Hybrid choice models, structure, identification and estimation

Thijs Dekker*
Institute for Transport Studies
University of Leeds, Leeds, UK
Stephane Hess, Andrew Daly
Institute for Transport Studies
University of Leeds, Leeds, UK
* E-mail: t.dekker@leeds.ac.uk

Abstract:

For many years, discrete choice models and the underlying Random Utility framework have been at the core of modelling various choice processes in transport, health and environmental economics [1,2,8]. Typically, these models use a set of exogenous covariates in explaining observed choices and possibly allow for heterogeneity in preferences across respondents. More recently, there has been growing interest in making use of additional information which is, however, not observable in the form of exogenous variables. This includes, for example, the role of attitudes, convictions and other personal traits. In this case, the analyst will only have indicators of these underlying latent variables at his disposal. Using these indicators directly in the model to explain the choice process might put the researcher at the risk of endogeneity bias and measurement error [e.g. 6,7]. A more natural approach is therefore to treat such indicators as a dependent variable affected by a latent variable. The latter then acts as an endogenous covariate in one (or more) choice models [e.g. 9]. Hybrid choice models, as a more general class of models, use latent variables to link various behavioural responses, including choice models, and allow to trace the impact of socio-economic characteristics and other context variables on these responses both directly and indirectly through the latent variables. This type of models is increasing in popularity due to its capability of enhancing the behavioural representation of the choice process [e.g. 4].

Estimation of hybrid choice models is complex, because the various individual model components are combined (and linked) in a unified framework after which the parameters are estimated simultaneously. Estimation of the individual model components is in most cases straightforward when estimated independently and depends on the response format, e.g. a simple ordered model might be used for Likert-scale responses. The introduction of latent variables, however, brings about identification issues [e.g. 3,5] and correlation structures between the various model components which are currently not fully understood in the literature. As a result, some empirical applications have estimated models that are over-specified and (or) incorrectly interpreted the role of

the parameters tracing the impact of the latent variable on the decision process. Other applications have not allowed for the full degree of flexibility, notably in terms of the correlation structure between the indicators.

We develop a generic modelling framework comprising three layers. The bottom layer consists of observed respondent characteristics and other context variables all serving as explanatory variables. The top layer represents various decision environments, being observed choices, responses to follow-up questions or other stated or revealed preferences. Initially, there are no links between the various decision environments. A direct link is established between the bottom and top layer measuring the direct impact of the explanatory variables in each decision environment. The middle layer introduces the unobserved latent variables as an alternative driver of the choice process. Hence, a latent variable is explained by the bottom layer and serves as an explanatory variable in one (or more) decision environment. The impact of a latent variable on the top layer is traced by means of a scaling parameter. Since this scaling parameter multiplies both the structural and the random part of the latent variable, correlation is introduced in the error structure of the various models in the top layer. The latter effect is generally unrecognized within the literature and affects the interpretation of the respective parameter. In fact, estimation may become problematic because the scaling parameter is overworked.

In this paper we revisit the identification requirements for various model types based on a range of response formats. Specifically, we look at the extent to which these identification requirements are affected by the introduction of latent variables. In this we extend the work initialized by Ben-Akiva et al. [3], Bolduc et al. [5] and Daly et al. [6] and highlight the importance of including explanatory variables for the various latent variables for identification. Moreover, it turns out identification is not affected by introducing correlation across various indicators or measurement equations. Such correlation can be expected when several indicators attempt to measure the same latent variable. By doing so, we are able to purely trace the impact of the latent variable on the decision process, without worrying about (potential) additional correlation introduced by its associated scaling parameter.

The second part of the paper looks into the application of hybrid choice models. Here, we illustrate that the general model structure is not affected by using alternative response formats, i.e. econometric models, in the top layer of the model. First, simulated data are applied to illustrate the various theoretical issues raised above and stress the risk of running into empirical identification issues. Second, two empirical datasets are analysed. The first dataset is from a public transport route choice study, containing a large number of ordered and binary indicators. The second dataset focuses on Willingness-To-Pay for flood risk reductions and combines both the use of a stated choice experiment with an open ended contingent valuation question in the same survey.

Overall, the paper aims at making a substantial contribution to understanding the connections between the different model parts in hybrid choice models in order to avoid unnecessary theoretical

and empirical identification issues. Its results are therefore of interest to a generic public interest in choice processes with underlying latent variables, or even broader choice processes which are linked to each other in various ways.

References:

[1] de Bekker-Grob, E.W. Ryan, M. and Gerard, K., "Discrete choice experiments in health economics: a review of the literature", Health Economics, 21 (2), 145-172, 2012.

[2] Ben-Akiva, M. and Lerman, S. "Discrete Choice Analysis: theory and application to travel demand", MIT Press, 1985.

[3] Ben-Akiva, M., Walker, J. L., Bernardino, A. T., Gopinath, D. A., Morikawa, T., and Polydoropoulou, A., "Integration of choice and latent variable models", Massachusetts Institute of Technology, Cambridge, MA, 1999.

[4] Bolduc, D. and Alvarez-Daziano, R., "On the estimation of hybrid choice models", in: Choice modelling: the-state-of-the-art and the state-of-practice, Hess, S. and Daly, A., Chapter 11, pp. 259-287, Emerald Group Publishing, 2010.

[5] Bolduc, D., Ben-Akiva, M., Walker, J. L., and Michaud, A., "Hybrid choice models with logit kernel: applicability to large scale models," in Integrated Land-Use and Transportation Model: Behaviroual Foundations, M. Lee-Gosselin & S. Doherty, eds., Elsevier, Oxford, pp. 275-302, 2005.

[6] Daly A., Hess S., Patruni B., Potoglou D. and Rohr C., "Using ordered attitudinal indicators in a latent variable choice model: a study of the impact of security on rail travel behaviour", *Transportation*, 39 (2), pp. 267-297, 2011.

[7] Dekker, T. Hess, S. Brouwer, R. and Hofkes, M.W. "Implicitly or explicitly uncertain?" , ITS working paper, University of Leeds, 2012.

[8] Hanley, N., Wright, R.E. and Adamowicz, V., "Using choice experiments to value the environment", Environmental and Resource Economics, Special Issue Frontiers of Environmental & Resource Economics: Testing the Theories, 11(3–4), 413–428, 1998.

[9] Rungie, C. Coote, L. and Louviere, J., "Structural choice modelling: theory and applications to combining choice experiments", Journal of Choice Modelling, 4(3), 1-29, 2011.