_1, \ldots, \alpha_J$ with $\alpha_1 + \cdots + \alpha_J = 1$ and probability measures μ_j on Θ such that
(4.2) $\mu_j(m_j) = 1$ for each j and $\mu = \sum_{j=1}^J \alpha_j \mu_j$.
(Here μ_j is not in general the conditional distribution of μ given m_j since a model may overlap with or include others.) Then, we assume that the prior probability for $(\theta_1, \ldots, \theta_K)$ on Θ^K is $\sum_{j=1}^J \alpha_j \mu_j^K$ where μ_j^K is the distribution for which $\theta_1, \ldots, \theta_K$ are i.i.d. μ_j. Thus $\theta_1, \ldots, \theta_K$ can be generated by first choosing a value of j with probabilities α_j, then taking the θ_i to be i.i.d. (μ_j). Note, however, that then m_j is not necessarily the best model: it may happen that the values of $\theta_1, \ldots, \theta_K$ all belong to some smaller model m_i, although this will be unlikely for large K.

The set of pairs (m_j, μ_j) of models m_j with priors μ_j satisfying (4.2) will be called *compatible* if whenever $m_i \subset m_j$, if $k_i < k_j$, then $\mu_j(m_i) = 0$, while if $k_i = k_j$, then $\mu_j(m_i) > 0$ and μ_i is the conditional distribution of μ_j given m_i.

Now, given a probability measure Q on Θ^K, the probability for it that all the θ_k are in m_j is $Q(m_j^K)$. Recall that N_j is the union of all m_r^K such that $m_r \subset m_j$ strictly, and $m_j^{(K)} := m_j^K \setminus N_j$. Then the probability that m_j is the best model is $Q(m_j^{(K)}) = Q(m_j^K) - Q(N_j)$.

Given the observations X_i^k, $i = 1, \ldots, n_k$, i.i.d. P_{θ_k}, let $x^{(k)} := (X_1^k, \ldots, X_{n_k}^k)$, $k = 1, \ldots, K$. Let $g_k(x^{(k)}, \phi) := g_k((X_1^k, \ldots, X_{n_k}^k), \phi) := \prod_{j=1}^{n_k} f(X_j^k, \phi)$. We then have a posterior probability for $\theta^{(K)} = (\theta_1, \ldots, \theta_K)$. The probability that m_j is the best model can be evaluated more explicitly and simply when Q is a posterior distribution for a prior distribution $\sum_{i=1}^J \alpha_i \mu_i^K$ and the (m_i, μ_i) are compatible. Let $\phi^{(K)} := (\phi_1, \ldots, \phi_K)$, $x = (x^{(1)}, \ldots, x^{(K)})$, and let $h_K(x, \phi^{(K)}) = \prod_{k=1}^K g_k(x^{(k)}, \phi_k)$ be the likelihood function for a sample x of observations. Then the posterior distribution is $\nu_x := \sum_{i=1}^J \alpha_i h_K(x, \cdot)\mu_i^K / D_x$ where the denominator D_x is the total mass of the measure in the numerator, and for a function $g \geq 0$ and measure μ, $g\mu$ is the measure given by $(g\mu)(A) := \int_A g \, d\mu$ for measurable sets A. The posterior probability that m_j is the best model is $\nu_x(m_j^{(K)})$, for which all terms in the sum give 0 except for the jth and the set I_j of those i such that $m_j \subset m_i$ with $\mu_i(m_j) > 0$ and $i \neq j$. By compatibility we can also omit the i in I_j, leaving only the $i = j$ term, if we replace α_j by $\beta_j := \alpha_j + \sum\{\alpha_i[\mu_i(m_j)]^K : i \in I_j\}$, so that

$$\nu_x(m_j^{(K)}) = \beta_j \int_{m_j^{(K)}} h_K(x, \phi^{(K)}) d\mu_j^K(\phi^{(K)}) / D_x. \tag{4.3}$$

We now turn to assumptions on the structure of the models m_j as subsets of Θ. We recall the notion of C^∞ manifold imbedded in $\mathbf{R}^d$ (Spivak (1979), pp. 38, 65). Let $\mathbf{R}^d$ be a Euclidean space. A set $m \subset \mathbf{R}^d$ will be called a C^∞

manifold-with-boundary of dimension $k \leq d$ imbedded in $\mathbf{R}^d$ if, for $k \geq 1$, for each point ϕ of m, there is a neighborhood V of ϕ in m and an open set U containing 0 in $\mathbf{R}^k$ such that there is a C^∞ 1-1 function ξ from U into $\mathbf{R}^d$ with derivative $k \times d$ matrix of full rank k everywhere on U, with $\xi(0) = \phi$, such that ξ is a homeomorphism onto its range $\xi(U)$ and such that $\xi^{-1}(V)$ is either (a) U, in which case we say ϕ is in the "interior" of m, or (b) $U \cap \{x : \ x_1 \geq 0\}$, in which case we say ϕ is on the "boundary" of m (e.g. Spivak (1965), p. 113). If the boundary is empty, a manifold-with-boundary is called a *manifold.* If $k < d$, the "interior" of an imbedded manifold-with-boundary m differs from its usual topological interior in $\mathbf{R}^d$, which is empty. For example, in $\mathbf{R}^2$, $\{(x, y) : y = 0, x \leq 1\}$ is a manifold-with-boundary whose boundary is the point $(1, 0)$.

For $k = 0$, we define a manifold of dimension 0 to be a finite set. A 0-dimensional connected manifold is a single point. We define the boundary of a finite set to be empty. We will assume:
(4.4) For some d, each m_j is a C^∞ k_j-dimensional connected manifold-with-boundary imbedded in $\mathbf{R}^d$, $k_j \geq 0$.

Next, we specialize to exponential families. Here Proposition 2.2 of Haughton (1988) (see also Corollary 2.2 of Poskitt (1987)) will be extended to the case of multiple independent data sets, and to the case where a model can be included in another of the same dimension. Specifically, we will make the following assumptions.
(4.5) For $k = 1, \ldots, K$, let $x^{(k)} = (X_1^k, \ldots, X_{n_k}^k)$ where X_i^k, $i = 1, \ldots, n_k$, are i.i.d. random variables from an exponential family in standard form with densities $f(x, \theta_k) = \exp(x \cdot \theta_k - b(\theta_k))$ with respect to a finite measure ν on $\mathbf{R}^d$.

Let Θ be the natural parameter space of the exponential family. Part of what we mean by "standard form" is that the interior IntΘ is a non-empty open set in $\mathbf{R}^d$. Our next assumptions are:
(4.6) Each θ_k is in IntΘ, $k = 1, \ldots, K$.
(4.7) The probability law μ_j on m_j is absolutely continuous and has everywhere strictly positive C^∞ density f_j with respect to $d\psi_1, \ldots, d\psi_{k_j}$ for any C^∞ parameterization $(\psi_1, \ldots, \psi_{k_j})$ on an open set in m_j.

If compatibility and (4.7) hold, and $m_i \subset m_j$ of the same dimension, then for any parameterization on an open set $V \subset m_j$, f_i on $m_i \cap V$ is f_j restricted to $m_i \cap V$ and renormalized by a constant multiple. Next, we assume:
(4.8) Each model m_j is included in (or equal to) a model m_i of the same dimension which is a manifold.

We calculate the posterior probability that the model m_j is best, given n_k observations forming the kth data set, $k = 1, \ldots, K$, and given the prior probabilities as described above: by (4.3) and (4.5) it equals

$$\nu_x(m_j^{(K)}) = \beta_j \int_{m_j^{(K)}} \exp\Big[\sum_{k=1}^{K}(\sum_{i=1}^{n_k} X_i^k)\cdot\phi_k - n_k b(\phi_k)\Big] d\mu_j(\phi_1)\cdots d\mu_j(\phi_K)/D_x,$$

where the denominator D_x depends on x and $n_1,\ldots,n_K$ but not on the model m_j. Let $n_* = (n_1,\ldots,n_K)$, $S(n_*,j) := S(n_*,j,x) := \log(\nu_x(m_j^{(K)}))$. Recall that a model m_j is *competitive* if the true $\theta^{(K)} = (\theta_1,\ldots,\theta_K)$ is in the closure of $m_j^{(K)}$. Then $\theta^{(K)}$ is in the closure of m_j^K, so θ_k is in the closure of m_j for all k. Since θ_k belongs to some model by (3.0), it must belong to m_j by (3.4). So $\theta^{(K)} \in m_j^K$. For a 2×2 contingency table as in Section 2.1, if independence does not hold, then $\mathcal{M}_4$ is competitive and $\mathcal{I}_2$ is not; if independence holds, $\mathcal{M}_4$ and $\mathcal{I}_2$ are both competitive. In the baseball example in Section 2.2, if all true means are zero (for several data sets), all models m_1, m_2, m_3, m_4 are competitive. If all true means are positive (in m_2), m_2 is competitive and m_1, m_3, m_4 are not competitive.

Definition. A model m_j will be called *fully competitive* if it is competitive and in the neighborhood of $\theta^{(K)}$, m_j^K is a manifold and N_j a finite union of lower-dimensional manifolds or manifolds-with-boundary. Let $B_{\infty,2}(\theta^{(K)},r) := \{\phi : |\eta(\phi_k)-\eta(\theta_k)| \le r \text{ for } k=1,\ldots,K\}$ where $\eta(\phi_k)$ and $\eta(\theta_k)$ denote local coordinates for ϕ_k and θ_k in a parameterization near θ_k, and $|x-y|$ denotes the Euclidean distance beween x and y in $\mathbf{R}^{k_j}$. The model m_j will be called *well competitive* if $\mu_j^K(B_{\infty,2}(\theta^{(K)},r)) > 0$ for all $r > 0$ and for some $\delta > 0$ $\mu_j^K(B_{\infty,2}(\theta^{(K)},r)\cap m_j^{(K)}) \ge \delta\mu_j^K(B_{\infty,2}(\theta^{(K)},r))$ for all r small enough.

Roughly speaking, for a well competitive model m_j, in a neighborhood of $\theta^{(K)}$, $m_j^{(K)}$ will occupy a wedge or cone with strictly positive solid angle at $\theta^{(K)}$, as opposed, for example, to a case where $m_j^{(K)}$ has a sharp "thorn" at $\theta^{(K)}$ like the set $0 \le y \le x^2$ at $(0,0)$. In our examples in Section 2, all competitive models are fully competitive, except for the baseball data when some true means are zero; then all competitive models are well competitive (see Figure 1: m_1 is well competitive but not fully competitive in this case). Under a C^∞ change of parameterizations, the Euclidean metric and thus the sets $B_{\infty,2}(\theta^{(K)},r)$ will change, but we note that the "well competitive" condition is preserved, possibly with a different δ.

The best model is always competitive and will be fully competitive under (3.5′).

The next assumption will be needed for part (E) of the following Theorem 4.1:
(4.9) For some β, $1 < \beta < \infty$, $n_i/n_k \le \beta^2$ for all i,k and n_i, n_k.

For any real-valued function f of a vector ψ of m real variables $\psi_1, \dots, \psi_m$, let $D_\psi f$ be the gradient $(\partial f/\partial \psi_1, \dots, \partial f/\partial \psi_m)$. $D^2_\psi f$ will be the matrix of second derivatives $\partial^2 f/\partial \psi_i \partial \psi_j$. If f is vector-valued, $f = (f_1, \dots, f_d)$, then $D_\psi f$ will mean the $d \times m$ matrix $\{\partial f_i/\partial \psi_j\}$. The implications of the following main theorem will be discussed after its statement:

Theorem 4.1. *Let K independent data sets be given, where the kth consists of n_k observations i.i.d. P_{θ_k}, and assume* (3.0), (3.1), (3.4), (3.5), (3.6), (4.1), (4.2)*, compatibility of the models, and* (4.4) *through* (4.8) *hold. Then the best model is fully competitive and:*
(A) *If m_i is not competitive then the probability that SBICR chooses it goes to* 0 *exponentially in n_k for some k.*
(B) *If m_j is fully competitive then we have as $n_1, \dots, n_K \to \infty$:*

$$S(n_*, j) = T(n_*, j) - \log D_x + \sum_{k=1}^{K} O_p(n_k^{-1/2}), \quad \textit{where}$$

$$T(n_*, j) = \log \beta_j + \sum_{k=1}^{K} \Big[n_k \sup_{\phi \in m_j} (\overline{X}_k \cdot \phi - b(\phi)) - \frac{1}{2} k_j \log(\frac{n_k}{2\pi}) + \log f_j(\overline{\theta}^{j,k}_{n_k}) \\ - \frac{1}{2} \log \det\{i_{rs}(\overline{\theta}^{j,k}_{n_k})\} \Big],$$

$\overline{X}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_i^k$, and $\overline{\theta}^{j,k}_{n_k}$ is the point of m_j, eventually defined and unique almost surely as $n_1, \dots, n_K \to \infty$, where the function $\phi \mapsto \overline{X}_k \cdot \phi - b(\phi)$ attains its maximum, $k = 1, \dots, K$, and i_{rs} is the Fisher information matrix of the model m_j in a parameterization η for m_j at $\overline{\theta}^{j,k}_{n_k}$. Here f_j is the density of μ_j with respect to $d\eta_1, \dots, d\eta_{k_j}$. The same assertion holds if $T(n_*, j)$ is replaced by $T_L(n_*, j)$ where the information matrix i_{rs} is replaced by $-L_{rs} = -D^2_\eta F_k$, where F_k is $1/n_k$ times the kth log likelihood function, i.e. $F_k(\eta) := \overline{X}_k \cdot \phi_j(\eta) - b(\phi_j(\eta))$, and where $\phi_j(\eta)$ is the point of $m_j \subset \mathbf{R}^d$ with coordinates η. Also:

$$\mathrm{SBICR}(m_j) = T(n_*, j) + O_p(1) = T_L(n_*, j) + O_p(1).$$

(C) *If m_γ is competitive, then*

$$S(n_*, \gamma) \le T(n_*, \gamma) - \log D_x + O_p(1) = \mathrm{SBICR}(m_\gamma) - \log D_x + O_p(1) \\ = \sum_{k=1}^{K} [n_k \sup_{\phi \in m_\gamma} (\overline{X}_k \cdot \phi - b(\phi)) - \frac{1}{2} k_\gamma \log n_k] - \log D_x + O_p(1).$$

(D) *For any model m_γ and competitive model m_j,*

$$\mathrm{SBICR}(m_\gamma) \le \mathrm{SBICR}(m_j) - \frac{1}{2}(k_\gamma - k_j) \sum_k \log n_k + O_p(1).$$

If m_γ *is also competitive then* $\mathrm{SBICR}(m_\gamma) = \mathrm{SBICR}(m_j) - \frac{1}{2}(k_\gamma - k_j)\sum_k \log n_k + O_p(1)$.

(E) *For any well competitive model* m_j*, if* (4.9) *holds,*

$$S(n_*, j) = \sum_{k=1}^{K}[n_k \sup_{\phi \in m_j}(\overline{X}_k \cdot \phi - b(\phi)) - \frac{1}{2}k_j \log n_k] - \log D_x + O_p(1).$$

Also, $S(n_*, j) = \mathrm{SBICR}(m_j) - \log D_x + O_p(1)$.

Brief sketch of the proof

(A) One shows that $m_i^{(K)} \subset \bigcup_k G_k$ where the G_k are such that the overall (for all K data sets) log likelihood on G_k is less than its value at $\theta^{(K)}$ by at least $c_k n_k/2$ except on an event with probabilities going to zero exponentially in $\min(n_1, \ldots, n_K)$ (this uses a theorem of Cramér (1938) on tail probabilities as improved by Petrov (1954)). But by (3.6), the difference in penalty functions cannot favor m_i over the best model m_j by more than $c_k n_k/3$, so (A) follows.

(B) One shows that if m_j is well competitive, the posterior probability $\nu_x(m_j^{(K)})$ defined in (4.3) equals:

$$(1 - R)\beta_j \int_{m_j^K} h_K(x, \phi^{(K)})d\mu_j^K(\phi^{(K)})/D_x,$$

where R is exponentially small in $\min(n_1, \ldots, n_K)$ with probabilities going to one as $n_1, \ldots, n_K \to \infty$. The domain of integration of interest is now the Cartesian product m_j^K, so that the integral on m_j^K splits into a product of integrals on m_j. For each integral on m_j, one can then apply the results in Haughton (1988) established by a Laplace expansion method which yields the terms in (B).

(C) By (4.8), m_γ is included in a model m_i which is a manifold and so that the dimensions of m_i and m_γ are equal. One then shows that

$$\nu_x(m_j^{(K)}) \leq \beta_j C_i \int_{m_i^K} h_K(x, \phi^{(K)})d\mu_i^K(\phi^{(K)})/D_x,$$

where C_i is a constant. The integral on m_i^K is then treated as in (B).

(D) follows easily from (3.3).

(E) The inequality $\leq$ between the left and right hand sides of (E) follows easily from (C). The difficulty is to prove the other direction $\geq$. The key is to obtain for each k lower bounds on balls of radius $C/n_k^{1/2}$ for the integrand h_K in (4.3); this is done with Taylor expansions, separately for the two cases where the M.L.E. $\hat{\theta}_k$ on m_j is in the interior of m_j, and where $\hat{\theta}_k$ is on the boundary of m_j (itself a manifold of dimension $k_j - 1$). The volume condition in the definition of a well competitive model applied to the balls of radius $C/n_k^{1/2}$ yields the terms in (E).

Comments on Theorem 4.1. Now let's see why Theorem 4.1 shows that SBICR procedures are approximations to Bayes procedures for priors satisfying the given assumptions. Non-competitive models have their probabilities of being chosen, either by SBICR or Bayes procedures, going to 0 exponentially in some sample size n_k. For fully competitive models, the terms that approach ∞ in both the SBICR criterion statistic and the logarithm of the numerator of the posterior probability agree, so that the difference of the statistics is bounded in probability. Better agreement is possible for specific priors as we saw in Section 2.1.

Clearly, no procedure not depending on priors can fit with all Bayes procedures (for priors of the assumed kind) without having possible errors corresponding to constants added to the logarithms of the posterior probabilities, since the different logs of (priors and thus the) posterior probabilities themselves can differ in this way.

Under our assumptions a competitive model other than the best model must have higher dimension than the best model. Thus the $O_p(1)$ bound in Theorem 3.1(D) will be duly dominated by the difference in penalties.

For a well competitive model, under (4.9) Theorem 3.1(E) shows that the SBICR procedure gives a criterion within $O_p(1)$ of deciding on the basis of the log of the posterior probability that a model is best.

Example 4.1. For models which are not well competitive, the SBICR procedure may not be very close to a Bayes procedure: let $m_1 = \{(x, y) : \ y \geq 0\}$ and $m_2 := \{(x, y) : y \geq x^4\}$, with $m_3 = \{(0, 0)\}$, and let $(0, 0)$ be the true parameter θ. Then m_1 and m_2 are both competitive but not fully so. For any K, $m_1^{(K)}$ is a rather thin set near ((0,0),...,(0,0)), so that posterior probabilities of neighborhoods of $\theta^{(K)}$ will be unusually small in relation to maximum likelihood. Here m_1 is competitive but not well competitive.

Example 4.2. It will be shown why condition (3.6) is needed in Theorems 3.1 and 4.1. Let the models consist of normal laws on $\mathbf{R}^2$ with unit covariance matrix having arbitrary mean in $\mathbf{R}^2$ for the second model m_2 and mean of the form $(\mu, 0)$ for the model m_1. Let the prior probability $\mu_1 = N(0, 1)$ on the x axis and $\mu_2 = N(0, I)$ on the plane. Let $\theta_1 = (0, 0)$ and $\theta_2 = (0, 1)$. Then m_2 is the best model, and m_1 is not even competitive. If $\log n_1 \to \infty$ faster than n_2, then it can be checked directly that asymptotically the Bayes and SBICR choices of "best" model will both be m_1. In this sense the Bayes and SBICR procedures are not consistent in such a case of widely different sample sizes. Note that a pair (θ_1, θ_2) of which just one is in a model of lower dimension would have probability 0 of occurring under a prior of the assumed kind $\alpha_1\mu_1^2 + \alpha_2\mu_2^2$.

Acknowledgements

Information criteria for multiple data sets arose in work of the first author with Dr. Marc Goldring and Prof. John Lisman (cf. Goldring and Lisman (1994)). We would also like to thank Robert Kass for a helpful conversation.

References

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

Berger, J. O. and Mortera, J. (1991). Interpreting the stars in precise hypothesis testing. *Internat. Statist. Rev.* **59**, 337-353.

Cramér, H. (1938). Sur un nouveau théorème-limite de la théorie des probabilités, in *Colloque consacré à la théorie des probabilités, Actualités Sci. Indust.,* **No. 736, 5-23,** Hermann, Paris.

Chatterjee, S., Handcock, M. and Simonoff, J. (1995). *A Casebook for a First Course in Statistics and Data Analysis.* Wiley, New York.

Goldring, M. A. and Lisman, J. E. (1994). Multi-step rhodopsin inactivation schemes can account for the size variability of single photon responses in *Limulus* ventral photoreceptors. *J. Gen. Physiol.* **103**, 691-727.

Good, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* M.I.T. Press, Cambridge, MA.

Good, I. J. (1976). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist.* **4**, 1159-1189.

Hannan, E. J. (1980). The estimation of the order of an ARMA process. *Ann. Statist.* **8**, 1071-1081.

Haughton, D. (1984). On the choice of a model to fit data from an exponential family. Ph.D. thesis, Mathematics, M.I.T.

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

Haughton, D. (1989). Size of the error in the choice of a model to fit data from an exponential family. *Sankhyā A* **51**, 45-58.

Haughton, D. (1991). Consistency of a class of information criteria for model selection in non-linear regression. *Commun. Statist. Theory and Methods* **20**, 1619-1629.

Haughton, D., Haughton, J. and Izenman, A. J. (1990). Information criteria and harmonic models in times series analysis. *J. Statistical Computation and Simulation.* **35**, 187-207.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. London A* **186**, 453-461.

Jeffreys, H. (1961). *Theory of Probability*, 3d ed. Clarendon Press, Oxford.

Johnson, N. L. and Kotz, S. (1972). *Distributions in Statistics: Continuous multivariate distributions.* Wiley, New York.

Kass, Robert E. (1989). The geometry of asymptotic inference. *Statist. Sci.* **4**, 188-219 (with discussion).

Nasci, R. S., Harris, C. W. and Porter, C. K. (1983). Failure of an insect electrocuting device to reduce mosquito biting. *Mosquito News* **43**, 180-184.

Petrov, V. V. (1954). Generalization of a limit theorem of Cramér (in Russian), *Uspekhi Mat. Nauk* **9**, 195-202.

Pietrusza, D. (1995). The business of baseball. In *Thorn and Palmer* (Eds., 1995), pp. 588-599.

Poskitt, D. S. (1987). Precision, complexity and Bayesian model determination. *J. Roy. Statist. Soc. B* **49**, 199-208.

Rasmussen, S. (1992). *An Introduction to Statistics with Data Analysis.* Brooks/Cole, Pacific Grove, CA.

Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica* **46**, 1273-1291.

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

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63**, 117-126.

Spivak, M. (1965). *Calculus on Manifolds.* Benjamin, New York.

Spivak, M. (1979). *A Comprehensive Introduction to Differential Geometry* **1**, 2nd ed. Publish or Perish, Berkeley, Calif.

Thorn, J. and Palmer, P. (Eds.) (1995). *Total Baseball*, 4th ed.; *The Official Encyclopedia of Major League Baseball* (2552 pp.). Viking Penguin, New York.

Woodroofe, M. (1982). On model selection and the arc-sine laws. *Ann. Statist.* **10**, 1182-1194.

Massachusetts Institute of Technology, Department of Mathematics, 77 Massachusetts Ave., Room 2-245, MIT, Cambridge, MA 02139-4307, U.S.A.

Department of Mathematical Sciences, Bentley College, 175 Forest St., Waltham, MA 02154, U.S.A.

(Received July 1995; accepted November 1996)