Sensitivity and Specificity of Information Criteria

John J. Dziak[1], Donna L. Coffman[2], Stephanie T. Lanza[3], and Runze Li[4]

[1]**The Methodology Center, The Pennsylvania State University**
[2]**The Methodology Center and College of Health and Human Development, The Pennsylvania State University**
[3]**The Methodology Center and College of Health and Human Development, The Pennsylvania State University**
[4]**Department of Statistics and The Methodology Center, The Pennsylvania State University**

ABSTRACT

Choosing a model with too few parameters can involve making unrealistically simple assumptions and lead to high bias, poor prediction, and missed opportunities for insight. Such models are not flexible enough to describe the sample or the population well. A model with too many parameters can fit the observed data very well, but be too closely tailored to it. Such models may generalize poorly. Penalized-likelihood information criteria, such as Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC), the Consistent AIC, and the Adjusted BIC, are widely used for model selection. However, different criteria sometimes support different models, leading to uncertainty about which criterion is the most trustworthy. In some simple cases the comparison of two models using information criteria can be viewed as equivalent to a likelihood ratio test, with the different models representing different alpha levels (i.e., different emphases on sensitivity or specificity; Lin & Dayton 1997). This perspective may lead to insights about how to interpret the criteria in less simple situations. For example, AIC or BIC could be preferable, depending on sample size and on the relative importance one assigns to sensitivity versus specificity. Understanding the differences among the criteria may make it easier to compare their results and to use them to make informed decisions.

Keywords: AIC, BIC, model selection, variable selection, information criteria, AICc, CAIC, sensitivity, specificity

INTRODUCTION

Many model selection techniques have been proposed in different settings (see, e.g., Miller, 2002; Pitt and Myung, 2002; Zucchini, 2000). Among other considerations, researchers must balance sensitivity (having enough parameters to adequately model the relationships among variables in the population) with specificity (not overfitting a model or suggesting nonexistent relationships). Several of the simplest and most common model selection criteria can be discussed in a unified way as log-likelihood functions with simple penalties. These include Akaike's Information Criterion (AIC; Akaike, 1973), the Bayesian Information Criterion (BIC; Schwarz, 1978), Bozdogan's consistent AIC (CAIC; Bozdogan, 1987), and the adjusted BIC (see Sclove, 1987). They consist of a goodness-of-fit term plus a penalty to control overfitting, and provide a standardized way to balance sensitivity and specificity.

Each of these simple criteria involves choosing the model with the best penalized log-likelihood (i.e., the highest value of $\ell - A_n p$ where ℓ is the log-likelihood, A_n is some constant or some function of the sample size n, and p is the number of parameters in the model). For historical reasons, instead of finding the highest value of ℓ minus a penalty, this is sometimes expressed as finding the lowest value of -2ℓ plus a penalty, i.e.,

$$-2\ell + A_n p. \tag{1}$$

Expression (1) is what Atkinson (1980) called the generalized information criterion. In this paper we refer to (1) as an *information criterion* (IC). Expression (1) is sometimes replaced in practice (e.g. Collins

Table 1. Summary of Information Criteria

Criterion	Penalty Weight	Emphasis	Consistent?	Likely Kind of Error
AIC	$A_n = 2$	Good future prediction	No	Overfitting
Adjusted BIC	$A_n = \ln\left(\frac{n+2}{24}\right)$	Depends on n	Yes	Depends on n
BIC	$A_n = \ln(n)$	Parsimonious model	Yes	Underfitting
CAIC	$A_n = \ln(n+1)$	Parsimonious model	Yes	Underfitting

Notes. Two variants of AIC are also mentioned in the text. The AIC3 is similar to AIC but uses $A_n = 3$ instead. The AIC_c in a regression equals $AIC+2(m+1)(m+2)/(n-m-2)$ where m is the number of included regression coefficients, including the intercept.

and Lanza, 2010) by the practically equivalent $G^2 + A_n p$, where G^2 is the deviance, that is, -2ℓ plus a function of the saturated model.

Expression (1) cannot be used directly in practice without first choosing A_n. Specific choices of A_n make (1) equivalent to AIC, BIC, adjusted BIC or CAIC. Thus, although motivated by different theories and goals, algebraically these criteria are only different values of A_n in (1), thus corresponding to different relative degrees of emphasis on parsimony (Claeskens and Hjort, 2008; Lin and Dayton, 1997). Because the different ICs often do not agree, the question often arises as to which is best to use in practice. In this paper we review this question by focusing on the similarities and differences among AIC, BIC, CAIC, and adjusted BIC, especially in view of an analogy between their different complexity penalty weights A_n and the α levels of hypothesis tests. In the following section we review AIC, BIC, CAIC, adjusted BIC, and related criteria. We then compare them in three simulated case studies involving latent class analysis (LCA), exploratory factor analysis, and multiple linear regression. Our goal is not to provide a comprehensive survey of model selection, but to demonstrate ways in which some intuitive and useful principles arise repeatedly in different contexts.

Common Penalized-Likelihood Information Criteria

Akaike's Information Criterion (AIC)

In this section we review some commonly used ICs. For convenience these ICs are summarized in Table 1. First, the AIC of Akaike (1973) sets $A_n = 2$ in (1). It estimates the relative Kullback-Leibler (KL) distance (a nonparametric distance measure) of the likelihood function specified by a fitted candidate model, from the unknown true likelihood function that generated the data. The fitted model closest to the truth in the KL sense would not necessarily be the model which best fits the observed sample, since the *observed* sample can often be fit arbitrary well by making the model more and more complex. Rather, the best KL model is the model that most accurately describes the population distribution and hence future samples from it. More specifically, the KL distance can be written as $E_t(\ell_t(y)) - E_t(\ell(y))$ where E_t is the expected value under the unknown true distribution function, ℓ is the log-likelihood of the data under the fitted model being considered, and ℓ_t is the log-likelihood of the data under the unknown true distribution. $E_t(\ell_t(y))$ will be the same for all models being considered, so KL is minimized by choosing the model with highest $E_t(\ell(y))$. The $\ell(y)$ from the fitted model is a biased measure of $E_t(\ell(y))$ because the same data are being used to estimate the model and to try to predict its fit to future data.

However, Akaike (1973) showed that an approximately unbiased estimate of $E_t(\ell(y))$ would be a constant plus $\ell - \mathrm{tr}(\hat{\mathbf{J}}^{-1}\hat{\mathbf{K}})$ (where $\mathbf{J}$ and $\mathbf{K}$ are two $p \times p$ matrices, described below, and tr() is the trace, or sum of diagonals). $\hat{\mathbf{J}}$ is an estimator for the covariance matrix of the parameters based on the second-derivatives matrix of ℓ in the parameters and $\hat{\mathbf{K}}$ is an estimator based on the cross-products of the first derivatives (see Claeskens and Hjort, 2008, pp. 26-7). Akaike showed that $\hat{\mathbf{J}}$ and $\hat{\mathbf{K}}$ are asymptotically equal for the true model, so that the trace becomes approximately p, the number of parameters in the model. For models that are far from the truth, the approximation may not be as good. However, they presumably have poor values of ℓ, so the precise size of the penalty is unimportant (Burnham and Anderson, 2002). The resulting expression $\ell - p$ suggests using $A_n = 2$ in (1) and concluding that fitted models with low

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

values of (1) will be likely to provide a likelihood function closer to the truth. AIC is discussed further by Burnham and Anderson (2002, 2004) and Kuha (2004).

Criteria Related to AIC. It may be that the crucial AIC approximation $\mathrm{tr}(\hat{\mathbf{J}}^{-1}\hat{\mathbf{K}}) \approx p$ is too optimistic and the resulting penalty for model complexity is too weak (Tibshirani and Knight, 1999; Hastie et al., 2001). In the context of regression and time series models, several researchers (e.g., Sugiura, 1978; Hurvich and Tsai, 1989; Burnham and Anderson, 2004) have suggested using a corrected version, $\mathrm{AIC_c}$, which applies a slightly heavier penalty that depends on p and n; it gives results very close to those of AIC unless n is not large relative to p. For small n, Hurvich and Tsai (1989) showed that $\mathrm{AIC_c}$ sometimes performs better than AIC. Also, in the context of mixture models such as LCA, some researchers (Andrews & Currim, 2003; Fonseca & Cardoso, 2007; Yang & Yang, 2007) have suggested $A_n = 3$ in (1) instead of 2. The latter is sometimes called "AIC3." There is little theoretical basis for AIC3, despite fairly good simulation performance.

The Deviance Information Criterion is beyond the scope of this paper and more sophisticated and computationally intensive than (1). However, it has some relationship to AIC (Claeskens and Hjort, 2008).

Also, some selection criteria are in fact asymptotically equivalent to AIC, at least for linear regression. These include Mallows' C_p, and leave-one-out cross-validation (Shao, 1997; Stone, 1977). The latter involves fitting the candidate model on many subsamples of the data, each excluding one subject, and observing the average squared error in predicting the extra response. Each approach is intended to correct a fit estimate for the artifactual inflation in observed performance caused by fitting a model and evaluating it with the same data. Model parsimony is not a motivating goal in its own right, but is a means to reduce unnecessary sampling error caused by having to estimate too many parameters relative to n. Thus, especially for large n, sensitivity is likely to be treated as more important than specificity. If parsimonious interpretation is of interest in its own right, another criterion such as BIC, described in the next section, might be more appropriate.

Schwarz's Bayesian Information Criterion (BIC)

In Bayesian model selection, a prior probability is set for each model M_i, and prior distributions are also set for the nonzero coefficients in each model. If we assume that one and only one model, along with its associated priors, is true, we can use Bayes' Theorem to find the posterior probability of each model given the data. Let $\Pr(M_i)$ be the prior probability set by the researcher, and $\Pr(\mathbf{y}|M_i)$ be the probability density of the data under M_i, calculated as the expected value of the likelihood function of $\mathbf{y}$ given the model and parameters, over the prior distribution of the parameters. According to Bayes' theorem, the posterior probability $\Pr(M_i|\mathbf{y})$ of a model is proportional to $\Pr(M_i)\Pr(\mathbf{y}|M_i)$. The degree to which the data support M_i over another model M_j is given by the ratio of the posterior odds to the prior odds: $(\Pr(M_i|\mathbf{y}) \,/\, \Pr(M_j|\mathbf{y})) \,/\, (\Pr(M_i)/\Pr(M_j))$. If we assume equal prior probabilities for each model then this simplifies to the "Bayes factor" (see Kass and Raftery, 1995): $B_{ij} = \Pr(M_i|\mathbf{y})/\Pr(M_j|\mathbf{y}) = \Pr(\mathbf{y}|M_i)/\Pr(\mathbf{y}|M_j)$, so that the model with the higher Bayes factor also has the higher posterior probability. Schwarz (1978) and Kass and Wasserman (1995) showed that in many kinds of models B_{ij} can be roughly approximated by $\exp(-\frac{1}{2}BIC_i + \frac{1}{2}BIC_j)$, where BIC (sometimes called SIC) is (1) with $A_n = \ln(n)$, especially if a certain "unit information" prior is used for the coefficients. Thus the model with the highest posterior probability is likely the one with lowest BIC. BIC is described further in Raftery (1995), Wasserman (2000), and Weakliem (1999).

BIC is sometimes preferred over AIC because BIC is "consistent." Assuming that a fixed number of models are available and that one of them is the true model, a consistent selector is one which will select the true model with probability approaching 100% as $n \to \infty$. In this context, the *true* model is the smallest adequate model (i.e., the single model that minimizes KL distance, or the smallest such model if there is more than one; Claeskens and Hjort, 2008). AIC is not consistent because it has a non-vanishing chance of choosing an unnecessarily complex model as n becomes large.

Researchers have proposed benchmarks for judging whether the size of a difference in AIC or BIC between models is practically significant (see Burnham and Anderson, 2004; Raftery, 1995). For example, an AIC difference between two models of less than 2 provides little evidence for one over the other; an AIC difference of 10 or more is strong evidence.

Criteria Related to BIC. Sclove (1987) suggested an adjusted BIC, abbreviated as ABIC or BIC*, based on the work of Rissanen (1978) and Boekee and Buss (1981). It uses $A_n = \ln((n+2)/24)$ instead of $A_n = \ln(n)$. This penalty will be much lighter than that of BIC, and may be lighter or heavier than that of AIC, depending on n. The unusual expression for A_n comes from Rissanen's work on model

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

selection for autoregressive time series models from a minimum description length perspective (see Stine, 2004, for a review). It is not immediately clear whether or not the same adjustment is still appropriate in different contexts. Also similar to BIC (despite its name) is the CAIC, the "corrected" or "consistent" AIC proposed by Bozdogan (1987), which uses $A_n = \ln(n) + 1$. (It should not be confused with the AIC_c discussed earlier.) This penalty tends to result in a more parsimonious model, and more underfitting, than BIC with $A_n = \ln(n)$. This A_n was chosen somewhat arbitrarily and described as one example of an A_n that would provide model selection consistency. However, any A_n proportional to $\ln(n)$ provides model selection consistency, so CAIC has no clear advantage over the better-known BIC.

1 INFORMATION CRITERIA IN SIMPLE CASES

The fact that the criteria described here all have the form of (1), except for different A_n, leads to two insights into how they will behave in various situations. First, *when comparing several different models of the same size* (e.g., different five-predictor subsets in regression subset selection) all criteria of the form (1) will always agree on which model is best. Each will select whichever model has the best fitted likelihood (the best fit to the observed sample, e.g., best R^2, lowest error estimate, lowest deviance). This is because only the first term in (1) will differ across the candidate models, so A_n does not matter.

Second, *for a nested pair of models, different ICs act like different α levels on a likelihood ratio test*. For comparing models of different sizes, when one model is a restricted case of the other, the larger model will typically offer better fit to the observed data at the cost of needing to estimate more parameters. The criteria will differ only in how they make this tradeoff (Lin and Dayton, 1997; Sclove, 1987). Thus, an IC will act like a hypothesis test with a particular α level (Söderström, 1977; Teräsvirta and Mellin, 1986; Pötscher, 1991; Claeskens and Hjort, 2008; Foster and George, 1994; Stoica et al., 2004; Leeb and Pötscher, 2009; van der Hoeven, 2005). Suppose a researcher will choose whichever of M_0 and M_1 has the better (lower) value of an IC of the form (1). This means that M_1 will be chosen if and only if $-2\ell_1 + A_n p_1 < -2\ell_0 + A_n p_0$, where ℓ_1 and ℓ_0 are the fitted maximized log-likelihoods for each model. Although it is interpreted differently, algebraically this comparison is the same as a likelihood ratio (LR) test (Leeb & Pötscher, 2009, p. 900; Pötscher 1991; Söderström, 1997; Stoica, Selén, and Li, 2004) rejecting M_0 if and only if

$$-2(\ell_0 - \ell_1) > A_n(p_1 - p_0). \tag{2}$$

The left-hand side is the classic LR test statistic (since a logarithm of a ratio of quantities is the difference in the logarithms of the quantities). Thus, in the case of nested models an IC comparison is mathematically a null hypothesis test with a different interpretation. The α level is specified indirectly through the critical value A_n; it is whatever proportion of the null hypothesis distribution of the LR test statistic is less than A_n.

For many kinds of models, the asymptotic H_0 distribution of $-2(\ell_0 - \ell_1)$ is asymptotically χ^2 with degrees of freedom (df) equal to $p_1 - p_0$. Consulting a χ^2 table and assuming $p_1 - p_0 = 1$, AIC ($A_n = 2$) becomes equivalent to a LR χ^2 test at an α level of about .16 (i.e., the probability of a χ^2_1 deviate being greater than 2). In the same situation, BIC (with $A_n = \ln(n)$) has an α level that depends on n. If $n = 10$ then $A_n = \ln(n) = 2.30$ so $\alpha = .13$. If $n = 100$ then $A_n = 4.60$ so $\alpha = .032$. If $n = 1000$ then $A_n = 6.91$ so $\alpha = .0086$, and so on. With moderate or large n, significance testing at the customary level of $\alpha = .05$ is an intermediate choice between AIC and BIC, because the square of a standard normal variate is a χ^2 variate with one degree of freedom, so that the two-sided critical z of 1.96 is equivalent to a critical χ^2 of $A_n = 1.96^2 \approx 4$.

For AIC, the power of the test increases with n, so that rejecting any given false null hypothesis is essentially guaranteed for sufficiently large n even if the effect size is tiny; however, the Type I error rate is constant and never approaches zero. On the other hand, BIC becomes a more stringent test (has a lower Type I error rate) as n increases. Thus, nonzero but tiny effects are less likely to lead to rejecting the null hypothesis for BIC than for AIC (see Raftery, 1995). Fortunately, even for BIC, the decrease in α as n increases is slow; thus power still increases as n increases, although more slowly than it would for AIC. Thus, for BIC, both the Type I and Type II error rates decline slowly as n increases, while for AIC (and for classical significance testing) the Type II error rate declines more quickly but the Type I error rate does not decline at all. This is intuitively why a criterion with constant A_n cannot be asymptotically consistent even though it may be more powerful for a given n (see Claeskens & Hjort, 2008; Kieseppä, 2003; Yang, 2005).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

Nylund et al. (2007) seem to interpret the lack of consistent selection as a flaw in AIC (Nylund et al., 2007, p. 556). An alternative view is that AIC attempts to find the model with good performance in some predictive sense. If n is small, then we may have to choose a smaller model to get more reliable coefficient estimates. However, if n is large, then standard errors will be small and one can afford to use a rich model. Thus, from an AIC perspective and for large n, Type II error (an underfit model) is considered worse than a Type I error (overfit model). In contrast, with BIC we are willing to take a higher risk of choosing too small a model, to improve our probability of choosing the true (smallest correct) model. *BIC considers Type I and Type II errors to be about equally undesirable* (Schwarz, 1978), while *AIC considers Type II errors to be more undesirable than Type I errors* unless n is very small. Neither perspective is always right or wrong. AIC is not a defective BIC, nor vice versa (see Kieseppä, 2003; Shibata, 1981, 1986).

In classical hypothesis testing, overfitting (Type I errors) are considered worse than Type II errors, because the former is considered an incorrect statement while the latter is simply a "failure to reject." However, for prediction, a Type II error can be quite harmful. One might almost say that for traditional hypothesis testing, a model is wrong when it is too large, while for estimation it is wrong when it is too small. Sometimes the relative importance of sensitivity or specificity depends on the decisions to be made based on model predictions. For example, in some environmental or epidemiological contexts, Type II error might be much more harmful to public health. Another way to characterize the comparison of two nested models is by analogy to a medical diagnostic test (see, e.g., Altman and Bland, 1994), replacing "Type I error" with "false positive" and "Type II error" with "false negative." AIC and BIC use the same data, but apply different cutoffs for whether to "diagnose" the smaller model as being inadequate. AIC is more sensitive (lower false negative rate) but BIC is more specific (lower false positive rate). The utility of each cutoff is determined by the consequences of a false positive or false negative, and by one's beliefs about the base rates of positives and negatives. Thus, AIC and BIC represent different sets of prior beliefs in a Bayesian sense (see Burnham and Anderson, 2004; Kadane and Lazar, 2004) or, at least, different judgments about the importance of parsimony.

Consider a simple classic scenario: a comparison of the means of two independent samples of equal size, assumed to have equal variance. One sample might be the control group and the other the experimental group in a randomized clinical trial. (See also Claeskens and Hjort, 2008, pp. 102-106, who work though an even simpler example in greater detail.) In our example two models are being compared:

$$\begin{aligned} M_0: &\quad y_{ij} = \beta_0 + e_{ij} \\ M_1: &\quad y_{ij} = \beta_0 + \beta_1 x_i + e_{ij} \end{aligned} \tag{3}$$

where the e_{ij} are independent normal errors with error variance σ^2, and x_i is dummy-coded 0 for the control group and 1 for the experimental group, so that β_1 is the effect of being in the experimental group. M_0 is nested within M_1 and has one fewer parameter, since M_0 sets $\beta_1 = 0$ in the expression for M_1. In a null hypothesis test such as a two-group independent-samples t-test, the smaller model M_0 becomes the null hypothesis H_0 and the larger model M_1 becomes the alternative hypothesis H_1. However, suppose an investigator decided to choose between these models using an IC. From the discussion in the previous subsection, the "test" based on an IC would be equivalent to a LR χ^2 test, with some α determined implicitly by A_n in (1). Here the implied LR χ^2 test would be approximately equivalent to a z-test (i.e., a naïve t-test treating the standard deviation as known; recall that a pooled variance t-statistic is asymptotically equivalent to a z-statistic, and the square of a z random variate is a χ^2.) Thus, using AIC in this case would be much like a t-test or z-test, although with an α of about $\Pr(\chi^2_1 > 2) \approx .16$ instead of the customary .05. Thus, if H_0 in (3) were true, AIC would erroneously favor H_1 about 16% of the time. In contrast, the α levels of the BIC-like criteria would be much smaller and would decrease as n increases. BIC and CAIC are more parsimonious than AIC at any reasonable n, that is, they have lower α and thus necessarily lower power. ABIC has an α which changes with n as BIC does, but is much less parsimonious (higher α) than original BIC at any n. Importantly, this simple example shows why no IC can be called best in an unqualified sense. Since choosing A_n in this simple situation is equivalent to choosing an α level for a significance test, the universally "best" IC cannot be defined any more than the "best" α.

For nested models differing by more than one parameter, one can still express an IC as an LR test using (2). The only difference is that the χ^2 distribution for finding the effective asymptotic α level now has df higher than 1. Thus for example, AIC might be less liberal when comparing a $(p+5)$-parameter model to a p-parameter model than it would be in comparing a $(p+1)$-parameter model to a p-parameter

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

model, since $\Pr(\chi^2_5 > 2 \times 5) \approx .075$ while $\Pr(\chi^2_1 > 2 \times 1) \approx .16$. Similarly, the α level for BIC also depends on the difference in number of parameters, but with moderate or high n it will still be lower than that of AIC. It is often difficult to determine the α value that a particular criterion really represents in a particular situation, for two reasons. First, even for regular situations in which a LR test is known to work well, the χ^2 distribution for the test statistic is asymptotic and will not apply well to small n. Second, in some situations the rationale for using an IC is, ironically, the failure of the assumptions needed for a LR test. That is, the test emulated by the IC will itself not be valid at its nominal α level anyway. Therefore, although the comparison of A_n to an α level is helpful for getting a sense of the similarities and differences among the ICs, simulations are required to describe exactly how they behave.

2 LATENT CLASS ANALYSIS CASE STUDY

A very common use of ICs by social and behavioral scientists is in selecting the number of components for a finite mixture model. Many kinds of models, such as latent class analysis (LCA; see Lazarsfeld and Henry, 1968; Collins and Lanza, 2010), suppose that the population is composed of multiple classes of a categorical latent variable. Before the parameters for these classes can be estimated, it is first necessary to determine the number of classes. Sometimes one might have a strong theoretical reason to specify the number of classes, but often this must be done using the data. In this section we consider LCA with categorical indicator variables as in Collins and Lanza (2010), although conceptually the problem is similar for some other finite mixture models.

A naïve approach would be to use LR or deviance (G^2) tests sequentially to choose the number of classes, and conclude that the k-class model is large enough if and only if the $(k+1)$-class model does not fit significantly better. The selected number of classes would be the smallest k that is not rejected. However, the assumptions for the supposed asymptotic χ^2 distribution in a LR test are not met in the setting of LCA, so that the p-values from those tests are not valid (see Lin & Dayton, 1997; MacLachlan & Peel, 2000). H_0 here is not nested in a regular way within H_1, since a k-class model is obtained from a $(k+1)$-class model either by constraining any one of the class sizes to a boundary value of zero or by setting the class-specific item-response probabilities equal between any two classes. Ironically, the lack of regular nesting structure that makes it impossible to decide on the number of classes with an LR test also invalidates approximations used in the AIC and BIC derivations (McLachlan & Peel 2000, pp. 202-212). From (2), comparing any two values of k using ICs is algebraically the same as using a naïve likelihood ratio test with some α level. Nonetheless, ICs are widely used in LCA and other mixture models. Asymptotically, when the true model is well-identified, BIC (and hence also AIC and ABIC) will at least not underestimate the true number of components (Leroux 1992; McLachlan & Peel 2000, p. 209).

Lin and Dayton (1997) compared the performance of AIC, BIC, and CAIC for LCA models in simulations. Instead of exploring the number of classes, they were comparing k-class models which differed in complexity (the number of parameters needed to describe the model from which the data was generated) due to different sets of simplifying assumptions. When a very simple model was used as the true model, BIC and CAIC were more likely to choose the true model than AIC, which tended to choose an unnecessarily complicated one. When a more complex model was used to generate the data and measurement quality was poor, AIC was more likely to choose the true model than BIC or CAIC, which were likely to choose an overly simplistic one. They explained that this was very intuitive given the differing degrees of emphasis on parsimony.

Other simulations have explored the ability of the ICs to determine the correct number of classes. In Dias (2006), AIC had the lowest rate of underfitting but often overfit, while BIC and CAIC practically never overfit but often underfit. AIC3 was in between and did well in general. The danger of underfitting increased when the classes did not have very different response profiles; in these cases BIC and CAIC almost always underfit. Yang (2006) reported that ABIC performed better in general than AIC (whose model selection accuracy never got to 100% regardless of n) or BIC or CAIC (which underfit too often and required large n to be accurate). Yang and Yang (2007) compared AIC, BIC, AIC3, ABIC and CAIC. When the true number of classes was large and n was small, CAIC and BIC seriously underfit, but AIC3 and ABIC performed better. Nylund et al. (2007) presented various simulations on the performance of various ICs and tests for selecting the number of classes in LCA, as well as factor mixture models and growth mixture models. Overall, in their simulations, BIC performed much better than AIC (which tended to overfit) or CAIC (which tended to underfit) (Nylund et al., 2007, p. 559). However, this does not

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

mean that BIC was the best in every situation. In most of the scenarios, BIC and CAIC almost always selected the correct model size, while AIC had a much smaller accuracy in these scenarios because of a tendency to overfit. In those scenarios, n was large enough so that the lower sensitivity of BIC was not a problem. However, in a more challenging scenario (8-item, unequally sized classes, $n = 200$ on p. 557) BIC essentially never chose the larger correct model and it usually chose one which was much too small. Thus, as Lin and Dayton (1997) found, BIC may select too few classes when the true population structure is complex but subtle (somewhat like a small but nonzero effect in our z-test example) and n is small. Wu (2009) compared the performance of AIC, BIC, ABIC, CAIC, naïve tests, and the bootstrap LR test in hundreds of different simulated scenarios. Performance was heavily dependent on the scenario, but the method that worked adequately in the greatest variety of situations was the bootstrap LR test, followed by ABIC and classic BIC. Wu (2009) speculated that BIC seemed to outperform ABIC in the most optimal situations because of its parsimony, but that ABIC seemed to do better in situations with smaller n or more unequal class sizes. It is not possible to say which IC is universally best, even in the idealized world of simulations. Rather, the true parameter values and n used when generating simulated data determine the relative performance of the ICs. Even within a given scenario, the relative performance of criteria depends on which aspect of performance is being discussed. Below we illustrate these ideas with simulated examples.

In a real-data analysis described in Collins and Lanza (2010, p. 12), adolescents ($n = 2087$) were asked whether or not, in the past year, they had participated in each of six kinds of behavior considered delinquent. On this basis, they used LCA to postulate four classes: non-delinquent (49% of adolescents), verbal antagonists (26% of adolescents), shoplifters (18%), and general delinquents (6%). We used this as the starting point for a simulation study to show how the behavior of ICs differs by penalty weight, sample size, and true model values. For each value of n in $100, 200, 400, 600, \ldots, 1800$, we generated 1000 random datasets each with sample size n, using the estimates from Collins and Lanza (2010) as the true population values. Latent class models specifying 1, 2, 3, 4, 5, or 6 classes were fit to these datasets, and the class sizes selected by AIC, BIC, AIC3, ABIC, and CAIC were recorded. In addition to the 4-class model described in Collins and Lanza (2010), an alternative 3-class scenario was also used to conduct another set of simulations. For this, we blended the last two classes into one class, with response probabilities determined by a weighted average. The smallest class has prevalence 24.5% in the three-class model, compared to about 6% in the four-class model.

The ICs were compared on four performance measures: proportion overfitting (choosing more than the correct number of classes), proportion underfitting (choosing less than the correct number of classes), proportion correct fitting (choosing the correct number of classes), and estimation error, measured as the square root of the mean (over replications and cells) of the squared estimation error for the cell probabilities of the underlying contingency table. To explain this fourth performance measure, recall that an LCA model can be seen as an approach to modeling a contingency table, in this case the $2 \times 2 \times 2 \times 2 \times 2 \times 2$ table of relative frequencies of each possible combination of responses to the six delinquency questions. Thus, we defined mean squared model error in this example as $SSE = \sum_j (\hat{P}_j - P_j)^2$ over all $2^6 = 64$ cells j in the contingency table, where $\hat{P}$ is the estimated probability of the cell and P is the true probability. We used this as a basis for a root-mean-squared error measure of LCA performance in estimating the contingency table: specifically, the average across replications of $\sqrt{SSE/2^6}$ for a given method at a given n. The result can be roughly interpreted as an average empirical standard error measure for the individual cell probabilities P_j. The data were generated in SAS and analyzed using PROC LCA (Lanza, Collins, Lemmon & Schafer, 2007; Lanza, Dziak, Huang, Xu, & Collins, 2011). The simulation results are shown as graphs with linear interpolation in Figures 1 and 2.

The results in the three-class scenario were mostly as expected. Underfitting rates were very high for small n but they quickly declined. BIC and CAIC had the highest underfitting rates. Overfitting rates were low, except for AIC and for ABIC with small n. In terms of mean squared error (MSE), all of the criteria chose models that estimated the contingency table well when n was large, but the more sensitive criteria (AIC, ABIC, AIC3) did better than the more specific criteria (BIC, CAIC) for smaller n. *If the true model is selected*, the estimates do not depend on which criterion was used to select the model. However, the more parsimonious criteria had poorer estimation quality *on average* because they underfit more often. The $\sqrt{MSE}$ values in Figure 2 appear very small, but recall that the cell proportions being estimated are themselves small (averaging $1/64$).

In the four-class scenario, the true model was much harder for the criteria to detect, largely because

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

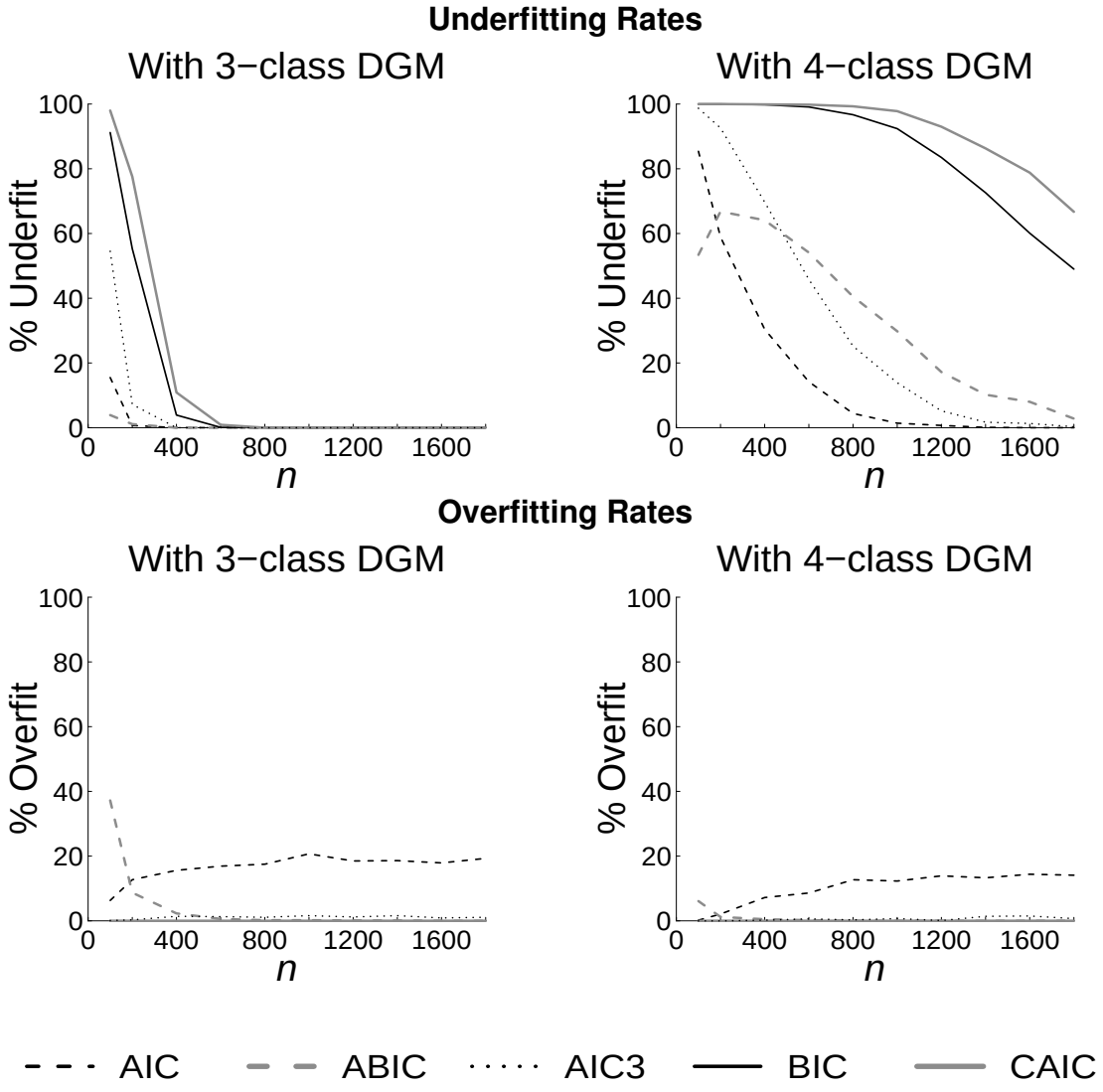

Figure 1. Underfitting and overfitting rates of information criteria for LCA example. "DGM" denotes data-generating model (simulated true model).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

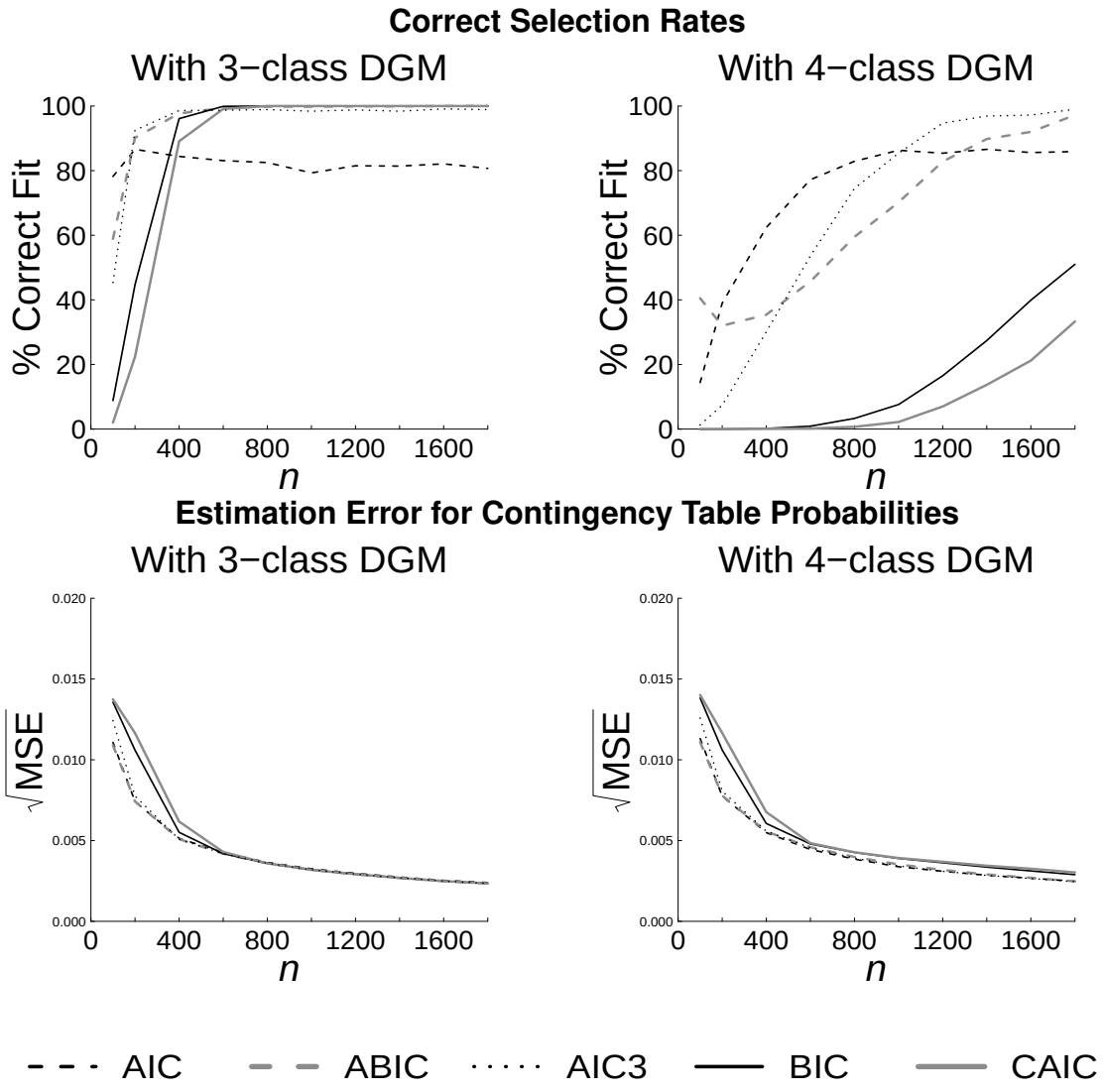

Figure 2. Correct model selection rates and root mean squared estimation error of information criteria for LCA example. "DGM" denotes data-generating model (simulated true model).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

one of the classes was so small (see Wu, 2009). Underfitting rates were much slower to decline in n; it was much harder to detect all four classes here than to detect all three classes earlier. Overfitting rates were very small except for AIC. Because of substantial underfitting, BIC and CAIC now had a much poorer accuracy for choosing the correct number of classes than did AIC, ABIC, or AIC3. For $n = 1400$, BIC and CAIC still underfit more than 50% of the time although AIC, ABIC or AIC3 usually selected the correct model. All criteria did a good job of estimating the underlying contingency table when n was large enough, although BIC and CAIC required a larger n.

The very limited simulation studies described here cannot provide general rules for all situations, but they provide useful heuristics. If the goal of having a sufficiently rich model to describe the heterogeneity in the population is more important than parsimony, or if some classes are expected to be small or similar to other classes but distinguishing among them is still considered important for theoretical reasons, then perhaps AIC, AIC3 or ABIC should be used instead of BIC or CAIC. If obtaining a few large and distinct classes is more important, then BIC is more appropriate. Sometimes, the AIC-favored model might be so large as to be difficult to use or understand, so the BIC-favored model is a better choice (e.g., in Chan, Leu, and Chen, 2007, BIC favored a mixture model with 5 classes and AIC favored at least 10; the authors chose the BIC-favored model by other considerations).

Another possibility is to use a bootstrap LR test. Unlike the naïve LR test, Nylund et al. (2007) showed empirically that the bootstrap LR test with a given α level does generally provide a Type I error rate at or below that specified level. Like any test or criterion, it still requires the choice of a tradeoff between sensitivity and specificity (i.e., an α level). However, it is good to be able to choose and know the α level being used in a straightforward way. Both Nylund et al. (2007) and Wu (2009) found that this bootstrap test seemed to perform somewhat better than the ICs in various situations. Mplus can perform such a test (Muthén and Muthén, 2007). A bootstrap LR test is also available as a SAS macro in conjunction with PROC LCA; both are available at *methodology.psu.edu*. The bootstrap LR test is beyond the scope of this paper, as are more computationally intensive versions of AIC and BIC, involving bootstrapping, cross-validation or posterior simulation (see McLachlan and Peel, 2000, pp. 204-212).

3 FACTOR ANALYSIS CASE STUDY

Another situation in which the behavior of ICs may be compared is in choosing the number of factors in an exploratory factor analysis. Akaike (1987) and Song and Belin (2007) further describe the use of AIC and BIC in factor analysis, and explore how to handle missing data in this context. It is beyond the scope of this paper to recommend for or against the use of ICs, as opposed to other methods such as the likelihood ratio χ^2 difference test (see Akaike, 1987; Hayashi et al., 2007) or retaining as many factors as there are eigenvalues greater than one (see Kaiser, 1960). However, exploratory factor analysis provides another straightforward setting for a simulated case study in which we can compare the behavior of the ICs.

Harman's classic "24 psychological tests" dataset (Harman, 1976) is a correlation matrix of psychometric test scores, which is available in the `datasets` library in R (R Development Core Team, 2010). We used it to create two scenarios, one with a 3-factor structure and the other with a 4-factor structure. For each scenario, we did an exploratory factor analysis on the original correlation matrix, specifying the desired number of factors, then used the correlation matrix implied by the estimates from this analysis as though it were a true population model, from which we generated random samples. For each n value in 50, 60, 70, …, 600 we simulated 1000 datasets of size n under both the 3-factor and 4-factor models. To each dataset we fit 1-, 2-, 3-, 4-, and 5-factor models and calculated the AIC, BIC, AIC3, ABIC and CAIC.

We calculated overfitting and underfitting rates for each criterion, as well as the square root of the mean (across replications and parameters) squared error (MSE) of estimation for estimating the parameters of the true 3- or 4-factor population covariance matrix. The overfitting rate was defined as the proportion of simulations in which an IC selected more factors than were the data-generating model. Underfitting rate, similarly, was the proportion of simulations in which an IC indicated fewer. The final performance measure, $\sqrt{\text{MSE}}$, tells how well the selected model describes the population. Simulations and graphs were done in R. The results are shown in Figure 3 and 4. Underfitting rates are initially very high for most of the criteria, but they quickly decline as n (thus statistical power) increases. For small n, BIC and CAIC were more likely to underfit than AIC, with ABIC and AIC3 intermediate. For n over 300 or 400, none of the methods underfit very often. Overfitting rates were practically zero for all of the methods except AIC, which had an overfitting rate around 15% regardless of n. ABIC and AIC3 performed quite well

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

in terms of overall model selection accuracy. For all methods, estimation error declined as n increased, although for small n it was lower (better) for AIC and ABIC than for BIC or CAIC. For $n > 300$, each IC performed about equally well.

In both the LCA and factor analysis examples, the MSE for moderate n was sometimes better for AIC than for BIC. This is because the cost to MSE of fitting too large a model is often less than that of fitting too small a model, as shown in Figure 5. In this figure, as long as the candidate model is at least as rich as the data-generating model (i.e., has at least as many classes or factors as the true data-generating model), estimation performance is good. The best performance (lowest MSE) is obtained from the exactly correct model size, but models that are slightly too rich might have performance that is almost as good. Overly simple models sometimes have much poorer performance because they represent an excessive constraint on the implied covariance estimate of the variables. As an extreme example, suppose an LCA investigator wrongly assumes that there is only one class, when in reality the sample comes from a multiple-class population. Under the one-class model, the standard LCA assumption that responses are independent conditional on the latent variable (see Collins and Lanza, 2010) becomes an assumption that all responses are *unconditionally* independent. It would be impossible for this model to give good predictions about how the variables are related in the population. For the opposite extreme, suppose the investigator assumes that there are very many classes, one for each observed response profile. The model would be useless in practice because it is as complex as the original dataset. However, the implied estimates of the population contingency table and of the correlation matrix of the indicator variables would be unbiased, since they would be based on a direct empirical estimate of the relative frequency of each possible response profile. That is, the population might still be described fairly accurately, although uninterpretably and with high sampling error.

4 MULTIPLE REGRESSION CASE STUDY

The previous examples involved situations in which a small sequence of models of increasing complexity were considered to model the covariance of a given set of variables. Thus, the candidate models could be treated as nested in at least an informal or conceptual way, even if the regularity conditions for a test of nested hypotheses might not be met. Thus, comparing the ICs in terms of sensitivity and specificity, as if they were tests, was straightforward. The question arises of how well this generalizes to situations with a much larger range of candidate models which are not all arranged in order, as in subset selection for regression (reviewed by Hocking, 1976; George, 2000; Miller, 2002).

One way to study subset selection is to consider a simple special case. If all of the predictors in a normal linear regression model are mutually uncorrelated (orthogonal) and responses have constant variance σ^2, then the contribution of each predictor to the fit can be considered independently. Thus subset selection with ICs becomes equivalent to significance tests on each predictor variable, and the different ICs become different alpha levels. That is, predictor j will be included if and only if it is significant using an F test with critical value A_n, or t test with critical value $\sqrt{A_n}$, from (1) (Foster and George, 1994, p. 1947). Then the different A_n can be interpreted as determining the α levels for a test, or the thresholds for how large a β estimate has to be to include it in the model.

In real research, the predictors are usually far from independent. Thus, model selection is not equivalent to a test of each coefficient separately. Since the variance proportions in the response variable accounted for by different variables now overlap, the importance of a given variable depends on what other variable is in the model. Specifically, how the importance of a predictor variable is determined depends on the selection method being used, for example, whether (1) marginal significance testing on the full model, (2) a forward, backward or combined stepwise approach, or (3) an all-subsets approach is used. Trimming the model using significance tests on each variable separately (see Freedman and Pee, 1989), although perhaps common, ignores the ways in which the importance of one variable may depend on which others are considered. In contrast, a stepwise approach does consider which other variables are included when evaluating a new variable. Since it considers nested pairs of variables at a time, stepwise selection by ICs is much like stepwise testing by t or F tests. A disadvantage of stepwise approaches is that they only evaluate a small fraction of the possible number of subsets available, those on or adjacent to the stepwise path used. Last, an all-subsets approach tries every model in the space, but since there are so many (2^p if there are p predictors available), the selection of the best may be strongly influenced by chance sampling variance. Thus, it is harder to describe how prediction methods will behave, except asymptotically as $n \to \infty$.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

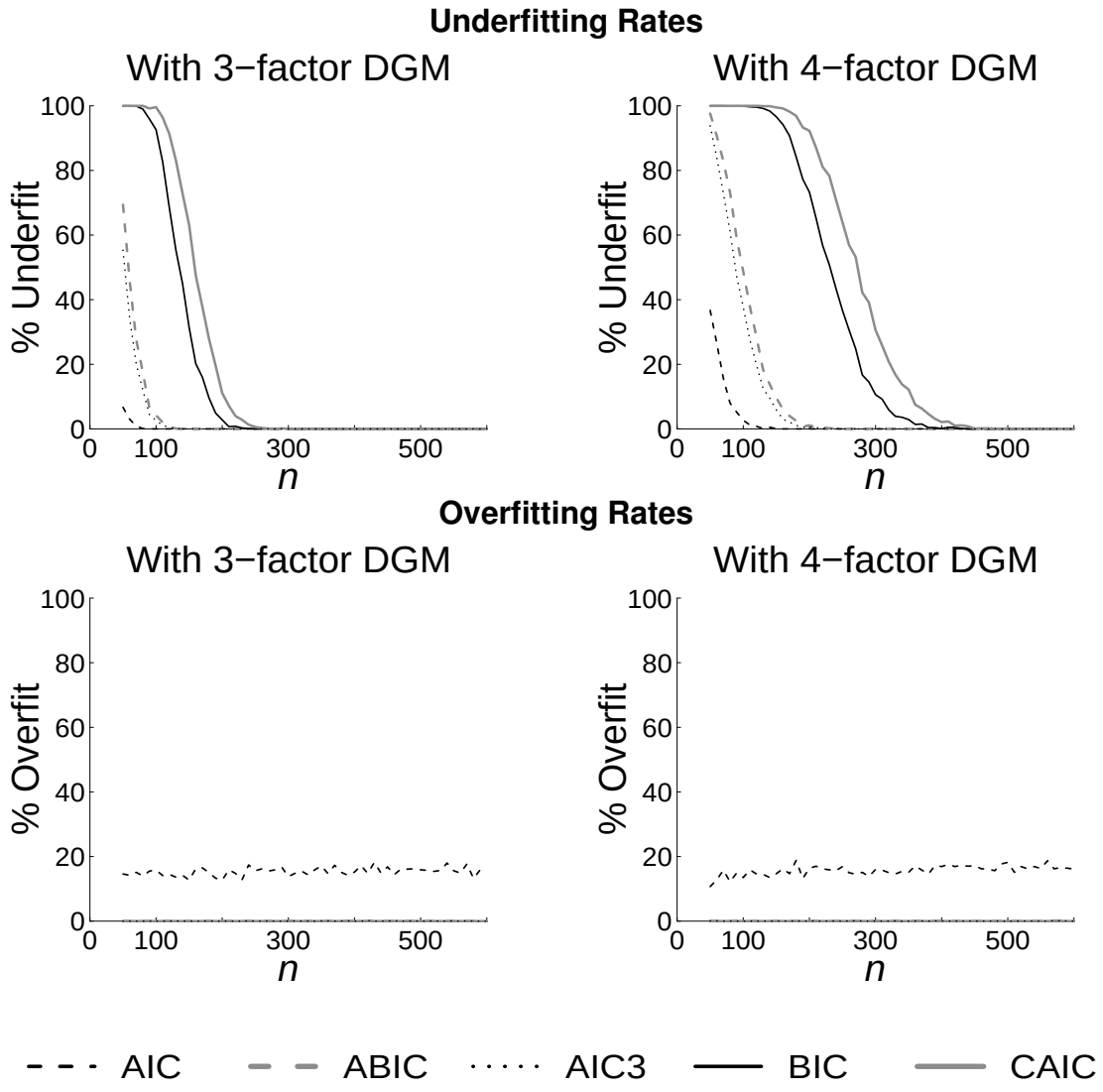

Figure 3. Underfitting and overfitting rates of information criteria for factor analysis example. "DGM" denotes data-generating model (simulated true model).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

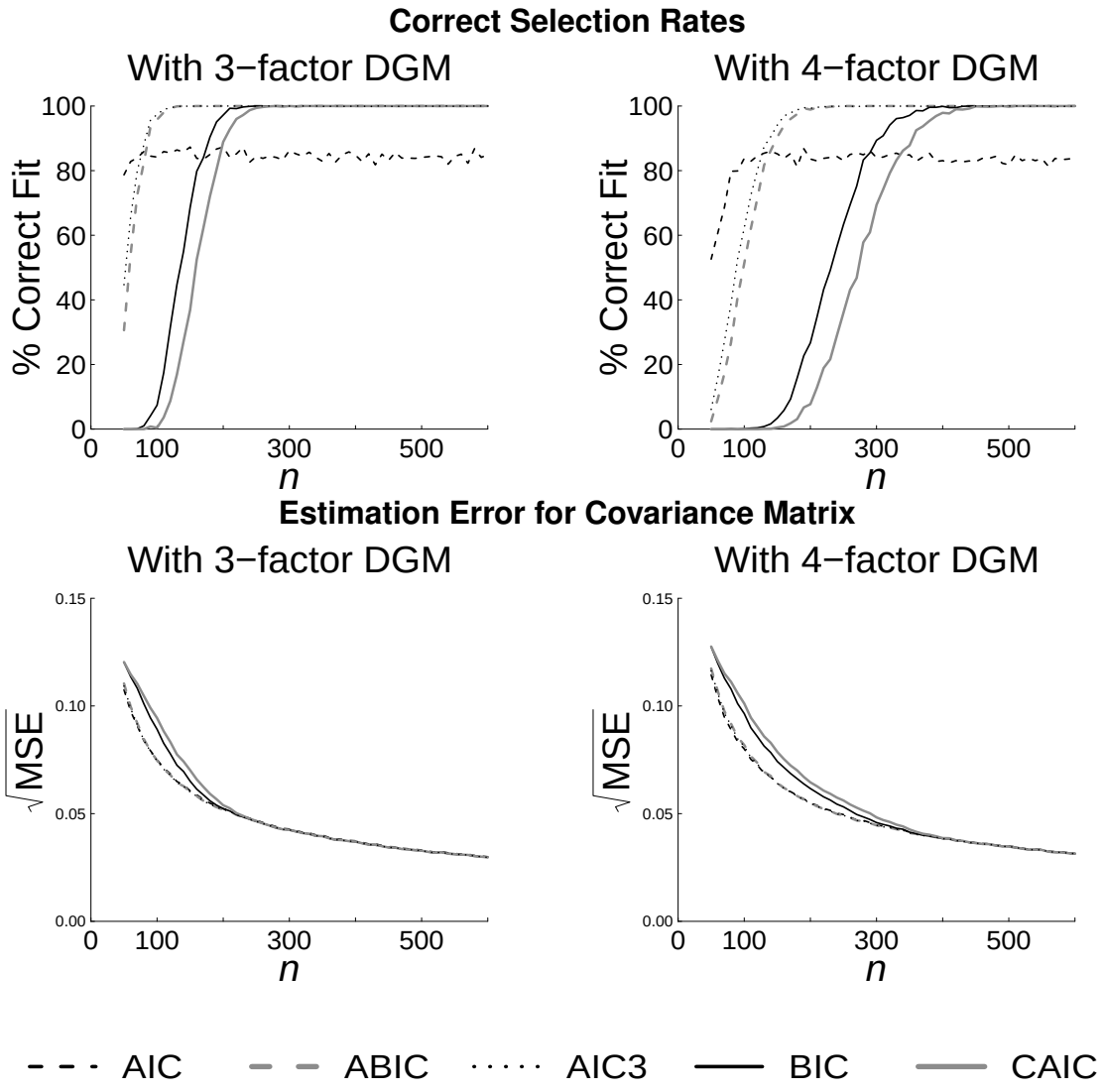

Figure 4. Correct model selection rates and root mean squared estimation error of information criteria for factor analysis example. "DGM" denotes data-generating model (simulated true model).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

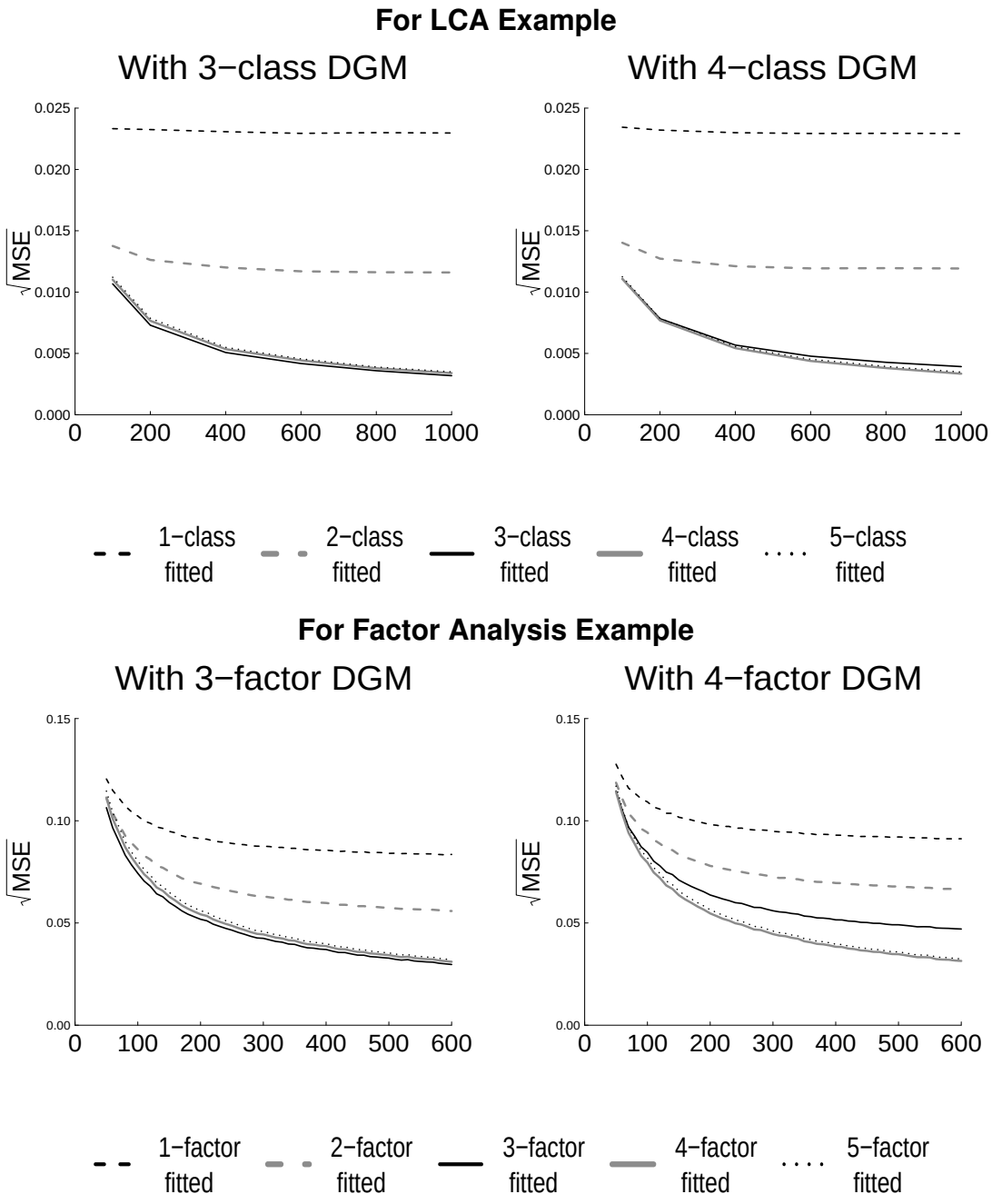

Figure 5. Estimation performance by models of various fixed sizes in the LCA and factor analysis examples. "DGM" denotes data-generating model (simulated true model).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

The asymptotic properties of the ICs for variable selection are elegantly reviewed by Shao (1997) and Yang (2005). Unfortunately, the relative superiority of AIC or BIC depends on the specific assumptions being made in the asymptotic scenario, in ways that can be challenging to interpret for real data when n is finite and the form of the true model unknown. If a small group of models are being compared and one is exactly true, then BIC will find that model with high probability for sufficiently high n. However, if a range of model sizes are being compared but no finite size provides an exactly correct parametric model, AIC will do better than BIC in terms of estimation error because BIC is less likely to choose a rich enough model to approximate the truth.

Furthermore, in all-subsets approaches with many predictors, there are so many models to evaluate (often more than the number of subjects available) that it is infeasible to claim that the best subset for the sample is the best for the population, even supposing such a unique best subset exists (see Burnham and Anderson, 2002; Chatfield, 1995; Zucchini, 2000). Therefore, ICs are sometimes used in assigning weights in a model averaging approach rather than in choosing a single best model (see Burnham and Anderson, 2002; Claeskens and Hjort, 2008). Subset selection techniques can be useful when the number of predictors is large and little prior information exists, but they have been criticized in the social sciences due to their automaticity and blindness to theoretical interpretability. Where theoretical information is available, it might be better to consider only a relatively few models considered meaningful, interpretable and credible (see Burnham and Anderson, 2002; Thompson, 1995).

To again explore the similarities and differences between the ICs, we consider a multiple linear regression simulation case study. To create the simulated datasets, we returned to Harman's example. We considered the problem of predicting the Arithmetic Problems score from ten other scores in a regression (General Information, Paragraph Comprehension, Word Classification, Word Meaning, Addition, Code, Counting Dots, Deduction, Problem Reasoning, Series Completion).

Under the covariance matrix provided by Harman's dataset, the regression coefficients for predicting Arithmetic Problems from the other ten variables of interest (an intercept-only model was assumed because the scores were all standardized) were approximately -0.0226, 0.1532, 0.0053, 0.1183, 0.3542, 0.0537, 0.0912, 0.0650, 0.0985, 0.0947. The regression coefficients for predicting Arithmetic Problems from only the two best predictors, Paragraph Comprehension and Addition, were 0.3392 and 0.4621. We thus constructed two scenarios for simulating datasets. In the "diffuse true model" scenario, the ten predictors are assumed to be normally distributed with mean 0 and have the same population covariance as the observed covariance in the dataset. The response is assumed to be normal, with a mean generated under the assumption that population regression coefficients are equal to the sample regression coefficients for the ten-predictor model for the original data. In the "sparse true model" scenario, the ten predictors have the same distribution, but the regression coefficients for the response are all 0 except Paragraph Comprehension at 0.3392 and Addition at 0.4621. Thus in the first scenario, there are many small but nonzero β values, so sensitivity is important. In the second, each β is either large or zero, so sensitivity is less vital and specificity is rewarded. In each scenario, the assumed unconditional population variance for each of the predictors was 1. The error variance for the response, conditional upon the predictors, was chosen at each replication so that the total sample variance of y was 1. For each of the two scenarios, and for each of $n = 50$, 100, 150, 200, 300, 400, ..., 1500, we simulated 2000 datasets. With 10 candidate predictors, there were $2^{10} = 1024$ possible subsets. For each of these subsets, a regression model was fit to each dataset and evaluated on AIC, BIC, ABIC, CAIC and AIC_c. (AIC3 was not used, since it seems to be primarily discussed for mixture models.) Also, for each selected model, its prediction performance (in terms of square root of mean squared error of prediction) in a sample of 10000 new cases was recorded. The results are shown in Figure 6.

Under the diffuse scenario, the true data-generating model had 10 predictors, but for small n all of the methods chose a much smaller subset. The number of predictors which they allowed into the model increased as n increased, but it was generally higher for AIC than for BIC and CAIC. ABIC was usually intermediate between AIC and BIC. AIC_c behaved almost the same as AIC because the number of predictors available was modest and n was not very small. Under the sparse scenario, both of the truly nonzero predictors were so strongly related to the response that even BIC and CAIC usually included them. Thus, BIC and CAIC generally chose the correct model. Here the relationship which was seen previously between the MSEs of the AIC-like and BIC-like criteria is surprisingly reversed, and the more specific criteria are more accurate because they happen to agree with the special structure of the true coefficients. Which scenario is more realistic is not clear and might depend on the context and field of

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

study.

We only considered linear regression in this paper. For nonparametric regression, both the predictors and the complexity of the relationship of each predictor to the response must be selected somehow. Irrelevant predictors may be much more harmful in nonparametric regression because each one may essentially involve fitting many parameters (see Hastie et al., 2001).

5 DISCUSSION

Most of the simulations shown here illustrated similar principles. AIC and similar criteria often risk choosing too large a model, while BIC and similar criteria often risk choosing too small a model. For small n, the most likely error is underfitting, so the criteria with lower underfitting rates, such as AIC, often seem better. For larger n, the most likely error is overfitting, so more parsimonious criteria, such as BIC, often seem better. Unfortunately, the point at which the n becomes "large" depends on numerous aspects of the situation. In simulations, the relative performance of the ICs at a given n depended on the nature of the "true model" (the distribution from which the data came). This finding is unhelpful for real data, where the truth is unknown. It may be more helpful to think about which aspects of performance (e.g. sensitivity or specificity) are most important in a given situation.

Our simulations were simplistic in some ways. We assumed there was a true model size for which the model assumptions fit exactly. Underfitting and overfitting could be defined as underestimating and overestimating the true number of classes or factors. For observed data for which models are only approximations to reality, more care is required in considering what it means for a model to be too small, correct, or too large (Burnham and Anderson, 2002, p. 32) Performance can be expressed in terms of a quantitative criterion such as MSE, avoiding the use of a "correct" size, but this may favor AIC-like over BIC-like criteria.

There is no obvious conclusion about whether or when to use ICs, instead of some other approach. Kadane and Lazar (2004) suggested that ICs might be used to "deselect" very poor models (p. 279), leaving a few good ones for further study, rather than indicating a single best model. One could use the ICs to suggest a range of model sizes to consider; for example, one could use the BIC-preferred model as a minimum size and the AIC-preferred model as a maximum, and make further choices based on other kinds of fit criteria, on theory, or on subjective inspection of the results (Collins & Lanza 2010). If BIC indicates that a model is too small, it may well be too small (or else fit poorly for some other reason). If AIC indicates that a model is too large, it may well be too large for the data to warrant. Beyond this, theory and judgment are needed.

ACKNOWLEDGMENTS

The authors thank Linda M. Collins for very valuable suggestions and insights which helped in the development of this paper. We also thank Michael Cleveland and Amanda Applegate for their careful review and recommendations.

Data analysis was done using the R 2.11 and SAS 9.1 statistical packages and graphs were done in R. The R software is copyright 2010 by The R Foundation for Statistical Computing. SAS software is copyright 2002-3 by SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

This report has been disseminated as Methodology Center Technical Report 12-119, June 27, 2012.

Vrieze (2012) published a paper in *Psychological Methods* which made some similar points to this report. Vrieze (2012) also points out that AIC and BIC differ only in their penalty weight and can be viewed as somewhat analogous to tests, that there is a tradeoff between the advantages and disadvantages of the AIC and BIC, that AIC and BIC are intended for different purposes and their performance is dependent on the performance measure and on the true model, that the BIC is "consistent" but at the price of a higher risk of underfitting, and that AIC may offer more accurate description of the population distribution but at the price of a nonvanishing risk of overfitting.

This research was supported by NIH grants P50 DA10075-14 and R21 DA24260 from the National Institute on Drug Abuse, and grant DMS-03048869 from the National Science Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Drug Abuse, the National Institutes of Health, or the National Science Foundation.

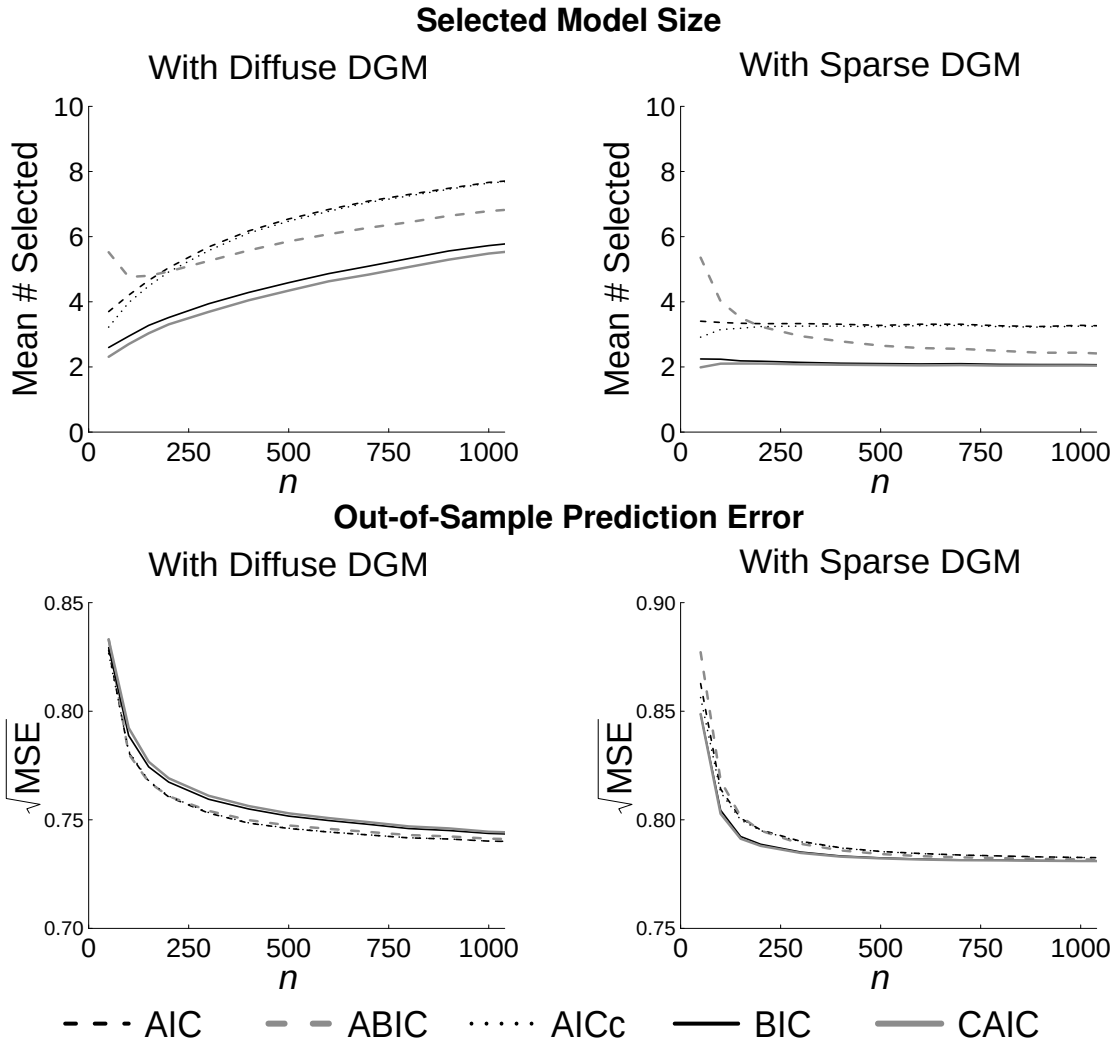

Figure 6. Average number of variables selected and root mean squared prediction error of the information criteria in the multiple regression example. "DGM" denotes data-generating model (simulated true model).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akademai Kiado, Budapest, Hungary.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52:317–332.

Altman, D. G. and Bland, J. M. (1994). Diagnostic tests 1: sensitivity and specificity. *British Medical Journal*, 308:1552.

Andrews, R. L. and Currim, I. S. (2003). A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research*, 40:235–243.

Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika*, 67:413–418.

Boekee, D. E. and Buss, H. H. (1981). Order estimation of autoregressive models. In *Proceedings of the 4th Aachen colloquium: Theory and application of signal processing*, pages 126–130.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52:345–370.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer-Verlag, New York, NY, 2nd edition.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33:261–304.

Chan, W.-H., Leu, Y.-C., and Chen, C.-M. (2007). Exploring group-wise conceptual deficiencies of fractions for fifth and sixth graders in Taiwan. *The Journal of Experimental Education*, 76:26–57.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A*, 158:419–466.

Claeskens, G. and Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge, New York, NY.

Collins, L. M. and Lanza, S. T. (2010). *Latent class and latent transition analysis for the social, behavioral, and health sciences*. Wiley, Hoboken, NJ.

Dias, J. G. (2006). Model selection for the binary latent class model: A Monte Carlo simulation. In Batagelj, V., Bock, H.-H., Ferligoj, A., and Žiberna, A., editors, *Data Science and Classification*, pages 91–99. Springer-Verlag, Berlin, Germany.

Fonseca, J. R. S. and Cardoso, M. G. M. S. (2007). Mixture-model cluster analysis using information theoretical criteria. *Intelligent Data Analysis*, 11:155–173.

Foster, D. P. and George, E. I. (1994). The Risk Inflation Criterion for multiple regression. *Annals of Statistics*, 22:1947–1975.

Freedman, L. S. and Pee, D. (1989). Return to a note on screening regression equations. *American Statistician*, 43:279–82.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association*, 95:1304–1308.

Harman, H. H. (1976). *Modern Factor Analysis*. University of Chicago, Chicago, IL, 3rd, rev. edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction*. Springer, New York, NY.

Hayashi, K., Bentler, P. M., and Yuan, K. (2007). On the likelihood ratio test for the number of factors in exploratory factor analysis. *Structural Equation Modeling*, 14:505–526.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, 32:1–49.

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

Kadane, J. B. and Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American Statistical Association*, 99:279–290.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20:141–151.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwartz criterion. *Journal of the American Statistical Association*, 90:928–34.

Kieseppä, I. A. (2003). AIC and large samples. *Philosophy of Science*, 70:1265–1276.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods and*

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

Research, 33:188–229.
Lanza, S. T., Collins, L. M., Lemmon, D. R., and Schafer, J. L. (2007). PROC LCA: A SAS procedure for latent class analysis. *Structural Equation Modeling*, 14:671–694.
Lanza, S. T., Dziak, J. J., Huang, L., Xu, S., and Collins, L. M. (2011). *PROC LCA & PROC LTA Users' Guide (Version 1.2.7 beta)*. The Methodology Center, Penn State, University Park, PA.
Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent structure analysis*. Houghton Mifflin, Boston.
Leeb, H. and Pötscher, B. M. (2009). Model selection. In Andersen, T., Davis, R. A., Kreiß, J.-P., and Mikosch, T., editors, *Handbook of Financial Time Series*. Springer, Berlin.
Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *Annals of Statistics*, 20:1350–1360.
Lin, T. H. and Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics*, 22:249–264.
McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley, New York.
Miller, A. J. (2002). *Subset selection in regression*. Chapman & Hall, New York, 2nd edition.
Muthén, L. K. and Muthén, B. O. (2007). *Mplus User's Guide*. Muthén & Muthén, Los Angeles, CA, 5th edition.
Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14:535–569.
Pitt, M. A. and Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, 6:421–425.
Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory*, 7:163–185.
R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25:111–163.
Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14:465–471.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.
Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52:333–43.
Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264.
Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, 68:45–54.
Shibata, R. (1986). Consistency of model selection and parameter estimation. *Journal of Applied Probability*, 23:127–141.
Söderström, T. (1977). On model structure testing in system identification. *International Journal of Control*, 26:1–18.
Song, J. and Belin, T. R. (2008). Choosing an appropriate number of factors in factor analysis with incomplete data. *Computational Statistics and Data Analysis*, 52:3560–3569.
Stine, R. A. (2004). Model selection using information theory and the MDL principle. *Sociological Methods & Research*, 33:230–260.
Stoica, P., Selén, Y., and Li, J. (2004). On information criteria and the generalized likelihood ratio test of model order selection. *IEEE Signal Processing Letters*, 11:794–797.
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.
Sugiura, N. (1978). Further analysis of the data by Akaike's Information Criterion and the finite corrections. *Communications in Statistics, Theory, and Methods*, A7:13–26.
Teräsvirta, T. and Mellin, I. (1986). Model selection criteria and model selection tests in regression models. *Scandinavian Journal of Statistics*, 13:159–171.
Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, 55:525–534.
Tibshirani, R. and Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society, B*, 61:529–546.
van der Hoeven, N. (2005). The probability to select the correct model using likelihood-ratio based criteria in choosing between two nested models of which the more extended one is true. *Journal of Statistical Planning and Inference*, 135:477–86.
Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). *Psychological*

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015

Methods, 17:228–243.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44:92–107.

Weakliem, D. L. (1999). A critique of the Bayesian Information Criterion for model selection. *Sociological Methods and Research*, 27:359–397.

Wu, Q. (2009). *Class extraction and classification accuracy in latent class models*. PhD thesis, Pennsylvania State University.

Yang, C. (2006). Evaluating latent class analysis models in qualitative phenotype identification. *Computational Statistics and Data Analysis*, 50:1090–1104.

Yang, C. and Yang, C. (2007). Separating latent classes by information criteria. *Journal of Classification*, 24:183–203.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? a conflict between model identification and regression estimation. *Biometrika*, 92:937–950.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, 44:41–61.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1103v1 | CC-BY 4.0 Open Access | rec: 20 May 2015, publ: 20 May 2015