Bayesian Model Selection in Structural Equation Models

Adrian E. Raftery
University of Washington [1]

August 28, 1991; revised February 18, 1992

[1]Adrian E. Raftery is Professor of Statistics and Sociology, DK-40, University of Washington, Seattle, WA 98195. This research was supported by NIH grant no. 5R01HD26330-02. The author is grateful to David Madigan for helpful discussions, to Ken Bollen, Herb Costner, Scott Long and Michael Sobel for very useful comments, and to Sandy Welsh for excellent research assistance.

Abstract

A Bayesian approach to model selection for structural equation models is outlined. This enables us to compare individual models, nested or non-nested, and also to search through the (perhaps vast) set of possible models for the best ones. The approach selects several models rather than just one, when appropriate, and so enables us to take account, both informally and formally, of uncertainty about model structure when making inferences about quantities of interest. The approach tends to select simpler models than strategies based on multiple P-value-based tests. It may thus help to overcome the criticism of structural equation models that they are too complicated for the data they are applied to.

Contents

1 Introduction

The class of structural equation models is broad, and its very generality makes model selection difficult. This is especially so when the underlying theory does not allow one to specify the model structure fairly completely in advance. In that case, the number of models initially entertained can be very large and some exploratory model selection strategy is necessary. A common approach is to use theory to specify an initial model, and then to use a sequence of tests based on P-values to decide whether the model should be simplified or expanded (e.g. Bollen, 1989; Long, 1983a,b).

There are many difficulties with such a strategy. The theory of P-values was developed for the comparison of two nested models. In a typical structural equation model application, there may be hundreds of substantively meaningful models, many of them non-nested. Wheaton's (1978) model for the sociogenesis of psychological disorders contains about 20 parameters, or arrows in the corresponding graph, some of which could be removed (see also Long, 1983b, Figure 4.1). Indeed, the key scientific issue may often be expressed in terms of *which* of the arrows can be removed. Thus, there may be up to about 2^{20}, or roughly one million competing models; this number will be smaller if theory or prior research enables us safely to assume in advance that some of the arrows are in or out, but will typically still be large. The sampling properties of the overall *strategy* in such

circumstances is unknown and may be very different from the properties of the individual tests that make it up (Miller, 1984; Fenech and Westfall, 1988).

Perhaps most fundamentally, conditioning on a single selected model ignores model uncertainty and so leads to underestimation of the uncertainty about the quantities of interest. This underestimation can be large, as was shown by Regal and Hook (1991) in the contingency table context and by Miller (1984) in the regression context. One bad consequence is that it can lead to decisions that are too risky (Hodges, 1987).

Even when there are only two models, M_0 and M_1, and they are nested, it is not clear that the P-value is a suitable criterion for model comparison. Indeed, Berger and Sellke (1987) and Berger and Delampady (1989) have argued that P-values and evidence are actually often in *conflict*; these articles are highly recommended to the reader.

Also, I have argued elsewhere (Raftery, 1986b) that significance tests based on P-values ask the wrong question. They ask whether the null model is "true", a question to which we already know that the answer is "no". A scientifically more relevant question is "Which model predicts the data better?" (i.e. under which model are the data more likely to have been observed?), and this leads to a different approach, the Bayesian one. One consequence is that tests based on P-values tend to reject the null model frequently with large samples of the sizes typically met with in sociology, even when M_0 predicts the data well; see Raftery (1986b) for a dramatic example with $n = 110{,}000$. By the same token, when there are many models, specification searches that consist of multiple P-value-based tests tend to yield models that are over-complicated.

An alternative that seems to overcome these problems is provided by the Bayesian approach, which is described in Section 2. In Section 3, the Bayesian approach is applied to structural equation modeling, model selection strategies are discussed, and an example is given. This approach is applicable whether the prior theory and research is strong, in which case the number of models considered will be small, or weak, in which case the number of models that could potentially be entertained is often enormous. In the latter case, the research is often exploratory.

The Bayesian approach tends to favor simpler, more parsimonious, models than the sequential P-value approach. This can lead to further simplification by allowing the removal of variables that now appear irrelevant to the research hypothesis, yielding less cluttered graphs and making interpretation easier. It has been said that structural equation modeling is a "castle built on sand", because it is based on models that are too complex for the data used to estimate them. On the other hand, structural equation

models do correspond well to the actual research hypotheses of social scientists. They also implement the famous advice of R.A. Fisher for the analysis of observational data, "Make your theories elaborate", so as to minimize the possibility of observed associations being spurious (i.e. due to unobserved variables), rather than causal (see also Blalock, 1979). I believe that Bayesian model selection for structural equation models may give us the best of both worlds: simple and interpretable models that are well supported by the data, and yet reflect the real research hypotheses and are based on sufficiently "elaborate" theories.

2 Bayesian Model Comparison

2.1 Comparing Two Models: Bayes Factors

We first consider the artificial situation where only two models, M_0 and M_1 are being compared. These may be any two competing structural equation models. This is relevant if, for example, we are interested in whether one particular parameter is zero or not, and we are prepared to assume that the model is otherwise well specified. The results are also useful as building blocks in the more general and realistic situation where there are many models.

The posterior odds for M_0 against M_1 given data D are

$$\frac{p(M_0|D)}{p(M_1|D)} = \left[\frac{p(D|M_0)}{p(D|M_1)}\right]\left[\frac{p(M_0)}{p(M_1)}\right] = B_{01}\lambda_{01}. \tag{1}$$

The first equality follows from Bayes' theorem. In equation (1), $\lambda_{01} = p(M_0)/p(M_1)$ is the *prior odds* for M_0 against M_1; it is often taken to be unity, representing "neutral" prior information that does not favor either model. The quantity $B_{01} = p(D|M_0)/p(D|M_1)$ is the *Bayes factor*, or ratio of posterior to prior odds, and $p(D|M_k)$ $(k = 0,1)$ is the *marginal likelihood* defined by

$$p(D|M_k) = \int p(D|\theta_k, M_k)p(\theta_k|M_k)d\theta_k, \tag{2}$$

where θ_k is the (vector) parameter of M_k, $p(\theta_k|M_k)$ is the prior distribution of θ_k and $p(D|\theta_k, M_k)$ is the likelihood.

The marginal likelihood $p(D|M_k)$ is also the *predictive probability* of the data given the model M_k, and so it measures how well the model predicts the data, which I would argue is the right criterion for model evaluation. Thus the Bayes factor measures how well M_0 predicts the data relative to M_1, and so is just the right quantity on which to base model

comparison. Note that this comparison does not depend on the assumption that either model is "true", unlike model comparisons based on P-values.

Finding B_{01} involves the multiple integral in equation (2), which is often of high dimension; this is not in general easy to evaluate analytically. However, good analytic approximations are available. A Taylor series expansion of the log-likelihood about the maximum likelihood estimator $\hat{\theta}_k$ shows that

$$2\log p(D|M_k) \approx 2\log p(D|\hat{\theta}_k, M_k) - \log|V_k| + 2\log p(\hat{\theta}_k|M_k) - \nu_k \log(2\pi), \tag{3}$$

with a relative error that is $O(n^{-\frac{3}{2}})$ (Chow, 1981). The first term on the right-hand side of equation (3) is twice the maximized log-likelihood, the second term is minus the log of the determinant of V_k, the variance matrix of $\hat{\theta}_k$, the third term is twice the log of the prior density at $\hat{\theta}_k$, and the fourth term is $-\log(2\pi)$ times $\nu_k = \dim(\hat{\theta}_k)$, the number of independent parameters in M_k. The first, second and fourth terms are readily obtained from the standard output of most statistical model-fitting programs. The third term is readily calculated and is generally negligible in moderate to large samples. Slightly more accurate approximations are available using the Laplace method (Raftery, 1988), but these are harder to calculate.

The matrix V_k/n tends to a constant matrix as $n \to \infty$, where n is the number of independent (scalar) observations that contribute to the likelihood. Thus the determinant of V_k/n, namely $|V_k/n| = |V_k|/n^{\nu_k}$, tends to a scalar constant, and so does its logarithm, $(\log|V_k| - \nu_k \log n)$. Thus, if we remove from the right-hand side of equation (3) all the terms that do not tend to infinity as $n \to \infty$, we obtain the cruder but simpler approximation

$$2\log p(D|M_k) \approx 2\log p(D|\hat{\theta}_k, M_k) - \nu_k \log n, \tag{4}$$

with a relative error of $O(n^{-1})$.

We thus have two approximations to the Bayes factor, namely, by equation (3),

$$-2\log B_{01} \approx L^2 - \log|V_0| + \log|V_1| - 2\log\left\{\frac{p(\hat{\theta}_0|M_0)}{p(\hat{\theta}_1|M_1)}\right\} + (\nu_1 - \nu_0)\log(2\pi), \tag{5}$$

where L^2 is the standard likelihood-ratio test statistic, and, by equation (4),

$$-2\log B_{01} \approx L^2 - (\nu_1 - \nu_0)\log n = BIC_{01}. \tag{6}$$

When $BIC_{01} > 0$, the criterion favors M_1, and when $BIC_{01} < 0$ it favors M_0. This result was first established by Schwarz (1978) for regression models and extended to log-linear

models by Raftery (1986a). It is usually stated in terms of twice the logarithm because of the connection with the likelihood-ratio test statistic, but one can recover the approximate Bayes factor itself from the equation

$$B_{01} \approx e^{-\frac{1}{2}BIC_{01}}. \tag{7}$$

In structural equation models, the Lagrange multiplier (LM) and Wald (W) tests are often used in place of the likelihood-ratio test because they can be much less expensive computationally.[1] The LM and W test statistics are asymptotically equivalent to L^2 (see Bollen, 1989, pp.293-296), and so may be used, at least roughly, in place of L^2 in equation (6). If M_0 is nested within M_1 and the only difference between them is that M_0 constrains one parameter of M_1 to be equal to zero, then the W test statistic is equal to t^2, where t is the relevant t-statistic. Thus, in that case, L^2 may be replaced by t^2 in equation (6).
For the comparison of two *nested* models, Jeffreys (1961, Appendix B) has suggested the following order of magnitude interpretation of the Bayes factor, B_{01}. (Corresponding approximate values of BIC_{01} from equation (7) are also shown.) If $B_{01} > 1$ ($BIC_{01} < 0$), the data favor M_0 and there is no evidence for the additional effects represented by M_1. If $1 \geq B_{01} > 10^{-1}$ ($0 \leq BIC_{01} < 4.6$), then there is weak evidence for M_1. If $10^{-1} \geq B_{01} > 10^{-2}$ ($4.6 \leq BIC_{01} < 9.2$), the evidence for M_1 is strong, while if $B_{01} \leq 10^{-2}$ ($BIC_{01} \geq 9.2$), the evidence for M_1 is conclusive. As a rough rule of thumb, I use BIC values of 0, 5 and 10 as cut-off points for the different "grades of evidence". Of course, the interpretation depends on the context. For example, Evett (1991) has argued that for forensic evidence alone to be conclusive in a criminal trial one would require posterior odds for M_1 (guilt) against M_0 (innocence) of at least 1000:1, rather than the 100:1 suggested by Jeffreys. This corresponds to a BIC value of about 14, but in such a matter one would also, presumably, want B_{01} to be calculated more accurately than by the BIC approximation.
The approximate t-values corresponding to different grades of evidence and different sample sizes are shown in Table 1. For minimal evidence this is $t = \sqrt{\log n}$, for "strong" evidence it is $t = \sqrt{\log n + 5}$, while for "conclusive" evidence it is $t = \sqrt{\log n + 10}$. Note that the critical values are larger than those based on P-values, and also that they increase with n.

2.2 Many Models: Accounting for Model Uncertainty

Suppose now that instead of just two models there are $(K + 1)$ models $M_0, M_1, \ldots, M_K$. In most studies, there is one, or perhaps a small number of quantities of primary interest such as the coefficient associated with a particular arrow in the graph that represents the

structural equation model; we denote such a quantity of interest by Δ. The full Bayesian solution to inference about Δ takes account explicitly of model uncertainty and is as follows. First we form the *posterior distribution* of Δ under each model M_k, namely the conditional distribution of Δ given the observed data D. If θ_k is the parameter of M_k (consisting of the identified parameters in the eight matrices that define the structural equation model), then the posterior distribution of θ_k is

$$p(\theta_k|D, M_k) = \frac{p(D|\theta_k, M_k)p(\theta_k|M_k)}{\int p(D|\theta_k, M_k)p(\theta_k|M_k)d\theta_k}, \tag{8}$$

and the posterior distribution of Δ is just the marginal distribution of Δ from equation (8), obtained by integrating out the other components of θ_k. The posterior distribution of Δ will usually be approximately normal if the sample size is reasonably large. It can often be well approximated by a normal distribution with mean $\hat{\Delta}_k$, the maximum likelihood estimator of Δ under M_k, and variance V_k, the (approximate) variance matrix of $\hat{\Delta}_k$; the precise form of the prior $p(\theta_k|M_k)$ often has little effect on the final inference. For further discussion of these results and the basic ideas of Bayesian inference see, for example, Edwards, Lindeman and Savage (1963), Box and Tiao (1973, ch.1) and Cox and Hinkley (1974, ch.10).

We then form the final posterior distribution of Δ as a weighted average of the posterior distributions of Δ under each of the models, weighted by their posterior model probabilities $p(M_k|D)$, namely

$$p(\Delta|D) = \sum_{k=0}^{K} p(\theta_k|D, M_k)p(M_k|D), \tag{9}$$

where

$$p(M_k|D) = \frac{p(D|M_k)p(M_k)}{\sum_{k=0}^{K} p(D|M_k)p(M_k)}. \tag{10}$$

If all the models have the same prior probability, then this may be approximated by

$$p(M_k|D) \approx \frac{e^{\frac{1}{2}BIC_{0k}}}{\sum_{k=0}^{K} e^{\frac{1}{2}BIC_{0k}}}, \tag{11}$$

where $BIC_{00} = 0$ and BIC_{0k} results from comparing M_k with M_0, considered as a baseline model.

Note that Bayes factors and BIC values for comparing two models may, in general, be computed by comparing each with a different, third model. Suppose that we want to

compare M_1 with M_2, and we compare each of them separately with M_0, then we have

$$B_{12} = \frac{B_{02}}{B_{01}}, \tag{12}$$

and

$$BIC_{12} = BIC_{02} - BIC_{01}. \tag{13}$$

An approximate combined point estimate and standard error can be calculated from the formulae:

$$E[\Delta|D] \approx \sum_{k=0}^{K} \hat{\Delta}_k p_k, \tag{14}$$

$$\mathrm{Var}[\Delta|D] \approx \sum_{k=0}^{K} (V_k p_k + \hat{\Delta}_k^2 p_k) - E[\Delta|D]^2, \tag{15}$$

where $p_k = p(M_k|D)$.

When the total number of models is very large, the sums over all models will be impractical to calculate. In Madigan and Raftery (1991), we argued that one should exclude from the sum

(a) models that are much less likely than the most likely model, say 10, 20 or 100 times less likely, corresonding to BIC differences of about 5, 6 or 10; and

(b) models containing effects for which there is no evidence, i.e. those that have more likely models nested within them.

We refer to the models that are left as falling within *Occam's window*, a generalization of the famous Occam's razor, or principle of parsimony in scientific explanation.

3 Application to Structural Equation Models

3.1 Bayes Factors for Structural Equation Models

The results in Section 2 apply quite generally to structural equation models. Note that in the definition of BIC, "n" is equal to the total number of (scalar) observations, namely $N(p+q)$ in the notation of Bollen (1989, Table 2.2), where N is the number of cases. Note also that for the calculation of BIC, most structural equation model software reports L^2 in terms of a comparison with the saturated model, M_S.

Suppose that we want to compare two models M_1 and M_2 and that for each model we have the likelihood ratio test statistic L_k^2 (relative to the saturated model), and the number of

degrees of freedom, df_k, equal to $\frac{1}{2}(p+q)(p+q+1) - \nu_k$ (k = 1,2). Then we may compare each model in turn with the saturated model M_S using

$$BIC_{kS} = L_k^2 - \mathrm{df}_k \log\{N(p+q)\}. \tag{16}$$

Thus, by equation (13) we may compare M_1 with M_2 using

$$BIC_{12} = BIC_{1S} - BIC_{2S}. \tag{17}$$

In what follows, when we write BIC without subscripts, we will be referring to BIC_{kS}. Note that M_k will be preferred to M_S if $BIC_{kS} < 0$. Thus BIC_{kS} may be regarded as a kind of "goodness of fit" test of M_k in the sense that if $BIC_{kS} > 0$, M_k will be "rejected" in favor of the saturated model. Note also that the more accurate approximation in equation (5) can be calculated fairly easily from the output of standard programs such as LISREL and EQS.

3.2 Search Strategy

How should we carry out the search for good models? The number of possible models can be huge and, in principle, each new model considered must be checked for identifiability, a non-trivial task.

It seems that the search should be limited from the outset by substantive considerations. One way to do this might be as follows. First, develop one or several "encompassing" models that are sufficiently elaborate to include all the postulated effects, and check that each of these models is identified. Then specify which arrows have to be in all the models considered, for substantive reasons or in light of previous empirical results. Then the search would be restricted to models that are nested within the "encompassing" models and that contain the "essential" arrows. Since these are all nested within identified models, they are themselves quite likely to be identified. However, this is not necessarily the case and needs to be checked for the preferred models.[2] I now outline two main types of search strategy.

3.2.1 Stepwise Strategies

Stepwise strategies have been used for many years for variable selection in regression, and have been extended to other contexts such as log-linear models (Goodman, 1971). They include forward selection, backward elimation and strategies that combine features of both. One possibility is a stepwise strategy that sequentially adds and deletes arrows in the structural equation model graph, with the criterion for making a change being that the

Bayes factor (or BIC) favor the change. Probably the most convenient criterion for this purpose is the approximation to BIC obtained by replacing L^2 by t^2 in equation (6) for deciding whether to remove an arrow, and the approximation that replaces L^2 by the LM test statistic or modification index (Bollen, 1989, p.293) for deciding whether to add an arrow. A good starting model could be found by first estimating the "encompassing" model, and including all the arrows that have t-values greater than the critical value for weak evidence, which, if BIC is used, is $\sqrt{\log n}$ (see Table 1).
Such a strategy would, hopefully, lead to the most likely single model, $M_{\max}$, say (although this is not guaranteed). One could then attempt to find the other models that lie in Occam's window by comparing $M_{\max}$ with models that differ from it by at most a few arrows, and retaining those models M_k such that $B_{\max,k}$ is less than some appropriate number such as 10, 20 or 100. Finally, models that have more likely models nested within them would be removed.

3.2.2 The Down–Up Algorithm

This algorithm was proposed by Madigan and Raftery (1991) in the context of graphical models for contingency tables, and it seems also to be applicable to structural equation models. It is a direct and efficient way of finding all the models in Occam's window. It proceeds through model space from larger to smaller models and then back again in a series of pairwise comparisons of nested models.
It eliminates models by applying the principle that if a model is rejected, then its submodels (i.e. the models nested within it) are also rejected. All models that have not been rejected are included in Occam's window. In a pairwise comparison of M_0 with M_1, where M_0 is nested within M_1, M_0 (and all its submodels) is rejected if $B_{01} < C^{-1}$, M_1 is rejected if $B_{01} > 1$, and both models are retained if $C^{-1} \leq B_{01} \leq 1$. Here C is a constant set equal to, for example, 10, 20 or 100. Generally, the smaller C is, the fewer models there are in Occam's window, but we have found that the results tend to be fairly insensitive to the precise choice of C.
We now outline the search technique. The search can proceed in two directions: "Up" from each starting model by adding arrows, or "Down" from each starting model by dropping arrows. When starting from a non-saturated, non-empty model, we first execute the "Down" algorithm. Then we execute the "Up" algorithm, using the models from the "Down" algorithm as a starting point. Experience to date suggests that the ordering of these operations has little impact on the final set of models. Let $\mathcal{A}$ and $\mathcal{F}$ be subsets of

model space $\mathcal{M}$, where $\mathcal{A}$ denotes the set of "acceptable" models and $\mathcal{F}$ denotes the models under consideration. For both algorithms, we begin with $\mathcal{A} = \emptyset$ and $\mathcal{F}$ =set of starting models.

Down Algorithm

1. Select a model M from $\mathcal{F}$
2. $\mathcal{F} \leftarrow \mathcal{F} - M$ (i.e. replace $\mathcal{F}$ by $\mathcal{F} - M$), and $\mathcal{A} \leftarrow \mathcal{A} + M$
3. Select a submodel M_0 of M by removing an arrow from M
4. Compute $B = \frac{p(M_0|D)}{p(M|D)}$
5. If $B > 1$ then $\mathcal{A} \leftarrow \mathcal{A} - M$ and if $M_0 \notin \mathcal{F}, \mathcal{F} \leftarrow \mathcal{F} + M_0$
6. If $C^{-1} \leq B \leq 1$ then if $M_0 \notin \mathcal{F}, \mathcal{F} \leftarrow \mathcal{F} + M_0$
7. If there are more submodels of M, go to 3
8. If $\mathcal{F} \neq \emptyset$, go to 1

Up Algorithm

1. Select a model M from $\mathcal{F}$
2. $\mathcal{F} \leftarrow \mathcal{F} - M$ and $\mathcal{A} \leftarrow \mathcal{A} + M$
3. Select a supermodel M_1 of M by adding an arrow to M
4. Compute $B = \frac{p(M|D)}{p(M_1|D)}$
5. If $B < C^{-1}$ then $\mathcal{A} \leftarrow \mathcal{A} - M$ and if $M_1 \notin \mathcal{F}, \mathcal{F} \leftarrow \mathcal{F} + M_1$
6. If $C^{-1} \leq B \leq 1$ then if $M_1 \notin \mathcal{F}, \mathcal{F} \leftarrow \mathcal{F} + M_1$
7. If there are more supermodels of M, go to 3
8. If $\mathcal{F} \neq \emptyset$, go to 1

Upon termination, $\mathcal{A}$ contains the set of potentially acceptable models. Finally, we remove all the models which have a more likely submodel, and those models M_k for which

$$\frac{p(M_{\max} \mid D)}{p(M_k \mid D)} > C. \tag{18}$$

$\mathcal{A}$ now contains the acceptable models, namely those in Occam's window.
The algorithm has been coded for *discrete* recursive causal models and also for graphical log-linear models. It is very efficient, carrying out about 3,000 model comparisons per minute on a workstation. Efficient computer implementation of the algorithm for structural equation models remains to be done.

4 Example

I now outline how some of these ideas might apply in an example. The data considered is that of Wheaton (1978) on the sociogenesis of psychological disorder (PD); this data was reanalyzed by Long (1983b). The main issue is whether low socio-economic status (SES) leads to PD, PD leads to low SES, both, or neither. Low SES causing PD is referred to as social causation, while PD leading to low SES is called social selection. One of Wheaton's data sets consisted of $N = 603$ individuals in Illinois for whom SES was measured at three different time points; two indicators of PD were also measured at each of the last two time points. Father's SES was also used, giving a total of eight observed variables and $n = 603 \times 8 = 4,824$, so that $\log n = 8.48$. Figure 1, reproduced from Long (1983b), shows the main variables and represents Long's model M_f, which was close to his preferred model for the data.
Twenty-two models for the data are shown in Table 2. Long (1983b, Table 4.2) fitted six models to the data, and the BIC values are shown in Table 2. Models M_a, M_b and M_c have positive BIC values, and so are unsatisfactory. Model M_d does fit slightly better than the saturated model, but model M_e fits much better. Model M_e is preferred to model M_f, even though the evidence is weak. Also, M_e is nested within M_f, so that if we considered only Long's six models to start with, only M_e would be in Occam's window, and we would not need to average over models to make inference about individual parameters. Also, M_e "fits" the data in the sense of being better than the saturated model (note that this is in conflict with the result based on P-values), but this does not preclude our search for better models. The remaining sixteen models document part of an implementation of the Down algorithm starting from model M_e, so that $\mathcal{F} = \{M_e\}$. Here we take $C = 20$, by analogy with the

popular 5% significance level for tests, so that $B_{01} < C^{-1}$ if $BIC_{01} > 6$. Thus we will reject the larger model M_1 if $BIC_{0S} < BIC_{1S}$, while we will reject the smaller model, M_0, if the difference between the BIC values is greater than 6, i.e. if $BIC_{01} = BIC_{0S} - BIC_{1S} > 6$. If BIC_{01} is between 0 and 6, both M_0 and M_1 will be retained for the time being.

Model 7 specifies that $\gamma_{31} = 0$ in model M_e, i.e. the γ_{31} arrow is removed. This model is preferred to M_e and is nested within it, so that M_e is now rejected and model 7 is retained for comparison with further models. Models 8–10 also represent simplifications of M_e that are preferred to it; thirteen other simplifications of M_e were also tried but were rejected and are not shown in Table 2. The comparison between models 11 and 13 is a case where both models were retained, with a BIC difference of 5 in favor of the larger model (although both were later rejected, model 11 in favor of model 15, and model 13 in the final phase of the algorithm, in favor of the best model, model 22).

In the end, only model 22 is in Occam's window, so that the result is fairly clear-cut. All the parameters in model 22 have highly significant t-values (all are greater than $\sqrt{\log n + 6} = 3.8$), none of the meaningful modification indices (LM test statistics) is significant according to BIC (i.e. greater than $\log n = 8.48$), and there are no outliers among the standardized residuals (each residual corresponds to one of the 36 sample covariances).

Model 22 is shown in Figure 2; it is considerably simpler than the model in Figure 1. Substantively, it leads to the conclusion that there is social causation but not social selection; the key feature that indicates this is the absence of β_{43}, the arrow from PD at time 2 (PD2) to SES at time 3 (SES3). Wheaton (1978) came to the same conclusion, but Long (1983b) pointed out that Wheaton's own model corresponds closely to model M_d, and that there β_{43} is significant, supporting the social selection hypothesis.

Why this conflict? It seems that the significance of β_{43} in M_d is an artefact due to the misspecification of M_d. As one moves towards better-fitting models, from M_d to M_e to model 15, $\hat{\beta}_{43}$ goes from -1.50 to -0.85 to a non-significant -0.47 ($t = -0.8$), while in models 16, 17 and 19, $\hat{\beta}_{43}$ remains close to -0.5 and non-significant. The non-significance of β_{43} is also indicated by the contrasts between models 19 and 22, and between models 15 and 18, which in each case favor the model that does not include β_{43}. In comparing M_d with model 22, note that model 22 not only fits substantially better in terms of L^2, but also uses seven less parameters. The estimates of all the parameters that are present in the favored model 22 remained quite stable across models, but this was not true for β_{43}.

A backward selection strategy starting from M_e that uses a sequences of P-value-based

significance tests at the 1% level would, like our approach, select model 22. However, a similar strategy at the 5% level would both choose model 20 rather than model 22, as would the AIC (Akaike, 1987), The only difference between the two models is that model 20 includes γ_{41} (the direct effect of Father's SES on SES at time 3), but model 22 does not. The P-value for γ_{41} in model 20 is 0.017, and even if one wishes to use P-values, a reasonable balance between power and significance would suggest a lower significance level with such a large sample size. In addition, model 20 is substantively unsatisfactory because it says that Father's SES has a direct on SES at time 3 but not at time 2. In any event, models 20 and 22 both lead to the same conclusions about the research question of primary interest.

5 Discussion

An approach to Bayesian model selection and accounting for model uncertainty in structural equation models has been described. This seems to hold out the promise of selecting simpler and more interpretable models, without compromising the richness of the structural equation model framework for representing substantive research hypotheses. When appropriate, several models are selected, rather than just one. Thus ambiguity about model structure is pinpointed and taken into account when making inference about quantities of interest.

Another common approach to model selection consists of selecting a single model based on an information criterion such as AIC (Akaike, 1987; Sclove, 1987; Bozdogan, 1987; Cudek and Browne, 1983). Note that BIC can be used in this way, as can the posterior probability or other approximations to it such as equation (5), to select the single model with the best value of the criterion; however this is not what is advocated here. A key difference between that approach and the one described here is that we select several models when appropriate and base inferences on all the selected models, thus taking account of uncertainty about model structure. While model uncertainty was not a crucial issue in the example discussed here, Madigan and Raftery (1991) gave several examples with multivariate data where it is important and failing to take account of it leads to inferior predictive performance as well as reduced scientific insight.

Notes

References

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

Berger, J.O. and Delampady, M. (1987) Testing precise hypotheses (with Discussion). *Statist. Sci.*, **2**, 317–352.

Berger, J.O. and Sellke, T. (1987) Testing a point null hypothesis: The irreconcilability of P values and evidence (with Discussion). *J. Amer. Statist. Ass.* **82**, 112–139.

Blalock, H.M. (1979) The presidential address: Measurement and conceptualization problems: The major obstacle to integrating theory and research. *Amer. Sociol. Rev.* **44**, 881-894.

Bollen, K.A. (1989) *Structural Equations with Latent Variables*. New York: Wiley.

Box, G.E.P. and Tiao, G.C. (1973) *Bayesian Inference in Statistical Analysis*. Reading, Mass.: Addison-Wesley.

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

Chow, G.C. (1981) A comparison of the information and posterior probability criteria for model selection. *J. Econometrics* **16**, 21–33.

Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*. London: Chapman and Hall.

Cudeck, R. and Browne, M.W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research* **18**, 147–167.

Edwards, W., Lindman, H. and Savage, L.J. (1963) Bayesian statistical inference for psychological research. *Psych. Rev.* **70**, 193–242.

Evett, I.W. (1991) Implementing Bayesian methods in forensic science. Paper presented at the Fourth Valencia International Meeting on Bayesian Statistics.

Fenech, A. and Westfall, P. (1988) The power function of conditional log-linear model tests. *J. Amer. Statist. Ass.* **83**, 198–203.

Goodman, L.A. (1971) The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics* **13**, 33–61.

Hodges, J.S. (1987) Uncertainty, policy analysis and statistics (with Discussion). *Statist. Sci.* **2**, 259–291.

Jeffreys, H. (1961) *Theory of Probability.* (3rd ed.), Oxford University Press.

Long, J.S. (1983a) *Confirmatory Factor Analysis*. Beverly Hills: Sage.

Long, J.S. (1983b) *Covariance Structure Models*. Beverly Hills: Sage.

Madigan, D. and Raftery, A.E. (1991) Model selection and accounting for model uncertainty in

graphical models using Occam's window. Technical Report no. 213, Department of Statistics, University of Washington.

Miller, A.J. (1984) Selection of subsets of regression variables (with Discussion). *J. R. Statist. Soc. (ser. A)* **147**, 389–425.

Raftery, A.E. (1986a) A note on Bayes factors for log-linear contingency table models with vague prior information. *J. R. Statist. Soc. (ser. B)* **48**,249–250.

Raftery, A.E. (1986b). Choosing models for cross-classifications. *Amer. Sociol. Rev.* **51**, 145–146.

Raftery, A.E. (1988) Approximate Bayes factors for generalised linear models. *Technical Report 121*, Department of Statistics, University of Washington.

Regal and Hook (1991) The effects of model selection on confidence intervals for the size of a closed population. *Statist. in Medecine* **10**, 717–721.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Sclove, S.L. (1987) Application of some model-selection criteria to some problems in multivariate analysis. *Psychometrika*, **52**, 333–344.

Wheaton, B. (1978) The sociogenesis of psychological disorder. *Amer. Sociol. Rev.* **43**, 383–403.

Table 1: Approximate minimum t-values for different grades of evidence and sample sizes

Evidence	Minimum BIC	$n =$			
		30	100	1000	10000
"Weak"	0	1.84	2.15	2.63	3.03
"Strong"	5	2.90	3.10	3.45	3.77
"Conclusive"	10	3.66	3.82	4.11	4.38

Table 2: Model comparison for Wheaton's data.

k	Model	L^2	df	BIC_{kS}
1.	M_a	142.6	13	32.3
2.	M_b	147.5	15	20.3
3.	M_c	131.0	13	20.7
4.	M_d	89.9	11	−3.4
5.	M_e	45.4	10	−39.4
6.	M_f	40.8	9	−35.5
7.	$M_e - \{\gamma_{31}\}$	45.4^t	11	−47.9
8.	$M_e - \{\gamma_{51}\}$	45.4^t	11	−47.9
9.	$M_e - \{\beta_{52}\}$	45.4^t	11	−47.9
10.	$M_e - \{\theta^{\epsilon}_{47}\}$	43.7^t	11	−49.6
11.	$M_e - \{\gamma_{31}, \gamma_{51}, \beta_{52}, \theta^{\epsilon}_{47}\}$	46.9	14	−71.8
12.	11. $-\{\beta_{43}\}$	47.8	15	−79.4
13.	11. $-\{\psi_{54}\}$	60.4	15	−66.8
14.	11. $-\{\beta_{43}, \psi_{54}\}$	60.6	16	−75.1
15.	11. with $\psi_{23} = \psi_{54}$	48.5	15	−78.7
16.	15. $-\{\gamma_{21}\}$	49.5^t	16	−86.2
17.	15. $-\{\gamma_{41}\}$	52.5^t	16	−83.2
18.	15. $-\{\beta_{43}\}$	49.1^t	16	−86.6
19.	15. $-\{\gamma_{21}, \gamma_{41}\}$	55.0^M	17	−89.2
20.	15. $-\{\gamma_{21}, \beta_{43}\}$	50.2^M	17	−94.0
21.	15. $-\{\gamma_{41}, \beta_{43}\}$	53.3^M	17	−90.9
22.	15. $-\{\gamma_{21}, \gamma_{41}, \beta_{43}\}$	55.9	18	−96.8

NOTE: The first six models are those fit in Table 4.2 of Long (1983b). The "$-\{\}$" notation indicates that the parameters inside the curly brackets have been constrained to equal zero. For example, model 11 is the same as model M_e with the additional constraints that γ_{31}, γ_{51}, β_{52}, and θ^{ϵ}_{47} are equal to zero. In the L^2 column, a superscript "t" indicates that the quantity shown is an approximation equal to the L^2 value for a larger model plus the square of the relevant t-statistic, while the superscript "M" indicates that the quantity shown is an approximation equal to the L^2 value for a smaller model minus the relevant LM test statistic or modification index.

Figure 1: Model M_f for data from Wheaton (1978), and definitions of the main variables. This is reproduced from Long (1983b), in which it is Figure 4.1.

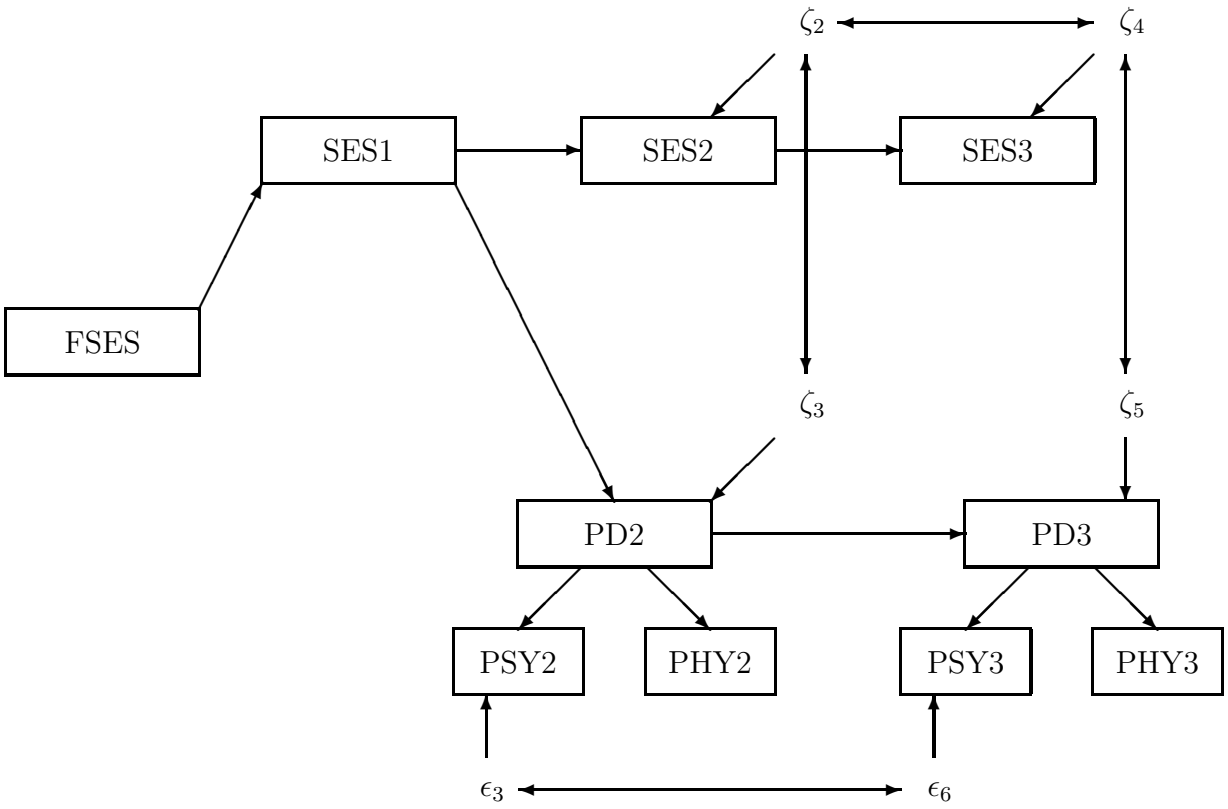

Figure 2: Model 22 from Table 2 for data from Wheaton (1978).