The PDF of this article has been modified
from its original version.

The major changes were as follows:
- Section numbers were added.
- The author's note and reference notes were removed.
- Table 2's header was altered.

AIC and BIC

Comparisons of Assumptions and Performance

JOUNI KUHA
London School of Economics

The two most commonly used penalized model selection criteria, the Bayesian information criterion (BIC) and Akaike's information criterion (AIC), are examined and compared. Their motivations as approximations of two different target quantities are discussed, and their performance in estimating those quantities is assessed. Despite their different foundations, some similarities between the two statistics can be observed, for example, in analogous interpretations of their penalty terms. The behavior of the criteria in selecting good models for observed data is examined with simulated data and also illustrated with the analysis of two well-known data sets on social mobility. It is argued that useful information for model selection can be obtained from using AIC and BIC together, particularly from trying as far as possible to find models favored by both criteria.

Keywords: *Bayesian inference; Kullback-Leibler divergence; mobility tables; model selection; parsimony; prediction*

1. INTRODUCTION

The most common approach to the comparison and selection of statistical models is standard significance testing of nested models. Such tests are theoretically well understood and almost universally established in quantitative data analysis. However, they also have some potentially undesirable properties. For example, when sample sizes are large, significance tests are sensitive to quite small deviations from the null hypothesis, so that in very large data sets, all reasonably parsimonious models may be rejected as having a statistically significant lack of fit. Standard tests are also mostly unsuitable for comparing nonnested models and provide little guidance for choosing between models that have not been rejected. Such limitations of significance tests have stimulated interest in other approaches

SOCIOLOGICAL METHODS & RESEARCH, Vol. 33, No. 2, November 2004 188-229
DOI: 10.1177/0049124103262065
© 2004 Sage Publications

to model selection. One common class of such alternatives, the so-called penalized model selection criteria, is the main topic of this article.

Throughout this article, we consider the following generic situation of comparing models for an observed sample of data D. Here D is of size n, where n is a quantity describing the amount of information in the sample. For independent data, n is simply the number of observations, but with appropriate modifications, it can also be defined in more general cases (e.g., when the data are autocorrelated or clustered). A number of possible models M_k for D are considered, with each model having a likelihood function $p(D|\theta_k; M_k)$, which is fully specified by parameter vector θ_k with p_k parameters. Often, it is more convenient to consider the log-likelihood $l(\theta_k) = \log p(D|\theta_k; M_k)$. This is maximized by the maximum likelihood estimate (MLE) $\hat{\theta}_k$ of θ_k. The task of model comparison is then to assess which of the models are in some sense adequate for the data and which one or ones could be chosen as the basis for interpretation, prediction, or other subsequent use.

Assessment of the candidate models can be carried out as a sequence of comparisons between pairs of models, here denoted by M_1 and M_2. An important special case of this, although not required in general, is that of nested models. In that case, the parameter vector of M_2 can be partitioned as $\theta_2 = (\alpha, \psi)$ in such a way that M_1 is of the same parametric form with $\theta_1 = (\alpha, \psi_0)$, where ψ_0 is some known fixed value, often $\psi_0 = 0$. Thus, M_1 is nested in M_2, and the comparison is similar to a classical testing problem with the null hypothesis $\psi = \psi_0$ tested against an unrestricted alternative M_2. A general test for this hypothesis is provided by the likelihood ratio test statistic $2[l(\hat{\theta}_2) - l(\hat{\theta}_1)]$, which is asymptotically distributed as χ^2 with $p_2 - p_1$ degrees of freedom when M_1 is the true model.

Penalized model selection criteria are statistics of the form

$$2[l(\hat{\theta}_2) - l(\hat{\theta}_1)] - a(p_2 - p_1), \tag{1.1}$$

where a is some known positive quantity. In the nested case, the first term is equal to the likelihood ratio test statistic, but the criteria can also be applied to nonnested models. It is often convenient to choose

M_2 to be the saturated model, so that the penalized criterion becomes $G_1^2 - a \cdot df_1$, where G_1^2 is the deviance for model M_1 and df_1 its degrees of freedom. Different nonsaturated models can then be compared by comparing their values of this statistic. The operational interpretation of criterion (1.1) is that the preferred model for D is M_1 if (1.1) is negative and M_2 if it is positive, with the possible modification that values close to zero are regarded as inconclusive. In practice, of course, model choice is rarely based solely on such decision rules but depends also on, for example, purposes of the analysis and subject matter information. Nevertheless, formal criteria such as (1.1) also have a role in the selection process, and we will concentrate on them here.

Statistics such as (1.1) are known as penalized criteria because of their form as the sum of two terms of a particular kind. The first term, the difference of the maximized log-likelihoods, reflects the fit of the two models to the observed data. This tends to favor larger models. In the case of nested models, the first term is always nonnegative and increases when more parameters are added to M_2 (i.e., when $p_2 - p_1$ increases). The second term, on the other hand, then becomes smaller. If the number of parameters is regarded as a measure of the complexity of a model, $p_2 - p_1$ is the increased complexity of M_2 over M_1, and the second term in (1.1) can be interpreted as a penalty for this increase. The two terms of (1.1) thus pull in opposite directions, apparently expressing a trade-off between fit and complexity of models. A parsimonious model is favored in such a comparison unless the improvement in fit achieved by a more complex one is sufficiently large.

Penalized criteria have, in principle, several attractions. They allow comparisons of nonnested as well as nested models. The penalty for large models offsets to some extent the large-sample behavior of significance tests, where simple models are increasingly likely to be rejected when n is large. This is most apparent when a in (1.1) is an increasing function of n, so that the penalty itself is strongest in large samples. Also, most of the penalized criteria are not, despite the apparent simplicity of (1.1), ad hoc statistics but based on explicit theoretical considerations. A selection based on such a criterion can thus be regarded, at least approximately, as a choice of the best model according to some underlying definition of optimality. This can be a

strong motivation for using particular forms of (1.1), provided that the optimality criterion itself is considered compelling and that the penalized criterion is a good estimate of it.

A variety of penalized criteria, obtained from different theoretical starting points, have been proposed in the statistical literature. The alphabet of such "information criteria" (IC) now includes at least AIC, BIC, DIC (Spiegelhalter, Best, Carlin, and van der Linde 2002), EIC (Ishiguro, Sakamoto, and Kitgawa 1997), FIC (Wei 1992), GIC (Nishii 1984), NIC (Murata, Yoshizawa, and Amari 1991), SIC (Schwarz 1978), and TIC (Takeuchi 1976), as well as other related criteria that have not yet received acronyms of their own. Here we will concentrate on two of these: the Bayesian information criterion BIC, also known as Schwarz's information criterion (SIC, Schwarz 1978), and AIC, which usually stands for Akaike's information criterion (Akaike 1973). These are defined as

$$\text{AIC} = 2[l(\hat{\theta}_2) - l(\hat{\theta}_1)] - 2(p_2 - p_1) \qquad \text{and} \qquad (1.2)$$

$$\text{BIC} = 2[l(\hat{\theta}_2) - l(\hat{\theta}_1)] - \log n(p_2 - p_1). \qquad (1.3)$$

Most of the other penalized criteria mentioned above are modifications or generalizations of these two, which are also the ones most often used in practice. Of the two, BIC is more commonly used in sociology (e.g., see, in addition to associated discussions, Raftery 1986, 1995; Weakliem 1999), while AIC is very popular in, say, econometrics.

The main purpose of this article is to describe and compare AIC and BIC as model selection criteria. In Sections 2 and 3, they are first considered separately, describing the objectives, assumptions, and derivations of each criterion. Simulations of two small examples are used to illustrate their performance in meeting those objectives. The criteria are then considered together and in comparison to each other. First, in Section 4, some broad similarities between them are observed, particularly considering the interpretation of their penalty terms and the reasons why these appear to reflect a preference for parsimonious models.

In Section 5, performance of the criteria in selecting good models for observed data is examined using simulation studies. Such a comparison is not straightforward, and even its relevance could be questioned, given that the two criteria are based on different

theoretical motivations and objectives. However, in a broader sense, AIC and BIC do have the same aim, that of identifying good models, even if they differ in their exact definitions of a "good model." Comparing them is thus justified, at least to examine how each criterion performs according to the criteria of success of the other or how they behave when both should prefer the same model. We will not be able to claim that one is consistently better than the other. Instead, it is argued that once we are familiar with the criteria and their properties, comparisons between them can provide useful guidance for model selection, both when the two criteria agree and when they disagree.

The use of BIC and AIC in model selection is illustrated with an analysis of two large data sets on social mobility, introduced below and continued in Section 6. The first set of data is that compiled by Hazelrigg and Garnier (1976) (denoted hereafter as HG) and analyzed by Grusky and Hauser (1984) and subsequently by several others (e.g., Raftery 1986, 1995; Xie 1992; Weakliem 1999). The data consist of mobility tables for $n = 113,556$ men from 16 nations (denoted below by N). In each table, a man's occupation (destination class, D) and his father's occupation (origin class, O) are classified as nonmanual, manual, or farm. The second data set is that of Erikson and Goldthorpe (1992b) (denoted hereafter as EG), which consists of $n = 76,076$ men from 9 nations. The origin and destination classes are categorized using a five-class version of the class schema presented by Erikson and Goldthorpe, and the men are also classified into four 10-year birth cohorts (C).

Table 1 gives basic statistics for some models fitted to these data sets. The first parts of the table show some basic log-linear models. In analyzing comparative mobility data, it is common to consider initially two simple hypotheses about class mobility: the so-called Lipset-Zetterberg (LZ) hypothesis (Lipset and Zetterberg 1959) that all industrialized societies have the same *absolute* mobility rates and the Featherman-Jones-Hauser (FJH) hypothesis (Featherman, Jones, and Hauser 1975) that the *relative* mobility rates are "basically " the same. In Table 1, both the LZ and FJH hypotheses are first operationalized in a strict form, where the relevant mobility rates are the same across all nations and cohorts. For the LZ hypothesis, this implies that both the associations (odds ratios) between O and D and their

TABLE 1: Fitted Models for Two Sets of Data on Social Mobility

Model	G^2	df	*BIC*	*AIC*
Hazelrigg-Garnier data set (n = 113, 556):				
Basic log-linear models				
HG.1 Conditional independence: (*NO, ND*)	42,970	64	42,225	42,842
HG.2 Strict LZ: (*N, OD*)	18,392	120	16,995	18,152
HG.3 Strict FJH: (*NO, ND, OD*)	1,329	60	630	1,209
HG.4 Full cross-national variation: (*NOD*)	0	0	0	0
Extended models				
HG.5 Quasi-symmetry	150	16	−36	118
HG.6 Explanatory	490	46	−43	398
HG.7 Uniform asymmetry	49	15	−125	19
HG.8 Farm inheritance asymmetry	26	14	−137	−2
Erikson-Goldthorpe data set (n = 76, 076)				
Basic log-linear models				
EG.1 Conditional independence: (*NCO, NCD*)	22,315	576	15,661	20,983
EG.2 Strict LZ: (*NC, OD*)	23,413	840	13,972	21,733
EG.3 Strict FJH: (*NCO, NCD, OD*)	1,175	560	−5, 119	55
EG.4 Full cross-national variation: (*NCO, NCD, NOD*)	584	432	−4, 271	−280
EG.5 Full cross-cohort variation: (*NCO, NCD, COD*)	1,083	512	−4, 671	59
EG.6 Both (EG.4 + EG.5): (*NCO, NCD, NOD, COD*)	502	384	−3, 814	−266
EG.7 Saturated model (*NCOD*)	0	0	0	0
Extended models				
EG.8 Core fluidity model	1,080	513	−4, 686	54
EG.9 National variant model	944	512	−4, 810	−80

SOURCE: Sources of the models include Grusky and Hauser (1984); Weakliem (1999); Erikson and Goldthorpe (1992b; EG.8 and EG.9 modified from their modelsx).
NOTE: BIC = Bayesian information criterion; AIC = Akaike's information criterion; LZ = Lipset-Zetterberg hypothesis; FJH = Featherman-Jones-Hauser hypothesis.

marginal distributions are the same in all nations and cohorts. The strict form of the FJH hypothesis entails that the marginal distributions may vary across N and C, but the OD association remains the same. This assumption is then relaxed in the next models (HG.3-4 and EG.4-7), which allow the OD associations to vary freely between nations and/or cohorts. For the HG data, this results in the saturated model.

All deviance values (G^2) and deviance differences for the log-linear models are strongly significant, so standard tests provide little help in these large samples. The BIC and AIC values in Table 1 compare each of the models to the saturated model, so that negative values indicate models judged to be better than the saturated one. The strict LZ model is a very bad fit for both data sets, according to both AIC and BIC. This is also true for the strict FJH model for the HG data set. For the EG data, on the other hand, this model (EG.3) is the best according to BIC, but not according to AIC, which prefers model EG.4, allowing free *NOD* associations. Both criteria agree that including cross-cohort variation in mobility rates (*COD*) is not necessary.

Results after the initial models are thus somewhat inconclusive. For the HG data, no improvement over the saturated model has been achieved. For the EG data, AIC proposes a much larger model than BIC. Both findings suggest that model search should continue. This will be done in Section 6, where we return to the example and the remaining models in Table 1.

2. MOTIVATION OF BIC

Good introductions to Bayesian statistical analysis are given, for example, by Gelman, Carlin, Stern, and Rubin (1995) and Carlin and Louis (1996). In brief, Bayesian estimation of the parameters θ_k of model M_k centers on the posterior distribution $p(\theta_k|D, M_k) \propto p(D|\theta_k, M_k)\,p(\theta_k|M_k)$, where the symbol $\propto$ indicates that the two sides of the expression are equal except for a proportionality constant, which does not depend on θ_k and thus has no effect on its estimation. The quantity $p(\theta_k|M_k)$ is the density function of a prior distribution for θ_k. The prior expresses our beliefs about possible values of θ_k before obtaining the data D. The expression for the posterior density shows how the prior beliefs are updated into posterior ones in light of the observed data. All statements about θ_k are then based on the posterior distribution. For example, its mode and mean are possible point estimates of θ_k.

One of the possible Bayesian approaches to model selection is based on comparing probabilities that each of the models under consideration is the true model that generated the observed data. Prior

values $p(M_k)$ for these probabilities are again assigned. After observing D, these are updated to posterior model probabilities $p(M_k|D) \propto p(D|M_k)p(M_k)$, where

$$p(D|M_k) = \int p(D|\theta_k, M_k)p(\theta_k|M_k)\, d\theta_k \tag{2.1}$$

is the marginal likelihood of model M_k, with the integral taken over the range of all possible values for θ_k.

The proportionality constant involved in $p(M_k|D)$ can be ignored if we are only interested in relative values of the model probabilities. Comparisons are then based on ratios such as

$$\frac{p(M_2|D)}{p(M_1|D)} = \frac{p(D|M_2)}{p(D|M_1)} \frac{p(M_2)}{p(M_1)}. \tag{2.2}$$

The expression on the left-hand side of (2.2) is known as posterior odds, while the ratio of the prior probabilities is the prior odds. Prior odds are updated to posterior ones by the ratio $\mathrm{BF}_{21} = p(D|M_2)/p(D|M_1)$. Known as the *Bayes factor*, it is the central quantity of this approach to Bayesian model comparison. Often, the prior odds are taken to be 1, in which case the Bayes factor is also equal to the posterior odds.

The Bayes factor is a measure of the evidence provided by the data in favor of M_2 over M_1. It is often convenient to consider the transformation $\mathrm{T}_B = 2 \log \mathrm{BF}_{21}$, which is on the same scale as the deviance statistic. Positive values of T_B indicate that M_2 is more likely to be the true model given the data and the particular prior distributions $p(\theta_k|M_k)$ used in the analysis, and the numerical value of T_B describes the strength of such evidence for or against M_1 and M_2. For example, one common rule of thumb recommends that T_B should be at least 2 in absolute value to be regarded as positive evidence for either model (for a thorough discussion of such rules and other issues related to Bayes factors, as well as further references, see, e.g., Kass and Raftery 1995; Raftery 1995).

The ratio (2.2) and T_B can, in principle, be calculated for any pair of models for D. These need not be nested and may, in general, be completely different in form and assumptions. For any two models, furthermore, there is an infinite number of possible prior distributions and thus of Bayes factors. For most of these, the integral in (2.1) will

not have a closed-form expression, so calculating the Bayes factor is often nontrivial. With some computational effort, it can, however, be well approximated by simulation-based numerical methods. An alternative is to use a closed-form approximation of it. One such expression is obtained by applying the so-called Laplace approximation (see, e.g., Tierney and Kadane 1986) to $p(D|M_2)$ and $p(D|M_1)$ defined by (2.1). This gives

$$\begin{aligned} \mathrm{T}_B \;\approx\; & 2[l(\tilde{\theta}_2) - l(\tilde{\theta}_1)] + 2[\lambda(\tilde{\theta}_2) - \lambda(\tilde{\theta}_2)] + \log|\Psi_2| \\ & - \log|\Psi_1| + (p_2 - p_1)\log(2\pi), \end{aligned} \tag{2.3}$$

where $\tilde{\theta}_k$ is the mode of the posterior distribution $p(\theta_k|D, M_k)$, $\lambda(\theta_k) = \log p(\theta_k|M_k)$, and $\Psi_k = -[\partial^2 l(\tilde{\theta}_k)/\,\partial\theta_k\,\partial\theta_k' + \partial^2\lambda(\tilde{\theta}_k)/\,\partial\theta_k\,\partial\theta_k']^{-1}$. This holds for most regular models and priors and requires only posterior quantities routinely available from a fully Bayesian analysis. The error of the approximation is of the rapidly diminishing order $O(n^{-1})$, so it is very accurate when n is moderately large.

The BIC criterion is also an approximation of T_B. For didactic purposes, it is useful to consider a slightly more general version of it, given by

$$\mathrm{T}_B \approx \mathrm{BIC}_e = 2[l(\hat{\theta}_2) - l(\hat{\theta}_1)] - (p_2 - p_1)\log(1 + n/n_0), \tag{2.4}$$

where n_0 is a positive constant that is discussed below. When $n_0 = 1$ and n is not very small, $\log(1 + n/n_0)$ is approximately $\log n$, and BIC_e becomes the BIC given in (1.3).

The statistic (2.4) is simpler than the Laplace approximation and is particularly convenient because it involves only standard output from maximum likelihood estimation. On the other hand, it involves no explicit specification of the prior distributions $p(\theta_k|M_k)$ and can thus only be a good approximation of T_B for some specific sets of priors. These priors implicitly assumed by BIC_e are ones where

$$p(\theta_k|M_k) \sim N(\hat{\theta}_k, (n/n_0)\hat{V}_k), \tag{2.5}$$

that is, the prior density for both M_1 and M_2 is that of a multivariate normal distribution with mean $\hat{\theta}_k$ and variance $(n/n_0)\hat{V}_k$, where $\hat{V}_k = -[\partial^2 l(\hat{\theta}_k)/\partial\theta_k\,\partial\theta_k']^{-1}$ is the estimated variance matrix of the MLE $\hat{\theta}_k$ obtained from the observed data D. Plugging these priors into (2.3)

shows that (2.4) is then equal to the Laplace approximation of T_B and so shares its large-sample accuracy. For other prior distributions, BIC_e is a less good (and possibly very bad) approximation of T_B.

The prior distributions underlying (2.4) are approximately comparable to estimated sampling distributions of the MLEs of θ_k that would be obtained from a hypothetical sample of n_0 observations. It is thus reasonable to call n_0 the *prior sample size*. It describes how informative the prior is relative to the information provided by D, with small values of n_0 corresponding to large prior variances. The standard BIC approximation is obtained when $n_0 = 1$, and the prior implied by this is often called the *unit information prior* (Kass and Wasserman 1995).

Two kinds of values of n_0 may be of interest. In the first case, n_0 is a constant (e.g., the $n_0 = 1$ leading to BIC), while the asymptotic considerations refer to increasing the sample size n. This is a common approach in Bayesian estimation. As n increases, the posterior density then comes to be dominated by the information from the data, and posterior estimates of the parameters become increasingly similar to the MLEs. In model selection, however, it can be argued that this approach is less natural and may penalize large models too heavily (see, e.g., Smith and Spiegelhalter 1980; Cox 1995; Stone 1979). Unlike in the case of parameter estimates, for example, the conclusions from a Bayes factor may disagree with those obtained from a significance test when n is large and n_0 is constant.[1] When such conflicts occur, they are such that the Bayes factor prefers a smaller model that is rejected by a significance test.

The argument for other choices of n_0 is that a comparison of two models is seriously contemplated only if there is real uncertainty about which of the two is correct. This implies that, in the nested case of Section 1, ψ is a priori either equal to ψ_0 or within a few estimation standard errors of it. The prior variance of ψ should then be proportional to n^{-1} and thus n_0 proportional to n. We will discuss the effects of changing the prior variance by varying n_0 in Section 4. Until then, we will mostly consider the BIC case of $n_0 = 1$.

Because the implicit prior distributions associated with the BIC_e approximation depend on the observed data, they clearly cannot represent genuine priors specified before collecting the data (except approximately in cases when the priors actually were obtained from

a training sample of n_0 observations; see, e.g., O'Hagan 1995). Their choice and the use of BIC_e (in practice, usually BIC) thus require careful justification. This can be provided in at least two rather different ways. First, BIC_e is also a good approximation of T_B for some other prior distributions. When the means and variances of the priors are sufficiently close to (2.5), the error of the approximation is of the larger but still asymptotically vanishing order $O(n^{-1/2})$ (see Raftery 1995, who outlines the proof for the case $n_0 = 1$; the general case is obtained with a straightforward modification).

An even more general result (presented by, e.g., Jeffreys 1961; Cox and Hinkley 1978; Kass and Vaidyanathan 1992; Kass and Wasserman 1995) applies when the models M_1 and M_2 are nested. For this, we first parametrize the models in terms of $\phi_2 = (\beta, \psi)$, $\phi_1 = (\beta, \psi_0)$, where β and ψ have the property of *null orthogonality* (for a definition of this, see Kass and Vaidyanathan 1992); this is typically different from the parameterization in terms of $\theta = (\alpha, \psi)$ introduced in Section 1. Then BIC_e is again an approximation of T_B with an error of order $O(n^{-1/2})$ if (1) the prior for ψ under M_2 is of the form (2.5), (2) the prior for β is the same under both models, and (3) the priors for β and ψ are independent of each other. The exact choice of the prior distribution for β, on the other hand, is essentially arbitrary. It can then be shown that these choices correspond approximately to priors for the θ-parameterization, which are comparable to (2.5) in that the prior variance matrix for ψ and the covariances between α and ψ are similar to those in (2.5).[2] In the case of nested models, the accuracy of the approximation is thus largely unaffected by changes in the prior distributions of the shared parameters α, as long as these are mutually consistent. The result is also relatively insensitive to the prior mean of ψ, at least when n_0 is small. What matters most is that the prior variance of ψ and the covariance structure between α and ψ should be reasonably similar to those of the data-dependent priors discussed above. We will briefly examine the effects of changes in these parts of the priors in the examples below.

The second argument supporting the use of BIC is more pragmatic than theoretically coherent. It is that, in practice, BIC is arguably often used as an essentially non-Bayesian statistic. If we are indeed carrying out a fully Bayesian analysis with explicitly specified prior distributions, it is nowadays usually possible to use numerical methods

or a Laplace approximation to calculate the actual Bayes factor quite accurately. There is thus little need for a rough approximation such as BIC, except as a quick initial calculation. On the other hand, BIC can also be easily calculated from the standard output of a purely likelihood-based analysis and can be used for model selection also in such an analysis. It is then of relatively little concern that the prior distributions involved in its Bayesian motivation remain entirely implicit. Nevertheless, it is reassuring to know that such a motivation exists, so that the criterion is not a purely ad hoc statistic. It is also worth noting that the implicit priors of BIC are at least consistent in that the same prior is associated with a given model, whatever the pair of models being compared. Thus, the set of comparisons between all possible pairs of candidate models is also consistent in this sense.

Even when BIC is used in such an implicitly and half-heartedly Bayesian way, it is still desirable that it should be defensible as an approximate Bayes factor. This could be challenged on several levels of generality. Most broadly, it could be argued that the priors implied by BIC are inappropriate and do not represent reasonable prior beliefs about the parameters. Such objections are discussed by Weakliem (1999; see also the associated discussion). He argues, first, that the unit information prior often differs from prior distributions that would be chosen by many researchers; in particular, it may be less informative than even the weakest prior that would be realistically contemplated. This may have important consequences because conclusions from Bayes factors are sensitive to the choice of the prior distributions, much more so than Bayesian estimates of parameters. Weakliem's second main comment concerns the implicit assumption of standard BIC that the prior distributions represent $1/n$ times the information about the parameters provided by the observed data. This is questionable because the amount of information in n observations depends on both the nature of the parameters and the configuration of the data. The unit information prior of BIC, rather than being a simple and unambiguous reference distribution, can thus represent very different levels of prior information in different circumstances.

The strength of these objections is undeniable. They reinforce the observation that BIC cannot replace a fully Bayesian model comparison with prior distributions chosen carefully for a specific problem. Nevertheless, a simple criterion such as BIC can still also be useful

(we could also consider BIC_e with different values of n_0, thus varying the informativeness of the implicit prior). Weakliem's (1999) and other similar comments should, however, be borne in mind in its interpretation. In particular, the second of the comments above shows that BIC may often imply a prior that is even less informative (and the penalty term effectively even stronger) than the unit information formulation suggests. This issue will be discussed further in the context of the social mobility example in Section 6.

A much narrower criterion of the adequacy of BIC is whether it is a good approximation of T_B with the specified priors. If that is not the case, different errors of approximation in different models might distort comparisons of models. With the unit information priors exactly as specified above, BIC differs from T_B only because it is an asymptotic approximation, so the difference should be small in large samples. More generally, we could consider its accuracy as an approximation of T_B given some other prior distributions. BIC cannot then be expected to perform well in general but should do so for priors that are reasonably close to the unit information prior. Below we consider this in two simple examples with different priors in which always $n_0 = 1$ (so that their overall informativeness is the same), but the prior means and variance structures may differ from those assumed by BIC.

EXAMPLE 1: NORMAL LINEAR MODELS

Suppose that the data $D = (y_1, \ldots, y_n)$ are n observations of a continuous response variable Y. We consider linear models in which M_k states that $D \sim N(X_k\theta_k, \sigma^2 I)$ for some known matrix X_k. Thus, under all models, y_i are independent with variance σ^2, which is assumed to be known. In particular, two nested models, M_1 and M_2, are considered. In M_1, X_1 is a column of 1s and $\theta_1 = \alpha$, and thus $\text{E}(y_i) = \alpha$ for all i. For M_2, X_2 has rows $(1, x_i)$ and $\theta_2 = (\alpha, \psi)$, so that $\text{E}(y_i) = \alpha + \psi x_i$, where x_i are known explanatory variables. The priors implied by BIC_e are normal distributions, with means $\bar{y} = n^{-1}\sum y_i$ and $\hat{\theta}_2 = (\hat{\alpha}_2, \hat{\psi})$ and variances σ^2/n_0 and $\sigma^2(n/n_0)(X_2'X_2)^{-1}$ for M_1 and M_2, respectively, where $\hat{\theta}_2$ denotes the least squares estimate. The null orthogonality of the parameters in M_2 mentioned in Section 2 is obtained by centering the explanatory

variable, so that $\bar{x} = 0$. In that case, $\hat{\alpha}_2$ and $\hat{\psi}$ are uncorrelated, and $\hat{\alpha}_2 = \bar{y}$ and its variance σ^2/n are the same as under M_1.

When these prior are used, T_B is for any sample size exactly equal to $BIC_e = \sum[\hat{e}_{1i}^2 - \hat{e}_{2i}^2]/\sigma^2 - \log(1 + n/n_0)$, where $\hat{e}_{ki} = y_i - \hat{y}_{ki}$ are the standard least squares residuals for the fitted models M_k. In other cases, numerical examples suggest that BIC_e is a reasonably good approximation of T_B as long as the priors are fairly close to the ones described above. This result is most sensitive to changes in the prior covariance between α and ψ. Numerical examples of this are not shown here because they are qualitatively similar to the ones in the next example.

EXAMPLE 2: LOG-LINEAR MODELS FOR A THREE-WAY TABLE

Here the data are the observed cell counts $D = (y_{111}, y_{112}, \ldots, y_{222})$ for a $2 \times 2 \times 2$ contingency table for three binary variables A, B, and C, with a sample size $n = \sum_{i,j,l} y_{ijl}$. Consider log-linear models in which y_{ijl} are independent Poisson variates with means μ_{ijl}, given by $\log \mu_{ijl} = \lambda_0 + \lambda_{A(i)} + \lambda_{B(j)} + \lambda_{C(l)} + \lambda_{AB(ij)} + \lambda_{AC(il)} + \lambda_{BC(jl)} + \lambda_{ABC(ijl)}$. Applying identifiability constraints in which any parameter with a subscript i, j, or l equal to 1 is set to zero, we can denote the remaining parameters by $\lambda_A = \lambda_{A(2)}$, $\lambda_B = \lambda_{B(2)}$, and so forth.

The two models to be compared are M_1, in which the only nonzero parameters are $\theta_1 = \alpha = (\lambda_0, \lambda_A, \lambda_B, \lambda_C)$ and M_2, which has parameters $\theta_2 = (\alpha, \lambda_{AB}) = (\alpha, \psi)$. Thus, under M_1, all the variables are mutually independent, while M_2 allows for an association between A and B. In both cases, the model for μ_{ijl} can be expressed as $\log \mu_{ijl} = x'_{(k)ijl}\theta_k$, with an appropriate choice of the vector $x_{(k)ijl}$. The variance matrices of the maximum likelihood estimates $\hat{\theta}_k$ are estimated by $\hat{\text{var}}(\hat{\theta}_k) = (X'_k D_k X_k)^{-1}$, where X_k is a matrix with rows $x'_{(k)ijl}$, and D_k is a diagonal matrix of the fitted values $\hat{\mu}_{ijl}^{(k)} = \exp(x'_{(k)ijl}\hat{\theta}_k)$. The prior distributions underlying the BIC_e approximation are derived from these as before.

For log-linear models, BIC_e is only a large-sample approximation of T_B. To examine both the accuracy of this approximation and its sensitivity to other choices of the prior distributions, we carried out a small simulation study. Tables of $n = 100$ observations were considered, with true models for μ_{ijl} given by different choices of

$\theta_2 = (\alpha, \psi)$. The value of α was either $(\lambda_0, 0, 0, 0)$ or $(\lambda_0, 1.5, 1.5, 1.5)$, where in both cases, $\lambda_0 = \log[n\,(\mu_{111}/\sum \mu_{ijl})] = \log(n\pi_{111})$. In the first of these cases (labeled "Even Data" in Table 2), expected counts μ_{ijl} are all fairly similar (exactly equal to 12.5 when $\psi = 0$), while in the second case ("Sparse Data"), they vary more widely (from 0.6 to 55 when $\psi = 0$). This difference is used to examine the accuracy of the asymptotic approximation in different circumstances (in other contexts, asymptotic results for contingency tables tend to be least accurate for sparse tables with some small frequencies). Three values of ψ were considered: 0, $0.2 = 2/\sqrt{n}$, and 1. Thus, model M_1 holds in the first case and M_2 in the other two, with the true value of ψ being relatively close to $\psi_0 = 0$ in the second but not in the third case. All six combinations of these α and ψ were considered, with 2,000 simulated data sets generated for each setting from a multinomial distribution with probabilities $\pi_{ijl} = \mu_{ijl}/\sum \mu_{ijl}$.

Values of T_B were calculated for four sets of prior distributions for θ_k, all taken to be normal. The prior means of θ_1 and θ_2 were either $\hat{\theta}_1 = \hat{\alpha}_1$ and $\hat{\theta}_2 = (\hat{\alpha}_2, \hat{\psi})$ or α^* and $(\alpha^*, 0)$, where α^* denotes the true value of α in each simulation. Similarly, the prior variances were either $(n/n_0)\hat{\mathrm{var}}(\hat{\theta}_k)$ or $(n/n_0)V_k$, where $n_0 = 1$, V_k is a diagonal matrix with diagonal elements equal to those of $(X_k' D_\mu X_k)^{-1}$, and D_μ denotes a diagonal matrix of the true values of μ_{ijl}. We consider all four combinations of these choices for the prior means and prior variance matrices, where in each case the first choice corresponds to the prior implied by BIC. The prior is indicated in the third column of Table 2 by stating which of the moments were as assumed by BIC.

BIC was calculated for each generated data set, together with the Laplace approximation (2.3) of T_B for each choice of prior distributions. Table 2 shows their averages over the 2,000 simulated data sets for each simulation setting. Also given are results for an estimate of T_B (labeled "Simul."), obtained using an importance sampling procedure proposed by Newton and Raftery (1994). This involves no asymptotic approximations and can be regarded as the true value of T_B apart from simulation variation controlled by the sample sizes (here 8,000) used for the importance sampling. Here the simulation standard error in the averages in Table 2 is less than 0.004.

Differences between the Laplace approximation and the simulation-based estimate are due to the large-sample nature of the

TABLE 2: Simulation Results for Approximations of $T_B = 2\log BF_{21}$ in the Log-Linear Example 2

			Even Data				*Sparse Data*			
			Mean				*Mean*			
	ψ	*BIC_e Prior?*	*Simul.*	*Lapl.*	*BIC*	*Percentage BIC Agrees*	*Simul.*	*Lapl.*	*BIC*	*Percentage BIC Agrees*
I	0	Both	−3.58	−3.59	−3.59	99.8	−3.61	−3.68	−3.68	99.9
		Mean	−5.15	−5.15	−3.59	97.8	−7.76	−7.82	−3.68	98.1
		Variance	−3.59	−3.60	−3.59	99.8	−3.62	−3.69	−3.68	99.9
		Neither	−5.16	−5.17	−3.59	97.8	−7.79	−7.85	−3.68	98.1
II	0.2	Both	−3.37	−3.38	−3.38	99.8	−3.52	−3.59	−3.59	99.9
		Mean	−5.05	−5.06	−3.38	97.4	−7.78	−7.84	−3.59	97.0
		Variance	−3.38	−3.40	−3.38	99.8	−3.53	−3.60	−3.59	100.0
		Neither	−5.05	−5.07	−3.38	97.4	−7.78	−7.84	−3.59	96.8
III	1	Both	1.67	1.65	1.65	98.7	−2.12	−2.23	−2.23	99.6
		Mean	−0.32	−0.34	1.65	83.6	−6.71	−6.80	−2.23	84.9
		Variance	1.56	1.54	1.65	98.7	−2.16	−2.26	−2.23	99.6
		Neither	−0.18	−0.20	1.65	84.6	−6.55	−6.64	−2.23	85.0

NOTE: Simul. = a simulation-based (importance sampling) estimate; Lapl. = the Laplace approximation (2.3); BIC = Bayesian information criterion (1.3). The mean is the average of these statistics in 2,000 simulated data sets, and "Percentage BIC Agrees" denotes the proportion of data sets in which BIC has the same sign as the simulation-based estimate. See the text for the parameter values used in the simulations.

approximation. Here the level of agreement is very high, even in the case of sparse tables. It thus seems that the small-sample error of the Laplace approximation is not a serious concern in these examples, even with the moderate sample size of 100. The same is true for BIC when it is otherwise accurate. As expected, BIC and the Laplace approximation are equal when the prior distributions are chosen appropriately. Furthermore, BIC here is mostly a good approximation of T_B also for the other priors considered. Its accuracy is essentially unaffected by small changes in the prior mean (note, however, that if n_0 were proportional to n, the results would be more sensitive in this respect). On the other hand, a change in the prior variance structure—here, specifically assuming prior independence between the parameters instead of a covariance structure mimicking that of the MLEs—clearly changes T_B and causes BIC to be inaccurate.

BIC does not always even need to be a highly accurate approximation for T_B if it is used mainly to identify which of the models has the highest posterior probability. In other words, what matters most is that the sign of T_B should be accurately determined. To examine this, Table 2 also shows the proportion of the simulations in which BIC had the same sign as the importance sampling estimate of T_B (i.e., the proportion of cases in which these two statistics would lead to the same conclusion about the choice of model). The agreement is very close—in most cases, nearly complete. The only cases in which there is some discrepancy are those with the wrong prior variance structure and $\psi = 1$, where BIC tends to prefer the smaller model too often. Thus, in summary, BIC performs well in these small examples, in the narrow sense of approximating T_B adequately. This result holds even when the prior distributions of the parameters differ to a small extent from those assumed by BIC.

3. MOTIVATION OF AIC

The AIC statistic (1.2) has a very different theoretical motivation from that of the Bayesian BIC. This can be described in two apparently different but essentially equivalent ways: either using a measure of similarity between candidate models and the true model for the data or in terms of the expected predictive performance of the models. These

motivations are briefly outlined here. More information, as well as a thorough account of and references to literature since the original work of Akaike (1973), is given by Burnham and Anderson (1998).

Suppose that the observed data D are generated by a distribution with a density function $f(D)$. This true model is regarded as unknown and unknowable. In particular, it is not assumed that it corresponds to any of the models under consideration, so that $f(D)$ does not need to be equal to any $p(D|\theta_k; M_k)$ with any value of θ_k. The aim of model selection can then not be to identify the true model exactly but to propose simpler models that are good approximations of it. Comparisons between the true model and candidate models M_k can be expressed in terms of some measure of distance between the densities $f(D)$ and $p(D|\theta_k; M_k)$. One such measure is the basis of AIC. It is the Kullback-Leibler (K-L) distance or discrepancy between the densities, which is defined as

$$\begin{aligned} I[f, p(\theta_k)] &= \int f(y) \log \left[\frac{f(y)}{p(y|\theta_k, M_k)} \right] dy \\ &= C - \mathrm{E}_y \log p(y|\theta_k; M_k), \end{aligned} \tag{3.1}$$

where $C = \mathrm{E}_y f(y)$. Here, y is a random variable of the same size as D from the true density f, and E_y (like all expectations below) denotes its expected value with respect to the true distribution.

An obvious strategy for model comparison would be to choose the model that is closest to the true model, as measured by the K-L discrepancy. This, however, cannot be implemented directly because several of the quantities in (3.1) remain unknown or ambiguous. First, C can never be known since it depends only on the true distribution. However, because it is the same for all of the candidate models M_k, it can be removed simply by considering differences of $I[f, p(\theta_k)]$ between models, which do not need to be nested. Only the differences are needed for comparison of the discrepancies and thus of models.

The second problem is that (3.1) is actually the discrepancy between the true model and one density of the type M_k with a specific parameter value θ_k. What is needed instead is something that represents the discrepancy between f and the whole family of distributions

comprising M_k. It is natural to define this as $I[f, p(\theta_k^*)]$, where θ_k^* is the value for which (3.1) is smallest among all possible values of θ_k. We call θ_k^* the "pseudo-true" value of θ_k. If M_k is in fact the true model with some value of θ_k, this must be equal to θ_k^*, and the discrepancy is zero.

Next, the unknown θ_k^* needs to be estimated. If M_k is the true model, the MLE $\hat{\theta}_k$ is a consistent estimate of θ_k^*. Remarkably, the same is true even when M_k is not correct and θ_k^* is thus only a pseudo-true value. This result provides a strong motivation for maximum likelihood (ML) estimation and suggests that the target quantity could be estimated by substituting MLEs for θ_k^* in (3.1). Two models could then be compared using the difference $I[f, p(\hat{\theta}_2^y)] - I[f, p(\hat{\theta}_1^y)] = \mathrm{E}_y[\log p(y|\hat{\theta}_2^y; M_2) - \log p(y|\hat{\theta}_1^y; M_1)]$, where $\hat{\theta}_k^y$ denotes the MLE of θ_k obtained from data y. However, one more problem is apparent in this. If M_1 is nested within M_2, $\log p(y|\hat{\theta}_2^y; M_2)$ can never be smaller than $\log p(y|\hat{\theta}_1^y; M_1)$, and thus the difference would never indicate preference for the smaller model M_1. This happens because the same set of data y is used both to estimate $\hat{\theta}_k^y$ and to judge the fit of the resulting model. This leads to an overoptimistic assessment of fit that would inevitably favor larger models. To avoid this, one more hypothetical extension of the situation is introduced by assuming that the MLE is based on a separate, independent sample of data x from the same true distribution f. Denoting this MLE by $\hat{\theta}_k^x$, we obtain a new target quantity

$$\begin{aligned} \mathrm{T}_A &= 2\mathrm{E}_x\{I[f, p_1(\hat{\theta}_1^x)] - I[f, p_2(\hat{\theta}_2^x)]\} \\ &= 2\mathrm{E}_x\mathrm{E}_y\,[\log p(y|\hat{\theta}_2^x, M_2) - \log p(y|\hat{\theta}_1^x, M_1)], \end{aligned} \tag{3.2}$$

which has again been multiplied by 2 so that estimates of it will be on the same scale as the likelihood ratio statistic. Both reasonable and estimable, this is the target quantity that AIC aims to estimate. When T_A is positive, the expected (over x) K-L discrepancy between M_2 and the true model f is smaller than that between M_1 and f and vice versa.

In Section 2, the target quantity T_B was fully determined by the models, prior distributions, and observed data. The task was then to *approximate* its numerical value as well as possible. Here, however, T_A is a property of the true distribution of D and is thus unknown.

However, it can be *estimated* from observed data. Under certain conditions, a good estimate of it is

$$\mathrm{T}_A \approx \mathrm{AIC}_e = 2[l(\hat{\theta}_2) - l(\hat{\theta}_1)] - (p_2 - p_1) \times (1 + n/n_0), \quad (3.3)$$

where $\hat{\theta}_k$ is again the MLE obtained from the actual data D. Here, n_0 denotes the size of the hypothetical training sample x (or, more precisely, the variance of $\hat{\theta}_k^x$ is assumed to be n/n_0 times the variance of $\hat{\theta}_k^y$). This is a more general version of the standard AIC (1.2), which is obtained as a special case when $n_0 = n$. The derivation of AIC is originally due Akaike (1973), and an accessible version of it is given by Burnham and Anderson (1998). The result (3.3) is obtained from the proof in Burnham and Anderson by introducing, at appropriate places, the above assumption about the variances of $\hat{\theta}_k^x$ and $\hat{\theta}_k^y$. Expression (3.3) will be used in Section 4, where it helps us to examine the interpretation of its penalty term. Otherwise, only the standard AIC statistic is considered in the examples below.

The definition of T_A in (3.2) involves two hypothetical data sets: one (x) that is used to estimate the parameters θ_k and one (y) that is used to judge the fit of the resulting model. This formulation leads naturally to the second interpretation of AIC, which does not make explicit use of the K-L discrepancy. Instead, it focuses on prediction. The log-likelihood $\log p(y|\hat{\theta}_k^x)$ can be regarded as a measure of how well model M_k with parameters estimated by $\hat{\theta}_k^x$ predicts new data y. This is seen most easily in the simple normal case of Example 1, in which the measure is essentially the mean squared error of prediction $\sum(y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is a fitted value for $\hat{y}_i$ based on a model estimated from x. The form of the log-likelihood is different in other models, but it can still be interpreted as a measure of goodness of prediction. Thus, a second interpretation of model selection using AIC_e is that it is based on estimating how well each model, with parameters estimated from n_0 observations, is expected to predict a new data set of n observations from the same distribution.

The second interpretation links AIC to other methods of model comparison, which are more explicitly predictive. For normal linear models, well-known statistics such as Mallow's C_p and adjusted R^2 are closely related to AIC, as all are essentially different

transformations of estimates of the mean squared error of prediction (see, e.g., Miller 2002). More generally, AIC is closely related to certain versions of cross-validation (Stone 1977).

As in the case of BIC_e, AIC_e is a good estimate of T_A only under certain conditions. First, it is again a large-sample estimate, where both n and n_0 in (3.3) are in general assumed to be large. Second, and more surprisingly, it assumes that the models under consideration are actually true. In the case of nested models, this amounts to a condition that the true model should be the smaller model M_1 or very close to it. This seems like a strange assumption, given that it was earlier specifically not required that any of the candidate models should be true. The motivation of this assumption, which may or may not be justified in a given comparison, is that it greatly simplifies the resulting estimate of T_A. The quantity $(1 + n/n_0)$ in (3.3) (i.e., the constant 2 in the standard AIC) is in fact an estimate of a certain property of the true distribution, which has exactly this value when the fitted models are true. Otherwise, the estimate is biased. On the other hand, it has a variance of zero. Other, less biased, estimates for the same quantity exist, but their variances must also be larger. Thus, the constant estimate used in AIC_e, besides being trivial to calculate, is likely to have a lower mean squared error than alternatives in many models in which its assumptions are at least roughly satisfied.

Both of the assumptions of AIC can be relaxed by using modified estimators. The true model assumption can be removed by replacing $(1 + n/n_0)$ with alternative, data-dependent quantities. Finite-sample bias can be reduced by using adjusted versions of AIC, proposed first by Sugiura (1978) and Hurvich and Tsai (1989). Results by them and others (see, e.g., Burnham and Anderson 1998) suggest that in small samples, the bias can be serious and that improvement obtained from adjusted estimates can often be substantial. A slight disadvantage of such estimates is that the small-sample adjustment needs to be derived specifically for each class of models considered. Here we will not consider these approaches further. Instead, we will try to assess how well the simple approximation performs in estimating T_A and how its different assumptions contribute to its bias. This is again done in the context of the two examples introduced in Section 2.

EXAMPLE 1: NORMAL LINEAR MODELS (CONTINUED)

To assess the performance of AIC, one must assume something about the true model. Here we consider the fairly restricted case in which y_i are independent normal random variables with means μ_i and variance σ^2, where σ^2 is known but μ_i can be of any general form. This includes also the models M_1 ($\mu_i = \alpha$) and M_2 ($\mu_i = \alpha + \psi x_i$) as special cases. In this case, AIC is, unusually, an exactly unbiased estimate of T_A, even when μ_i are not given by either of the models under consideration. Then $\mathrm{T}_A = \mathrm{E}_D(AIC) = \sum[e_{1i}^2 - e_{2i}^2]/\sigma^2 - (p_2 - p_1)$, where $e_{ki} = \mu_i - \tilde{\mu}_{ki}$ and $\tilde{\mu}_{ki} = X_k(X_k'X_k)^{-1}X_k'\mu_i$ are the best linear prediction of μ_i obtainable with the explanatory variables in X_k.

Neither of the two sources of bias in AIC thus exist in this example. However, it is also subject to a third kind of error. As shown in (3.2), the target quantity T_A is the expected value of a function of data y and x drawn from the true distribution, whereas AIC_e in (3.3) is an estimate of it based on a single data set D from that distribution. That is comparable to estimating the mean of a distribution by the value of one observation from it. Consequently, AIC has a variance that will not be reduced to zero by increasing n. This is easiest to see in the case in which M_1 actually holds. Then $e_{1i} = e_{2i}$, $\mathrm{T}_A = -(p_2 - p_1) = -1$, and $\mathrm{AIC} = 2[l(\hat{\theta}_2) - l(\hat{\theta}_1)] - 2$, which is distributed as $\chi_1^2 - 2$. There is thus a probability of about 0.157 that AIC will be positive, leading to the choice of the larger model M_2. With this probability, a single data set of n observations drawn from the distribution is such that parameter estimates from it will incorrectly make M_2 appear a better predictor of future data than M_1.

EXAMPLE 2: LOG-LINEAR MODELS
FOR A THREE-WAY TABLE (CONTINUED)

A simulation study of AIC was carried out, with 100,000 simulated tables of $n = 100$ or $n = 1,000$ observations generated for each parameter setting. The true models were all log-linear with the same two values of $\alpha = (\lambda_0, \lambda_A, \lambda_B, \lambda_C)$ (leading to "even" and "sparse" data), as in the simulation in Section 2. Five choices for the other parameters were considered. Models I, II (with $\psi = \lambda_{AB} = 2/\sqrt{n}$), and III were the same as in Table 2, while in Models IV and V, the only

other nonzero parameters were $\lambda_{BC} = 1$ and $\lambda_{ABC} = 1$, respectively. Thus, Cases IV and V are ones in which neither M_1 nor M_2 holds. In Case IV, M_1 is closer to the true model, while M_2 is generally closer to it in Case V.

Table 3 shows the means and standard deviations of AIC calculated for the 100,000 simulated data sets. It also shows for each case an estimate of the true value of T_A, estimated from (3.2) by averaging more than 100,000 sets of values of x and y generated separately from the true distribution. This has a simulation standard error of less than 0.01 for the smaller values and at most about 0.05 when $\hat{T}_A = 51.61$. The table also shows the proportion of simulations in which AIC had the same sign as the true T_A.

The difference between T_A and the sample mean of AIC reflects the bias in AIC from the two sources discussed above. In Model I, both M_1 and M_2 hold, and only the small-sample bias should be present. This bias is noticeable but fairly small, even with $n = 100$, and smaller still when $n = 1,000$. In Cases II through V, in which one or both of the two models are incorrect, the failure of this assumption of AIC also introduces some bias. However, the fact that the observed biases are not very different from those of Model I suggests that this bias is not substantial. In these simple examples, AIC thus seems to be a nearly unbiased estimate of T_A.

The main discrepancy between AIC and T_A arises, as in Example 1 above, not from average bias but from the variability of AIC across samples. Thus, it is always possible that the sign of T_A, estimated by AIC from a single sample, will be incorrect. In Case I, in which model M_1 holds, the error rate is close to the 15.7 percent we would expect from the χ^2_1 distribution. In the other cases, the proportion of agreement varies, depending on the balance between the mean and variability of AIC. The error rate is highest in cases when T_A is close to zero (i.e., in which the models are difficult to distinguish because their predictive performances are fairly similar).

4. PARALLELS BETWEEN AIC AND BIC

The AIC and BIC criteria are derived from very different theoretical starting points and are thus, to some extent, incommensurable.

TABLE 3: Simulation Results for AIC as an Estimate of T_A in the Log-Linear Example 2

		n = *100*				n = *1,000*			
		AIC				*AIC*			
Model	*Data*	T_A	*Mean*	*Standard Deviation*	*Percentage Agree*	T_A	*Mean*	*Standard Deviation*	*Percentage Agree*
I	Even	−1.08	−0.99	1.44	84.4	−1.01	−1.00	1.42	84.4
	Sparse	−1.03	−1.07	1.27	85.6	−1.04	−0.99	1.43	84.2
II	Even	−0.82	−0.73	1.76	79.3	−0.76	−0.74	1.73	79.0
	Sparse	−0.97	−0.98	1.48	84.3	−0.95	−0.91	1.54	82.3
III	Even	4.14	4.28	4.81	81.1	51.61	51.58	14.57	100
	Sparse	0.04	0.31	2.74	38.9	7.99	8.04	6.75	91.8
IV	Even	−1.11	−0.98	1.44	84.0	−1.03	−1.00	1.42	84.4
	Sparse	−0.75	−1.25	1.20	89.8	−1.07	−0.99	1.43	84.0
V	Even	1.13	1.26	3.33	53.6	21.55	21.55	9.56	100
	Sparse	−0.07	0.07	25.92	64.9	6.82	6.88	6.24	89.3

NOTE: T_A denotes a more precise simulation estimate of T_A; the mean and standard deviation are the average and standard deviation of the Akaike's information criterion (AIC) in 100,000 simulated tables, and "Percentage Agree" is the proportion of simulations in which AIC had the same sign as T_A. See the text for the parameter values of the true models.

Nevertheless, some similarities and connections between them can also be observed. First, both are penalized criteria of the form (1.1). In Sections 2 and 3, we have considered the extended versions BIC_e (2.4) and AIC_e (3.3), both involving an additional quantity n_0 (which is defined differently in the two cases). We can use this to interpret any penalized criterion (1.1) both in a Bayesian way as BIC_e and as a predictive criterion AIC_e. Specifically, a criterion with penalty coefficient a corresponds to BIC_e with $n_0 = n/(e^a - 1)$ and also to AIC_e with $n_0 = n/(a - 1)$. In particular, the standard BIC (1.3) is equal to AIC_e with $n_0 = n/(\log n - 1)$, and the standard AIC (1.2) can also be interpreted as BIC_e with $n_0 = n/(e^2 - 1) \approx n/6.4$. Thus, AIC corresponds to a Bayes factor in which the prior sample size is proportional to n. Similarly, use of BIC is comparable to a predictive analysis in which the training sample is substantially smaller than the sample to be predicted (since $1/\log n$ tends to zero as n increases). In both cases, the value of n_0 implied by one criterion is somewhat unusual in the context of the other.

Another kind of comparison between the criteria is highlighted by writing their target quantities as

$$\begin{aligned} \text{T}_A &= 2\{\text{E}_y[\text{E}_{\hat{\theta}_2^x} \log p(y|\hat{\theta}_2^x; M_2)] \\ &\quad - \text{E}_y[\text{E}_{\hat{\theta}_1^x} \log p(y|\hat{\theta}_1^x; M_1)]\} \qquad \text{and} \qquad (4.4) \\ \text{T}_B &= 2\{\log \text{E}_{\theta_2} p(D|\theta_2; M_2) \\ &\quad - \log \text{E}_{\theta_1} p(D|\theta_1; M_1)\}, \qquad (4.5) \end{aligned}$$

where the second expectations in (4.4) are with respect to the sampling distributions of the MLEs $\hat{\theta}_k^x$ and the expectations in (4.5) with respect to the prior distributions of θ_k. The expressions are (apart from the order of logs and expectations) of broadly the same form. The similarity is enhanced by the fact that the prior distributions assumed for θ_k in BIC_e are identical to the sampling distributions of $\hat{\theta}_k^x$ used to obtain AIC_e.

As one interpretation of (4.4) in terms of prediction, we could try to consider a similar view of (4.5). However, this lacks the second expectation over future data y. Instead, a (rather contrived) predictive interpretation of (4.5) and BIC_e is that they assess how well the prior distributions predict the data D that were actually observed. As an alternative to this, it would also be possible to consider

the same prediction problem as in AIC_e in a Bayesian way. The target quantity is then based on differences of $\mathrm{E}_y \log p(y|D;\, M_k) = \mathrm{E}_y \log \mathrm{E}_{\theta_k} p(y|\theta_k;\, M_k)$, where the expectation over θ_k is with respect to its posterior distribution given D. When the prior distribution for θ_k is relatively uninformative, this posterior is approximately the same as the sampling distribution of $\hat{\theta}_k$. It is then not surprising that this can be used to obtain model selection criteria that are comparable to AIC (Spiegelhalter et al. 2002). Other Bayesian predictive criteria such as this in more general settings have been considered by Laud and Ibrahim (1995) and Gelfand and Ghosh (1998).

A third parallel between AIC_e and BIC_e is observed by comparing the origins of their penalty terms. As discussed in Section 1, the second term in criteria such as (2.4) and (3.3) can be interpreted as a penalty for the increased complexity of the larger model in a comparison. However, the derivations of BIC_e and AIC_e in Sections 2 and 3 did not involve any reference to parsimony or introduce any explicit preference for it. It is thus of some interest to examine the origins of the penalty terms in more detail, to see if this provides any insight into the idea of simplicity of models and the reasons for why it should be desirable in modeling observed data. These explanations turn out to be essentially analogous for AIC_e and BIC_e, despite their different motivations.

Consider first the Bayesian criterion (2.4), which approximates T_B. The central quantity is the marginal likelihood $p(D|M_k)$ given by (2.1). This expresses the likelihood of data D given model M_k. Since M_k is the family of distributions $p(D|\theta_k;\, M_k)$ for all possible θ_k rather than any single value of it, $p(D|M_k)$ is calculated by integrating over θ_k, as shown in (2.1). In other words, the marginal likelihood is essentially a (continuous) weighted average of the likelihoods $p(D|\theta_k;\, M_k)$ at different values of θ_k, with the weights given by values of the prior density $p(\theta_k|M_k)$. The penalty for complexity arises because, all other things being equal, a larger model is disadvantaged in this calculation because it has to spread the prior probability more thinly over a larger dimensional parameter space, leaving less prior mass for the regions supported by the data. This is illustrated in Figure 1, which refers to the following one-parameter example.

Consider a further simplified version of the normal linear Example 1 of Sections 2 and 3. Suppose that the constant model M_1 is

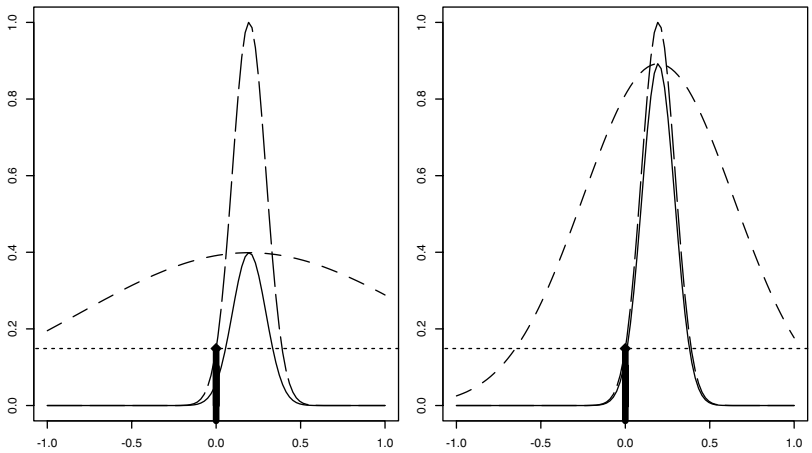

Figure 1: Bayesian Model Comparison in a Simple Normal Example Described in the Text

NOTE: The long dashed line shows the likelihood as a function of mean parameter α for a single data set of $n = 100$ observations, the short dashed line is a prior density for α (normal with mean at sample mean of data and variance $1/n_0$, where $n_0 = 1$ on the left and $n_0 = 5$ on the right), and the solid line is their product. The black diamond shows the value of the likelihood at $\alpha = 0$.

compared to an even simpler model M_0, in which the mean α is fixed at zero. Thus, model M_0 has no unknown parameters, while M_1 has one. Figure 1 shows the quantities involved in $p(D|M_k)$ for this example for a single data set of $n = 100$ observations generated from a distribution satisfying model M_1 with $\alpha = 0.2$ and $\sigma^2 = 1$. The likelihood function $p(D|\alpha; M_1) = p(D|\alpha)$ is shown (scaled to have a maximum of 1) by the long dashed lines. It is maximized at the MLE $\hat{\alpha} = \bar{y}$, which is here approximately 0.2. Model M_0 effectively involves a prior distribution that places probability 1 on $\alpha = 0$, and thus $p(D|M_0)$ is equal to $p(D|0)$. This is shown by the black diamonds in Figure 1. For model M_1, the prior for α is taken as normal with mean $\bar{y}$ and variance σ^2/n_0, where $n_0 = 1$ in the left-hand plot and $n_0 = 5$ on the right. The prior density is shown by the short dashed line and the product $p(D|\alpha, M_1)p(\alpha|M_1)$ by the solid curve. The marginal likelihood $p(D|M_1)$ is thus given by the area under the solid curve.

The difference between the two models is that α is a free parameter in M_1 but fixed at 0 in M_0. For any data set, there are values of α that

provide a better fit to D than $\alpha = 0$. In Figure 1, these are the ones for which the likelihood is higher than the dotted horizontal line. All other values of α, on the other hand, give a worse fit than $\alpha = 0$. The integrated value $p(D|M_1)$ depends on the relative sizes of these sets of good and bad values of α and, crucially, on the values of the prior $p(\alpha|M_1)$ assigned to them. In the left-hand plot, where $n_0 = 1$, the prior variance is large, and the prior distribution gives fairly high weights even to values that are far from $\hat{\alpha}$. As a result, T_B is here -0.81, indicating that, overall, M_0 is a more likely explanation for the data. On the right, where $n_0 = 5$, the prior values decrease faster away from the central region of the parameter space, high-likelihood values of α dominate (2.1), and the conclusion is that M_1 is slightly supported over M_0, with T_B equal to 0.77. This also illustrates that even such apparently minor changes in the prior can change the conclusion of the comparison. Such sensitivity is one reason why values of T_B close to zero should not be treated as decisive evidence for either model.

The penalty term of BIC_e and the results of the Bayesian comparison also depend on the sample size n. When it increases, the likelihood becomes increasingly peaked around the maximum likelihood estimate and decreases quickly around it. If, at the same time, the prior variance remains fixed, increasing n has the same effect in Figure 1 as decreasing n_0 with n fixed. In both cases, those values of α not supported by the data account for an increasingly large portion of the prior probabilities $p(\alpha|M_1)$, and it becomes more likely that $p(D|M_0)$ will be larger than $p(D|M_1)$. The large prior variance implied by $n_0 = 1$ is thus a less innocuous choice in model comparison than it would be in parameter estimation. This is part of the motivation for the argument (discussed in Section 2) that it might often be more reasonable to take n_0 to be proportional to n.

A parallel argument can be made for AIC_e, both in this example and in general. It can be illustrated by a plot similar to Figure 1, apart from the scale and shape of the curves. In it, we would replace $p(D|\alpha)$ with $\mathrm{E}_y \log p(y|\hat{\alpha}^x)$ and $p(\alpha|M_1)$ with $p(\hat{\alpha}^x|M_1)$, the sampling distribution of the MLE $\hat{\alpha}^x = \bar{x}$ based on a sample of n_0 observations. The target quantity is then the integral of the product of these over the range of $\hat{\alpha}^x$. For model M_0, this is simply $\mathrm{E}_y \log p(y|0)$. Again, some samples x will result in an estimate of α, which gives a better expected prediction of y than using $\alpha = 0$, but other values of $\hat{\alpha}^x$

are worse predictors. The balance of these and thus the value of T_A depend on the sampling variance of $\hat{\alpha}^x$, which is σ^2/n_0. The value of $n_0 = 100$ implied by standard AIC corresponds to $n_0 \approx 27$ for BIC_e in Figure 1. Thus, AIC, as always, penalizes the larger model much less than BIC. In fact, here $T_A = 3$, strongly favoring M_1.

In summary, in this simple example, as well as more generally, the penalty term of BIC_e in a nested comparison arises from prior uncertainty about the values of the additional parameters ψ of the larger model. Similarly, in a predictive comparison implied by AIC_e, the penalty is due to uncertainty in the estimates of those additional parameters. In both cases, the crucial issue is that the model is a family of distributions with different values of the parameter ψ, not just with a single value of it. Some of those ψ are better than the fixed value $\psi = \psi_0$ used by the smaller model, but many are much worse. In effect, the larger model gives us more choice of distributions but also more chances of getting that choice wrong. The result of the model comparison depends on the balance between the good and the bad choices given the available information. If the likely values of the additional parameters ψ are not well enough established by that information, the correct decision is to "play safe" and prefer the simpler model.

5. COMPARISONS OF PERFORMANCE

The examples of Sections 2 and 3 suggest that AIC and BIC often do very well in estimating or approximating their target quantities. This, however, is ultimately of lesser interest than the question of their usefulness in practice, as measured by their success in consistently selecting good models for observed data. This is a much more complicated question both to ask and to answer. Some of its difficulty lies in simply defining what is meant by a "good model." As we have seen in Sections 2 and 3, AIC and BIC represent two rather different answers to this question. The aim of the Bayesian approach motivating BIC is to identify the models with the highest probabilities of being the true model for the data, assuming that one of the models under consideration is true. The derivation of AIC, on the other hand, explicitly denies the existence of an identifiable true model and instead uses expected

prediction of future data as the key criterion of the adequacy of a model. These two aims—identification of the true model and prediction of new data—are the two criteria considered here. Their contrast as aims of BIC and AIC has often been pointed out, for example, by Atkinson (1981) and Chow (1981).

Some general theoretical results about the properties of the selection criteria are available. Not surprisingly, one of the main implications of them is that different criteria tend to perform well according to the definitions of success built into their derivations. For identification of the true model, the key concept is *consistency*, which here means that the probability of selecting the true model from a set of candidates tends to 1 as n increases if the true model is one of the models under consideration and the true parameter value θ^* remains fixed. BIC is typically a consistent model selector (see, e.g., Shibata 1976; Nishii 1984). In general, a penalized criterion can only be consistent if its penalty term is a fast enough increasing function of n (Hannan and Quinn 1979). Thus, AIC is not consistent, as it always has some probability of selecting models that are too large (note, however, that in finite samples, adjusted versions of it can behave much better in this respect; see, e.g., Hurvich and Tsai 1989). This result is a compelling argument for the use of consistent criteria, if we believe in the existence of a simple true model, but almost irrelevant otherwise.

The most general large-sample results on predictive performance concern *asymptotic efficiency* of model selection criteria in cases when the dimensionality of the true model is infinite or increases with the sample size. Such a condition is not as far-fetched as it may seem but arguably represents most real situations. The processes that generate a set of observed data can rarely be represented *exactly* by any fixed finite-dimensional model. For example, each individual observation may follow a slightly different model with some specific features. Even if the true model were actually finite, it might not correspond to any finite version of the candidate models (such as when the candidates are polynomial regression models and the true model is of some other nonlinear form, which can only be represented exactly by a polynomial of an infinite degree). In such cases, a model selection criterion is asymptotically efficient if the expected mean squared error of predictions from models selected by it is the smallest

possible in large samples. AIC is asymptotically efficient in this sense, while BIC is not (Shibata 1981).

Asymptotic results do not necessarily give a full picture of the performance of the criteria in small samples, so simulation studies are also useful. Here we conduct them for the two examples considered earlier, applying both criteria of success to both AIC and BIC. It should be noted, however, that as the choices of settings for model comparison are infinite, the results are at best suggestive of more general findings. Much larger simulations are needed to assess the performance of selection criteria more broadly. One impressive study of this kind was carried out by McQuarrie and Tsai (1998), who considered several model selection criteria for a wide range of models and types of data. In most of their simulations, the true model was included in the candidates, so BIC outperformed AIC. The sample sizes were mostly small, so a small-sample adjustment of AIC also did well, sometimes even better than BIC and most of the other criteria. As expected, AIC performed best in large-sample simulations in which the true model was not included.

EXAMPLE 1: NORMAL LINEAR REGRESSION (CONTINUED)

Table 4 shows results of a simulation study in which $n = 100$ or $n = 1,000$ values of x_i were first generated from a uniform distribution with values between –1 and 1. Keeping these x_i fixed, 100,000 sets of y_i were then generated from a normal distribution with mean $\mu_i = \alpha + \psi x_i + \phi x_i^2$ and variance $\sigma^2 = 1$. Six choices for $\theta^* = (\alpha, \psi, \phi)$ were considered, as shown in the first column of Table 4. Cases I to III correspond to Models I to III in the log-linear Example 2. In Case I, $\mu_i = 0$ is constant, so model M_1 is the true model. In Cases II and III, $\mu_i = \psi x_i$, so that the linear model M_2 is true. In Case II, $\psi = 2/\sqrt{n}$, so that true model is also relatively close to M_1, while in Case III, $\psi = 1$, and the difference between the models is clear. In the last three cases, the true model is neither M_1 nor M_2 but a quadratic in x_i, with mean of the form $\mu_i = x_i^2 + \psi x_i - 1$. This means that the values of μ_i at –1 and 1 are $-\psi$ and ψ, respectively, so ψ effectively determines how much linear trend there is in μ_i over the range of x_i in these data. In Case IV, $\psi = 0$, so that there is curvature but no linear trend, while the trend is small in Case V and large in Case VI.

TABLE 4: Simulation Results of Model Choice in the Normal Linear Example 1, With 100,000 Simulated Data Sets for Each Setting

		Percentage Selecting True Model				*Percentage Selecting Better Prediction*					
		n = *100*		n = *1,000*		n = *100*			n = *1,000*		
Parameters	*True*	*AIC*	*BIC*	*AIC*	*BIC*	T_A	*AIC*	*BIC*	T_A	*AIC*	*BIC*
I (0, 0, 0)	M_1	84.3	96.8	84.3	99.1	−1.00	84.3	96.8	−1.00	84.3	99.1
II (0, $2n^{-1/2}$, 0)	M_2	39.4	15.8	40.3	7.2	0.30	39.4	15.8	0.35	40.3	7.2
III (0, 1, 0)	M_2	100	100	100	100	30.88	100	100	339.52	100	100
IV (−1, 0, 1)						−0.99	84.1	96.8	−0.92	82.5	98.9
V (−1, $2n^{-1/2}$, 1)						0.70	46.2	20.4	0.52	43.1	8.1
VI (−1, 1, 1)						33.15	100	100	360.39	100	100

NOTE: "Parameters" shows the parameters (α, ψ, ϕ) in the true model $\mu_i = E(y_i|x_i) = \alpha + \psi x_i + \phi x_i^2$ (see the text for further details of the models). The middle block of columns shows the proportion of simulations in which Akaike's information criterion (AIC) or the Bayesian information criterion (BIC) selected the true model if it is one of the models considered. The right-hand block of columns shows the proportion of simulations in which the model with the better expected predictive performance (as measured by T_A) was chosen.

The table shows two kinds of proportions of agreement over the 100,000 simulations. For Cases I to III, in which one of the two models is correct, the proportion in which AIC or BIC favors the correct model is given. Also shown are proportions of the samples in which the criteria select the model favored by T_A. The true value of T_A for each model is also shown; negative values of T_A indicate that the smaller model M_1 gives better expected predictions (as measured by T_A) and vice versa. When either M_1 or M_2 is the true model, the same model is here also the one favored by T_A. As the results in Table 4 are broadly similar to the ones in the next example, they are discussed together below.

EXAMPLE 2: LOG-LINEAR MODELS FOR A THREE-WAY TABLE (CONTINUED)

Table 5 shows the results of a similar simulation in the log-linear example. The models considered are the same as in Table 3. Here we can see that in the case of true Model II, the model with the slightly better expected prediction is actually not the true model M_2, a somewhat counterintuitive result that is possible in finite samples.

Considering the results in Tables 4 and 5, we focus on performance in identifying the model favored by T_A. In the one case (Model II in Table 5), in which this differed from an available true model, both criteria still tended to favor the model with smaller T_A (note that this does not violate the consistency of BIC, which does not hold when the difference between the models is of order $O(n^{-1/2})$). Both criteria perform well when the difference between the models is clear. When the smaller model is clearly the better (Cases I and IV in Table 4; Cases I, II, and IV in Table 5), BIC selects it almost always and AIC with only its inevitable error rate (here about 15.7 percent but less if degrees of freedom are larger). When the larger model M_2 is clearly the better predictor (Models III and VI in Table 4 and Models III and V with a larger sample and even data in Table 5), both criteria again select that model essentially always. However, the preference for M_2 must be much stronger for this to happen than the corresponding preference for M_1. In the rest of the simulations, the models are very close, or M_2 is only moderately better as a predictor. Both criteria perform rather badly in these cases. The difficulty is, of course, not surprising, and the conclusion in practice would most likely be that

TABLE 5: Simulation Results of Model Choice in the Log-Linear Example 2, With 100,000 Simulated Data Sets for Each Setting

			Percentage Selecting True Model				*Percentage Selecting Better Prediction*					
			n = *100*		n = *1,000*		n = *100*			n = *1,000*		
Model	*Data*	*True*	*AIC*	*BIC*	*AIC*	*BIC*	T_A	*AIC*	*BIC*	T_A	*AIC*	*BIC*
I	Even	M_1	84.4	96.6	84.4	99.2	−1.08	84.4	96.6	−1.01	84.4	99.2
	Sparse		85.6	97.7	84.2	99.1	−1.03	85.6	97.7	−1.04	84.2	99.1
II	Even	M_2	20.8	5.7	21.0	1.7	−0.82	79.3	94.3	−0.76	79.0	98.3
	Sparse		15.7	3.4	17.7	1.2	−0.97	84.3	96.6	−0.95	82.3	98.8
III	Even	M_2	81.1	56.3	100	100	4.14	81.1	56.3	51.61	100	100
	Sparse		38.9	16.2	91.8	62.4	0.04	38.9	16.2	7.99	91.8	62.4
IV	Even						−1.11	84.0	96.7	−1.03	84.4	99.1
	Sparse						−0.75	89.8	98.1	−1.07	84.0	99.1
V	Even						1.13	53.6	26.1	21.55	100	98.3
	Sparse						−0.07	64.9	86.2	6.82	89.3	55.6

NOTE: The simulation settings are the same as in Table 3, and the entries are as in Table 4.

the data do not provide enough information to distinguish between the models.

In the next section, we will consider the implications of the simple observation that the choice is easiest when the two criteria agree. In such cases, the selection will be wrong only when both criteria are wrong. Suppose first that the smaller model M_1 should in fact be preferred. Then AIC and BIC will incorrectly agree only if even BIC selects too large a model. In the small examples considered above, as well as more generally, this seems to be quite rare. If, on the other hand, the larger model should be selected but both prefer the smaller one, even AIC has selected too small a model. This occurred more frequently in Tables 4 and 5 but only in cases when the two models were very similar. This suggests, tentatively, that finding a model that is favored by both criteria is fairly reassuring. The selection is then likely to be best of the candidates or not much worse in terms of predictive performance. Results are less clear-cut when the criteria disagree on the best model. Both criteria may recommend models that could reasonably be regarded as less good than available alternatives, and we cannot in general conclude that one or the other is consistently more likely to be correct. When they do err, broadly speaking, AIC tends to favor models that are too large and BIC models that are too small. Thus, an optimistic interpretation of these results is that even a disagreement at least suggests bounds for the range of acceptable models.

6. THE SOCIAL MOBILITY EXAMPLES REVISITED

We will finish by returning to the social mobility example introduced in Section 1, using it to illustrate the implications of what can be learned from using AIC and BIC together. The key question of model choice in these data sets concerns the association (*OD*) between origin class and destination class and whether this varies between nations (simpler models of conditional independence and the strict LZ hypothesis are rejected by all criteria, and AIC and BIC agree that allowing for between-cohort variation in *OD* in the EG data is not necessary). Among the basic log-linear models, the relevant ones are those (HG.3 and EG.3) in which the *OD* association

parameters are the same in all nations and those (HG.4 and EG.4) that allow unrestricted variation in them across nations. Models between these two extremes can also be defined. In particular, we can consider models in which the pattern of *OD* associations is more parsimonious than in the unrestricted interaction model for each country, but the parameters of this pattern are allowed to vary between countries. Erikson and Goldthorpe (1992b) have argued that this is an expression of the FJH hypothesis in a less strict form and one more in keeping with its original formulation, which describes "basic" (rather than exact) similarity in patterns of mobility across nations.

Four such patterned association models for the HG data are shown in Table 1. Two of them were proposed by Grusky and Hauser (1984): a quasi-symmetry model (HG.5) and a model (HG.6) in which the parameters of the quasi-symmetry association, rather than varying freely across nations, depend on five explanatory variables. The remaining two models are modifications of the quasi-symmetry model suggested by Weakliem (1999) to allow for more flexible associations. Model HG.7 introduces a single additional asymmetry parameter, which is the same for all nations, while HG.8 adds a second asymmetry parameter related to farm inheritance.

Two additional models are considered for the EG data. These are motivated by models proposed by Erikson and Goldthorpe (1992b). They first defined what they called their model of core social fluidity. This is a topological model incorporating effects related to hierarchy, inheritance, sector, and affinity in the class structure. This was then extended by introducing national variations (based on country-specific considerations and inspection of residuals from the core model) in which some of the levels of the topological models are defined slightly differently for individual countries. Models EG.8 and EG.9 are variants of the core model and the national variant model, respectively. They are, however, not identical to the models by EG, who defined the models for a seven-class version of their class schema rather than five as used here. The models considered here have been modified by assigning level parameters to the combined categories in a somewhat ad hoc way motivated by the original models.[3]

As noted above, model choice is easiest when AIC and BIC agree on the preferred model. This is then unlikely to be far from the best of the candidate models. Such a choice is also very robust in the theoretical

context of both AIC and BIC. For example, interpreting both BIC and AIC as approximate Bayes factors (cf. Section 4), agreement between them implies that the choice is insensitive to quite dramatic changes in the informativeness of the priors of the model parameters.

When the two criteria do not agree on the preferred model, one conclusion that could be drawn is that it is worthwhile pursuing model selection further. This parallels comments by Raftery (1995), who argues that model search should continue when there is a large discrepancy between models deemed appropriate by BIC and standard significance tests. In large samples, AIC often narrows this range of models if it favors smaller models than significance tests. Further search may then uncover a model preferred by both criteria. Even when such a model is not found, the comparisons will usually rule out many candidates. In particular, the search will often suggest boundaries for the set of acceptable models, with BIC indicating the smallest acceptable models and AIC the largest models that need to be considered further.

For the HG data, conclusions about the extended models HG.5 to HG.7 are discrepant, with only BIC favoring them over the saturated model. The farm inheritance asymmetry model HG.8, on the other hand, gives the smallest value of both BIC and (just barely) AIC (note also that the p value for its deviance is 0.026, so the model is borderline adequate even according to a goodness-of-fit test). So a model acceptable to both criteria is found in this example. The results suggest that the data can be represented by a nonsaturated model, as long as it allows for enough asymmetry in the mobility patterns. This agrees with the less strict interpretation of the FJH hypothesis.

For the EG data, there are several pairs of comparisons in which AIC and BIC prefer the same model. If we use this to rule out some models, three candidates remain. BIC prefers the strict FJH model EG.3 and AIC model EG.4, which specifies unconstrained cross-national variation in mobility rates. The results are thus not conclusive. Choosing one of EG.3 and EG.4 would imply a claim that AIC has here overfitted the data or that BIC has overpenalized complexity and recommended too simple a model. For example, it might be argued, following the comments by Weakliem (1999) discussed in Section 2, that the priors implied by BIC are here too weak. The additional parameters included in EG.4 but not in EG.3 correspond

to *OD* associations estimated separately for each nation. Thus, the number of observations that contribute information about each such parameter is not n but, at most, the sample size n_j for a single nation. This implies that the unit information prior is roughly comparable to information from n_j/n rather than $n_0 = n/n = 1$ observations. Here these numbers range from 0.02 to 0.37, which implies rather less prior information than the straightforward interpretation of the unit information prior.

The national variant model EG.9 is intermediate between EG.3 and EG.4. It is the second best of the candidates according to both AIC and BIC and preferable to the saturated model according to both of them. It could thus be regarded as a reasonable compromise between the other two models. The remaining disagreement could also again be regarded as encouragement to continue the search for better fitting parsimonious models for the mobility rates. For example, the version of the national variant model used here is a rather crude modification of Erikson and Goldthorpe's model to the five-class case, and a more theoretically informed specification of such a model would quite possibly improve the fit (for an example of such a modification—in that case, from 7 classes to 12—see Erikson and Goldthorpe 1992a).

7. CONCLUSIONS

Both AIC and BIC provide well-founded and self-contained approaches to model comparison, although with different motivations and partially different objectives. Both are typically (although not invariably) good approximations of their own theoretical target quantities. Often, this also means that they will identify good models for observed data, but both criteria can still fail in this respect, even in the very simple examples considered above. Because of this, an approach of using the two criteria together (as well as significance tests and perhaps other indices of model fit) has been advocated here. When the criteria agree on the best model, this provides reassurance on the robustness of the choice. Even disagreement usually rules out many models and provides bounds for the set of adequate models while also suggesting that the model search should continue.

Relying on only one information criterion for model selection would produce apparently more conclusive results than using multiple ones, as each criterion is designed to identify the candidate model estimated to be the best in a particular well-defined sense. This is a very desirable property if we are convinced that one criterion is clearly the most appropriate and its performance consistently at least as good as that of other methods. Here, however, no such claims have been made. As described above, there are theoretical results showing that model selection based on one or the other of the criteria is optimal under some conditions. Different conditions, however, lead to different conclusions, and arguably none of them captures the full complexity of real model selection problems. For example, the decision on which criterion to favor depends strongly on what we think of the nature of true models for variables considered in the social sciences. Perhaps many of them are close to simple models with a fairly small number of strong effects, as assumed by the Bayes factor approach. Or perhaps they are instead mostly collections of many effects, maybe changing over time or across individuals, which would emphasize the importance of a prediction-based approach. Answers to such questions can only come over time, from repeated studies on the same topics. This provides an independent assessment of the relative merits of different methods of model selection, by allowing us to evaluate the reproducibility of selected models and the accuracy of predictions from them. When such information is not available, it seems prudent to adopt a cautious attitude that makes use of a battery of selection methods instead of relying too much on any one of them.

NOTES

1. For example, consider the approximate Bayes factor (2.4) for the case of two nested models with $p_2 - p_1 = 1$. The first term is the likelihood ratio test for the point null hypothesis $\psi = \psi_0$. Denote its observed value by G^2_{12}. When this is large enough, the conclusion from the test would be to reject the null hypothesis in favor of the unrestricted alternative $\psi \neq \psi_0$. However, for any value of G^2_{12}, the Bayes factor would instead favor the smaller model if $n/n_0 > \exp(G^2_{12}) - 1$ (i.e., if n was large enough or n_0 small enough). This is an example of the so-called "Lindley's paradox" (see Lindley 1957; Shafer 1982, as well as the associated discussion). It does not occur if n_0 is proportional to n. The result is closely related to the reasons for the apparent preference for parsimonious models by penalized criteria such as the Bayesian information criterion (BIC), as discussed in Section 5.

2. An outline of the proof of this result is available from the author upon request.

3. Specifically, for the class containing Classes I to III in the five-class schema, the parameters for Classes I and II in the seven-class topological model were used, and parameters for Class VIIb were used for the combined class of IVc and VIIb.

REFERENCES

Akaike, H. 1973. "Information Theory and an Extension of the Maximum Likelihood Principle." Pp. 267–81 in *Second International Symposium on Information Theory*, edited by B. N. Petrov and F. Csaki. Budapest: Akademiai Kaido. (Reprinted, with an introduction by J. deLeeuw, in *Breakthroughs in Statistics, Volume I*, edited by Samuel Kotz and Norman L. Johnson. New York: Springer, 1992, pp. 599–624.)

Atkinson, A. C. 1981. "Likelihood Ratios, Posterior Odds and Information Criteria." *Journal of Econometrics* 16:15–20.

Burnham, K. P. and D. R. Anderson. 1998. *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York: Springer.

Carlin, B. P. and T. A. Louis. 1996. *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall.

Chow, G. C. 1981. "A Comparison of the Information and Posterior Probability Criteria for Model Selection." *Journal of Econometrics* 16:21–33.

Cox, D. R. 1995. "The Relation Between Theory and Application in Statistics (With Discussion)." *Test* 4:207–61.

Cox, D. R. and D. V. Hinkley. 1978. *Problems and Solutions in Theoretical Statistics*. London: Chapman & Hall.

Erikson, R. and J. H. Goldthorpe. 1992a. "The CASMIN Project and the American Dream." *European Sociological Review* 8:283–305.

———. 1992b. *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Oxford, UK: Oxford University Press.

Featherman, D. L., F. L. Jones, and R. M. Hauser. 1975. "Assumptions of Social Mobility Research in the U.S.: The Case of Occupational Status." *Social Science Research* 4:329–60.

Gelfand, A. E. and S. K. Ghosh. 1998. "Model Choice: A Minimum Posterior Predictive Loss Approach." *Biometrika* 85:1–11.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 1995. *Bayesian Data Analysis*. London: Chapman & Hall.

Grusky, D. B. and R. M. Hauser. 1984. "Comparative Social Mobility Revisited: Models of Convergence and Divergence in 16 Countries." *American Sociological Review* 49:19–38.

Hannan, E. J. and B. G. Quinn. 1979. "The Determination of the Order of an Autoregression." *Journal of the Royal Statistical Society B* 41:190–95.

Hazelrigg, L. E. and M. A. Garnier. 1976. "Occupational Mobility in Industrial Societies: A Comparative Analysis of Differential Access to Occupational Ranks in Seventeen Countries." *American Sociological Review* 41:498–511.

Hurvich, C. M. and C.-L. Tsai. 1989. "Regression and Time Series Model Selection in Small Samples." *Biometrika* 76:297–307.

Ishiguro, M., Y. Sakamoto, and G. Kitagawa. 1997. "Bootstrapping Log Likelihood and EIC, an Extension of AIC." *Annals of the Institute of Statistical Mathematics* 49:411–34.

Jeffreys, H. 1961. *Theory of Probability*. 3d ed. Oxford, UK: Oxford University Press.

Kass, R. E. and A. E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90:773–95.

Kass, R. E. and S. K. Vaidyanathan. 1992. "Approximate Bayes Factors and Orthogonal Parameters, With Application to Testing Equality of Two Binomial Proportions." *Journal of the Royal Statistical Society B* 54:129–44.

Kass, R. E. and L. Wasserman. 1995. "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion." *Journal of the American Statistical Association* 90:928–34.

Laud, P. W. and J. G. Ibrahim. 1995. "Predictive Model Selection." *Journal of the Royal Statistical Society B* 57:247–62.

Lindley, D. V. 1957. "A Statistical Paradox." *Biometrika* 44:187–92.

Lipset, S. M. and H. L. Zetterberg. 1959. "Social Mobility in Industrial Societies." Pp. 11–75 in *Social Mobility in Industrial Society*, edited by S. M. Lipset and R. Bendix. Berkeley: University of California Press.

McQuarrie, A. D. R. and C.-L. Tsai. 1998. *Regression and Time Series Model Selection*. Singapore: World Scientific.

Miller, A. J. 2002. *Subset Selection in Regression*. 2d ed. London: Chapman & Hall.

Murata, N., S. Yoshizawa, and S. Amari. 1991. "A Criterion for Determining the Number of Parameters in an Artificial Neural Network Model." Pp. 9–14 in *Artificial Neural Networks: Proceedings of ICANN-91*, vol. 1, edited by T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas. Amsterdam: North Holland.

Newton, M. A. and A. E. Raftery. 1994. "Approximate Bayesian Inference With the Weighted Likelihood Bootstrap (With Discussion)." *Journal of the Royal Statistical Society B* 56:3–48.

Nishii, R. 1984. "Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression." *Annals of Statistics* 12:758–65.

O'Hagan, A. 1995. "Fractional Bayes Factors for Model Comparison (With Discussion)." *Journal of the Royal Statistical Society B* 57:99–138.

Raftery, A. E. 1986. "Choosing Models for Cross-Classifications." *American Sociological Review* 51:145–46.

———. 1995. "Bayesian Model Selection in Social Research (With Discussion)." Pp. 111–63 in *Sociological Methodology 1995*, edited by P. V. Marsden. Cambridge, MA: Blackwell.

Schwarz, G. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6:461–64.

Shafer, G. 1982. "Lindley's Paradox (With Discussion)." *Journal of the American Statistical Association* 77:325–51.

Shibata, R. 1976. "Selection of the Order of an Autoregressive Model by Akaike's Information Criterion." *Biometrika* 63:117–26.

———. 1981. "An Optimal Selection of Regression Variables." *Biometrika* 68:45–54.

Smith, A. F. M. and D. J. Spiegelhalter. 1980. "Bayes Factors and Choice Criteria for Linear Models." *Journal of the Royal Statistical Society B* 42:213–20.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde. 2002. "Bayesian Measures of Model Complexity and Fit (With Discussion)." *Journal of the Royal Statistical Society B* 64:583–639.

Stone, M. 1977. "An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion." *Journal of the Royal Statistical Society B* 39:44–47.

———. 1979. "Comments on Model Selection Criteria of Akaike and Schwartz." *Journal of the Royal Statistical Society B* 41:276–78.

Sugiura, N. 1978. "Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections." *Communications in Statistics A (Theory and Methods)* A7:13–26.

Takeuchi, K. 1976. "Distribution of Informational Statistics and a Criterion of Model Fitting." *Suri-Kagaku* (*Mathematical Sciences*) 153:12–18 (in Japanese).
Tierney, L. and J. B. Kadane. 1986. "Accurate Approximations for Posterior Moments and Marginal Densities." *Journal of the American Statistical Association* 81:82–86.
Weakliem, D. L. 1999. "A Critique of the Bayesian Information Criterion for Model Selection." *Sociological Methods & Research* 27:359–97.
Wei, C. Z. 1992. "On Predictive Least Squares Principles." *Annals of Statistics* 20:1–42.
Xie, Y. 1992. "The Log-Multiplicative Layer Effect Model for Comparing Mobility Tables." *American Sociological Review* 57:380–95.

Jouni Kuha is a lecturer in statistics and research methodology at the London School of Economics. His research interests include statistical analysis of problems involving measurement error and missing data, statistical model selection, and statistical methodology in the social sciences. Recent publications include A Nonparametric Approach to Matched Pairs With Missing Data *and* Covariate Measurement Error in Quadratic Regression.