Bayesian Estimation and Testing of Structural Equation Models

by

R. Scheines, H. Hoijtink, and A. Boomsma

October 1995

Report CMU-PHIL-66

Philosophy
Methodology
Logic

Pittsburgh, Pennsylvania 15213-3890

Bayesian Estimation and Testing of Structural Equation Models

Richard Scheines, Herbert Hoijtink, and Anne Boomsma

Department of Philosophy

Carnegie Mellon University, USA

and

Department of Statistics and Measurement Theory

University of Groningen, The Netherlands

We thank David Spiegelhalter for suggesting applying the Gibbs sampler to structural equation models to the first author at a workshop in Wiesbaden. We thank Ulf Böckenholt, Marijtje van Duijn, Clark Glymour, Ivo Molenaar, Steve Klepper, and Tom Snijders for helpful discussions, mathematical advice, and critiques of earlier drafts of this paper.

The Gibbs sampler we report on in this paper is implemented in the TETRAD III program, which as of the date of submission is still in development. Information about TETRAD III or requests for reprints should be sent to Richard Scheines at the Dept. of Philosophy, Carnegie Mellon University, Pgh, PA, 15213.

Email: R.Scheines@andrew.cmu.edu.

Abstract

The Gibbs sampler can be used to obtain samples of arbitrary size from the posterior distribution over the parameters of a structural equation model given covariance data and a prior distribution over the parameters. Point estimates, standard deviations and interval estimates for the parameters can be computed from these samples. If the prior distribution over the parameters is uninformative, the posterior is proportional to the likelihood, and asymptotically the inferences based on the Gibbs sample are the same as those based on the maximum likelihood solution, e.g., output from LISREL or EQS. In small samples, however, the likelihood surface is not multivariate normal and in some cases not even unimodal. Nevertheless, the Gibbs sampler draws a sample from the true posterior distribution over the parameters regardless of the sample size and the shape of the likelihood surface. With an informative prior distribution over the parameters, it can be used to estimate underidentified models, as we illustrate on a simple errors-in-variables model.

Key Words: Structural equation modeling, Gibbs sampler, Bayesian inference, Posterior predictive p-values.

1. Introduction

This paper shows that when the Gibbs sampler is used for structural equation models (SEMs), two issues of practical interest can be addressed. First, posterior distributions over the parameters and the posterior predictive p-value of fit statistics can be approximated to arbitrary precision, even for small samples. Second, prior knowledge may be incorporated to a fuller extent within the Bayesian framework of inference than within the classical framework.

In the Bayesian approach, posterior distributions can be approximated by the Gibbs sampler or by normal distributions based on maximum likelihood estimates. In what follows we compare both statistical procedures, and evaluate their merits in structural equation modeling. As an introduction and for notation, we first briefly review ML-estimation and Bayesian statistical inference.

1.1 Maximum Likelihood Estimation

Let X = (xi, x_2, ..., x_N) be a set of N normally and independently distributed random variables x = (xi, x_2,..., x_p)^f, with expectation E{x} = (I and variance-covariance matrix £ = £(0_{pop})> where the elements of X depend on the values of t ≤ p(p+l)/2 unknown population parameters 6_{pop} = (61, 0_2, ..., 6_t)^f. 2(0_{pO}p) represents the structural equation model (SEM) in the population. Under the assumption that the probability density of x is p-variate normal, i.e., x ~ N_p{|i,£(0_{pO}p)}, maximum likelihood estimates 0ML of the unknown parameter vector G_{pop} can be obtained. For structural equation models 0_ML can be calculated using programs like LISREL (Jøreskog & Sørbom, 1993b) and EQS (Bentler, 1989).

Without loss of generality it is assumed that there is no interest in estimating first order moments. Then, for estimation purposes the sample covariance matrix S (p x p), where S is an unbiased estimate of £ based on a sample of observations X (N x p), is a sufficient statistic.

Let $p(\mathbf{X}|\theta)$ denote the joint probability density function of $\mathbf{X}$. If $p(\mathbf{X}|\theta)$ is regarded as a function of θ, given the observations $\mathbf{X}$, it is called the likelihood function of θ given $\mathbf{X}$, i.e.,

$$L(\theta|\mathbf{X}) = p(\mathbf{X}|\theta) \ . \tag{1}$$

To be more specific, for a given $\mathbf{X}$ the likelihood function is by definition any function of θ proportional to $p(\mathbf{X}|\theta)$; it is thus defined up to a multiplicative constant. Given the sample covariance matrix $\mathbf{S}$ the likelihood can be expressed as (cf. Anderson, 1958, p. 157)

$$L(\theta|\mathbf{S}) = |\Sigma(\theta)|^{-(N-1)/2} \exp[\{-(N-1)/2\} \ tr[\mathbf{S}\{\Sigma(\theta)\}^{-1}] \ , \tag{2}$$

and thus it follows that the log-likelihood

$$\log L(\theta|\mathbf{S}) = -\{(N-1)/2\} \ \{\log|\Sigma(\theta)| + tr[\mathbf{S}\{\Sigma(\theta)\}^{-1}]\} \ . \tag{3}$$

Standard ML-estimation of a structural equation model by the LISREL program, for example, uses an iterative Davidon-Fletcher-Powell (DFP) algorithm which minimizes a function of the log-likelihood:

$$F_{ML}[S;\Sigma(\theta)] = \log|\Sigma(\theta)| + tr[\mathbf{S}\{\Sigma(\theta)\}^{-1}] - \log|\mathbf{S}| - p \ , \tag{4}$$

where p is the number of observed variables.

It is well established theoretically that asymptotically, as N goes to infinity, the sampling distribution of θ_{ML} is $N_t(\theta_{pop}, J^{-1}(\theta))$, where $J(\theta)$ is the expected Fisher information matrix (cf. Tanner, 1993, p. 16). This implies that the marginal sampling distribution of a single parameter estimate θ_{ML} is asymptotically $N(\theta_{pop}, AVAR\{\theta_{ML}\})$, where AVAR denotes the diagonal element of $J^{-1}(\theta)$ corresponding to the parameter at hand (cf. Bollen, 1989, p. 468f.).

Thus, because of the asymptotic normality of 0ML, sampling theoretical statistical inferences can be made with respect to individual unknown model parameters $0po_P$. Also, the sampling distribution of the so called $\%^2$ goodness-of-fit statistic, $(N-1)F_ML[S;£(0)]$, which is used for testing the hypothesized model against a saturated model, asymptotically has a central chi-squared distribution with p(p+l)/2 - t degrees of freedom, given that the model holds.

Such an approach reflects a frequentist point of view: $0po_P$ is considered fixed, but 0ML is a vector of random variables. The probabilities involved refer to the frequency with which different values of parameter estimates (arising from sets of data *other* than those which are actually observed) could* occur for some *fixed but unknown* value of a population parameter $6po_P$ (cf. Box & Tiao, 1973, p. 72). In this paper, however, the emphasis will be on a Bayesian approach.

1.2BayesianStatisticalInference

In a Bayesian framework statistical inferences are associated with *different* values of parameters which could have given rise to the *fixed* set of data which has actually been observed (cf. Box & Tiao, 1973, p. 72). The main interest is in the posterior density of the vector of random variables 0 given the sample data X, which for continuous variables is defined as

$$p(0|X) = p(X|0)\ p(0) / J\ p(X|0)\ p(0)\ d0$$

$$-\ p(X|0)\ p(0)\ . \qquad (5)$$

Here p(0) is the prior distribution over 0, expressing what is known about 0 before any knowledge of X. In constrast, the posterior distribution p(0|X) expresses the result of changing p(0) to take sample data X into account. Given that L(0|X) = p(X|0), see (I), it follows that (5) can be expressed as

$$p(\theta|\mathbf{X}) \propto L(\theta|\mathbf{X})\, p(\theta) \ , \qquad (6)$$

showing that $p(\theta|\mathbf{X})$ is proportional to the product of the likelihood $L(\theta|\mathbf{X})$ and the prior $p(\theta)$. As emphasized by Box and Tiao (1973, p. 11), the likelihood function plays an important role here: 'It is the function through which the data **X** modifies prior knowledge of θ; it can therefore be regarded as representing information about θ coming from the data.'

Depending on the amount of prior knowledge relative to the information in the sample, the posterior distribution can be dominated by the likelihood or by the prior. If an uninformative ('improper') prior $p(\theta) = c$ is used, where c is a real constant, the posterior distribution $p(\theta|\mathbf{X})$ is proportional to the likelihood function, i.e.,

$$p(\theta|\mathbf{X}) \propto L(\theta|\mathbf{X}) \ , \qquad (7)$$

and thus, apart from a normalizing constant, the posterior distribution function is equal to the likelihood function. Also, the marginal posterior distribution functions would be proportional to the marginal likelihood functions: $p(\theta|\mathbf{X}) \propto L(\theta|\mathbf{X})$.

If, on the other hand, an informative prior distribution is used, and in this paper it is assumed throughout that in such a case $p(\theta)$ has a multivariate normal distribution $N_t(\mu_0,\Sigma_0)$, then when N is small enough for the prior to make more than a negligible contribution to the posterior the marginal posterior distribution functions $p(\theta|\mathbf{X})$ are not proportional to the corresponding marginal likelihood functions $L(\theta|\mathbf{X})$; (6) holds, not (7).

It is also well established theoretically that asymptotically the posterior density $p(\theta|\mathbf{X})$ converges to normality. That is, the posterior density (likelihood) is proportional to the multivariate normal density $N(\theta_{ML}, I^{-1}(\theta|\mathbf{X}))$, where $I(\theta|\mathbf{X})$ is the observed Fisher information matrix (cf. Tanner, 1993, p. 16).

/.3 *Finite Sample Size and Prior Knowledge*

The ML-estimation theory used in structural equation modeling is asymptotic theory. The same holds for generalized least squares (GLS), and weighted least squares or asymptotic distribution free estimation (WLS or ADF). Thus, for making proper statistical inferences the sample size N must be large.

Several robustness studies show that sample size matters for the behaviour of these estimation methods for SEM. See, for example, Bearden, Sharma and Teel (1982), Boomsma (1982, 1983), Baldwin (1986), Chou, Bentler and Satorra (1991), Hu, Bentler and Kano (1992), and Yung and Bentler (1994). From such research it may roughly be concluded (neglecting the effect of model complexity and other interacting factors) that the behaviour of ML and GLS is.fairly robust if the sample size N is as large as 300, say, but definitely not so if N < 50, say. On the other hand, ADF estimation requires huge sample sizes to obtain proper parameter estimates.

The variances and covariances of parameter estimates are also often incorrectly estimated in small sample studies. Structural equation modeling programs like LISREL and EQS use a sample estimate of the asymptotic variances AVAR(GML), which may differ substantially from the true variance of $0_M L$ given a small sample (Boomsma, 1983). As a consequence, if the sample size is small the sampling distribution of the (standardized) parameter estimates is unknown, and often cannot be estimated well by applying formulas based on asymptotic theory.

Further, the distribution of likelihood-ratio fit statistics cannot be trusted in small samples. For almost any sample size, the distribution of each of the numerous fit indices currently available is almost completely unknown. See Jöreskog and Sörbom (1993a), or Hu and Bentler (1995), for an overview.

In short, asymptotic estimation theory, like ML-estimation, is inappropriate in structural equation modeling when the sample size is small. A solution to this problem is to work with posterior density functions p(6|X), which can be numerically approximated to arbitrary precision at any sample size with Markov Chain Monte Carlo methods, and

in particular with the Gibbs sampler. This procedure will be compared with the normal approximation to the posterior obtained from ML-estimation. We now give a brief outline of the Gibbs sampler and the comparison to be made.

1.4 The Gibbs sampler and ML-approximations

With the Gibbs sampler (Geman & Geman, 1984), joint and marginal posterior distributions, $p(\theta|\mathbf{X})$ and $p(\theta|\mathbf{X})$, can be approximated to arbitrary precision for any sample size N without having to know a closed form expression of the posterior distribution. All that is required is a closed form expression for the prior distribution $p(\theta)$ and for the likelihood of each parameter θ_i conditional on the other parameters and the data, i.e., $L(\theta_i|\theta_1,...,\theta_{i-1},\theta_{i+1},...,\theta_t,\mathbf{S})$. Casella and George (1992) give a good introduction to the Gibbs sampler; more technical treatments can be found in Tanner (1993), and in Volume 55 (No. 1) of the Journal of the Royal Statistical Society, Series B, for example.

The Gibbs sampler can be used either by the Bayesian or the classical, sampling theory oriented statistician. For the Bayesian, the Gibbs sampler is mainly used to approximate marginal posterior distributions, whereas for the classical statistician it is mainly used to approximate (marginal) likelihood functions (cf. Casella & George, 1992, p. 173).

In the posterior distribution $p(\theta|\mathbf{X})$, two statistics of interest are the mean, denoted as θ_{EAP}, where EAP means expected a posteriori, and the standard deviation of $(\theta|\mathbf{X})$ around θ_{EAP}, denoted as $SD(\theta_{EAP})$. Knowledge of the posterior marginal densities also allows the inspection of order statistics θ_α, defined such that $p(\theta \leq \theta_\alpha \mid \mathbf{S}) = \alpha$. The median, for example (denoted θ_{MDAP}), is $\theta_{.50}$. These order statistics allow the inspection of central regions of a specific size, say 90%, e.g., regions from $\theta_{.05}$ to $\theta_{.95}$. Also, knowing the posterior density allows the inspection of the fit of the model by procedures using posterior predictive p-values (cf. Section 2.2; see also Gelman, Meng & Stern, 1994; Rubin & Stern, 1994).

The limiting normal approximation of the likelihood [i.e., the approximation of $L(\theta|\mathbf{X})$ by $N_t(\theta_{ML}, I^{-1}(\theta|\mathbf{X}))$, see Section 1.2] can be viewed as a normal approximation of $p(\theta|\mathbf{X})$. If the sample size is large, this is reasonable even with an informative prior distribution, because, 'as $N \rightarrow \infty$, the likelihood dominates the prior distribution, so we could just use the likelihood alone to obtain the mode and curvature for the normal approximation.' (Gelman, et al., 1995, p. 92)

As the sample size N increases, θ_{ML} converges to the mode of the marginal posterior density, and the estimated asymptotic standard error of θ_{ML}, denoted as $ASD(\theta_{ML})$, converges to the standard deviation of θ in the posterior density.

Thus in large samples we expect the Gibbs sampler and the familiar ML normal theory approximation of the posterior density (likelihood) to produce almost exactly the same numerical quantities for corresponding statistics, even though their interpretation is different (cf. Box & Tiao, 1973, Chapter 2). In examples considered below (Section 3), we compare the two approaches, not only on large samples with an uninformative prior, but also on small samples with and without an informative prior. Asymptotic theory tells us to expect the same results for large N, but we expect the results to diverge as the sample size decreases. Since the Gibbs sampler approximates the true posterior at any sample size, we expect results based on its approximation of the posterior to be better than those obtained from an approximation of the posterior based on ML-estimation.

1.5 Outline of the Paper

In Section 2 the Gibbs sampler is explained. A detailed outline is given of how to obtain a sample from the joint and marginal posterior distributions $p(\theta|\mathbf{X})$ and $p(\theta|\mathbf{X})$. Statistics of interest and a model check using posterior predictive p-values are also discussed. Section 3 presents several examples in which we use Gibbs sampling techniques on structural equation models.

In Sections 3.1 and 3.3, parameter estimates for Wheaton, et al.'s (1977) Stability of Alienation model computed by the Gibbs sampler are compared with maximum likelihood estimates computed by LISREL 8. The two statistical approaches are compared both, for large (Section 3.1) and for small sample size (Section 3.3). The results show an interesting characteristic of the posterior densities p(6|X), or the likelihood functions L(G|X): the absence of unimodality. In Section 3.2 this multimodality is discussed and illustrated at some length. The uncertainties that remained from previous discussions around this theme (see for example Rubin & Thayer, 1982, 1983; Bentler & Tanaka, 1983) seem to be resolved.

In Section 3.4 we explore one adwtage of using a probability measure to incorporate prior knowledge into parameter estimation. SEM programs like LISREL allow the user to incorporate prior knowledge, but only in the form of inequality constraints, linear and non-linear constraints on parameters, and interval restrictions on their values. While these sorts of constraints can be incorporated directly into the prior distribution p(0) (cf. Section 2.1; Box and Tiao (1973, pp. 67-69); Smith and Roberts (1993, p. 12f.), so can probabilistic information about the parameters, which allows the Gibbs sampler to provide informative estimates for underidentified models. An example of such an approach is given in Section 3.4, where an underindentified errors-in-variables model is estimated using the Gibbs sampler.

Using the $\%^2$ goodness-of-fit test, Section 3.5 illustrates how the Gibbs sampler can be used to compute posterior predictive p-values (cf. Section 2.2). The final section of paper details some of the questions that the research we report on here suggests but does not answer.

2. The Gibbs Sampler

2.1 *Obtaining a Sample from* p(O|S)

The Gibbs sampler (Casella and George, 1992; Tanner, 1993; Smith and Roberts, 1993) is an iterative procedure that, after convergence, renders a dependent sample from p(0|S). In each iteration m=l,...,M, each of the model parameters is sampled from its posterior conditional on the current values of the other parameters, the inequality constraints appropriate for the parameter at hand, and the data S.

For m=0, initial values are assigned to each of the model parameters 0j, j=l,...,t. In TETRAD in, each parameter is giyeil an initial value by the user. If the prior is informative, the prior mean for each parameter is used as the starting value, i.e., 6° = |i<). Subsequently for m=l,...,M, and j=l,...,t, G^1 is sampled from

$$p(\theta_j \mid .) = p(\theta_j | \theta_1^m, \ldots, \theta_{j-1}^m, \theta_{j+1}^{m-1}, \ldots, \theta_t^{m-1}, LB_j, UB_j, S) \quad , \tag{8}$$

where LB and UB denote the lower and upper bound respectively that are appropriate for 0j. A few examples: if a parameter is a variance, the lower bound is zero and the upper bound is °°. If a researcher decides that parameter 2 has to be larger than parameter 1 and smaller than parameter 3, the lower and the upper bound are G_1^and $G_3^{TM''1}$ respectively. If a parameter is unconstrained, the lower and upper bounds are -°° and °°, respectively.

A sample from (8) can be obtained using a combination of inverse probability sampling and rejection sampling (Gelman, Carlin, Stern, and Rubin, 1995, Chapter 10). The resulting procedure is summarized in steps a) through f) that follow. Steps a) and b) describe how to obtain values for the parameters of a normal approximation of (8). Steps c) through e) describe how inverse probability sampling may be used to obtain a random draw from the normal approximation of (8) truncated below and above by LB and UB respectively. Finally, in step f) rejection sampling is used to decide whether the random

draw obtained is accepted or rejected: the closer the approximate and the true density of the random draw, the larger the probability that it will be accepted.

a) Find the value MAX of θ_j that maximizes (8), and compute its asymptotic variance (AVAR) as 1/I(MAX), where I(MAX) denotes the observed Fisher information for (8) evaluated at θ_j = MAX.

b) Use MAX and AVAR, multiplied with a factor D, as parameters in a normal approximation of $p(\theta_j|.)$, denoted by $p^*(\theta_j|.)$. The variance of the normal approximation is multiplied by a factor D, to make (almost) sure that the left hand side of (12) is never larger than 1.0 (the upperbound of a uniform deviate). Experience until now indicates that for $D \geq 2$ the left hand side of (12) is rarely larger than 1.0.

c) Compute the probabilities that $\theta_j \leq LB_j$ and $\theta_j \geq UB_j$ from the normal approximations:

$$\alpha_{LB} = \int_{-\infty}^{LB_j} p(\theta_j|.)\, d\theta_j \ , \tag{9}$$

and

$$\alpha_{UB} = \int_{UB_j}^{\infty} p(\theta_j|.)\, d\theta_j \ , \tag{10}$$

d) Generate a uniform random deviate u[0,1].

e) Use inverse probability sampling to obtain a draw from the admitted region of the normal approximation, i.e., compute θ_j^m such that

$$u\,(1 - \alpha_{LB} - \alpha_{UB}) = \int_{-\infty}^{\theta_j^m} p^*(\theta_j|.)\, d\theta_j \ . \tag{11}$$

f) Use rejection sampling to decide whether the draw obtained in step e) will be accepted, or if step e) has to be repeated. I.e., generate a uniform random deviate v[0,1], and if

$$[p(\theta_j^m \mid .) / p^*(\theta_j^m \mid .)]\ [p^*(MAX|.) / p(MAX|.)] > v\ , \qquad (12)$$

accept θ_j^m, else repeat step e).

2.2 Statistics for the Posterior

Using θ and θ' as generic symbols to represent any of the parameters in θ, the expected a posteriori estimates (θ_{EAP}) are approximated in the K remaining elements of the Gibbs sample (see Section 2.3, paragraph 1) by

$$\theta_{EAP} \approx \sum_{k=1}^{K} \theta^k / K\ . \qquad (13)$$

The median a posteriori (θ_{MDAP}) estimates of the model parameters are given by the 50-th percentile of the K sampled values of θ.

The elements from the posterior covariance matrix of the parameters centered around the expected a posteriori estimates ($COV(\theta, \theta')$) are given by

$$COV(\theta, \theta') \approx \sum_{k=1}^{K} (\theta^k - \theta_{EAP})(\theta'^k - \theta'_{EAP})/K\ . \qquad (14)$$

The $\approx$ in (13) and (14) reflects that the accuracy with which the summations in (13) and (14) approximate the corresponding integrals over uni- and bivariate marginals of (2) depends on K. The same holds for the accuracy with which the median a posteriori estimate approximates the 50-th percentile of the corresponding marginal of (2).

With a constant prior, the posterior density of θ is proportional to the likelihood of θ. This implies that the finite sample covariance matrix of θ_{ML} can be approximated by

(14), with 0_{EAP} replaced by 6ML. Maximum likelihood estimates of the model parameters based on multivariate normality can be computed using LISREL, for example.

The Gibbs sampling approach should not be used to compute the maximum likelihood estimate 0ML, however. Since the dimensionality of 0 is usually large, a sample of only K values of 0 provides too crude a grid in the parameter space to have any confidence that the sampled value with the highest likelihood is also the maximum likelihood.

2.3 Convergence and Autocorrelation

There is not yet a generally agreed upon method to decide whether the sequence generated by the Gibbs sampler has converged or not. See, for example, Gelman and Rubin (1992) and subsequent discussions. To avoid strong dependencies among subsequent draws from p(q|S) (see below), we will retain only every r-th iteration from the iterative sequence described above. These iterations will be indexed k=l,...,K. Inspection of the mean, median, standard deviation, and 5-th and 95-th percentile of the distribution of each parameter, across four sequences of K/4 iterations, will be used to determine whether the Gibbs sampler has converged (the resulting numbers are similar) or not (the resulting numbers are substantially dissimilar or are mildly dissimilar but show an increasing or decreasing trend). Convergence will be discussed for several of the examples presented in Section 3.

The Gibbs sampler does not render independent draws of 0 from its posterior density. It is clear from the iterative sequence that 0^m depends on $Q^{m\sim l}$, ...,0°, i.e., subsequent draws are dependent. If r is chosen such that the multiple correlation between each of the elements in 0^m and all of the elements in $0^{m\,r}$ is small, then the resulting sample will be (approximately) linearly independent. Note however, that this does not imply the absence of nonlinear dependencies. So far, the dependence structure for subsequent draws can only be determined exactly for very simple models. Structural

equation models do not belong to the class of simple models. As a consequence, the best safeguard against strong nonlinear dependencies is the use of a large value for r.

Experience up to now indicates that (for all practical purposes) linear and nonlinear dependencies may be ignored. Parameter estimates, standard deviations, covariance matrices, posterior predictive p-values are virtually the same whether computed from K iterations with lag r > 1, or K iterations with r = 1 (provided that K is a large number, and, provided that the iterative sequence has converged). An explanation for this feature may be the following. Due to the dependence of adjacent iterations, the Gibbs sampler 'over-samples[1] *certain* regions of the parameter space if the number of iterations is small, but it 'over-samples[1] *each* region of the parameter space if the number of iterations is large. If for each region of the parameter space, the degree of 'over-sampling[1] is proportional to the posterior density of that region, the result is virtually indistinguishable from an independent sample.

2.4 Goodness of Fit Statistics and Posterior Predictive p-values

In this section it will be explained how posterior predictive (or Bayesian) p-values (Gelman, Meng and Stern, 1996; Rubin, 1984; Rubin and Stern, 1994) can be used to evaluate any goodness-of-fit statistic that is a function of the model parameters and the observed data. As an example, we use the so called $\%^2$ goodness-of-fit statistic:

$$LR(S,Z(0)) = (N-1)\, F_{ML}[S;E(e)] \ , \qquad (15)$$

where the discrepancy function FML is defined as in (4).

Let S(6) denote a covariance matrix drawn pseudo-randomly (with appropriate N) from S(G).[1] The posterior predictive p-value of the $\%^2$ goodness-of-fit statistic can be written as:

---

[1] We use the Monte Carlo Generator in TETRAD II (Scheines, et, al., 1994, chp. 13) to draw pseudo-random samples from a given SEM.

$$\text{p-value} = \int_{\theta} \Pr\{LR(S,\Sigma(\theta)) < LR(S(\theta),\Sigma(\theta))\}\ p(\theta|S)\ d\theta\ ,$$

$$\approx \sum_{k=1}^{K} \Pr\{LR(S,\Sigma(\theta^k)) < LR(S(\theta^k),\Sigma(\theta^k))\} / K\ , \qquad (16)$$

$$\approx \sum_{k=1}^{K} \sum_{z=1}^{Z} I_z/ZK,$$

where, $I_z = 1$ if $LR(S,\Sigma(\theta^k)) < LR(S_z(\theta^k),\Sigma(\theta^k))$, and 0 otherwise. The integral in (16) is approximated using a summation over a grid of K values of θ sampled from $p(\theta|S)$. The Pr(.) in (16) is approximated by the probability observed in z=1,...,Z data matrices (with sample covariance matrices $S_z(\theta^K)$ simulated conditional on each value of θ^k.

The principle underlying (16) can be explained as follows. If $\Pr\{LR(S,\Sigma(\theta_{pop})) < LR(S(\theta_{pop}),\Sigma(\theta_{pop}))\}$ is smaller than say 5 percent, then we would decide that the model is specified incorrectly. The population values of the model parameters are never known, but their posterior density is. The Bayesian solution is to use the average of $\Pr\{LR(S,\Sigma(\theta)) < LR(S(\theta),\Sigma(\theta))\}$ using $p(\theta|S)$ as a weight function. If the posterior predictive p-value is small (say 0.05 or less), the observed values of the fit statistics are usually worse than the corresponding simulated values of the fit statistics. This implies that the model used does not provide an accurate description of the observed data set.

3. Examples

This section gives several examples in which the Gibbs sampler implemented in TETRAD III is used to sample from the posterior distribution over the parameters of a structural equation model. We begin by comparing LISREL's maximum likelihood estimates of the parameters of the Stability of Alienation model (Wheaton, et al., 1977) with estimates based on the Gibbs sample. We discuss the problems that arise for maximum likelihood estimation when the likelihood surface is multimodal and the sample

size is small, which we illustrate with the Alienation model and with a simpler, more analytically accessible model. We then illustrate how diffuse but not totally flat prior distributions over the parameters eliminate multimodality in the posterior. Next we use a simple errors-in-variables model to illustrate how the Bayesian approach to estimation can handle underidentified models. Finally, we illustrate posterior predictive p-values.

3.1 The Stability of Alienation: Large Sample

Consider a longitudinal structural equation model developed by Wheaton, et al., (1977) to investigate the stability of social alienation (Figure 1).

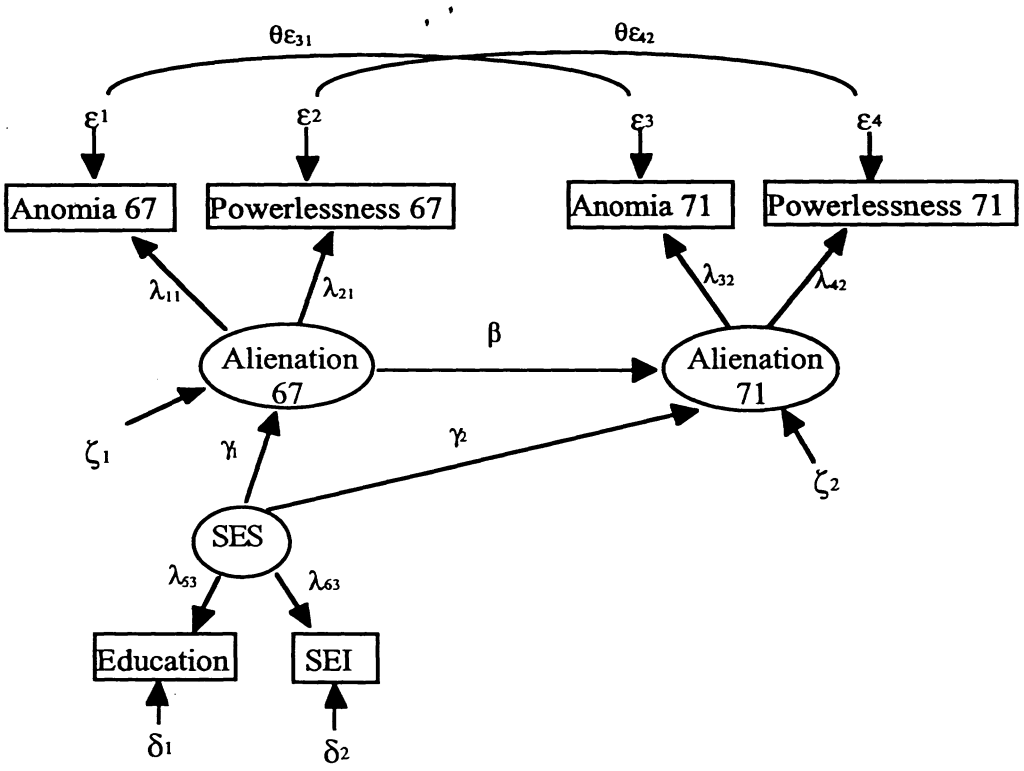

Figure 1. The Stability of Alienation model.

The purpose of this study was to estimate the effect that a given level of social alienation in 1967 (Alienation 67) had on the level of social alienation in 1971 (Alienation 71), controlling for socioeconomic status (SES). Thus measurement models were

constructed for the latent variables, and the central purpose of the study was to estimate the parameter p.

In practice the population parameters of the Alienation model are unknown. Since our purpose is methodological, however, it is preferable to work with a model in which the population is known. We thus varied the sample covariance matrix S reported in Wheaton, et al.'s paper slightly such that S = £(0_{pO}p) = £(9ML)- We then did a series of experiments in which we varied the sample size but not the elements of S. The population parameters are given in Table 1 and the covariance matrix S = 2XG$_{pO}$p) = £(GML) i^{n} Table 2. In the first experiment we used a very large sample size (N = 20,000), and compared the results of LISREL 8 (Jöreskog and Sörbom, 1993) and the Gibbs sampler in TETRAD III.

Parameter	Value	Parameter	Value	Parameter	Value
$0e_{u}$	4.730	$0e_{42}$	0.340	V11	4.850
$6e_{22}$	2.570	Xn	1.000	V22	4.090
$0e_{33}$	4.400	X21	0.980	ϕ	6.810
ee_{44}	3.070	^32	1.000	Yi	-0.570
es_{n}	2.800	*42	0.920	Y2	-0.230
05_{22}	2.649	^53	1.000	p	0.610
$0e_{31}$	1.620	X«	0.522		

Table 1. Popluation parameters for the Stability of Alienation model.

In all of the estimation experiments involving this model, the latent error variances (Yii>Y22>(!)) were fixed at their population values, one factor loading for each latent (A,$_{n}$,

A,32, ^53) was constrained to be strictly positive, and the remaining parameters were given population starting values.

Anomia 67	11.7926					
Powerlessness 67	6.9213	9.3529				
Anomia 71	6.8209	5.0969	12.5674			
Powerless 71	4.7849	5.0292	7.5140	9.9829		
Education	-3.8817	-3.8041	-3.9341	-3.6194	9.6100	
SH	-2.0262	-1.9857	-2.0536	-1.8893	3.5548	4.5045

Table 2. S = Z(e_{pop}) = *Z(QML)* for the Stability of Alienation model.

TETRAD III produced a Gibbs sample of size M=10,000 from p(6|S) that, although it is autocorrelated (the multiple correlation coefficients, as decribed in Section 2.3 but only for autocorrelation with the other structural parameters, were 0.7, 0.5, and 0.5 for (3, Yi, and y_2 respectively), rendered the same results as a normal approximation to the posterior using 6_ML and $AVAR(0_ML)$ as obstained by LISREL 8. Figure 2 shows a histogram with an accompanying normal curve of the 10,000 values of the structural parameter p that the Gibbs procedure sampled from p(0|S).

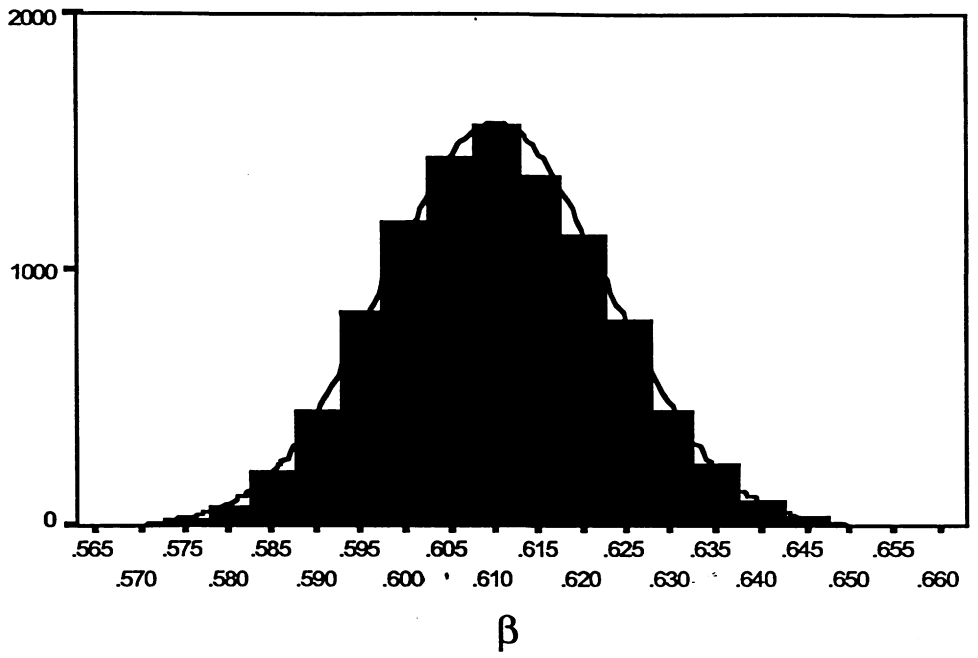

Figure 2. Frequency distribution for the Gibbs sample from the posterior marginal over β. M=10,000, N=20,000, and β_{pop} = 0.61.

To eliminate the autocorrelation in the 10,000 original draws, we kept every 10th iteration to produce 1,000 final draws (K=1,000). The multiple correlation coefficients in the resulting sub-sample were 0.11, 0.12, and 0.06, for β, γ_1, and γ_2 respectively.

To confirm that the final 1,000 iterations are from a sequence that has converged, we calculated θ_{EAP} and $SD(\theta_{EAP})$ for each of the model's three structural parameters (β, γ_1, and γ_2) within each of four blocks (250 iterations each). The results are given in Table 3.

	β		γ_1		γ_2	
Block	θ_{EAP}	$SD(\theta_{EAP})$	θ_{EAP}	$SD(\theta_{EAP})$	θ_{EAP}	$SD(\theta_{EAP})$
1	0.608	0.012	-0.571	0.011	-0.230	0.012
2	0.610	0.011	-0.570	0.010	-0.230	0.011
3	0.609	0.013	-0.569	0.011	-0.230	0.011
4	0.611	0.012	-0.571	0.011	-0.229	0.010

Table 3. Convergence analysis for the structural parameters in the Stability of Alienation model. M=10,000 and N=20,000.

Since θ_{EAP} and $SD(\theta_{EAP})$ vary by at most 0.003 between any two blocks, we conclude that the sequence has converged. Table 4 gives point estimates θ_{EAP}, θ_{MDAP} and the measure of spread $SD(\theta_{EAP})$ for the final sample of 1,000 draws, and compares them to the corresponding output from ML estimation as computed by LISREL 8. It is worth noting that eliminating the autocorrelation in the Gibbs sample had no effect on these numbers, i.e., the same results held for all 10,000 iterations and the retained 1,000.

	θ_{pop}	θ_{EAP}	θ_{MDAP}	θ_{ML}	$SD(\theta_{EAP})$	$ASD(\theta_{ML})$
γ_1	-0.570	-0.570	-0.570	-0.570	0.011	0.011
γ_2	-0.230	-0.230	-0.230	-0.230	0.011	0.011
β	0.610	0.610	0.610	0.610	0.012	0.013

Table 4. Final Gibbs estimates vs. ML estimates for the Stability of Alienation Model. K=1,000 and N=20,000.

$ASD(\theta_{ML})$, which is calculated by LISREL 8 and other SEM software packages, assumes that the likelihood is normal and is thus just an asymptotic approximation of the

standard deviation. For all practical purposes, however, when N=20,000 the asymptote has been reached and the approximation is exact. As Boomsma (1983) has shown on precisely this model, however, the approximation badly breaks down at small samples, e.g., N=50, and inferences based on ASD(6ML) can be wildly overconfident. With enough iterations, i.e., when K is large enough, the Gibbs sample will converge in distribution to the exact posterior density no matter what the sample size, and thus given convergence, SD(6EAP) is exact at any sample size.

3.2 The Likelihood Surface Can be Multimodal

Even with a flat prior, at small sample? the posterior is sometimes quite different than one would expect from asymptotic ML theory. To illustrate, we repeated the study above with N=50. What emerged was at first disturbing but eventually illuminating. The marginal posterior distribution of some of the parameters had more than one mode and were very diffuse relative to the asymptotic approximation obtained from the ML solution. In what follows we show that for certain SEMs, the likelihood surface is indeed multimodal, and because the problem is interesting in its own right, we pause to discuss it before examining the small sample results for the Alienation model.

Maximum likelihood estimators use an iterative search algorithm (cf. Section 1.1) in order to find the value of 0 where L(9|S) is maximal. With only one starting point for 6, such searches are reliable only if the surface of L(0|S) is unimodal with respect to 6. Although many authors have expressed concern that in certain cases the likelihood surface is not unimodal (Rubin & Thayer 1982, 1983; Bentler and Tanaka, 1983), as far as we know no one has shown a clear violation of unimodality or characterized conditions under which the surface is or is not unimodal.

The likelihood is usually written as a function of 6, but it can also be written as a function of what Lehmann (1959, p. 51) calls the "natural" parameters 0_{nat}- He proves that the natural parameter space is convex, and Brown (1986, Lemma 5.3, p. 146) proves

that the log-likelihood as a function of convex parameters is strictly concave. Thus the likelihood is unimodal in 0_{Mt}.

Unfortunately, the elements of 0_{Mt} are complicated (and totally *unnatural)* functions of the elements of 0, and it is prohibitively difficult to even write these functions out for models with more than three measured variables. In a simple factor model with two indicators (Figure 3) these functions are analytically accessible, however.

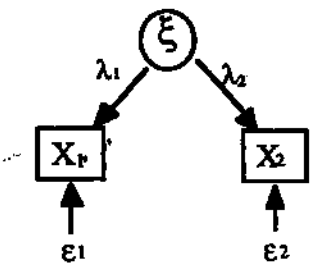

Figure 3. **Simple factor model with two indicators.**

Assuming that Cov($e_p e_2$) = Cov^e^) = Cov(e_p^) = 0, and fixing 6_u=Var(e_1) and X_f at 1.0, we arrive at a model with three free parameters: <0_{22}=Var(e_2), \$_2$, c()=Var(^)>, whose population values respectively are: <1.0, 1.4, 1.0>. Table 5 gives the population covariance matrix £(0^), both symbolically and numerically.

X_1	$\lambda_1^2\phi + \theta_{11} = 2.00$	
X_2	$\lambda_1\lambda_2\phi = 1.40$	$\lambda_2^2\phi + \theta_{22} = 2.96$

Table 5. Population covariance matrix for the simple factor model in Figure 3.

If the likelihood surface is unimodal given the covariances in Table 5, then the first order partial derivatives of the log-likelihood function for N=50 with respect to each of the free

parameters θ_j, $\frac{\partial \ln(L(\theta|X)}{\partial \theta_j}$ should vanish at only one point in the parameter space: θ_{pop} = $<\theta_{22}, \lambda_2, \phi> = <1.0, 1.4, 1.0>$. As it turns out, however, for this case they also vanish at $\theta_{alt} = <2.96, -1.0571, <10^{-16}>$. We obtained θ_{alt} by deriving the first order partial derivatives symbolically (called normal equations when they are set equal to 0), and then using Mathcad 4.0 (Mathsoft, 1993) to find two solutions to the normal equations: θ_{pop} and θ_{alt}. Whereas the likelihood is locally maximal at θ_{pop}, it is locally minimal at θ_{alt}. That is, $L(\theta_{alt}|S) \leq L(\theta_b|S)$ for any θ_b in the local neighborhood around θ_{alt}. We confirmed this with LISREL as follows. We gave LISREL the model in Figure 3 and the covariance matrix in Table 5, set N=50, and asked it to find θ_{ML} with two different sets of starting values:

$$\theta_{s1} = <2.96, -1.0569, 0.00001>$$

$$\theta_{s2} = <2.96, -1.0572, 0.00001>$$

Starting from θ_{s1}, LISREL found $\theta_{ML} = \theta_{pop}$, but starting from θ_{s2}, LISREL iterated away from θ_{pop} and in the end failed to converge.

Plotting the likelihood surface makes the multimodality vivid. Figure 4 shows a 3-dimensional plot of the likelihood surface against λ_2 and ϕ, with all other parameters fixed at their population values. The function is calculated and plotted at each point where the grid lines intersect, and extrapolated between such points. Although the values of λ_2 and ϕ proceed monotonically from back to front and left to right, and thus the topology is accurate, in order to provide an informative visualization of the surface we had to vary the scale substantially at different parts of the grid. That is, the numerical interval between plotted values of λ_2 and ϕ is not constant.

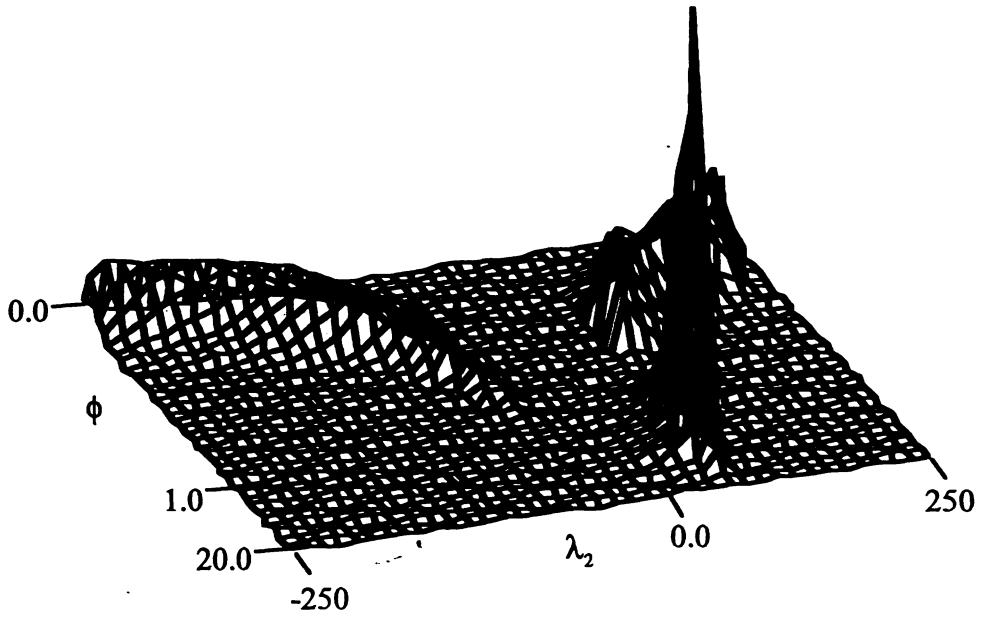

Figure 4. The likelihood surface for the model in Figure 3 and covariance matrix in Table 5 plotted against λ_2 and ϕ.

The spike in the middle is over the population values of λ_2 and ϕ, the ridge running back and to the left corresponds to values of ϕ approaching 0 and values of λ_2 getting more and more negative, and the ridge running back and to the right correspond to values of ϕ approaching 0 and values of λ_2 getting more and more positive. In fact the left and right hand ridges never peak, but continue to rise gently until they meet the edge of the parameter space where ϕ hits 0. This is also evident from LISREL 8, which converges to the population mode when given starting values for the parameters near the large central mountain, e.g., $\lambda_2 = 3.0$ and $\phi = 0.5$. However, if the parameters are started solidly within the left hand ridge, e.g., $\lambda_2 = -150$ and $\phi = 0.01$, then LISREL lowers ϕ and λ_2 in each iteration until it hits the iteration limit and then reports convergence failure. We conjecture that at least some convergence problems in SEM programs are the result of starting the

iterative estimation procedure in a region of the parameter space in which the likelihood surface slopes up towards a mode at a boundary of the parameter space.

The Gibbs sampler shows how the multimodality in the joint likelihood impacts the posterior marginal distributions. We ran TETRAD **III** for 50,000 iterations on the covariance matrix in Table 5 with N=50, a flat prior, and population starting values for all the parameters. We kept every **25th iteration** and on the **remaining 2,000** draws performed the same analysis **for convergence and autocorrelation that we described** in **the** previous examples. The **multiple correlation coefficients were all below 0.1. They** were 0.071, 0.096, and **0.088., for G_{22}, ϕ, and X_2 respectively. Table 6 and Table 7 show** the convergence results. For each of foui; blocks with 500 draws each, Table 6 shows the point estimates GEAP and 0MDAP> and Table 7 the measures of spread SD(0EAP), Q_{O5}, and 0.95, where 6_a is defined such that $p(0 \leq 0_a \mid S) = a$.

	0_{22}		ϕ		λ_2	
Block	**0EAP**	**0MDAP**	**0EAP**	**®MDAP**	@EAP	**0MDAP**
1	1.534	1.248	0.880	0.921	0.437	1.381
2	1.495	1.270	0.867	0.917	0.282	1.341
3	1.520	1.235	0.829	0.894	-0.991	1.385
4	1.463	1.238	0.865	0.906	0.329	1.366

Table 6. Convergence results for point estimates of the parameters of the model in Figure 3-

0MDAP is in general less sensitive to outliers, multimodality, and violations of normality than is 0EAP> and is thus a more robust but less critical convergence criterion than ©EAP- The results in Table 6 confirm this. Whereas ^EAP* fc>r example, fluctuates fairly wildly from block to block, X^MDAP is quite stable. Similarly, Q_{Q5} and $0._{95}$ are less

sensitive to outliers, multimodality, and violations of normality than is SD(G_{EAP}), and thus comprise a more robust but again less critical convergence criterion than does SD(6_{EAP}). In Table 7, for example, SD(A_{7EAP}) fluctuates much more dramatically than does either X^.95 or X^.05.

	0_{22}			•			λ_2		
Block	**SD(6_{EAP})**	**0.05**	**0.95**	SD(8EAP)	**0.05**	**0.95**	SDCOEAP)	**0.05**	**0.95**
1	1.479	0.196	3.445	**0.570**	**0.005**	**1.805**	**17.97**	-3.87	9.31
2	1.317	0.174	3.204	0.589	0.003	1.777	41.56	-6.22	10.27
3	1.562	0.126	3.451	' 0.561	0.002	1.675	33.17	-7.97	10.42
4	1.367	0.182	3.162	0.561	0.004	1.759	21.59	-4.31	9.28

Table 7. Convergence results for measures of spread of the parameters of the model in Figure 3.

Since fluctuations in 9_{EAP} and in SD(G_{EAP}) are caused by outliers in one of the extreme modes, convergence would only be verified using these statistics if K is enormous. For our purposes, we consider the iterative sequence to be converged if 6MDAP> 0.05, and $9_{.95}$ (which are robust against outliers) are stable over the four blocks.

The histogram of ⁜ in the retained 2,000 draws (Figure 5) has a shape proportional to the marginal posterior p(<(>|S), and, because we used a flat prior, to the marginal likelihood L(<j)|S). This histogram is clearly not unimodal, with at least one mode near the population value of 1.0 and another against the edge of the parameter space at 0.

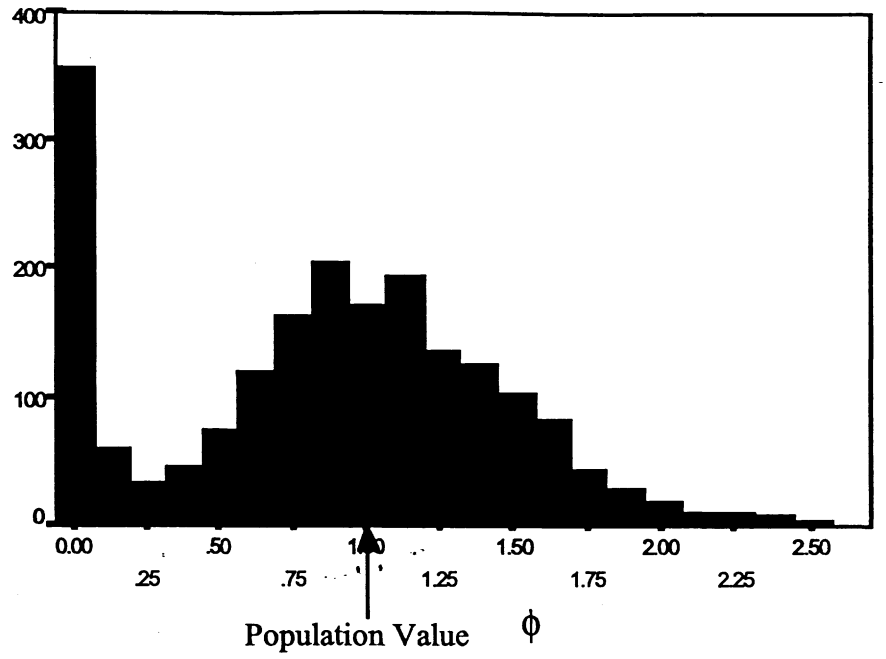

Figure 5. Histogram of the values of ϕ sampled from the posterior over the parameters in the model in Figure 3.

We suggest the following explanation. $L(\phi|S)$ is not unimodal because two elements in the implied covariance matrix (see Table 5) involve the product of ϕ and λ_2. When λ_2 has a high absolute value, e.g., -100, ϕ must be very close to 0 for $\lambda_2\phi$ and $\lambda_2^2\phi$ to remain low, which they must be in order for the implied covariances to be near those observed.

3.3 The Stability of Alienation: Small Sample Results

We now examine the Alienation case when the sample size is small. Treating again the covariance matrix in Table 2 as a sample, but with N=50, we ran the Gibbs sampler in TETRAD III for 100,000 iterations, keeping every 50th to end up with a final sample of 2,000 values from $p(\theta|S)$ that showed satisfactory convergence and autocorrelation. We also used LISREL 8 to compute a maximum likelihood estimate of the model's parameters from the same covariance matrix with N=50. The point of Wheaton, et al.'s study was to

estimate the stability of social alienation, which is the parameter β. We therefore focus our analysis on $p(\beta|S)$ and on LISREL's ML estimate β_{ML}. Table 8 shows the wild discrepancy between LISREL's results and those based on the final 2,000 values sampled from $p(\beta|S)$.

β_{ML}	β_{MDAP}	β_{EAP}	$ASD(\beta_{ML})$	$SD(\beta_{EAP})$	$\beta_{.025}$	$\beta_{.975}$
.610	1.439	-21.695	0.22	213.9	-499.8	499.7

Table 8. A comparison of the estimates of β in the Stability of Alienation model: Gibbs vs. Maximum Likelihood. N=50.

Inferences about β_{pop} supported by the two analyses are completely at odds. What is particularly striking is that $SD(\beta_{EAP})$ is approximately 1,000 times larger than $ASD(\beta_{ML})$. β_{ML} is almost three times as big as its standard error $ASD(\beta_{ML})$, and thus according to asymptotic maximum likelihood estimation theory we can reject the null hypothesis that β_{pop} is negative or 0 at a significance level of 0.05. Any sensible inference applied to $p(\beta|S)$ would conclude that from this data we know almost nothing about β_{pop}, let alone its sign.

The reason for the discrepancy in the two analyses is the multimodality of the likelihood function for the Alienation model. The Gibbs procedure samples from the entire posterior, and when there are many modes and the sample size is small, $ASD(\theta_{ML})$ is a poor approximation of the diffusion in the marginal posteriors. The histogram in Figure 6 suggests that $p(\beta|S)$ is tri-modal.

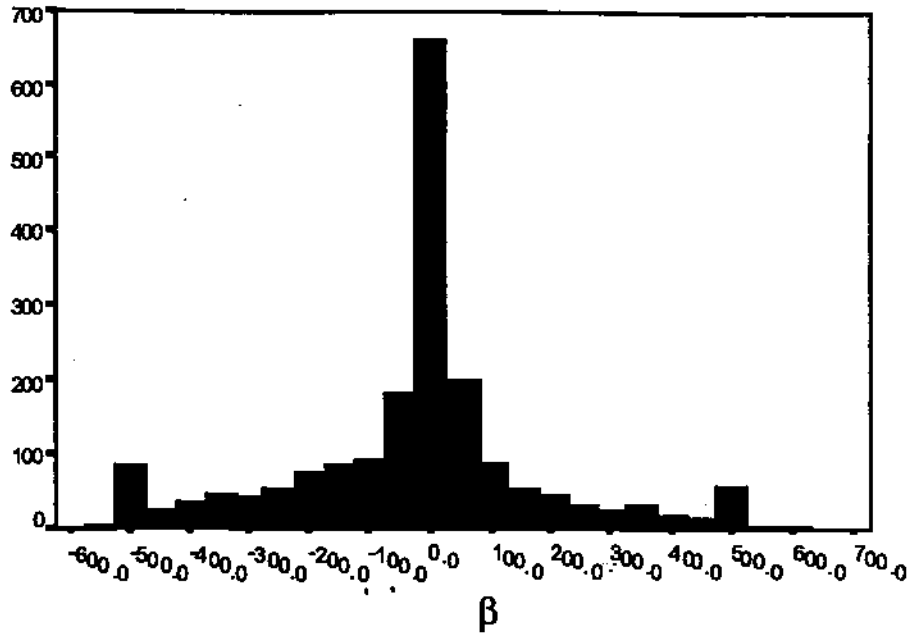

Figure 6. Histogram of the values of (3 sampled from the posterior over the parameters in the Stability of Alienation Model, N=50.

Which analysis is more reasonable? With 17 parameters to estimate from 50 observations, and absolutely no information about the ranges that the population parameters are likely to occupy, then it would be surprising if we *could* learn much of anything about the parameters.

In fact what makes the ML estimate of **p** "significant" in this case is a tacit use of prior information that in effect treats the likelihood function "as if' it were unimodal. It is of course desirable to rule out alternative modes for substantive reasons. Given Wheaton, et al's. design, for example, we simply would not believe that ($3_{pop} \sim -500$, which is where one of its modes lies in the Gibbs sample from p((3|S) (Figure 6). As it turns out, incorporating even a very loose prior over the parameters to rule-out nonsensical modes causes the posterior to collapse toward the ML solution. To us this indicates that whenever inferences are made about ML estimates on a problem in which the likelihood is multimodal, prior information about the parameters is involved, yet tacitly.

To illustrate, we ran the Gibbs sampler again on the Alienation model with the relatively diffuse normal prior over the model's parameters given in Table 9.

	μ_0	σ_0
All free error variances	**2.5**	**1.414**
All factor loadings	**1.0**	**4.000**
p	**0.5**	**4.000**
Yi	**-0.5**	**4.000**
γ_2	**-0.5**	**4.000**

Table 9. Prior pistribution over the parameters in the Stability of Alienation model.

The final sample from the posterior was assembled by keeping every 50th out of 100,000 iterations, leaving acceptable levels of autocorrelation and solid evidence of convergence. In Table 10 we list point estimates and the bounds of a 95% central interval around the point estimates for P, y_u and 72- In the LISREL solution, the point estimate is 0_{ML}, and the 2.5 percentile is 6_{ML} - 2 ASD(0_{ML}). The point estimates for the Gibbs results are the values of 6MDAP, and the 2.5 and 97.5 percentiles for the Gibbs sample are **0.025 and 6.975 respectively.**

		Point Estimate	2.5 percentile	97.5 percentile
$\beta_{pop} = 0.61$	LISREL	0.61	0.17	1.05
	Gibbs - Flat Prior	1.44	-499.78	499.69
	Gibbs - Loose Prior	0.62	0.12	1.36
$\gamma_{1pop} = -0.57$	LISREL	-0.57	-1.07	-0.07
	Gibbs - Flat Prior	0.67	-3.47	5.78
	Gibbs - Loose Prior	-0.57	-1.10	-0.18
$\gamma_{2\,pop} = -0.23$	LISREL	-0.23	-0.69	0.23
	Gibbs - Flat Prior	2.98	-405.50	462.50
	Gibbs - Loose Prior	-0.24	-0.82	0.30

Table 10. Alternative estimates for the structural parameters in the Stability of Alienation model, N=50.

The difference in the Gibbs results for a flat and loose prior are dramatic. For β, the size of the central 95% interval in the marginal posterior shrinks more than 800 fold from almost 1,000 to 1.24. This collapse in the posterior distribution is in part an effect of eliminating sampling from the alternative modes, because the posterior central 95% interval of 1.24 is much smaller than the central 95% interval in the prior (four times $\sigma_0(\beta) = 16.0$).

As the sample grows large, the alternative modes become small enough to ignore, so techniques which assume they do not exist like ML estimation are perfectly reasonable. At small N, however, it seems that they cannot be avoided, and the quantities calculated from an ML solution on the basis of asymptotic theory can be wildly off. On the other hand, when multimodality exists and the sample size is small enough for it to matter, then

even small amounts of prior knowledge can have a big effect on bringing the posterior back toward a solution consistent with the usual assumptions about unimodality. Inferences based on the ML solution seem to use such knowledge tacitly, and in practice it might well be reasonable to do so, but in our perspective it is always better to make assumptions explicit. In the Bayesian perspective, the sort of prior knowledge that serves to eliminate sampling from alternative modes is a step in that direction.

3.4 Estimating Underidentified Models

Substantial prior information about the parameters exists in many research contexts. The sign of a factor loading is often known, the results of previous research can provide precise prior information about parameter values and their standard errors, and in contexts of repeated measurements beliefs about parameter covariation can be warranted. If there is a lot of data at hand, e.g., N=1,000, a prior distribution has little or no effect on the posterior. If the sample size is small, however, e.g., N = 50, the prior can make a large difference.

An informative prior distribution over the parameters can also make it possible to estimate the parameters of an underidentified model. Virtually every introductory book on structural equation models routinely warns readers to ensure that all the parameters in their models are identifiable, i.e., uniquely determined from the measured data given the statistical assumptions and the discrepancy function being minimized. This is good practical advice, but since nature has no apparent reason to prefer systems whose models are identified, it is a maxim that has no obvious connection to the truth. Further, identification comes with a price: assumptions must be made which sometimes have little theoretical justification. Thus it is desirable to develop estimation techniques which explicitly incorporate uncertainty about identifying assumptions. Bayesian estimation is one such possibility.

A simple structural equation model involving two measured variables serves to illustrate the problem and a Bayesian solution to it. Suppose we wish to estimate the

influence of Lead Exposure (LE) on IQ-scores in some population of children. We might specify the following model, where X is a measure of LE, and all variables are expressed as deviations from their mean:

$$IQ = p\,x + e_{IQ} \tag{17}$$

The parameters of this model are p, Var(X), and Varteq). If routine statistical assumptions are satisfied, then these parameters are identified. They will only be *scientifically* informative, however, if other assumptions hold as well. The estimate of p reflects how unit changes in sample values of X inform us about changes in the expectation of IQ. What we hope scientifically is that this is *also* how a unit change in **LE *that we produced by an outside manipulation* would affect the expectation of IQ. To** move from the regression results to this sort of conclusion we need at least to assume that there are no unmeasured variables besides LE that confound the relation between X and IQ. Further, we must also assume that X perfectly measures Lead Exposure.

Although both of these assumptions seem highly unlikely to hold even approximately in the actual world, let us for the moment accept that no other factors besides LE confound the relation between X and IQ, and consider only the measurement error assumption. Since Lead Exposure is notoriously difficult to measure accurately, this assumption is questionable and a better model of what is going on is the standard errors-in-variables formulation in which LE is measured by X with error e_x:

$$IQ = pLE + e_{IQ}$$

$$\mathbf{X = LE + } e_x \tag{18}$$

Even with the usual assumptions, which now include Cov(LE,£$_{IQ}$) = Cov(LE,£x) = COV(£IQ,£X) = 0, this model is underidentified, i.e., for any implied covariance matrix £(90 that minimizes a discrepancy function, e.g., equation (4), of the implied and observed

covariances, there are an infinity of 0j^ such that £(6$) = £(0j). Several strategies have been explored for augmenting the errors-in-variables model to identify it, the most common being instrumental variables, but all require assumptions in addition to the ones already made for this model.

A better approach, in our view, was suggested by Klepper and Learner (1984). Although one cannot uniquely determine p even from population data on just IQ and X, they show that p can be bounded in the population just from assuming that all variances are strictly positive. Klepper (1988) has also shown how more restrictive bounds on Var(£x) in turn further narrow the bounds on p.

Klepper's strategy is really Bayesian in spirit. The more prior information you have about how much or how little measurement error you face, the tighter the bounds on the parameter of interest p. In this spirit, we used a prior distribution over Var(£x) to make the inference about p fully Bayesian. We used the following multivariate normal population model, which we chose for simplicity:

$$IQ = -0.657\,LE + 6IQ\,,$$

$$X = LE + \varepsilon_X\,,$$

$$Var(LE) = Var(£x) = Var(e_{IQ}) = 1.0\,, \qquad (19)$$

$$Cov(LE,6x) = Cov(e_{IQ},e_x) = Cov(ei_Q,£x) = 0\,.$$

For this model, $E(0) = \begin{bmatrix} 2.0 & 1 \\ -.657 & 1.4322 \end{bmatrix}$ Although there are four free parameters (Var(£x), Var(e_{IQ}), Var(LE), and P) let us assume that the only meaningful prior knowledge concerns the amount of measurement error for Lead Exposure, that is, the ratio of Var(e_x) to Var(LE). Without loss of generality we will suppose that our observed covariance matrix S = Z(0). The variance of X is 2.0, and since Var(X) = Var(£x) +

Var(LE), we can specify that half of X's variance is due to measurement error by assuming that $Var(\varepsilon_X) = 1.0$. Table 11 shows the multivariate normal prior (the variances are truncated below at 0) used for this example (Σ_o is diagonal).

	Prior	
Parameter	μ_o	σ_o
$Var(\varepsilon_X)$	1.0	0.1
$Var(\varepsilon_{IQ})$	1.0	4.0
Var(LE)	1.0	4.0
β	-1.0	4.0

Table 11. Prior distribution over the parameters of the model in equation 19.

By setting $\sigma_o(Var(\varepsilon_X))$ to 0.1 in the prior, $p(0.8 \leq Var(\varepsilon_X) \leq 1.2) = 0.95$, which corresponds to the belief that approximately between 40% and 60% of the variance in our measure of Lead Exposure is noise. We gave TETRAD III the covariance matrix S with N=100, and ran the Gibbs sampler for 100,000 iterations. The sequence of iterations converged quickly, and after keeping every 50th iteration there was almost no autocorrelation left in the sequence. The multiple correlation coefficients, again as described in section 2.3, are 0.025, 0.056, 0.022, and 0.045 for $Var(\varepsilon_X)$, $Var(\varepsilon_{IQ})$, Var(LE), and β respectively. Table 12 shows that the sequence has converged.

6_{EAP}

Block	Var(ex)	Varfog)	Var(LE)	p
1	**1.013**	**1.020**	1.033	**-0.687**
2	**1.007**	**1.043**	1.050	**-0.685**
3	**1.004**	**1.057**	1.053	**-0.672**
4	**1.012**	**1.018**	**1.026**	**-0.703**

$SD(6_{EAP})$

Block	Vai(ex)	$Var(e_{IQ})$	Var(LE)	p
1	**0.101**	0.227	0.314	0.235
2	0.099	0.218	0.319	0.226
3	0.099	0.230	0.311	0.226
4	**0.101**	0.229	0.311	0.228

Table 12. Convergence results for the Gibbs sample from the posterior over the parameters in the model in equation 19.

Figure 7 contains a histogram for the values of (3 in the final Gibbs sample of 2,000 draws. Since the distribution is not normal, we use the median for a point estimate and the central 95% interval for inference. In this case PMDAP = -0.660, and the 95% central interval of **p(P|S)** ranges from -1.090 to -0.384. Thus from this model, the data, and our prior, we can conclude that there is a substantial negative influence of Lead Exposure on IQ.

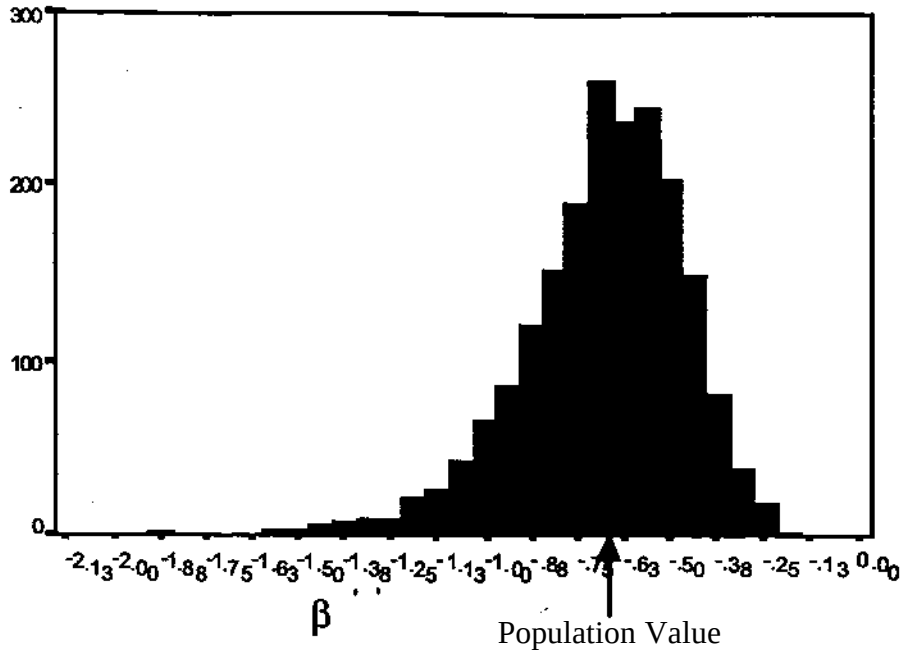

Figure *1.* Histogram of the values of (3 sampled from the posterior over the parameters in the model in equation 19.

The example shows how Bayesian parameter estimation can convert uncertain beliefs about incidental model parameters into useful posterior knowledge about parameters of interest. In this case probabilistic knowledge about the measurement error infecting X, our measure of Lead Exposure, was converted into uncertain but useful knowledge about the dependence of IQ on Lead Exposure, which is what the model is in service of estimating.

The model in equation 19 contains four independent free parameters to estimate from only three data points. Thus the degrees of freedom are -1, and a prior which is informative about at least one of the parameters serves to "identify" the model. Had we specified a still more complicated model (Figure 8) in which Intelligence was a latent variable measured by IQ-scores, then in order to estimate the parameters we would need a prior that was informative about at least two of the parameters.

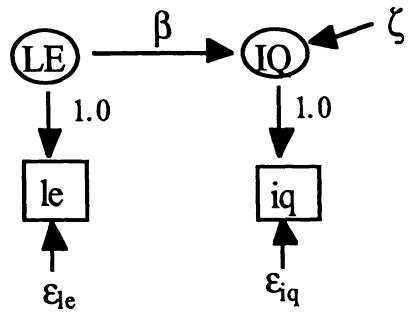

Figure 8. Full errors-in-all-variables model for the effect of Lead Exposure on Intelligence.

3.5 Posterior Predictive p-values for the χ^2 Goodness-of-fit Test

In Section 2.4, we explained how posterior predictive p-values can be computed. In this section we apply these ideas to the χ^2 goodness-of-fit statistic (15). This statistic is calculated by LISREL and other SEM programs and is used in what is called the χ^2 goodness-of-fit test, whose p-value is the probability that the value of the χ^2 goodness-of-fit statistic with appropriate degrees of freedom is as large or larger than the one observed. In general, a test is said to be exact if the probability of making a Type I error is exactly α (Good, 1994, p. 16). Since the χ^2 goodness-of-fit statistic is distributed as χ^2 only asymptotically, it is well known that the test based on it is not exact at small sample sizes, e.g., N=50.

For small samples, the p-value of the χ^2 goodness-of-fit statistic calculated on the basis of asymptotic theory is probably different from the posterior predictive p-value. The latter can be computed for any sample size, i.e., no reference to asymptotic properties is necessary for the computatioin of the posterior predictive p-value.

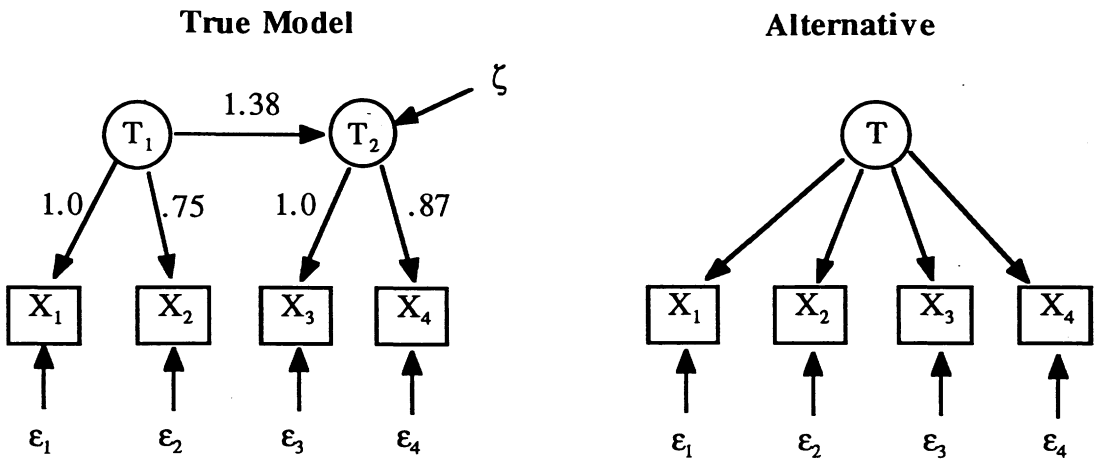

Figure 9. The two models used to illustrate posterior predicitve p-values.

We used TETRAD III to draw two pseudo-random samples from the true model in Figure 9, one with N=50 and one with N=500 (Table 13).

Population

X_1	2.0000			
X_2	0.7500	1.5625		
X_3	1.3800	1.0350	3.9044	
X_4	1.2006	0.9004	2.5268	3.1983

	N=50				N=500			
X_1	1.9084				2.0839			
X_2	0.6117	1.7885			0.7572	1.4982		
X_3	0.9807	1.1704	3.5978		1.5154	1.1129	4.3547	
X_4	0.9926	0.7173	2.3755	3.0669	1.2591	0.9668	2.9162	3.4920

Table 13. Covariance matrices for the posterior predictive p-value example.

To calculate p-values for the goodness-of-fit test on the basis of asymptotic theory, we ran LISREL on each model for both samples (Table 14).

		d.f.	χ^2	p-value
N=50	True Model	1	2.34	0.126
	Alternative	2	3.45	0.178
N=500	True Model	1	0.35	0.554
	Alternative	2	20.16	< 0.001

Table 14. LISREL goodness of fit results for the two models in Figure 9 on the two sample covariance matrices in Table 13.

Using a significance level $\alpha = 0.05$, the asymptotic test was able to separate the true model from the alternative at N=500, but could not do so at N=50. We then ran the Gibbs sampler in TETRAD III on each model on each sample with a flat prior. In each study we computed 50,000 iterations and kept every 50th to retain 1,000 draws. In each case convergence and autocorrelation were satisfactory. We computed the posterior predictive p-value for a model with $Z = 5$ (cf. Section 2.4). The first two columns of Table 15 show the results and repeat the p-values from LISREL for comparison.

		p-values		
		LISREL	Gibbs-Flat Prior	Gibbs-Loose Prior
N=50	True Model	0.126	0.030	0.355
	Alternative	0.178	0.025	0.047
N=500	True Model	0.554	0.545	
	Alternative	< 0.001	< 0.001	

Table 15: Comparative p-values for the two models in Figure 9 on the two sample covariance matrices in Table 13.

At N=500, LISREL and Gibbs with a flat prior yield nearly indistinguishable p-values, but at N=50 the results differ considerably. The posterior predictive p-values have descended into the rejection region (< 0.05) for both models. At N=50, the posterior predicitve p-values from the Gibbs sample produced with a flat prior are smaller than those computed on the basis of asymptotic theory by LISREL at least in part because the likelihood surface is multimodal for both models. At N=50 the alternative modes were visited with enough frequency to matter, whereas at N=500 the sampler spends almost all of its time within the mode that contains 0_ML- Iterations 0^k that are sampled from an alternative mode drive down the posterior predicitve p-value. That is, because the covariance matrices $L(6^k)$ are normally quite "distant" from S, PrlLRCS,!^)) < LR(S(9^k),E(e^k))} is very low.

Our experience so far, which is limited to a small number of experiments, indicates that when the likelihood is multimodal, the posterior predicitve p-value is typically low for very small sample sizes, even when the model is correctly specified and the sample S is close to £($0p_Op$). Thus, before further research, we are not confident that the Gibbs p-value as computed in (16) is an exact test in multimodal cases with a flat prior and small sample size.

Since using even a very loose prior effectively eliminates sampling from the alternative modes, in order to examine the effect of the multimodality on the p-values computed from the Gibbs sampler at N= 50, we ran TETRAD HI on both models with a loose prior in which (io(0) = 0ML, Cov_0(6i,9j) = 0 for i*j, and a_o(6i) = 6 for all i. Again we drew a sample of 50,000 from the posterior, and again kept every 50th draw. The results of the autocorrelation and convergence analyses were satisfactory. The computed posterior predictive p-values (the third column in Table 15) changed little for the alternative model (0.025 to 0.047), but rose dramatically for the true model (0.030 to 0.355), allowing us to reject the alternative and accept the true model with a = 0.05, contrary to either the asymptotic $\%^2$ goodness-of-fit test or the posterior predictive p-values from a Gibbs sample produced with a flat prior.

4. Discussion

A Bayesian approach to structural equation modeling has recently begun to draw attention. The focus, however, has been on finding better theoretical ways to handle uncertainty over a class of models rather than on handling uncertainty over the many possible values of the parameters in a single model. Raftery, (1993, 1994), for example, attempts to use Bayesian theory to solve problems concerning model comparison that present difficulties for the normal hypothesis testing regime. Geiger and Heckerman (1994) explore using Bayesian methods for automated model search. Raftery (1994) has also investigated incorporating model uncertainty into inferences about individual parameters.

Our research, however, uses Bayesian theory to estimate the parameters of a single SEM. Because the likelihood dominates the posterior at large sample sizes, the Bayesian and standard frequentist approaches diverge practically only when the sample size is small. Data sets in real research are of practical necessity often small, however, so this seems an important area to explore.

The prime benefits of the Bayesian approach seem to be an exact approximation of the posterior over the parameters and over fit statistics at any sample size, instead of an asymptotic approximation known to be off in the small sample, and the ability to estimate a wider class of models (underidentified) by incorporating even small amounts of prior knowledge.

In this paper we confined ourselves primarily to examples in which a single sample was analyzed. In future work we hope to explore the frequentist properties of the Bayesian approach to estimation and testing. That is, we intend to draw several samples with small N from $\Sigma(\theta_{pop})$, and estimate the model's parameters and test the model on each sample in order to determine the frequentist properties of the methods proposed in this paper.

References

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. New York: Wiley.

Baldwin, B.O. (1986). *The effects of structural model misspecification and sample size on the robustness of LISREL maximum likelihood parameter estimates*. Unpublished doctoral dissertation, Department of Administrative and Foundational Services, Lousiana State University.

Bearden, W.O., Sharma, S., & Teel, J.E. (1982). Sample size effects on chi-square and other statistics used in evaluating causal models. *Journal of Marketing Research, 19*, 425-430.

Bentler, P.M. (1989). *EQS: Structural equations program manual* (Version 3.0). Los Angeles, CA: BMDP Statistical Software.

Bentler, P.M., & Tanaka, J.S. (1983). Problems with EM algorithms for ML factor analysis. *Psychometrika, 48*, 247-251.

Bollen, K. (1989). *Structural equations with latent variables*. New York: Wiley.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K.G. Joreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction* (Part I, pp. 149-173). Amsterdam: North-Holland.

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and nonnormality.* Amsterdam: Sociometric Research Foundation, (doctoral dissertation, Rijksuniversiteit Groningen)

Box, G.E.P., & Tiao, G.C. (1973). *Bqyesian inference in statistical analysis.* Reading, MA: Addison-Wesley.

Brown, L.D. (1986). *Fundamentals of statistical exponential families with applications in statistical decision theory.* Hayward, CA: Institute of Mathematical Statistics.

Casella, G., & George, E.I. (1992). Explaining the Gibbs sampler. *The American Statistician,* 46, 167-174.

Chou, C.-P., Bentler, P.M., & Satona, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology* 44, 347-357.

Geiger, D., & Heckerman, D. (1994). Learning Gaussian Networks. Microsoft Technical Report, MSR-TR-94-lCf. Redmond, WA.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Gelman, A., & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science,* 7,457-511.

Gelman, A., Meng, X.-L., & Stern, H.S. (1993). *Bayesian model invalidation using tail area probabilities.* Unpublished manuscript, Department of Statistics, Harvard University.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions,and the Bayesian restoration of images. *IEEE Transactions on. Pattern Analysis and Machine Intelligence, 6, 721-741.*

Glymour, C, Schemes, R., Spirtes, P., & Kelly, K. (1987). *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling.* Orlando, FL: Academic Press.

Good, P. (1994). *Permutation Tests.* New York, Springer-Verlag

Hu, L.-T., & Bentler, P.M. (1995). Evaluating model fit. In R.H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Thousand Oaks, CA: Sage.

Hu, L.-T., Bentler, P.M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112,* 351-362.

Jöreskog, K.G., & Sörbom, D. (1993a). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Chicago, IL: Scientific Software International.

Jöreskog, K.G., & Sörbom, D. (1993b). *LISREL 8: User's reference guide.* Chicago, IL: Scientific Software International.

Klepper, S. (1988). Regressor diagnostics for the classical errors-in-variables model. *Journal of Econometrics,* 37, 225-250.

Klepper, S., & Learner, E. (1984). Consistent sets of estimates for regressions with errors in all variables. *Econometrica,* 52,163-183.

Lehmann, E.L. (1959). *Teting statistical hypotheses.* New York: Wiley.

MathSoft (1993). *Mathcad 4.0. User's guide.* Windows version. Cambridge, MA.

Raftery, A.E. (1993). Bayesian model selection in structural equation models. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 163-180). Newbury Paik, CA: Sage.

Raftery, A. E. (1994). Bayesian model selection in social research. Working Paper no. 94-12, Center for Studies in Demograph and Ecology, Univ. of Washington.

Rubin, D.B. (1984). Bayesian justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics,* 12,1151-1172.

Rubin, D.B., & Stern, H.S. (1994). Testing in latent class models using a posterior predictive check distribution. In A. von Eye & C.C. Clogg (Eds.), Latent variables *analysis: Applications for developmental research* **(pp. 420-438). Thousand Oaks,** CA: Sage.

Rubin, D.B., & Thayer, D.T. (1982). EM algorithms for ML factor analysis. *Psychometrika, 47,* 69-76.

Rubin, D.B., & Thayer, D.T. (1983). More on EM for ML factor analysis. *Psychometrika, 48,* 253-257.

Scheines, R., Spirtes, P., Glymour, G., & Meek, C. (1994). *TETRAD II: Tools for causal modeling. User's manual.* Hilisdale, NJ: Erlbaum.

Smith, A.F.M., & Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B,* 55, 3-23.

Spirtes, P., Glymour, C, & Scheines, R. (1993). *Causation, prediction, and search.* New York: Springer.

Tanner, M.A. (1993). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions* (2^{nd} ed.). New York: Springer.

Wheaton, B., Muthén, B., Alwin, D., & Summers, G. (1977). Assessing reliability and stability in panel models. In D.R. Heise (Ed.), *Sociological Methodology 1977* (pp. 84-136). San Francisco: Jossey-Bass.

Yung, Y.-F., & Bentler, P.M. (1994). Bootstrap-corrected ADF test statistics. *British Journal of Mathematical and Statistical Psychology, 47,* 63-84.