Journal of Multivariate Analysis 135 (2015) 43–58

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

A Bayesian method for analyzing combinations of continuous, ordinal, and nominal categorical data with missing values

Xiao Zhang[a,*], W. John Boscardin[b], Thomas R. Belin[c], Xiaohai Wan[d,e], Yulei He[f], Kui Zhang[e]

[a] Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, United States

[b] Department of Medicine, University of San Francisco, United States

[c] Department of Biostatistics, UCLA Jonathan and Karin Fielding School of Public Health, United States

[d] AstraZeneca Pharmaceuticals LP, United States

[e] Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, United States

[f] Office of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, United States

ARTICLE INFO

Article history:
Received 3 September 2013
Available online 13 December 2014

AMS subject classification:
62H99

Keywords:
Multivariate probit model
Multinomial probit model
MCMC

ABSTRACT

From a Bayesian perspective, we propose a general method for analyzing a combination of continuous, ordinal (including binary), and categorical/nominal multivariate measures with missing values. We assume multivariate normal linear regression models for multivariate continuous measures, multivariate probit models for correlated ordinal measures, and multivariate multinomial probit models for multivariate categorical/nominal measures. Then we assume a multivariate normal linear model on the continuous vector comprised of continuous variables and those underlying normal variables for ordinal variables from multivariate probit models and for categorical variables from multinomial probit models. We develop a Markov chain Monte Carlo (MCMC) algorithm to estimate unknown parameters including regression parameters, cut-points for ordinal data from the multivariate probit models, and the covariance matrix encompassing both continuous variables and the underlying normal latent variables. Combining the continuous variables and the normal latent variables allows us to model combinations of continuous, ordinal, and categorical multivariate data simultaneously. The framework incorporates flexible priors for the covariance matrix, provides a foundation for inference about the underlying covariance structure, and imputes missing data where needed. The method is illustrated through simulated examples and two real data applications.

1. Introduction

Multivariate measures and longitudinal data arise in many fields of science. There is a long history of methodological development for analyzing multivariate continuous data from both classical and Bayesian perspectives, e.g. [31,63,34,62,15,59,11].

Statistical methods for analyzing multivariate ordinal (or polytomous) data including multivariate binary (or dichotomous) data have also been established. Generalized estimating equations (GEE) methods have propelled the development

[*] Corresponding author.
E-mail address: xiao.zhang@cshs.org (X. Zhang).

http://dx.doi.org/10.1016/j.jmva.2014.11.007
0047-259X/© 2014 Elsevier Inc. All rights reserved.

of classical methods for analyzing multivariate ordinal data, such as Zeger and Liang [66], Liang and Zeger [35], Prentice [52], Miller et al. [45], Qu et al. [53]. From a Bayesian perspective, Chib and Greenberg [9], Nandram and Chen [48], Liu [38], and Edwards and Allenby [18] analyzed multivariate ordinal data in the setting of the multivariate probit model.

Compared with multivariate continuous data and multivariate ordinal data, analyzing multivariate nominal categorical data is much less familiar. Liang et al. [36] performed multivariate regression analyses for categorical data using GEE with intensive computation. The multinomial probit model, which was developed for univariate nominal categorical data, has been generalized to the multinomial multiperiod probit model for multiple categorical data, e.g., [42,22,23,55,27]. To release the restriction on the covariance matrix of the multinomial multiperiod probit model and to generalize the multinomial probit model, Zhang et al. [69] proposed the multivariate multinomial probit model for multivariate nominal data with the multinomial probit model as a special case.

In practice, sometimes it is inevitable to analyze mixtures of multivariate continuous, ordinal and nominal measures, or various combinations of these three types of measures. This research field can be in general divided into several branches: (1) joint modeling using direct likelihood estimation and GEE methods; (2) joint modeling using general location models; (3) latent variable models (structural equation modeling, Bayesian latent variable models); (4) Other alternatives.

Joint modeling methods are to derive the joint distributions of the mixed outcomes and then to use GEE or quasi-likelihood methods to make statistical inference. Related references can be found in [6,54,24]. Modeling mixed measures using GEE include Zeger et al. [67], Legler et al. [33], and Spiess [61].

The general location model Olkin and Tate [50] has been popularized in analyzing mixed continuous and categorical data through specifying multinomial models for categorical variables and conditional multivariate normal models for continuous variables with different means across cells from those categorical variables and a common covariance matrix across cells. Using the general location model to analyze mixed measurements can be found in [37,19,59,13]. Liu and Rubin [39] extended the common covariance matrix to allow different, but proportional covariance matrices and replace the multivariate normal distribution specified for continuous variables by multivariate t distribution.

The structural equation model [30,47] has also been a popular tool to analyze mixed measures, as has path analysis [7], which refers to a similar use of linear models without latent variables. It assumes a continuous latent variable underlying several observed measures describing a common concept. The linear equation links the latent variable and the observed variables, i.e., factor analysis model is assumed for latent variables and the observed variables. Using the structural equation models to analyze mixed data types can be referred to Muthén [47], Arminger and Küsters [2], Shi and Lee [60], Lee and Zhu [32].

Using latent variable models is another active area for analyzing mixed measurement. Sammel et al. [58] proposed a latent variable mixed effects model by assuming a latent variable, linearly linked to the observed covariates, for each subject and the distribution for each type of measurement given the latent variable is from an exponential family. EM algorithm is applied to estimate the unknown parameters and the latent variables. Dunson [17] proposed Bayesian latent variable models for clustered mixed outcomes. The outcomes and the latent variables are linked through a known function and the latent variables are assumed to follow an exponential family. Adding random effect variables into the means of the specified exponential distribution accounts the correlated structure for mixed outcomes. Bayesian sampling algorithm can be derived to make statistical inference. Moustaki and Knott [46] proposed a latent trait model for various type of measurements. They assumed the distributions for mixed measurements given the latent variables are from the exponential family and estimate the unknown parameters and the latent variables using the maximum likelihood method. O'Malley et al. [51] combined the general location model and the latent trait model for mixed outcomes. Daniels and Normand [12] added latent variables to the model through the mean functions to estimate the correlations among different types of measurements. Goldstein et al. [25] proposed multilevel models for mixed data types. Weiss et al. [64] analyzed mixed outcomes through assuming various exponential distributions based on the type of the outcomes and then linearly linked the unknown mean functions with random effect variables to count for the correlated structure.

Besides the above general areas for analyzing the mixed measurements, there are other alternative methods, such as Miglioretti [44] used latent transition regression models for mixed outcomes and de Leon and Wu [14] proposed a copula-based regression models for a bivariate mixed outcome.

In this manuscript, we propose a joint modeling method using latent variables for mixed measures analyzed from a Bayesian perspective. We assume the multivariate probit model for multivariate ordinal (including binary) data. This means that there is an underlying normal latent variable for each ordinal outcome. We further assume the multivariate multinomial probit model for multivariate nominal data, where for each nominal outcome with p levels, there would be $p - 1$ normal latent variables. The multivariate multinomial probit model generalizes the multinomial probit model for the univariate nominal data. Detailed discussion about the multivariate multinomial probit model can be referred to Zhang et al. [69]. Then we combine these latent variables from multivariate ordinal and nominal measures with the multivariate continuous measures. We assume a multivariate linear model on those combined latent variables and the continuous measures and develop an MCMC algorithm to estimate the unknown parameters including the covariance matrix for the latent variables and the normal continuous variables.

Statistical methods for incomplete mixed data types are not well developed in the literature, and our paper offers expanded flexibility and generality. Although GEE methods could be considered in the present context, our proposed approach provides a foundation for drawing inferences relevant to the correlation structure of the data, and as such arguably represents an advance over GEE as well. Whereas Zhang et al. [69] describes an incomplete-data method for multivariate nominal