Bayesian Estimation and Testing of Structural Equation Models*

Richard Scheines
Dept. of Philosophy
Carnegie Mellon University, USA

Herbert Hoijtink
Dept. of Methodology and Statistics
University of Utrecht, The Netherlands

Anne Boomsma
Dept. of Statistics, Measurement Theory, and Information Technology
University of Groningen, The Netherlands

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 (SEM) 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 Gaussian and in some cases contains local maxima. Nevertheless, the Gibbs sample comes from the correct 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, the posterior can be used to make inferences about the parameters of underidentified models, as we illustrate on a simple errors-in-variables model.

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

---

*We thank David Spiegelhalter for suggesting applying the Gibbs sampler to structural equation models to the first author at a 1994 workshop in Wiesbaden. We thank Ulf Böckenholt, Chris Meek, Marijtje van Duijn, Clark Glymour, Ivo Molenaar, Steve Klepper, Thomas Richardson, Teddy Seidenfeld, and Tom Snijders for helpful discussions, mathematical advice, and critiques of earlier drafts of this paper.

Information 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.

1. Introduction

With modern computers and the Gibbs sampler, a Bayesian approach to structural equation modeling (SEM) is now possible. Posterior distributions over the parameters of a structural equation model can be approximated to arbitrary precision with the Gibbs sampler, even for small samples. Being able to compute the posterior over the parameters allows us to address several issues of practical interest. First, prior knowledge about the parameters may be incorporated into the modeling process. Second, we need not rely on asymptotic theory when the sample size is small, a practice which has been shown to be misleading for inference and goodness-of-fit tests in SEM (Boomsma, 1983; Hoogland & Boomsma, in press). Third, the class of models that can be handled is no longer restricted to just-identified or over-identified models. Whereas each identifying assumption must be taken as given in the classical approach, in a Bayesian approach some of these assumptions can be specified with perhaps more realistic uncertainty. Each of these practical advantages is illustrated with data in section 3.

The paper is organized as follows. In the remainder of this section, we review maximum likelihood estimation (ML), Bayesian statistical inference, and introduce notation. In section 2 we explain how the Gibbs sampler can be applied to obtain a sample from the posterior distribution over the parameters of a SEM. We present statistics that can be used to summarize marginal posterior densities, as well as model checks using posterior predictive p-values. In section 3 we illustrate these techniques with two examples, the classic Stability of Alienation model (Wheaton, Muthén, Alwin, and Summers, 1977) and the effect of cumulative environmental lead exposure on IQ in children. We use the Alienation model to compare classical and Bayesian estimation on large and small samples, and we use the lead and IQ example to illustrate how a Bayesian strategy handles underidentified models. In the final section of the paper, we discuss general methodological issues.

1.1 Maximum Likelihood Estimation

The Gibbs sampler is not the only way to compute an approximation of the posterior distribution over the parameters of a SEM. One can also use normal distributions based

on maximum likelihood (ML) estimates. In what follows we compare both statistical approaches and evaluate their merits for SEM. As an introduction and for notation, we briefly review ML-estimation and Bayesian statistical inference.

Let $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)'$ be a set of N normally and independently distributed random variables $\mathbf{x} = (x_1, ..., x_p)'$, with expectation $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. The matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is a continuously differentiable matrix valued function of the parameter vector $\boldsymbol{\theta} = (\theta_1 , ..., \theta_t)'$, whose elements θ_j are the values of $t \leq p(p+1)/2$ unknown parameters. $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ represents the structural equation model in the population. Without loss of generality, we have no interest in first order moments. In that case, the sample covariance matrix **S** (p x p) is a sufficient statistic for estimation, where **S** is an unbiased estimate of $\boldsymbol{\Sigma}$ based on a sample of observations **X** (N x p). Hereafter, all densities and probabilites that are a function of **X** will be written as a function of **S**, the sufficient statistic for X. Under these assumptions, the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{ML}$ of the unknown parameter vector $\boldsymbol{\theta}$ can be obtained.

Let $p(\mathbf{S}|\boldsymbol{\theta})$ denote the joint probability density function of **S**. If $p(\mathbf{S}|\boldsymbol{\theta})$ is regarded as a function of $\boldsymbol{\theta}$, given the observations **S**, it is called the likelihood function of $\boldsymbol{\theta}$ given **S**, i.e., $L(\boldsymbol{\theta}|\mathbf{S}) = p(\mathbf{S}|\boldsymbol{\theta})$. Given the sample covariance matrix **S**, the log-likelihood can be expressed as

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

and thus in standard ML-estimation the following function of the log-likelihood is minimized:

$$F_{ML}[\mathbf{S},\boldsymbol{\Sigma}(\boldsymbol{\theta})] = \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + tr[\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})] - \log|\mathbf{S}| - p . \quad (2)$$

Programs like LISREL (Jöreskog & Sörbom, 1993) calculate $\hat{\boldsymbol{\theta}}_{ML}$ and estimates of the observed information matrix $I(\hat{\boldsymbol{\theta}}_{ML}|\mathbf{S})$, and thus estimates of asymptotic standard errors of each parameter estimate $\hat{\theta}_{j,ML}$, denoted as $SE(\hat{\theta}_{j,ML})$.

1.2 Bayesian Statistical Inference

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 occurred (cf. Box & Tiao, 1973, p. 72). In that context the focus is on the posterior density of $\boldsymbol{\theta}$ given the sample covariance matrix **S**, which, for normal variables is defined as

$$p(\boldsymbol{\theta}|\mathbf{S}) = p(\mathbf{S}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) / \int p(\mathbf{S}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta} \quad \propto \quad p(\mathbf{S}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \ . \tag{3}$$

Here $p(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$, expressing what is known about $\boldsymbol{\theta}$ before any knowledge of **S**. In constrast, the posterior distribution $p(\boldsymbol{\theta}|\mathbf{S})$ expresses the result of changing $p(\boldsymbol{\theta})$ to take the sample data into account. Given that $L(\boldsymbol{\theta}|\mathbf{S}) = p(\mathbf{S}|\boldsymbol{\theta})$, it follows that (3) can be expressed as

$$p(\boldsymbol{\theta}|\mathbf{S}) \propto L(\boldsymbol{\theta}|\mathbf{S})\, p(\boldsymbol{\theta}) \ . \tag{4}$$

Depending on the amount of prior knowledge relative to the sample information, the posterior distribution can be dominated by the likelihood or by the prior. If an uninformative ('improper') prior $p(\boldsymbol{\theta}) = c$ is used, where c is a real constant, the posterior distribution is proportional to the likelihood function, i.e., $p(\boldsymbol{\theta}|\mathbf{S}) \propto L(\boldsymbol{\theta}|\mathbf{S})$. If on the other hand an informative prior distribution is used, and in this paper it is assumed throughout that in such a case $p(\boldsymbol{\theta})$ has a multivariate normal distribution $N(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$ truncated below zero for variances, in small samples the posterior distribution $p(\boldsymbol{\theta}|\mathbf{S})$ is not proportional to the likelihood function $L(\boldsymbol{\theta}|\mathbf{S})$. Note that, for each variance, a normal truncated below zero is similar to an inverse chi-square, and allows the user to specify approximately the same prior knowledge.

Asymptotically, the posterior density $p(\boldsymbol{\theta}|\mathbf{S})$ converges to the likelihood, which, under appropriate regularity conditions, is proportional to the multivariate normal density $N(\hat{\boldsymbol{\theta}}_{ML}, I^{-1}(\hat{\boldsymbol{\theta}}_{ML}|\mathbf{S}))$ (cf. Tanner, 1993).

To summarize, there are at least two types of approximations to the posterior distribution over the parameters of a SEM: 1) a normal-based maximum likelihood approximation, i.e., $N(\hat{\boldsymbol{\theta}}_{ML}, I^{-1}(\hat{\boldsymbol{\theta}}_{ML}|\mathbf{S}))$, which can be obtained from LISREL, for

example, and 2) an approximation based on a sample from p(**q**|**S**) computed by the Gibbs sampler.

1.3 Finite Sample Size

The ML-estimation theory used in SEM is asymptotic theory. The same holds for other estimation methods, like generalized least squares (GLS) and weighted least squares (WLS). 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 SEM estimators, see for instance Bearden, Sharma and Teel (1982), Boomsma (1982, 1983), Baldwin (1986), Chou, Bentler and Satorra (1991), Hu, Bentler and Kano (1992),Yung and Bentler (1994), and Hoogland and Boomsma (in press). From such research it may roughly be concluded that, in order to obtain proper parameter estimates, the behaviour of ML, GLS and WLS is not robust for small N. More importantly, the (co)variances of parameter estimates are often incorrectly estimated in small sample studies, especially by the WLS method. As a consequence, for small N the sampling distribution of (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 is not known for small N. For almost any sample size the distribution of many fit indices that happen to be available is almost completely unknown; see Hu and Bentler (1995) or Boomsma (1996), for an overview.

In summary, it is not appropriate to use asymptotic estimation theory in SEM when the sample size is small. One strategy is to use the posterior distribution over the parameters instead of the asymptotic sampling distribution of the ML-estimator.

1.4 The Gibbs Sampler and ML-approximations

Joint and marginal posterior distributions, p(**q**|**S**) and p(q|**S**), can be numerically approximated to arbitrary precision, for any finite sample size N, with Markov Chain Monte Carlo (MCMC) methods, and in particular with a single-component Metropolis-Hastings algorithm, a specific case of which is the Gibbs sampler (Geman & Geman, 1984; Chib & Greenberg, 1995, p. 332).

If the sample size is large, the limiting normal approximation of the likelihood (i.e., the approximation of $L(\boldsymbol{\theta}|\mathbf{S})$ by $N(\hat{\boldsymbol{\theta}}_{ML}, I^{-1}(\hat{\boldsymbol{\theta}}_{ML}|\mathbf{S}))$ is a legitimate approximation of $p(\boldsymbol{\theta}|\mathbf{S})$, 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, Carlin, Stern, & Rubin, 1995, p. 92)

As the sample size N increases, the ML-estimate $\hat{\theta}_{j,ML}$ converges numerically to the mode of the marginal posterior density, and its estimated standard error, $SE(\hat{\theta}_{j,ML})$, converges to the standard deviation of θ_j in the posterior normal density, denoted as $SD(\theta_j)$.

Thus in large samples the Gibbs sampler and the normal theory ML-approximation of the posterior density (likelihood) should produce almost exactly the same numerical quantities for corresponding statistics, though their interpretation will be different (cf. Box & Tiao, 1973). We expect these quantities to diverge as sample size decreases, however.

In comparing both approaches it will be clear from the examples that Gibbs' sampling has a number of advantages over the normal ML-approximation.

a. Asymptotic inference is not needed. We do not have to rely on normal approximations of the posterior. The procedure works for all sample sizes.
b. Knowing the posterior density allows inspection of the fit of the model by posterior predictive p-values (see Gelman, Meng & Stern, 1996; Rubin & Stern, 1994; Meng, 1994).
c. Prior knowledge can be incorporated flexibly. Inequality restrictions can be implemented in the sampling procedure in such a way that not only the parameter estimates, but the estimated standard errors and interval estimates as well, are bound to those restrictions.
d. The user may get information about multimodality in marginal posterior densities, which is undetectable by standard procedures.

e. The posterior for the parameters of an underidentified model can be obtained by using the Gibbs sampler with an informative prior, but not with a normal ML-approximation.

2. Posterior Inference Based on the Gibbs Sampler

The Gibbs sampler is an iterative procedure that, after it has converged, renders a dependent sample from p($\boldsymbol{\theta}$|**S**). In each iteration m=1,...,M, each parameter 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 sample covariance matrix **S**. An accessible but detailed introduction to the Gibbs sampler can be found in Casella and George (1992), more elaborate discussions in Gelfand and Smith (1990), Gilks, Richardson, and Spiegelhalter (1996), Tierney (1993), and Smith and Roberts (1993).

The parameter vectors in the Gibbs sample can be used to compute characteristics of the marginal posterior density. Among other things, expected a posteriori estimates, median a posteriori estimates, posterior standard deviations, central credibility intervals, and the posterior covariance matrix of the parameters may be computed.

2.1 Initial Values

The iterative process begins by assigning an initial value (m=0) to the model parameters $\boldsymbol{\theta}$. We use a subscript to index the parameter, and a superscript to index the iteration. Thus, the jth parameter in the mth iteration is written as θ_j^m. If the prior distribution is informative (in which case it is a multivariate normal truncated below zero for variance parameters), then the mean in the prior is used as the starting value, i.e., $\boldsymbol{\theta}^0 = \boldsymbol{\mu}_0$. If the prior is uninformative, then initial values may be chosen relatively arbitrarily, e.g., zero for a path coefficient, and one for a variance.

2.2 Sampling the Conditional Posterior of Each Parameter

In each iteration, each parameter is sampled in a fixed order from its posterior conditional upon the current values of the other parameters and the data. Parameter θ_j in iteration m, for example is sampled from:

$$p(\theta_j \mid .) = p(\theta_j \mid \theta_1^m, \ldots, \theta_{j-1}^m, \theta_{j+1}^{m-1}, \ldots, \theta_t^{m-1}, LB_j, UB_j, \mathbf{S}) \quad , \tag{5}$$

where LB_j and UB_j are lower and upper bounds for θ_j, respectively. This is what Chib and Greenberg (1995, p. 332) call the Gibbs sampler. Note that (5) is conditional upon the current values of the other parameters, which for some parameters is the value sampled in iteration m and for others the value sampled in the previous iteration (m-1).

"Fixed parameters" are left at their initial value and never updated, although they are still conditioned on when evaluating (5) for a free parameter. Parameters may be subjected to inequality constraints with respect to constants or with respect to each other. A few examples: if a parameter is a variance, the lower bound is zero and the upper bound is ∞. If parameter 2 has to be larger than parameter 1 and smaller than parameter 3, the lower and the upper bounds are θ_1^m and θ_3^{m-1} respectively. If a parameter is unconstrained, the lower and upper bounds are -∞ and ∞, respectively.

The conditional posterior (5) is similar to (4) with all parameters fixed at their current values except θ_j, i.e.,

$$p(\theta_j \mid .) \propto L(\theta_1^m, \ldots, \theta_{j-1}^m, \theta_j, \theta_{j+1}^{m-1}, \ldots, \theta_t^{m-1} \mid \mathbf{S})\ p(\theta_j) \tag{6}$$

The likelihood in (6) corresponds to (1), and this is the part of our use of MCMC that is specific to SEM. Since the SEM version of the conditional posterior in (6) is not necessarily proportional to a commonly used distribution like a normal or a chi-squared, it cannot be sampled from by using standard computational procedures. We draw samples from (6) by using a combination of inverse probability sampling and rejection sampling (Gelman, et al., 1995, pp. 302-305). The idea is to first approximate (6) by a standard distribution (in our case a normal) that can be sampled using standard procedures, and then to adjust this distribution by rejecting draws in proportion to how the approximation differs from (6).

2.3 Inverse Probability and Rejection Sampling

2.3.1 The Approximating Distribution

The first step in rejection sampling is the choice of a distribution that is approximately proportional to (6), and from which a pseudo-random sample can be easily obtained. We use a normal distribution with mean (denoted as Mode) equal to the mode of (6) and variance equal to cV, where V is computed as the inverse of the observed Fisher information evaluated at Mode, and c is the "stretch," which will be explained below. This distribution will be denotod by prox(θ_j |Mode,cV). We compute the mode of (6) with Brent's method in one dimension (Press, Teukolsky, Vetterling and Flannery, 1992, pp. 395-398). The variance V is computed as

$$V = \frac{.00005}{\log p(\text{Mode}|\ .) \ - \ \log p(\text{Mode}+.01|\ .)} \quad . \tag{7}$$

This formula can be obtained by approximating (6) using a second order Taylor expansion around Mode (see Gelman, et al., 1995, p. 95).

2.3.2 Rejection Sampling

In our current implementation of the Gibbs sampler as described here, which is available in TETRAD III[1] and was used for the illustrations in this paper, we repeatedly sample from the normally distributed prox(θ_j|Mode,cV) until a "draw" is within the bounds imposed by LB_j and UB_j.

After a draw v from prox(θ_j|Mode,cV) ~ N(Mode,cV) is within the upper and lower bounds, we then correct for the approximation by keeping v with probability proportional to the ratio of the real and appropriately normalized approximating distributions, evaluated at v. This is valid only when the approximating distribution "covers" the true distribution (see the figure presented by Gelman, et al., 1995, p. 304). We insure this in two ways. First, the variance of the approximating distribution is multiplied with a "stretch factor" c (experience untill now indicates that c=2 is usually fine). Second, the

[1] TETRAD III is available at: http://hss.cmu.edu/philosophy/TETRAD/tetrad.html

approximating distribution is multiplied by the ratio of the conditional posterior (6) and the approximating distributions evaluated at Mode. This ensures that the true conditional posterior density and the approximating density are equal at the mode.

The value v drawn from the approximating distribution is thus accepted as a draw from (6) with probability

$$\frac{p(v \mid .)}{\text{prox}(v \mid \text{Mode}, cV)\ \dfrac{p(\text{Mode} \mid .)}{\text{prox}(\text{Mode} \mid \text{Mode}, cV)}} \quad . \tag{8}$$

Note that (8) can only be interpreted as a probability if it is less than or equal to 1.0, which is not always the case. Rejection sampling undersamples from those regions of v in which (8) exceeds 1.0. To minimize this, one always accepts v as a draw when (8) exceeds 1.0 and tries to form an approximating distribution such that the regions in which (8) exceeds 1.0 are small and far out in the tails of both the approximating and real distributions.

2.4 Convergence and Dependence

In practice there is no generally agreed upon method to decide whether a Gibbs sequence has converged or not. See, for example, Gelman and Rubin (1992) and subsequent discussions, e.g., MacEachern and Berliner (1994).

We use the following procedure to assess convergence. First, we retain only every 25th or 50th iteration from the original sequence ($\mathbf{q}^1$,..,$\mathbf{q}^M$) described above (the rationale behind this step will be explained below). These iterations are indexed $\mathbf{q}^k$, k=1,...,K. We inspect the mean, median, standard deviation, and 5th and 95th percentile of the sample from the marginal posterior distribution over each parameter across each of four sequences of K/4 iterations. If the resulting numbers are similar, we judge the sampler to have converged, if they are dissimilar or are mildly dissimilar but show an increasing or decreasing trend we judge the sampler to have not converged. All the examples in section 3 converged quickly and solidly.

The Gibbs sampler usually requires a "burn in" period before it converges in distribution to the true posterior. In our examples (see section 3), burn in was always

almost instantaneous. When the initial segment of a Gibbs sequence has not burned in, the obvious solution is to discard the initial segment of iterations occurring prior to convergence, and analyze only the draws after burn-in.

The Gibbs sampler does not render independent draws from the posterior. It is clear from the sampling scheme described above that the draws in each iteration depend on the draws obtained in the previous iterations. Currently it is not clear if this dependence is a problem with respect to making inferences about the posterior. See, for example, Gelman, et al., (1995, p. 330). Some authors propose to use only every 50th iteration to achieve approximate independence (Zeger & Karim, 1991). Our experience on SEM models is that, as long as a sequence has converged and the number of iterations retained is substantial, it makes no practical difference if we keep all or every 25th or every 50th iteration. To be safe, however, in all our examples we use every 25th or 50th.

2.5 Posterior Inference

The marginal posteriors can be used to make inferences with respect to the parameters of the SEM under investigation. For each parameter θ_j, the mean (expected a posteriori, $\theta_{j,EAP}$) or the median (median a posteriori, $\theta_{j,MDAP}$) can be used as point estimates, and the posterior standard deviation ($SD(\theta_{j,EAP})$), or the 95% central credibility interval ($\theta_{j,.025}$ - $\theta_{j.975}$) can be used as a measure of our uncertainty about these point estimates. Since we do not have the posterior directly but only a Gibbs sample from it, these quantities cannot be computed directly, but can be closely approximated (the quality of the approximation depends on the number of retained iterations K) by calculating their sample analogues, i.e., the Gibbs sample mean $\hat{\theta}_{j,EAP}$, sample median $\hat{\theta}_{j,MDAP}$, sample standard deviation $SD(\hat{\theta}_{j,EAP})$, and the 95% sample credibility interval ($\hat{\theta}_{j,.025}$ - $\hat{\theta}_{j,.975}$). Furthermore, the posterior covariance matrix may be estimated by computing the covariance matrix of the sample of K parameter vectors. Note finally, that plots of univariate marginal distributions are easily constructed.

The mode of the marginals in the posterior (the maximum a posteriori, or $\theta_{j,MAP}$) is the only important quantity that cannot be easily estimated from the Gibbs sample. In the

case where the prior distribution is uninformative or "swamped" by the likelihood, then $\hat{\theta}_{j,MAP}$ can be obtained using standard SEM software like LISREL.

2.6 Goodness-of-Fit Statistics From Posterior Predictive p-values

The likelihood ratio goodness-of-fit statistic for SEM:

$$LR[\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})] = (N-1)\,[\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \mathrm{tr}[\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})] - \log|\mathbf{S}| - p\,] \quad , \tag{9}$$

is known to be distributed as χ^2 only asymptotically, and can be substantially non χ^2 for finite samples (Bollen, 1989). Thus p-values (tail-area probabilities) for the goodness-of-fit statistic based on the χ^2 distribution can also be way off. In this section we explain how posterior predictive p-values (Rubin, 1984; Meng, 1994; Gelman, Meng and Stern, 1996) can be used in SEM to evaluate the likelihood ratio goodness-of-fit statistic without relying on asymptotics. The classical p-value based on (9) is

$$\text{p-value} = p\{LR[\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})] < LR[\mathbf{S}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})] \mid H, \boldsymbol{\theta}\}\,, \tag{10}$$

where $\mathbf{S}(\boldsymbol{\theta})$ denotes a covariance matrix drawn randomly, with appropriate N, from $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, and H denotes the null hypothesis, i.e., the SEM specified holds in the population. Because the population parameters $\boldsymbol{\theta}$ are in practice unknown, and are thus "nuisance parameters," it is not possible to evaluate (10) (see Meng, 1994).

The solution implemented in standard SEM software like LISREL and EQS is to use $\hat{\boldsymbol{\theta}}_{ML}$ for $\boldsymbol{\theta}$, which gives an approximation of (10) that is asymptotically correct. The posterior predictive p-value replaces (10) by (11):

$$\begin{aligned} \text{p-value} &= p(LR[\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})] < LR[\mathbf{S}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})] \mid \mathbf{S}, H) \\ &= \int_{\theta} p\,(LR[\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})] < LR[\mathbf{S}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})] \mid H, \boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{S})\, d\boldsymbol{\theta} \quad . \end{aligned} \tag{11}$$

By using $p(\boldsymbol{\theta}|\mathbf{S})$ the nuisance parameter $\boldsymbol{\theta}$ from (10) is integrated out of (11). Note that (11) is not an asymptotic approximation. The p-value defined in (11) can be approximated from the Gibbs sample with (12):

$$\text{p-value} \approx \sum_{k=1}^{K} p\{LR(\mathbf{S},\boldsymbol{\Sigma}(\mathbf{q}^k)) < LR(\mathbf{S}(\mathbf{q}^k),\boldsymbol{\Sigma}(\mathbf{q}^k))\} / K$$

$$\approx \sum_{k=1}^{K} \sum_{z=1}^{Z} I_{kz}/KZ, \tag{12}$$

where the indicator variable $I_{KZ} = 1$ if $LR(\mathbf{S},\boldsymbol{\Sigma}(\mathbf{q}^k)) < LR(\mathbf{S}_z(\mathbf{q}^k),\boldsymbol{\Sigma}(\mathbf{q}^k))$, and 0 otherwise. The integral in (11) is approximated using a summation over K values of $\mathbf{q}$ sampled from $p(\mathbf{q}|\mathbf{S})$, where K is the number of values of $\mathbf{q}$ sampled from $p(\mathbf{q}|\mathbf{S})$. The inequality p(.<.) in (11) and (12) is approximated by the proportion observed in z = 1,..,Z sample covariance matrices $\mathbf{S}_z(\mathbf{q}^k)$ drawn pseudo-randomly from the population determined by $\mathbf{q}^k$. All standard SEM software can now draw psuedo-random covariance matrices from a parameterized SEM or a given population covariance matrix. For a detailed account of how we implemented this, see (Scheines, et al., 1994, chapter 13).

3. Examples

This section discusses examples in which the Gibbs sampler implemented in TETRAD III is used to draw a sample from the posterior distribution over the parameters of a structural equation model. We use a classic LISREL model of the stability of social alienation to compare maximum likelihood estimates with estimates based on the Gibbs sample when N is large and small. We then consider an example in which we specify an informative prior distribution over the amount of measurement error in an "underidentified" errors-in-all-variables model of the effect of lead exposure on the IQ of children, concluding that lead's effect is indeed deleterious.

3.1 The Stability of Alienation

Consider a classic longitudinal structural equation model developed by Wheaton, Muthén, Alwin, and Summers, (1977) to investigate the stability of social alienation (Figure 1), where measured indicators are boxed and latent variables are enclosed in ovals. Anomia and powerlessness are scales constructed from survey questions, education is years of school, and SEI a socio-economic index constructed from several factors, e.g., income, job status, etc.

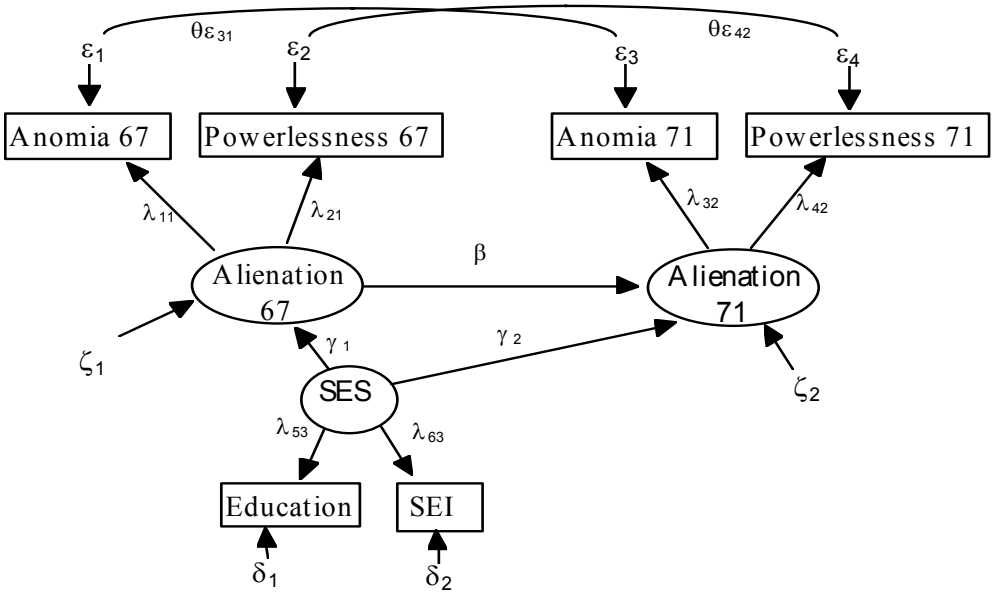

Figure 1: The Stability of Alienation model.

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). Measurement models were constructed for the latent variables, and the central purpose of the study was to estimate the parameter β. Using the sample covariance matrix **S** reported in Wheaton, et al.'s paper, we compared estimates obtained first with EQS (Bentler, 1995) and second from the Gibbs sampler in TETRAD III.

Assuming that the observed variables are multivariate normal (which is done in the original analysis), and using an improper "flat" prior $p(\boldsymbol{\theta}) = c$, in which case $p(\boldsymbol{\theta}|\mathbf{S}) \propto L(\boldsymbol{\theta}|\mathbf{S})$, we computed a Gibbs sample of size M=25,000 from $p(\boldsymbol{\theta}|\mathbf{S})$, using initial values equal to the ML-estimates, i.e., $\boldsymbol{\theta}^0 = \hat{\boldsymbol{\theta}}_{ML}$. We kept every 25th iteration to produce an approximately independent 1,000 final draws (K=1,000). Table 1 gives, for each structural parameter, a point estimate $\hat{\theta}_{j,EAP}$ and a measure of the marginal posterior diffusion $SD(\hat{\theta}_{j,EAP})$ for the retained subsample of 1,000 draws, as well as the corresponding quantities calculated from asymptotic theory by EQS. The relevant properties of the posterior are nearly identical to those computed by EQS from the normal approximation to the posterior using $\hat{\boldsymbol{\theta}}_{ML}$ and $SE(\hat{\boldsymbol{\theta}}_{ML})$.

Table 1. Gibbs vs. ML-estimates for the Stability of Alienation Model. M=25,000, K=1,000, and N=932.

	$\hat{\theta}_{j,EAP}$	$\hat{\theta}_{j,ML}$	SD($\hat{\theta}_{j,EAP}$)	SE($\hat{\theta}_{j,ML}$)
γ_1	-0.579	-0.575	0.057	0.056
γ_2	-0.226	-0.227	0.055	0.052
$\hat{\beta}$	0.608	0.607	0.052	0.051

As we discussed in section 2.6, the Gibbs sample can be used to compute a posterior predictive p-value based on the likelihood ratio goodness-of-fit statistic. To again compare results based on the Gibbs sample to those calculated by the standard LISREL or EQS approach, we calculated the posterior predictive p-value (with Z=5) and the p-value for the χ^2 goodness-of-fit statistic as computed by EQS. Using the sample covariance matrix **S** reported in Wheaton, we consider the full model in Figure 1 above, and also a submodel of Figure 1 in which all error terms are uncorrelated. For the full model, EQS gave $p(\chi^2) = 0.315$, whereas the posterior predictive p-value was 0.447. For the uncorrelated error model, EQS gave a p-value based on the χ^2 of 0.00, and the posterior predictive p-value was also 0.00.

We repeated the study with N set artificially to 20,000, and the estimates, standard errors, and p-values became virtually identical.

As Boomsma (1983) has shown (in fact using the Wheaton et al., model), inferences based on SE($\hat{\theta}_{ML}$) can be wildly overconfident in the small sample. The reason for this is that at small sample sizes the asymptotic approximation of the likelihood surface is sometimes quite different from the actual likelihood. To illustrate, we repeated the study above with a pseudo-random sample $\mathbf{S_{50}}$ (N=50) drawn by TETRAD III from the population defined by $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_{ML})$ for Wheaton et al.,'s original data and model (Table 2).

Table 2. Sample Covariance Matrix $\mathbf{S_{50}}$

Anomia 67	14.302					
Powerlessness 67	7.064	8.296				
Anomia 71	8.563	4.700	16.253			
Powerless 71	6.881	5.624	8.425	10.169		
Education	-4.834	-4.829	-6.271	-5.838	12.894	
SEI	-2.081	-2.486	-2.700	-2.563	3.417	2.808

What emerged was at first disturbing but eventually illuminating. The marginal posterior distributions for some of the parameters had more than one mode and were very diffuse relative to the asymptotic approximation obtained from the ML solution. For certain SEMs, including Wheaton's model, the likelihood surface indeed has more than one local maximum (Scheines, Boomsma, and Hoijtink, 1997), and it is for this reason that the approximation of the posterior by a maximum likelihood estimator is so poor. Table 3 shows the wild discrepancy between EQS's results and those based on a subsample (K=1,000 and M=10,000) values sampled from $p(\beta|\mathbf{S}_{50})$.

Table 3. A comparison of the estimates and standard errors of β in the Stability of Alienation model: Gibbs sampling vs. ML. M=10,000, K=1,000, and N=50.

$\hat{\beta}_{ML}$	$\hat{\beta}_{MDAP}$	$\hat{\beta}_{EAP}$	$SE(\hat{\beta}_{ML})$	$SD(\hat{\beta}_{EAP})$	$\hat{\beta}_{.025}$	$\hat{\beta}_{.975}$
0.493	0.830	6.650	0.228	45.701	-54.92	128.30

The inferences about β from ML and Bayesian estimation are completely at odds when N is small, e.g., 50. What is particularly striking is that $SD(\hat{\beta}_{EAP})$ is approximately 200 times larger than $SE(\hat{\beta}_{ML})$, even though for the original sample at N=932 these quantities are almost identical. The estimate $\hat{\beta}_{ML}$ is over twice as big as its standard error $SE(\hat{\beta}_{ML})$, and thus according to asymptotic maximum likelihood estimation theory we can reject the null hypothesis that β is negative or 0 at a significance level of 0.05. From the Gibbs sample $p(\beta|\mathbf{S}_{50})$, however, we know almost nothing about β, let alone its sign.

In sum, although asymptotic ML-estimation provides a very good approximation of the posterior over the parameters when the sample size is large, e.g., 500, it gives a very poor approximation when N is small, e.g, 50.

3.2 Underidentified Models: Lead and IQ

In a 1985 article in *Science*, Needleman, Geiger and Frank reanalyzed data they had previously collected on the effect of lead exposure on the verbal IQ score of 221 suburban white children. After eliminating approximately 35 potential confounders with backwards stepwise regression, they settled on regressing child's IQ on lead exposure, controlling for measures of genetic factors, environmental stimulation, and physical factors that might compromise the child's cognitive endowment. Using the Build Module in TETRAD II (Scheines, et al., 1994), we were able to eliminate all the physical factor variables with almost no predictive loss (Scheines, 1997). The final set of variables we used are as follows:

ciq	the child's verbal IQ score
lead	the measured concentration of lead in the child's baby teeth
med	the mother's level of education, in years
piq	the parent's IQ scores

Standardizing all the measured variables (which we do throughout this analysis), the regression solution is as follows, with t-statistics in parentheses:

$$\hat{ciq} = -.177\ \text{lead} + .251\ \text{med} + .253\ \text{piq} \quad .$$
$$\qquad\quad (2.89) \qquad (3.50) \qquad (3.59)$$

All coefficients are significant at 0.05, $R^2 = .243$, and the estimates are very close to those obtained by including the physical factor variables (see Scheines, 1997).

As Klepper (1988) points out, however, the measured regressor variables are really proxies that almost surely contain substantial measurement error. Although an errors-in-all-variables SEM (Figure 2) seems a more reasonable specification, unless we know precisely the amount of measurement error for each regressor, this model is underidentified.

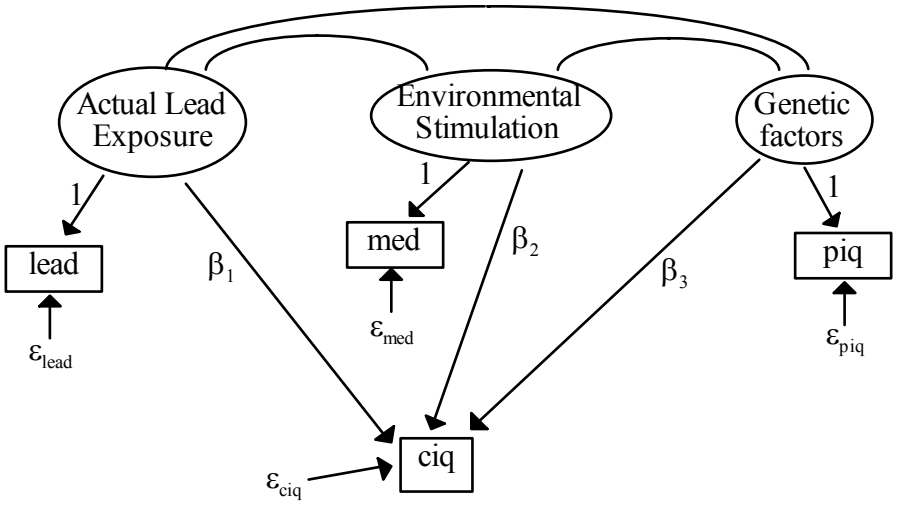

Figure 2: Errors-in-all-variables model for Lead's influence in IQ. Measured variables are boxed, and latent variables enclosed in ovals.

Several strategies have been discussed for handling models of this type and underidentified models in general. One is instrumental variable estimation (Bollen, 1989, p. 110), another is a sensitivity analysis (Greene & Ernhart, 1993) and still another is to bound parameters rather than produce a point estimate for them (Klepper & Leamer, 1984). An additional strategy, made possible by the Gibbs sampler, is Bayesian estimation. In this section we illustrate the Bayesian alternative, and in section 4.1 we briefly discuss the different strategies.

If we standardize the measured variables in the model shown in Figure 2, then the amount of measurement error for lead, which measures Actual Lead Exposure, and for med, which measures Environmental Stimulation, and for piq, which measures Genetic factors, is parameterized by $\text{Var}(\varepsilon_{lead})$, $\text{Var}(\varepsilon_{med})$, and $\text{Var}(\varepsilon_{piq})$, respectively. Since the model implies that Var(lead) = Var(Actual Lead Exposure) + $\text{Var}(\varepsilon_{lead})$, for example, and we are constraining Var(lead) to unity, then if we were to set $\text{Var}(\varepsilon_{lead}) = 0.25$, we would be asserting that 25% of the variance of measured lead comes from measurement error, while 75% comes from Actual Lead Exposure.

In this case, and many others like it, there is reasonable prior information about the amount of measurement error present, but it is not specific enough to assign a unique value to the parameters associated with measurement error. Needleman pioneered a

technique of inferring cumulative lead exposure from measures of the accumulated lead in a child's baby teeth. Between 0% and 40% of the variance in Needleman's proxy is probably from measurement error, with 20% a conservative best guess. For the measures of environmental stimulation and genetic factors, we are less confident, so we will guess that between 0% and 60% of the variance in med and piq is from measurement error, with 30% as our best guess. To translate these speculations into a prior, we specified a normal prior (truncated below zero) in which the mean is set to our best guess and the standard deviation half the distance to the extremity of our guess.

Table 4. Prior distribution over the parameters in the errors-in-all-variables model.

Parameter	Mean (μ_0)	Standard Deviation (σ_0)
Var(ε_{led})	0.20	0.10
Var(ε_{med})	0.30	0.15
Var(ε_{piq})	0.30	0.15
Other 10 Parameters	Comparable Regression value	4.00

For example, the mean in our prior for Var(ε_{med}) is 0.30, and our standard deviation is 0.15. Table 4 summarizes the marginal distributions for our mutlivariate normal prior (truncated below zero for variance parameters), and in our prior we assume there is no covariation between parameters. For all non-measurement error parameters, we used the comparable regression estimate as a mean in the prior, and a standard deviation of 4.0. For example, for β_1, we used a mean in the prior of -0.177, and standard deviation of 4.0. With such a high standard deviation, the prior is effectively uninformative about the 10 non-measurement error parameters.

Using this prior, and the mean values in the prior for inital values in the Gibbs sequence, we produced 50,000 iterations with the Gibbs sampler in TETRAD III. The sequence converged immediately. The histogram in Figure 3 shows the shape of the marginal posterior over β_1, the crucial coefficient representing the influence of actual lead exposure on children's IQ.

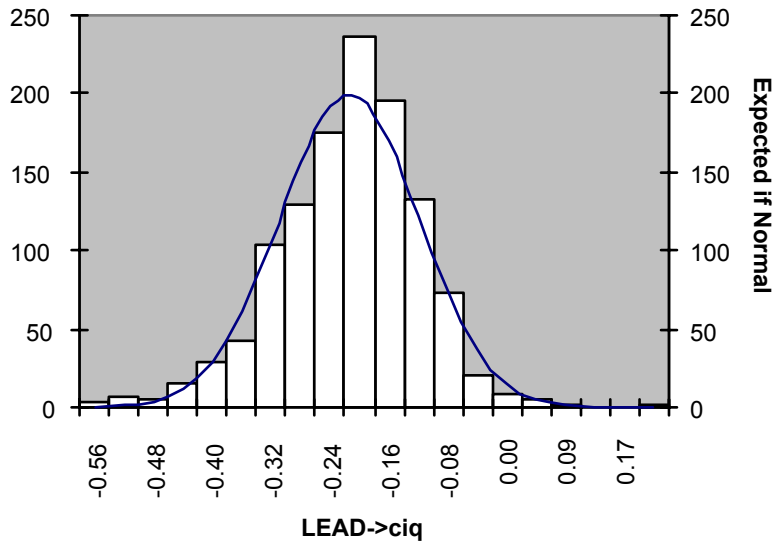

Figure 3. Histogram of relative frequency of β_1 in Gibbs sample. M=50,000, K=1,000, and N=221

The results support Needleman's original conclusion, but do not require the unrealistic assumption of zero measurement error. The Bayesian point estimate of the effect of Actual Lead on IQ, $\hat{\beta}_{1,EAP}$, is -0.215, and since the central 95% region of its marginal posterior lies between -0.420 and -0.038, we conclude that exposure to environmental lead is indeed deleterious conditional on this model and our prior uncertainty as specified.

4. Discussion

In this section we consider some of the methodological points that arise in applying Bayesian estimation and testing to SEM.

4.1 Underidentified Models

Virtually every introductory book on SEM 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. To make matters concrete, consider the errors-in-variables model of lead and IQ in Figure 2. Although our original regression model involving these measured variables is just identified, it seems almost certain that the measured regressors are in fact proxies for the real causal quantities of interest, which are indeed measured with error. Incorporating this fact into the model's specification, however, produces an underidentified model.

As we noted above, several strategies have appeared in the statistical and social science literature for handling underidentified models, in particular errors-in-variables models. One solution, popular especially in econometrics, is instrumental variable estimation. For each true regressor X_i^* measured by X_i with error, one finds another variable that "has no direct impact on the dependent variable, but has a correlation with the explanatory variable and no correlation with the disturbance term" (Bollen, 1995, p. 110). Such a variable will indeed allow us to consistently estimate the coefficient relating the true explantory variable X_i^* to the dependent variable Y, but the estimator now depends crucially on at least two extra identifying assumptions. To use instrumental variable estimation on the model in Figure 2, we would need to find three such variables.

In a sensitivity analysis (Greene & Ernhart, 1993), one fixes enough free parameters to identify the model. One then sets these parameters at a variety of levels, and then plots the estimates for the parameter of interest (and a 95% confidence interval around the estimate, for example) as a function of these other parameters. In the lead case, the free parameters might be the measurement error parameters, and the parameter of interest β_1. One then looks for the dependence of the estimated parameter of interest (and its standard error) on the parameters fixed. The researcher must then decide if prior knowledge can reasonably bound the parameters manipulated in the analysis into regions such that the parameter of interest is on one side of a threshold. Just this strategy is taken by Greene and Ernhart (1993), and their findings are consistent with ours. A sensitivity analysis avoids eliciting a full prior (in fact it minimizes the amount of prior knowledge required), but it can be difficult to apply when the parameter of interest is a relatively complicated function of the parameters varied. Researchers will rarely, for example, be able to bound four parameters into any but the simplest sort of region in a four-dimensional parameter

space. Most analyses report the dependence between the parameter estimate and the manipulated parameters one parameter at a time, which can be substantially misleading.

A similar strategy is to bound the parameters in an underidentified linear errors-in-all-variables model directly. Klepper and Leamer (1984), for example, proved that in certain circumstances the parameters in such models can be bounded just from assuming that the variance-covariance matrix is positive semi-definite. In other circumstances, bounds on some parameters can be extracted from bounds on others, in which case this strategy is similar to the sensitivity analysis strategy. Klepper (1988) has extended this technique and made it practical by sequentially probing the user's prior knowledge for the committments necessary for a bounding solution. Applying Klepper's technique to Needleman's data, we found that we must be willing to bound the measurement error of lead, med, and piq at 0.710, 0.465, and 0.457 respectiviely. Bounding the amount of measurement error for Actual Lead Exposure at 71% seems reasonable, but bounding it below 50% for Environmental Stimulation seems a bit suspect. The main difficulty with this technique, however, is that it does not admit inference -- it applies to population data and thus is forced to treat the sample data as if it were population data.

In the Bayesian strategy for handling underidentified models, no exact identifying assumptions are necessary (as in instrumental variable estimation), and no exact bounding levels are necessary (as in Klepper's strategy or sensitivity analysis). One need only specify a prior, approximate the posterior, and make inferences based on the posterior as we did in the lead and IQ case. On the other hand, in many cases background knowledge is weak, and pretending to capture this uncertainty by elliciting a well defined prior probability distribution can be more wishful thinking than good science.

If the model specified is underidentified, which is not the case in instrumental variable estimation, then all of these strategies attempt to leverage imperfect prior knowledge about some model parameters into imperfect but useful knowledge about others. In the Bayesian strategy it might seem strange that we can sharpen the information on a parameter, e.g., β_1 in the lead and IQ case, when in large samples the same parameter would have a flat posterior distribution (because the model is underidentified and because the likelihood dominates the prior in large samples). It is not the case, however, that the likelihood surface over an underidentified parameter need be

entirely flat. Rather it must have a flat region at its peak in the likelihood surface, which will dominate the posterior in the large sample. Klepper and Leamer (1984) show, however, that the region where the likelihood is maximal is flat but bounded, and not flat over the entire likelihood surface.

4.2 The Posterior Predictive Check

The posterior predictive check that we implemented was suggested by Rubin (1984) and elaborated by Meng (1994) and Gelman, Meng and Stern (1996). Although not a purely Bayesian test of model fit, the posterior predictive p-value is a clever hybrid between a classical and Bayesian approach to model testing.

In a fully Bayesian approach, one puts a prior distribution over the models under consideration, collects data, and computes the posterior over these models. This approach has been applied to SEM by Raftery and Madigan (Madigan & Raftery, 1991; Raftery 1993, 1994, 1996). Raftery's thrust has been to analytically approximate posterior probabilities with the Bayes Information Criterion.

In the classical approach to SEM model testing, one calculates a p-value for a model by computing a measure of discrepancy between the observed **S** and an estimate of the implied covariance matrix, e.g., the likelihood ratio test in (9), and comparing this discrepancy to a reference distribution of discrepancies, e.g., the χ^2 with the appropriate degrees of freedom.

There are two practical problems with the classical approach when applied to SEM. First, even if the population parameters $\boldsymbol{\theta}$ are known, the reference distribution is only known asymptotically. This can be overcome by simulation or bootstrap methods, however. In the simulation solution, for example, one specifies $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and draws any number of pseudo-random samples from $\boldsymbol{\theta}$ and forms the reference distribution of discrepancies empirically. Several SEM programs now perform this computation for the likelihood ratio test, e.g., EQS.

The second problem is that for fixed N the reference distribution of discrepancies is not invariant under different values of the population parameters $\boldsymbol{\theta}$, i.e., the test is not pivotal. The posterior predictive check addresses this problem by incorporating uncertainty over $\boldsymbol{\theta}$ into the p-value. It forms a reference distribution of discrepancies by

mixing *all* the reference distributions determined by different values of **q**, in proportion to the density of **q** in the posterior.

Since it produces a p-value, however, in the end the posterior predictive check resorts to a frequentist justification, and it is still an open question how it will fare in SEM when compared systematically with a large simulation study to the classical p-value and other alternatives.

4.3 Multimodality, Asymptotics, and Tacit Prior Information

In ML-estimation of SEMs from a Bayesian point of view the posterior computed from asymptotic theory is by definition Gaussian and thus unimodal. When the sample size is small, however, the actual likelihood surface and thus the posterior is for some models multimodal. As the sample grows large, the alternative modes become small enough to ignore, so techniques which assume they do not exist are perfectly reasonable. At small N the possibility of multimodality cannot be ignored, however, 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 in some cases small amounts of prior knowledge can have a big effect on bringing the posterior back to unimodality (Scheines, et al., 1997).

4.4 Multivariate Normality

Although in this paper we assume that the measured variables **X** are distributed as multivariate normal, there is no need to do so in the Bayesian approach in general and in the Gibbs sampler. The only requirement for using these techniques is that one be able to evaluate the (conditional) likelihood L(**q**|**X**) and the prior p(**q**) for any value of **q**. In SEMs with latent variables and continuous **X**, we know how to do this when **X** is multivariate normal but not otherwise. Extending the distributions over non-normal continuous **X** for which we can evaluate L(**q**|**X**) in SEM is therefore an important research topic.

If the measured variables are discrete, but are thought to be projections of underlying variables distributed as multivariate normal, then we can also evaluate L(**q**|**X**); see Muthen (1984).

Another class of causal models that have received substantial attention in the last several years are Bayesian networks (Pearl, 1988; Spirtes, Glymour, & Scheines, 1993; Jensen, 1996). If all the variables in a Bayesian network are measured, discrete and distributed multinomially, then the likelihood function can be evaluated (Heckerman & Geiger, 1994), and the Gibbs sampler used profitably. Geiger, Heckerman, and Meek (1996) have recently pushed the discrete variable Bayesian network technology forward to include latent variables.

References

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. (1995). *EQS: Structural equations program manual* (Version 5.0). Encino, CA: Multivariate 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.

Bollen, K. (1995) An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika*, 61, 109-121.

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)

Boomsma, A. (1996). De adequaatheid van covariantiestructuurmodellen: een overzicht van maten en indexen [The adequacy of structural equation models: An overview of statistics and indices]. *Kwantitatieve Methoden, 52,* 7-52.

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

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

Chib, S. & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician, 49,* 327-335

Chou, C.-P., Bentler, P.M., & Satorra, 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., and Meek, C. (1996). *Asymptotic Model Selection for Directed Networks with Hidden Variables* (Microsoft Technical Report MSR-TR-96-07). Microsoft Research.

Gelfand, A.E., & Smith, A.M.F. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85,* 398-409.

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

Gelman, A., Meng, X.-L., & Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*.

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

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.

Greene, T. and Ernhart, C. (1993). Dentine lead and intelligence prior to school entry: A statistical sensitivity analysis. *Journal of Clinical Epidemiology, 46,* 323-329.

Heckerman, D., & Geiger, D. (1995). *Likelihoods and Parameter Priors for Bayesian Networks.* (Technical Report MSR-TR-95-54). Microsoft Research.

Hoogland, J.J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26,* 329-368.

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.

Jensen, F.V. (1996). *An introduction to Bayesian networks*. New York: Springer Verlag

Jöreskog, K.G., & Sörbom, D. (1993). *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., & Leamer, E. (1984). Consistent sets of estimates for regressions with errors in all variables. *Econometrica*, *52,* 163-183.

Lee, S.-Y. (1981). A Bayesian approach to confirmatory factor analysis. *Psychometrika*, *46,* 153-160.

MacEachern, S.N., & Berliner, L.M. (1994). Subsampling the Gibbs sampler. *The American Statistician*, *48,* 188-190.

Madigan, D., and Raftery, A. E. (1991). *Model selection and accounting for model uncertainty in graphical models using Occam's window* (Technical Report #213). Washington, DC: University of Washington, Department of Statistics.

Meng, X.L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22,* 1142-1160.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*, *49,* 115-132.

Needleman, H., Geiger, S., and Frank, R. (1985). Lead and IQ Scores: A Reanalysis, *Science*, *227,* 701-704.

Press, S.J. (1989). *Bayesian statistics: Principles, models, and applications.* New York: Wiley.

Press, S.J., & Shigemasu, K. (1989). Bayesian inference in factor analysis. In Gleser, L.J., Perlman, M.D., Press, S.J., & Sampson, A.R. (Eds.), *Contributions to probability and statistics: Essays in honor of Ingram Olkin* (pp. 271-287). New York: Springer.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992). *Numerical Recipes in Fortran*. Cambridge: Cambridge University Press.

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 Park, CA: Sage.

Raftery, A. E. (1994). *Bayesian model selection in social research* (Working Paper No. 94-12). University of Washington, Center for Studies in Demography and Ecology.

Raftery, A. E. (1996). Hypothesis testing and model selection. In W.R. Gilks, S. Richardson, & D. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice* (pp. 163-187). London: Chapman & Hall.

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. (1997). Estimating Latent Causal Influence: TETRAD II Model Selection and Bayesian Parameter Estimation. *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*. D. Madigan, ed., January 1997.

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

Scheines, R., Boomsma, A., Hoijtink, H. (1997). *The mulitmodality of the likelihood function in structural equation models* (Technical Report CMU-87-Phil). Pittsburgh, PA: Carnegie Mellon University, Department of Philosophy.

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.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, *22,* 1701-1762.

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.

Zeger, S.L., and, Karim, M.R. (1991). Generalized linear models with random effect; a Gibbs sampling approach. *Journal of the American Statistical Association*, *86,* 79-86.