Maximum likelihood and Bayesian estimation for nonlinear structural equation models

Melanie M Wall
Columbia University

Maximum likelihood and Bayesian estimation for nonlinear structural equation models

Melanie Wall
Email: melanie@biostat.umn.edu,
phone: (612)625-2138, fax: (612)626-0660
Division of Biostatistics
School of Public Health
University of Minnesota
A460 Mayo Building, MMC 303
Minneapolis, MN 55455-0378, USA

December 31, 2007

1 Introduction

Structural equation modeling (SEM) began at its roots as a method for modeling **linear** relationships among latent variables. The well-known software for SEM name LISREL (Jöreskog and Sörbom, 1996) stands for "**Linear** Structural Relations". But, in many cases, the restriction to linearity is not adequate or flexible enough to explain the phenomena of interest. For example, if the slope between two continuous latent variables is directly affected or "moderated" by a third continuous latent variable, this relationship which can be modeled via a cross-product term between the two latent variables, cannot be estimated via the traditional SEM methods. The difficulty is that traditional estimation methods appropriate for fitting linear structural models are focused on minimization of a discrepancy function between the observed and modeled covariance matrix and this cannot be extended in a straightforward way to handle nonlinear structural models. That is, estimation of parameters in a nonlinear structural model cannot be accomplished using only the sample covariance matrix of the observed data.

Kenny and Judd (1984) introduced the first statistical method aimed at producing estimates of parameters in a nonlinear structural equation model (specifically a quadratic or cross-product structural model with a linear measurement model). The basic idea of Kenny and Judd (1984) was to create new "observed variables" by taking products of existing variables and then using them as additional indicators of the nonlinear terms in the model. The method as described by Kenny and Judd (1984) resulted in many tedious constraints on the model covariance matrix. Despite the cumbersome modeling restrictions, the product indicator method of Kenny and Judd (1984) was possible to implement in existing *linear* structural equation modeling software programs (e.g. LISREL).

The idea pioneered by Kenny and Judd (1984) of creating products of observed indicators to serve as new indicators of latent quadratic and latent interaction terms attracted methodological discussions and alterations by a number of papers, including: Hayduk (1987), Ping (1995, 1996a, 1996b,1996c), Jaccard and Wan (1996), Jöreskog and Yang (1996, 1997), Li et al. (1998), several papers within the book edited by Schumacker and Marcoulides (1998), Li et al. (2000), Algina and Moulder (2001), Wall and Amemiya (2001), Moulder and Algina (2002), and Wen et al. (2002). Marsh et al. (2004) give an excellent comparison of these product indicator methods for estimating a structural model with a latent interaction. They categorize the different estimation approaches as: "constrained" using the Algina and Moulder (2001) adjustment to the Jöreskog and Yang (1996) method, "partially constrained" using the GAPI approach of Wall and Amemiya (2001), and "unconstrained" which is newly introduced in the same paper Marsh et al. (2004). The partially and unconstrained methods are shown to produce good parameter estimates even under scenarios in their simulation study where the distribution of the exogenous factors were not normal. Indeed, the product indicator techniques are a workable solution for estimation of simple quadratic or interaction structural models, and the techniques can be implemented in existing linear structural equation modeling software. On the other hand, the rather ad-hoc step of creating new product indicators can not be extended to more general nonlinear models limiting its potential usefulness as a general method.

Given a parametric form for the nonlinear structural equation model and distributional assumptions for the latent variables and errors, it is possible to write down a likelihood function and hence theoretically it should be possible to perform maximum likelihood or Bayesian estimation for the parameters. The problem up until recently has been one of computational difficulty; the nonlinearities in the model create a likelihood which does not have closed analytic form. Over the last 20 years though there have been great advances in the statistical computation methods for maximizing intractable likelihoods and generating from intractable posterior distributions. Building on these computational methods, there is a growing literature focused on using direct maximum likelihood and Bayesian methods for estimation specifically for different forms of nonlinear structural equation models: using full maximum likelihood there is, e.g., Klein, et al. (1997), Klein and Moosbrugger (2000), Amemiya and Zhao (2001), Lee and Zhu (2002), Lee and Song (2003a), and Lee et.al (2003); and using Bayesian methods there is, e.g., Wittenberg and Arminger (1997), Arminger and Muthén (1998), Zhu and Lee (1999), Lee and Zhu (2000), Song and Lee (2002), Lee and Song (2003b), Lee et al. (2007), and Lee (2007).

The implementation of the maximum likelihood and Bayesian methods for a nonlinear structural equation model will be the focus of this chapter. Section 2 and 3 present the linear and nonlinear structural equation model, respectively. Section 4 generally describes maximum likelihood and Bayesian estimation and briefly characterizes some of the statistical computation algorithms useful for implementing them for the nonlinear SEM. Section 5 describes implementation in existing software and Section 6 presents two worked examples of nonlinear SEM's and demonstrates their estimation (with code given) in SAS Proc NLMIXED (for maximum likelihood estimation), Winbugs (for Bayesian estimation), and Mplus (for maximum likelihood estimation specifically for the cross-product). Particular attention is paid to the care needed in the interpretation for the cross-product SEM model. Finally, section 7 is left for discussion.

2 Linear structural equation model

To present the nonlinear structural equation model it is useful to first consider the traditional **linear structural equation model**. Given a vector of p observed variables $\mathbf{Z}_i$ for the i^{th} individual in a sample of size n and a vector of q latent variables $\boldsymbol{f}_i$ such that $\boldsymbol{f}_i = (\boldsymbol{\eta}_i', \boldsymbol{\xi}_i')'$ where $\boldsymbol{\eta}_i$ are the d endogenous latent variables and $\boldsymbol{\xi}_i$ are the $q-d$ exogenous latent variables, the linear structural equation model is:

$$\begin{aligned} \boldsymbol{Z}_i &= \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\boldsymbol{f}_i + \boldsymbol{\epsilon}_i && (1) \\ \boldsymbol{\eta}_i &= \boldsymbol{\gamma}_0 + \boldsymbol{\Gamma}_1\boldsymbol{\eta}_i + \boldsymbol{\Gamma}_2\boldsymbol{\xi}_i + \boldsymbol{\delta}_i && (2) \end{aligned}$$

where in the measurement model (1), the matrices $\boldsymbol{\lambda}_0$ $(p \times 1)$ and $\boldsymbol{\Lambda}$ $(p \times q)$ contain fixed or unknown scalars describing the linear relation between the observations $\boldsymbol{Z}_i$ and the common latent factors $\boldsymbol{f}_i$, and $\boldsymbol{\epsilon}_i$ represents the $(p \times 1)$ vector of random measurement error independent of $\boldsymbol{f}_i$ such that $E(\boldsymbol{\epsilon}_i) = \mathbf{0}$ and $Var(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$ with fixed and unknown scalars in $\boldsymbol{\Psi}$; and in the structural model (2) it is assumed the equation errors $\boldsymbol{\delta}_i$ have $E(\boldsymbol{\delta}_i) = \mathbf{0}$, $Var(\boldsymbol{\delta}_i) = \boldsymbol{\Delta}$ and are independent of the $\boldsymbol{\xi}_i$ as well as independent of $\boldsymbol{\epsilon}_i$ in (1), and the matrices $\boldsymbol{\gamma}_0$ $(d \times 1)$, $\boldsymbol{\Gamma}_1$ $(d \times d)$, $\boldsymbol{\Gamma}_2$ $(d \times (q-d))$, and $\boldsymbol{\Delta}$ $(d \times d)$ are fixed or unknown scalars. Furthermore, it is assumed that the diagonal of $\boldsymbol{\Gamma}_1$ is zero and that $(\mathbf{I} - \boldsymbol{\Gamma}_1)$ is invertible so that the structural model can be solved explicitly for each element of $\boldsymbol{\eta}$. Additionally, a common restriction placed on the measurement model to ensure identifiability is the errors-in-variables parameterization where q of the observed variables are each fixed to be equal to one of the q different latent variables plus measurement error. For a thorough discussion of identifiability in linear structural equation models see e.g. Bollen (1989).

Given an $(m \times 1)$ vector of observed exogenous covariates $\mathbf{X}_i$ for each individual, it is straightforward to extend (1)-(2), to include observed predictors in either or both of the measurement and structural model, i.e.

$$\begin{aligned} \boldsymbol{Z}_i &= \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda} \boldsymbol{f}_i + \boldsymbol{\Lambda}_x \mathbf{X}_i + \boldsymbol{\epsilon}_i & (3) \\ \boldsymbol{\eta}_i &= \boldsymbol{\gamma}_0 + \boldsymbol{\Gamma}_1 \boldsymbol{\eta}_i + \boldsymbol{\Gamma}_2 \boldsymbol{\xi}_i + \boldsymbol{\Gamma}_x \mathbf{X}_i + \boldsymbol{\delta}_i \ . & (4) \end{aligned}$$

Non-zero elements of the $(p \times m)$ matrix $\boldsymbol{\Lambda}_x$ are typically interpreted as an indication of lack of measurement invariance in the way that $\boldsymbol{Z}_i$ measures the latent variables $\boldsymbol{f}_i$. The elements of $\boldsymbol{\Gamma}_x$ represent the regression-type relationship between the observed covariates and the endogenous latent variables after controlling for the the other endogenous and exogenous predictors in the model.

The commonly assumed linear link relationship between observed and latent variables in the measurement model (1) is useful for observed variables $\boldsymbol{Z}_i$ measured on a continuous scale with latent factors $\boldsymbol{f}_i$ hypothesized on a continuous scale. When the p elements of the observed vector $\mathbf{Z}_i$ are not all continuously distributed, the linear model (1) relating the latent factors to the observed variables is not appropriate. Traditionally, “latent trait models” or “item response theory models” (Van Der Linden and Hambleton, 1997) have been used to model continuous latent factors with observed variables that are all ordered categorical. The most common assumption underlying these models is that given the $\boldsymbol{f}_i$, the elements of $\mathbf{Z}_i$ are independent. Muthen (1984) developed a general “underlying variable” or “latent response” framework for fitting structural equation models that involve mixed categorical and continuous observed indicators of the continuous latent variables. Incorporating observed covariates, and using the language of generalized linear models (McCullagh and Nelder (1989)), a generalized linear latent variable model has also been introduced that allows for both continuous and categorical outcomes as measures for latent factors (e.g. Takane and de Leeuw, (1987), Bartholomew and Knott (1999), Sammel et al. (1997), Moustaki and Knott (2000), Moustaki (2003), Skrondal and Rabe-Hesketh (2004), and Huber et al. (2004))

For observed variables $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, \ldots Z_{ip})'$ on an individual i, the **generalized linear structural equation model** relating the observed vector $\mathbf{Z}_i$ to the latent factors $\boldsymbol{f}_i$ is

$$\begin{aligned} P(\mathbf{Z}_i | \boldsymbol{f}_i, \mathbf{X}_i) &= P(\mathbf{Z}_i | \boldsymbol{f}_i, \mathbf{X}_i, \boldsymbol{\Lambda}, \boldsymbol{\Lambda}_x, \boldsymbol{\Psi}) & (5) \\ \boldsymbol{\eta}_i &= \boldsymbol{\gamma}_0 + \boldsymbol{\Gamma}_1 \boldsymbol{\eta}_i + \boldsymbol{\Gamma}_2 \boldsymbol{\xi}_i + \boldsymbol{\Gamma}_x \mathbf{X}_i + \boldsymbol{\delta}_i & (6) \end{aligned}$$

where typically it is assumed $P(\mathbf{Z}_i | \boldsymbol{f}_i, \mathbf{X}_i, \boldsymbol{\Lambda}, \boldsymbol{\Lambda}_x, \boldsymbol{\Psi}) = \prod_{j=1}^{p} P(Z_{ij} | \boldsymbol{f}_i, \mathbf{X}_i, \boldsymbol{\Lambda}, \boldsymbol{\Lambda}_x, \boldsymbol{\Psi})$, and $P(Z_{ij} | \boldsymbol{f}_i, \mathbf{X}_i, \boldsymbol{\Lambda}, \boldsymbol{\Lambda}_x, \boldsymbol{\Psi})$ is a distribution from an exponential family (with $\boldsymbol{\Psi}$ as the scale parameter when appropriate) and can be of different type for each j. For example, if Z_{ij} is a binary random variable then $P(Z_{ij} | \boldsymbol{f}_i, \mathbf{X}_i, \boldsymbol{\Lambda}, \boldsymbol{\Lambda}_x)$ can be taken to be the Bernoulli distribution with success probability given through the inverse logit, i.e. $E(Z_{ij} | \boldsymbol{f}_i, \mathbf{X}_i) = 1/(1 + exp(-(\boldsymbol{\lambda}_{0j} + \boldsymbol{\lambda}_j \boldsymbol{f}_i + \boldsymbol{\lambda}_{xj} \mathbf{X}_i)))$ and no distinct scale parameter. Or, if Z_{ij} is continuously distributed then $P(Z_{ij} | \boldsymbol{f}_i, \mathbf{X}_i, \boldsymbol{\Lambda}, \boldsymbol{\Lambda}_x, \boldsymbol{\Psi})$ can be taken to be Normal with conditional mean given through the linear link, i.e. $E(Z_{ij}|\boldsymbol{f}_i, \mathbf{X}_i) = \lambda_{0j} + \boldsymbol{\lambda}_j \boldsymbol{f}_i + \boldsymbol{\lambda}_{xj}\mathbf{X}_i$ with a distinct scale parameter, $Var(Z_{ij}|\boldsymbol{f}_i, \mathbf{X}_i) = \Psi_j$. Notice that the structural equation model system (5)-(6) may be referred to as nonlinear due to the possible nonlinear link in the measurement model. However, the structural model (6), which express how the latent variables are related to one another, is still linear and thus this model is NOT typically referred to as a nonlinear structural equation model.

3 Nonlinear structural equation model

The linear structural equation model (1)-(2) and its extensions to include covariates (3)-(4) and the generalized linear structural equation model (5)-(6) have been extensively studied and used in the literature. But even the more general model (5)-(6) is still quite limited due to the restriction of linearity in the structural model (6). The **nonlinear structural model** provides a more general formulation of the structural model that allows for nonlinearities in the following way

$$\boldsymbol{\eta}_i = \mathbf{H}(\boldsymbol{\eta}_i, \boldsymbol{\xi}_i, \mathbf{X}_i; \boldsymbol{\Gamma}) + \boldsymbol{\delta}_i \tag{7}$$

where $\mathbf{H}$ is a $(d \times 1)$ vector function with unknown parameters $\boldsymbol{\Gamma}$, and $\boldsymbol{\delta}_i$ is random equation error independent of $\boldsymbol{\xi}_i$ with $E(\boldsymbol{\delta}_i) = 0$ and $Var(\boldsymbol{\delta}_i) = \boldsymbol{\Delta}$ such that $\boldsymbol{\Delta}$ is a $(d \times d)$ matrix of fixed or unknown scalars. Note $\mathbf{H}$ is a function of both $\boldsymbol{\eta}_i$ and $\boldsymbol{\xi}_i$ and so it is assumed that $\mathbf{H}$ is such that there are no elements of $\boldsymbol{\eta}_i$ which are functions of themselves. Furthermore, as described by Wall and Amemiya (2007), in order for the nonlinear structural equation model to be identifiable, it is necessary that (7) can be re-written in "reduced form", i.e. such that endogenous factors are functions only of exogenous factors, observed covariates and errors (i.e. not functions of other endogenous factors). This requirement is similar to that in the linear structural model (2) that $(\mathbf{I} - \boldsymbol{\Gamma}_1)$ be invertible. A wide variety of nonlinear structural models satisfy this form.

A general sub-class of (7) which are identifiable includes models nonlinear in the parameters and recursively nonlinear in the endogenous variables. The recursiveness enables the model to be written in reduced form thus making it identifiable, for example,

$$\eta_{1i} = \gamma_{10} + \gamma_{11} \exp(\gamma_{12}\eta_{2i} + \gamma_{13}\xi_{1i}) + \delta_{1i}. \tag{8}$$

$$\eta_{2i} = \gamma_{20} + \gamma_{21}\xi_{1i} + \delta_{2i} \quad . \tag{9}$$

Notice that (8) represents η_{1i} as a nonlinear function of another endogenous variable, η_{2i}, yet it is possible to rewrite this recursive system in reduced form, so that $\eta_{1i} = \gamma_{10} + \gamma_{11} \exp(\gamma_{12}(\gamma_{20} + \gamma_{21}\xi_{1i} + \delta_{2i}) + \gamma_{13}\xi_{1i}) + \delta_{1i}$ is only a function of exogenous factors and errors. Hence the model can be identified.

A simple but useful sub-class of (7) is

$$\boldsymbol{\eta}_i = \boldsymbol{\gamma}_0 + \boldsymbol{\Gamma}_1\boldsymbol{\eta}_i + \boldsymbol{\Gamma}_2\boldsymbol{\xi}_i + \boldsymbol{\Gamma}_3\mathbf{X}_i + \boldsymbol{\Gamma}_4\mathbf{g}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i \tag{10}$$

where the setup is the same as in the linear case except for the addition of the $(d \times r)$ $\boldsymbol{\Gamma}_4$ matrix of fixed or unknown scalars and the $\mathbf{g}(\boldsymbol{\xi}_i) = (g_1(\boldsymbol{\xi}_i), g_2(\boldsymbol{\xi}_i), \ldots g_r(\boldsymbol{\xi}_i))'$ which represents an $(r \times 1)$ vector function of known nonnlinear functions of the exogenous latent variables. The structural model (10) is accurately described as linear in endogenous variables, additive nonlinear in exogenous variables, and linear in parameters. This class of nonlinear structural model and particularly its special cases of the polynomial and specifically the second order interaction polynomial is the one that has been almost exclusively examined in the literature up to this point.

Some examples of nonlinear structural models that are encompassed by (10) are:
a cubic polynomial model with d = 2, (q-d) = 1, r = 2

$$\begin{aligned} \eta_{1i} &= \gamma_{10} + \gamma_{11}\eta_{2i} + \gamma_{12}\xi_{1i} + \gamma_{13}\xi_{1i}^2 + \gamma_{14}\xi_{1i}^3 + \gamma_{15}X_i + \delta_{1i} \quad (11) \\ \eta_{2i} &= \gamma_{20} + \qquad\qquad \gamma_{21}\xi_{1i} + \gamma_{22}\xi_{1i}^2 + \gamma_{23}\xi_{1i}^3 + \gamma_{24}X_i + \delta_{2i} \quad (12) \end{aligned}$$

and a simple cross-product "interaction" model with d = 1, (q-d) = 2, r = 1

$$\eta_{1i} = \gamma_0 + \gamma_1\xi_{1i} + \gamma_2\xi_{2i} + \gamma_3\xi_{1i}\xi_{2i} + \delta_{1i} \quad . \qquad (13)$$

Given a specific nonlinear structural model (7), the **nonlinear structural equation model** is completed by combining it with one of the measurement models (1), (3), or (5) described above. That is, for individual i, the joint distribution of the observed response vector $\mathbf{Z}_i$ and the random latent variables $\mathbf{f}_i$ conditional on the observed exogenous covariates $\mathbf{X}_i$ and parameters $\boldsymbol{\theta}$ can be written as

$$\begin{aligned} P(\mathbf{Z}_i, \boldsymbol{f}_i | \mathbf{X}_i, \boldsymbol{\theta}) &= P(\mathbf{Z}_i | \boldsymbol{f}_i, \mathbf{X}_i, \boldsymbol{\theta}_m) P(\boldsymbol{f}_i | \mathbf{X}_i, \boldsymbol{\theta}_f) \\ &= P(\mathbf{Z}_i | \boldsymbol{\eta}_i, \boldsymbol{\xi}_i, \mathbf{X}_i, \boldsymbol{\theta}_m) P(\boldsymbol{\eta}_i, \boldsymbol{\xi}_i | \mathbf{X}_i, \boldsymbol{\theta}_f) \\ &= P(\mathbf{Z}_i | \boldsymbol{\eta}_i, \boldsymbol{\xi}_i, \mathbf{X}_i, \boldsymbol{\theta}_m) P(\boldsymbol{\eta}_i | \boldsymbol{\xi}_i, \mathbf{X}_i; \boldsymbol{\theta}_s) P(\boldsymbol{\xi}_i | \mathbf{X}_i, \boldsymbol{\theta}_\xi) \quad (14) \end{aligned}$$

where $\boldsymbol{\theta}_m$ represents the measurement model parameters (i.e. $\boldsymbol{\lambda}_0, \boldsymbol{\Lambda}, \boldsymbol{\Lambda}_x, \boldsymbol{\Psi}$), and $\boldsymbol{\theta}_f$ represents all the parameters governing the distribution of the factors including the structural model parameters $\boldsymbol{\theta}_s$ from the nonlinear structural model (i.e. $\boldsymbol{\Gamma}, \boldsymbol{\Delta}$) and the parameters $\boldsymbol{\theta}_\xi$ describing the distribution of the exogenous factors conditional on the exogenous covariates. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_m, \boldsymbol{\theta}_s, \boldsymbol{\theta}_\xi)$ and note that the three sets of parameters from the three parts of the model are all distinct. Notice that (14) represents a general probabilistic model relating observed and latent variables and encompasses nonlinear SEM, generalized linear SEM, and linear SEM.

In Section 6, estimation via maximum likelihood and the Bayesian method for the nonlinear SEMs with specific structures (8)-(9) and (13) will be demonstrated using existing software.

4 Maximum Likelihood and Bayesian Estimation

In a paper focused on estimation of the second order polynomial structural model with a latent interaction similar to (13), Lee, Song, and Poon (2004) conclude that "At present, it is not convenient for general users to apply the Bayesian or the exact maximum likelihood (ML) approaches [to this model] due to the following reasons: (a) estimates cannot be obtained by existing software, (b) the underlying theory involves unfamiliar recent developments of sophisticated tools in statistical computing, and (c) the computational burdens are heavier than those approaches that can be implemented in user-friendly structural equation modeling programs." While (b) may be true in that the statistical computing tools may not be familiar to everyone, ML and Bayesian methods can now be readily implemented for not only the simple second order nonlinear structural model, but for the more general nonlinear structural model (7). This section generally describes the ML and Bayesian estimation methods and in the next 2 sections their implementation for nonlinear SEM in existing software is demonstrated.

Maximum likelihood estimation (which is a frequentist statistical method) and Bayesian statistical methods are both **model-based** inference methods. That is, given a parametric probabilistic statistical model, like the nonlinear SEM in (14), the ML and Bayesian methods are two ways to find the "best fitting" model from a specific class of models based on a particular dataset. Both the ML and Bayesian methods are focused on obtaining information about $\boldsymbol{\theta}$ and perhaps $\boldsymbol{f}_i$, but how they go about it is somewhat different. Fundamentally, the difference is that the ML method considers $\boldsymbol{\theta}$ to be a fixed unknown constant while the Bayesian method considers $\boldsymbol{\theta}$ to come from some random prior distribution. In a broader context, philosophical debates between the frequentist and Bayesian paradigms have existed for decades in statistics with no one winner, see e.g. Little (2006) for a nice discussion of the spectrum of this debate. Moreover, for the nonlinear SEM, it is not the intention of the current paper to present one method as superior to the other but instead simply to describe them and demonstrate that the major practical hurdle for both - intensive computational algorithms - has been alleviated by some existing softwares.

4.1 Maximum Likelihood method

Given the joint distribution (14) of the observed response vector $\mathbf{Z}_i$ and the random latent variables $\mathbf{f}_i$, the likelihood function (or the "marginal likelihood of $\boldsymbol{\theta}$") associated with an i.i.d. sample $\mathbf{Z}_1 \ldots \mathbf{Z}_n$ with observed covariates $\mathbf{X}_1 \ldots \mathbf{X}_n$ is

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} P(\mathbf{Z}_i|\mathbf{X}_i, \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \int P(\mathbf{Z}_i, \boldsymbol{f}_i|\mathbf{X}_i, \boldsymbol{\theta}) \partial \boldsymbol{f}_i \\ &= \prod_{i=1}^{n} \int P(\mathbf{Z}_i|\boldsymbol{f}_i, \mathbf{X}_i; \boldsymbol{\theta}_m) P(\boldsymbol{f}_i|\mathbf{X}_i, \boldsymbol{\theta}_f) \partial \boldsymbol{f}_i \\ &= \prod_{i=1}^{n} \int P(\mathbf{Z}_i|\boldsymbol{\eta}_i, \boldsymbol{\xi}_i, \mathbf{X}_i, \boldsymbol{\theta}_m) P(\boldsymbol{\eta}_i|\boldsymbol{\xi}_i, \mathbf{X}_i; \boldsymbol{\theta}_s) P(\boldsymbol{\xi}_i|\mathbf{X}_i, \boldsymbol{\theta}_\xi) \partial \boldsymbol{\xi}_i \qquad (15) \end{aligned}$$

Note that the likelihood is not a function of the random latent variables $\mathbf{f}_i$ since they are "marginalized" out of the expression through the integral. Furthermore, note that the likelihood is a function of the fixed parameters $\boldsymbol{\theta}$ and the goal of the maximum likelihood procedure is to find the single value of $\boldsymbol{\theta}$ that maximizes the function, in other words, the value which "most likely" generated the given the data.

In the special case when the measurement and structural models, $P(\mathbf{Z}_i|\boldsymbol{\eta}_i, \boldsymbol{\xi}_i, \mathbf{X}_i, \boldsymbol{\theta}_m)$, and $P(\boldsymbol{\eta}_i|\boldsymbol{\xi}_i, \mathbf{X}_i, \boldsymbol{\theta}_s)$ are both linear as in (3)-(4), and the errors, $\boldsymbol{\epsilon}_i$, $\boldsymbol{\delta}_i$, and exogenous factors

$\boldsymbol{\xi}_i$ are assumed to be normally distributed, then the joint distribution (14) is multivariate normal. This means that when $\boldsymbol{\xi}_i$ is marginalized (integrated) out in (15), the distribution of $P(\mathbf{Z}_i|\mathbf{X}_i, \boldsymbol{\theta})$ is also simply multivariate normal. Thus, the likelihood is simply the joint distribution of i.i.d. multivariate normal variables which is a closed form analytic function of the observed sample mean and covariance matrix along with the modeled mean and covariance matrix. In other words, the traditional linear structural equation model leads to a nice closed form for (15). It is this closed form multivariate normal likelihood that has been the backbone of linear SEM.

On the other hand, if either the measurement or structural model have nonlinear relationships in their conditional means or else the underlying exogenous factors or error terms are not normally distributed, then the integral in (15) will no longer have an analytic solution in general. Herein lies the difficulty since in this case it is necessary to contend with the integral while trying to maximize the likelihood. Generally, this is not a simple numerical computational task with one clear best, most accurate, fastest, solution. But fortunately several modern statistical computational techniques have been developed particularly suited to this sort of likelihood function involving possibly multidimensional integration over latent quantities. Two general classes of computational methods for addressing this are: to approximate the integral in the likelihood (15) or to sidestep the integral in (15) by employing the expectation maximization (EM) algorithm (Dempster et al. 1977). For a more comprehensive look at the details of different computational methods for performing maximum likelihood see, e.g. Skrondal and Rabe-Hesketh (2004) Chapter 6.

One class of computational techniques are based on direct approximation to the integrated likelihood. In the case when the exogenous factors $P(\boldsymbol{\xi}_i|\mathbf{X}_i, \boldsymbol{\theta}_\xi)$ and the nonlinear structural model equation errors $\boldsymbol{\delta}_i$ can be assumed to be normally distributed, the integral in (15) can be approximated by an adaptive Gaussian quadrature method. Then given a closed form approximation to the integral, the likelihood can be approximated in a closed form. With the closed-form approximation for the likelihood, the maximization of it can be carried out through a quasi-Newton algorithm. Much statistical work has been done for comparing computational methods using different approximations to the integral for maximum likelihood estimation of nonlinear mixed effects models (e.g. Pinheiro and Bates 1995). The nonlinear SEM can be considered a kind of nonlinear mixed effects model (Patefield 2002), hence the computation techniques relevant for ML estimation in nonlinear mixed effects models are applicable to nonlinear SEM.

It is possible to consider the latent variables $\mathbf{f}_i$ as missing data and hence this suggests the use of the Expectation Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) for maximum likelihood estimation. Notice that if we were able to observe the latent variables (i.e. if they were not missing), the maximum likelihood estimation of $\boldsymbol{\theta}$ would be very straightforward. The so-called "complete data likelihood" treats the $\boldsymbol{f}_i$ as if they were observed and is taken as

$$L_{\text{complete}}(\boldsymbol{\theta}) \;=\; \prod_{i=1}^{n} P(\mathbf{Z}_i, \boldsymbol{f}_i|\mathbf{X}_i, \boldsymbol{\theta}) \,. \tag{16}$$

The basic idea of the EM algorithm is that rather than maximize the likelihood (15) directly (often referred to as the "observed data likelihood"), instead, the EM algorithm iteratively maximize the expected conditional "complete data log likelihood" conditional on the observed data and the most recent estimates of the parameters. At each iteration step the expected conditional complete data log likelihood is formed (E-step) and it is maximized with respect to $\boldsymbol{\theta}$ (M-step). Then the new "estimate" of $\boldsymbol{\theta}$ is used in the next iteration to again form the E-step and this new conditional expectation function is then maximized again. This procedure is continued until the new "estimate" of $\boldsymbol{\theta}$ is within some very small increment from the previous estimate. The estimate of $\boldsymbol{\theta}$ that results from convergence of the EM algorithm has been proven to be the maximum likelihood estimator, i.e. the value that maximizes the "observed data likelihood" (15).

If there is a nice closed form for the E-step (i.e. the expected conditional complete data log likelihood) then the algorithm is usually straightforward because "nice forms" usually can be maximized pretty easily. But because of the necessarily nonnormal distribution of $\boldsymbol{\eta}_i$ arising from any nonlinear function in the structural model (7), difficulty arises in the integration of the E-step since no closed form is available. Klein, et al. (1997) and Klein and Moosbrugger (2000) proposed a mixture distribution to approximate the nonnormal distribution arising specifically for the interaction model (13) and used this to adapt the EM algorithm to produce maximum likelihood estimators in that special case.

Stochastic versions of the EM algorithm have been implemented for more general forms of the nonlinear structural equation model than just the interaction. Briefly we list some recent works in these methods. Taking the distribution of $P(\boldsymbol{\xi}_i; \boldsymbol{\theta}_\xi)$ to be normally distributed, Amemiya and Zhao (2001) performed maximum likelihood for the general nonlinear model using the Monte Carlo EM algorithm. Lee and Zhu (2002) addressed the intractable E-step by using the Metropolis-Hastings algorithm and conditional maximization in the M-step. This same computational framework for producing ML estimates was then used by Lee et al. (2003) in the case of ignorably missing data. The method was then further extended to the case where the observed variables $\mathbf{Z}_i$ may be both continuous or polytomous (Lee and Song, 2003a) assuming the underlying variable structure with thresholds relating the polytomous items to the continuous factors.

Once the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is obtained from any of the computational methods above, the estimate of the asymptotic covariance matrix (and hence the standard errors) can be obtained from the observed information matrix (i.e. the negative inverse of the Hessian of the log-likelihood evaluated at $\widehat{\boldsymbol{\theta}}$). The Hessian is often straightforward to obtain as a by-product of the maximum likelihood procedure.

From a frequentist perpsective there is a distinction between estimation and prediction. Fixed parameters are **estimated** and unknown random quantities are **predicted**. Since the latent variables in the nonlinear structural equation model (14) are taken as random quantities from a distribution $P(\boldsymbol{f}_i|\mathbf{X}_i, \boldsymbol{\theta}_f)$, only the parameters $\boldsymbol{\theta}_f$ governing their distribution are estimated by maximum likelihood, not the quantities for $\boldsymbol{f}_i$ themselves. But, as random unknown quantities, prediction of the $\boldsymbol{f}_i$ can be performed using the expected conditional mean of $\boldsymbol{f}_i$ given the data and the resulting MLE's $\widehat{\boldsymbol{\theta}}$, i.e.

$$\hat{\boldsymbol{f}}_i = E(\boldsymbol{f}_i|\mathbf{Z}, \mathbf{X}, \widehat{\boldsymbol{\theta}}) \quad . \tag{17}$$

The values (17) are called "empirical Bayes" predictions and are not to be confused with the Bayesian method described in the next section. The reason they have Bayes in the name is that Bayes rule (i.e. $P(A|B) = [P(B|A)P(A)]/P(B)$) is used to form the conditional probability of $P(\boldsymbol{f}_i|\mathbf{Z}, \mathbf{X}, \widehat{\boldsymbol{\theta}})$. The empirical Bayes predicted values or "factor score estimates" are a function only of data (since the MLE $\widehat{\boldsymbol{\theta}}$ is a function only of data) and not of any "prior" distribution for $\boldsymbol{\theta}$.

4.2 Bayesian method

In the maximum likelihood method above, the elements of $\boldsymbol{\theta}$ were considered fixed parameters in the population and $\mathbf{f}_i$ were random latent variables coming from some distribution $P(\mathbf{f}_i|\mathbf{X}_i, \boldsymbol{\theta}_f)$. As random latent variables, the $\mathbf{f}_i$ were not explicitly estimated in the ML procedure, in particular, they were "marginalized out" by integration. Once the ML estimator for $\boldsymbol{\theta}$ was obtained, the predicted values for the $\mathbf{f}_i$ could be obtained as in (17). In the Bayesian method there need not be any distinction between random latent variables and parameters; all unobserved quantities can be considered parameters and all parameters are considered random. That is, in the Bayesian method, the $\mathbf{f}_i$ are also considered parameters and the parameter $\boldsymbol{\theta}$ is assigned a distribution called a prior distribution $P(\boldsymbol{\theta})$.

Thus, now given the addition of a prior distribution $P(\boldsymbol{\theta})$, we extend the joint distribution of the nonlinear structural equation model in (14) - which was conditional on $\boldsymbol{\theta}$ - into the fully Bayesian joint model of the data and the parameters, i.e.,

$$\begin{aligned} P(\mathbf{Z}_i, \boldsymbol{f}_i, \boldsymbol{\theta}|\mathbf{X}_i) &= P(\mathbf{Z}_i, \boldsymbol{f}_i|\mathbf{X}_i, \boldsymbol{\theta})P(\boldsymbol{\theta}) \\ &= P(\mathbf{Z}_i|\boldsymbol{f}_i, \mathbf{X}_i, \boldsymbol{\theta}_m)P(\boldsymbol{f}_i|\mathbf{X}_i, \boldsymbol{\theta}_f)P(\boldsymbol{\theta}) \\ &= P(\mathbf{Z}_i|\boldsymbol{\eta}_i, \boldsymbol{\xi}_i, \mathbf{X}_i, \boldsymbol{\theta}_m)P(\boldsymbol{\eta}_i|\boldsymbol{\xi}_i, \mathbf{X}_i, \boldsymbol{\theta}_s)P(\boldsymbol{\xi}_i|\mathbf{X}_i, \boldsymbol{\theta}_\xi)P(\boldsymbol{\theta}_m, \boldsymbol{\theta}_s, \boldsymbol{\theta}_\xi). \quad (18) \end{aligned}$$

The model (18) may be referred to as a "hierarchical Bayesian model" signifying that some parameters depend in turn on other parameters. Because the latent variables $\mathbf{f}_i = (\boldsymbol{\eta}_i, \boldsymbol{\xi}_i)$ are now considered parameters, and since they are dependent on other "higher level" parameters $\boldsymbol{\theta}$ which are more appropriately called "hyperparameters" the model is so-called "hierarchical". The hierarchical description is meant to reflect the interdependence of randomness at different levels. Specifically, the observed variables $\mathbf{Z}_i$ are dependent on the parameters $\mathbf{f}_i = (\boldsymbol{\eta}_i, \boldsymbol{\xi}_i)$ and $\boldsymbol{\theta}_m$, and in turn, the parameters $\boldsymbol{\eta}_i$ are dependent on parameters $\boldsymbol{\xi}_i$ and $\boldsymbol{\theta}_s$, and the parameters $\boldsymbol{\xi}_i$ are dependent on parameters $\boldsymbol{\theta}_\xi$. Finally, the hyperparameter $\boldsymbol{\theta} = (\boldsymbol{\theta}_m, \boldsymbol{\theta}_s, \boldsymbol{\theta}_\xi)$ in turn has its own hyperprior distribution $P(\boldsymbol{\theta}_m, \boldsymbol{\theta}_s, \boldsymbol{\theta}_\xi)$ not dependent on anything else and typically fully specified by the user. As mentioned in the the introduction to this section, the fundamental distinction between the ML approach and the Bayesian approach is the reliance on a prior distribution for $\boldsymbol{\theta}$.

Skrondal and Rabe-Hesketh (2004) Chapter 6.11.4 nicely describes different motivations for choosing prior distributions. The first motivation which is the one that can be considered the 'truly Bayesian' motivation is to specify a prior distribution for the parameters that reflects informed knowledge based on past experience about the parameter. The other motivations are what Skrondal and Rabe-Hesketh refer to as "pragmatic". The prior can be specified to ensure that estimates are constrained to be within the parameter space, they can be specified to aid identification, and probably most commonly, they can be specified as non-informatively as possible simply to provide a mechanism for generating a posterior distribution that will mostly be governed by the likelihood and have negligible influence from the prior. Returning to the prior $P(\boldsymbol{\theta}_m, \boldsymbol{\theta}_s, \boldsymbol{\theta}_\xi)$ in (18) for the nonlinear structural equation model, there are several possible ways to specify this joint prior, but the most straightforward is to assume independence among all the parameters and specify the typical pragmatic non-informative distributions to the specific elements. That is, we assume $P(\boldsymbol{\theta}_m, \boldsymbol{\theta}_s, \boldsymbol{\theta}_\xi) = P(\boldsymbol{\lambda}_0)P(\boldsymbol{\Lambda})P(\boldsymbol{\Lambda}_x)P(\boldsymbol{\Psi})P(\boldsymbol{\Gamma})P(\boldsymbol{\Delta})P(\boldsymbol{\theta}_\xi)$ and we choose highly variable normal distributions for the "regression coefficient" type parameters, i.e. $\boldsymbol{\lambda}_0$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Lambda}_x$, and $\boldsymbol{\Gamma}$ and disperse inverse gamma distributions for the "variance" parameters $\boldsymbol{\Psi}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\theta}_\xi$.

The main target then of Bayesian inference is the posterior distribution of the parameters given the observed data $\mathbf{Z} = (\mathbf{Z}_1 \ldots \mathbf{Z}_n)$ and $\mathbf{X} = (\mathbf{X}_1 \ldots \mathbf{X}_n)$. The posterior is obtained by Bayes rule. It is possible to focus specifically on the posterior for just the parameters $\boldsymbol{\theta}$, i.e.,

$$P(\boldsymbol{\theta}|\mathbf{Z}, \mathbf{X}) \quad = \quad \frac{\prod_{i=1}^n \int P(\mathbf{Z}_i|\boldsymbol{f}_i, \mathbf{X}_i, \boldsymbol{\theta}_m)P(\boldsymbol{f}_i|\mathbf{X}_i, \boldsymbol{\theta}_f)\partial \boldsymbol{f}_i P(\boldsymbol{\theta})}{\int \prod_{i=1}^n \int P(\mathbf{Z}_i|\boldsymbol{f}_i, \mathbf{X}_i, \boldsymbol{\theta}_m)P(\boldsymbol{f}_i|\mathbf{X}_i, \boldsymbol{\theta}_f)\partial \boldsymbol{f}_i P(\boldsymbol{\theta})\partial\boldsymbol{\theta}} \tag{19}$$

or following the data augmentation idea of Tanner and Wong (1987) it may be useful to include the latent variables into consideration as parameters and focus on the posterior jointly for $\boldsymbol{f}$ and $\boldsymbol{\theta}$, i.e.,

$$P(\boldsymbol{f}_1 \ldots \boldsymbol{f}_n, \boldsymbol{\theta}|\mathbf{Z}, \mathbf{X}) \quad = \quad \frac{\prod_{i=1}^n P(\mathbf{Z}_i|\boldsymbol{f}_i, \mathbf{X}_i, \boldsymbol{\theta}_m)P(\boldsymbol{f}_i|\mathbf{X}_i, \boldsymbol{\theta}_f)P(\boldsymbol{\theta})}{\int\int \prod_{i=1}^n P(\mathbf{Z}_i|\boldsymbol{f}_i, \mathbf{X}_i, \boldsymbol{\theta}_m)P(\boldsymbol{f}_i|\mathbf{X}_i, \boldsymbol{\theta}_f)P(\boldsymbol{\theta})\partial\boldsymbol{f}\partial\boldsymbol{\theta}} \quad . \tag{20}$$

Notice that one way to interpret the numerator of (19) is that it is equal to the likelihood (15) times the prior for $\boldsymbol{\theta}$. Similarly, the numerator of (20) can be seen as the complete data likelihood (16) times the prior for $\boldsymbol{\theta}$. Or another way to see the numerator of (20) is as the the likelihood for $\boldsymbol{f}_i$ and $\boldsymbol{\theta}_m$ (i.e. $P(\mathbf{Z}_i|\boldsymbol{f}_i, \mathbf{X}_i, \boldsymbol{\theta}_m)$) times the prior for $\boldsymbol{f}_i$, $P(\boldsymbol{f}_i|\mathbf{X}_i, \boldsymbol{\theta}_f)$, times the prior for $\boldsymbol{\theta}$. Computationally it is often more straightforward to deal with (20) rather than (19) because it does not include an integral in the numerator (similar to the contrast between using the EM algorithm with (16) rather than maximizing the likelihood (15) directly). When describing the Bayesian method, it is common to point out that the posterior distribution updates prior "knowledge" about parameters using information in the observed data found through the likelihood. That is, depending on which likelihood is being described, we can loosely say that the posterior distribution is proportional to the likelihood times the prior(s). Note the denominator of the posterior is a constant ("normalizing constant"), hence the previous statement is proportional rather than equal. The posterior distribution contains all relevant information about the unknown parameters and so summaries of the posterior distribution are the main focus of Bayesian inference.

Bayesian inference is based entirely on the posterior distribution and summaries of it. But for many models beyond just the basic ones, there is not a closed analytic form for the posterior, so calculation of expected means and quantiles cannot be done directly. Bayesian computation, i.e. computing quantities from posterior distributions, is a well-developed and still actively growing research field in statistics (Carlin and Louis 2000, Chapter 5). Analytic approximations to the posterior, e.g. via Laplace's method (Tierney and Kadane, 1986) and numerical integration methods (e.g. Geweke 1989) based on conventional Monte Carlo integration have been developed. But the real explosion of applications of Bayesian methods followed the advent of Markov Chain Monte Carlo (MCMC) methods for drawing samples from the joint posterior distribution via the Metropolis-Hastings algorithm (Hastings 1970) and its special case the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith 1990). Details of these MCMC methods are described in standard Bayesian textbooks, e.g. Carlin and Louis (2000) and Congdon (2001), and specifically for latent variable models in Lee (2007) and Skrondal and Rabe-Hesketh (2004) (Chapter 6.11).

MCMC methods are particularly useful for sampling from distributions that entail multidimensional integrals. The basic idea of MCMC is to sample from a distribution by constructing a Markov chain that has the desired distribution as its equilibrium distribution. Rather than sampling directly from the joint posterior distribution, the MCMC methods sample the conditional posteriors of the individual parameters conditional on the last sampled value of all the other parameters and the data. These full conditional distributions often have forms that can be simulated from straightforwardly. Samples are then drawn iteratively from the chain and after a sufficiently large number of iterations (say T) when the chain has converged to its equilibrium distribution (in this case the joint posterior), the continued draws from the chain represent simulated "observations" of the parameters from the posterior. Then, by continuing to take a large number of additional samples from the chain after it has converged (at iteration T), a simulated (empirical) sample of the posterior distribution is produced and can be used to perform any desired inference. Typically the expected mean of the posterior is computed by taking the empirical mean of the MCMC samples and is treated as the Bayesian "estimate" of the parameter. Similarly, the standard deviation of the posterior samples is the standard error and quantiles can be calculated corresponding to some desired credible intervals.

The T draws from the chain that are needed to allow the Markov chain to reach its equilibrium at the joint posterior are discarded and are often referred to as the "burn in" samples. Before convergence, the draws do not represent samples from the joint posterior and thus are not useful to keep. There are recommendations for monitoring the convergence of the chain in order to know how big T should be (Gelman 1996) but there is no one best solution. A common technique is to generate multiple chains with different starting values and decide that convergence has occurred when the chains (which all started at different places) are mixed well together indicating they have reached the same equilibrium distribution.

5 Implemention in existing software

The original commercial softwares developed specifically for SEM, e.g. LISREL (Jöreskog and Sörbom, 1996), AMOS (Arbuckle, 1995), EQS (Bentler, 1985), SAS Proc CALIS (SAS, 2002) were all developed for linear structural equation models of the form (1)-(2) or (3)-(4). In all of these softwares, maximum likelihood estimation is performed assuming $P(\boldsymbol{Z}|\boldsymbol{f}, \mathbf{X}, \boldsymbol{\theta}_m)$ and $P(\boldsymbol{f}|\mathbf{X}, \boldsymbol{\theta}_f)$ with linear links are multivariate normally distributed. In addition, different forms of least squares estimation based on the covariance matrix are available in these softwares. For the specific case when the observed variables $\mathbf{Z}$ are ordered categorical, the generalized linear SEM (5)-(6) is often seen in the literature referred to as a "structural equation model with categorical variables" and can also be fit using the original linear SEM software, via the "underlying variable" approach that utilizes polychoric correlations (Muthén 1984). This limited information approach can result in biased estimators when the data generating distribution is far from the normal one assumed for polychoric correlations

(DiStefano (2002), Huber et al (2004)). Full maximum likelihood estimation for the generalized linear SEM (5)-(6) with observed variables from any exponential family (Moustaki and Knott (2000), Rabe-Hesketh et al (2002)) requires more advanced computational algorithms that handle numerical integration like those described in Section 4.1 and these have been implemented in the highly flexible latent variable modeling softwares Mplus (Muthén and Muthén, 2007) and STATA's GLLAMM (Rabe-Hesketh et al 2004).

But what about existing software for nonlinear SEM? For this we turn to what might be best described as general flexible modeling statistical software rather than programs designed specifically for SEM. SAS Proc NLMIXED (SAS Version 8.0 or later) and Winbugs (Spiegelhalter et al, 2002) are highly flexible computational softwares that are equipped to perform, respectively, maximum likelihood or Bayesian estimation for general nonlinear models including random latent variables. Both of these softwares can be used to perform estimation for nonlinear SEM with very general functional forms of the nonlinearity in (7). The method of Gaussian quadrature approximation followed by quasi-Newton maximization can be implemented in PROC NLMIXED and the MCMC method for Bayesian posterior inference is implemented in Winbugs. In SAS NLMIXED it is necessary in the model to specify the exogenous factors and structural equation errors to be normally distributed, while in Winbugs all the factors and errors in the model can be specified to be almost any distribution. Both softwares take longer to run as the number of latent factors increases and both are expected to be less numerically precise in terms of assessing convergence as the number of factors increase. This is indeed a limitation of these current computational statistical methods, not a limitation of the softwares to implement them. The use of stochastic EM algorithms for maximum likelihood and the use of approximation methods for Bayesian analysis both mentioned in Section 4 may lead to improvements in numerical stability of estimation but at present these computational methods have not been implemented directly in any convenient software (to the knowledge of the author).

For the special case of the second-order interaction structural model (13), the adapted EM algorithm method of Klein and Moosbrugger (2000) is implemented for maximum likelihood estimation in Mplus version 3.0 and later. Similar to Proc NLMIXED, in the Mplus implementation, it is necessary to specify the exogenous factors and structural equation errors to be normally distributed. Furthermore, as mentioned in the introduction there are product indicator methods which can be used for the special case of the second-order interaction structural model using existing linear SEM programs. But, while the linear SEM softwares employ maximum likelihood, the actual likelihood being maximized when using the product indicator method is not (15) but instead an ad-hoc constrained version of a linear model likelihood. Those methods will not be considered further here.

Both Proc NLMIXED and Mplus (for the interaction model) will produce empirical Bayes factor score estimates (17). In Winbugs, the estimates of the latent factors are easily obtained since the user can specify which of the posterior distributions of the parameters (including the underlying factors) should be displayed in the output.

Finally, it is important to comment on the need for starting values. All of the statistical computation methods described in Section 4 either for maximizing a likelihood or generating Bayesian posterior samples require initial values for the parameter estimates being sought. As the models become more complex (as in the nonlinear structural equation model), providing good starting values is often very important to facilitate the algorithm reaching convergence.

SAS Proc NLMIXED and Mplus will provide default starting values of $\boldsymbol{\theta}$ if the user does not give them. The default starting values used by Mplus are constructed with the specific function of SEM parameters in mind, for example, for the measurement error variance parameters $\boldsymbol{\Psi}$, it uses half the sample variance of the observed variables. In contrast, Proc NLMIXED by default sets all initial parameter values to 1. In the Bayesian setting, recall that both $\boldsymbol{\theta}$ and $\boldsymbol{f}_i$ are parameters. In Winbugs, the user must specify starting values for $\boldsymbol{\theta}$, but the starting values for $\boldsymbol{f}_i$ can be chosen by the software which is helpful since the number of $\boldsymbol{f}_i$ increases with the sample size.

6 Demonstrations of fitting nonlinear SEM

In this section we will demonstrate the use of SAS Proc NLMIXED and Winbugs for a very general nonlinear structural equation model, one that it is nonlinear in parameters and nonlinear in endogenous variables (8)-(9), and we will demonstrate their use along with Mplus for the ubiquitous interaction SEM (13). Implentation of Mplus for a similar interaction model can also be seen in the Mplus Userguide (Muthen and Muthen 2007) Example 5.13. A similar implementation of Winbugs for the second-order cross product model has been previously demonstrated by Lee et al. (2007) and Lee (2007) Chapter 8. In a slightly different set-up, the implentation of PROC NLMIXED for nonlinear SEM has been previously demonstrated by Patefield (2002).

6.1 Nonlinear SEM Example 1 - Nonlinear in parameters and nonlinear in endogenous variables

We generate n=500 independent observations from the nonlinear structural model (8)-(9) where the three latent factors are measured via a linear measurement model (1) with simple structure and 3 observed variables $\boldsymbol{Z}$ for each of the factors (resulting in 9 observed variables). The exogenous factor ξ_{1i}, and errors $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\delta}_i$ are generated as normal variates. Specifically, the nonlinear SEM example is shown below with parameter names denoted. True parameter values are shown in Table 1.

$$\begin{aligned}
\begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \\ Z_6 \\ Z_7 \\ Z_8 \\ Z_9 \end{pmatrix} &= \begin{pmatrix} \lambda_{01} \\ \lambda_{02} \\ \lambda_{03} \\ \lambda_{04} \\ \lambda_{05} \\ \lambda_{06} \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \Lambda_{11} & 0 & 0 \\ \Lambda_{21} & 0 & 0 \\ 0 & \Lambda_{32} & 0 \\ 0 & \Lambda_{42} & 0 \\ 0 & 0 & \Lambda_{53} \\ 0 & 0 & \Lambda_{63} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \eta_1 \\ \eta_2 \end{pmatrix} + \boldsymbol{\epsilon} & (21) \\
diag(Var\boldsymbol{\epsilon}) &= (\Psi_1, \Psi_2, \Psi_3, \Psi_4, \Psi_5, \Psi_6, \Psi_7, \Psi_8, \Psi_9) & (22) \\
E(\xi_1) &= \mu_{\xi_1}, \quad Var(\xi_1) = \phi_1 & (23) \\
\eta_1 &= \gamma_{10} + \gamma_{11} \exp(\gamma_{12}\eta_2 + \gamma_{13}\xi_1) + \delta_1 & (24) \\
\eta_2 &= \gamma_{20} + \gamma_{21}\xi_1 + \delta_2 & (25)
\end{aligned}$$

The model being fit to this data using SAS Proc NLMIXED and Winbugs will be the same nonlinear structural model that generated the data with the same simple structure linear measurement model with correctly specified zero elements. Generated data and computer code (as shown below and in Appendix A) for both SAS Proc NLMIXED and Winbugs are also available at
http://www.biostat.umn.edu/∼melanie/NONLINEARSEM/index.html.

Here is the code for SAS PROC NLMIXED

```
data a; infile "C:data1forsas.txt"; input dummy id z1-z9; run;

proc nlmixed data = a;

parms psi1 = .30 psi2 = .06 psi3 = .129 psi4=.53 psi5 = .24 psi6 = .15
psi7 = .64 psi8 = .80 psi9 = .96;

****** Specify the CONDITIONAL means of each observed variables
z1-z9 given the random latent variables ksi1, eta1, eta2;

     mu1   = lam10+ lam11*ksi1   ;
     mu2   = lam20+ lam21*ksi1   ;
     mu3   = lam30+ lam32*eta1   ;
     mu4   = lam40+ lam42*eta1   ;
     mu5   = lam50+ lam53*eta2   ;
     mu6   = lam60+ lam63*eta2   ;
     mu7   =              ksi1   ;
     mu8   =              eta1   ;
     mu9   =              eta2   ;

******** Specify the nonlinear structural model;

 eta2 = gam20 + gam21*ksi1 + delta2;
 eta1 = gam10 + gam11*exp(gam12*eta2 + gam13*ksi1) + delta1;

****** Write out the log of the joint distribution of the observed
data z1-z9 CONDITIONAL on the random factors ksi1, eta1, eta2;

partofloglike = -.5*log(ps1) - (1/(2*ps1)) * (z1 - mu1)**2
                -.5*log(ps2) - (1/(2*ps2)) * (z2 - mu2)**2
                -.5*log(ps3) - (1/(2*ps3)) * (z3 - mu3)**2
                -.5*log(ps4) - (1/(2*ps4)) * (z4 - mu4)**2
                -.5*log(ps5) - (1/(2*ps5)) * (z5 - mu5)**2
                -.5*log(ps6) - (1/(2*ps6)) * (z6 - mu6)**2
                -.5*log(ps7) - (1/(2*ps7)) * (z7 - mu7)**2
                -.5*log(ps8) - (1/(2*ps8)) * (z8 - mu8)**2
                -.5*log(ps9) - (1/(2*ps9)) * (z9 - mu9)**2;
```

```
model dummy ~ general(partofloglike);

******** Specify the exogenous random terms in the latent factor distribution;

random ksi1 delta1 delta2 ~ normal([muksi1, 0, 0],
                                   [phi1,
                                       0, ddelta1,
                                       0,       0,   ddelta2]) subject = id;
bounds ps1-ps9>=0, phi1>=0, ddelta1-ddelta2>=0;
run;
```

The *parms* statement sets the initial value of the specified parameters. Here the $\mathbf{\Psi}$ parameters are set to be equal to half of the sample variance of the respective observed variables. All other parameter starting values are set to 1 by default when not specified by the user. The *dummy* variable listed on the left-hand side of the *model* statement is fixed at 1 for all observations and is just used as a place holder since all the data z1-z9 are already in the *partofloglike* statement. The *general* function syntax requires there to be some variable name on the left-hand side of the tilde, hence the inclusion of *dummy.* The form of the *partofloglike* comes from the linear measurement model which assumes that conditional on the factors, the 9 observations are uncorrelated and normally distributed. We emphasize that this does not mean the observations are normally distributed (they are certainly not normal because of the nonlinear structural model), but that conditionally they are normally distributed, in other words, the measurement errors $\boldsymbol{\epsilon}_i$ are normally distributed, with diagonal $\mathbf{\Psi}$ matrix. It is possible to specify other nonlinear link distributions by writing out their respective distributional forms.

Here is the code for the model in Winbugs (also see Appendix A):

```
model{

for (i in 1:500){

#Specify the measurement model
 z[i,1] ~dnorm(mu[i,1],psiinv[1])
 z[i,2] ~dnorm(mu[i,2],psiinv[2])
 z[i,3] ~dnorm(mu[i,3],psiinv[3])
 z[i,4] ~dnorm(mu[i,4],psiinv[4])
 z[i,5] ~dnorm(mu[i,5],psiinv[5])
 z[i,6] ~dnorm(mu[i,6],psiinv[6])
 z[i,7] ~dnorm(mu[i,7],psiinv[7])
 z[i,8] ~dnorm(mu[i,8],psiinv[8])
 z[i,9] ~dnorm(mu[i,9],psiinv[9])
     mu[i,1] <- lam0[1]+ lam1[1]*ksi1[i]
     mu[i,2] <- lam0[2]+ lam1[2]*ksi1[i]
     mu[i,3] <- lam0[3]+ lam1[3]*eta1[i]
     mu[i,4] <- lam0[4]+ lam1[4]*eta1[i]
     mu[i,5] <- lam0[5]+ lam1[5]*eta2[i]
     mu[i,6] <- lam0[6]+ lam1[6]*eta2[i]
     mu[i,7] <-              ksi1[i]
     mu[i,8] <-              eta1[i]
     mu[i,9] <-              eta2[i]

#Specify the nonlinear structural model
eta2[i] <- gam[1] + gam[2]*ksi1[i] + delta2[i]
eta1[i] <- gam[3] + gam[4]*exp(gam[5]*eta2[i] + gam[6]*ksi1[i]) + delta1[i]

#Specify the random parts of the latent factor distributions
ksi1[i]~dnorm(muksi,phi1inv)
delta1[i]~dnorm(0,ddelta1inv)
delta2[i]~dnorm(0,ddelta2inv)
}

###priors for Psi
 for (t in 1:9){
    psiinv[t]~dgamma(.001,.001)
    psi[t]<- 1/psiinv[t]}

####priors for lam0 and lam1
 for (k in 1 : 6) {
    lam0[k] ~ dnorm(0.0, 0.0001)
    lam1[k] ~ dnorm(0.0, 0.0001)}

####priors for gamma
for (j in 1 : 6) {
    gam[j] ~ dnorm(0.0, 0.0001)}

####priors for muksi and phi1 and ddelta1 ddelta2
muksi ~ dnorm(0.0, .0001)

phi1inv~dgamma(.001,.001)
phi1<-1/phi1inv

ddelta1inv~dgamma(.001,.001)
ddelta1<-1/ddelta1inv

ddelta2inv~dgamma(.001,.001)
ddelta2<-1/ddelta2inv
}
```

Notice that Winbugs parameterizes in terms of the precision (i.e. the inverse of the variance) in the *dnorm* function which explains the use of inverses for specifying the variances. In addition, for Bayesian inference it is necessary to specify prior distributions for all the model parameters. A normal prior with very large variance (e.g. *dnorm(0,.0001)*) is a typical "non-informative" prior for regression-type coefficients and the (*dgamma(.001,.001)*) prior on the precision leads to a diffuse "non-informative" prior on the variances.

In SAS, once the data is read in, the program is ran simply by executing the code above. Winbugs uses a point-and-click interface for specifying the different parts of the model including the details of the MCMC computations. A detailed step-by-step outline for executing the model in Winbugs is given in Appendix A. To fit this one dataset with this model, SAS Proc NLMIXED took 3 minutes and 57 seconds, and Winbugs took 7 minutes and 22 seconds (for 14,000 MCMC iterations) on the same laptop computer with Intel Core 2CPU, 2.16 GHz processor. Table 1 presents the true values for the parameters used in generating data for this example and the resulting maximum likelihood estimates and standard errors from SAS Proc NLMIXED, and the Bayesian posterior means and standard errors from Winbugs.

Notice that for this sample of n=500 observations, both estimation procedures are very close to one another with estimates and standard errors very similar. In fact, for most of the measurement model parameters they are practically identical. The similarity between maximum likelihood estimation and the Bayesian method is to be expected since with non-informative priors and large numbers of observations the posterior means in the Bayesian method are essentially the same as the maximum likelihood estimates. Further, we note that both methods yield confidence intervals ('credible intervals' in the Bayesian context) that cover the true value for all parameters, particularly both methods find that the γ_{13} is not different from zero indicating it is not needed in the nonlinear exponential (as is the truth for that variable).

For this particular example, both PROC NLMIXED and Winbugs were able to converge in a reasonable amount of time using the crude starting values given (i.e. simply specifying the $\mathbf{\Psi}$ starting values to be half the sample variance of the respective observed variables and then allowing all other parameter starting values to be set to 1). It is often necessary to provide better starting values particularly for the nonlinear structural model parameters due to the complexity of the model and perhaps variability of the data. For a nonlinear SEM such as the one examined here, this could be accomplished by taking the observed variables z_7, z_8, and z_9 - each of which was identified to be equal to one of the three latent variables plus error- and performing a nonlinear regression similar to (8)-(9) but taking the observed variables to be equal to their respective latent variables. This can be accomplished using SAS Proc NLIN or any other software that performs nonlinear regression. Then the estimates from the regression can provided crude estimates which can then be used as starting values for the nonlinear structural model parameters.

6.2 Nonlinear SEM Example 2 - The interaction model

Due to its potential useful interpretation as a moderating effect, the special case (13) which includes a simple latent cross-product (interaction) between two exogenous latent variables has been studied extensively. As described in the introduction there is a large literature surrounding product indicator methods (following Kenny and Judd (1984)) aimed specifically at estimation for this second-order interaction model.

Maximum likelihood and Bayesian solutions can be accomplished for this latent interaction model by making only slight changes to the SAS Proc NLMIXED and Winbugs code presented in the previous section. In this section, we demonstrate that maximum likelihood estimates can also be obtained using Mplus in the special case of the latent interaction model. It is important to note that the maximum likelihood computational method used in SAS Proc NLMIXED differs from that used by Mplus and so even though both aim to obtain maximum likelihood estimates, in fact, both are only as good as their respective computational approximations. While it is beyond the scope of this chapter to provide a detailed comparison of numerical accuracy and computational speed between Mplus and SAS Proc NLMIXED for the latent interaction model, it is expected that Mplus should perform generally better (more accurate and faster) for the interaction model than Proc NLMIXED as the computational algorithm in Mplus is tailored to the specific form of the interaction structural model. Furthermore, the code in Mplus is more concise due to the simple syntax available for specifying the measurement model (using the 'by' command).

For demonstration we generate data from the interaction model (13) taking the exogenous factors and errors to be normally distributed. Similar to the measurement model (21) used in the previous subsection (except here we have two exogenous variables ξ_1 and ξ_2 and only one endogenous variable η_1), we generate three observed variables for each of the three latent variables using a linear measurement model with simple structure, resulting in 9 observed variables for $\mathbf{Z}$. Specifically, the generating model is such that ξ_1 and ξ_2 are normally distributed with means 5 and 2 respectively, both with variance 1 and correlation equal to 0.5. Then $\eta_1 = .2 + .1\xi_1 + .1\xi_2 + .2\xi_1\xi_2 + \delta$, with δ normally distributed and $Var(\delta)$ taken so that $R^2 = .5$. The $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are chosen in the measurement model so all observed variables have reliability of .75. A sample of size n=500 is generated.

Here is the Mplus code for performing maximum likelihood estimation of the latent interaction model to $\mathbf{Z}$:

```
Data: file is "C:\mplusatroot\interactiondata";

variable: names z1-z9; usevariables are z1-z9;

Analysis: type = random;
          algorithm = integration;
Model:

! putting [z7@0 z8@0 z9@0] fixes the intercept for the z7, z8, z9 variables
! in the measurement model to zero, the other measurement model intercepts
! are estimated and are called the NU parameters in the output

       ksi1 by z7 z1 z2 ;
       ksi2 by z8 z3 z4 ;
       eta1 by z9 z5 z6 ;
       [z7@0 z8@0 z9@0];

! the next command using 'xwith' creates the latent interaction term
! which is then named ks1xks2
       ks1xks2 | ksi1 xwith ksi2;
```

```
! the next line specifies the interaction structural model
       eta1 on ksi1 ksi2 ks1xks2;

! putting [ksi1* ksi2* eta1*] tells Mplus to freely estimate a mean for
! ksi1 and ksi2 and an intercept for the structural model of
! eta1 on ksi1 ksi2 and ks1xks2, respectively
       [ksi1* ksi2* eta1*];
```

Table 2 displays the results from fitting one dataset of size n=500 from the cross product model as specified above using the three different softwares. For brevity only the structural model parameters are presented. Similar to the example presented in Section 6.1, the results from these three procedures lead to very similar parameter estimates and in all cases identical conclusions in terms of statistical significance of parameters (with the interaction coefficient γ_{13} estimate being the only one statistically significant). SAS Proc NLMIXED and MPlus both took approximately 45 seconds to converge and Winbugs took approx 4 minutes (for 14,000 MCMC iterations) on the same laptop computer with Intel Core 2CPU, 2.16 GHz processor. The same starting values were used for all three procedures and they were taken to be the default starting values used in Mplus. The one exception was that Winbugs would not converge even after 30,000 iterations if starting values of 0 were given for the means of both ξ_{1i} and ξ_{2i} (recall their true means were 5 and 2 respectively). Recall that Winbugs treats the individual latent factors as parameters and by default provides starting values for all 3×500 of them by using the starting values given for the $\boldsymbol{\theta}$ parameters and generating from the prior distribution for the latent factors. Since the starting values of zero for the means of ξ_{1i} and ξ_{2i} are far from the truth, this leads to all the 3×500 starting values for the latent factors to be far from the truth, hence the problem. To remedy this, better starting value for the means of ξ_{1i} and ξ_{2i} were used by using the sample mean of the respective observed variables z7 and z8 identified to be direct measures of the exogenous factors plus error. This lead to the final result presented.

While typically inference for the structural model parameters (as shown in Table 2) are the main output of interest, it is possible to also obtain predicted values for the underlying latent factors using all three software. In Winbugs, the estimates of the latent factors can be obtained by including the factors as a node to be updated in the output. As described earlier, using the ML procedure, factor score estimates are obtained as empirical Bayes estimates using (17) and this is done in both Proc NLMIXED and Mplus. In SAS Proc NLMIXED, the lines of code added to obtain factor score estimates for each of the three factors are:

```
predict ksi1 out = save1;
predict ksi2 out = save2;
predict eta1 out = save3;
```

where the predicted values and prediction intervals for each individuals three latent variables are stored in the respective datasets, save1-save3. In Mplus, the lines of code added are:

```
SAVEDATA: file is "C:\outfscores.dat";
          save = fscores;
```

where a file called “outfscores.dat” will be created containing the factor score estimates.

For the example dataset considered here, the correlation between the factor score estimates for each of the 3 factors obtained from Proc NLMIXED, Mplus and Winbugs were identical out to 3 decimal places. Hence, in addition to very similar estimates for $\boldsymbol{\theta}$, the three procedures give nearly identical estimates for the individual latent factors. Moreover a comparison between the factor score estimates and the true generated latent variables finds a correlation of .95 for both ξ_1 and ξ_2 and .94 for η_1. This strong similarity is governed (as it would be also in a linear SEM) by the reliability of the individual observed variables and the number of items measuring each factor. Here each observed variable had a true reliability for its respective factor of .75 and there were 3 observed variables for each factor, so it is expected the similarity between the true factors and the predicted ones would be high.

In the last part of this section, the importance of paying attention to means and intercepts in nonlinear SEM is emphasized because means and intercepts in nonlinear SEM can directly effect interpretation. As Moosbrugger et al. (1998) pointed out, linear transformation (e.g. mean centering) to the exogenous latent variables have profound effects on the structural coefficients. "In a structural equation with a latent interaction effect, the parameters γ_1 and γ_2 do not represent constant effects of the latent variables. In contrast to structural equation models without latent interaction terms, the structural parameters γ_1 and γ_2 are not independent of translations of the latent variables, whereas the latent interaction effect γ_3 is unaffected by the scale translation. Therefore, the parameters γ_1 and γ_2 must be interpreted in relation to the scaling chosen for the latent variables ξ_1 and ξ_2. Again, one should not interpret the parameters γ_1 and γ_2 on their own, but interpret the way in which the linear relationship between η and ξ_1 is moderated by ξ_2."

Figure 1 demonstrates that knowing the coefficients in the structural relationship (in this case $\eta_1 = .2 + .1\xi_1 + .1\xi_2 + .2\xi_1\xi_2$) is not alone enough to describe the nature of the interaction relationship. It is also necessary to use information about the means of the latent variables when interpreting the interaction relationship. The three rows of plots in Figure 1 represent a model with the same coefficients in the structural model but with the means of the latent variables differing across the three rows, i.e. $E(\xi_1, \xi_2) = \{(0, 0), (5, 5), (5, 2)\}$, respectively. Each row leads to a different interpretation of the relationship. In the first row it is seen that when one or the other exogenous variables is fixed at a low value, the relationship between the other variable and the outcome, η_1, is negative. Whereas for high fixed values of either variable, the relationship between the other, respective, variable and the outcome is positive. In the second row, similar to the first, the relationships are symmetric in that the way f1 and f2 both relate to f3 is the same. This is an artifact of both the means being the same as well as the coefficents in the interaction model for f1 and f2 being the same. Note that in the third row, where the means of the two variables differ, there are different relationships with the outcome. For low values of ξ_2, there is little or no increase in η_1 when ξ_1 increases, whereas if there is a large value of ξ_2 present then we expect to see large increases in η_1 when ξ_1 increases. On the other hand, for fixed values of ξ_1, there is always an increase in η_1 as ξ_2 increases. The increase is just larger in some cases (e.g. ξ_1 high) than others (ξ_1 low).

It is common for structural equation modeling software by default to fit mean centered data, implying that the means of all the latent factors are zero. For a linear structural equation model this has no effect on the resulting coefficients, but for a nonlinear structural model it can. In the Mplus code presented above, the two lines [z7@0 z8@0 z9@0]; and [ksi1* ksi2* eta1*]; were added specifically so that intercepts and means of the latent factors would be estimated. The effect of dropping these two lines is to defer to the default setting of fitting mean centered data which implies latent factor means fixed to zero.

Continuing the example of the interaction model, we demonstrate the one-to-one relationship between the coefficients in a model with the means of latent variables left free to be estimated as compared to one where they are fixed to zero. Take $f_1^* = f_1 - \mu_1$, $f_2^* = f_2 - \mu_2$, $f_3^* = f_3 - \mu_3$, then $f_3 = \gamma_0 + \gamma_1 f_1 + \gamma_2 f_2 + \gamma_3 f_1 f_2$ can be rewritten as

$$\begin{aligned} f_3^* + \mu_3 &= \gamma_0 + \gamma_1(f_1^* + \mu_1) + \gamma_2(f_2^* + \mu_2) + \gamma_3(f_1^* + \mu_1)(f_2^* + \mu_2) && (26) \\ f_3^* &= (-\mu_3 + \gamma_0 + \gamma_1\mu_1 + \gamma_2\mu_2 + \gamma_3\mu_1\mu_2) && (27) \\ & \quad +(\gamma_1 + \gamma_3\mu_2)f_1^* + (\gamma_2 + \gamma_3\mu_1)f_2^* + \gamma_3 f_1^* f_2^* && (28) \\ &= -\gamma_3 Cov(f_1, f_2) + (\gamma_1 + \gamma_3\mu_2)f_1^* + (\gamma_2 + \gamma_3\mu_1)f_2^* + \gamma_3 f_1^* f_2^* \ . && (29) \end{aligned}$$

Notice that the coefficient of the interaction term is invariant, that is, whether we work with the f or the f^* variables, we get γ_3 as the coefficient for the cross-product. This is useful since it implies that in order to test the interaction term equal to zero, it does not matter whether the means of the latent factors are fixed to zero or not. The coefficients for f_1 and f_2, though are different depending on centering. One implication of the invariance of the coefficients of f_1 and f_2 is that testing these coefficients equal to zero depends on what is assumed about the mean of the factors and so is not particularly interesting on its own.

Consider fitting the same data generated for the cross-product model above again based on mean centered data. Similar Mplus code is used but the constraint [$z7@0$ $z8@0$ $z9@0$] and the line [$ksi1*$ $ksi2*$ $eta1*$] are dropped leaving Mplus to go with the default of mean centering all the data. We find that the estimate for the interaction term γ_3 (i.e. 0.213 with s.e. (.093)) is exactly the same as before as in Table 2, and significant regardless of whether means and intercepts are included. On the other hand, completely different conclusions would be made about the significance of γ_1 and γ_2 in the model without intercepts. Now, $\hat{\gamma}_1 = .339(.117)$ and $\hat{\gamma}_2 = 1.188(.129)$ are both highly significant whereas they were not different from zero in the previous parameterization. We point out though that plots like those in Figure 1 of the two different looking results (in terms of having different coefficients) would actually look the same when plotted, only the center of the scale on the axes would be different. Remember these two models are equivalent as they are just re-parameterizations of one another, so ultimately the interpretation should be the same. It is recommended that plots similar to those shown in Figure 2 which take into account the mean value of the latent factors be presented to explain the results from interaction models rather than just relying on the sign and magnitude of the coefficients in the structural model.

7 Conclusion

The major complication introduced by nonlinear terms in the structural model is that estimation of the parameters in the SEM can no longer be accomplished by the well known, often used linear SEM estimation method of modeling the observed data as multivariate normal and hence comparing the observed covariance matrix $\mathbf{S}$ to a model covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. The sample covariance matrix of the observed data $\mathbf{S}$ is no longer a sufficient statistic for the model parameters once nonlinear terms are added to the structural model. This is because the observed data is no longer multivariate normal (as a consequence of the nonlinear term), hence we need more information from the data than that simply provided by the covariance matrix.

As a result, in order to perform maximum likelihood or Bayesian inference, it is necessary to use more sophisticated statistical computation algorithms. Fortunately some of these algorithms are currently available in commercial software and the nonlinear structural equation model can be fit with maximum likelihood using SAS Proc NLMIXED (or Mplus for the simple interaction model) and within a Bayesian framework using Winbugs.

In this chapter it was demonstrated that the maximum likelihood and Bayesian procedures give very similar results for the examples presented. This similarity is expected to be the case more generally whenever non-informative priors are used in the Bayesian setting. As Hill (1990) describes "besides varying interpretation of probability, the only essential difference between the schools is in the model itself", that is, compare the model (14) used for the maximum likelihood frequentist procedure to the model (18) used for the Bayesian method which includes the addition of a prior. The fact that the posterior distribution for a parameter in the Bayesian setting is proportion to the likelihood function used in maximum likelihood times the prior should give intuition that there will not be much difference in the results as long as the likelihood (i.e. the observed data) provides a lot more information than the prior. Thus, the choice between the Bayesian method and the frequentist maximum likelihood estimation method, in the opinion of this author, is one of implementation convenience not of superiority of one method over the other. As models become more complex, the computational algorithms needed to fit them become more intensive and thus for all practical purposes the methods which will get used are the one for which there are user-friendly efficient softwares to do them.

While maximum likelihood and Bayesian methods provide appropriate inference when the distributional assumptions of the underlying factors and errors are correct, and, as shown in this paper there are softwares capable of performing them, they may provide severely biased results when these non-checkable distributional assumptions are incorrect. It is important that in addition to improving algorithms and software for performing maximum likelihood and Bayesian inference, that statistical methods continue to be developed that are robust to distributional assumptions. Wall and Amemiya (2000, 2003) introduced a two-stage method of moments (2SMM) procedure for fitting (10) when the nonlinear $\mathbf{g}(\boldsymbol{\xi}_i)$ part consists of general polynomial terms. The 2SMM produces consistent estimators for the structural model parameters for virtually any distribution of the observed indicator variables where the linear measurement model holds. The procedure uses factor score estimates in a form of nonlinear errors-in-variables regression and produces closed-form method of moments type estimators as well as asymptotically correct standard errors. Moreover, Wall and Amemiya (2007) present a pseudo-likelihood approach for the general nonlinear structural equation model (7) that weakens distributional assumptions of underlying exogenous factors by allowing them to be mixtures of normal distributions. A method called efficient method of moments EMM has also been introduced for nonlinear SEM (Lyhagen 2007) as a way to robustify violations of the distributional assumptions for the underlying exogenous factors and errors.

Finally, it is our hope that now that estimation for these nonlinear structural equation models is implementable within commercial software that they will be applied to real theories motivating their need.

References

Algina, J., & Moulder, B. C. (2001). A note on estimating the JoreskogYang model for latent variable interaction using LISREL 8.3. *Structural Equation Modeling*, 8, 4052.

Amemiya Y and Zhao Y (2001). Estimation for nonlinear structural equation system with an unspecified distribution. *Proceedings of Business and Economic Statistics Section, the Annual Meeting of the American Statistical Association* (CD-ROM).

Arbuckle JL (1995) *AMOS for Windows: Analysis of moment structures (Version 4.0).* Chicago: SmallWaters.

Arminger, G. and Muthén, B. (1998) A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika*, 63(3), 271-300.

Bartholomew, D.J., and Knott, M. (1999). *Latent Variable Models and Factor Analysis*, 2nd. ed., Kendall's Library of Statistics.

Bentler PM (1985) *Theory and implementation of EQS: A structural equations program.* Los Angeles: BMDP Statistical Software

Bollen KA (1989). *Structural Equations with Latent Variables.* New York, Wiley.

Carlin B and Louis T (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edition CRC Press

Congdon P (2001). *Bayesian Statistical Modelling*, Wiley.

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

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling* 9(3), 327346.

Geman S and Geman D (1984). Stochastic relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Gelfand A and Smith A (1990) Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85, pp. 398-409.

Gelman, A. (1996). Inference and monitoring convergence. Pages 131-143 in W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds. *Markov Chain Monto Carlo in Practice.* Chapman and Hall/CRC, Boca Raton, Florida.

Geweke J, 1989. Bayesian inference in economietric models using Monte Carlo integration. *Econometrica*, 57, 1317-1340.

Hasting WK (1970) Monte Carlo Sampling Methods Using Markov Chains and Their Applications,*Biometrika*, 57(1):97-109, 1970.

Hayduck LA (1987). *Structural Equation Modeling with LISREL: Essentials and Advances.*

Hill JR (1990). A general framework for model-based statistics. *Biometrika*, 77(1), 115-126.

Huber P, Ronchetti E, Victoria-Feser M (2004) Estimation of generalized linear latent variable models, *Journal of the Royal Statistical Society Series B*, 66, 893-908.

Jaccard, J., and Wan, C.K. (1996). *LISREL approaches to interaction effects in multiple regression*, Sage.

Joreskog KG and Sorbom D (1996) *LISREL 8 user's reference guide.* Chicago: Scientific Software International.

Jöreskog KG and Yang F (1996). Non-linear structural equation models: The Kenny-Judd model with interaction effects. In G.A. Marcoulides and R.E. Schumacker (Eds.) *Advanced Structural Equation Modeling: Issues and Techniques*, 57-88.

Jöreskog KG and Yang F (1997). Estimation of interaction models using the augmented moment matrix: Comparison of asymptotic standard errors. In W. Bandilla and F. Faulbaum (Eds.) *SoftStat '97 Advances in Statistical Software 6*, 467-478.

Kenny DA and Judd CM (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, 96(1), 201-210.

Klein, A., Moosbrugger, H., Schermelleh-Engel, K., Frank, D. (1997). A new approach to the estimation of latent interaction effects in structural equation models. In W. Bandilla and F. Faulbaum (Eds.) *SoftStat '97 Advances in Statistical Software 6*, 479-486.

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, 65, 457474.

Lee SY (2007) *Structural Equation Modeling: A Bayesian Approach.* Wiley.

Lee SY and Song XY (2003a). Maximum likelihood estimation and model comparison of nonlinear structural equation models with continuous and polytomous variables. *Computational Statistics and Data Analysis*, 44, 125-142.

Lee, S.Y., Song, X.Y. (2003b). Model comparison of nonlinear structural equation models with fixed covariates. *Psychometrika*, 68(1), 27-47.

Lee SY, Song XY and Lee JCK (2003). Maximum likelihood estimation of nonlinear structural equation models with ignorable missing data *Journal of Educational and Behavioral Statistics*, 28(2), 111-134.

Lee SY, Song XY, and Poon WY (2004). Comparison of approaches in estimating interaction and quadratic effects of latent variables. *Multivariate Behavioral Research*, 39, 37-67.

Lee SY, Song XY, and Tang NS (2007) Bayesian Methods for Analyzing Structural Equation Models With Covariates, Interaction, and Quadratic Latent Variables, *Structural Equation Modeling*, 14(3), 404-434.

Lee SY and Zhu HT (2000). Statistical analysis of nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, 53, 209-232.

Lee SY and Zhu HT (2002). Maximum likelihood estimation of nonlinear structural equation models. *Psychometrika*, 67(2), 189-210.

Li, F. Z., Duncan T. E., Acock, A. (2000). Modeling interaction effects in latent growth curve models. *Structural Equation Modeling*, 7, 497533.

Li, F., Harmer, P., Duncan, T., Duncan, S., Acock, A., and Boles, S. (1998). Approaches to testing interaction effects using structural equation modeling methodology. *Multivariate Behavioral Research*, 33(1), 1-39.

Little R (2006) Calibrated Bayes: A Bayes/Frequentist Roadmap, *The American Statistician*, 60(3), 1-11.

Lyhagen J (2007) Estimating Nonlinear Structural Models: EMM and the Kenny-Judd Model, *Structural Equation Modeling*, 14(3), 391-403.

Marsh, H.W., Wen, Z., Hau, K.T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods*, 9(3), 275-300.

McCullagh and Nelder (1989) *Generalized Linear Models*, 2nd edition Chapman and Hall CRC

Moosbrugger H, Schermelleh-Engel K, Klein A (1998) Methodological problems of estimating latent interaction effects, *Methods of Psychological Research Online* 1997, Vol 2(2) Internet: http://www.pabst-publishers.de/mpr

Moulder B. C., & Algina, J. (2002). Comparison of methods for estimating and testing latent variable interactions. *Structural Equation Modeling*, 9, 119.

Moustaki I, (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *British J. Math. Statist. Psych.* v56. 337-357.

Moustaki I and Knott M (2000). Generalized latent trait models. *Psychometrika.* v65. 391-411.

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

Muthén LK and Muthén BO (1998-2007). *Mplus User's guide. Version 5.* Los Angeles: Muthen and Muthen.

Patefield M (2002). Fitting non-linear structural relationships using SAS procedure NLMIXED, *Journal of the Royal Statisitical Society: Series D (The Statistician)*, 51(3), 355-366.

Ping, R.A. (1995). A parsimonious estimating technique for interaction and quadratic latent variables. *Journal of Marketing Research*, 32, 336-347.

Ping, R.A. (1996a). Latent variable interaction and quadratic effect estimation: A two-step technique using structural equation analysis. *Psychological Bulletin*, 119, 166-175.

Ping, R.A. (1996b). Latent variable regression: A technique for estimating interaction and quadratic coefficients. *Multivariate Behavioral Research*, 31, 95-120.

Ping, R.A. (1996c). Estimating latent variable interactions and quadratics: The state of this art. *Journal of Management*, 22, 163-183.

Pinheiro J and Bates D (1995). Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model *Journal of Computational and Graphical Statistics*, Vol. 4, No. 1, 12-35.

Rabe-Hesketh S, Pickles A, and Scrondal A (2004) On web at www.gllamm.org. *GLLAMM Manual.* UC Berkely Division of Biostatistics Working Paper Series. Working Paper 160.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2002) Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2, 121.

Sammel, M., Ryan, L., and Legler, J. (1997). Latent variable models for mixed discrete and continuous outcomes, *JRSS-B*, 59, 667-678.

SAS Institute Inc (2002), SAS Version 9.1. Cary, NC: SAS Institute Inc.

Schumacker, R. and Marcoulides, G. (Eds) (1998) *Interaction and nonlinear effects in structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum Associates.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Song XY and Lee SY (2002). A Bayesian approach for multigroup nonlinear factor analysis. *Structural Equation Modeling*, 9(4), 523-553.

Spiegelhalter DJ, Thomas A, Best NG and Lunn D (2002). *WinBugs user manual (Version 1.4).* Cambridge, UK: MRC Biostatistics Unit

Takane, Y. and de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52, 393-408.

Tanner M and Wong W (1987) The calculation of posterior distributions by data augmentation (with discussion) *Journal of the American Statistical Association*, 82, 528-550.

Tierney L and Kadane J (1986) and numerical integration techniques (similar to those used Accurate Approximations for Posterior Moments and Marginal Densities, *Journal of the American Statistical Association*, 81, 82-86.

Van Der Linden, W.J. and Hambleton, R.K. (1997) *Handbook of modern item response theory*, Springer

Wall, M.M. and Amemiya, Y. (2000). Estimation for polynomial structural equation models. *Journal of the American Statistician*, 95, 929-940.

Wall, M.M. and Amemiya, Y. (2001). Generalized appended product indicator procedure for nonlinear structural equation analysis. *Journal of Educational and Behavioral Statistics*, 26(1), 1-29.

Wall, M.M. and Amemiya, Y. (2003). A method of moments technique for fitting interaction effects in structural equation models. *British Journal of Mathematical and Statistical Psychology*, 56, 47-63.

Wall M.M. and Amemiya, Y. (2007) "Nonlinear structural equation modeling as a statistical method" In *Handbook of Latent Variable and related Models*, ed Sik-Yum Lee, Chapter 15, 321-344, Elsevier, The Netherlands.

Wen, Z., Marsh, H.W., Hau, K.T. (2002). Interaction effects in growth modeling: A full model. *Structural Equation Modeling*, 9(1), 20-39.

Wittenberg, J., and Arminger, G. (1997). Bayesian non-linear latent variable models-specification and estimation with the program system BALAM. In W. Bandilla and F. Faulbaum (Eds.) *SoftStat '97 Advances in Statistical Software 6*(487-494).

Zhu, H.T and Lee, S.Y. (1999). Statistical analysis of nonlinear factor analysis models. *British Journal of Mathematical and Statistical Psychology*, 52, 225-242.

Table 1: True parameter values and parameter estimates from SAS Proc NLMIXED and Winbugs from the nonlinear SEM Example 1 (21)-(25).

	truth	ML estimates (s.e) Proc NLMIXED	Bayesian estimates (s.e.) Winbugs
parameters for measurement model (21)-(22) - $\boldsymbol{\theta}_m$			
λ_{01}	0	0.035 (0.023)	0.034 (0.024)
λ_{02}	0	-0.001 (0.010)	-0.001 (0.010)
λ_{03}	0	0.003 (0.037)	0.003 (0.037)
λ_{04}	0	-0.063 (0.074)	-0.065 (0.076)
λ_{05}	0	-0.013 (0.022)	-0.013 (0.022)
λ_{06}	0	-0.018 (0.018)	-0.018 (0.018)
Λ_{11}	0.7	0.657 (0.026)	0.661 (0.026)
Λ_{21}	0.3	0.286 (0.011)	0.287 (0.011)
Λ_{32}	0.4	0.393 (0.018)	0.394 (0.018)
Λ_{42}	0.8	0.817 (0.037)	0.818 (0.037)
Λ_{53}	0.5	0.506 (0.020)	0.506 (0.021)
Λ_{63}	0.4	0.394 (0.016)	0.395 (0.016)
Ψ_1	0.163	0.157 (0.014)	0.158 (0.014)
Ψ_2	0.030	0.029 (0.003)	0.029 (0.003)
Ψ_3	0.080	0.084 (0.007)	0.085 (0.007)
Ψ_4	0.320	0.308 (0.027)	0.312 (0.027)
Ψ_5	0.125	0.115 (0.010)	0.117 (0.010)
Ψ_6	0.080	0.076 (0.006)	0.076 (0.006)
Ψ_7	0.333	0.267 (0.027)	0.272 (0.027)
Ψ_8	0.500	0.465 (0.040)	0.469 (0.041)
Ψ_9	0.500	0.525 (0.043)	0.530 (0.044)
parameters for exogenous factors (23) - $\boldsymbol{\theta}_\xi$			
μ_ξ	0	-0.042 (0.050)	-0.041 (0.050)
ϕ_1	1	1.008 (0.082)	1.004 (0.083)
parameters for structural model (24) - $\boldsymbol{\theta}_s$			
γ_{10}	0.5	0.587 (0.135)	0.588 (0.142)
γ_{11}	1.0	0.917 (0.144)	0.916 (0.150)
γ_{12}	0.6	0.642 (0.070)	0.652 (0.067)
γ_{13}	0.0	-0.019 (0.051)	-.021 (0.045)
$Var(\delta_1)$	0.25	0.248 (0.035)	0.250 (0.036)
parameters for structural model (25) - $\boldsymbol{\theta}_s$			
γ_{20}	0.0	0.028 (0.050)	0.029 (0.055)
γ_{21}	1.0	0.940 (0.051)	0.938 (0.050)
$Var(\delta_2)$	0.5	0.505 (0.055)	0.508 (0.056)

Table 2: True parameter values and parameter estimates from Mplus, SAS Proc NLMIXED and Winbugs from the nonlinear SEM Example 2 in Section 6.2, i.e., the interaction model (13)

		ML estimates (s.e)		Bayesian estimates (s.e.)
parameter	truth	Mplus	Proc NLMIXED	Winbugs
γ_{10}	0.2	0.968 (.893)	0.984 (.930)	0.971 (.917)
γ_{11}	0.1	-0.079 (.195)	-0.083 (.204)	-.079 (.201)
γ_{12}	0.1	0.131 (.481)	0.122 (.480)	0.133 (.477)
γ_{13}	0.2	0.213 (.093)	0.215 (.094)	0.213 (.093)
$Var(\delta)$	3.05	3.354 (.333)	3.353 (.323)	3.385 (.334)

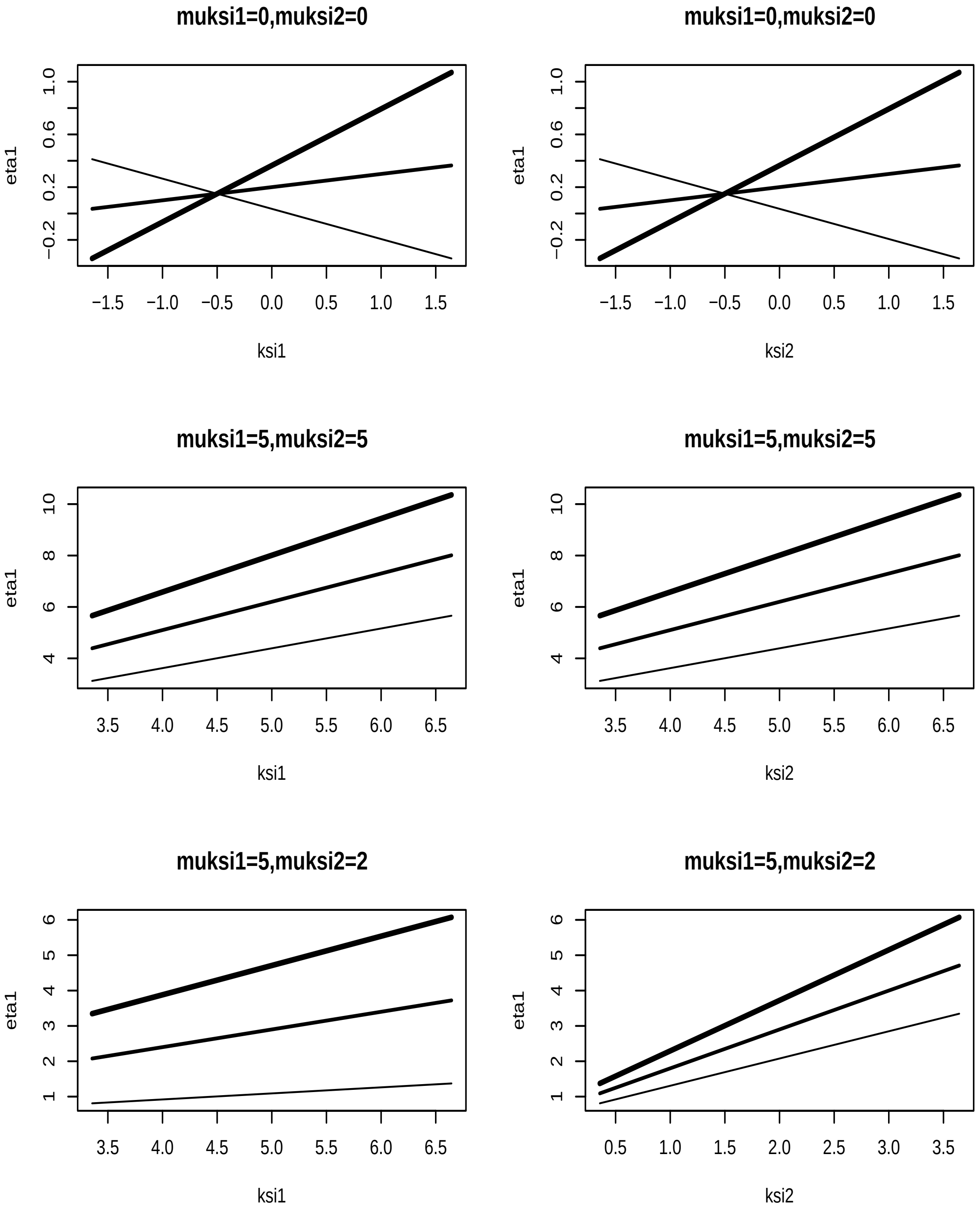

Figure 1: Plots of interaction relationship $\eta_1 = .2 + .1\xi_1 + .1\xi_2 + .2\xi_1\xi_2$ for different true means of exogenous factors (variances of f_1 and f_2 are one). In each plot of η_1 on ξ_1, the lines represent the relationship at low (5th percentile, thinnest line), median, and high (95th percentile, thickest line) fixed values of ξ_2. The respective relation is shown in each plot of η_1 on ξ_2 for fixed values of ξ_1. The only difference is that each row has a different mean value for ξ_1 and ξ_2: row1 (0,0), row2 (5,5), row3 (5,2).