Transportation Research Part B 78 (2015) 341–363

Contents lists available at ScienceDirect

Transportation Research Part B

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

Introducing non-normality of latent psychological constructs in choice modeling with an application to bicyclist route choice

Chandra R. Bhat [a,b,*], Subodh K. Dubey [a,1], Kai Nagel [c,2]

[a] The University of Texas at Austin, Department of Civil, Architectural and Environmental Engineering, 301 E. Dean Keeton St. Stop C1761, Austin, TX 78712, United States
[b] King Abdulaziz University, Jeddah 21589, Saudi Arabia
[c] Institute for Land and Sea Transport (ILS), Transport System Planning and Traffic Telematics, TU Berlin Secr SG 12, Salzufer 17-19, D-10587 Berlin, Germany

ARTICLE INFO

Article history:
Received 19 September 2014
Received in revised form 15 April 2015
Accepted 20 April 2015
Available online 27 May 2015

Keywords:
Multivariate skew-normal distribution
Multinomial probit
ICLV models
MACML estimation approach
Bicyclist route choice

ABSTRACT

In the current paper, we propose the use of a multivariate skew-normal (MSN) distribution function for the latent psychological constructs within the context of an integrated choice and latent variable (ICLV) model system. The multivariate skew-normal (MSN) distribution that we use is tractable, parsimonious in parameters that regulate the distribution and its skewness, and includes the normal distribution as a special interior point case (this allows for testing with the traditional ICLV model). Our procedure to accommodate non-normality in the psychological constructs exploits the latent factor structure of the ICLV model, and is a flexible, yet very efficient approach (through dimension-reduction) to accommodate a multivariate non-normal structure across all indicator and outcome variables in a multivariate system through the specification of a much lower-dimensional multivariate skew-normal distribution for the structural errors. Taste variations (i.e., heterogeneity in sensitivity to response variables) can also be introduced efficiently and in a non-normal fashion through interactions of explanatory variables with the latent variables. The resulting model we develop is suitable for estimation using Bhat's (2011) maximum approximate composite marginal likelihood (MACML) inference approach. The proposed model is applied to model bicyclists' route choice behavior using a web-based survey of Texas bicyclists. The results reveal evidence for non-normality in the latent constructs. From a substantive point of view, the results suggest that the most unattractive features of a bicycle route are long travel times (for commuters), heavy motorized traffic volume, absence of a continuous bicycle facility, and high parking occupancy rates and long lengths of parking zones along the route.

1. Introduction

Economic choice modeling has continually seen improvements and refinements in specification, partly because of the availability of new techniques to estimate models. One such development is the incorporation of random taste heterogeneity

* Corresponding author at: The University of Texas at Austin, Department of Civil, Architectural and Environmental Engineering, 301 E. Dean Keeton St. Stop C1761, Austin, TX 78712, United States. Tel.: +1 512 471 4535; fax: +1 512 475 8744.

E-mail addresses: bhat@mail.utexas.edu (C.R. Bhat), subbits@gmail.com (S.K. Dubey), nagel@vsp.tu-berlin.de (K. Nagel).

[1] Tel.: +1 512 471 4535; fax: +1 512 475 8744.
[2] Tel.: +49 30 314 23308; fax: +49 30 314 26269.

http://dx.doi.org/10.1016/j.trb.2015.04.005
0191-2615/

(i.e., taste variations in response to explanatory variables) across decision makers using discrete (non-parametric) or continuous (parametric) or mixture (combination of discrete and continuous) random distributions for model coefficients. Such a specification also leads to correlations across alternative utilities when one or more random coefficients appear in the utility specifications of multiple alternatives. Early examples included studies by Revelt and Train (1996) and Bhat (1997), and there have now been many applications of this approach, using (primarily) latent-class multinomial logit and mixed multinomial logit formulations. A second development is the explicit consideration of latent psychological constructs (such as attitudes, perceptions, values and beliefs) within the context of economic choice models, which has the advantage (over the random taste heterogeneity approach) that it imparts more structure to the underlying choice process based on theoretical concepts and notions drawn from the psychology field. Additionally, it provides the opportunity to efficiently introduce random taste variations and the concomitant correlations across alternative utilities (we will come back to this latter point, which we believe has been less discussed and less exploited in the literature to date). This second development, commonly referred to as integrated choice and latent variable (ICLV) models (Ben-Akiva et al., 2002; Bolduc et al., 2005), may be viewed as a variation of the traditional structural equation methods (SEMs) (see, for example, Muthen, 1978, 1984) to accommodate an unordered-response outcome. Specifically, the traditional SEM includes a structural equation model for the latent variables (as a function of exogenous variables) as well as a measurement equation model that relates latent variables to observed continuous, binary, or ordered-response indicator variables. The ICLV model, conceptually speaking, adds an unordered-response outcome variable that may be considered as another indicator variable in the measurement component of the traditional SEM (except that the measurement component typically does not include exogenous variables, while the unordered-response choice variable is modeled as a function of exogenous variables).

Another area of intense research in the recent past, but originating more from the statistical field, is the consideration of non-normal distributions in modeling data. This has been spurred by the increasing presence of multi-dimensional data that potentially exhibit non-normal features such as asymmetry, heavy tails, and even multimodality. Parametric approaches to accommodate non-normality span the gamut from finite mixtures of normal distributions to skew-normal distributions (and more general skew-elliptical distributions) to mixtures of skew-normal distributions (and mixtures of more general skew-elliptical distributions). Some recent applications include Pyne et al. (2009), Lachos et al. (2010), Contreras-Reyes and Arellano-Valle (2013), Riggi and Ingrassia (2013), Lin et al. (2013) and Vrbik and McNicholas (2014). Many of these recent studies use either a multivariate skew-normal or a skew-t distribution as the basis for accommodating non-normality, with different proposals on how to characterize these skew distributions (see Lee and McLachlan (2013) for a recent review and synthesis of the many different proposals). In the context of the multivariate skew-normal distribution, broadly speaking, there are two forms – restricted and unrestricted, with what Lee and McLachlan characterize as "extended" and "generalized" being relatively minor generalizations of the restricted and unrestricted forms. However, it is well recognized now that the underlying basis for all of the different proposals for the multivariate skew-normal distribution originate in the pioneering work of Azzalini and Dalla Valle (1996). Arellano-Valle and Azzalini (2006) provided a unified framework to characterize the many other proposals since Azzalini and Dalla Valle (1996), and showed how their unified skew-normal (SUN) distribution includes all other proposals as special cases. Thus, in this research, we will maintain notations that correspond to the SUN distribution.

In the current paper, we bring together the two developments discussed above – the ICLV model structure and the treatment of non-normality through a multivariate skew-normal or MSN distribution specification. In particular, we allow the latent constructs in the ICLV model to be skew-normal. After all, there is no theoretical basis for specifying these constructs as normal (as is typically assumed in the literature); thus, there is substantial appeal in specifying a more general non-normal specification that is then characterized empirically. To our knowledge, this is the first probit-kernel based ICLV model proposed in the econometric literature, which has several important features.[3] First, it recognizes the very real possibility that latent variables are non-normally distributed after conditioning on exogenous variables. Imposing normality when the structural errors in the latent variable relationship with exogenous variables are non-normal can render the parameter estimates inconsistent in the measurement equations corresponding to binary or ordinal indicators, as well as in the unordered outcome model (this is because of the non-linear nature of the relationship with the latent variable; see Geweke and Keane, 1999; Caffo et al., 2007; Wall et al., 2012). Of course, this inconsistency will permeate into the coefficients of the structural component (relating the latent variables to exogenous variables) because these structural coefficients are being implicitly estimated through the relationship embedded in the measurement equations. Incorrectly imposing normality will also lead, in general, to inefficient estimation in all of the ICLV model components and can lead to incorrect inferences. Second, our proposal to include non-normality exploits the latent factor structure of the ICLV model. That is, our approach constitutes a flexible, yet very efficient approach (through dimension-reduction) to accommodate a multivariate non-normal structure across all indicator and outcome variables through the specification of a much lower-dimensional multivariate skew-normal distribution for the structural errors. This leads to parsimony in the additional parameters introduced because of non-normality. Third, taste variations

[3] Brey and Walker (2011) is the only other study we are aware of that considers a non-normal distribution within an ICLV model. However, they use a single latent variable, and their approach adds to convergence problems to what is already a difficult convergence problem in the context of a logit-kernel based formulation. Specifically, the integrand in their integration involves an increasingly complicated mixing. Even in the typical normally-mixed latent variables, convergence is not easy and it is not uncommon for the estimation to simply not converge (see Alvarez-Daziano and Bolduc, 2013; Bhat and Dubey, 2014). On the other hand, our flexible skew-normal distribution, combined with our proposed estimation technique, does make the estimation simpler and easier in a probit-based kernel context, even with multiple latent variables.