A Comparison of Bayes Factor Approximation Methods Including Two New Methods

March 10, 2011

Abstract

Bayes Factors play an important role in comparing the fit of models ranging from multiple regression to mixture models. Full Bayesian analysis calculates a Bayes Factor from an explicit prior distribution. However, computational limitations or lack of an appropriate prior sometimes prevent researchers from using an exact Bayes Factor. Instead, it is approximated, often using Schwarz's (1978) Bayesian Information Criterion (BIC), or a variant of the BIC. In this paper we provide a comparison of several Bayes Factor approximations, including two new approximations, the SPBIC and IBIC. The SPBIC is justified by using a scaled unit information prior distribution that is more general than the BIC's unit information prior, and the IBIC approximation utilizes more terms of approximation than in the BIC. In a simulation study we show that several measures perform well in large samples, that performance declines in smaller samples, and that SPBIC and IBIC can provide improvement to existing measures under some conditions, including small sample sizes. We then illustrate the use of the fit measures in an empirical example from the crime data of Ehrlich (1973). We conclude with recommendations for researchers.

Keywords: Bayes Factor · Empirical Bayes · Information Criterion · Laplace Approximation · Model Selection · Scaled Unit Information Prior · Variable Selection

1 Introduction

Bayes Factors play an increasingly important role in comparing fit in a variety of statistical models ranging from multiple regression (Spiegelhalter and Smith 1982; Carlin and Chib 1995) to mixture models (Richardson and Green 1997). Under a full Bayesian analysis it is possible to calculate the exact Bayes Factor that derives from an explicit prior distribution. However, in practice we often develop model selection criteria based on approximations of Bayes Factors, either because of computational limitations or due to difficulty of specifying reasonable priors. The most popular choice in this class of model selection criteria is the Schwarz (1978) Bayesian Information Criterion (BIC). The BIC's popularity is understandable in that it has several desirable features. First, it is readily calculable using only the values of the likelihood function, the model degrees of freedom, and the sample size. Second, the BIC permits a researcher to compare the fit of nested and non-nested models (Raftery 1995), which is not typical using other traditional procedures such as Likelihood Ratio tests. The researcher can use the BIC to calculate the strength of evidence in favor of particular models relative to other models using guidelines such as those of Jeffreys (1939), as updated by Kass and Raftery (1995) and Raftery (1995).

These valuable aspects depend on the accuracy of the BIC as an approximation to a Bayes Factor that will aid the selection of the best model. The justification for the BIC as an approximation rests either on an implicit unit information prior distribution or on large sample properties and the dominance of terms not involving the prior distribution for the parameters (Raftery 1995). Furthermore, the sample size, N, that enters the calculation of BIC is ambiguous in some situations. For example, the appropriate N in multilevel (hierarchical) models and log-linear models is not always clear (Kass and Raftery 1995). Censoring in survival analysis and overdispersion in clustered survey data also raise the problem of how to define the effective sample size (Fitzmaurice 1997; Volinsky and Raftery 2000).

Interest in BIC and alternatives to it among sociological methodologists runs high. For example, see Raftery (1995) and Weakliem (1999) and their discussions for highly informative debates on the merits and limits of the BIC in sociological research. In practice, the BIC has become a standard means by which sociologists evaluate and compare competing models. However, alternative approximations are possible and potentially superior. Few studies provide a comprehensive comparison of BIC and its alternatives.

Our paper has several purposes. First, we present several varieties of Bayes Factor approximations, including two new approximations—the SPBIC and the IBIC—which are based on modifications of two different theoretical derivations of the standard BIC. Second, we evaluate these approximations as a group: under what conditions is Bayes Factor approximation, of whatever variety, appropriate? A third goal is to assess whether the choice of approximation matters: do all BF approximations perform essentially the same, or are there conditions under which we should prefer some over others? Relative performance is assessed using an extensive simulation study of regression modeling across a range of conditions common in social science data, as well as via an empirical example using Ehrlich's (1973) well-known crime data. Our ultimate aim is to contribute toward a more informed consensus around the use of Bayes Factor approximations in sociological research.

The next section of the paper reviews the concept of the Bayes Factor. Section 3 derives the BIC using a special prior distributions, and proposes the SPBIC using a modified prior. Section 4 provides an alternative justification for the BIC based on a large sample rationale, and then presents two variants: the existing HBIC, and a new approximation, the IBIC. In Section 5 we provide a simple example showing the steps in calculating the various approximations. The simulation study comparing the fit indices and an empirical example are given in Section 6, which is followed by our conclusions.

2 Bayes Factor

In this section we briefly describe the Bayes Factor and the Laplace approximation. Readers familiar with this derivation may skip this section. For a strict definition of the Bayes Factor we have to appeal to the Bayesian model selection literature. Thorough discussions on Bayesian statistics in general and model selection in particular are in Gelman et al. (1995), Carlin and Louis (1996), and Kuha (2004). Here we provide a very simple description.

We start with data $\boldsymbol{Y}$ and hypothesize that they were generated by either of two competing models M_1 and M_2.[1] Moreover, let us assume that we have prior beliefs about the data generating models M_1 and M_2, so that we can specify the probability of data coming from model M_1, $P(M_1) = 1 - P(M_2)$. Now, using Bayes' theorem, the ratio of the posterior probabilities is given by

$$\frac{P(M_1|\boldsymbol{Y})}{P(M_2|\boldsymbol{Y})} = \frac{P(\boldsymbol{Y}|M_1)}{P(\boldsymbol{Y}|M_2)} \frac{P(M_1)}{P(M_2)}.$$

The left side of the equation can be interpreted as the posterior odds of M_1 versus M_2, and it represents how the prior odds $P(M_1)/P(M_2)$ is updated by the observed data $\boldsymbol{Y}$. The factor $P(\boldsymbol{Y}|M_1)/P(\boldsymbol{Y}|M_2)$ that is multiplied to obtain the posterior from the prior is known as the Bayes Factor, which we will denote as

$$B_{12} = \frac{P(\boldsymbol{Y}|M_1)}{P(\boldsymbol{Y}|M_2)}. \tag{1}$$

Typically, we will deal with parametric models M_k, which are described through model parameters, such as $\boldsymbol{\theta}_k$. So the marginal likelihoods $P(\boldsymbol{Y}|M_k)$ are evaluated using

$$P(\boldsymbol{Y}|M_k) = \int P(\boldsymbol{Y}|\boldsymbol{\theta}_k, M_k)P(\theta_k|\ M_k)d\theta_k, \tag{2}$$

where $P(\boldsymbol{Y}|\theta_k, M_k)$ is the likelihood under M_k, and $P(\theta_k|M_k)$ is the prior distribution of θ_k. A closed form analytical expression of the marginal likelihoods is difficult to obtain, even with completely specified priors (unless they are conjugate priors). Moreover, in most cases $\boldsymbol{\theta}_k$ will be high-dimensional and a direct numerical integration will be computationally intensive, if not impossible. For this reason we turn to a way to approximate this integral given in (2).

2.1 Laplace Method of Approximation

The Laplace Method of approximating Bayes Factors is a common device (see Raftery 1993; Weakliem 1999; Kuha 2004). Since we also make use of the Laplace method for our proposed approximations, we will briefly review it here.

[1] Generalization to more than two competing models is achieved by comparing the Bayes Factor of all models to a "base" model and choosing the model with highest Bayes Factor

The Laplace method of approximating $\int P(\boldsymbol{Y}|\boldsymbol{\theta}_k, M_k)P(\theta_k|\ M_k)d\theta_k \ = \ P(\boldsymbol{Y}|M_k)$ uses a second order Taylor series expansion of the $\log P(\boldsymbol{Y}|M_k)$ around $\widetilde{\boldsymbol{\theta}}_k$, the posterior mode. It can be shown that

$$\log P(\boldsymbol{Y}|M_k) = \log P(\boldsymbol{Y}|\widetilde{\boldsymbol{\theta}}_k, M_k) + \log P(\widetilde{\boldsymbol{\theta}}_k|M_k) + \frac{d_k}{2}\log(2\pi) - \frac{1}{2}\log|-H(\widetilde{\boldsymbol{\theta}}_k)| + O(N^{-1}), \quad (3)$$

where d_k is the number of distinct parameters estimated in model M_k, π is the usual mathematical constant, $H(\widetilde{\boldsymbol{\theta}}_k)$ is the Hessian matrix

$$\frac{\partial^2 \log[P(\boldsymbol{Y}|\boldsymbol{\theta_k}, M_k)P(\theta_{\mathbf{k}}|M_k)]}{\partial\theta_{\mathbf{k}}\partial\theta_{\mathbf{k}}'},$$

evaluated at $\boldsymbol{\theta_k} = \widetilde{\boldsymbol{\theta}}_k$.and $O(N^{-1})$ is the "big-O" notation indicating that the last term is bounded in probability from above by a constant times N^{-1} (see, for example, Tierney and Kadane 1986; Kass and Vaidyanathan 1992; Kass and Raftery 1995).

Instead of the Hessian matrix we will be working with different forms of information matrices. Specifically, the observed information matrix is denoted by

$$I_O(\theta) = -\frac{\partial^2 \log[P(\boldsymbol{Y}|\boldsymbol{\theta_k}, M_k)]}{\partial\theta_{\mathbf{k}}\partial\theta_{\mathbf{k}}'},$$

whereas the expected information matrix I_E is given by

$$I_E(\theta) = -E\left[\frac{\partial^2 \log[P(\boldsymbol{Y}|\boldsymbol{\theta_k}, M_k)]}{\partial\theta_{\mathbf{k}}\partial\theta_{\mathbf{k}}'}\right].$$

Let us also define the average information matrices $\bar{I}_E \ = \ I_E/N$ and $\bar{I}_O \ = \ I_O/N$. If the observations come from an i.i.d. distribution we have the following identity

$$\bar{I}_E = -E\left[\frac{\partial^2 \log[P(Y_i|\boldsymbol{\theta_k}, M_k)]}{\partial\theta_{\mathbf{k}}\partial\theta_{\mathbf{k}}'}\right],$$

the expected information for a single observation. We will make use of this property later.

In large samples the Maximum Likelihood (ML) estimator, $\widehat{\theta}_k$, will generally be a reasonable approximation of the posterior mode, $\widetilde{\boldsymbol{\theta}}_k$. If the prior is not that informative relative to the likelihood, we can write

$$\log P(\boldsymbol{Y}|M_k) = \log P(\boldsymbol{Y}|\widehat{\theta}_k, M_k) + \log P(\widehat{\theta}_k|M_k) + \frac{d_k}{2}\log(2\pi) - \frac{1}{2}\log|I_O(\widehat{\theta}_k)| + O(N^{-1}), \quad (4)$$

where $\log P(\boldsymbol{Y}|\widehat{\theta}_k, M_k)$ is the log likelihood for M_k evaluated at $\widehat{\theta}_k$, $I_O(\widehat{\theta}_k)$ is the observed information matrix evaluated at $\widehat{\theta}_k$ (Kass and Raftery 1995) and $O(N^{-1})$ is the "big-O" notation that refers to a term bounded in probability to be some constant multiplied by N^{-1}. If the expected information matrix, $I_E(\widehat{\theta}_k)$, is used in place of $I_O(\widehat{\theta}_k)$, then the approximation error is $O(N^{-1/2})$ and we can write

$$\log P(\boldsymbol{Y}|M_k) = \log P(\boldsymbol{Y}|\widehat{\theta}_k, M_k) + \log P(\widehat{\theta}_k|M_k) + \frac{d_k}{2}\log(2\pi) - \frac{1}{2}\log\left|I_E(\widehat{\theta}_k)\right| + O(N^{-1/2}) \quad (5)$$

If we now make use of the definition, $\bar{I}_E = I_E/N$ and substitute it in (5), we have

$$\log P(\boldsymbol{Y}|M_k) = \log P(\boldsymbol{Y}|\widehat{\theta}_k, M_k) + \log P(\widehat{\theta}_k|M_k) + \frac{d_k}{2}\log(2\pi) - \frac{d_k}{2}\log(N) - \frac{1}{2}\log\left|\bar{I}_E(\widehat{\theta}_k)\right| + O(N^{-1/2}). \quad (6)$$

This equation forms the basis of several approximations to Bayes Factors, which we discuss below.

3 Approximation through Special Prior Distributions

The BIC has been justified by making use of a unit information prior distribution of θ_k. Our first new approximation, SPBIC, builds on this rationale using a more flexible prior. We first derive the BIC, then present our proposed modification.

3.1 BIC: The Unit Information Prior Distribution

We assume for model M_K (the full model, or the largest model under consideration), the prior of $\boldsymbol{\theta}_K$ is given by a multivariate normal density

$$P(\boldsymbol{\theta}_K|M_K) \sim N\left(\boldsymbol{\theta}_K^*, \boldsymbol{V}_K^*\right) \tag{7}$$

where $\boldsymbol{\theta}_K^*$ and $\boldsymbol{V}_K^*$ are the prior mean and variance of θ_K. In the case of BIC we take $\boldsymbol{V}_K^* = \bar{I}_0^{-1}$, the average information defined in Section 3. This choice of variance can be interpreted as setting the prior to have approximately the same information as the information contained in a single data point.[2] This prior is thus referred to as the *Unit Information Prior* and is given by

$$P(\boldsymbol{\theta}_K|M_K) \sim N\left(\boldsymbol{\theta}_K^*, \left[\bar{I}_O(\hat{\boldsymbol{\theta}}_K)\right]^{-1}\right). \tag{8}$$

For any model M_k nested in the full model M_K, the corresponding prior $P(\theta_k|M_k)$ is simply the marginal distribution obtained by integrating out the parameters that do not appear in M_k. Using this prior in (4), we can show that

$$\begin{aligned} \log P(\boldsymbol{Y}|M_k) &\approx \log(P(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}_k, M_k)) - \frac{d_k}{2}\log(N) - \frac{1}{2N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T I_0(\hat{\boldsymbol{\theta}}_k)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ &\approx \log(P(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}_k, M_k)) - \frac{d_k}{2}\log(N). \end{aligned} \tag{9}$$

Note that the the error of approximation is still of the order $O(N^{-1})$. Comparing models M_1 and M_2 and multiplying by -2 gives the usual BIC

$$BIC = 2(l(\hat{\boldsymbol{\theta}}_2) - l(\hat{\boldsymbol{\theta}}_1)) - (d_2 - d_1)\log(N) \tag{10}$$

where we use the more compact notation of $l(\hat{\boldsymbol{\theta}}_k) \equiv \log(P(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}_k, M_k))$ for the log likelihood function at $\hat{\boldsymbol{\theta}}_k$.

[2] Note that unlike the expected information, even if the observations are i.i.d., the average observed information $\bar{I}_O(\widehat{\theta}_k)$, may not be equal to the observed information of any single observation.

An advantage of this unit information prior rationale for the BIC is that the error is of order $O(N^{-1})$ instead of $O(1)$ when we make no explicit assumption about the prior distribution. However, some critics argue that the unit information prior is too flat to reflect a realistic prior (e.g., Weakliem 1999).

3.2 SPBIC: The Scaled Unit Information Prior Distribution

The SPBIC, our first alternative approximation to the Bayes Factor, is similar to the BIC derivation except instead of using the unit information prior, we use a *scaled* unit information prior that allows the variance to differ and permits more flexibility in the prior probability specification. The scale is chosen by maximizing the marginal likelihood over the class of normal priors. Berger (1994), Fougere (1990), and others have used related concepts of partially informative priors obtained by maximizing entropy. In general, these maximizations involve tedious numerical calculations. We ease the calculation by employing components in the approximation that are available from standard computer output. We first provide a Bayesian interpretation of our choice of the scaled unit information prior and then provide an analytical expression for calculating this empirical Bayes prior.

We have already shown how the BIC can be interpreted as using a unit information prior for calculating the marginal probability in (4). This very prior leads to the common criticism of BIC being too conservative towards the null model, as the unit prior penalizes complex models too heavily. The prior of BIC is centered at $\boldsymbol{\theta}^*$ (parameter value at the null model) and has a variance scaled at the level of unit information. As a result this prior may put extremely low probability around alternative complex models (for discussion, see Weakliem 1999; Kuha 2004; Berger et al. 2006). One way to overcome this conservative nature of BIC is to use the ML estimate, $\hat{\theta}$, of the complex model M_k, as the center of the prior $P(\theta_k|M_k)$. But this prior specification puts all the mass around the complex model and thus favors the complex model heavily. A *scaled unit information prior* (henceforth denoted as SUIP) is a less aggressive way to be more favorable to a complex model. The prior is still centered around θ^*, so that the highest density is at θ^*, but we choose its scale, c_k, so that the prior flattens out to put more mass on the alternative models. This places higher prior probability than the unit information prior on the space of complex models, but at the same time prevents complete concentration of probability on the complex model.[3] This effectively leads to choosing a prior in the class of Bayes Factor-based information criterion that is most favorable to the model under consideration, placing the SUIP in the class of Empirical Bayes priors.

To obtain the analytical expression for the appropriate scale, we start from the marginal probability given in (2). Instead of using the more common form of the Laplace approximation applied jointly to the likelihood and the prior, we use the approximate normality of the likelihood, but choose the scale of the variance of the multivariate normal prior so as to maximize the marginal likelihood. As a result, our prior distribution is not fully specified by the information matrix; rather the variance is a scaled version of the variance of the unit information prior defined in (8). This prior has the ability to adjust itself through a scaling factor, which may differ significantly from unity in several situations (e.g., highly collinear covariates, dependent observations).

The main steps in constructing the SPBIC are as follows. First we define the scaled unit information prior as

$$P_k(\boldsymbol{\theta}_k) \sim N\left(\boldsymbol{\theta}_k^*, \frac{\left[\bar{I}_O(\hat{\boldsymbol{\theta}}_k)\right]^{-1}}{c_k}\right), \tag{11}$$

where $\bar{I}_O = \frac{1}{N} I_O$ and c_k is the scale factor which we will evaluate from the data. Note that c_k may vary across different models. Using the above prior in (11) and the generalized Laplace integral approximation described in Appendix A, we have

$$\log P(\boldsymbol{Y}|M_k) \approx l(\hat{\boldsymbol{\theta}}_k) - \frac{1}{2} d_k \log\left(1 + \frac{N}{c_k}\right) - \frac{1}{2}\frac{c_k}{N + c_k}\left[(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)^T I(\hat{\boldsymbol{\theta}}_k)(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)\right]. \tag{12}$$

Now we estimate c_k. Note that we can view (12) as a likelihood function in c_k, and we can write

[3] Specifically, the prior is still centered around θ^*, and the distribution being a normal, the highest density is still at θ^*, but we divide the variance $\bar{I}_O(\hat{\boldsymbol{\theta}})$ by the computed constant c_k (see (11)), which is ideally less than 1 for competing models favorable to the data. This adjustment inflates the variance, thus flattening out the density to put relatively more mass on a competing alternative model. While comparing a range of alternatives the scaled prior puts higher prior probability than the unit information prior on each complex model that conforms to the data, but at the same time prevents complete concentration of probability on those complex models.

$L(c_k|\boldsymbol{Y}, M_k) = \int L(\boldsymbol{\theta}_k)\pi(\boldsymbol{\theta}_k|c_k)d\boldsymbol{\theta}_k = L(\boldsymbol{Y}|M_k)$. The optimum c_k is given by its ML estimate, which can be obtained as follows. Note that there are two cases for estimating c_k, with the choice in a particular application depending on the scaling of the variance in the unit information prior.

$$\begin{aligned} 2l(c_k) &= -d_k \log\left(1+\frac{N}{c_k}\right) + 2l(\hat{\boldsymbol{\theta}}_k) - \frac{c_k}{N+c_k}\hat{\boldsymbol{\theta}}_k^T I_O(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k \\ \Longrightarrow \quad 2l'(c_k) &= d_k \frac{N}{c_k(N+c_k)} - \frac{\hat{\boldsymbol{\theta}}_k^T I_O(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k N}{(N+c_k)^2} \\ \Longrightarrow \quad 0 &= N(N+c_k)d_k - N\ c_k\ \hat{\boldsymbol{\theta}}_k^T I_O(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k \\ \Longrightarrow \quad \frac{\hat{c_k}}{N} &= \begin{cases} \frac{d_k}{\hat{\boldsymbol{\theta}}_k^T I_O(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k - d_k} & \text{if } d_k < \hat{\boldsymbol{\theta}}_k^T I_O(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k, \\ \infty & \text{if } d_k \geq \hat{\boldsymbol{\theta}}_k^T I_O(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k. \end{cases} \end{aligned}$$

Using this c_k we arrive at the SPBIC measures under the two cases.

Case 1: When $d_k < \hat{\boldsymbol{\theta}}_k^T I_O(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k$, we have

$$\begin{aligned} \log P(\boldsymbol{Y}|M_k) &= l(\hat{\boldsymbol{\theta}}_k) - \frac{d_k}{2} + \frac{d_k}{2}(\log d_k - \log(\hat{\boldsymbol{\theta}}_k^T I_O(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k)) \qquad (13) \\ \Longrightarrow SPBIC &= 2(l(\hat{\boldsymbol{\theta}}_2) - l(\hat{\boldsymbol{\theta}}_1)) - d_2\left(1 - \log\left[\frac{d_2}{\hat{\boldsymbol{\theta}}_2^T I_O(\hat{\boldsymbol{\theta}}_2)\hat{\boldsymbol{\theta}}_2}\right]\right) \\ &\quad + d_1\left(1 - \log\left[\frac{d_1}{\hat{\boldsymbol{\theta}}_1^T I_O(\hat{\boldsymbol{\theta}}_1)\hat{\boldsymbol{\theta}}_1}\right]\right) \qquad (14) \end{aligned}$$

Case 2: Alternatively, when $d_k \geq \hat{\boldsymbol{\theta}}_k^T I_O(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k$, as $c_k = \infty$, the prior variance goes to 0 so the prior distribution is a point mass at the mean, $\boldsymbol{\theta}^*$. Thus we have

$$\log P(\boldsymbol{Y}|M_k) = l(\hat{\boldsymbol{\theta}}_k) - \frac{1}{2}\hat{\boldsymbol{\theta}}_k^T I_O(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{\theta}}_k,$$

which gives

$$SPBIC = 2(l(\hat{\boldsymbol{\theta}}_2) - l(\hat{\boldsymbol{\theta}}_1)) - \hat{\boldsymbol{\theta}}_2^T I_O(\hat{\boldsymbol{\theta}}_2)\hat{\boldsymbol{\theta}}_2 + \hat{\boldsymbol{\theta}}_1^T I_O(\hat{\boldsymbol{\theta}}_1)\hat{\boldsymbol{\theta}}_1. \qquad (15)$$

4 Approximation through Eliminating $\mathrm{O}(1)$ Terms

An alternative derivation of the BIC does not require the specification of priors. Rather, it is justified by eliminating smaller order terms in the expansion of the marginal likelihood with arbitrary priors (see (6)). After providing the mathematical steps of deriving the BIC using (6), we show how approximations of the marginal likelihood with fewer assumptions can be obtained using readily available software output.

4.1 BIC: Elimination of Smaller Order Terms

One justification for the Schwarz (1978) BIC is based on the observation that the first and fourth terms in (6) have orders of approximations of $O(N)$ and $O(\log(N))$, respectively, while the second, third, and fifth terms are $O(1)$ or less. This suggests that the former terms will dominate the latter terms when N is large. Ignoring these latter terms, we have

$$\log P(\boldsymbol{Y}|M_k) = \log P(\boldsymbol{Y}|\widehat{\theta}_k, M_k) - \frac{d_k}{2}\log(N) + O(1). \tag{16}$$

If we multiply this by -2, the BIC for, say, M_1 and M_2 is calculated as

$$BIC = 2[\log P(\boldsymbol{Y}|\widehat{\theta}_2, M_2) - \log P(\boldsymbol{Y}|\widehat{\theta}_1, M_1)] - (d_2 - d_1)\log(N). \tag{17}$$

Equivalently using $l(\hat{\boldsymbol{\theta}}_k)$, the log-likelihood function in place of the $\log P(\boldsymbol{Y}|\widehat{\theta}_k, M_k)$ terms in (17) we have

$$BIC = 2[l(\hat{\boldsymbol{\theta}}_2) - l(\hat{\boldsymbol{\theta}}_1)] - (d_2 - d_1)\log(N). \tag{18}$$

The relative error of the BIC in approximating a Bayes Factor is $O(1)$. This approximation will not always work well, for example when N is small, but also when sample size does not accurately summarize the amount of available information. This assumption breaks down for example when the explanatory variables are extremely collinear or have little variance, or when the number of parameters increase with sample size (Winship 1999; Weakliem 1999).

4.2 IBIC and Related Variants: Retaining Smaller Order Terms

Some of the terms for the approximation of the marginal likelihood that are dropped by the BIC can be easily calculated and retained. One variant called the HBIC retains the third term in equation (5) (Haughton 1988). A simulation study by Haughton, Oud, and Jansen (1997) found that this approximation performs better in model selection for structural equation models than does the usual BIC.

The second term can also be retained by calculating the estimated expected information matrix, $I_E(\widehat{\theta}_k)$, which is often available in statistical software for a variety of statistical models. Taking advantage of this and building on Haughton, Oud, and Jansen (1997), we propose a new approximation for the marginal likelihood, omitting only the second term in (6).

$$\log P(\boldsymbol{Y}|M_k) = \log P(\boldsymbol{Y}|\widehat{\theta}_k, M_k) - \frac{d_k}{2}\log\left(\frac{N}{2\pi}\right) - \frac{1}{2}\log\left|\bar{I}_E(\widehat{\theta}_k)\right| + O(1).$$

For models $M_1$and M_2 and multiplying by -2, this leads to a new approximation of a Bayes Factor, the *Information matrix-based Bayesian Information Criterion* (IBIC).[4] This is given by

$$IBIC = 2[l(\hat{\boldsymbol{\theta}}_2) - l(\hat{\boldsymbol{\theta}}_1)] - (d_2 - d_1)\log\left(\frac{N}{2\pi}\right) - \log\left|\bar{I}_E(\widehat{\theta}_2)\right| + \log\left|\bar{I}_E(\widehat{\theta}_1)\right|. \qquad (20)$$

Of the three approximations that we have discussed, IBIC includes the most terms from equation (5), leaving out just the prior distribution term $\log P(\widehat{\theta}_k|M_k)$.

5 Calculation Examples

To further illustrate the calculation of these Bayes Factor approximations, next we provide a simple example using generated data.[5] Consider a multiple regression model with two independent

[4] Another approximation due to Kashyap (1982), and given by

$$2[l(\hat{\boldsymbol{\theta}}_2) - l(\hat{\boldsymbol{\theta}}_1)] - (d_2 - d_1)\log(N) - \log\left|\bar{I}_E(\widehat{\theta}_2)\right| + \log\left|\bar{I}_E(\widehat{\theta}_1)\right| \qquad (19)$$

is very similar to our proposed IBIC. The IBIC incorporates the estimated expected information matrix at the parameter estimates for the two models being compared.

[5] The data generated for this example are available online.

variables where only the first independent variable is in the true model and the sample size is N = 50. Suppose we estimate three models: one with only $x1$, one with only $x2$, and the last with both $x1$ and $x2$. Table 1 contains the ingredients for the approximations, where the rows give the necessary quantities from the regression output.[6] The three columns give the values of these components for each of the three models.

[Insert Table 1 here]

Using these numbers and the formulas we can calculate each approximation. For instance, the BIC formula is $-2l(\hat{\boldsymbol{\theta}}) + d\log(N)$. Reading from Table 1 for the predictor $x1$ (Column 1):

$$-2 \times (-87.21) + 2 \times 3.91 = 182.24$$

The SPBIC formula is $-2l(\hat{\boldsymbol{\theta}})+d\left(1-\log\left(\frac{d}{\hat{\boldsymbol{\theta}}^T I(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\theta}}}\right)\right)$ where the values of its components are taken from Table 1 and listed below in its formula.

$$-2 \times (-87.21) + 2 \times (1 - \log(2/270.74)) = 186.24$$

In a similar fashion, we can calculate the remaining approximations. For this particular model we observe that SPBIC and IBIC achieve their minimum for the smaller model including predictor $x1$ alone, whereas all other information criterion achieve their minimum for the larger model including both predictors. Thus the choice of approximation can have consequences for model selection.

6 Numerical Examples

In this section we examine the BF approximations to better understand their performances in choosing the correct models in multiple regression under different conditions. As part of this

[6] Recall from above that the calculation of our Bayes Factor approximations requires several or more parts of the output that accompanies a regression analysis. These are the degrees of freedom of the regression model, the estimated log-likelihood, the estimates of the parameters, the observed information matrix, the mean expected information matrix and the log of its determinant, $\log(N)$, and $\log(N/2\pi)$.

analysis, we compare the performance of SPBIC and IBIC to each other and to the BIC and HBIC. Our first example is based on a simulation study following Fan and Li (2001) and Hunter and Li (2005), where our objective is to select the correct variables in a linear regression. However, we vary the design factors more extensively than in Fan and Li (2001) and Hunter and Li (2005), so as to capture conditions common in social science data, including a wide range of sample sizes, r-squares, and number of and degree of correlation among variables . The second example details variable selection results for a frequently analyzed dataset on crime rates (Ehrlich 1973).

6.1 Simulation: Variable Selection in Linear Regression

Regression models remain a common tool in sociological analyses. Ideally, subject matter expertise and theory would provide the complete specification of the explanatory variables required to explain a dependent variable. But in practice there is nearly always some uncertainty as to the best specification of the multiple regression model. Our simulation is designed to examine this common problem. In this example we simulate the following linear model

$$Y = \boldsymbol{X}\beta + \epsilon$$

$$\text{with design Matrix } \boldsymbol{X} \sim \mathcal{N}(0, \Sigma_x), \;\; \Sigma_x = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}.$$

We consider the following experimental design for our simulation: eight candidate covariates for the regression; four different levels of model complexity, 2, 4, 6, and 7 variables in the true model; five different sample sizes, N = 50, N = 100, N = 500, N = 1,000, and N = 2,000; three levels of variance explained, R^2 = 0.30, R^2 = 0.60 and R^2 = 0.90; and two levels of correlation between the explanatory variables, $\rho = 0.25$ and $\rho = 0.75$.[7] We also performed the simulations using two additional covariance matrices with non-identical correlations and using lower R^2 values (0.05,

[7] We also tried sample sizes of 30, 60, and 200, and manipulations of the error term distribution between a normal, t with three degrees of freedom, and exponential with $\lambda = 1$. All of these cases produced results similar to those reported here.

0.10, and 0.20). The results with the additional covariance matrices are presented below, while those with the additional R^2 values are available in the online appendix. All coefficients are set to one throughout the simulations and each complexity/sample size/R^2/ρ combination of conditions was run 500 times, producing a total of 120,000 iterations.

The task of choosing the true model is a challenging one in that there could be anywhere from zero to all eight covariates in the true model. This results in 256 possible regression models to choose among for each combination of simulation conditions. We perform all subset regression model selections using each of these criteria and record the percentage of times the correct model was selected.[8] We consider a correct selection to be a case in which the value of a measure for the true model is the lowest or is within 2 units of the best-fitting model. The value of 2 is chosen based on the Jeffreys-Raftery (1995, page 139) guidelines that suggest a difference of less than 2 signifies only a small difference in fit between two models. In essence, we consider models whose difference in fit is less than 2 as essentially tied in their fit. However, our results are quite similar if we consider a correct choice to be only a clear and exact selection of the true model.

6.1.1 Results

The simulation results are given in Figures 1 and 2. In each graph, the x-axes represent increasing sample sizes and the y-axes represent the percentage of cases in which the correct model was selected. The rows of graphs present the outcomes with R^2 = 0.90, 0.60 , and 0.30, respectively, and the two left and two right columns of graphs present results with ρ = 0.25 and 0.75, respectively. Figure 1 presents the findings with 2 and 4 covariates in the true model and Figure 2 presents results with 6 and 7 covariates.

[Insert Figure 1 here]

[8] We opt to include a true model in our simulation design because it provides a clear and easily understood test of the relative performance of the various model selection criteria under ideal conditions. Also, this design is consistent with how sociologists typically approach model selection in practice. However in most empirical research in social science, no fitted model can realistically be expected to be "correct," but rather all models are better or worse approximations of reality. One direction for future research could be to conduct simulations that employ other approaches to model selection. For example, Burnham and Anderson (2004) conduct model averaging rather than search for the one best model in comparing the relative performance of AIC and BIC and Weakliem (2004) recommends hypothesis testing for the direction of effects rather than their presence or absence.

[Insert Figure 2 here]

We first consider the conditions under which *all* BF approximations perform well or perform poorly in selecting the true model. The conditions that improve the performance of all BF approximations are large samples sizes ($N \geq 500$), higher R^2s, lower collinearity, and fewer covariates. For instance, if the R^2 is 0.60 and there are only four or fewer true covariates in the model, and $N \geq 500$, then using any one of the BF approximations will lead to the true model over 90% of the time even with high collinearity (ρ = 0.75). Or if the R^2 is very high (0.90) all BF approximations approach 100% accuracy as long $N \geq 500$. On the other hand, all the BF approximations perform poorly when the R^2 is more modest (0.30), the collinearity is high (ρ = 0.75), and the number of covariates in the true model is larger (6 or more). In these latter conditions, none of the BF approximations works well.

Overall, increasing the unexplained variance or the correlation between covariates proves to be detrimental to all of the fit statistics. In addition, these deteriorating effects are more severe in the more complex models shown in Figure 2, especially when the sample size is small. Another finding from these simulations is that model complexity exerts differing effects on performance depending on values of the other conditions. For example, when R^2 = 0.90 and ρ = 0.25, there is little change in performance with increases in the number of variables in the true model (top-left panels of both figures). However, when R^2 = 0.60 and ρ = 0.75, performance drops considerably across all four fit statistics as the model becomes more complex (compare the middle right panels of Figure 1 to the middle right panels of Figure 2). Under the worst conditions—6 or 7 covariates, R^2 = 0.30, ρ = 0.75—all of the criteria perform poorly even when N = 2,000 (bottom-right panels in Figure 2).

Though the similarity of performance is noteworthy, there are situations where some BF approximations outperform others. In many, but not all cases, the SPBIC and IBIC exhibit better performance than BIC and HBIC when the sample size is small. In Figure 1, when R^2 = 0.90, ρ = 0.25, and the sample size is 50, the SPBIC and IBIC select the true model over 90% of the time, while the next best choice, BIC, is closer to 80%. HBIC performs the worst in this case.

Overall, these simulations show that there are conditions when all BF approximations perform well and other conditions where none works well. Under the former conditions it matters little which BF approximation is selected since all help to select the correct model. Under the latter condition when all BF approximations fail, it does not make sense to use any of them. None of the fit statistics can overcome a combination of a low R^2, high collinearity, large number of covariates, and modest sample size.

There are simulation conditions where the BF approximations depart from each other. Under these situations, the SPBIC and IBIC are generally more accurate. For example, with a large R^2 (0.90) and smaller sample size, the SPBIC and IBIC tend to outperform the BIC and HBIC. The BIC would be next in its overall accuracy followed by the HBIC.

6.1.2 Additional Simulation Conditions

The simulations in the previous section keep identical correlations among the covariates (ρ = 0.25 or 0.75). Here we show results that suggest that our findings are not dependent on keeping the correlations the same.[9] We performed the simulations with two additional covariance matrices:

$$\text{Matrix 1} = \begin{bmatrix} 1.0 & 0.3 & 0.3 & 0.3 & -0.2 & -0.2 & -0.2 & -0.2 \\ 0.3 & 1.0 & 0.3 & 0.3 & -0.2 & -0.2 & -0.2 & -0.2 \\ 0.3 & 0.3 & 1.0 & 0.3 & -0.2 & -0.2 & -0.2 & -0.2 \\ 0.3 & 0.3 & 0.3 & 1.0 & -0.2 & -0.2 & -0.2 & -0.2 \\ -0.2 & -0.2 & -0.2 & -0.2 & 1.0 & 0.3 & 0.3 & 0.3 \\ -0.2 & -0.2 & -0.2 & -0.2 & 0.3 & 1.0 & 0.3 & 0.3 \\ -0.2 & -0.2 & -0.2 & -0.2 & 0.3 & 0.3 & 1.0 & 0.3 \\ -0.2 & -0.2 & -0.2 & -0.2 & 0.3 & 0.3 & 0.3 & 1.0 \end{bmatrix}$$

[9] We also simulated with a covariance matrix in which the correlations were drawn randomly from a uniform distribution ranging from 0 to 1. Conclusions were unchanged.

$$\text{Matrix 2} = \begin{bmatrix} 1.0 & 0.7 & 0.7 & 0.7 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.7 & 1.0 & 0.7 & 0.7 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.7 & 0.7 & 1.0 & 0.7 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.7 & 0.7 & 0.7 & 1.0 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 1.0 & 0.7 & 0.7 & 0.7 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.7 & 1.0 & 0.7 & 0.7 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.7 & 0.7 & 1.0 & 0.7 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.7 & 0.7 & 0.7 & 1.0 \end{bmatrix}$$

Figures 3 and 4 present the same results as Figures 1 and 2 from the previous section, but with the new covariance matrices. The results contained in the Figures are quite similar to those shown in the prior figures, so we will not discuss them further.

[Insert Figure 3 here]

[Insert Figure 4 here]

We also performed the simulations with lower R^2 values than in the previous simulations: R^2 = 0.05, 0.10, and 0.20. To conserve space, we present these figures in the online appendix. The main point to note is that the pattern observed for declining R^2s occurs here as well. The worse case for finding the true model occurs when R^2 and N are small and the number of covariates that should be included in the model is large.

6.2 Variable Selection in Crime Data

We turn now to a widely-used empirical example from Ehrlich (1973) that tests whether deterrence affects crime rates in 47 states in the United States. Like our simulation example, this is a regression problem where we want to choose the best variables from among a set of possible covariates, but unlike our simulation we do not know the true model. However, we can compare our results to those of others who have analyzed these data. We apply our two newly developed model selection criteria (SPBIC and IBIC) and BIC to an analysis of a number of models applied to these data. This example has properties that are common in sociological analyses that use states or nations as the unit of analysis: a relatively small N, and potentially high R-squared—exactly the conditions under which we expect SPBIC and IBIC to perform better than the standard BIC. [10]

6.2.1 Description

This dataset originates with Ehrlich (1973), who had data on 47 states in 1960. The original data contained some errors which were corrected by Vandaele (1978). Following Raftery (1995), we use this corrected data for our analysis (the data are available at http://www.statsci.org/data/general/uscrime.txt). The variables for the analysis are listed in Table 2. Following Vandaele (1978), we use a natural log transformation of all variables except the dummy variable for the South.

[Insert Table 2 here]

Table 3 lists a series of models ($M1$ to $M16$), which we derived from previous results by others who have used these data. The model number represent the variables included in the model where the variable numbers come from Table 2. For instance, $M1$ refers to Model 1, which contains % males 14 to 24 (1), mean years of schooling (3), police expenditure in 1960 (4), and income inequality (13). An analogous interpretation holds for the other models.

To narrow down the subset of models to which we fit our various fit criteria, we draw on Raftery's (1995) "Occam's Window" analysis of the same data, which accounts for model uncertainty by selecting a range of candidate models within the subset of all possible models, and then uses this subset to estimate the posterior probability that any given variable is in the true model. Variables 1, 3, 4, and 13 were selected with near certainty based on Raftery's Occam's Window analysis, and so we include these variables in all of the models 1–16. Model 1 includes only these baseline variables, whereas Models 2–16 add all possible combinations of 4 additional variables. Raftery's results indicated considerable model uncertainty for the significance of percent non-white (9) and the unemployment rate for urban males aged 35–39 (11). We also include the two vari-

[10] We exclude the HBIC results from our presentation of results for this example because it performed more poorly on average in the simulations than did either the more familiar BIC or our two new approximations.

ables of primary theoretical interest: probability of imprisonment (14) and average time served (15). These variables are measures of deterrence, which was the focus of Ehrlich's original work on predicting the crime rate. We excluded other variables, either because previous researchers found little evidence of their importance, or because they are highly collinear with variables we included. Finally, Models 17 and 18 are the original models fit by Ehrlich, and Model 19 includes all variables in the data set.

[Insert Table 3 here]

6.2.2 Results

Table 4 lists the SPBIC, IBIC, and BIC values for all 19 models from Table 3. Lower values indicate better fitting models than higher values for all three measures. Model $M1$ is the best fitting model for SPBIC and IBIC, but is only ranked eleventh in fit using the BIC. The BIC measure has model $M16$ as the best fitting model whereas the SPBIC and IBIC rank this model fifteenth and fourteenth, respectively. These results indicate that these different methods of approximating Bayes Factors can lead to a different ranking of models, though the rankings given by SPBIC and IBIC are closer to each other than either is to BIC.

[Insert Table 4 here]

An interesting aspect of these results relates to Ehrlich's (1973) theory of deterrence. The probability of imprisonment (14) and the average time served in prison (15) are the two key deterrence variables in Ehrlich's model. The top five fitting models according to SPBIC and IBIC do not include average time served in prison (15) and only some of these include the probability of imprisonment (14). Raftery's (1995) analysis of the same data using the BIC also calls into question whether average time served in prison belongs in the model, but supports including the probability of imprisonment. Furthermore, Ehrlich's (1973) theoretically specified models—$M17$ and $M18$—rank poorly on all three fit measures.

Finally, Tables 5 and 6 present the coefficient estimates and standard errors from each of the models. Note that the magnitude of effects differs between the selected models. For example, the marginal effect of $\%$ males 14–24 in $M1$ (the SPBIC and IBIC selection) is only 73% the size of the effect estimated in $M16$ (BIC's selection).

[Insert Table 5 here]

[Insert Table 6 here]

We recognize that Ehrlich's (1973) analysis has provoked much debate and we do not make strong claims as to the truth of the selected models particularly since there might be other important determinants of crime not included in any of these models. Rather we present our findings as an illustration of the possibility of incongruent results with standard model selection techniques. Other things being equal, based on our theoretical derivation and our simulation results we suggest that SPBIC and IBIC should be preferred to BIC especially given the modest sample size of this empirical example and relatively high variance explained.[11]

7 Conclusions

Bayes Factors are valuable tools to aid in the selection of nested and non-nested models. However, exact calculations of Bayes Factors from fully Bayesian analyses that include explicit prior distributions are relatively rare. Instead, researchers more commonly use approximations to Bayes Factors, the most well-known being the BIC. In this paper we provided a comparison of BIC and several alternatives, including two new approximations to Bayes Factors. One, the SPBIC, uses a scaled unit information prior that gives greater flexibility than the implicit unit information prior that underlies the BIC. The second, the IBIC, incorporates more terms from the standard Laplace approximation than does the BIC in estimating Bayes Factors.

From a practitioners' standpoint, both the SPBIC and IBIC are straightforward to calculate for any application in which software outputs an expected or observed information matrix. This is possible in many software packages for generalized linear models and structural equation models,

[11] We recognize that the closeness of values on these fit measures suggest that the model uncertainty approach of Raftery (1995) could be useful for this example. However, in practice most sociologists continue to look for the "best" or "true" model.

among others. But unlike our example of variable selection in linear models (Section 6), the expected and the observed information matrices may not be analytically the same. In models where both the observed and the expected information matrix are available, it is desirable to use the observed information matrix as it provides a more accurate approximation (for details see Efron and Hinkley 1978; Kass and Vaidyanathan 1992; Raftery 1996). In Appendix B we illustrate how to calculate SPBIC and IBIC in regression models using Stata and R.

In addition to proposing two new BF approximations, a contribution of this paper is to compare the performance of these and the BIC and HBIC in their accuracy of selecting valid models. In our simulation of variable selection in regression, we found that no BF approximation was accurate under all conditions. Indeed, with large samples, and modest to large R^2s, it practically does not matter which of the BF approximations a researcher chooses since they are all highly accurate. On the other hand, if the R^2 is modest, collinearity is high, and a half dozen or more variables belong in the model, then none of the BF approximations works well. When there was a departure in performance, we did find that SPBIC and IBIC performed modestly better in smaller sample sizes, followed by the BIC and then the HBIC.

Based on these results, we would advise researchers with large samples and high R^2s to choose a BF approximation that is most readily available. On the other hand, our results suggest that the SPBIC and IBIC might be more useful than BIC when the sample is smaller. If the R^2 and N are low and it is likely that a half-dozen or more covariates are in the true model, then none of the BF approximations should be used.

Though our simulation experiments were informative about the performance of the SPBIC, IBIC, BIC, and HBIC, they are far from the final word. We need to examine additional empirical data and simulation designs to better understand these ways of estimating Bayes Factors. In addition, it would be valuable to study multilevel, mixture, and structural equation models to assess the accuracies of these approximations to Bayes Factors in these different types of models.

A Generalized Laplace Approximation

First, let us fix some notations. For a vector valued function $a(\boldsymbol{\theta})$, let $H(\boldsymbol{\theta}) = -\dfrac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$ denote the negative Hessian matrix. In the case $a(\boldsymbol{\theta}) = l(\boldsymbol{\theta}_k)$, is a log likelihood function; $I_O(\boldsymbol{\theta}) = H(\boldsymbol{\theta})$, is the observed information matrix, based on N observations. Also let b be the pdf of a normal with mean $\boldsymbol{\theta}^*$ and variance , V^* such that $b(\boldsymbol{\theta}) = \phi(\boldsymbol{\theta}; \boldsymbol{\theta}^*, V^*)$.

Proposition 1 *Using the above notations*

$$\int_{\boldsymbol{\theta}} \exp[a(\boldsymbol{\theta})] b(\boldsymbol{\theta}) d\boldsymbol{\theta} \approx a(\hat{\boldsymbol{\theta}}) \frac{|V^*|^{\frac{1}{2}}}{\left|H^{-1}(\hat{\boldsymbol{\theta}}) + V^*\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \left(H^{-1}(\hat{\boldsymbol{\theta}}) + V^*\right)(\boldsymbol{\theta}^* - \boldsymbol{\theta}^*)\right)$$

Proof. Applying Taylor series expansion only on $a(\boldsymbol{\theta})$ we get

$$\int_{\boldsymbol{\theta}} \exp[a(\boldsymbol{\theta})] b(\boldsymbol{\theta}) d\boldsymbol{\theta} \approx \exp\left(a(\hat{\boldsymbol{\theta}})\right) (2\pi)^{\frac{p}{2}} \left|H^{-1}(\hat{\boldsymbol{\theta}})\right|^{\frac{1}{2}} \int_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, H^{-1}) b(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{A.1}$$

Further using convolution of normals we have

$$\int_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, H^{-1}) b(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}; \boldsymbol{\theta}_*, H^{-1}) \phi(\boldsymbol{\theta}; \boldsymbol{\theta}^*, V^*) d\boldsymbol{\theta} = \phi(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}^*, H^{-1} + V^*). \tag{A.2}$$

Substituting (A.2) in (A.1) we get Proposition 1.

B Calculating SPBIC and IBIC in Stata and R

The following code provides examples for calculating SPBIC and IBIC in Stata and R.

B.1 Stata

Note that it is easiest to use the `glm` procedure because it automatically produces the necessary parts of the computation: the log-likelihood (`e(ll)`), number of parameters (`e(k)`), number of observations (`e(N)`), coefficients (`e(b)`), and covariance matrix of the coefficients (`e(V)`).

```
* Create data
set seed 10000
set obs 500
gen x1 = invnorm(uniform())
gen x2 = invnorm(uniform())
gen x3 = invnorm(uniform())
gen x4 = invnorm(uniform())
gen y = x1 + x2 + x3 + x4 + invnorm(uniform())

* Model fitting
glm y x1 x2 x3 x4

** Calculate SPBIC, Case 1
matrix info = inv(e(V))
matrix thinfoth = e(b)*info*e(b)’
scalar spbic1 = -2*e(ll) + e(k)*(1-log(e(k)/el(thinfoth, 1, 1)))
scalar list spbic1

** Calulate SPBIC, Case 2
scalar spbic2 = -2*e(ll) + el(thinfoth, 1, 1)
scalar list spbic2

** Calculate IBIC
scalar ibic = -2*e(ll) + e(k)*log(e(N)/(2*_pi)) + log(det(info))
scalar list ibic
```

B.2 R

Both SPBIC and IBIC can be calculated from output from the `lm()` function in R.

```
# Create data
set.seed(10000)
x1 <- rnorm(500)
x2 <- rnorm(500)
x3 <- rnorm(500)
x4 <- rnorm(500)
```

```
y <- x1 + x2 + x3 + x4 + rnorm(500)

# Model fitting
model <- lm(y ~ x1 + x2 + x3 + x4)

# Calculate SPBIC, Case 1
info <- solve(vcov(model))
thinfoth <- model$coef %*% info %*% model$coef
pspbic.1 <- as.numeric(length(model$coef)*(1-log(length(model$coef)/(thinfoth))))
spbic.1 <- -2*as.numeric(logLik(model)) + pspbic.1

# Calculate SPBIC, Case 2
pspbic.2 <- thinfoth
spbic.2 <- -2*as.numeric(logLik(model)) + pspbic.2

# Calculate IBIC
idet <- log(det(info))
pibic <- length(model$coef)*log(length(y)/(2*pi)) + idet
ibic <- -2*as.numeric(logLik(model)) + pibic
```

References

Berger, James O. 1994. "An Overview of Robust Bayesian Analysis." *Test (Madrid)* 3(1):5–59.

Berger, James O., Surajit Ray, Ingmar Visser, Ma J. Bayarri and W. Jang. 2006. Generalization of BIC. Technical report University of North Carolina, Duke University, and SAMSI.

Burnham, Kenneth P. and David R. Anderson. 2004. "Multimodel Inference: Understanding AIC and BIC in Model Selection." *Sociological Methods & Research* 33(2):261–304.

Carlin, Bradley P. and Siddhartha Chib. 1995. "Bayesian Model Choice via Markov Chain Monte Carlo Methods." *Journal of the Royal Statistical Society, Series B, Methodological* 57(3):473–484.

Carlin, Bradley P. and Thomas A. Louis. 1996. *Bayes and Empirical Bayes Methods for Data Analysis*. New York: Chapman and Hall.

Efron, Bradley and David V. Hinkley. 1978. "Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information." *Biometrika* 65(3):457–483.

Ehrlich, Isaac. 1973. "Participation in Illegitimate Activities: A Theoretical and Empirical Investigation." *Journal of Political Economy* 81(3):521–565.

Fan, Jianqing and Runze Li. 2001. "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties." *Journal of the American Statistical Association* 96(456):1348–1360.

Fitzmaurice, Garrett M. 1997. "Model Selection with Overdispersed Data." *The Statistician* 46(1):81–91.

Fougere, P. 1990. Maximum Entropy and Bayesian Methods. In *Maximum Entropy and Bayesian Methods*, ed. P. Fougere. Dordrecht, NL: Kluwer Academic Publishers.

Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin. 1995. *Bayesian Data Analysis*. London: Chapman & Hall.

Haughton, Dominique M. A. 1988. “On the Choice of a Model to Fit Data From an Exponential Family.” *The Annals of Statistics* 16(1):342–355.

Haughton, Dominique M. A., Johan H. L. Oud and Robert A. R. G. Jansen. 1997. “Information and Other Criteria in Structural Equation Model Selection.” *Communications in Statistics, Part B – Simulation and Computation* 26(4):1477–1516.

Hunter, David R. and Runze Li. 2005. “Variable selection using MM algorithms.” *Annals of Statistics* 33(4):1617–1642.

Jeffreys, Harold. 1939. *Theory of Probability*. New York: Oxford University Press.

Kashyap, Rangasami L. 1982. “Optimal Choice of AR and MA Parts in Autoregressive Moving Average Models.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 4(2):99–104.

Kass, Robert E. and Adrian E. Raftery. 1995. “Bayes Factors.” *Journal of the American Statistical Association* 90(430):773–795.

Kass, Robert E. and Suresh K. Vaidyanathan. 1992. “Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of Two Binomial Proportions.” *Journal of the Royal Statistical Society, Series B: Methodological* 54(1):129–144.

Kuha, Jouni. 2004. “AIC and BIC: Comparisons of Assumptions and Performance.” *Sociological Methods & Research* 33(2):188–229.

Raftery, Adrian E. 1993. Bayesian Model Selection in Structural Equation Models. In *Testing Structural Equation Models*, ed. Kenneth A. Bollen and J. Scott Long. Newbury Park, CA: Sage pp. 163–180.

Raftery, Adrian E. 1995. *Sociological Methodology*. Cambridge, MA: Blackwell chapter Bayesian Model Selection in Social Research (with Discussion), pp. 111–163.

Raftery, Adrian E. 1996. “Approximate Bayes Factors and Accounting for Model Uncertainty in Generalised Linear Models.” *Biometrika* 83(2):251–266.

Richardson, Sylvia and Peter J. Green. 1997. “On Bayesian Analysis of Mixtures with an Unknown Number of Components.” *Journal of the Royal Statistical Society, Series B, Methodological* 59(4):731–758.

Schwarz, Gideon. 1978. “Estimating the Dimension of a Model.” *Annals of Statistics* 6(2):461–464.

Spiegelhalter, D. J. and A. F. M. Smith. 1982. “Bayes Factors for Linear and Loglinear Models with Vague Prior Information.” *Journal of the Royal Statistical Society, Series B, Methodological* 44(3):377–387.

Tierney, Luke and Joseph B. Kadane. 1986. “Accurate Approximations for Posterior Moments and Marginal Densities.” *Journal of the American Statistical Association* 81(393):82–86.

Vandaele, Walter. 1978. Participation in Illegitimate Activities: Ehrlich Revisited. In *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*, ed. Alfred Blumstein, Jacqueline Cohen and Daniel Nagin. Washington, D.C.: National Academy of Sciences.

Volinsky, Chris T. and Adrian E. Raftery. 2000. “Bayesian Information Criterion for Censored Survival Models.” *Biometrics* 56(1):256–262.

Weakliem, David L. 1999. “A Critique of the Bayesian Information Criterion for Model Selection.” *Sociological Methods & Research* 27(3):359–397.

Weakliem, David L. 2004. “Introduction to the Special Issue on Model Selection.” *Sociological Methods & Research* 33(2):261–304.

Winship, Christopher. 1999. “Editor’s Introduction to the Special Issue on the Bayesian Information Criterion.” *Sociological Methods & Research* 27(3):355–358.

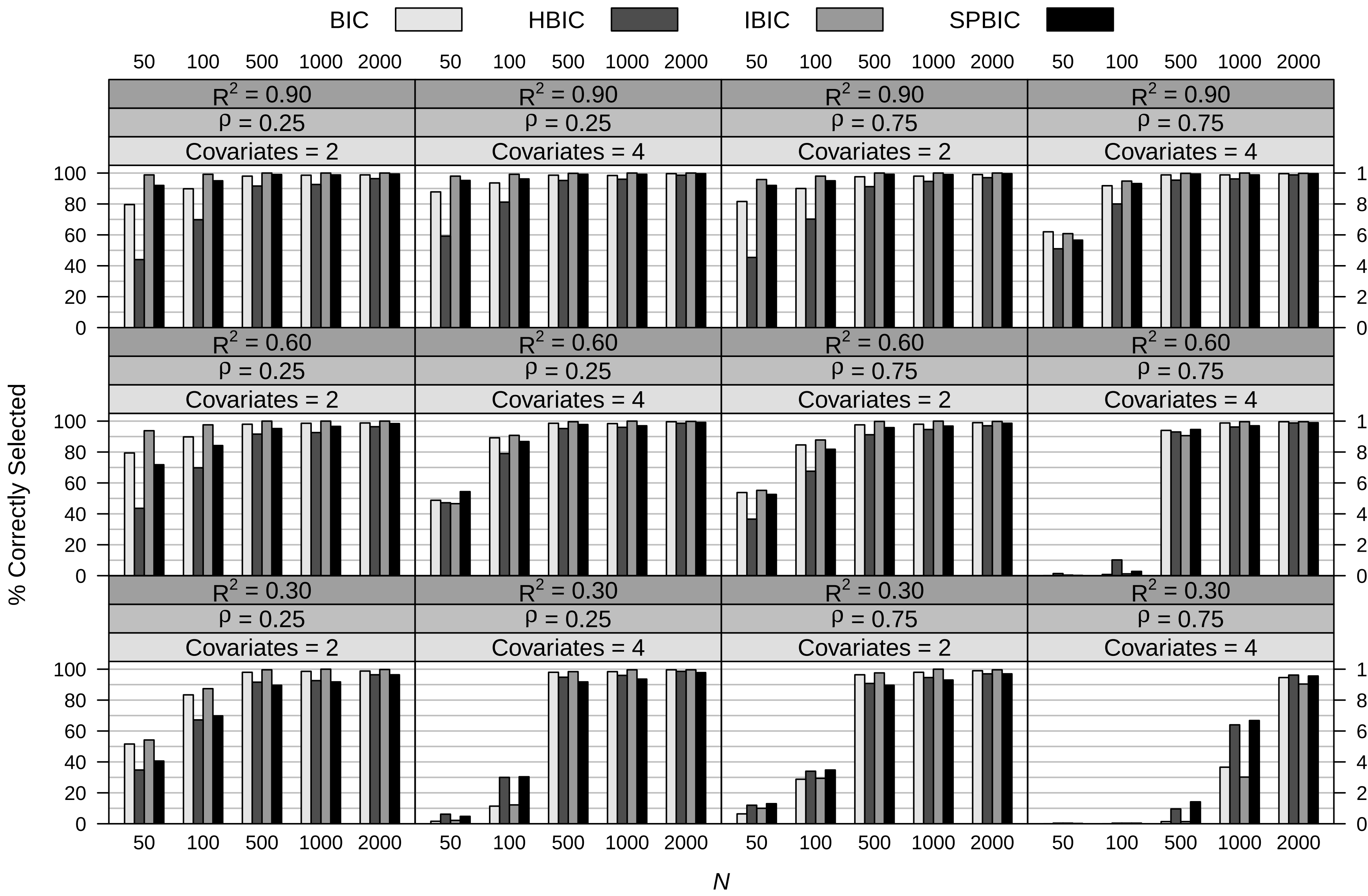

Figure 1: Simulation Results with 2 and 4 Covariates in the True Model

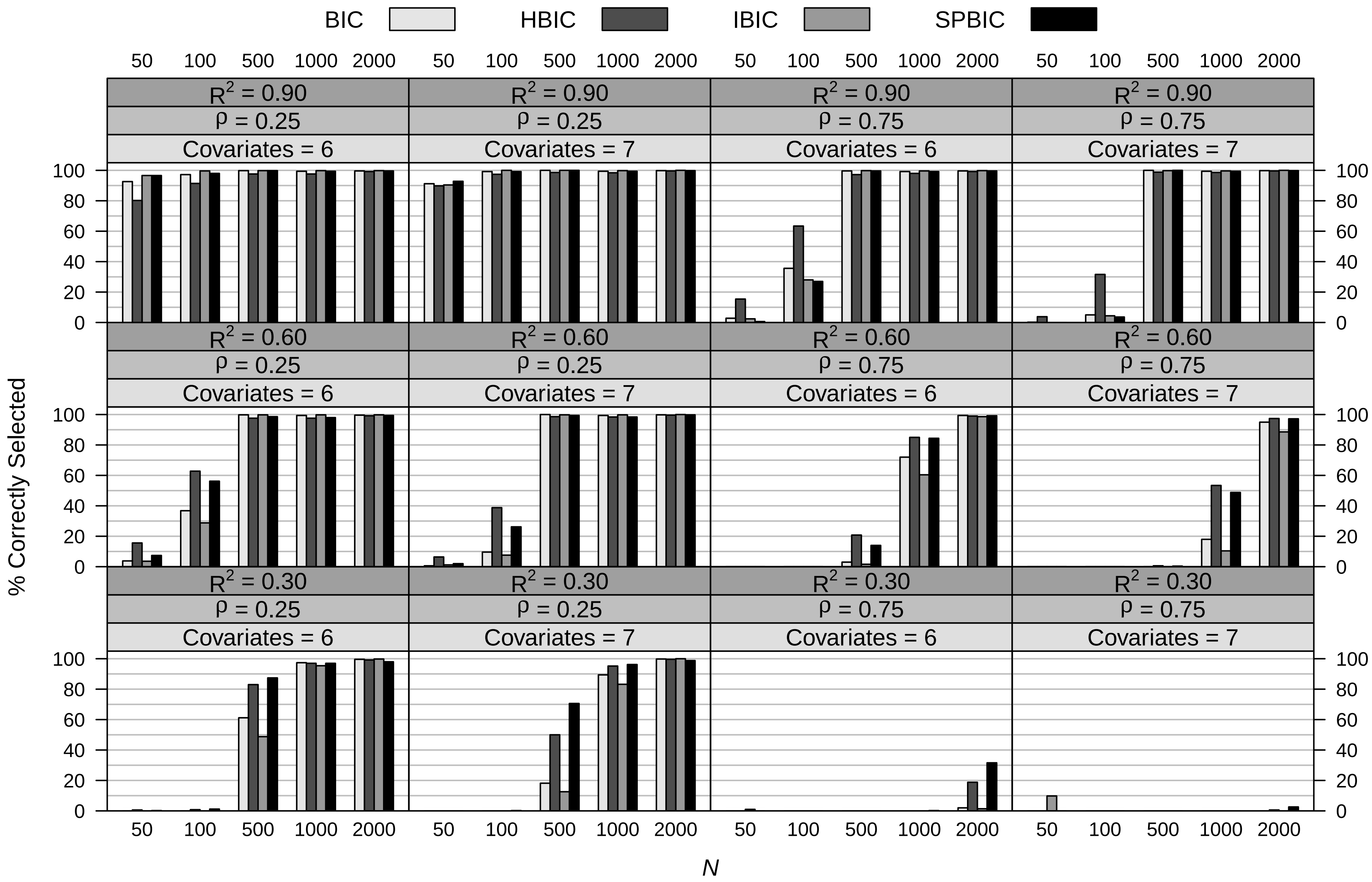

Figure 2: Simulation Results with 6 and 7 Covariates in the True Model

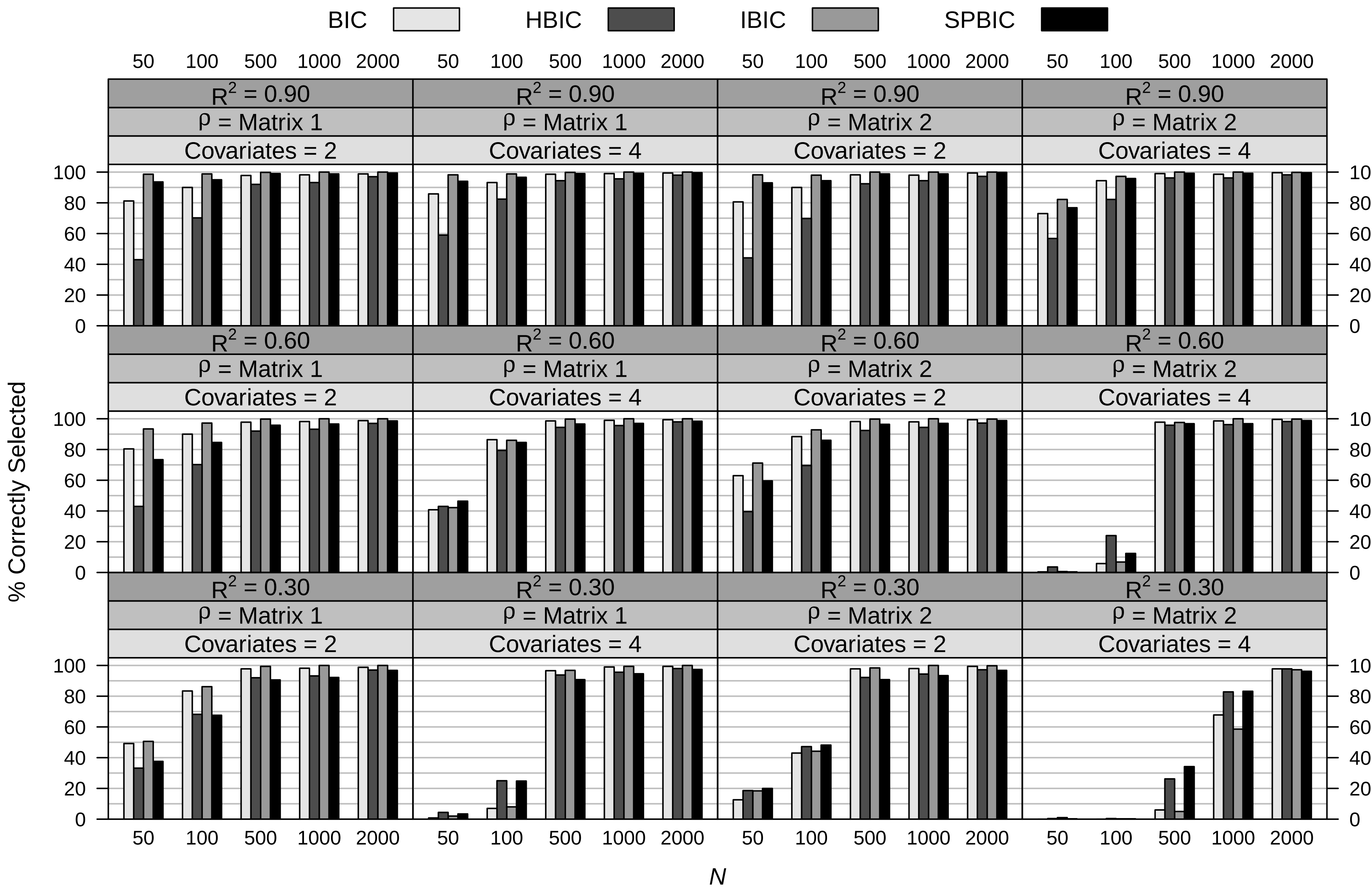

Figure 3: Simulation Results with 2 and 4 Covariates in the True Model and Different Covariance Structures

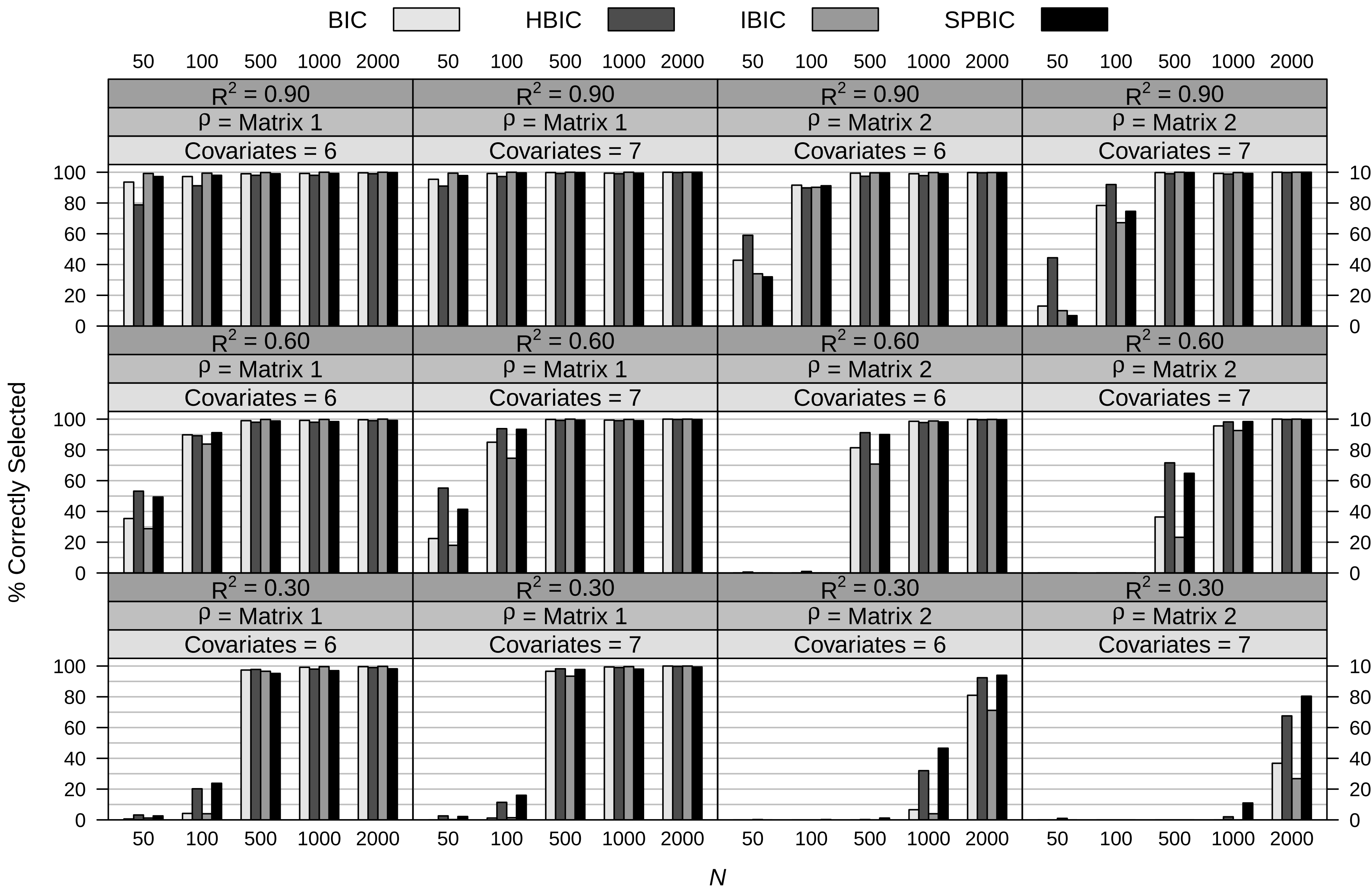

Figure 4: Simulation Results with 6 and 7 Covariates in the True Model and Different Covariance Structures

Table 1: Bayes Factor Approximation Calculation Examples

Statistic	$x1$	$x2$	$x1$ and $x2$
Log Likelihood	-87.21	-125.37	-84.80
N	50	50	50
$\log(N)$	3.91	3.91	3.91
$\log \frac{N}{2\pi}$	2.77	2.77	2.77
d	2	2	3
Information Matrix $\bar{I}(\widehat{\theta})$	$\begin{pmatrix} 25.04 & 3.93 \\ 3.93 & 25.23 \end{pmatrix}$	$\begin{pmatrix} 5.44 & 0.42 \\ 0.42 & 9.32 \end{pmatrix}$	$\begin{pmatrix} 27.00 & 4.23 & 2.08 \\ 4.23 & 27.20 & 22.79 \\ 2.08 & 22.79 & 46.25 \end{pmatrix}$
Regression Coefficients $\widehat{\theta}$	(0.545, 3.147)	(0.94, 1.29)	(0.522, 3.501, $-$0.419)
$\hat{\boldsymbol{\theta}}^T I(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\theta}}$	270.74	21.26	296.67
$\left\lvert \bar{I}(\widehat{\theta}) \right\rvert$	6.42	3.92	9.87
SPBIC	186.24	257.47	186.38
IBIC	184.99	258.82	185.70
HBIC	178.57	254.90	175.83
BIC	182.24	258.57	181.34

Table 2: Variable Number and Name for Crime Data from Ehrlich (1973) and Vandaele (1978)

Number	Variable Name
1	% males 14 to 24
2	Southern state dummy variable
3	Mean years of education
4	Police expenditures in 1960
5	Police expenditures in 1959
6	Labor force participation rate
7	Number of males per 1000 females
8	State population
9	Number of nonwhites per 1000 population
10	Unemployment rate of urban males 14 to 24
11	Unemployment rate of urban males 35 to 39
12	GDP
13	Income inequality
14	Probability of imprisonment
15	Average time served in state prisons

Table 3: List of Models to be Compared

Model Number	Included Variable Numbers
$M1$	1, 3, 4, 13
$M2$	1, 3, 4, 9, 13
$M3$	1, 3, 4, 11, 13
$M4$	1, 3, 4, 9, 11, 13
$M5$	1, 3, 4, 13, 14
$M6$	1, 3, 4, 9, 13, 14
$M7$	1, 3, 4, 11, 13, 14
$M8$	1, 3, 4, 9, 11, 13, 14
$M9$	1, 3, 4, 13, 15
$M10$	1, 3, 4, 9, 13, 15
$M11$	1, 3, 4, 11, 13, 15
$M12$	1, 3, 4, 9, 11, 13, 15
$M13$	1, 3, 4, 13, 14, 15
$M14$	1, 3, 4, 9, 13, 14, 15
$M15$	1, 3, 4, 11, 13, 14, 15
$M16$	1, 3, 4, 9, 11, 13, 14, 15
$M17$	9, 12, 13, 14, 15
$M18$	1, 6, 9, 10, 12, 13, 14, 15
$M19$	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

Table 4: SPBIC, IBIC, and BIC for Models $M1$ to $M19$

Model	SPBIC	IBIC	BIC
$M1$	35.36	12.66	4.10
$M2$	41.62	18.17	4.96
$M3$	38.83	12.90	1.77
$M4$	45.04	18.52	2.59
$M5$	38.85	14.31	1.79
$M6$	43.04	17.89	0.25
$M7$	42.00	14.39	-0.97
$M8$	46.08	18.00	-2.71
$M9$	43.83	18.25	7.49
$M10$	50.13	23.96	8.58
$M11$	46.94	18.29	4.83
$M12$	53.29	24.18	5.98
$M13$	46.21	18.39	3.97
$M14$	47.39	19.00	-1.14
$M15$	49.74	18.98	1.69
$M16$	51.19	19.97	-3.26
$M17$	56.02	36.98	21.47
$M18$	75.57	42.52	26.90
$M19$	102.08	40.50	14.69

Table 5: Model Results in Crime Data (M1–M9)

	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$	$M7$	$M8$	$M9$
% Males	76.02*	84.11*	101.98*	110.59*	79.69*	77.89*	105.02*	103.79*	73.22*
14–24	(34.42)	(37.44)	(35.32)	(38.13)	(32.62)	(35.69)	(33.30)	(36.22)	(34.61)
Education	166.05*	156.97*	203.08*	193.70*	160.15*	162.14*	196.47*	197.77*	178.32*
	(45.80)	(48.79)	(47.42)	(50.05)	(43.42)	(46.43)	(44.75)	(47.43)	(47.72)
Police Exp.	129.80*	134.16*	123.31*	127.85*	121.23*	120.07*	115.02*	114.25*	126.54*
(1960)	(14.38)	(16.35)	(14.16)	(16.00)	(14.06)	(16.69)	(13.75)	(16.20)	(14.82)
Income	64.09*	67.93*	63.49*	67.52*	68.31*	67.50*	67.65*	67.11*	64.93*
Inequality	(15.27)	(16.78)	(14.68)	(16.12)	(14.56)	(15.95)	(13.94)	(15.27)	(15.32)
Nonwhites		−3.08		−3.24		0.71		0.48	
		(5.36)		(5.15)		(5.35)		(5.13)	
Male Unemployment			91.36*	91.75*			89.37*	89.28*	
35–39			(43.41)	(43.74)			(40.91)	(41.43)	
Prob. of					−3867.27*	−3936.28*	−3801.84*	−3848.21*	
Imprisonment					(1596.55)	(1697.27)	(1528.10)	(1625.39)	
Avg. Time									4.63
Served									(4.97)
Intercept	−4249.22*	−4345.72*	−5243.74*	−5349.29*	−4064.57*	−4038.99*	−5040.50*	−5022.44*	−4451.80*
	(858.51)	(881.56)	(951.16)	(972.88)	(816.28)	(848.33)	(899.84)	(931.58)	(886.92)
N	47	47	47	47	47	47	47	47	47
R^2	0.70	0.70	0.73	0.73	0.74	0.74	0.77	0.77	0.71

Note: Cell entries report coefficient estimates with standard errors in parentheses for Models 1–9, as described in Tables 2 and 3. * $p < 0.05$.

Table 6: Model Results in Crime Data (M10–M19)

	$M10$	$M11$	$M12$	$M13$	$M14$	$M15$	$M16$	$M17$	$M18$	$M19$
% Males	81.74*	99.30*	108.37*	81.86*	78.38*	106.66*	103.95*		67.53	87.83*
14–24	(37.57)	(35.41)	(38.16)	(33.25)	(36.03)	(33.88)	(36.60)		(49.59)	(41.71)
Education	168.98*	216.22*	206.60*	152.04*	154.97*	189.41*	191.50*			188.32*
	(50.47)	(49.15)	(51.56)	(47.02)	(48.80)	(48.29)	(49.87)			(62.09)
Police Exp.	131.08*	119.85*	124.58*	121.99*	119.65*	115.70*	113.96*			192.80
(1960)	(16.69)	(14.57)	(16.30)	(14.29)	(16.86)	(13.99)	(16.38)			(106.11)
Nonwhites	−3.28		−3.44		1.51		1.14	15.72*	13.91*	4.20
	(5.37)		(5.15)		(5.61)		(5.38)	(6.14)	(6.57)	(6.48)
Income	69.03*	64.36*	68.66*	68.39*	66.68*	67.73*	66.44*	69.31*	68.42*	70.67*
Inequality	(16.84)	(14.71)	(16.15)	(14.70)	(16.17)	(14.08)	(15.50)	(25.02)	(25.32)	(22.72)
Avg. Time	4.75	4.83	4.96	−2.76	−3.20	−2.31	−2.65	−11.28	−10.59	−3.48
Served	(5.02)	(4.77)	(4.81)	(5.78)	(6.07)	(5.54)	(5.83)	(7.73)	(8.03)	(7.17)
Male Unemployment		92.22*	92.66*			88.72*	88.43*			167.80
35–39		(43.40)	(43.71)			(41.36)	(41.90)			(82.34)
Prob. of				−4400.84*	−4633.60*	−4249.76*	−4426.07*	−7282.04*	−6829.52*	−4855.27*
Imprisonment				(1962.11)	(2164.70)	(1880.67)	(2077.10)	(2818.62)	(2884.84)	(2272.37)
GDP								0.43*	0.47*	0.10
								(0.10)	(0.11)	(0.10)
Labor Force									711.09	−663.83
Participation									(1212.69)	(1469.73)
Male Unemployment									635.86	−5827.10
14–24									(2592.84)	(4210.29)
South										−3.80
										(148.76)
Police Exp.										−109.42
(1959)										(117.48)
Males per										17.41
1000 Females										(20.35)
State										−0.73
Population										(1.29)
Intercept	−4559.37*	−5464.36*	−5582.08*	−3918.67*	−3840.76*	−4911.09*	−4848.99*	−2234.91*	−3849.58*	−5984.29*
	(911.05)	(975.53)	(998.08)	(879.05)	(935.15)	(960.73)	(1015.68)	(1013.16)	(1548.59)	(1628.32)
N	47	47	47	47	47	47	47	47	47	47
R^2	0.71	0.74	0.74	0.74	0.74	0.77	0.77	0.52	0.54	0.80

Note: Cell entries report coefficient estimates with standard errors in parentheses for Models 10–19, as described in Tables 2 and 3. * $p < 0.05$.