Computational Statistics and Data Analysis 68 (2013) 111–128

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Bayesian diagnostics of transformation structural equation models

Ji Chen, Pengfei Liu, Xinyuan Song *

Department of Statistics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

ARTICLE INFO

Article history:
Received 19 March 2013
Received in revised form 3 June 2013
Accepted 10 June 2013
Available online 28 June 2013

Keywords:
Latent variables
Logarithm of Bayes factor
Perturbation schemes
Nonparametric transformation
MCMC method

ABSTRACT

In the behavioral, social, psychological, and the medical sciences, the most widely used models in assessing latent variables are the structural equation models (SEMs). However, most of the existing statistical methods for analyzing SEMs have been developed for normally distributed data. Transformation SEMs are useful tools for tackling the non-normality of multidimensional data and simultaneously revealing the interrelationships among latent variables. The main objective of this paper is to develop a Bayesian diagnostic procedure for transformation SEMs. The first- and second-order local inference measures are established with the objective functions that are defined based on the logarithm of Bayes Factor. Markov chain Monte Carlo (MCMC) methods with the Bayesian *P*-splines approach are developed to compute the local influence measures and to estimate nonparametric transformations, latent variables, and unknown parameters. Compared with conventional maximum likelihood-based diagnostic procedures, the Bayesian diagnostic approach could detect outliers and/or influential points in the observed data, as well as conduct model comparison and sensitivity analysis through various perturbations of the data, sampling distributions, or prior distributions of the model parameters. Simulation studies reveal the empirical performance of the proposed Bayesian diagnostic procedure. An actual data set that is extracted from a study based on the Hong Kong Diabetes Registry is used to illustrate the application of our methodology.

1. Introduction

Latent variables that are measured by several observed variables to reflect their characteristics are highly common in practical research. Structural equation models (SEMs) are important tools in analyzing interrelationships among latent and observed variables. In the past years, SEMs have been widely applied to many disciplines, including business, education, medicine, as well as social and psychological sciences. One of the crucial assumptions of basic SEMs is the normality of the observed data. Given that this assumption is often untrue in practice, the problem of tackling non-normality has received much research attention. For instance, Bentler (1983) and Browne (1984) developed the asymptotically distribution free methods. Shapiro and Browne (1987) as well as Kano et al. (1993) proposed robust methods based on the multivariate-*t* distribution. Lee and Xia (2006, 2008) established more general robust models through random weights with the *t* and the normal distribution, respectively. Yung (1997), Dolan and van der Maas (1998), and Song and Lee (2012) discussed finite mixture modeling approach with a fixed number of components.

Another research direction moves toward transforming response variables such that the resulting model satisfies the model assumptions. Box and Cox (1964) provided a family of power transformations, the Box–Cox transformation, to address non-normality. Since then, various extensions and other parametric transformation families have been developed

* Corresponding author. Tel.: +852 3943 7929; fax: +852 2603 5188.
E-mail address: xysong@sta.cuhk.edu.hk (X. Song).
URL: http://www.sta.cuhk.edu.hk/xysong/ (X. Song).

0167-9473/$ – see front matter
http://dx.doi.org/10.1016/j.csda.2013.06.012

under both the maximum likelihood (ML) and the Bayesian framework. However, all the aforementioned statistical models/methods are parametric. In many circumstances, restricting models to be parametric limits the scope of statistical inference. For instance, robust methods based on the t-distribution effectively manage symmetrically heavy-tailed non-normal data, but are less efficient for skewed and other nonsymmetric data. To relax the parametric assumption, several semiparametric methods have been proposed. Yang and Dunson (2010) developed semiparametric SEMs, in which the distributions of latent variables or random errors were assumed unknown and were formulated through a truncated or centered Dirichlet process with stick-breaking priors; see also Song et al. (2010) and Chow et al. (2011). However, these semiparametric methods still encounter problems when the observed data are highly non-normal. Recently, Song and Lu (2012) developed a Bayesian nonparametric transformation SEM. Their extensive simulation studies and analyses of real data examples showed that the transformation SEM could effectively manage extremely non-normal data.

On the other hand, results of statistical inference could be misleading if they are unduly influenced by several model inputs such as data, model specification, and assumptions. Local influence analysis (Cook, 1986) examines whether the statistical results are sensitive to such minor perturbations of model inputs. A modification of a model input that severely influences the key analysis results is certainly a cause for concern. Therefore, local influence analysis has been an important research topic in applied statistics. Cook (1986) applied a unified method to linear regression analysis, and emphasized the generality and applicability of the method to a wide variety of problems with smooth likelihoods and perturbations. Since this pioneer work, local influence analysis has been applied to many statistical models, including various kinds of latent variable models. For example, Tanaka and Odaka (1989) thoroughly studied the sensitivity analysis in the context of factor analysis. Kwan and Fung (1998) studied local influence for specific restricted likelihoods under the factor analysis model. Lee and Wang (1996) developed influence measures for the more general covariance structure models. Zhu and Lee (2001) extended Cook's approach to incomplete data models. Lee and Tang (2004) applied the local influence approach to nonlinear SEMs.

Although the above mentioned methods have been demonstrated useful for statistical diagnostics of different models, they are mainly developed based on the ML approach. Very few studies have been conducted in the context of the Bayesian approach. Recently, Zhu et al. (2011) proposed a geometric approach to conduct Bayesian influence analysis. Given that outliers and/or influential points are usually not obvious for non-normal data, developing sound statistical diagnostic methods for such data is challenging and important. Motivated by the work of Zhu et al. (2011), we develop a Bayesian diagnostic procedure for the local influence analysis of transformation SEMs. Compared with conventional ML-based diagnostic procedures, the Bayesian approach could detect outliers and/or influential points in the observed data, as well as conduct model comparison and sensitivity analysis through various perturbations of the data, sampling distributions, or the prior distributions of model parameters. Latent variables and nonparametric transformation functions make the statistical analysis of the proposed model computationally demanding. The sampling-based feature of the Bayesian approach is especially appealing for the analysis of such complex models. However, to the best of our knowledge, no study has focused on the Bayesian diagnostics of transformation SEMs.

We illustrate our methodology through a multi-center prospective cohort study on the risk factors of osteoporotic fractures in older people. The main goal of this study is to examine the influence of serum concentration, precursors, and metabolites of sex hormones on bone mineral density (BMD). A total of 1460 Chinese men aged 65 years and above were recruited through private solicitation as well as public advertising from community centers and public housing estates. Prior medical knowledge suggests that multiple indicators rather than a single observed variable could reflect BMD and its determinants, such as estrogen, androgen, precursors, and metabolites. For instance, BMD is usually measured at both spine and hip, thereby reflected by the combined spine and hip BMD. Estrogen is likewise simultaneously measured by several indicators, including estrone (E1), estrone sulfate (E1-S), and estradiol (E2), which maintain female secondary sex characters. Considering the aforementioned nature of the data and high non-normality of some variables (see their histograms in Fig. 5), we propose a transformation SEM to investigate how the above determinants influence the level of BMD. Given that outliers or influential points are usually not obvious for non-normally distributed multidimensional data, we propose to use the Bayesian diagnostic procedure to detect outliers and/or influential observations, as well as conduct the model comparison and sensitivity analysis.

This paper is organized as follows. Section 2 defines the transformation SEM. The Bayesian P-splines approach for estimating the nonparametric transformation functions is also briefly described. Section 3 introduces the Bayesian diagnostic procedure, as well as the first- and second-order local influence measures with objective functions that are defined based on the logarithm of the Bayes factor. Similarly discussed are a number of perturbation schemes and the posterior computations of the associated Bayesian local influence measures. In Section 4, we conduct simulation studies to demonstrate the empirical performance of the proposed methodology. Section 5 reports the result of an actual application of the methodology on type 2 diabetes and its important determinants based on the Hong Kong Diabetes Registry. Section 6 concludes the paper. Technical details are provided in Appendices A and B.

2. Transformation structural equation models

2.1. Model description

For $i = 1, 2, \ldots, n$, let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})^T$ be a vector of observed variables and $\boldsymbol{\varpi}_i = (\varpi_{i1}, \ldots, \varpi_{iq})^T$ be a vector of latent variables, where $p > q$. The transformation SEM is defined by