Sensitivity and Specificity of Information Criteria

John J. Dziak,[1*]

Donna L. Coffman,[2]

Stephanie T. Lanza,[3]

Runze Li,[4]

Lars S. Jermiin[5]

October 22, 2018

[1] The Methodology Center, The Pennsylvania State University, University Park, PA, 16802, USA,

[2] Department of Epidemiology and Biostatistics, Temple University, Philadelphia, PA, 19122, USA,

[3] The Methodology Center, The Edna Bennett Pierce Prevention Research Center, and Department of Biobehavioral Health, The Pennsylvania State University, University Park, PA, 16802, USA,

[4] Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA, 16802, USA,

[5] Research School of Biology, Australian National University, Acton, ACT 2601, Australia; School of Biology & Environment Science, University College Dublin, Belfield, Dublin 4, Ireland; Earth Institute, University College Dublin, Belfield, Dublin 4, Ireland

* The corresponding author is John J. Dziak, `jjd264@psu.edu`.

Abstract

Information criteria (ICs) based on penalized likelihood, such as Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC), and sample-size-adjusted versions of them, are widely used for model selection in health and biological research. However, different criteria sometimes support different models, leading to discussions about which is the most trustworthy. Some researchers and fields of study habitually use one or the other, often without a clearly stated justification. They may not realize that the criteria may disagree. Others try to compare models using multiple criteria but encounter ambiguity when different criteria lead to substantively different answers, leading to questions about which criterion is best. In this paper we present an alternative perspective on these criteria that can help in interpreting their practical implications. Specifically, in some cases the comparison of two models using ICs can be viewed as equivalent to a likelihood ratio test, with the different criteria representing different alpha levels and BIC being a more conservative test than AIC. This perspective may lead to insights about how to interpret the ICs in more complex situations. For example, AIC or BIC could be preferable, depending on the relative importance one assigns to sensitivity versus specificity. Understanding the differences and similarities among the ICs can make it easier to compare their results and to use them to make informed decisions.

1 Introduction

Many model selection techniques have been proposed for different settings (for reviews see Miller, 2002; Pitt and Myung, 2002; Zucchini, 2000; Johnson and Omland, 2004). Among other considerations, researchers must balance sensitivity (suggesting enough parameters to accurately model the patterns, processes, or relationships in the data) with specificity (not suggesting nonexistent patterns, processes, or relationships). Several of the simplest and most popular model selection criteria can be discussed in a unified way as log-likelihood functions with simple penalties. These include Akaike's Information Criterion (Akaike, 1973, AIC), the Bayesian Information Criterion (Schwarz, 1978, BIC), the sample-size-adjusted AIC or AIC_C of Hurvich and Tsai (1989), the "consistent AIC" (CAIC) of Bozdogan (1987), and the sample-size-adjusted BIC (ABIC) of Sclove (1987) (see Table 1). Each of these ICs consists of a goodness-of-fit term plus a penalty to reduce the risk of overfitting, and each provides a standardized way to balance sensitivity and specificity.

Applying an IC involves choosing the model with the best penalized log-likelihood: that is, the highest value of $\ell - A_n p$, where ℓ is the log-likelihood, A_n is a constant or a 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 often expressed as finding the lowest value of -2ℓ plus a penalty:

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

and we follow that convention here. Expression (1) is what Atkinson (1980) called the generalized information criterion; in this paper we simply refer to (1) as an IC. Expression (1) is sometimes replaced in practice by the practically equivalent $G^2 + A_n p$, where G^2 is the deviance, defined as twice the difference in log-likelihood between the current model and the saturated model, that is, the model with the most parameters which is still identifiable (e.g., Collins and Lanza, 2010).

In practice, Expression (1) cannot be used directly without first choosing A_n. Specific choices of A_n make (1) equivalent to AIC, BIC, ABIC or CAIC. Thus, although motivated by different theories and goals, algebraically these criteria are only different values of A_n in (1), corresponding to different relative degrees of emphasis on parsimony, that is, on the number of free parameters in the selected model (Claeskens and Hjort, 2008; Lin and Dayton, 1997; Vrieze, 2012). 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 examine this question by focusing on the similarities and differences among AIC, BIC, CAIC, and ABIC, especially in view of an analogy between their different complexity penalty weights A_n and the α levels of hypothesis tests. We especially focus on AIC and BIC, which have been extensively studied theoretically (Pötscher, 1991; Atkinson, 1980; Kuha, 2004; Zhang, 1993; Ding *et al.*, 2018; Shao, 1997; Kadane and Lazar, 2004; Vrieze, 2012), and which are often used not only in their own right but as tuning criteria to improve the performance of more complex model selection techniques (e.g., in high-dimensional regression variable selection Wu and Ma, 2015; Narisetty and He, 2014; Wang

Table 1: Summary of Common Information Criteria

Criterion	Penalty Weight	Emphasis	Likely Kind of Error
Non-consistent criteria			
AIC	$A_n = 2$	Good prediction	Overfitting
Consistent criteria			
ABIC	$A_n = \ln\left(\frac{n+2}{24}\right)$	Depends on n	Depends on n
BIC	$A_n = \ln(n)$	Parsimony	Underfitting
CAIC	$A_n = \ln(n) + 1$	Parsimony	Underfitting

et al., 2007). The AIC and BIC are widely used in many important applications in bioinformatics, including in molecular phylogenetics (Posada, 2008; Darriba *et al.*, 2012; Jayaswal *et al.*, 2014; Kalyaanamoorthy *et al.*, 2017; Lefort *et al.*, 2017).

In the following section we review the motivation and theoretical properties of these ICs. We then discuss their application to a common application of model selection in medical, health and social scientific applications: that of choosing the number of classes in a latent class analysis (e.g., Collins and Lanza, 2010). Finally, we propose practical recommendations for using ICs to extract valuable insights from data while acknowledging their differing emphases.

Common Penalized-Likelihood Information Criteria

In this section we review some commonly used ICs. Their formulas, as well as some of their properties which we describe later in the paper, are summarized for convenience in Table 1.

Akaike's Information Criterion (AIC)

First, the AIC Akaike (1973) sets $A_n = 2$ in (1). It estimates the relative Kullback-Leibler (KL) divergence (a nonparametric measure of difference between distributions) of the likelihood function specified by a fitted candidate model, from the likelihood function governing the unknown true process that generated the data. The fitted model closest to the truth in the KL sense would not necessarily be the model that 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 or the process that produced the data. Such a model would not necessarily have the lowest error in fitting the data already observed (also known as the training sample) but would be expected to have the lowest error in predicting future data taken from the same population or process (also known as the test sample). This is an example of a bias-variance tradeoff (see, e.g., Hastie *et al.*, 2001).

Technically, the KL divergence 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. This is intuitively understood as the difference between the estimated and the 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 $E_t(\ell(y))$, especially if p is large, because a model with many parameters can generally be fine-tuned to appear to fit a small dataset well, even if its structure is such that it cannot generalize to describe the process that generated the data. Intuitively, this means that if there are many parameters, the fit of the model to the originally obtained data (training sample) will seem good regardless of whether the model is correct or not, simply because the model is so flexible. In other words, once a particular dataset is used to estimate the parameters of a model, the fit of the model on that sample is no longer an independent evaluation of the quality of the model. The most straightforward way to address this fit inflation would be by cross-validation on a new sample, but AIC and similar criteria attempt to achieve something similar when there is no other sample (see Shao, 1993, 1997).

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 diagonal elements). $\hat{\mathbf{J}}$ is an estimator for the covariance matrix of the parameters, based on the matrix of second derivatives of ℓ in each of 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, poor models presumably have poor values of ℓ, so the precise size of the penalty is less important (Burnham and Anderson, 2002). The resulting expression $\ell - p$ suggests using $A_n = 2$ in (1) and concluding that fitted models with low 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. When n is small or p is large, 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 when n is large relative to p. For small n, Hurvich and Tsai (1989) showed that $\mathrm{AIC_C}$ sometimes performs better than AIC. Theoretical discussions of model selection often focus on the advantages and disadvantages of AIC versus BIC, and $\mathrm{AIC_C}$ gets little attention because it is asymptotically equivalent to AIC. However, this equivalence is subject to the assumption that p is fixed and n becomes very large. Because in many situations p is comparable to n or larger, $\mathrm{AIC_C}$ may deserve more attention in future work.

Some other selection criteria are asymptotically equivalent to AIC, at least for linear regression. These include Mallows' C_p (see George, 2000), leave-one-out cross-validation (Shao, 1997; Stone, 1977), and the generalized cross- validation (GCV) statistic (see Golub *et al.*, 1979; Hastie *et al.*, 2001). Leave-one-out cross-validation involves fitting the candidate model on many subsamples of the data, each excluding one subject (i.e., participant or specimen), and observing the average squared error in predicting the extra response. Each approach is intended to correct a fit estimate for the artificial inflation in observed performance caused by fitting a model and evaluating it with the same data, and to find a good balance between bias caused by too restrictive a model and excessive variance caused by too rich a model (Hastie *et al.*, 2001). 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.

The Deviance Information Criterion used in Bayesian analyses (Spiegelhalter *et al.*, 2002; Gibson *et al.*, 2018) is beyond the scope of this paper because it cannot be expressed as a value of A_n in Expression (1). However, it has some relationship to AIC and has an analogous purpose (Claeskens and Hjort, 2008; Ando, 2013).

Other ICs are named after AIC but do not derive from the same theoretical framework, except that they share the form (1). For example, some researchers (Andrews and Currim, 2003; Fonseca and Cardoso, 2007; Yang and Yang, 2007) have suggested using $A_n = 3$ in expression (1) instead of 2. The use of $A_n = 3$ is sometimes called “AIC3.” There is no statistical theory to motivate AIC3, such as minimizing KL divergence or any other theoretical construct, but on an *ad hoc* basis it has fairly good simulation performance in some settings, being stricter than AIC but not as strict as BIC. Also, the CAIC, the “corrected” or “consistent” AIC proposed by Bozdogan (1987), uses $A_n = \ln(n)+1$. (It should not be confused with the AIC_C discussed above.) This penalty tends to result in a more parsimonious model and more underfitting than AIC or BIC, and it is therefore not very similar to AIC. This value of A_n was chosen somewhat arbitrarily as an example of an A_n that would provide model selection consistency, a property described below in the section for BIC. However, any A_n proportional to $\ln(n)$ provides model selection consistency, so CAIC has no real advantage over the better-known and better-studied BIC (see below), which also has this property.

Schwarz’s Bayesian Information Criterion (BIC)

In Bayesian model selection, a prior probability is set for each model M_i, and prior distributions (often uninformative priors for simplicity) 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 let $\Pr(\mathbf{y}|M_i)$ be the probability density of the data given

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:

$$\frac{\frac{\Pr(M_i|\mathbf{y})}{\Pr(M_j|\mathbf{y})}}{\frac{\Pr(M_i)}{\Pr(M_j)}}.$$

If we assume equal prior probabilities for each model, this simplifies to the "Bayes factor" (see Kass and Raftery, 1995):

$$B_{ij} = \frac{\Pr(M_i|\mathbf{y})}{\Pr(M_j|\mathbf{y})}$$

so that the model with the highest Bayes factor also has the higher posterior probability. Schwarz (1978) and Kass and Wasserman (1995) showed that, for many kinds of models, B_{ij} can be roughly approximated by $\exp(-\frac{1}{2}BIC_i + \frac{1}{2}BIC_j)$, where BIC equals Expression (1) with $A_n = \ln(n)$, especially if a certain "unit information" prior is used for the coefficients. The use of Bayes factors has been argued to be more interpretable than that of significance tests in some practical settings (Raftery, 1996; Goodman, 2008; Beard *et al.*, 2016) although with some caveats (see Gigerenzer and Marewski, 2015; Murtaugh, 2014). Thus the model with the highest posterior probability is likely to be the one with lowest BIC. BIC is described further in Raftery (1995) and Wasserman (2000), but critiqued by Gelman and Rubin (1995) and Weakliem (1999), who find it to be an oversimplification of Bayesian methods. BIC can also be called the Schwarz criterion.

BIC is sometimes preferred over AIC because BIC is "consistent" (e.g., Nylund *et al.*, 2007). Assuming that a fixed number of models are available and that one of them is the true model, a consistent selector is one that selects the true model with probability approaching 100% as $n \to \infty$ (see Rao and Wu, 1989; Zhang, 1993; Shao, 1997; Yang, 2005; Claeskens and Hjort, 2008). The existence of a true model here is not as unrealistically dogmatic as it sounds (Burnham and Anderson, 2004; Kuha, 2004). Rather, the *true* model can be defined as the smallest adequate model, that is, the single model that minimizes KL divergence, or the smallest such model if there is more than one (Claeskens and Hjort, 2008). There may be more than one such model because if a given model has a given KL divergence from the truth, any more general model containing it will have no greater distance from the truth. This is because there is some set of parameters for which the larger model becomes the model nested within it. However, the theoretical properties of BIC are better in situations in which a model with a finite number of parameters can be treated as "true" (Shao, 1997).

AIC is not consistent because it has a non-vanishing chance of choosing an unnecessarily complex model as n becomes large. The unnecessarily complex model would still closely approximate the true distribution but would use more parameters than necessary to do so. However, selection consistency involves some performance tradeoffs when n is modest, specifically, an elevated risk of poor performance caused by underfitting (see Shibata, 1986; Shao, 1997; Pötscher, 1991; Vrieze, 2012). In general, the strengths of AIC and BIC cannot be combined by any single choice of A_n (Leeb, 2008; Yang, 2005). However, in some cases it is possible to construct a more complicated model selection approach that uses aspects of both (see Ding *et al.*, 2018).

Criteria Related to BIC. Sclove (1987) suggested a sample-size-adjusted BIC, variously abbreviated as ABIC, SABIC, 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 selection for autoregressive time series models from a minimum description length perspective (see Stine, 2004). It is not clear whether or not the same adjustment is still theoretically appropriate in different contexts, but in practice it is sometimes used in latent class modeling and seems to work fairly well (see Nylund *et al.*, 2007; Tein *et al.*, 2013).

2 Information Criteria in Simple Cases

The above shows that AIC and BIC differ in theoretical basis. They also often disagree in practice, generally with AIC indicating models with more parameters and BIC with less. This has led many researchers to question whether and when a particular value of the "magic number" A_n (Bozdogan, 1987) can be chosen as most appropriate. Two special cases – comparing equally sized models and comparing nested models – each provide some insight into this question.

First, *when comparing different models of the same size* (i.e., number of parameters to be estimated), all ICs of the form (1) will always agree on which model is best. For example, in regression variable subset selection, suppose two models each use five covariates. In this case, any IC will select whichever model has the highest likelihood (the best fit to the observed sample) after estimating the parameters. This is because only the first term in Expression (1) will differ across the candidate models, so A_n does not matter. Thus, although the ICs differ in theoretical framework, they only disagree when they make different tradeoffs between fit and model size.

Second, *for comparing a nested pair of models, different ICs act like different α levels on a likelihood ratio test* (LRT). 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 ICs will differ only in how they make this bias-variance 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; van der Hoeven, 2005; Vrieze, 2012; Murtaugh, 2014).

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 $-2\ell_1+A_np_1 < -2\ell_0+A_np_0$, where ℓ_1 and ℓ_0 are the fitted maximized log-likelihoods for each model. Although the comparison of models is interpreted differently in the theoretical frameworks used to justify AIC and BIC (Aho *et al.*, 2014; Kuha, 2004), algebraically this comparison is the same as a LRT (Söderström, 1977; Stoica *et al.*, 2004; Pötscher, 1991). That is, M_0 is rejected if and only if

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

The left-hand side is the LRT 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 an LRT with a different interpretation. The α level is specified indirectly through the critical value A_n; it is the proportion of the null hypothesis distribution of the LRT statistic that is less than A_n.

For many kinds of models, the asymptotic null-hypothesis 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 LRT 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. Thus when $p_1 - p_0 = 1$, significance testing at the customary level of $\alpha = .05$ is often an intermediate choice between AIC and BIC, corresponding to $A_n = 1.96^2 \approx 4$. However, as $p_1 - p_0$ becomes larger, all ICs become more conservative, in order to avoid adding many unnecessary parameters unless they are needed. Table 2 shows different effective α values for two values of $p_1 - p_0$, obtained using the R (R Development Core Team, 2010) code `1-pchisq(q=An*df,df=df,lower.tail=TRUE)` where `An` is the A_n value and `df` is $p_1 - p_0$. AIC$_\text{C}$ is not shown in the table because its penalty weight depends both on p_0 and on p_1 in a slightly more complicated way, but will behave similarly to AIC for large n and modest p_0.

The property of selection consistency can be intuitively understood from this perspective. For AIC, as for hypothesis tests, the power of a test increases with n. Thus, rejecting any given false null hypothesis is practically 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 decreasing Type I error rate) as n increases. The power increases more slowly (i.e., the Type II error rate decreases more slowly) than for AIC or for fixed-α hypothesis tests because the test is becoming more stringent, but now the Type I error rate is also decreasing. Thus, nonzero but practically negligible departures from a model are less likely to lead to rejecting the model for BIC than for AIC (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

Table 2: Alpha Levels Represented By Common Information Criteria

n	AIC	ABIC	BIC	CAIC
Assuming $p_1 - p_0 = 1$				
10	0.15730	1.00000	0.12916	0.06917
50	0.15730	0.37923	0.04794	0.02667
100	0.15730	0.22902	0.03188	0.01791
500	0.15730	0.08121	0.01267	0.00723
1000	0.15730	0.05339	0.00858	0.00492
5000	0.15730	0.02085	0.00352	0.00204
10000	0.15730	0.01404	0.00241	0.00140
100000	0.15730	0.00389	0.00069	0.00040
Assuming $p_1 - p_0 = 10$				
10	0.02925	1.00000	0.01065	0.00027
50	0.02925	0.65501	0.00002	< 0.0001
100	0.02925	0.15265	< 0.0001	< 0.0001
500	0.02925	0.00074	< 0.0001	< 0.0001
1000	0.02925	0.00005	< 0.0001	< 0.0001
5000	0.02925	< 0.0001	< 0.0001	< 0.0001
10000	0.02925	< 0.0001	< 0.0001	< 0.0001
100000	0.02925	< 0.0001	< 0.0001	< 0.0001

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 and Hjort, 2008; Yang, 2005; Kieseppä, 2003).

Nylund *et al.* (2007) seem to interpret the lack of selection consistency as a flaw in AIC (Nylund *et al.*, 2007, p. 556). However, the real situation is more complicated; AIC is not a defective BIC, nor *vice versa* (see Kieseppä, 2003; Shibata, 1981, 1986; Pötscher, 1991; Vrieze, 2012). Likewise, the other ICs mentioned here are neither right nor wrong, but are simply choices (perhaps thoughtful and perhaps arbitrary, but still technically valid choices). Since choosing A_n for a model comparison is closely related to choosing an α level for a significance test, the universally "best" IC cannot be defined any more than the "best" α; there will always be a tradeoff. Thus, debates about whether AIC is generally superior to BIC or *vice versa*, will be fruitless.

For non-nested models of different sizes, neither of the above simple cases hold; furthermore, these complex cases are often those in which ICs are most important because a LRT cannot be performed. However, it remains the case that A_n has a powerful effect on the tradeoff between the likelihood term and the penalty on the number of parameters, hence the tradeoff between good fit to the observed data and parsimony.

Almost by definition, there is no universal best way to decide how to make a tradeoff. Sometimes the relative importance of sensitivity or specificity depends on the decisions to be made based on model predictions. For example, in theoretical research Type I error is considered to be more serious because it is a false statement rather than simply a failure to reject a null hypothesis. However, in some environmental or epidemiological decision-making contexts, the decision corresponding to Type II error might be much more harmful to public health than that which would correspond to a Type I error, requiring increased attention to uncertainty about the adequacy of null hypothesis (Peterman, 1990; Andorno, 2004). In this way, one could characterize the comparison of models 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 could be seen as representing 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. For example, although AIC has favorable theoretical properties for choosing the number of parameters needed to approximate the shape of a nonparametric growth curve in general (Shao, 1997), in a particular application with such data Dziak *et al.* (2015) argued that BIC would give more interpretable results. They argued this because the curves in that context were believed likely to have a smooth and simple shape, as they represented averaged trajectories of an intensively measured variable on individuals with diverse individual experiences and because deviations from the trajectory could be modeled using other aspects of the model.

As a caveat, if a researcher wishes to consider practical consequences of decisions based on model choices directly, it may be much more satisfactory to explicitly use Bayesian decision theory rather than simply choosing a value of Expression A_n in (1) (see, e.g., Claxton *et al.*, 2000; Gelman and Rubin, 1995). Also, in practice it is often difficult to determine the α value that a particular criterion really represents, for two reasons. First, even for regular situations in which a LRT 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 LRT. 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. In the section below we review simulation results from a common application of ICs, namely the selection of the number of latent classes (empirically derived clusters) in a dataset.

3 The Special Case of Latent Class Analysis

A common use of ICs is in selecting the number of components for a latent class analysis (LCA). LCA is a kind of finite mixture model (essentially, a model-based cluster analysis; McLachlan and Peel, 2000; Lazarsfeld and Henry, 1968; Collins and Lanza, 2010). LCA assumes that the population is a "mixture" of multiple classes of a categorical latent variable. Each class has different parameters that define the distributions of observed items, and the goal is to account for the relationships among items by defining classes appropriately. In this section we consider LCA as described in Collins and Lanza (2010), although ICs are also used for more complex mixture models and clustering applications (e.g., Wang *et al.*, 2012; Ye *et al.*, 2015). LCA is very similar to cluster analysis, but is based on maximizing an explicitly stated likelihood function rather than focusing on a heuristic computational algorithm like k-means. Also, some authors use the term LCA only when the observed variables are also categorical, and use the term "latent profile analysis" for numerical observed variables, but we ignore this distinction here. LCA is also closely related to latent transition (LTA) models (see Collins and Lanza, 2010), an application of hidden Markov models (see, e.g., Eddy, 2004) that allows changes in latent class membership, conceptualized as transitions in an unobserved Markov chain. LCA models are sometimes used in combination with other models, such as in predicting class membership from genotypic or demographic variables, or predicting medical or behavioral phenotypes from class membership (e.g., Lubke *et al.*, 2012; Dziak *et al.*, 2016; Bray *et al.*, 2018). To fit an LCA model or any of its cousins, an algorithm such as EM (Dempster *et al.*, 1977; McLachlan and Peel, 2000; Gupta and Chen, 2010) is often used to alternatively estimate class-specific parameters and predict subjects' class membership given those parameters. The user must specify the number of classes in a model, but the true number of classes is generally unknown. (Nylund *et al.*, 2007; Tein *et al.*, 2013). Sometimes one might have a strong theoretical reason to specify the number of classes, but often this must be done using data-driven model selection.

A naïve approach would be to use likelihood ratio (LR) or deviance (G^2) tests sequentially to choose the number of classes and to conclude that the k-class model is large enough if and only if the $(k+1)$-class model does not fit the data significantly better. The selected number of classes would be the smallest k that is not rejected when compared to the $(k+1)$-class model. However, the assumptions for the supposed asymptotic χ^2 distribution in a LRT are not met in the setting of LCA, so that the p-values from those tests are not valid (see Lin and Dayton, 1997; McLachlan and Peel, 2000). The reasons for this are based on the fact that 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. That is, an meaningful k-class model is not obtained simply by setting a parameter to zero in a $(k+1)$ class model in the way that, for example, a more parsimonious regression model is obtained by constraining certain coefficients in a richer model to zero. Ironically, the lack of regular nesting structure that makes it impossible to decide on the number of classes with an LRT has also been shown to invalidate the mathematical approximations used in the AIC and BIC derivations in the same way (McLachlan and Peel, 2000, pp. 202-212). Nonetheless, ICs are widely used in LCA and other mixture models. This is partly due to their ease of use, even without a firm theoretical basis. Fortunately, there is at least an asymptotic theoretical result showing that, when the true model is well-identified, BIC (and hence also AIC and ABIC) will have a probability of underestimating the true number of classes that approaches 0 as sample size tends to infinity (Leroux, 1992; McLachlan and Peel, 2000, p. 209).

Lin and Dayton (1997) did an early simulation study comparing the performance of AIC, BIC, and CAIC for choosing which assumptions to make in constructing constrained LCA models, a model selection task which is somewhat but not fully analogous to choosing the number of classes. 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. Interpreting these results, Dayton (1998) suggested that AIC tended to be a better choice in LCA than BIC, but recommended computing and comparing both.

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 and were therefore easy to mistakenly lump together; 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). Fonseca and Cardoso (2007) similarly suggested AIC3 as the preferred selection criterion for categorical LCA models.

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 mean that BIC was the best in every situation. In most of the scenarios considered by Nylund *et al.* (2007), 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 with a small sample and unequally sized classes, (Nylund *et al.*, 2007, p. 557), BIC essentially never chose the larger correct model and it usually chose one that 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 (for example, a small but nonzero difference between the parameters of a pair of classes) and n is small. Wu (2009) compared the performance of AIC, BIC, ABIC, CAIC, naïve tests, and the bootstrap LRT in hundreds of simulated scenarios. Performance was heavily dependent on the scenario, but the method that worked adequately in the greatest variety of situations was the bootstrap LRT, followed by ABIC and classic BIC. Wu (2009) argued 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. Dziak *et al.* (2014) also concluded that BIC could seriously underfit relative to AIC for small sample sizes or other challenging situations. In latent profile analysis, Tein *et al.* (2013) found that BIC and ABIC did well for large sample sizes and easily distinguishable classes, but AIC chose too many classes, and no method performed well for especially challenging scenarios. In a more distantly related mixture modeling framework involving modeling evolutionary rates at different genomic sites, Kalyaanamoorthy *et al.* (2017) found that AIC, AIC_C, and BIC worked well but that BIC worked best.

Despite all these findings, 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. For small n, the most likely error in a simulation 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. Furthermore, all of these findings have limited usefulness in 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.

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 distinctly interpretable 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. In these cases, the BIC-favored model is clearly the better practical choice. For example, in Chan *et al.* (2007) BIC favored a mixture model with 5 classes, and AIC favored at least 10; the authors felt that a 10-class model would be too hard to interpret. In fact, it may be necessary for theoretical or practical reasons to choose a number of classes even smaller than that suggested by BIC. This is because it is important to choose the number of classes based on their theoretical interpretability, as well as by excluding any models with so many classes that they lead to a failure to converge to a clear maximum-likelihood solution (see Collins and Lanza, 2010; Pohle *et al.*, 2017; Bray and Dziak, 2018).

An alternative to ICs in latent class analysis and cluster analysis is the use of a bootstrap test (see McLachlan and Peel, 2000). Unlike the naïve NRT, Nylund *et al.* (2007) showed empirically that the bootstrap LRT with a given α level does generally provide a Type I error rate at or below that specified level. Both Nylund *et al.* (2007) and Wu (2009) found that this bootstrap test seemed to perform somewhat better than the ICs in various situations. The bootstrap LRT 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). Also beyond the scope of this paper are mixture-specific selection criteria such as the normalized entropy criterion

(Biernacki *et al.*, 1999) or integrated completed likelihood (Biernacki and Celeux, 2000; Rau and Maugis, 2018). However, the basic ideas in this article will still be helpful in interpreting the implications of some of the other selection methods. For example, like any test or criterion, the bootstrap LRT still requires the choice of a tradeoff between sensitivity and specificity (i.e., by selecting an α level).

4 Discussion

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. If BIC selects the largest and most general model considered, it is worth thinking about whether to expand the space of models considered (since an even more general model might fit even better), and similarly if AIC chooses the most parsimonious.

AIC and BIC each have distinct theoretical advantages. However, a researcher may judge that there may be a practical advantage to one or the other in some situations. For example, as mentioned earlier, in choosing the number of classes in a mixture model, the true number of classes required to satisfy all model assumptions is sometimes quite large, too large to be of practical use or even to allow coefficients to be reliably estimated. In that case, BIC would be a better choice than AIC. Additionally, in practice, one may wish to rely on substantive theory or parsimony of interpretation in choosing a relatively simple model. In such cases, the researcher may decide that even the BIC may have indicated a model that is too complex in a practical sense, and may choose to select a smaller model that is more theoretically meaningful or practically interpretable instead (Pohle *et al.*, 2017; Bray and Dziak, 2018). This does not mean that BIC overfit. Rather, in these situations the model desired is sometimes not the literally true model but simply the most useful model, a concept which cannot be identified using fit statistics alone but requires subjective judgment. Depending on the situation, the number of classes in a mixture model may either be interpreted a true quantity needing to be objectively estimated, or else as a level of approximation to be chosen for convenience, like the scale of a map. Still, in either case the question of which patterns or features are generalizable beyond the given sample remains relevant (c.f. Li and Marron, 2005).

A larger question is whether to use ICs at all. If ICs indeed reduce to LRTs in simple cases, one might wonder why ICs are needed at all, and why researchers cannot simply do LRTs. A possible answer is flexibility. Both AIC and BIC can be used to concurrently compare many models, not just a pair at a time, or to weight the estimates obtained from different models for a common quantity of interest. These weighting approaches use either AIC or BIC but not both, because AIC and BIC are essentially treated as different Bayesian priors. While currently we know of no mathematical theoretical framework for explicitly combining both AIC and BIC into a single weighting scheme, a sensitivity analysis could be performed by comparing the results from both. AIC and BIC can also be used to choose a few well-fitting models, rather than selecting a single model from among many and assuming it to be the truth (Kuha, 2004). Researchers have also 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; Murtaugh, 2014); for example, an AIC or BIC difference between two models of less than 2 provides little evidence for one over the other; an AIC or BIC difference of 10 or more is strong evidence. These principles should not be used as rigid cutoffs Murtaugh (2014), but as input to decision making and interpretation. Kadane and Lazar 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 future study; for example, in some cases one might use the BIC- preferred model as a minimum size and the AIC-preferred model as a maximum. AIC and BIC can also both be used for model averaging, that is, estimating quantities of interest by combining more than one model weighted by their plausibility (see Posada and Crandall, 2001; Posada and Buckley, 2004; Claeskens and Hjort, 2008; Johnson and Omland, 2004; Gelman and Rubin, 1995; Hoeting *et al.*, 1999; Burnham and Anderson, 2004). Despite these many worthwhile options, it is still important to remember an automatic and uncritical use of an IC is no more insightful than an automatic and uncritical use of a p-value.

Lastly, both AIC and BIC were developed in situations in which n was assumed to be much larger than p. None of the ICs discussed here were specifically developed for situations such as those found in many genome-wide association studies predicting disease outcomes, in which the number of participants (n) is often smaller than the number of potential genes (p), even when n is in the tens of thousands. The ICs can still be practically useful in this setting (e.g., Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013). However, sometimes they might need to be adapted (see, e.g., Chen and Chen, 2008; Pan *et al.*, 2016; Mestres *et al.*, 2018). More research in this area would be worthwhile.

Acknowledgements

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 for his careful review and recommendations on an earlier version of this paper. John Dziak thanks Frank Sabach for encouragement in the early stages. Amanda Applegate is also thanked for her editorial assistance. Lars Jermiin thanks the University College Dublin for its generous hospitality.

A previous version of this report has been disseminated as Methodology Center Technical Report 12-119, June 27, 2012, and as a preprint at `https://peerj.com/preprints/1103/`. The earlier version of the paper contains simulations to illustrate the points made. Simulation code is available at `http://www.runmycode.org/companion/view/1306` and results at `https://methodology.psu.edu/media/techreports/12-119.pdf`.

Funding

This research was supported by NIH grant P50 DA039838 from the National Institute on Drug Abuse. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Drug Abuse or the National Institutes of Health.

References

Aho, K. A., Derryberry, D., and Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, **95 3**, 631–6.

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

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

Ando, T. (2013). Predictive Bayesian model selection. *American Journal of Mathematical and Management Sciences*, **31**, 13–38.

Andorno, R. (2004). The precautionary principle: A new legal standard for a technological age. *Journal of International Biotechnology Law*, **1**, 11–19.

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.

Beard, E., Dienes, Z., Muirhead, C., and West, R. (2016). Using Bayes factors for testing hypotheses about intervention effectiveness in addictions research. *Addiction*, **111**, 2230–2247.

Biernacki, C. and Celeux, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.*, **22**, 719–725.

Biernacki, C., Celeux, G., and Govaert, G. (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters*, **20**, 267–272.

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.

Bray, B. C. and Dziak, J. J. (2018). Commentary on latent class, latent profile, and latent transition analysis for characterizing individual differences in learning. *Learning and Individual Differences*.

Bray, B. C., Dziak, J. J., Patrick, M. E., and Lanza, S. T. (2018). Inverse propensity score weighting with a latent class exposure: Estimating the causal effect of reported reasons for alcohol use on problem alcohol use 15 years later. *Prevention Science*.

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.

Chen, J. and Chen, Z. (2008). Extended bayesian information criterion for model selection with large model spaces. *Biometrika*, **95**, 759–771.

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

Claxton, K., Lacey, L. F., and Walker, S. G. (2000). Selecting treatments: a decision theoretic approach. *Journal of the Royal Statistical Society, A*, **163**, 211–225.

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

Cross-Disorder Group of the Psychiatric Genomics Consortium (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet*, **381**(9875), 1371 – 1379.

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: More models, new heuristics and parallel computing. *Nature Methods*, **9**, 772.

Dayton, C. M. (1998). *Latent class scaling analysis*. Sage, Thousand Oaks, CA.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, **39**, 1–38.

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

Ding, J., Tarokh, V., and Yang, Y. (2018). Bridging AIC and BIC: A new criterion for autoregression. *IEEE Transactions on Information Theory*, **64**, 4024–4043.

Dziak, J. J., Lanza, S. T., and Tan, X. (2014). Effect size, statistical power and sample size requirements for the bootstrap likelihood ratio test in latent class analysis. *Structural Equation Modeling*, **21**, 534–552.

Dziak, J. J., Li, R., Tan, X., Shiffman, S., and Shiyko, M. (2015). Modeling intensive longitudinal data with mixtures of nonparametric trajectories and time-varying effects. *Psychological Methods*, **20 4**, 444–69.

Dziak, J. J., Bray, B. C., Zhang, J.-T., Zhang, M., and Lanza, S. T. (2016). Comparing the performance of improved classify-analyze approaches in latent profile analysis. *Methodology*, **12**, 107–116.

Eddy, S. R. (2004). What is a hidden Markov model? *Nature Biotechnology*, **22**, 1315–1316.

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.

Gelman, A. and Rubin, D. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology*, **25**, 165–173.

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

Gibson, G. J., Streftaris, G., and Thong, D. (2018). Comparison and assessment of epidemic models. *Statistical Science*, **33**, 19–33.

Gigerenzer, G. and Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, **41**, 421–440.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215–223.

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hepatology*, **45**, 135–140.

Gupta, M. R. and Chen, Y. (2010). Theory and use of the EM algorithm. *Foundations and Trends in Signal Processing*, **4**, 223–296.

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

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, **14**, 382–417.

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

Jayaswal, V., Wong, T. K. F., Robinson, J., Poladian, L., and Jermiin, L. S. (2014). Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. *Systematic Biology*, **63**, 726–742.

Johnson, J. B. and Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in ecology and evolution*, **19**, 101–8.

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

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. F. K., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, **14**, 587–589.

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 Research*, **33**, 188–229.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent structure analysis*. Houghton Mifflin, Boston.

Leeb, H. (2008). Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. *Bernoulli*, **14**, 661–690.

Lefort, V., Longueville, J. E., and Gascuel, O. (2017). SMS: Smart model selection in PhyML. *Molecular Biology and Evolution*, **34**, 2422–2424.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *Annals of Statistics*, **20**, 1350–1360.

Li, R. and Marron, J. S. (2005). Local likelihood SiZer map. *Sankhyā: The Indian Journal of Statistics*, **67**, 476–98.

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.

Lubke, G. H., Stephens, S. H., Lessem, J. M., Hewitt, J. K., and Ehringer, M. A. (2012). The chrna5/a3/b4 gene cluster and tobacco, alcohol, cannabis, inhalants and other substance use initiation: Replication and new findings using mixture analysis. *Behavioral Genetics*, **42**, 636–646.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley, New York.

Mestres, A. C., Bochkina, N., and Mayer, C. (2018). Selection of the regularization parameter in graphical models using network characteristics. *Journal of Computational and Graphical Statistics*, **27**, 323–333.

Miller, A. J. (2002). *Subset selection in regression*. Chapman & Hall, New York, 2nd edition.

Murtaugh, P. A. (2014). In defense of p values. *Ecology*, **95**, 611–617.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Annals of Statistics*, **42**, 789–817.

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.

Pan, R., Wang, H., and Li, R. (2016). Ultrahigh-dimensional multiclass linear discriminant analysis by pairwise sure independence screening. *Journal of the American Statistical Association*, **111**, 169–179.

Peterman, R. M. (1990). The importance of reporting statistical power: the forest decline and acidic deposition example. *Ecology*, **71**, 2024–2027.

Pitt, M. A. and Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, **6**, 421–425.

Pohle, J., Langrock, R., van Beest, F. M., and Schmidt, N. M. (2017). Selecting the number of states in hidden markov models: Pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, **22**, 270–293.

Posada, D. (2008). jModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253–1256.

Posada, D. and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: Advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, **53**, 793–808.

Posada, D. and Crandall, K. A. (2001). Selecting the best-fit model of nucleotide substitution. *Systematic Biology*, **50**, 580–601.

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.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, **83**, 251–266.

Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika*, **76**, 369–374.

Rau, A. and Maugis, C. (2018). Transformation and model choice for rna-seq co-expression analysis. *Briefings in bioinformatics*, **19 3**, 425–436.

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. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, **88**, 486–494.

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.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, B*, **64**, 583–639.

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.

Tein, J.-Y., Coxe, S., and Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling*, **20**, 640–657.

Teräsvirta, T. and Mellin, I. (1986). Model selection criteria and model selection tests in regression models. *Scandinavian Journal of Statistics*, **13**, 159–171.

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 Methods*, **17**, 228–243.

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94 3**, 553–568.

Wang, Y., Xu, M., Wang, Z., Tao, M., Zhu, J., Wang, L., Li, R., Berceli, S. A., and Wu, R. (2012). How to cluster gene expression dynamics in response to environmental signals. *Briefings in bioinformatics*, **13 2**, 162–74.

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, C. and Ma, S. (2015). A selective review of robust variable selection with applications in bioinformatics. *Briefings in bioinformatics*, **16 5**, 873–83.

Wu, Q. (2009). *Class extraction and classification accuracy in latent class models*. Ph.D. 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.

Ye, M., Wang, Z., Wang, Y., and Wu, R. (2015). A multi-poisson dynamic mixture model to cluster developmental patterns of gene expression by rna-seq. *Briefings in Bioinformatics*, **16**(2), 205–215.

Zhang, P. (1993). On the convergence rate of model selection criteria. *Communications in Statistics: Theory and Methods*, **22**, 2765–2775.

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