Published in final edited form as:
Psychol Methods. 2012 June ; 17(2): 228–243. doi:10.1037/a0027127.

Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)

Scott I. Vrieze
University of Minnesota Minneapolis VA Medical Center

Abstract

This article reviews the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) in model selection and the appraisal of psychological theory. The focus is on latent variable models given their growing use in theory testing and construction. We discuss theoretical statistical results in regression and illustrate more important issues with novel simulations involving latent variable models including factor analysis, latent profile analysis, and factor mixture models. Asymptotically, the BIC is consistent, in that it will select the true model if, among other assumptions, the true model is among the candidate models considered. The AIC is not consistent under these circumstances. When the true model is not in the candidate model set the AIC is effcient, in that it will asymptotically choose whichever model minimizes the mean squared error of prediction/estimation. The BIC is not effcient under these circumstances. Unlike the BIC, the AIC also has a minimax property, in that it can minimize the maximum possible risk in finite sample sizes. In sum, the AIC and BIC have quite different properties that require different assumptions, and applied researchers and methodologists alike will benefit from improved understanding of the asymptotic and finite-sample behavior of these criteria. The ultimate decision to use AIC or BIC depends on many factors, including: the loss function employed, the study's methodological design, the substantive research question, and the notion of a true model and its applicability to the study at hand.

Much theory in psychology has been based on results from observational data, as opposed to running true experiments. This is especially true in psychology subspecialties such as clinical, counseling, social, personality, I/O, and developmental psychology, where in many cases running true experiments can be challenging if not impossible (Meehl, 1978; Cronbach & Meehl, 1955). Observational data must be used, and statistical techniques have been developed by which associations and trends can be modeled in diverse and sophisticated ways. The convenience of ever-faster personal computing and statistical software suites has made it possible to fit models that were previously intractable. Fortunately or unfortunately, the myriad of possible models highlights the importance of choosing good models and discarding bad ones. The problem of choosing a model (or set of models) is the focus of this review.

This paper is organized into three parts. (a) We discuss briefly the use of models in psychology. (b) We describe general results from the statistical literature on the AIC (Akaike, 1974) and the BIC (Schwarz, 1978). The BIC is consistent in model selection: it is guaranteed to select the true model as the sample size grows, as long as the true model is

Direct correspondence to Scott I. Vrieze at University of Minnesota, 75 East River Road, Minneapolis, MN 55455 or vrie0006@umn.edu.
The author expresses deep thanks to William M. Grove, Matt McGue, William G. Iacono, Niels G. Waller, Jeffery D. Long, Kristian Markon and four anonymous reviewers for insightful discussions and comments on earlier drafts of this article.

among the candidate models being considered (and other assumptions). The AIC asymptotically selects the model that minimizes mean squared error of prediction or estimation. The AIC also minimizes maximum possible risk in finite sample sizes. (c) Two simulation studies are provided to illustrate these issues.

This paper is meant to be thought-provoking for applied researchers and methodologists alike. The purpose is to summarize and illustrate important properties of the AIC and the BIC so that each can be used in a more informed manner. In our experience applied articles rarely defend their use of a particular model selection criterion, but instead give a reference or two to seminal (and too often mathematically inaccessible) articles about it. The articles leave to the reader to determine why one criterion was used and not another. In an extreme example, we have recently seen an article that used four criteria (two of which were linear functions of each other), and stated that the model for which three of the four criteria agreed would be selected as the best model. This approach does not reconcile the differences between selection criteria (except in the case where all criteria agree). Reconciling the differences between AIC and BIC appears to be very diffcult, if not impossible (e.g., Y. Yang, 2005).

The emphasis here is on latent variable modeling, which includes such models as factor, latent profile/class, factor mixture, biometric, mixed-effects/random-effects, growth, and latent change models, among others. A brief search for the dysjunction of these terms in titles and abstracts on the psychINFO database (plus “latent variable” and “structural equation model”) yielded 47,307 hits. Of these, 30,773 were attributable to the term “factor analysis.” Not only are latent variable models used extensively in psychology to address theoretical questions, but there also appears to be a heavy data-driven reliance on fit criteria like the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), perhaps for a lack of applicability of more traditional fit measures (such as analysis of residuals in regression).

Statistical Models in Psychology

A statistical model is a mathematically precise way to summarize properties of measurements and their relationships with one another. A simple example is linear regression, where one variable is predicted by a weighted linear function of some other variable(s). While this is a perfectly legitimate model, mathematically speaking, it may not do a very good job of describing the actual relationship between two variables. For example, perhaps the relationship is not linear, is not continuous, or is moderated, mediated, or suppressed by another (possibly unmeasured) variable. Scientific conclusions based on a mis-specified model can be erroneous and highly misleading (Mosteller & Tukey, 1977; Breiman, 2001).

While the methods of model selection discussed below are entirely applicable to choosing between regression models, the present review is concerned more directly with latent variable model selection. Multivariate measurement and structural equation models typically take the form of multivariate regressions, except that the predictor variables are latent (unobserved). Fit statistics such as the AIC and BIC are used with latent variable models, perhaps because these models are often estimated by maximum likelihood estimation and most statistical software suites used in psychology report them as rough and ready fit indices (e.g., AMOS, Mplus, Mx, OpenMx, LISREL, SAS, Latent Gold).

Cronbach and Meehl (1955) noted long ago the potential of latent variable models (e.g., factor analysis) to address aspects of construct validation and theory formulation. Perhaps the most common example of a statistical model relating latent variables to observed variables is factor analysis (FA; Bollen, 1989; Mulaik, 2010). In FA, correlations between

directly measured variables are taken as evidence that those variables share a common cause or set of causes. Absence of correlation indicates they share no common cause. The directly measured variables (e.g., height, score on a questionnaire) are termed *observed variables*, or OVs. In factor analysis the common cause (correlation) is modeled as one or more continuous unobserved variables (latent variables, or LVs) that cause the OVs.

Factor analysis has a long history in intelligence assessment (Spearman, 1904), and has been widely applied to the study of, for example, personality (Digman, 1990) and descriptive psychopathology (Krueger, 1999). The factor model is a ubiquitous analytical tool in many areas of psychology, both for proposing and testing psychological theory (e.g., the Big 5 theory of personality) and in measurement applications such as scale development and test construction. Specializations of factor analysis include biometric models for multivariate analysis of the heritability of a variable (Neale & Cardon, 1992), and growth/random e ects models for analysis of panel or longitudinal repeated-measures designs (Muthén, 2002).

Other LV models abound and, if fitted to the same set of OVs, can be compared with one another through using the AIC and the BIC. Examples include latent profile analysis (LPA) and latent class analysis (Lazarsfeld & Henry, 1968; Heinen, 1996). These posit categorical, rather than continuous, LVs. The latent profile model, for example, would be expected to better fit data arising from a categorically distributed causal input (e.g., a present/absent disease pathogen). In fact, latent profile models and factor models can both be fit to the same set of OVs, and one can determine whether the LVs are better represented as categorical (as in LPA) or continuous (as in FA). In more recent developments, factor mixture models (FMM) combine factor analysis and latent profile analysis (McLachlan & Peel, 2000; Muthén, 2008; Bauer & Curran, 2004). An FMM assumes the population is composed of discrete latent classes, each of which is further composed of a factor model. The overall model functions much like a multiple-group factor analysis, but group status is inferred from the data rather than known a priori. FMM is much more flexible than the latent profile (or class) model, in that individuals maintain within-class systematic individual differences. To take a psychopathological example, it may be that there exist two classes: a class of people with an alcohol use disorder and a class of people without one. The factor mixture model would allow alcoholics and non-alcoholics to vary in their severity of alcohol problems. This is not possible in an LPA approach, where all systematic individual differences are caused by admixture of homogeneous classes.

A diversity of other latent variable models exist (such as Poisson models for count data, survival models for evaluating the time elapsed until some event occurs, etc.), but the above give the reader a sense of the possibilities. Muthén (2002) gives an accessible and wide overview of the various applications of latent variable models.

Similarities between latent profile analysis (which has categorical LVs) and factor analysis (which has continuous LVs) illustrate a general point about latent variable model fitting. Many fit statistics are measures of the divergence between the OV covariance matrix (the saturated covariance matrix of the observed variables) and the model-predicted covariance matrix (that which can be deduced from the latent variable model). Measures of the differences between the observed covariance matrix and the model-estimated matrix are justified when comparing relative fit between factor models. However, when comparing factor models to class models the comparison between observed and predicted covariance matrices is useless. In fact, a class model with m classes will fit a covariance matrix generated from a factor model with $m - 1$ factors, and will do so just as well as would a factor model with $m - 1$ factors (Molenaar & van Eye, 1994). In the class model the covariance is assumed to arise from there being two or more subgroups of individuals and the observed variables imperfect discriminators of class status.

When comparing models with continuous and categorical latent variables it is therefore preferable to determine fit based on a likelihood function, which depends on higher order moments of the data rather than just means and covariances (all models discussed thus far can be estimated via maximum likelihood estimation). This is one reason to use the AIC and BIC: they allow for comparison of non-nested models, even when the likelihood function is different for the different classes of models. Non-nested here simply means that two models are not specific cases of a more general model. A straightforward nested comparison might be between a one-factor and two-factor model. The one-factor model is the same as the two-factor model, except that all loadings from the second factor are fixed at zero. A straightforward example of a non-nested comparison would be between factor and class models. Making comparisons on the basis of higher order moments requires large, but not prohibitively large, amounts of data. For example, sample sizes in the thousands are preferable for comparisons of simple non-nested models (Lubke & Muthén, 2005). We expect that comparisons of complex non-nested models (e.g., with hundreds of estimated parameters) would require increasingly large sample sizes.

Loss and Risk Functions

In practice one must select a criterion, or set of criteria, by which to measure the utility of a model. The AIC and BIC are only important in the context of one or more loss functions. Different models, and model selection criteria like the AIC and BIC, can perform better or worse depending on the chosen loss function.

Loss functions are measures of divergence, and there are as many loss functions as there are ways to measure divergence. To take a straightforward applied example, in some studies the only concern is accurate prediction. In this case minimization of cross-validated squared error of prediction may be a logical choice for the loss function. Cross-validation samples are used because in applied work we do not have omniscient access to the true future outcomes. In notation, we might write this as a sum of squared errors, or $SS_{cv}=\Sigma_i(y_i-\widehat{y_i})^2$, where y_i is the outcome variable for the ith person in the cross-validation sample, and $\widehat{y_i}$ is the model-estimated value for y in that sample. The model that minimizes SS_{cv} would be selected as the best model. The more familiar "mean squared error" is just the expectation of our squared error loss function above. Cross-validated mean squared error, or $MSE_{cv}=E(y-\widehat{y})^2$, is termed a "risk function." Risk functions are simply expectations of loss functions.

Other defensible functions abound. Sometimes the question is whether one can identify the true model. That is, there is a set of candidate models M, one of which generated the data at hand (denoted the "true" model, or M_0), and the loss function $L(M)$ is whether the model selection procedure can identify it. In notation, this loss function might be written as:

$$L(M)=\begin{cases}1 & \text{if the true model } M_0 \text{ is selected,}\\ 0 & \text{otherwise.}\end{cases}$$

We denote this as "zero/one" loss throughout. For an extensive but accessible discussion of loss functions see Schaid (2010).

In some cases the goal of a research project is to accurately estimate model parameters. The question here is whether a parameter is zero (that is, not statistically significant), or whether it is statistically significantly different from zero. A simple example of parameter estimation would be in estimating a correlation, where it is typically of substantive interest whether a correlation is zero or not. One might imagine a mean squared error loss function relating the

estimated correlation parameter to its true value, i.e., $E(r - \widehat{r})^2$. Statistical significance testing is often used to determine whether the parameter is indistinguishable from zero, usually with the sacred threshold of .05. However, this can be conceptualized as a model selection problem where a null model (i.e., $r = 0$) is compared to a model where the parameter is freely estimated ($(r=\widehat{r})$), and the two models are subsequently compared for fit. Either the AIC or the BIC could be used to accomplish this task.

In the behavioral genetics literature, for example, the effect of common environment often drops out as non-significant (Plomin & Daniels, 1987; Burt, 2009) according to traditional levels of significance. On this basis investigators often exclude the shared environmental effect altogether (i.e., fix it to zero) instead of estimating it. A different index of model fit (other than likelihood ratio tests with traditional .05 levels of significance) may have called for inclusion of the common environment effect and thus resulted in more accurate parameter estimation.

The same model selection procedure will not minimize all possible loss functions. Some loss functions are more appropriate in some situations. For example, zero/one loss may be more applicable in epidemiological contexts, where describing a population model is the goal. In other contexts, such as predicting future violence, minimization of cross-validated MSE of prediction would be more appropriate. As we shall see, neither the AIC or BIC will efficiently minimize all loss functions. Given this, we argue that one should identify their loss function(s) and then decide whether using the AIC or BIC is more appropriate in that context.

Definitions of AIC and BIC

There is a large analytical and Monte Carlo literature in statistics related to maximum likelihood model selection criteria (Kadane & Lazar, 2004). We discuss analytical properties of the AIC and the BIC in the context of regression. These properties are expected to hold in latent variable models specified by multivariate regressions, as long as estimation procedures proceed appropriately, which we assume throughout this article. Some Monte Carlo literature also exists, and a portion of it is discussed below.

Both the AIC and the BIC can be derived from a model's likelihood function and resulting maximum likelihood estimate (MLE). Using the likelihood as a model selection criterion in itself is not advised, as it is well-known to select overly parameterized models (Gelfand & Dey, 1994). To see this, imagine fitting a curve to N data points in two-dimensional space. A polynomial of degree $N - 1$ can fit the data perfectly, and maximize the likelihood of the observed data. Obviously, the drawback of a highly-parameterized model is in application to a novel dataset. Overfitted models are tailored to random and systematic error in the training data set which are different in cross-validation data, resulting in fit shrinkage in new samples. Consequently, the problem of model selection can be construed as a trade-off between fit (e.g., the MLE) and model complexity.

This trade-off is immediately apparent in the forms of the AIC and BIC. A general form of the AIC and BIC can be given as:

$$\text{AIC or BIC} = -2l_Y(\widehat{\tau}) + \alpha\kappa \quad (1)$$

where $\widehat{\tau}$ are the optimized model parameters, and $l_Y(\widehat{\tau})$ is the log of the likelihood of those parameters given the data $\boldsymbol{Y}$, κ is the total number of estimated model parameters (i.e., the number of elements in $\widehat{\tau}$). Note that we often denote $l_Y(\widehat{\tau})$ by the much simpler

"log(MLE)." α is a penalty coefficient and is responsible for the entire difference between the AIC and the BIC.

Both the AIC and BIC correct the maximum likelihood estimate by adding a function of the number of model parameters κ. The AIC has $\alpha = 2$ while the BIC has $\alpha = \log(N)$, where N is the number of observations contributing to the sum in the likelihood equation. The number of model parameters κ is taken to be a direct indicator of model complexity; the more parameters the more flexible and complex the model. Despite their obvious similarity, the virtues of AIC and BIC are quite different and, in some aspects, impossible to reconcile (Y. Yang, 2005).

Similarities between AIC, BIC, and Likelihood Ratio Tests

When models are nested, decisions based on AIC and BIC can be interpreted through a comparison with likelihood ratio tests. Imagine the simple scenario where one estimates a covariance. The purpose is to estimate the magnitude of the covariance and determine whether it is zero, on the one hand (i.e., not statistically significant), or nonzero. The likelihood ratio test would compare the log of maximum likelihood estimate for the freely estimated model (denoted MLE_1) to a restricted model with the covariance parameter fixed at zero (MLE_0). If the likelihood of MLE_0 is not significantly less than the likelihood of MLE_1, it is concluded that the covariance is not significantly different from zero. The metric involved is usually −2 log(MLE), where − 2 log(MLE_0) − (−2 log(MLE)) is distributed as χ^2 with a critical p-value of .05 corresponding to a difference in likelihood of 3.84 (i.e., $\chi^2(3.84, df = 1) = .05$).

In our example the full model has one estimated parameter (i.e., the covariance—we are ignoring the mean here for simplicity). The restricted model—with covariance equal to zero —has zero estimated parameters because we fix the covariance to be zero and refrain from estimating it. The restricted model would be selected if the AIC of that model is lower than the AIC of the full model. This would only happen if (−2 log(MLE_0) + 2κ_0) − (−2 log(MLE_1) + 2κ_1) is positive, and since we know $\kappa_1 = 1$ and $\kappa_0 = 0$ we know that $2\kappa_1 - 2\kappa_0 = 2 - 0 = 2$; that is, it will only be positive if log(MLE_0) − log(MLE_1) is greater than 2. Hence, we will conclude that the covariance parameter is zero at a statistical significance level of $\chi^2(2, 1) = .16$, a threshold that does not sit well with many peer reviewers. This high level of statistical significance is due to the AIC's relatively small penalty of 2κ. A greater penalty will yield more stringent levels of significance. The BIC, which for $N > 7$ has a greater penalty than the AIC, will give a lower significance level. If in our example there were 300 observations contributing to the covariance estimation, then the significance level implied by using the BIC would be .017. If we had 10000 observations the significance level would be .002.

That the penalty of BIC increases with N makes statistical significance more and more diffcult to achieve as the sample size increases. This attribute of the BIC is a direct consequence of its more general consistency property, to be discussed next. The consistency property of the BIC means that it is guaranteed to select the true model as the sample size grows infinitely large. In fact, Any criterion with a fixed statistical significance cutoff (such as with the AIC or a $p < .05$ threshold) will, with nonzero probability, include as significant parameters that are actually zero, no matter the sample size. Unlike the AIC, the BIC controls for this problem by making the parameter inclusion threshold more stringent as the sample size grows (i.e., the BIC's Type-1 error rate goes to zero whereas the AIC's does not).

Consistency, Effciency, and Minimax Properties: Theoretical Motivations for AIC and BIC

AIC and its Motivation from Kullback-Leibler Divergence

The AIC was derived by Akaike (1974) as an estimate of expected relative Kullback-Leibler (K-L) divergence. K-L divergence is just one kind of loss function (a familiar analogue would be Euclidian distance). It measures the distance, so to speak, between a candidate model and the true model—the closer the distance, the more similar the candidate to the truth. K-L divergence is just one kind of divergence measure, but the relative merits of K-L divergence are argued forcefully in Burnham and Anderson (2003, 2005), and has a somewhat intuitive interpretation in information theory. Imagine a distribution $g(\mathbf{y})$ and an encoding p of that distribution in bits. Now imagine not having p, but rather a different code q that is based on $f(\mathbf{y})$, which is not equal to $g(\mathbf{y})$. K-L divergence is the extra number of bits required to describe $g(\mathbf{y})$ with q compared to p (Cover & Thomas, 2006), and represents the informational ine ciency of using q instead of p. K-L divergence takes the form

$$\mathrm{KL}(g \| f) = \int g(\mathbf{y}) \log \frac{g(\mathbf{y})}{f(\mathbf{y})} \mathrm{d}\mathbf{y}, \quad (2)$$

where $g(\mathbf{y})$ is taken to be the true model probability density function (p.d.f.) and $f(\mathbf{y})$ the candidate model p.d.f. The K-L divergence cannot be known without knowledge of the true distribution $g(\mathbf{y})$, but in model comparisons this fact is inconsequential as $g(\mathbf{y})$ is the unchanging true distribution and does not depend on the sample at hand. The *relative* differences between candidate models will be the same whether the truth, $g(\mathbf{y})$, is known or not. The AIC implicitly estimates the divergence between the true model and the candidate model. This explains why K-L divergence is explicitly defined for two models and the AIC for only one. Because the true model is unknown, the absolute divergence between a candidate model and the true model is also unknown, but the relative differences *between* models can be used to rank order models according to their expected K-L divergence. The candidate model with the lowest AIC also has the lowest expected K-L divergence, even though the actual K-L divergence is unknown (because the true model is unknown). It is the expected, relative, K-L divergence that can be shown to be estimated by the AIC under certain assumptions (Burnham & Anderson, 2003; Chow, 1981; Bozdogan, 1987). An accessible derivation of the AIC as an estimator of K-L divergence is given in Burnham and Anderson (2003, pp. 362-371). We do not reproduce it here, but instead highlight one important feature of the general derivation.

The AIC is only a consistent estimator of K-L divergence if the true model is among the models under consideration (Burnham & Anderson, 2003). This is because AIC is a specific form of a more general estimator, the Takeuchi Information Criterion (TIC; Takeuchi, 1976; Chow, 1981). The $\boldsymbol{\kappa}$ in AIC is only a correct penalty of $l_{\mathbf{y}}(\widehat{\tau})$ if the true model is in the candidate model set. If the true model is not among the candidates then $\boldsymbol{\kappa}$ is biased. The TIC replaces $\boldsymbol{\kappa}$ in Equation (1) by a trace function of the product of expected Fisher Information matrices $\mathbf{J}_f\mathbf{J}_g^{-1}$, where the expectations are taken with respect to candidate model $f(\mathbf{y})$ and true model $g(\mathbf{y})$. If candidate model $f(\mathbf{y})$ happens to be the true model, then asymptotically these matrices are identical, and the trace reduces to $\mathrm{tr}\left(\mathbf{J}_f\mathbf{J}_g^{-1}\right) = \mathrm{tr}(I_{\kappa+\kappa})$, where I is the identity matrix (Burnham & Anderson, 2003, p. 368). When the true model is not under consideration the trace term may deviate from $\boldsymbol{\kappa}$. The TIC is appropriate for model selection regardless of whether the true model is a candidate; however, the TIC's modified penalty may be susceptible to high variability. This has led some (e.g., Shibata, 1983, 1989) to recommend the AIC even when the true model is not under consideration, as estimation

error of TIC may be more problematic than bias of AIC. The issue is unsettled in general, as the bias incurred by the AIC will be situation-specific.

An additional consideration when using AIC is its relatively poorer performance when κ is large relative to N. Burnham and Anderson (2003) recommend using a variant of AIC termed AIC_c, which takes the form $\mathrm{ACI}_c = ACI + \frac{2\kappa(\kappa+1)}{N-\kappa-1}$. When κ of the largest candidate model under consideration is small relative to N the AIC and AIC_c converge to the same value. Note that this additional bias correction is precise for multivariate normal OVs and linear regression, but the exact correction will differ for different models. McQuarrie and Tsai (1998) highly recommend AIC_c based on extensive simulation work.

BIC and its Bayesian Motivation and Interpretation

The Bayesian Information Criterion (BIC) has a theoretical motivation in Bayesian statistical analysis, especially the Bayes Factor (Kass & Raftery, 1995; Kass & Wasserman, 1995; Kass & Vaidyanathan, 1992; Kuha, 2004). In this section we change our notation slightly, and use the vertical bar "|" to denote a conditional probability distribution. We use this change to indicate explicitly the Bayesian view that model parameters are random variables whose probability distributions index degree of knowledge or belief about the true unknown fixed parameter value.

Bayesian estimation of parameters $\boldsymbol{\tau}_\ell$ for some model M_ℓ in light of observed data $\mathbf{y}$ takes the following form:

$$p(\tau_\ell|\mathbf{y},\ M_l) \propto p(\mathbf{y}|\tau_\ell,\ M_\ell)\, p(\tau_\ell|\ M_\ell) \tag{3}$$

Here we see that the posterior probability distribution of the parameters on the left-hand side is proportional to the probability distribution of the data given the parameters and model, as well as the prior probability distribution of the parameters given the model in the first place, before any data were observed. $p(\boldsymbol{\tau}_\ell|M_\ell)$, the prior probability distribution of the parameters given the model, must be specified *a priori* by the researcher.

Equation (3) determines how prior beliefs about the parameter values should be updated in light of observed data. There is nothing devious about Bayesian data analysis, and these equations are derived from basic algebraic equalities. The typical objection to Bayesian analysis does not involve the formalism. Instead, it is often objected that there is no reasonable way to determine an appropriate prior probability distribution, a predicament which may never be resolved. The selection of a prior probability distribution does affect Bayesian conclusions, and choosing a prior is not to be taken lightly (for a direct discussion of priors and BIC, see Weakliem, 1999).

Equation (3) can be modified to adjudicate among candidate models to select that model with the highest posterior probability of being correct. First, one selects a model family M_ℓ and prior probability distributions for $p(\boldsymbol{\tau}_\ell)$ and $p(M_\ell)$. These priors are then updated to posteriors $p(M_\ell|\mathbf{y})$ after observing $\mathbf{y}$. The parameters $\boldsymbol{\tau}_\ell$ of model M_ℓ are necessary to describe the model, but are integrated out for purposes of arriving at the quantity of interest, namely the posterior probability of M_ℓ conditional on the observed data:

$$p(M_\ell|\mathbf{y}) \propto p(\mathbf{y}|M_\ell)\, p(M_\ell) \tag{4}$$

where $p(M_\ell)$ is the prior probability for the ℓth model. To obtain $p(\mathbf{y}|M_\ell)$ we integrate $\boldsymbol{\tau}_\ell$ out of the right-hand side of Equation (3),

$$p(\mathbf{y}|M_\ell) = \int_{\tau_\ell \in \mathbf{T}_\ell} p(\mathbf{y}|\tau_\ell,\ M_\ell)\, p(\tau_\ell|M_\ell)\, \mathrm{d}\tau_\ell,$$

where $p(\mathbf{y}|\boldsymbol{\tau}_\ell, M_\ell)$ is the likelihood (as opposed to the log likelihood). Instead of maximizing the likelihood, Bayesian analysis considers all possible values over the range of the prior and calculates an average over them, hence the integral with respect to $\boldsymbol{\tau}_\ell$. We can see, then, that any model and parameters can be estimated, the parameter estimates are subsequently integrated out, and a final probability of model ℓ calculated. If we are only interested in the relative evidence for one model over another we can ignore the proportionality constant in Equations (3) and (4) (which are the same for all models) and take the ratio of evidence for one model over another:

$$\mathrm{BF} = \frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} \frac{p(M_1)}{p(M_2)}. \tag{5}$$

This ratio is called the Bayes Factor (BF), and it gives the weight of evidence in favor of model 1 over model 2, conditional on the prior probability (prior to any data being gathered) that model 1 is correct compared to model 2. In other words, it gives the odds ratio for model 1 over model 2. The Bayes Factor is statistically consistent, in that it will with probability 1.0 select the true model when N is large (Kass & Raftery, 1995), if the true model happens to be under consideration. A general discussion of Bayes Factors can be found in Gelman, Carlin, Stern, and Rubin (2004). To use the BIC to obtain an estimate for the BF for model 1 versus model 2 one would take the difference between the BIC for model 1 and the BIC for model 2. Because the BIC is on a logarithmic scale, it actually estimates 2 log(BF). Because the BIC estimates the BF, we can use the BIC itself to get posterior odds that model 1 is more probable than model 2, by exponentiating the difference between the BIC for model 1 and 2 (e.g., $e^{BIC_{M1} - BIC_{M2}}$). The exponentiation is necessary because of the BIC's logarithmic scale.

The BIC is usually given with correction $\log(N)\boldsymbol{\kappa}$, but it is useful to consider a more general version of it,

$$\mathrm{BIC}^* = 2l(\widehat{\tau}|\mathbf{Y}) - \log\left(1 + N/N_{prior}\right)\kappa \tag{6}$$

where N is the sample size and N_{prior} represents the sample size on which the prior is based (Atkinson, 1981; Kuha, 2004). That is, N_{prior} is the evidential weight of the prior. If N is large and $N_{\mathrm{prior}} = 1$ then BIC* ≈ BIC. Unlike the more transparent Bayes Factor, where the prior probability of the ℓth model must be specified, no prior probabilities on the parameters or the models are explicitly acknowledged in BIC. However, if BIC is an estimate of BF it must assume some prior, just as the BF does. (Keep in mind that there are as many different Bayes Factors for a single data set and model as there are possible prior probability p.d.f.'s.) The BIC is a better estimate of the BF under some priors than others—one presumes that applied researchers are tacitly assuming those priors for which the BIC works well, if they wish to use a posterior probability interpretation of BIC.

Applied researchers may have some expert judgement as to the distribution of model priors $p(M_\ell)$ as well as the parameters of each model $p(\boldsymbol{\tau}_\ell|M_\ell)$. For example, on the basis of previous research some ensconced model may have garnered evidence over some other. In that case one might assign a larger prior to the ensconced model, and lesser priors to competing models. However, and especially in social sciences, it may be diffcult to construct a defensible prior about the model parameters $\boldsymbol{\tau}$, such as factor loadings,

intercepts, mixing proportions, etc. In many applications there is very little known *a priori* about parameter values, and in that situation Kass and Wasserman (1995) argue that a defensible prior for τ_ℓ is multivariate normal

$$p(\tau_\ell|M_\ell)\sim\text{Norm}\left(\widehat{\tau}_\ell;\frac{N}{N_{\text{prior}}}\widehat{\Sigma}_\ell\right) \quad (7)$$

where $\widehat{\tau}_\ell$ is the maximum likelihood estimate of the model parameters and $\widehat{\Sigma}_\ell$ the estimated variance matrix of $\widehat{\tau}_\ell$, both of which are estimated given the observed data $\boldsymbol{y}$. Under these priors the BIC difference between the candidate model and true model will approximate 2 log(BF) with error of $O(N^{-1})$. That is, the BIC's error of estimation is asymptotically zero; as N increases the BIC more nearly equals 2 log(BF). However, the BIC is not so accurate in other circumstances. In general, such when the prior is poorly chosen, the BIC has error $O(1)$ in approximating the BF (Kass & Vaidyanathan, 1992; Kass, Tierney, & Kadane, 1990; Raftery, 1996; Tierney & Kadane, 1986). The “Big O” notation $O(1)$ indicates that error in approximation is not proportional to N. That is, the BIC would not converge to 2 log(BF), no matter how large the sample size is.

Of note is the role of N_{prior} in Equation (6) and (7). If N_{prior} is large, then the prior variance is small, indicating that probability density in the prior is concentrated around $\widehat{\tau}_\ell$. If N_{prior} is small, then the prior probability distribution's variance is large, and assigns probability density more evenly across the parameter space. In particular, if $N_{\text{prior}} = 1$, as Kass and Wasserman (1995) suggest, nearly nothing is known about the model or parameters prior to data collection and analysis. In other words, it quantifies the beliefs of a naive researcher. The quantification of naivete is precisely why Kass and Wasserman (1995) claim the BIC is useful as a rough-and-ready model fit statistic.

A final word about the BIC and BF: the interpretation of N in the BIC. Kass and Wasserman (1995, p. 993) note that N must be chosen carefully and, in fact, is the number of observations that contribute to the likelihood function. In regression, for example, N becomes the number of data points that contribute to the summation that appears in the formula for the likelihood. However, other models are not quite so straightforward. Pauler (1998) has shown that in repeated measures designs the effective N is a function of the correlations between observations. She provides some evidence that in mixed-effects models (which includes factor models and growth curves) N should be the number of subjects, regardless of repeated measurements. Raftery (1986) shows that in log-linear models for contingency tables N is the sum of the counts, not the number of cells. In survival analysis Raftery, Madigan, and Volinsky (1995) determined that N should be the total number of *uncensored* observations (e.g., deaths).

Consistency of BIC versus Efficiency of AIC

BIC receives its greatest theoretical motivation through its consistency property, which can be stated roughly as follows: As N grows large the BIC selects the true model with probability approaching 1. We already noted this property with respect to Bayes Factors above. A powerful result like this requires a host of assumptions. In this situation some important assumptions are (a) the true model is under consideration; (b) the true model's dimension (denoted κ_0) remains fixed as N grows; and (c) the number of parameters in the true model is finite. (Nishii, 1984; Shibata, 1983; Shao, 1997). Unlike the BIC, the AIC is not consistent under these circumstances.

The AIC will fail to select the true model with non-vanishing probability as N grows large, even when the true model is under consideration. To be consistent, a model selection

criterion must have a penality that approaches infinity as N approaches infinity (strong consistency, or convergence almost surely, requires the penalty to increase with N no more slowly than log log(N) (Hannan & Quinn, 1979)). The AIC's fixed penalty of 2 is non-increasing in N. Note that the problem here is not with overparameterized and structurally *incorrect* models, but with overly parameterized correct models. That is, the AIC will select, with nonzero probability, a more general case of the true model than is necessary (Shao, 1997). The consistency property makes BIC quite attractive, as it allows the researcher to conduct what amounts to data-driven theory building. One can collect data, fit models, and compare models with BIC, knowing that as N grows large the true model will be selected. There is a caveat, however, as this approach would require knowledge that the true model is in the candidate model set, which some have argued is never the case (Anderson & Burnham, 2002; Burnham & Anderson, 2003; McDonald, 2010).

The situation changes, however, when the number of parameters in the true model is infinite or increases with increasing N, or the true model is not in the candidate model set. When the true model is of infinite dimension (i.e., an infinity of parameters would be required to perfectly model the true functional form), or is not included in the candidate model set, it is diffcult to conceptualize consistency—there is no possibility of selecting the true model. In these circumstances one considers efficiency (or minimization) for some loss function such as mean squared error of prediction or estimation. If the number of parameters in the true model is infinite, or increases with increasing N, or if the true model is not in the candidate model set, then the AIC is asymptotically efficient in mean squared error of estimation/ prediction and in K-L divergence, whereas the BIC is not (Berger, Ghosh, & Mukhopadhyay, 2003; Chow, 1981; K. Li, 1987; Shibata, 1981, 1983; Stone, 1979). The BIC is efficient when the true model is among the candidates. In this context, efficiency just means that the loss function of interest is asymptotically minimized—that it is as small as possible given the candidate models.

In general, it appears that the BIC is preferred when there exists a fixed, finite dimensional true model, while the AIC is preferable when the true model is too complex to estimate parametrically (Shao, 1997). These differences are relevant to the distinction between parametric and non-parametric modeling. In parametric modeling the true function is assumed to be from a parametric family of functions with a fixed finite number of parameters, and models are fitted within the confines of that parametric model family. The BIC is preferable when the true model is assumed to be parametric and similar to candidate parametric models. In non-parametric modeling the true function is assumed to be too complex to adequately model with a known parametric function, and candidate functions are chosen that provide the best trade-off between bias and variance error in minimizing the loss function. As a general rule of thumb, the AIC is preferable in this case (Shao, 1997).

Minimizing Maximum Risk

The BIC consistency results described above are pointwise; that is, they hold for some true model as N grows large. To state this differently, pointwise asymptotics tell us how to expect the BIC and AIC to behave for given true and candidate models, with given parameter values, when the sample size grows large. A drawback of pointwise asymptotic theory is that the rate of convergence can depend on the true value of the parameters, and finite samples behave in sometimes unexpected ways. For some models and parameter values the BIC will perform more poorly in rate of convergence than for others. Interestingly, when the true model has a fixed finite number of parameters and N is finite, the BIC can possess worst-case risk greater than AIC. AIC, on the other hand, can minimize maximum risk in model selection (Y. Yang, 2005, 2007; Barron, Birgé, & Massart, 1999; Atkinson, 1980, 1981). In other words, while AIC may not be asymptotically consistent in selecting the true model, when N is finite it will not select a really bad model—but the BIC

may. This property is termed *minimax rate of convergence in risk*; AIC minimizes the maximum possible risk in model selection.

The AIC's minimax rate of convergence in risk (and the BIC's lack of it) is a highly general result in regression and density estimation, and holds under more general circumstances than required for consistency properties discussed above. For example, the AIC is minimax-rate optimal regardless of whether the true model is finitely or infinitely dimensional; it is optimal whether the true model is among the candidates or not; and it extends to highly non-linear models, such as functions defined in Sobolev spaces (Y. Yang, 2005, 2007; Barron et al., 1999; Y. Yang, 1999). Y. Yang (2007) provides a very accessible example of minimax-optimal convergence rates with simple linear regression, where the AIC, BIC, and $p < .05$ significance tests are compared. Given the general results in regression, we expect minimax properties of the AIC to extend to factor analysis. Whether the results apply to mixture models such as latent profile or factor mixture models is, to our knowledge, less clearly stated in the statistical literature (e.g., see, Burnham & Anderson, 2003). This remains an area for future analytical and simulation work.

Supplementing Analytical Results: Monte Carlo Investigations in Latent Variable Models

Monte Carlo studies of the AIC and BIC are useful supplements to analytical work. Asymptotic results describe what happens in the limit, as N grows arbitrarily large. Determining the rate of convergence is much more difficult, and usually simulations are necessary for guidance in applied problems (where the sample size is not arbitrarily large!).

Most frequently, simulations of latent variable model fit focus on a single criterion (e.g., AIC), and evaluate the probability that it selects a data-generating (and therefore “true”) model. Loss functions other than zero/one, such as K–L divergence, MSE of estimation/prediction, have not been studied in this literature.

To our knowledge five simulation studies have been published that directly compare the utility of AIC and BIC in latent variable model selection, and all were done with zero/one loss functions (Nylund, Asparouhov, & Muthén, 2007; Lubke & Neale, 2006; F. Li, Cohen, Kim, & Cho, 2009; C. Yang & Yang, 2007; Henson, Reise, & Kim, 2007).

Lubke and Neale (2006) studied the usefulness of AIC and BIC (among other criteria) in model selection for continuous and categorical latent variable models. The parameter space was small. In particular, the authors considered the effect of large class separation (distance between class means on the OVs) and class proportions on the ability of the AIC and BIC to select the data-generating model. All OVs were continuous. Data-generating models included two and three class latent class models (denoted 2C and 3C, respectively); one and two factor FA models (denoted 1F, 2F, and 3F, respectively), and FMMs with 2 classes/one factor and two classes/two factors (denoted 2C1F and 2C2F, respectively).

The relative performance of the AIC and BIC depended on three things: (a) the complexity of the true model, (b) class separation (low versus high separation), and (c) class proportions. For example, when the data-generating model was 2C, 3C, 1F or 2F the BIC outperformed the AIC. When erring, the AIC overfitted. When the data-generating model was a more complex FMM, the AIC outperformed the BIC; indeed, the BIC often had zero probability of selecting the true FMM. When erring, the BIC underfit. The AIC was especially impressive when class sizes were quite disparate (e.g., class proportions of .9 and .1) or classes were not well separated. According to this study it appears that the BIC pays a penalty for its consistency property, in that moderate to small effects will not be

included in the BIC-selected model unless the sample size is large (this is also noted in Y. Yang, 2007).

Nylund et al. (2007) focused on zero/one loss in selecting the correct number of classes in models with categorical latent variables, and found essentially the same trend as in Lubke and Neale (2006). However, Nylund et al. (2007) explored models with generally larger effects than Lubke and Neale (2006), and found more support for BIC. As we might expect from the analytical results reported above, the BIC will outperform the AIC in selecting the true model when the effects are large and the true model is under consideration. C. Yang and Yang (2007) evaluated the AIC and BIC (and many other criteria) under similar circumstances as Nylund et al. (2007), but with a more in-depth focus on latent class analysis. Again, the BIC did relatively well compared to the AIC when the data was generated under a simple model (e.g., one or two classes) and when class separations were larger. The AIC outperformed the BIC in selecting the true model when it was complex and/or class separations were smaller.

Henson et al. (2007) evaluated the AIC and BIC with factor mixture models and continuous observed variables. The data was always generated with two classes, each composed of a single latent factor. Then FMMs of one, two, and three classes were fit to the data. The usual story was told. When erring, the AIC tended to overfit (select a three-class model) and the BIC underfit (selecting the one-class model). The AIC did relatively better when effects were small (e.g., mixing proportions were .9 and .1 for the classes) and relatively worse when effects were large (e.g., mixing proportions at 50/50). In a study of FMMs with categorical variables (mixture IRT models), F. Li et al. (2009) obtained similar results, although found much greater support for the BIC in the parameter space studied there. The true model was varied from a one-parameter to a three-parameter IRT model, the number of OVs was varied from 15 to 30 items, sample sizes were 600 or 1200, and true models had one to 4 classes. The BIC correctly selected the true model 100% of the time when there were 1–3 classes, regardless of other conditions. The BIC only faltered when the true model was as complex as possible, three-parameter models with four classes. In that case the AIC outperformed the BIC.

These Monte Carlo studies show some strong trends. First, the authors clearly placed a premium on selecting the true model (zero/one loss). Second, the true models are all simple and have relatively large effects (small number of classes/factors). The smallest mixing proportion, for example, was 0.10. The BIC outperformed the AIC under those circumstances in which one would expect it to do so. Namely, when the true model is under consideration, selection of the true model is the measure of success, and the effects are relatively large (e.g., simple true models). When effects are smaller or the true model more complex the BIC performs relatively worse, and can be inferior to the AIC even when the true model is under consideration. These trends existed in all five articles reviewed. These themes are now expanded in two novel simulations.

Novel Simulations To Illustrate Asymptotic and Finite-Sample Properties

The simulations are broken into two parts. First, we evaluate AIC's and BIC's performance in a simple model selection scenario where the true model is in the candidate model set. We simulate data from a two-factor model and then fit one-, two-, and three-factor models. The performance of AIC and BIC is measured under three common loss functions: (a) zero/one loss, (b) accuracy in estimating the true covariance matrix (for which we calculate, for each element in the covariance matrix, the deviation from the estimated value and the true value, and take the average of all such deviations), and (c) minimizing maximum risk (minimaxity) in estimating the true covariance.

In the second simulation the true model is not in the candidate model set and is highly non-linear and complex. While technically the model is parametric, it is meant to reflect a complex non-parametric situation. Factor and class models are fit and selected by AIC and BIC. The performance of AIC and BIC is measured by only one loss function: estimating the true covariance matrix.

1. Performance of the AIC and BIC When the True Model is Parametric and Simple: Investigating Multiple Loss Functions

Simulation data was generated from a two-factor model with nine OVs. In the data-generating true model, loadings on the first factor (F_1) were in every case fixed at .85, .8, .75, 0, 0, 0, 0, 0, 0. Loadings on the second factor (F_2) were in every case fixed at zero for OVs 1–3 and 7–9. The loadings from F_2 onto OVs 4–6 were varied from 0 to .60, in increments of .01. This represents the major manipulation of the first simulation. It allows one to understand the behavior of the model selection criteria under a variety of effects, from very small to very large. Keep in mind that the true model was a one-factor model when Factor F_2's loadings were all zero, and a two-factor model when its loadings were anywhere between .01 and .60. In short, we arranged to have the true model range from a barely two-factor model (with very small effects, or loadings, onto F_2) to a clear-cut two-factor model (with very large effects, or loadings, onto F_2).

Fifty datasets were generated for each F_2 loading value (0 – .60), and for sample sizes of N = 500, 1000, 2000, and 5000. Three candidate models were then fit to these generated data: a one factor model with estimated loadings on OVs 1–3 and loadings on OVs 4–9 fixed at zero. A two-factor model with loadings from OVs 1–3 onto F_1 freely estimated, OVs 4–6 onto F_2 freely estimated, and loadings on OVs 7–9 fixed at zero. The three-factor model was a further extension of the two-factor model with loadings from OVs 7–9 onto yet another factor F_3 freely estimated. This is a simple model selection scenario. The true model is always in the candidate model set, but its loadings from OVs 4–6 on F_2 vary from 0 to .60. Since the true model is in the candidate model set, we expect BIC to have an advantage over AIC in this scenario. However, based on the Monte Carlo work discussed above, in these finite sample sizes we might expect the AIC to outperform the BIC when the effects (loadings onto F_2) are relatively small but nonzero.

It is important to notice that this simulation is unlike those reported in the previous section about existing Monte Carlo work. Here the focus is finite-sample performance across a wide range of parameter values. We are interested in overall performance, for each sample size, and for each of the possible parameter values. One should not necessarily expect the pointwise asymptotic properties to hold for all aspects of this simulation.

Zero/One Loss—Zero/one loss performance is depicted for the range of F_2 loadings (from 0 to .60) for each sample size in Figure 1. When the true model was a one-factor model (when loadings onto $F_2 = 0$) the BIC outperformed the AIC, for all sample sizes. When the true model was a two-factor model (loadings on $F_2 > 0$) the AIC outperformed the BIC up to a point, but then the BIC surpassed the AIC in performance. When the effects were small (i.e., loadings onto F_2 were small) the BIC's penalty of $\log(N)\kappa$ was too harsh, and it underfitted relative to the AIC by overly selecting the one factor model. However, the chart also belies the AIC's tendency to overfit. For example, even when $N = 5000$, the AIC never selected the two factor model with 1.0 probability, even when the effects were quite large (i.e., loadings on F_2 are .6). In these cases the AIC was selecting the three-factor model approximately 16% of the time whereas the BIC demonstrated its consistency property by selecting the true two-factor model 100% of the time. However, this only occurred at $N =$

5000 when the effects were large (> .15). When effects of F_2 were small the BIC ignored them and selected the one-factor model.

Note that model selection based on the AIC and BIC can be interpreted like statistical significance tests on the F_2 factor loadings. When the sample size was small (e.g., $N = 500$), the loadings had to be relatively large, or the AIC and BIC ignored them as non-significant. The BIC had less tolerance than the AIC for small loadings (i.e., requires a lower p-value), something that is necessary in a consistent model selection criterion. Unless the sample size was greater than 5000, the BIC ignored (true) factors with loadings less than .10. As the sample size grows arbitrarily large, the BIC will select models with factors that have arbitrarily small nonzero loadings. The AIC never entirely ignores factors with small loadings, regardless of the sample size.

MSE of Estimation and Minimizing Maximum Risk—Using the same simulation set-up, we now illustrate in Figure 2 minimization of the maximum risk in estimating the true covariance matrix. The MSE of estimating the covariance matrix is plotted in the top row of the figure. Based on results for zero-one loss in Figure 1, one might predict that the AIC will minimize loss, relative to the BIC, whenever the data-generating model is the two-factor model, which it is in every case except when loadings onto $F_2 = 0$. We see in the top row of Figure 2 that this was not the case. It was the BIC, not the AIC, that had lower risk when loadings are small. The BIC ignored the small loadings by selecting the one-factor model. This behavior is desirable when the true values of the loadings are small because the variability in parameter estimation, combined with the very small true effect of those parameters on the estimated covariance matrix, tend to result in more misfit when the effects are estimated than if they are ignored. The BIC ignored these small effects, whereas the AIC estimated them, and the BIC had concomitant lower risk than the AIC. This trend occured up to a point (about .27 when $N = 500$). After that point, the BIC persisted in selecting the one-factor model to its own detriment, at the expense of MSE, whereas the AIC did not. For $N = 500$ this is about .27 – .41. After that the BIC started selecting the two-factor (true) model just like the AIC and had similar risk. For those high loadings (> .4 for $N = 500$) the BIC actually did better than the AIC, because it always correctly selected the two-factor model, whereas the AIC occasionally selected the three-factor model and was again overfitting at the expense of the MSE.

Perhaps more interesting is the trend plotted in the bottom row of Figure 2. To obtain these figures, we re-expressed the information from the top row of panels. For each F_2 loading we have an estimate of the risk associated with using the AIC (AIC-risk) and the risk associated with using the BIC (BIC-risk). To obtain the relative risk of using the AIC compared to the BIC we subtracted the AIC-risk from the smaller of the AIC-risk or BIC-risk (this is the black line in the bottom row of figures). To obtain the relative risk of using the BIC compared to the AIC we subtracted the BIC-risk from the smaller of the AIC-risk and the BIC-risk (the red line). This gives us the relative risk of each criterion with respect to the other. Consider the first plot, where $N = 500$. At first, the AIC has worse relative risk. The AIC's relative risk hovers around .005 and then decreases to zero at an F_2 loading of about .27. It decreases to zero because this is the point where the AIC and BIC curves cross in the top row of Figure 2; it is the point where the BIC and AIC perform equally in risk. From .27 to about .41 the AIC's relative risk is zero, because AIC is minimizing MSE in the top row of Figure 2. The BIC, on the other hand, has much worse risk (relative to AIC) for loadings between .27 and approximately .41. For higher loadings the AIC again performs worse than the BIC (because it is selecting the three-factor model and overfitting). Of note is that, across the entire range of F_2 loadings, the BIC incurs the maximum possible risk for each sample size. That is, the BIC is responsible for the largest hump in each plot in the lower

array of Figure 2. In our scenario the AIC minimizes the maximum possible risk, and has the minimax property discussed above.

Ironically, as the sample size grows the maximum possible risk of BIC becomes larger relative to AIC's maximum possible risk. That is, even though the BIC is a consistent model selection criterion it performs increasingly worse in relative risk as N grows (but remains finite). We can compute the maximum possible risk of the BIC relative to the AIC by dividing the maximum of the BIC by the maximum of the AIC. For $N = 500$ this value is 4.29; for $N = 1000$ the value is 4.56; when $N = 2000$ the value increases to 5.29. When $N = 5000$ the value jumps to 8.02. If the trend displayed in Figure 2 continues as N increases, and AIC's maximum relative risk (relative to BIC) reduces at a rate much faster than the BIC's, then the maximum possible relative risk of BIC for large N is very large. This result is well-known in regression and density estimation (Y. Yang, 2007; Barron et al., 1999; Foster & George, 1994), although this is to our knowledge the first discussion of it in a latent variable model.

To summarize, the BIC pays a heavy penalty in maximum risk for its consistency property. Despite its conistency it clearly can fail in finite sample sizes, even in the unlikely and extremely favorable circumstances where the true model is under consideration, such as our simulation.

2. Performance of the AIC and BIC When the True Model is Highly Non-Linear and Complex

Now we describe a separate simulation where the true model is nonlinear. The intent here is for the true model to be, practically speaking, non-parametric. It is difficult to simulate data from a truly non-parametric model, so we use a highly non-linear model as the best alternative. The true model was generated as follows, with sample sizes of 500, 1000, 2000, and 5000. First, we generated three independent variables v_1, v_2, v_3:

$$\upsilon_1 \tilde{} \text{Norm}\,(\text{mean} = 6, \text{var} = .4)\,,$$

$$\upsilon_2 \tilde{} \text{Norm}\,(\text{mean} = 9, \text{var} = .8)\,,$$

$$\upsilon_3 \tilde{} \text{Norm}\,(\text{mean} = 1, \text{var} = .5)\,.$$

These three variables were then used to generate correlated observed variables using the following regressions:

$$o\upsilon_1 = \upsilon_1 + \upsilon_2 + .5\upsilon_3 + 2^{\upsilon_1} x,$$

$$o\upsilon_2 = \upsilon_1 + \upsilon_2 + x\sin(\upsilon_3)\,,$$

$$o\upsilon_3 = \upsilon_1 + x\upsilon_2^{1.5} + \upsilon_3$$

$$o\upsilon_4=\upsilon_1+.01x\upsilon_1^2+.01x\upsilon_1^3+\upsilon_2+x\sin(\upsilon_3).$$

The variable x was increased from 0 to 3 by increments of .25. This was intended to shift the model from being practically linear to being highly non-linear. Of course, technically, these are parametric models, but they are quite complex parametric functions, so much so that we consider them practically non-parametric for this simulation.

To these data were fit a series of non-nested models including a 1-factor model (1F), latent profile models with 2–7 classes (denoted 2C, 3C, and so on), and two factor mixture models, one with two classes each composed of a factor (2C1F), and another with three classes each composed of a factor (3C1F). In the FMM loadings and intercepts were fixed to be equal across classes for simplicity and to aid in optimization. The range of factor and class models provides flexibility for selecting models that can better capture non-linearity in the data, as compared to simply fitting linear factor models with varying numbers of factors. AIC and BIC were used to select the best model, then the mean squared error of estimating the covariance matrix was computed (similar to that displayed in the top row of figures in Figure 2).

To give the reader a sense of how realistic our simulation data might be we plotted in Figure 3 a scatterplot matrix for the simulated data when $x = 1$. Correlations between the OVs range from about .3 to about .8, and the univariate density estimates appear highly normal. We see some non-linearity in the off-diagonals, where a non-parametric loess smoother has been applied to each scatterplot (in red, with the red dotted lines representing 95% confidence intervals).

Presenting figures or tables displaying what models were selected for each sample size and each x-value would take up too much space. Instead, a major trend is highlighted only for $N = 1000$—the trend is similar for other sample sizes. For $x = 0$ and $N = 1000$ the BIC selected the 2C1F FMM 100% of the time. For $x = .25$ the BIC selected the 1F (86%) and the 6C class model (14%). For $x > .25$ the BIC invariably selected the 1F factor model, even though the extent of nonlinearity increased substantially in x and $o\upsilon_2$ became noticeably platykurtic. The AIC, on the other hand, showed higher variability in model selection. When $x = 0$ it selected the 2C1F FMM (96%) and the 3C1F (4%). When $x = .25$ it selected the 7C class model (100%). When x=.50 it selected the 1F (48%), 7C (40%) and 2C1F (12%). When $x = .75$ the AIC selected 1F (88%), 7C (4%), 2C1F (4%) and 3C1F (4%). The pattern for each $x > .75$ was largely similar to that for $x = .75$, at least for $N = 1000$. Clearly, the AIC will tend to select more complex models than the BIC. Perhaps unexpected is the BIC's tendency under these particular circumstances to select the simplest model available, despite increasingly complex observed data. A clear potential drawback to the AIC in this scenario (and we expect in general) is the wide variability in its model choice. Under the same data-generating scenario the AIC waffles significantly in its model choice and, in some cases, it may be that the BIC's bias (toward the simpler model) may be preferable to AIC's higher variability. The bias versus variance trade-off is just another complication of model selection, and choosing between AIC and BIC.

Mean squared error of estimating the covariance matrix is plotted in Figure 4, displayed the same way as in the top array of figures in Figure 2. There are several things to note in this figure. First, in general, the AIC performs increasingly better than the BIC as the data become more nonlinear. The improvement, at least in this circumscribed example, appears to hit an asymptote quite early at $x = .5$. Second, as the sample size grows, AIC tends to improve (slightly) relative to BIC, and does so at smaller x-values. It is important to

understand that this simulation is extremely limited, and there are non-parametric scenarios that can be modeled extremely well with simple parametric models and the BIC will outperform the AIC in those circumstances, despite the true model being non-linear and not in the candidate model set. Whether the AIC or the BIC is a better choice depends on the true effects in the data and what candidate models are under consideration. One could very easily simulate data that is non-linear, but still can be well-captured with linear models. That simulation would result in conclusions favorable to the BIC despite the true model not being among the candidates. The extent to which the true model is non-parametric may guide the choice between AIC and BIC. In fact, the notion of the true model is extremely important in the applied use of the AIC and BIC. We turn to this issue next.

How The True Model Influences The Choice Between AIC and BIC

Up to this point we have not discussed the notion of the "true model." It is clear from the preceding that one's approach to model selection, and the criteria used for it, depend on the true model. In statistics "true model" has a technical definition: it is the model that generated the data. This is how the term is used throughout this article. Asymptotic and simulation results thus far have depended on the status of the true model. Does it contain large or small effects? Is it in the candidate set? Is it of finite dimension? Does its dimension increase with increasing N? These are all questions to consider before using the AIC or BIC.

Here we spend a few paragraphs explicating the notion of a true model aware that this issue is extremely complex and likely will never be fully resolved. One often hears the adage "[A]ll models are wrong, but some are useful" (Box & Draper, 1987, p. 424). The statistician D. Cox (1995, p. 456) drives the point home more forcefully:

> [I]t does not seem helpful just to say that all models are wrong… The idea that complex physical, biological, or sociological systems can be exactly described by a few formulae is patently absurd.

Some argue that the truth in principle *cannot* be fully modeled, and offer analytical information-theoretic arguments to support this claim (Kolmogorov, 1968; Rissanen, 1987). When one reflects on the notion of a true model, and considers the extreme complexity of psychological systems (depending on the most complex human organ, the brain), it becomes less clear how a true model could be devised, much less that it has already been devised at this stage in our science. In this respect, where the true model captures the entire complexity of human behavior (or some other physical system), it seems useless to speak of a true model.

Even if the true model could in principle be under consideration, the complexity of candidate models may change with increasing N, increased scientific knowledge, and increasingly clever research designs. Fisher (1922) stated "More or less elaborate forms will be suitable according to the volume of the data." That is, one expects the candidate model pool to change as N grows large. Imagine fitting a regression of order 50 to a data set of a few hundred or a few million. High-order models are more defensible with large data sets.

We venture to suggest that no statistician or psychologist ever believes they have a true model under consideration, if by that we mean a complete listing of all relevant variables, all non-zero weightings and interactions between them, and precisely specified error distributions. Instead, psychological theories, and the statistical models they imply, possess levels of verisimilitude, with no model (or theory) having perfect verisimilitude (Meehl, 1990). Models are simplifications of reality, and the effect of unmeasured variables is presumed to be captured by the error distribution. That said, one can imagine some models, although not comprehensive models of behavior, as being, practically speaking, true.

The applicability of a practically true model depends on the type of study and the research goals. In some cases a simple model may be legitimate and justifiable (even if the error distribution is not perfectly correctly specified). Sometimes particular parametric statistical models are entirely reasonable—imagine modeling a coin toss with the binomial distribution (analogous to any event with a dichotomous outcome, such as treatment success or failure), or the number of shots it takes to make a basket in basketball with a geometric distribution. In these cases, where the physical properties of the system are well-understood and the statistical question well-defined, one may have good reason to consider the true model as practically parametric. In other cases, such as modeling the covariance structure of personality via factor analysis, it may not be justifiable to assume that the model(s) under consideration are appropriate, as the physical system under study is much less well understood (at least compared to a coin toss).

When the physical system is well-defined (e.g., in a true experiment where observations are randomized) the primary objective of the analysis can often be parameter estimation and computing confidence intervals (Cox, 1977). When the physical system is largely unknown then the correct model is much less clear. Sometimes the data are too sparse to support a useful model. That does not mean some specific candidate model is not scientifically useful, but it is much less clear if the model assumptions are satisfied, and model selection might be conducted to help ensure that the model under consideration is a useful one for the purposes of the scientific investigation at hand.

Considering the status of the true model is relevant to choosing between the AIC and BIC. Consider non-experimental observational data, where the physical system is not understood, there are a few large effects, many small tapering effects, and a small sample size. The true model here may be non-parametric, but there also may exist a parametric model that captures the vast majority of systematic variation with only a few parameters to estimate. Pragmatically, the parametric model is here the best model and an appropriate loss function would be MSE (not zero/one loss). However, due to the large effects in the data a researcher would be tempted to use the BIC to select it. Note this is consistent with our simulations (e.g., in Figure 2), where the BIC outperformed the AIC when effects were very small (tapering effects) and when effects were very large. The BIC would ignore small (but true) effects in the data, and obtain a smaller MSE in estimating the covariances. As the sample size increases, it may become reasonable to expand the model to capture the small effects, since estimation of more parameters becomes feasible with increasing N. As the sample size becomes quite large, the best approach may become practically non-parametric because one can now with confidence model some of the many small tapering effects, and the same researcher might discard the BIC in favor of using the AIC.

Now imagine a different scenario, where there are no large effects, many moderate effects, and few small effects. Here a non-parametric model may well be more capable of capturing moderate effects in the data, even at small sample sizes, and the AIC would be preferred to minimize MSE. Again, note that this is consistent with our results in Figure 2, where the AIC outperformed the BIC when the effects were moderate, but not when they were very large or very small. As the sample size increases it may be that the parameters of moderate size can be estimated with greater precision, and the best model becomes, practically speaking, parametric, with the BIC as the criterion of choice.

Choosing between the AIC and BIC in these situations depends on knowledge of the true model, which is difficult to have in practice. A more recent development in model selection is to use the data to guide the selection of a model selection criterion like AIC or BIC. For example, Liu and Yang (in press) have devised a *parametricness index* that converges to ∞ when the data are governed by a parametric model and 0 when they are governed by a non-

parametric model. The index provides a data-driven way to choose between AIC and BIC in the instant data set, and may resolve to some extent problems associated with assuming the true model exists (or does not). The parametricness index is an example of adaptive model selection criteria, a more recent movement in statistical theory where the choice of a model selection criterion is itself data-driven (Kadane & Lazar, 2004).

Summary and Conclusions

LV models can be used to suggest and test psychological theories, including etiological theories. For example, if class models fit a patient population well, then continuously distributed etiologies (e.g., blood pressure) may be less likely to cause the disease. Of course, today's best LV models will in the future be abandoned for better models, using better measures, and substantiated by more verisimilar substantive psychological theory. The change may be stark, such as switching from a continuous to a discrete model. This is not undesirable, but a reflection of the progress of the science. At each stage model results should guide substantive scientific considerations, and new scientific insights should influence future modeling.

AIC, BIC, and other model selection criteria like them, are motivated primarily by theoretical research in regression and density estimation. This work indicates that the BIC is consistent in selecting the true model when that model is a candidate (and other important assumptions). The AIC minimizes useful risk functions when the true model is not a candidate, or the candidate model is extremely complex. Even if the true model is under consideration the BIC is not a clear choice. This is the point about minimaxity above. In finite sample sizes (like those actually used in all psychological research) the BIC can perform worse than the AIC, even when the true model is under consideration. It should be stressed that this appears true even in large of sample sizes, and the bigger the sample the worse the problem. It is unclear at present whether all these asymptotic results extend to mixture models such as latent profile or factor mixture modeling, that is an area for future simulation and analytical work.

The choice between the AIC and BIC depends on one's notion of the true model. If the true model is assumed to be complex, with large, moderate, and small effects, and the candidate models oversimplifications, then the AIC may be preferred to the BIC. This depends on the complexity of the true data-generating process, the scientific background knowledge about the physical processes under study, and the quality and adequacy of the candidate models. To betray our bias, we expect the true model is quite complex in many areas of psychology, and is at best woefully modeled by current approaches. This is perhaps a defensible position merely given the lack of predictive and explanatory power of current models in personality, I/O, clinical, social, counseling, developmental, or individual differences psychology.

Simulation studies are still in their infancy, although it appears from such work that the BIC performs better than the AIC under zero/one loss when the data-generating model has a few large effects and simple candidate models are under consideration. This result is foreseeable from the asymptotic properties described above for regression. In future work we suggest that methodologists place Monte Carlo results within the context of asymptotic work.

Future Monte Carlo work is necessary to understand the behavior of AIC and BIC in finite sample sizes. Like the present report, this work should consider multiple loss functions, larger portions of the parameter space, and more complex true and candidate models. It should consider cases where the true model is not in the candidate model set. Such work would be more informative to the applied scientist, and help remind methodologists and applied researchers alike that science deals with extremely complex realities and quirky data sets. Real data analysis is extremely challenging with at least one prominent

psychometrician intentionally describes model fitting as an “art,” not a science (MacDonald, 2010).

Acknowledgments

The author expresses deep thanks to William M. Grove, Matt McGue, William G. Iacono, Niels G. Waller, Jeffrey D. Long, Kristian Markon and four anonymous reviewers for insightful discussions and comments on earlier drafts of this article.

References

Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19:716–723.

Anderson D, Burnham K. Avoiding pitfalls when using information-theoretic methods. Journal of Wildlife Management. 2002; 66:912–918.

Atkinson A. A note on the generalized information criterion for choice of a model. Biometrika. 1980; 67:413–418.

Atkinson A. Likelihood ratios, posterior odds, and information criteria. Journal of Econometrics. 1981; 16:15–20.

Barron A, Birgé L, Massart P. Risk bounds for model selection by penalization. Probability Theory and Related Fields. 1999; 113:301–413.

Bauer DJ, Curran PJ. The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. Psychological Methods. 2004; 9

Berger J, Ghosh J, Mukhopadhyay N. Approximations and consistency of Bayes factors as model dimension grows. Journal of Statistical Planning and Inference. 2003; 112:241–258.

Bollen, K. Structural equations with latent variables. Wiley; New York, NY: 1989.

Box, GE.; Draper, NR. Empirical model-building and response surfaces. Wiley; New York: 1987.

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

Breiman L. Statistical modeling: The two cultures. Statistical Science. 2001; 16:199–215.

Burnham, K.; Anderson, D. Model selection and multimodel inference: A practical-theoretic approach. Springer-Verlag; New York, NY: 2003.

Burnham K, Anderson D. Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods & Research. 2005; 33:261–304.

Burt SA. Rethinking environmental contributions to child and adolescent psychopathology: A meta-analysis of shared environmental influences. Psychological Bulletin. 2009; 135:608–637. [PubMed: 19586164]

Chow GC. A comparison of the information and posterior probability criteria for model selection. Journal of Econometrics. 1981; 16:21–33.

Cover, T.; Thomas, J. Elements of information theory. 2nd ed.. Wiley-Interscience; New York, NY: 2006.

Cox DR. The role of significance tests. Scandinavian Journal of Statistics. 1977; 4:49–70.

Cox DR. Model uncertainty, data mining, and statistical inference (with discussion). Journal of the Royal Statistical Society. 1995; 158:455–456.Series A

Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin. 1955; 52:281–302. [PubMed: 13245896]

Digman J. Personality structure: Emergence of the five-factor model. Annual Review of Psychology. 1990; 41:417–440.

Fisher RA. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. 1922; 222:309–368.Series A

Foster DP, George EI. The risk inflation criterion for multiple regression. Annals of Statistics. 1994; 22:1947–1975.

Gelfand A, Dey D. Bayesian model choice: Asymptotics and exact calculations. Journal of the Royal Statistical Society. 1994; 56:501–514.Series B

Gelman, A.; Carlin, J.; Stern, H.; Rubin, D. Bayesian data analysis. 2nd ed.. Chapman & Hall; Boca Raton, FL: 2004.

Hannan E, Quinn B. The determination of the order of an autoregression. Journal of the Royal Statistical Society. 1979; 41:190–195.Series B

Heinen, T. Latent class and discrete latent trait models: Similarities and differences. Sage; Thosand Oaks, CA: 1996.

Henson JM, Reise SP, Kim KH. Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. Structural Equation Modeling. 2007; 14:202–226.

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

Kass R, Raftery A. Bayes factors. Journal of the American Statistical Association. 1995; 90:773–795.

Kass, R.; Tierney, L.; Kadane, J. The validity of posterior asymptotic expansions based on Laplace's Method. In: Geisser, S.; Hodges, J.; Press, S.; Zellner, A., editors. Bayesian and likelihood methods in statistics and econometrics. North-Holland; New York, NY: 1990. p. 473-488.

Kass R, Vaidyanathan S. Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. Journal of the Royal Statistical Society. 1992; 54:129–144.Series B

Kass R, Wasserman L. reference bayesian test for nested hypothesis and its relationship to the schwarz criterion. Journal of the American Statistical Association. 1995; 90:928–934.

Kolmogorov A. Three approaches to the quantitative definition of information. International Journal of Computer Mathematics. 1968; 2:157–168.

Krueger R. The structure of common mental disorders. Archives of General Psychiatry. 1999; 56:921–926. [PubMed: 10530634]

Kuha J. AIC and BIC: Comparisons of assumptions and performance. Sociological Methods Research. 2004; 33:188–229.

Lazarsfeld, P.; Henry, N. Latent structure analysis. Houghton Mifflin; Boston, MA: 1968.

Li F, Cohen AS, Kim S-H, Cho S-J. Model selection methods for mixture dichotomous IRT models. Applied Psychological Measurement. 2009; 33:353–373.

Li K. Asymptotic optimality for c_p, c_l, cross-validation and generalized cross-validation: Discrete index set. Annals of Statistics. 1987; 15:958–975.

Liu W, Yang Y. Parametric or nonparametric? A parametricness index for model selection. Annals of Statistics. in press.

Lubke GH, Muthén BO. Investigating population heterogeneity with factor mixture models. Pyschological Methods. 2005; 10:21–39.

Lubke GH, Neale MC. Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? Multivariate Behavioral Research. 2006; 41:499–532.

McDonald RP. Structural models and the art of approximation. Perspectives on Psychological Science. 2010; 5:675–686.

McLachlan, G.; Peel, D. Finite mixture models. Wiley-Interscience; New York: 2000.

McQuarrie, A.; Tsai, C. Regression and time series model selection. World Scientific; Danvers, MA: 1998.

Meehl PE. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology. 1978; 46:806–834.

Meehl PE. Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant using it. Psychological Inquiry. 1990; 1:108–141.

Molenaar, P.; van Eye, A. On the arbitrary nature of latent variables. In: van Eye, A.; Clogg, C., editors. Latent variable analysis: Applications for developmental research. Sage; Thousand Oaks: 1994. p. 226-242.

Mosteller, F.; Tukey, JW. Data analysis and regression. Addison-Wesley; Reading, MA: 1977.

Mulaik, S. Foundations of factor analysis. 2nd ed.. Chapman & Hall; Boca Raton, FL: 2010.

Muthén B. Beyond sem: General latent variable modeling. Behaviormetrika. 2002; 29:81–117.

Muthén, B. atent variable hybrids: Overview of old and new models. In: Hancock, G.; Samuelsen, K., editors. Latent variable mixture models. Information Age Publishing, Inc; Charlotte, NC: 2008.

Neale, M.; Cardon, L. Methodology for genetic studies of twins and families. Kluwer Academic Publishers; Dordrecht: 1992.

Nishii R. Asymptotic properties of criteria for selection of variables in multiple regression. Annals of Statistics. 1984; 12:758–765.

Nylund K, Asparouhov T, Muthén B. Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. Structural Equation Modeling. 2007; 14:535–569.

Pauler D. The Schwarz criterion and related methods for normal linear models. Biometrika. 1998; 85:13–27.

Plomin R, Daniels D. Why are children in the same family so different from one another? Behavioral and Brain Sciences. 1987; 10:1–60.

Raftery A. A note on Bayes Factors for log-linear contingency table models with vague prior information. Journal of the Royal Statistical Society. 1986; 48:249–250.Series B

Raftery A. Approximate Bayes Factors and accounting for model uncertainty in generalised linear models. Biometrika. 1996; 83:251–266.

Raftery, A.; Madigan, D.; Volinsky, C. Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). In: Bernardo, J.; Berger, J.; Dawid, A.; Smith, A., editors. Bayesian statistics. Vol. 5. Oxford University Press; Oxford, U.K.: 1995. p. 323-350.

Rissanen J. Stochastic complexity. Journal of the Royal Statistical Society. 1987; 9:223–239.Series B

Schaid DJ. Genomic similarity and kernel methods 1: Advancements by building on mathematical and statistical foundations. Human Heredity. 2010; 70:109–131. [PubMed: 20610906]

Schwarz G. Estimating the dimension of a model. Annals of Statistics. 1978; 6:461–464.

Shao J. An asymptotic theory for linear model selection (with discussion). Statistica Sinica. 1997; 7:221–242.

Shibata R. An optimal selection of regression variables. Biometrika. 1981; 68:45–54.

Shibata R. Asymptotic mean e ciency of a selection of regression variables. Annals of the Institute of Statistical Mathematics. 1983; 35:415–423.

Shibata, R. Statistical aspects of model selection. In: Willems, J., editor. From data to model. Springer-Verlag; London, UK: 1989. p. 215-240.

Spearman C. General intelligence, objectively determined and measured. American Journal of Psychology. 1904; 15:201–293.

Stone M. Comments on model selection criteria of Akaike and Schwarz. Journal of the Royal Statistical Society. 1979; 41:276–278.Series B

Takeuchi K. Distribution of informational statistics and a criterion of model fitting. Mathematical Sciences. 1976; 153:12–18.

Tierney L, Kadane J. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association. 1986; 81:82–86.

Weakliem DL. A critique of the Bayesian Information Criterion for model selection. Sociological Methods & Research. 1999; 27:359–397.

Yang C, Yang C. Separating latent classes by information criteria. Journal of Classification. 2007; 24:183–203.

Yang Y. Model selection for nonparametric regression. Statistica Sinica. 1999; 9:475–499.

Yang Y. Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. Biometrika. 2005; 92:937–950.

Yang Y. Prediction/estimation with simple linear models: Is it really that simple? Econometric Theory. 2007; 23:1–36.

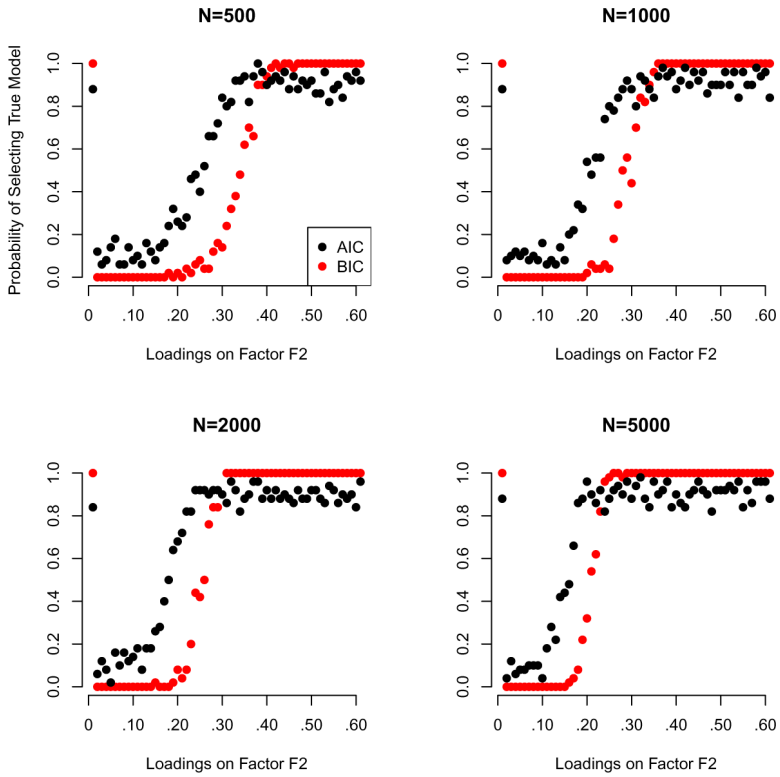

Figure 1.

AIC and BIC performance in selecting the true model, when the true model's effect sizes range from very small to large. The effect size is a factor loading that varies from zero to .6. Thus, the factor F_2 loading along the x-axis is the true loading. When the loading is zero, then the true model is a one-factor model, and BIC outperforms the AIC in selecting the one factor model (this occurs once in each panel). When the loading is nonzero the true model is a two-factor model, and plotted here is the probability that the AIC (or BIC) selected the two-factor model. Despite the BIC's consistency property, the AIC outperforms it for a range of loadings.

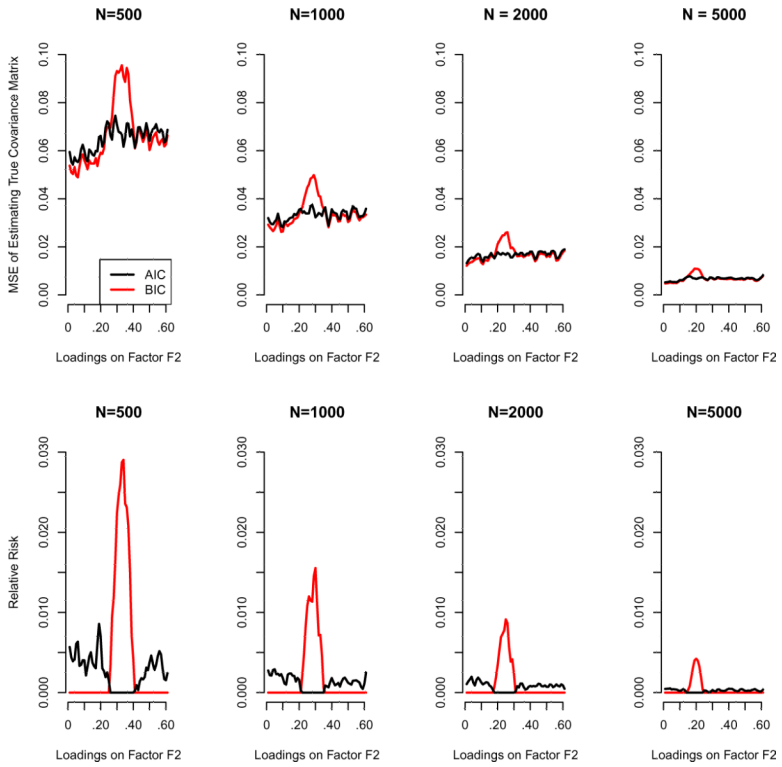

Figure 2.

AIC and BIC performance in minimizing mean squared error of estimating the true covariance matrix (in the upper array of four plots). These plots are created from the same simulation used to create Figure 1. Notice that the BIC outperforms the AIC for lower loading values. For lower loading values the BIC is selecting the one-factor model but the AIC is selecting the two-factor model (as can be seen in Figure 1). This is due to the higher penalty of log(*N*) that the BIC places on the more complex two-factor model. The upshot is that the BIC ignores the effect of these very small loadings by selecting the one factor model. This works in its favor because it outperforms the AIC in MSE. As the loadings of the data-generating two-factor model increase the BIC persists in selecting the one-factor model to its detriment, and the AIC begins outperforming it in MSE. This occurs up to the point where the loadings are too large for the BIC to ignore, and it begins outperforming the AIC again because it starts selects the true, two-factor model, every time, whereas the AIC errs at times and selects the three factor model. In the lower array is plotted the relative risk in mean squared error, which is a re-expression of the information in the top array of panels. It is simply the AIC minus the minimum of the AIC and BIC (plotted in black) or BIC minus the minimum of the AIC or BIC (plotted in red). The BIC yields the maximum possible risk in each sample size (has the highest value in each of the lower array of plots), whereas the AIC minimizes the maximum possible risk. Each plot had a loess smoother with a small span applied, to aid in visual interpretation.

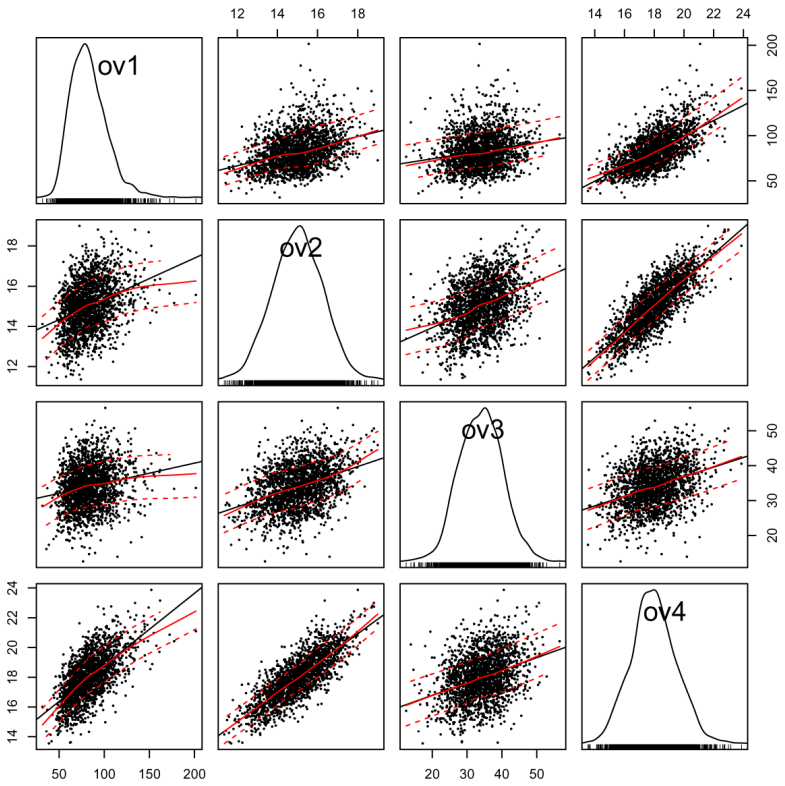

Figure 3.
Scatterplot matrix of observed data for the non-linear simulation when $x = 1$, where x is the degree of nonlinearity described in the text. Correlations range from about .3 to about .8. The red lines are loess-smoothed regressions, and indicate the extent to which the regression is non-linear.

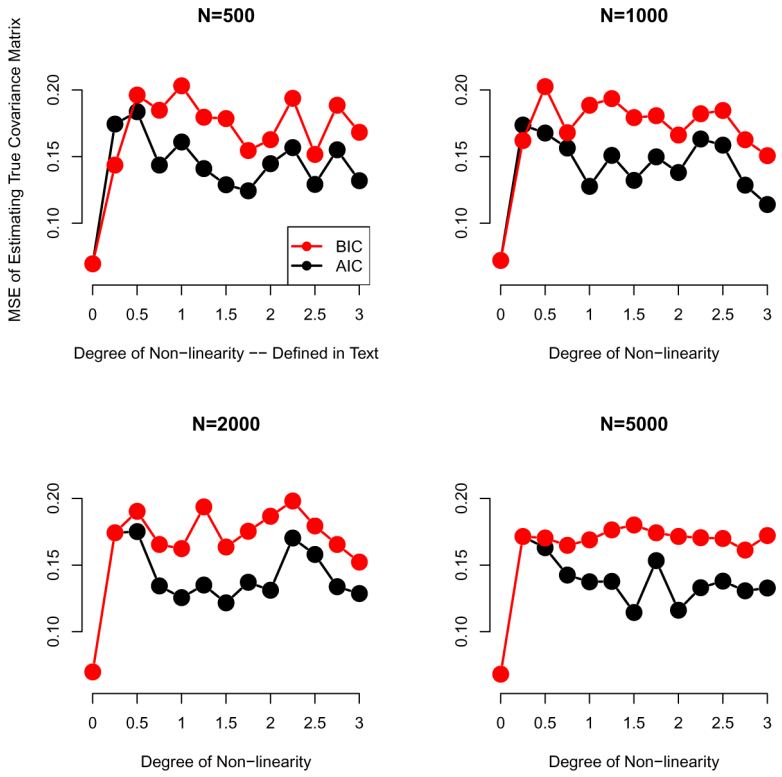

Figure 4.

AIC and BIC performance when the true model is not in the candidate model set. The risk function here is minimizing mean squared error of estimating the true covariance matrix under increasing amounts of non-linearity. When the data is linear (degree of non-linearity is zero) the AIC and BIC perform equally well. For small amounts of non-linearity the BIC outperforms the AIC for $N = 500$ and $N = 1000$ because the AIC is overly sensitive to inconsequential small true effects. As the degree of non-linearity increases the BIC persists in underfitting (selecting the one-factor model) much to its detriment, as the AIC begins outperforming it in MSE. See text for details on how the non-linearity was manipulated. Each dot represents an average MSE for 50 replications.