Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

RprobitB: Bayes Estimation of Discrete Choice Behavior Heterogeneity via Probit Models in R

Lennart Oelschläger
Bielefeld University

Dietmar Bauer
Bielefeld University

Abstract

RprobitB is an R package for Bayes estimation of probit models with a special focus on modeling choice behavior heterogeneity. In comparison to competing packages it places a focus on approximating the mixing distribution via a latent mixture of Gaussian distributions and thereby providing a classification of deciders. It provides tools for data management, model estimation via Markov Chain Monte Carlo Simulation, diagnostics tools for the Gibbs sampling and a prediction function. This paper demonstrates the functionalities of **RprobitB** on known choice datasets and compares estimation results across packages.

Keywords: discrete choice, probit models, heterogeneity, Bayes estimation, R.

1. Introduction

The multinomial probit model is one of the most widely-used statistical models to explain the choices that individuals make among a discrete set of alternatives, which is of central interest in many scientific areas, for example in transportation and marketing. In many such choice scenarios it is reasonable to assume, that the preferences of the decision makers are non-homogeneous. Based on personal characteristics, deciders generally weight attributes like time and cost differently. Heterogeneity in choice behavior can be modeled using mixing distributions for the coefficients. Recently, Oelschlaeger and Bauer proposed a new instrument for approximating the underlying mixing distribution that combines Bayes estimation and semi-parametric methods. This paper presents the implementation of the methodology in the R package **RprobitB**.

Traditionally, discrete choice models are interpreted as random utility models, including the multinomial logit (MNL) and the multinomial probit (MNP) model as the most prominent members. The MNL model affords straightforward analysis but suffers from the well-known

independence of irrelevant alternatives assumption. In contrast, the MNP model avoids this assumption, which however comes at the price of more complex parameter estimation, cf. Train (2009). In their basic form, these models often fail to take into account heterogeneity of individual deciders, cf. Train (2009), Chapter 6, or Train (2016). A concrete example of heterogeneous preferences is constituted by the value of travel time, cf. Cirillo and Axhausen (2006). Modeling heterogeneity in preferences is indispensable in such cases and has been elaborated in both the MNL and the MNP model by imposing mixing distributions on the coefficients, cf. Train (2009) and Bhat (2011).

Specifying these mixing distributions is an important part of the model selection. In absence of alternatives, it has been common practice so far to try different types of standard parametric distributions (including the normal, log-normal, uniform and tent distribution) and to perform a likelihood value-based model selection, cf. Train (2009), Chapter 6. Aiming to capture correlation patterns across parameters, Fountas, Anastasopoulos, and Abdel-Aty (2018) and Fountas, Pantangi, Hulme, and Anastasopoulos (2019) apply multivariate normal mixing distributions in their probit models, which however comes at the price of imposing the rather strong normality assumption on their parameters.

In order to alleviate these restrictions Train (2016) proposes a non-parametric approach based on grid methods. Building on the ideas of Train (2016) and Bhat and Lavieri (2018) recently Bauer, Büscher, and Batram (2019) introduced procedures for non-parametrically estimating latent class mixed multinomial probit models where the number of classes is chosen iteratively in the algorithm. These procedures have been demonstrated to be useful in reasonable sized cross-sectional data sets. However, for large panel data sets with a significant number of choice occasions per person, the approach is numerically extremely demanding in particular due to its non-parametric nature and has to deal with the curse of dimensionality.

In the Bayesian framework Scaccia and Marcucci (2010) presents the idea to estimate latent class logit models with a fixed prespecified number of Gaussian components. This approach does not require the maximization of the likelihood while at the same time it allows for approximation of the underlying mixing distribution. The same idea has also been applied to probit models, cf. Xiong and Mannering (2013) for an analysis of adolescent driver-injury data. In both cases however, the specification of the number of latent classes is based only on a trial-and-error strategy.

Oelschlaeger and Bauer presents a more flexible approach that combines the ideas of a Bayesian framework, approximating the mixing distribution through a mixture of normal distributions and updates on the number of latent classes within the algorithm analogously to Bauer *et al.* (2019). As a consequence, the procedure unites the benefits of a reduced numerical complexity for the estimation compared to the non-parametric likelihood maximization approach and the ability to approximate any mixing distribution. Presenting simulation results on artificial test cases, it is shown that the approach is capable of approximating the underlying mixing distributions and thereby guiding the specification of mixing distributions for real-world applications.

This packages adds to the line of discrete choice software packages in R in the following way: Its focus is entirely on Bayesian estimation, thereby it differs from the packages Rchoice. Furthermore, it places a focus on modeling choice behaviour heterogeneity by approximating the underlying mixing distribution through a latent mixture of normal distributions. The method is explained in detail in Oelschlaeger and Bauer.

In this article we present the methodology, give an overview over the functionality of the package and apply the package to datasets. Some of them were already analysed and we aim to reconstruct their findings. In addition, we added two datasets that are especially appropriate for the RprobitB package in modeling choice behaviour heterogeneits. The first one is a dataset of contraception choice from the German family panel pairfam. It contains repeated observations of males and femals over several years having different social demographics and relationship status choosing different means of contraception. This choice a priori can be considered to be very hetereogenous and dependent on factors not directly observable by the researcher. The second application deals with the opening choice of chess players depending on their and their openents playing strenght measured in the popular measure system Elo, their gender and nationality. Like the contraception example, this choice a priori can be considers to depend on psychological factors that are not directly observable by the researcher. By applying the functionality of this package we demonstrate how we are able to classify the players into different categories of playing style.

2. The method

3. Applications

3.1. Choice of contraception

3.2. Chess opening

4. Conclusion

This paper addressed the problem of specifying mixing distributions in the multinomial probit model with a panel data setting, constituting an important part of the model selection for which the literature does not provide much guidance so far. In the absence of alternatives, many applications of the mixed multinomial probit model rely on different types of standard parametric distributions for modelling heterogeneity in preferences in combination with a likelihood-value based model selection. This course of action is very restrictive and imposes strong assumptions on the distribution of the model parameters that could potentially lead to misspecification and biased estimates.

We proposed a new approach that improves the current specification strategies in several ways: First, our approach does not require any distributional assumption, since the latent class setting is flexible enough to approximate practically any distribution shape and allowing for any correlation pattern. Second, the weight-based updating scheme ensures that the number of latent classes does not need to be prespecified. Third, the imposed Bayesian estimation framework avoids many numerical problems that occur in the frequentist approach. Most notably, no likelihood function has to be evaluated nor approximated. Comparing the numerical estimation speed to the non-parametric frequentist approach of Bauer *et al.* (2019), we found that our implementation of the Bayesian approach is at least 10 times faster. The

improvement becomes more distinct for panel data settings with a high number of choice occasions. This is due to the fact that for given total sample size NT a large T is beneficial for the Bayesian approach as then the number of vectors $\beta_n,\ n = 1, ..., N$ is comparatively small, while in the frequentist approach calculating the likelihood becomes more challenging for increasing the number T of choice situations faced by each of the N individuals. On the other hand, the grid based frequentist approach of Bauer *et al.* (2019) can potentially achieve a better approximation (especially of point masses) due to the relatively high number of latent classes. However, this approach requires that a suitable grid is set prior to the estimation with a specification of upper bounds for the coefficients. Additionally, the curse of dimensionality plays a crucial role, which is less of a burden in the Bayesian approach. Note that for a fully specified parametric structure these concerns do not play such a big role also for the frequentist approach.

Our simulation results verified that the proposed approach achieves good approximations of the mixing distribution in common choice modelling situations, where the underlying heterogeneity cannot be captured by standard parametric distributions. It would be interesting to apply the approach also to empirical data in the future. Additionally, further research on how to properly address sign-restricted coefficients is required.

Computational details

The results in this paper were obtained using R 4.1.3 with the **RprobitB** 1.0.0.9000 package. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/.

Acknowledgments

This work has been financed partly by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 356500581 which is gratefully acknowledged.

References

Bauer D, Büscher S, Batram M (2019). "Non-parameteric estiation of mixed discrete choice models." *Second International Choice Modelling Conference in Kobe*.

Bhat C (2011). "The Maximum Approximate Composite Marginal Likelihood (MACML) Estimation of Multinomial Probit-Based Unordered Response Choice Models." *Transportation Research Part B: Methodological*, **45**.

Bhat C, Lavieri P (2018). "A new mixed MNP model accommodating a variety of dependent non-normal coefficient distributions." *Theory and Decision*, **84**.

Cirillo C, Axhausen K (2006). "Evidence on the distribution of values of travel time savings from a six-week diary." *Transportation Research Part A: Policy and Practice*, **40**.

Fountas G, Anastasopoulos PC, Abdel-Aty M (2018). "Analysis of accident injury-severities using a correlated random parameters ordered probit approach with time variant covariates." *Analytic Methods in Accident Research*, **18**.

Fountas G, Pantangi SS, Hulme KF, Anastasopoulos PC (2019). "The effects of driver fatigue, gender, and distracted driving on perceived and observed aggressive driving behavior: A correlated grouped random parameters bivariate probit approach." *Analytic Methods in Accident Research*, **22**.

Scaccia L, Marcucci E (2010). "Bayesian Flexible Modelling of Mixed Logit Models." *Proceedings from the 19th International Conference on Computational Statistics*.

Train K (2009). *Discrete choice methods with simulation*. 2. ed. edition. Cambridge Univ. Press.

Train K (2016). "Mixed logit with a flexible mixing distribution." *Journal of choice modelling*, **19**.

Xiong Y, Mannering FL (2013). "The heterogeneous effects of guardian supervision on adolescent driver-injury severities: A finite-mixture random-parameters approach." *Transportation Research Part B: Methodological*, **49**.

Affiliation:

Lennart Oelschläger
Department of Business Administration and Economics
Bielefeld University
Postfach 10 01 31
E-mail: lennart.oelschlaeger@uni-bielefeld.de

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd