

Simulation Study to Describe Bayesian Analysis of Nonlinear Structural Equation Modeling

Ferra Yanuar¹ & Aidinil Zetra²
ferrayanuar@sci.unand.ac.id, aidinil@soc.unand.ac.id

¹Department of Mathematics, Faculty of Mathematics and Natural Science, Andalas University, Kampus Limau Manis, 25163, Padang – Indonesia

²Department of Political Science, Faculty of Social and Political Science, Andalas University, Kampus Limau Manis, 25163, Padang – Indonesia

Abstract. Structural equation modeling (SEM) has widely used in many disciplines, such as economic, politic or health. Nonlinear structural equation modeling, as part of SEM, also has been developing analytically but still limited. In this method, the parameter models are estimated using conjugate prior in Bayesian approach. In nonlinear SEM, the models are specified including quadratic forms and/or interactions of latent variables. Posterior mean and posterior variance of the parameters are estimated using iteration approach since it is difficult to estimate those parameters model using analytical approach. The iteration approach used here is Markov Chain Monte Carlo (MCMC) method with Gibbs sampling. The simulation study is done to illustrate the proposed estimation methods for nonlinear model. A group of 300 data are generated to demonstrate the implementation of the proposed method. This study resulted that the proposed nonlinear SEM model could be accepted based on criteria of goodness of fit model.

Keywords: Bayesian analysis, nonlinear SEM, simulation study, structural equation modeling.

1 Introduction

Structural equation modeling (SEM) is a common and powerful tool in psychological, behavioral and social research. Traditionally, the latent variables in the structural equation model are assumed to have linear relations to each others [1], [2]. But recently, it has been recognized that nonlinear terms of independent latent variable for assessing dependent latent variables are useful in developing meaningful and more correct structural equation.

The estimation of nonlinear latent variables in the structural equation modeling (SEM) approach has increasing attention now adays. Several research scientists in social science, education and else often would like to consider models which contain interaction and/or quadratic terms in the latent variables. The simple quadratic model relating to this problem could be presented by following equation [3]

$$\eta = \beta_0 + \beta_1 \xi + \beta_2 \xi^2 + \zeta, \quad (1)$$

where $\zeta \sim \mathcal{N}(0, \psi)$ is a error term which is uncorrelated with the unobserved random variables or exogeneous latent variables $\xi \sim \mathcal{N}(0, \Phi)$. Regression coefficients are symbolled by β_0, β_1 and β_2 . Endogenous latent variable or η is assumed conditionally normal distribution given ξ with $\eta \sim \mathcal{N}(\beta_0 + \beta_1 \xi + \beta_2 \xi^2, \psi)$ [4].

Due to the complex distributions associated with the nonlinear latent variables, methods for analyzing such this structural equation modeling are become more difficult [5]. Several methods for the analysis of nonlinear SEM have been published, such as distribution analytic approaches [6], the product indicator approaches [7], moment based approaches [8] and Bayesian approaches [9]. More recently, methods that used the LISREL program have been proposed to analyzed some nonlinear structural equation models with interaction terms of latent variables.

This article presents a Bayesian approach to analyze nonlinear SEM with ordered categorical data. The purpose of this study is to model the nonlinear based on SEM model including the problem of nonnormal data as well as with the problem of small sample sizes. To estimate the posterior distribution of the parameters model, we do iteration approach by using Markov Chain Monte Carlo methods such as Gibbs sampler [10], [11], [12].

2 Materials and Methods

In structural equation model, the latent variable is unobserved variable that is measured by several observed variables or indicator variables. To construct the nonlinear in structural equation model, we assume to formulate an interaction in structural equation modeling between an endogenous latent variable, denoted by η with two exogeneous latent variables, symbolized by ξ_1 and ξ_2 . Let an endogenous latent variable η that is measured by one observed variable y , meanwhile exogeneous latent variables are measured by two observed variables each, denoted by x_1, x_2 and x_3, x_4 . Basically, the measurement equation models could be given by these following equations [13], [14]

$$x_1 = \lambda_1 \xi_1 + \epsilon_1, x_2 = \lambda_2 \xi_1 + \epsilon_2, x_3 = \lambda_3 \xi_2 + \epsilon_3, x_4 = \lambda_4 \xi_2 + \epsilon_4, \quad (2)$$

and

$$y = \lambda_5 \eta + \epsilon_5. \quad (3)$$

Parameter λ_j , for $j = 1, 2, \dots, 5$ are unknown loading factors that describe the relationships between exogeneous and endogeneous latent variables with observed variables. It is assumed that $\xi = (\xi_1, \xi_2)^T$ is distributed as $N(0, \Phi)$, where the covariance matrix Φ contains unknown parameters ϕ_{11} , ϕ_{21} and ϕ_{22} , whereas $\epsilon = (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)^T$ is distributed as $N(0, \Psi_\epsilon)$, where Ψ_ϵ is a diagonal covariance matrix with elements $\psi_\epsilon(i, i)$, $i = 1, \dots, 4$. The vectors ξ and ϵ are assumed to be uncorrelated.

Meanwhile the formula for structural equation model is assumed by following several different models of nonlinear latent variables in the structural equation modeling.

$$\begin{aligned} M1 : \eta &= \gamma_1 \xi_1 + \gamma_2 \xi_2 + \gamma_3 \xi_1 \xi_2 + \zeta, \\ M2 : \eta &= \gamma_1 \xi_1 + \gamma_2 \xi_2 + \gamma_3 \xi_1^2 + \gamma_4 \xi_2^2 + \zeta, \\ M3 : \eta &= \gamma_1 \xi_1 + \gamma_2 \xi_2 + \gamma_3 \xi_1 \xi_2 + \gamma_4 \xi_1^2 + \gamma_5 \xi_2^2 + \zeta, \end{aligned} \quad (4)$$

where γ_k , for $k = 1, \dots, 5$ are unknown regression coefficients between exogeneous latent variables ξ s to the endogenous latent variables η . Variable ζ is error term distributed as $N(0, \Psi_\zeta)$ and is uncorrelated with the ξ s. The model M1 involves marginal and interaction effects of the latent variables. The model M2 involves marginal and quadratic effects for each exogeneous latent variable. Meanwhile the model M3 is more complicate since it involves marginal, interaction effect and quadratic effect.

The best model will be considered in our simulation study based on the smallest value of DIC (Deviance Information Criteria) with the formula denoted by [15], [16]:

$$DIC = -2\log(p(\mathbf{X}|\theta)) + 2p, \quad (5)$$

with p is the number of parameters to be estimated.

Let the parameter vector θ contains all unknown parameters to be estimated. The Bayesian estimation method is used to estimate mean and standard error of all unknown parameters in θ . The Bayesian approach is well recognized as the attractive approach to analyze a wide variety of models [17]. In a Bayesian approach, unknown parameter $\theta = (\lambda, \Phi, \Psi_\epsilon, \Psi_\zeta)$ is considered as random variable which has prior probability density $p(\theta)$ that is equivalent with probability density function. In this study we use conjugate prior for all parameter estimated which then be used to construct posterior density[5].

$$\begin{aligned} \Psi_\epsilon, \Psi_\zeta &\sim \text{Gamma}(\alpha, \beta), \\ (\Lambda \mid \psi^{-1}) &\sim N(\Lambda, \psi\mathbf{H}), \\ \Phi^{-1} &\sim W_q(R_0, \rho_0). \end{aligned} \quad (6)$$

Let the observed data matrices be $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ and the observed data likelihood is denoted by $p(\mathbf{X}|\theta)$. In Bayesian estimation approach which is based on the posterior density function, it incorporates all the available information related to sample, denoted by $p(\mathbf{X}|\theta)$, and parameter to be estimated, or $p(\theta)$, in the estimation process. This statement is presented by following formula [18]

$$p(\theta|\mathbf{X}) \sim p(\theta) p(\mathbf{X}|\theta). \quad (7)$$

For nonlinear SEM analysis, to make work easier, it incorporates \mathbf{Z} as the matrix of all latent variables in the model into joint posterior distribution, denoted by $p(\theta, \mathbf{Z}|\mathbf{X})$. This joint posterior distribution can be adequately approximated by the empirical distribution of $\{(\theta^{(m)}, \mathbf{Z}^{(m)}); m = 1, \dots, M\}$ simulated from $p(\theta, \mathbf{Z}|\mathbf{X})$. Markov Chain Monte Carlo method with Gibbs Sampler algorithm is used to obtain sample by generating $\{(\theta^{(m)}, \mathbf{Z}^{(m)}); m = 1, \dots, M\}$. This sample can be regarded as simulated from the posterior distribution $p(\theta, \mathbf{Z}|\mathbf{X})$. Bayesian estimate of the unknown parameter θ is defined as the posterior mean of $p(\theta, \mathbf{Z}|\mathbf{X})$. Posterior mean represents the central tendency of θ in its posterior distribution, thus posterior mean of θ is estimated by this formula [19]

$$\hat{\theta} = \frac{\sum_{m=1}^M \theta^{(m)}}{M} \quad (8)$$

The posterior variance is estimated by

$$\text{Var}(\hat{\theta}|\mathbf{X}) = \frac{\sum_{m=1}^M (\theta^{(m)} - \hat{\theta})(\theta^{(m)} - \hat{\theta})^T}{M-1}. \quad (9)$$

Where the consistent estimator for standard error estimates of elements in $\hat{\theta}$ is calculated from the square roots of the corresponding diagonal elements of $\text{Var}(\hat{\theta}|\mathbf{X})$ [20]

3 Results and Discussions

In this study, we do such a literature review and simulation study to examine the empirical performances of the estimation approaches for analyzing nonlinear structural equation modeling as defined by the measurement equations and structural equation in Section 2 above.

3.1 Bayesian Estimation of Nonlinear SEM with Ordered Categorical Data

Following is the identification of the nonlinear SEM model with Bayesian approach. In this present study, a set of true population values of the unknown parameters were considered. We allow that components in observed variables $\mathbf{X} = (x_1, \dots, x_n)$ and $\mathbf{Z} = (z_1, \dots, z_n)$ both are ordered categorical. Then, we also let $\mathbf{Y} = (y_1, \dots, y_n)$ be the latent continuous data and $\mathbf{\Omega} = (\omega_1, \dots, \omega_n)$ be latent variables. Based on the technique of data augmentation, the observed data $[\mathbf{X}, \mathbf{Z}]$ are then augmented with the latent data $[\mathbf{Y}, \mathbf{\Omega}]$ in the posterior analysis. By implementing Markov Chain Monte Carlo (MCMC) through Gibbs sampler method, the unknown parameter θ will be obtained via this iteration process. At the m th iteration with current values $\alpha^{(m)}, \theta^{(m)}, \Omega^{(m)}$ and $Y^{(m)}$, we do[9]

Generate $\Omega^{(m+1)}$ from $p(\Omega | \alpha^{(m)}, \theta^{(m)}, Y^{(m)}, \mathbf{X}, \mathbf{Z})$;

Generate $\theta^{(m+1)}$ from $p(\theta | \Omega^{(m+1)}, \alpha^{(m)}, Y^{(m)}, \mathbf{X}, \mathbf{Z})$;

Generate $(\alpha^{(m+1)}, Y^{(m+1)})$ from $p(\alpha, Y | \theta^{(m+1)}, \Omega^{(m+1)}, \mathbf{X}, \mathbf{Z})$

In order to derive the conditional distributions involved in above equations, some natural assumptions will be imposed on the prior distributions of θ . Let $S = (s_1, \dots, s_n)$ with $s_i = (x_i, y_i)$. Then let θ_ϵ be the unknown parameters in μ, λ and Ψ_ϵ that are associated with measurement equation above and let θ_ω be the unknown parameters in λ_ω, Φ and Ψ_ζ that are associated with structural equation above. It is assumed here that prior distribution of θ_ϵ and θ_ω are independent of each other. Thus, we could assume here that $p(\theta) = p(\theta_\epsilon)p(\theta_\omega)$. Therefore conditional distribution corresponding to Ω, θ and α, Y are given by

$$p(\Omega | \theta, \alpha, Y, \mathbf{X}, \mathbf{Z}) = p(\Omega | \theta, \alpha, Y, \mathbf{X}) = p(\Omega | \theta, S),$$

$$p(\theta | \Omega, \alpha, Y, \mathbf{X}, \mathbf{Z}) = p(\theta_\epsilon, \theta_\omega | \Omega, S) = p(\theta_\epsilon | \Omega, S)p(\theta_\omega | \Omega, S),$$

$$p(\alpha, Y | \theta, \Omega, \mathbf{X}, \mathbf{Z}) = p(\alpha, Y | \theta, \Omega, Z).$$

Based on equations above, it can be noted that given S, the model with the ordered categorical data is the same as the model with the continuous data.

Conditional distributions involved in $p(\theta | \Omega, \alpha, Y, \mathbf{X}, \mathbf{Z})$ are the Gamma distribution, normal distribution and inverted Wishart distribution. It is not difficult to obtain observations by MCMC simulating steps. The same ways then could be implemented to simulate observations from $p(\Omega | \theta, \alpha, Y, \mathbf{X}, \mathbf{Z})$ and $p(\alpha, Y | \theta, \Omega, \mathbf{X}, \mathbf{Z})$. Bayesian estimated for Ω and θ can be obtained via a sufficiently large number of simulated observations collected by Gibbs sampler approach.

3.2 Illustrative Example

Our illustrative example here is to demonstrate the Bayesian method in modeling nonlinear structural equation model using generated data. A group 300 data is generated randomly using R package. There are three exogenous latent variables and one endogenous latent variables in the hypothesis model. The first exogenous latent variable is measured by six indicator variables. Meanwhile the second and third exogenous latent variable are measured each by two indicator variables. The endogenous latent variables is measured by three indicator variables. The responses of the respondents are available in five Likers scale start from 1 to 5.

Nonlinear structural equation modeling then applied to this data. Following are the result from several models which are tested in this study. The choice of the best model is based on the smallest value of DIC. Table 1 presents the proposed models nonlinear SEM and DIC value for each model. The estimation process is supported by WinBUGS software [2], [21], [22].

Table 1. Several Model Nonlinear SEM

No.	Model	DIC Value
1	$\pi = \gamma_1 x_1 + \gamma_2 \xi_1 + \gamma_3 \xi_2 + \gamma_4 \xi_3 + \gamma_5 \xi_3^2 + \gamma_6 \xi_2 \xi_3 + \zeta$	7472,740
2	$\pi = \gamma_1 x_1 + \gamma_2 \xi_1 + \gamma_3 \xi_2 + \gamma_4 \xi_3 + \gamma_5 \xi_2^2 + \zeta$	7479,930
3	$\pi = \gamma_1 x_1 + \gamma_2 \xi_1 + \gamma_3 \xi_2 + \gamma_4 \xi_3 + \gamma_5 \xi_3^2 + \zeta$	7472,090
4	$\pi = \gamma_1 x_1 + \gamma_2 \xi_1 + \gamma_3 \xi_2 + \gamma_4 \xi_3^2 + \zeta$	7476,310

Based on Table 1, the model 3 yield the smallest value of DIC that is 7472,740. Thus we prefer to choose model 3 as the best model. Following are the result for analysis of this model.

The structural equation for the proposed model 3 is denoted by

$$\hat{\pi} = -0,074x_1 + 1,146\xi_1 - 0,352\xi_2 + 0,143\xi_3 + 0,592\xi_3^2.$$

Meanwhile the values of the unstandardized coefficient of factor loading and the associated standard errors for each indicator variable in the measurement equations obtained based on nonlinear SEM are presented in Table 2. It is clear from Table 2 that all indicator variables that we hypothesized as predictors are significantly related to their respective latent variable.

Table 2. Coefficient Regressions in Measurement Model

Latent Variable	Indicator Variable	Estimate Bayesian Nonlinear SEM (SE)
First Exogenous LV (ξ_1)	(X_{11})	1
	(X_{21})	0.844 (0.054)*
	(X_{31})	0.772 (0.057)*
	(X_{41})	1.175 (0.115)*
	(X_{51})	1.127 (0.092)*
	(X_{61})	0.988 (0.087)*
Second Exogeneous LV (ξ_2)	(X_{72})	1
	(X_{82})	0.878 (0.073)*
Third Exogeneous LV (ξ_3)	(X_{93})	1
	(X_{103})	0.945 (0.084)*
Endogeneous LV (η)	(X_{114})	1
	(X_{124})	0.779 (0.074)*
	(X_{134})	0.790 (0.075)*

*Significant at 5% level

We then test the convergence statistics for all parameters of interest by plotting sequences of observations corresponding to some parameters generated by three different initial values. All plots indicate that the algorithm converged in less than 30000 iteration. In Figure 1, we present several plots for illustrative purposes.

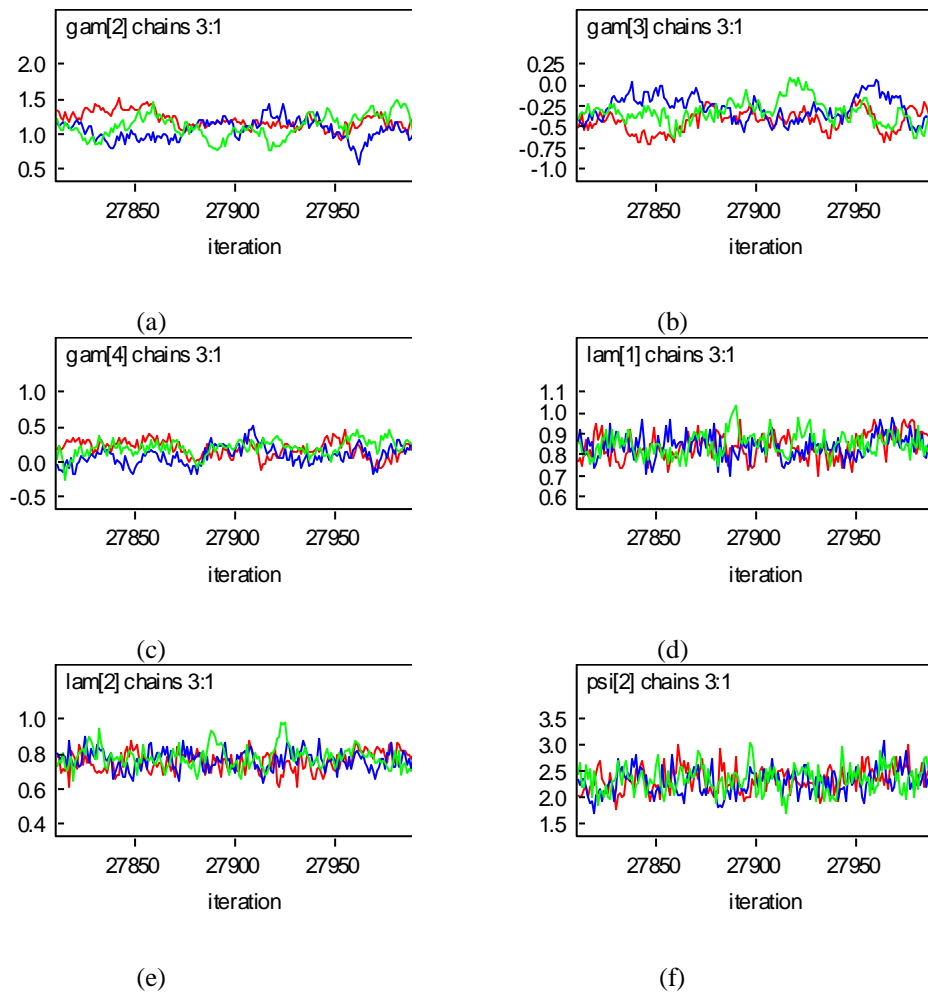


Figure 1. Several trace plots of selected parameters

The density plots of all parameter are also checked. Figure 2 presents several selected density plots. These figures inform us that all density plots have normal distribution

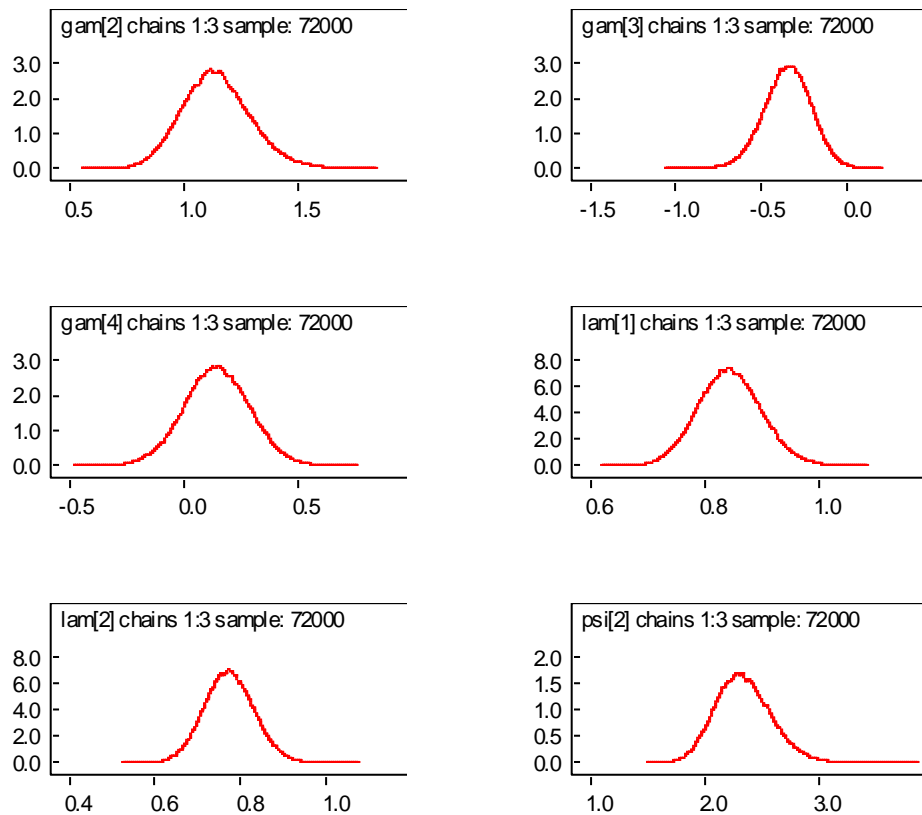


Figure 2. Several density plots of selected parameters

4 Conclusions

The main purpose of the present study is to demonstrate the value of the Bayesian SEM in handling nonlinear and nonnormal data. The strength of SEM is its ability to describe the relationship among latent variables as well as between the indicator variables and the corresponding latent variable [1]. These relationships could be linear or nonlinear. The problem in this study is to construct nonlinear SEM with ordered categorical data using Bayesian approach. Generated data is used to illustrate the use of the proposed method. The analysis in this study is implemented under WinBUGS version 1.4, a flexible and free download tool.

In this study, the models are put in a Bayesian analysis using conjugate priors. In nonlinear SEM, the model are specified including quadratic forms and/or interactions of latent variables mean and posterior variance of the parameters are estimated using iteration approach since it is difficult to estimate the parameter model using analytical approach. The iteration approach used here is Markov Chain Monte Carlo (MCMC) method with Gibbs sampling. This

study resulted that the proposed model could yield acceptable model based on several criteria of goodness of fit model.

References

- [1] F. Yanuar, K. Ibrahim, and A. A. Jemain, "On the application of structural equation modeling for the construction of a health index," *Environ. Health Prev. Med.*, vol. 15, no. 5, pp. 285–291, 2010.
- [2] F. Yanuar, "The use of Uninformative and informative prior distribution in Bayesian SEM," *Glob. J. Pure Appl. Math.*, vol. 11, no. 5, pp. 3259–3264, 2015.
- [3] C. F. Falk, "Are Robust Standard Errors the Best Approach for Interval Estimation With Nonnormal Data in Structural Equation Modeling?," *Struct. Equ. Model.*, vol. 00, pp. 1–23, 2017.
- [4] T. Y. Thanoon and R. Adnan, "Comparison between Bayesian structural equation models with ordered categorical data," *AIP Conf. Proc.*, vol. 1750, no. June, 2016.
- [5] A. Kelava, B. Nagengast, and H. Brandt, "A Nonlinear Structural Equation Mixture Modeling Approach for Nonnormally Distributed Latent Predictor Variables," *Struct. Equ. Model.*, vol. 21, no. 3, pp. 468–481, 2014.
- [6] A. G. Klein and B. O. Muthén, "Quasi-Maximum Likelihood Estimation of Structural Equation Models With Multiple Interaction and Quadratic Effects," *Multivariate Behav. Res.*, vol. 42, no. 4, pp. 647–673, 2008.
- [7] A. Kelava and H. Brandt, "A general non-linear multilevel structural equation mixture model," *Front. Psychol.*, vol. 5, pp. 1–16, 2014.
- [8] A. Mooijaart and P. M. Bentler, "An alternative approach for nonlinear latent variable models," *Struct. Equ. Model.*, vol. 17, no. 3, pp. 357–373, 2010.
- [9] F. Yanuar, K. Ibrahim, and A. Aziz Jemain, "Bayesian structural equation modeling for the health index," *J. Appl. Stat.*, vol. 40, no. 6, pp. 1254–1269, 2013.
- [10] S. Depaoli and J. P. Clifton, "A Bayesian Approach to Multilevel Structural Equation Modeling With Continuous and Dichotomous Outcomes," *Struct. Equ. Model.*, vol. 22, no. 3, pp. 327–351, 2015.
- [11] Z. Ke, Q. Zhang, and X. Tong, "Bayesian Meta-Analytic SEM: A One-Stage Approach to Modeling Between-Studies Heterogeneity in Structural Parameters," *Struct. Equ. Model.*, vol. 00, no. 00, pp. 1–23, 2018.
- [12] F. Yanuar, A. Zetra, C. Muharisa, and D. Devianto, "Bayesian Quantile Regression Method to Construct the Low Birth Weight Model," *J. Phys. Confrence Ser.*, vol. 1245, pp. 1–7, 2019.
- [13] S.-Y. Lee, X.-Y. Song, and J. C. K. Lee, "Maximum Likelihood Estimation of Nonlinear Structural Equation Models with Ignorable Missing Data," *J. Educ. Behav. Stat.*, vol. 28, no. 2, pp. 111–134, 2003.
- [14] S. Y. Lee and X. Y. Song, "Maximum likelihood estimation and model comparison of nonlinear structural equation models with continuous and polytomous variables," *Comput. Stat. Data Anal.*, vol. 44, pp. 125–142, 2003.
- [15] X.-Y. Song and S. Y. Lee, *Basic and Advanced Bayesian Structural Equation Modeling*. United Kingdom: A John Wiley & Sons, 2012.
- [16] M. K. Cain and Z. Zhang, "Fit for a Bayesian: An Evaluation of PPP and DIC for Structural Equation Modeling," *Struct. Equ. Model.*, vol. 26, no. 1, pp. 39–50, 2018.
- [17] S. Y. Lee, X. Y. Song, and N. S. Tang, "Bayesian methods for analyzing structural

- equation models with covariates, interaction, and quadratic latent variables,” *Struct. Equ. Model.*, vol. 14, no. 3, pp. 404–434, 2007.
- [18] F. Yanuar, “The Estimation Process in Bayesian Structural Equation Modeling Approach,” *J. Phys. Conf. Ser.*, 2014.
- [19] C. Muharisa, F. Yanuar, and D. Devianto, “Simulation Study of the Using of Bayesian Quantile Regression in Non- normal Error,” *Cauchy - J. Mat. Murni dan Apl.*, vol. 5, no. November, pp. 121–126, 2018.
- [20] S. Lee, X. Y. Song, and W. Poon, “Comparison of Approaches in Estimating Interaction and Quadratic Effects of Latent Variables Comparison of Approaches in Estimating Interaction and Quadratic Effects of Latent Variables Wai-Yin Poon,” *Multivariate Behav. Res.*, vol. 39, no. 1, pp. 37–41, 2004.
- [21] F. Yanuar, “The estimation process in Bayesian structural equation modeling approach,” *J. Phys. Conf. Ser.*, vol. 495, no. 1, 2014.
- [22] A. Rahmadita, F. Yanuar, and D. Devianto, “The Construction of Patient Loyalty Model Using Bayesian Structural Equation Modeling Approach,” *Cauchy - J. Mat. Murni dan Apl.*, vol. 5, no. 2, pp. 73–79, 2018.