Rodenburger, Daniel

Article — Published Version

Refining analytic approximation based estimation of mixed multinomial probit models by parameter selection

Metrika

Provided in Cooperation with:
Springer Nature

Suggested Citation: Rodenburger, Daniel (2023) : Refining analytic approximation based estimation of mixed multinomial probit models by parameter selection, Metrika, ISSN 1435-926X, Springer, Berlin, Heidelberg, Vol. 87, Iss. 4, pp. 411-425, https://doi.org/10.1007/s00184-023-00920-6

This Version is available at:
https://hdl.handle.net/10419/305217

Metrika (2024) 87:411–425
https://doi.org/10.1007/s00184-023-00920-6

Refining analytic approximation based estimation of mixed multinomial probit models by parameter selection

Daniel Rodenburger[1]

Received: 11 January 2022 / Accepted: 6 July 2023 / Published online: 28 July 2023

Abstract
Applying analytic approximations for computing multivariate normal cumulative distribution functions has led to a substantial improvement in the estimability of mixed multinomial probit models, both in terms of accuracy and especially in terms of computation time. This paper makes a contribution by presenting a possible way to improve the accuracy of estimating mixed multinomial probit model covariances based on the idea of parameter selection using cross-validation. Comparisons to the MACML approach indicate that the proposed parameter selection approach is able to recover covariance parameters more accurately, even when there is a moderate degree of independence between the random coefficients. The approach also estimates parameters efficiently, with standard errors tending to be smaller than those of the MACML approach, which can be observed by means of a real data case.

Keywords Analytic approximation · Cross-validation · Discrete choice models · Mixed multinomial probit · Parameter selection

1 Introduction

Statistical discrete choice methods are an effective way to describe, understand and predict individual choice behavior. The decision makers may be individuals, households, firms, or any other decision-making entity, and the alternatives may represent competing products, courses of action, or other options or elements about which a decision must be made. These methods are used in empirical social research and beyond, for example in the transport sector (Train 2009; Savolainen et al. 2011).

Many different approaches can be considered to explain the role of individual features and to predict choice probabilities when individuals choose from a finite set of discrete alternatives. A widely used model in applied econometrics of discrete choice

✉ Daniel Rodenburger
daniel.rodenburger@uni-jena.de

[1] University of Jena, Jena, Germany

analysis is the multinomial logit model introduced by Luce (1965) and McFadden (1973). Although this approach has a simple and elegant structure, it also requires the assumption of independence from irrelevant alternatives, which means that the ratio of the choice probabilities of two alternatives is independent of the characteristics of the other alternatives in the choice set. As a result, extensions to the multinomial logit model were introduced that relaxed the assumption of independently and identically distributed errors across alternatives. Common extensions are the class of generalized extreme value models proposed by McFadden (1978) and the multinomial probit model, which allow for comparatively flexible error covariance structures up to certain limits of identifiability (Train 2009). In mixed multinomial probit modeling, random coefficients can be considered which are able to reflect random taste variation across decision makers.

However, in the absence of a closed form likelihood due to resulting multidimensional normal integrals within the process of model building, this essential advantage was usually counterbalanced by computationally burdensome simulation-based estimation approaches for mixed multinomial probit models in the past, especially when considering large choice set and many choice occasions. A way to overcome this drawback is to apply analytic approximations for multidimensional normal integrals. Instead of the actual likelihood in its non-closed form, an analytic approximation of this likelihood is used, which can be maximized efficiently over a given parameter space using gradient-based standard optimization methods. This is proposed by Joe (1995), who refines the analytic approximation method of Solow (1990), but without testing it in the context of estimating probit models. Bhat (2011) takes up this idea with the introduction of his marginal approximate composite maximum likelihood (MACML) approach for mixed multinomial probit models, which is based on the analytic approximation method of Solow (1990) and Joe (1995) (SJ).

In a companion simulation study Bhat and Sidharthan (2011) examine the performance of the MACML approach in the context of estimating parameters of mixed multinomial probit models for cross-sectional and panel data. The results indicate that the analytic approximation based estimators provide parameter estimates, which are very close to the true parameter values. The reported estimation errors are smaller than those of the applied maximum simulated likelihood approach (MSL). Furthermore, the MACML approach provides smaller standard errors and is noted to be much faster[1] than the considered MSL approach.

Patil et al. (2017) provide a simulation evaluation where different estimation techniques for a mixed multinomial probit model with five alternatives are compared. They consider cross-sectional and panel data sets for scenarios with and without correlation among the random coefficients and compare a range of different approaches: the MACML approach; the Geweke–Hajivassiliou–Keane (GHK) simulator with Halton sequences and full information maximum likelihood; the GHK approach implemented in conjunction with the composite marginal likelihood estimation approach; the GHK approach with sparse grid nodes and weights, implemented in conjunction with the composite marginal likelihood estimation approach; and a Bayesian Markov Chain

[1] Bhat and Sidharthan (2011) report that, for the case of five random coefficients, the MACML approach is about 50 times faster than the considered maximum simulated likelihood approach.

Monte Carlo approach (MCMC). The results suggest that the MACML approach provides the best performance in terms of estimation accuracy and estimation time for all data-generating settings among the considered approaches. Batram and Bauer (2019) provide similar results compared to the MSL approach as part of a simulation evaluation for a cross-sectional mixed multinomial probit model with six alternatives. In addition, they consider different analytic approximation methods for the MACML approach: the SJ approximation method; the method of Mendell and Elston (1974) (ME) together with a modification of this method by Trinh and Genz (2015) (BME); and a selected method of Bhat (2018) (TVBS), who presents four new algorithms.

A closer look at the MACML results shows that there are differences between the parameters in terms of estimation errors: the estimation of the covariances between the random coefficients results in higher estimation errors than the estimation of their mean values (Bhat and Sidharthan 2011; Patil et al. 2017; Batram and Bauer 2019). To avoid singularities during optimization, the correlations between the random coefficients are not estimated directly, but via the corresponding Cholesky decomposition. Here again differences in the estimation errors become apparent. While the results of Bhat and Sidharthan (2011) and Batram and Bauer (2019) suggest that the estimation errors are smaller when there is no correlation between two random coefficients than when they are correlated, on closer examination the opposite seems to be the case. This apparent contradiction results from the fact that Bhat and Sidharthan (2011) and Batram and Bauer (2019) report an absolute percentage bias for each parameter, where the estimated absolute bias is adjusted by the true parameter value. Although this procedure may be justified with regard to the comparability of estimation errors, it is not applicable in the case of no correlation between two random coefficients. In these cases Bhat and Sidharthan (2011) and Batram and Bauer (2019) waive the adjustment[2] and report the difference in absolute value between mean estimate and the true value.

As a result, for Bhat and Sidharthan (2011) it appears at first that the reported mean estimation errors in the case of no correlation between two random coefficients are smaller than when they are correlated, 0.029 versus 0.076. However, if only mean absolute estimation errors are considered for all parameters to be estimated, this impression seems to be reversed, 0.029 versus 0.023. The absolute errors in the case of the true zero Cholesky parameters are thus on average over 25% higher than those of the non-zero entries. This observation is confirmed by Batram and Bauer (2019), where those errors are on average more than 30% larger. For ME, BME, and TVBS the absolute errors for the true zero Cholesky parameters are even higher: 35%, 45%, and almost 50%, respectively. Batram and Bauer (2019) provide standard and normalized MACML estimation results, where in the latter case choice proportions are normalized to sum to one. This approach results in a substantial increase in computation time. However, it also leads to significantly more accurate estimation results for ME and BME. The performance of SJ and TVBS are more or less unaffected by the normalization. The reported values refer to standard MACML estimation results for SJ and TVBS and normalized MACML estimation results for ME and BME.

[2] Since Patil et al. (2017) choose the same presentation of estimation errors, this probably applies to their study as well.

This paper introduces a refined method for estimating the covariances between the random coefficients of the mixed multinomial probit model for cross-sectional data. In particular, the estimation of the zero entries of the Cholesky decomposition shall be improved by parameter selection, where the selection mechanism is based on the idea of cross-validation.

The remainder of this study is structured as follows: Sect. 2 presents the mixed multinomial probit model for cross-sectional data. Section 3 introduces the proposed parameter selection approach. The theoretical part of this study is followed by an application part in Sect. 4, in which the effectiveness of the previously introduced parameter selection approach is examined by means of its ability to recover parameters from finite samples. Section 4.1 presents the experimental design. The estimation results are provided in Sect. 4.2 and allow for comparisons to the MACML approach. Other estimation approaches are not considered based on the results of Patil et al. (2017). The application part is closed with a real data case, which is located in an online appendix.[3] Sect. 5 concludes.

2 Mixed multinomial probit model

In the following the mixed multinomial probit model for cross-sectional data is presented. The notation refers to Bhat (2011).

If individuals $\{1, \ldots, Q\}$ are considered who are choosing from a set of alternatives $\{1, \ldots, I\}$, then the utility that individual q associates with alternative i can be modeled as $u_{qi} = \beta_q^\intercal x_{qi} + \epsilon_{qi}$, where x_{qi} denotes a K-dimensional vector of exogenous characteristics of alternative i and β_q is its individually specific valuation. For each individual $q \in \{1, \ldots, Q\}$ the coefficient vector β_q is assumed to be drawn from a multivariate normal distribution, i.e. $\beta_q \sim \mathcal{N}(b, \Omega)$. Thus, the mixed multinomial probit model overcomes limitations of the ordinary probit model by allowing for random taste variation. Furthermore, ϵ_{qi} is assumed to be independent and identically distributed with $\epsilon_{qi} \sim \mathcal{N}(0, \frac{1}{2})$ for model identification purpose, since the decision on one particular alternative is scale invariant. The variance is set to one-half to normalize the variances of the latent utility differences errors, which are typically not identified. This can be seen in the following.

In the case where individual q decides to choose alternative m, the corresponding utility differences are given by $y_{qim} := u_{qi} - u_{qm} = \beta_q^\intercal (x_{qi} - x_{qm}) + \epsilon_{qi} - \epsilon_{qm}$, where $i = 1, \ldots, I$ and $i \neq m$. Note that individual q chooses alternative m, if and only if $y_{qim} < 0$ holds for all $i \neq m$. In practice, y_{qim} can not be observed and $y^*_{qim} := \mathbb{1}_{\{y_{qim} < 0\}}$ has to be examined instead. The likelihood contribution of individual q for choosing alternative m depends upon the proportion of K to I. If $K < I - 2$, it is convenient to state the likelihood as

[3] The online appendix can be found at the following https://doi.org/10.5281/zenodo.8104188.

$$L_q(b, \Omega) = \int_{\mathbb{R}^K} \left(\int_{\mathbb{R}} \left[\prod_{i \neq m} \left[\Phi \left(\left[-\sqrt{2}(\beta^\intercal z_{qim}) \right] + \lambda \right) \right] \right] \phi(\lambda) d\lambda \right) \phi_K(\beta \mid b, \Omega) d\beta,$$

where $z_{qim} := x_{qi} - x_{qm}$, $\lambda := \sqrt{2}\epsilon_{qm}$, and $\phi_K(\beta \mid b, \Omega)$ denotes the K-variate normal density function at β with mean b and positive definite covariance matrix Ω. The standard univariate normal density and cumulative distribution function are denoted by $\phi(\cdot)$ and $\Phi(\cdot)$, respectively. In case of $K > I - 2$, let $y_{qm} := \left(y_{qm1}, \ldots, y_{qm(m-1)}, y_{qm(m+1)}, \ldots, y_{qmI}\right)^\intercal \in \mathbb{R}^{I-1}$, $Z_{qm} := \left(x_{q1} - x_{qm}, \ldots, x_{q(m-1)} - x_{qm}, x_{q(m+1)} - x_{qm}, \ldots, x_{qI} - x_{qm}\right) \in \mathbb{R}^{K \times (I-1)}$, and the error differences $\epsilon_{qm} := \left(\epsilon_{q1} - \epsilon_{qm}, \ldots, \epsilon_{q(m-1)} - \epsilon_{qm}, \epsilon_{q(m+1)} - \epsilon_{qm}, \ldots, \epsilon_{qI} - \epsilon_{qm}\right)^\intercal \in \mathbb{R}^{I-1}$. The vector of utility differences can then be expressed as $y_{qm} = Z_{qm}^\intercal \beta_q + \epsilon_{qm}$. Since β_q and ϵ_{qm} are assumed to be normally distributed, y_{qm} is also normally distributed with expectation $\mathbb{E}(y_{qm}) = Z_{qm}^\intercal b$ and covariance matrix

$$\mathbb{V}(y_{qm}) = Z_{qm}^\intercal \Omega Z_{qm} + \begin{pmatrix} 1 & 0.5 & \ldots & 0.5 \\ 0.5 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0.5 \\ 0.5 & \ldots & 0.5 & 1 \end{pmatrix} =: \Sigma_q.$$

Then the likelihood contribution of individual q for choosing alternative m is given by $L_q(b, \Omega) = \Phi_{I-1}\left(-Z_{qm}^\intercal b \mid 0, \Sigma_q\right)$, where $\Phi_{I-1}\left(-Z_{qm}^\intercal b \mid 0, \Sigma_q\right)$ denotes the $(I-1)$-variate normal cumulative distribution function at $-Z_{qm}^\intercal b$ with zero mean and covariance matrix Σ_q, whose invertibility is guaranteed by $K > I - 2$. However, due to singularity problems during the estimation process, a direct estimation of the entries of Ω is usually avoided. Instead, the entries of L, the lower triangular matrix of the Cholesky decomposition $\Omega = LL^\intercal$, are estimated (Bhat and Sidharthan 2011; Patil et al. 2017; Batram and Bauer 2019). These can be merged into the vector $l := \left(L_{11}\ L_{21}\ L_{22}\ L_{31}\ \ldots\ L_{KK}\right)$. Thus, $L_q(b, \Omega)$ becomes $L_q(b, l)$.

Given the likelihood contribution for individual q who chooses alternative m, the log-likelihood can be stated as $\mathcal{L}(b, l) = \sum_{q=1}^{Q} \mathcal{L}_q(b, l)$, where $\mathcal{L}_q(b, l) := \log\left(L_q(b, l)\right)$ and alternative m now represents the chosen alternative for every individual. These alternatives may of course vary across individuals. Since the normal cumulative distribution function cannot be expressed in closed form, the estimation of b and l is usually based on simulation approaches, which typically suffer from computational loads. This shortcoming can be addressed by the use of analytic approximation methods (Joe 1995; Bhat 2011). While Bhat and Sidharthan (2011) and Patil et al. (2017) rely on the SJ approach, Batram and Bauer (2019) additionally consider ME, BME, and TVBS. Among these approaches, SJ provides a very good ratio between computation time and accuracy. Hence, SJ is also applied in this paper.

3 Parameter selection approach

The estimation method proposed in the following is intended to improve the estimation of true zero Cholesky parameters of the mixed multinomial probit model outlined in Sect. 2. The approach is based on the idea of parameter selection and cross-validation.

Consider again the case $K > I - 2$ from Sect. 2, where K denotes the number of exogenous characteristics for each alternative $i \in \{1, \ldots, I\}$ together with the corresponding log-likelihood

$$\mathcal{L}(b, l) = \sum_{q=1}^{Q} \mathcal{L}_q(b, l) = \sum_{q=1}^{Q} \log\left(L_q(b, l)\right),$$

where $L_q(b, l) = \Phi_{I-1}\left(-Z_{qm}^{\mathsf{T}} b \mid 0, \Sigma_q\right)$ as described in Sect. 2. Now let $|l|$ denote the length of l and consider $p \in \{0, 1\}^{|l|}$. The set of individuals $\{1, \ldots, Q\}$ is randomly divided into c subsets $Q_1, \ldots, Q_c$. Then for $j = 1, \ldots, c$ one can obtain $\left(b^j, l_p^j\right) = \operatorname{argmax} \sum_{q \notin Q_j} \mathcal{L}_q(b, l)$ s.t. $l_i = 0$ if and only if $p_i = 0$ for $i = 1, \ldots, |l|$. This maximization problem can be handled by quasi-newton procedures using SJ and yields estimations $\hat{b}^j$ and $\hat{l}_p^j$ for $j = 1, \ldots, c$. Finally, a c-fold cross-validation score can be obtained by

$$\mathcal{F}(p) = \frac{1}{c} \sum_{j=1}^{c} \sum_{q \in Q_j} \mathcal{L}_q\left(\hat{b}^j, \hat{l}_p^j\right).$$

Maximizing $\mathcal{F}(p)$ with respect to p yields a suitable binary representation $p^* = \operatorname{argmax} \mathcal{F}(p)$ of the Cholesky parameters to be selected before maximizing the log-likelihood $\mathcal{L}(b, l)$, where the i-th Cholesky parameter is selected if and only if $p_i^* = 1$.

In order to solve this non-linear binary maximization problem, an appropriate optimization method has to be chosen. Arora et al. (1994) and Alves and Climaco (2007) provide reviews of methods for discrete-integer-continuous variable optimization. However, the very nature of the optimization problem does neither allow for sequential linearization nor any relaxation approaches. It is apparently not possible to provide any gradient information. Hence, a stochastic search method is applied, namely a binary genetic algorithm (Holland 1992; Goldberg 1989).

This optimization procedure is a metaheuristic over a given parameter space. It is motivated by genetics and the natural process of selection and evolution. A binary genetic algorithm provides some advantages: it does not require convexity of the optimization problem and is, at least in theory, capable of finding a global optimum. On the other hand, it can be computationally quite burdensome, especially if the evaluation of the objective function is costly. Furtunately, due to the SJ approximation approach, the c-fold cross-validation score can be evaluated quite efficiently for reasonable values[4]

[4] The hyperparameter c does not require optimization and can be based on the available computational capacity. Larger values lead to a better adaptation, which is marginally decreasing. In the extreme case, the stochastic component can be eliminated completely by leave-one-out cross-validation.

of c. In this paper, $c = 5$ is chosen because this value fits the available computational capacity and corresponds to the widely used ratio of training and test data splitting in machine learning, which means that 80% of the data is used for training and 20% for testing. The reason for this split is based on the well-known Pareto principle. However, this is only a widely used rule of thumb. Research on the optimality of the data splitting ratio has not yet led to a consensus (Picard and Berk 1990; Dobbin and Simon 2011; Afendras and Markatou 2019).

Regarding genetic optimization, the objective function is usually denoted as fitness. In the present case, the parameter space is given by $\{0, 1\}^{|I|}$ and the fitness corresponds to the c-fold cross-validation score $\mathcal{F}(p)$. The entire optimization procedure is based on the bio-inspired operators: selection, crossover, and mutation.

At each step of the optimization procedure there is a current population consisting of m different binary-valued vectors $p^1, \ldots, p^m \in \{0, 1\}^{|I|}$. The selection of a member $p^{'} \in \left\{p^1, \ldots, p^m\right\}$ is made with a certain probability according to its fitness regarding the current population. The selection operator copies vectors from the current population to the next, with numerous different strategies for implementing this selection process, cf. Haupt and Haupt (2004) for a review.

In a second step, the crossover operator splits and merges two selected vectors of the current population to exchange characteristics. This entails randomly selecting a start and end position on a pair of mating vectors and exchanging the sub-vector of 0's and 1's between these positions on one vector with that from the mating vector.

Finally, the mutation operator completes the genetic refinement process. It corresponds to selecting a few members of the population, randomly determining positions on the selected vectors, and switching the 0 or 1 at these positions. This step safeguards the process from locking into a local optimum during selection and crossover.

The outlined steps are repeated for successive populations until no further improvement of the fitness is attainable. The member with the highest level of fitness in the last population is the optimum vector p^*, cf. Haupt and Haupt (2004) for a more detailed discussion.

4 Accuracy comparison between the proposed parameter selection approach and the original estimation method

Within the following subsections, the ability of the proposed parameter selection approach to recover parameters from finite samples in cross-sectional mixed multinomial probit models is evaluated. The results are compared to those of the original MACML estimation approach by Bhat (2011).

4.1 Experimental design

The underlying utility function of the multinomial mixed probit model outlined in Sect. 2 is given by $u_{qi} = \beta_q^{\mathsf{T}} x_{qi} + \epsilon_{qi}$. According to the simulation studies of Bhat and Sidharthan (2011) and Patil et al. (2017), two 5-dimensional cross-sectional designs are taken into account. Thus a set of 5 alternatives is considered together with 5 inde-

pendent exogeneous variables. For each choice occasion q and each alternative i the realizations of these exogeneous variables are given by x_{qi}. They are drawn from a standard univariate normal distribution. Likewise, the error terms ϵ_{qi} are drawn from a univariate normal distribution with variance of one-half to normalize the variances of the latent utility error differences. The random coefficient vector β_q follows a multivariate normal distribution with mean b and covariance matrix $\Omega = LL^\intercal$, where L denotes the lower Cholesky matrix of the corresponding decomposition. The covariance parameters are not estimated directly, but using the corresponding lower Cholesky matrix, cf. Section 2.

The covariance matrix design Ω_{Bhat} is considered first, since it has already been well-discussed in the literature and therefore serves as a benchmark case (Bhat and Sidharthan 2011; Patil et al. 2017). In this example, 15 Cholesky parameters have to be estimated, of which 4 have a true value of zero, corresponding to a proportion of more than 25%:

$$\Omega_{\text{Bhat}} = \begin{pmatrix} 1.00 & -0.50 & 0.25 & 0.75 & 0.00 \\ -0.50 & 1.00 & 0.25 & -0.50 & 0.00 \\ 0.25 & 0.25 & 1.00 & 0.33 & 0.00 \\ 0.75 & -0.50 & 0.33 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{pmatrix}.$$

The corresponding lower Cholesky matrix L_{Bhat} has a similar structure due to the arrangement of the variables in Ω_{Bhat}. The block-diagonal structure is preserved by the Cholesky decomposition:

$$L_{\text{Bhat}} = \begin{pmatrix} 1.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.50 & 0.87 & 0.00 & 0.00 & 0.00 \\ 0.25 & 0.43 & 0.87 & 0.00 & 0.00 \\ 0.75 & -0.14 & 0.24 & 0.60 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{pmatrix}.$$

Block-diagonal covariance matrix designs are natural restrictions of the general unstructured covariance matrix design in mixed multinomial probit modeling (Baragatti 2011). In some cases of seemingly unstructered covariance matrices, reordering of the underlying variables can be considered to impose a block-diagonal structure. In addition to Ω_{Bhat}, another more taylor-made block-diagonal covariance matrix design is considered to test the parameter selection procedure:

$$\Omega_{\text{Block}} = \begin{pmatrix} 1.00 & 075 & 0.00 & 0.00 & 0.00 \\ 0.75 & 1.00 & -0.50 & 0.00 & 0.00 \\ 0.00 & -0.50 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 & 0.50 \\ 0.00 & 0.00 & 0.00 & 0.50 & 1.00 \end{pmatrix}.$$

In this case, also 15 Cholesky parameters have to be estimated, of which now 7 have a true value of zero, corresponding to a proportion of almost 50%. The corresponding

lower Cholesky matrix is given by:

$$L_{\text{Block}} = \begin{pmatrix} 1.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.75 & 0.66 & 0.00 & 0.00 & 0.00 \\ 0.00 & -0.76 & 0.65 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.50 & 0.87 \end{pmatrix}.$$

For every covariance matrix design considered the corresponding random coefficient mean values are the same and chosen to be $b = \begin{pmatrix} 1.5 & -1.0 & 2.0 & 1.0 & -2.0 \end{pmatrix}^{\mathsf{T}}$ due to Bhat and Sidharthan (2011) and Patil et al. (2017).

In each case, 20 random data sets of size 2500 are generated, corresponding to 2500 individual choice occasions.

For the considered experimental designs MACML benchmark estimates are provided using SJ. In addition, the analytic approximation based estimator using the SJ approximation method together with the genetic algorithm parameter selection procedure outlined in Sect. 3 is applied. Here the cross-validation parameter is set to $c = 5$, which corresponds to a widely used ratio of training and test data splitting in machine learning as discussed in Sect. 3.

Bhat and Sidharthan (2011), Patil et al. (2017) and Batram and Bauer (2019) provide relative estimation errors by computing mean parameter estimates across data sets and reporting an absolute percentage bias for each parameter, i.e. $|(\text{mean estimate} - \text{true value})/\text{true value}| \cdot 100$. In doing so, they weight the deviation of the mean estimated value from the true value by its size which, from a theoretical point of view, facilitates comparisons between different parameter values. However, from a more technical perspective, it is not immediately clear why normalization should be required, since larger parameter values are not necessarily accompanied by increased biases. The results of Bhat and Sidharthan (2011) and Batram and Bauer (2019) show that this is not the case and that they are not even accompanied by increased standard errors. Moreover, it is not possible to compute an absolute percentage bias for true zero parameters, which prevents the opportunity for appropriate comparisons in these cases.

In contrast, this study does not only provide average estimation results, but also focuses on the empirical cumulative distributions of the absolute errors, i.e. $|\text{estimate} - \text{true value}|$, aggregated in pairs according the two estimation methods considered. Aggregation is performed over parameters and data sets for each covariance matrix design, where zero and non-zero parameters are considered separately to examine the effects of the parameter selection procedure. Wilcoxon–Mann–Whitney paired tests with large sample normal approximation are used to test for differences in the results (Wilcoxon 1945; Mann and Whitney 1947).

Furthermore, mean absolute estimation errors (MAE) are reported for each parameter, i.e. $|\text{estimate} - \text{true value}|$ averaged across data sets. These errors are easily traceable and yield properly comparable results also for zero parameters. The result is likely to be a higher reported MAE for each parameter compared to an absolute bias due to Bhat and Sidharthan (2011) and Patil et al. (2017), since the triangle inequality applies. In addition, MAE standard deviations are reported. Sensitivity and specificity

Table 1 Estimation results for L_{Bhat} for SJ without and with parameter selection (cross-validation parameter $c = 5$), including mean absolute errors (MAE) and standard deviations based on 20 data sets and 2500 observations per data set

Parameter	True value	SJ without parameter selection			SJ with parameter selection		
		Mean Estimate	MAE	MAE SD	Mean Estimate	MAE	MAE SD
b_1	1.500	1.546	0.081	(0.084)	1.544	0.093	(0.094)
b_2	− 1.000	− 1.071	0.097	(0.054)	− 1.055	0.061	(0.049)
b_3	2.000	2.164	0.194	(0.139)	2.149	0.189	(0.121)
b_4	1.000	1.012	0.057	(0.047)	1.006	0.064	(0.053)
b_5	− 2.000	− 2.151	0.173	(0.136)	− 2.134	0.162	(0.131)
l_{11}	1.000	0.924	0.119	(0.060)	0.917	0.121	(0.075)
l_{21}	− 0.500	− 0.499	0.114	(0.101)	− 0.504	0.094	(0.067)
l_{22}	0.866	0.863	0.138	(0.128)	0.856	0.126	(0.095)
l_{31}	0.250	0.215	0.109	(0.071)	0.192	0.143	(0.087)
l_{32}	0.433	0.332	0.138	(0.070)	0.331	0.133	(0.080)
l_{33}	0.866	0.928	0.112	(0.068)	0.938	0.135	(0.094)
l_{41}	0.750	0.674	0.116	(0.088)	0.683	0.115	(0.072)
l_{42}	− 0.144	− 0.361	0.217	(0.106)	− 0.234	0.190	(0.090)
l_{43}	0.237	0.250	0.101	(0.077)	0.221	0.140	(0.091)
l_{44}	0.601	0.752	0.199	(0.108)	0.776	0.199	(0.119)
l_{51}	0.000	0.105	0.142	(0.100)	0.072	0.073	(0.139)
l_{52}	0.000	0.018	0.106	(0.066)	0.028	0.029	(0.101)
l_{53}	0.000	− 0.020	0.097	(0.068)	− 0.002	0.019	(0.043)
l_{54}	0.000	0.040	0.109	(0.075)	0.007	0.024	(0.062)
l_{55}	1.000	0.956	0.107	(0.080)	0.952	0.114	(0.078)
Mean across Parameters		–	0.135	(0.100)	–	0.126	(0.103)
Mean time		0.117	–	–	40.393	–	–
(SD)		(0.039)			(10.489)		
Sensitivity		–	–	–	0.825	–	–
Specificity					0.858		

values are provided to give information about the proportion of correctly neglected and selected parameters, respectively.[5]

4.2 Results

Table 1 provides the estimation results for the covariance matrix design due to Bhat and Sidharthan (2011) and Patil et al. (2017).

[5] All computations are performed on a laptop computer with 6 CPU cores with a clock speed between 600 and 3220 MHz and 16 GB of RAM.

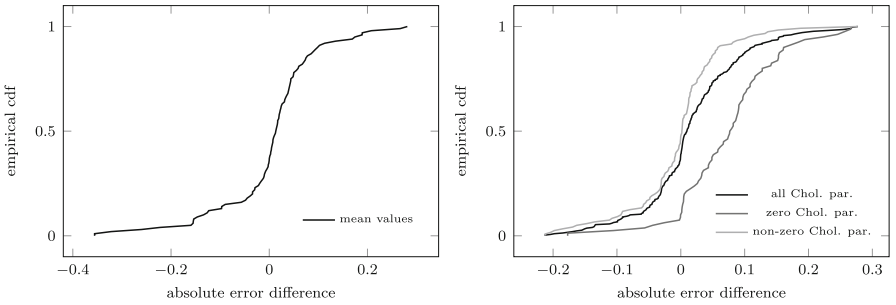

Fig. 1 Empirical cumulative distribution functions of the absolute error differences between SJ estimates without and with parameter selection for L_{Bhat}, estimated by trust region with cross-validation parameter $c = 5$ for 20 data sets and 2500 observations per data set split up according to mean values (l.) and Cholesky parameters (r.)

The difference in the accuracy of the estimation results in terms of absolute errors across all parameters is just over 7% in favor of the parameter selection procedure. However, a closer look at the true zero Cholesky entries shows that there are considerable differences. In the case of the SJ approximation without parameter selection, the mean absolute error here is 0.113, whereas in the case of the parameter selection procedure, the mean absolute error is only 0.036, which is less than one third. Small differences in favor of the parameter selection procedure are also apparent in the estimated mean values. Figure 1 shows the empirical cumulative distribution functions of the absolute error differences between SJ without and with parameter selection.

The differences in estimated mean values are significant (s = 3129, z = 2.075, p-value = 0.019). This also applies to the Cholesky parameters (s = 28520, z = 3.953, p − value = 0.000). However, this observation is solely due to the differences in the true zero Cholesky values (s = 3036, z = 6.789, p-value = 0.000). There is little difference in the true non-zero entries (s = 3635, z = 0.013, p-value = 0.990). The mean absolute errors are almost the same here.

The more accurate estimation results of the parameter selection procedure are accompanied by a significant increase in computation time, which is given in minutes and shows the main shortcoming of the approach. The values for sensitivity and specificity are well above 80% even with this comparatively small scaling of the experimental design.

Table 2 provides the estimation results for the covariance matrix Ω_{Block}. As before, the difference in the accuracy of the estimation results in terms of absolute errors across all parameters is in favor of the parameter selection approach. However, the difference is now over 20%. Again, the greatest differences can be seen in the true zero Cholesky parameters. In the case of the SJ approximation without parameter selection, the mean absolute error here is now 0.097, whereas in the case of the parameter selection procedure, the mean absolute error is now only 0.017, which is less than 20%. Figure 2 shows again the empirical cumulative distribution functions of the absolute error differences between SJ without and with parameter selection.

The significant differences in the estimated Cholesky values (s = 33709, z = 7.404, p-value = 0.000) are now again due to the fact that the differences in the true zero

Table 2 Estimation results for L_{Block} for SJ without and with parameter selection (cross-validation parameter $c = 5$), including mean absolute errors (MAE) and standard deviations based on 20 data sets and 2500 observations per data set

Parameter	True value	SJ without parameter selection			SJ with parameter selection		
		Mean Estimate	MAE	MAE SD	Mean Estimate	MAE	MAE SD
b_1	1.500	1.651	0.158	(0.145)	1.662	0.176	(0.152)
b_2	− 1.000	− 1.082	0.106	(0.106)	− 1.091	0.117	(0.105)
b_3	2.000	2.148	0.169	(0.148)	2.1688	0.206	(0.190)
b_4	1.000	1.136	0.138	(0.125)	1.142	0.145	(0.111)
b_5	− 2.000	− 2.177	0.194	(0.207)	− 2.204	0.221	(0.213)
l_{11}	1.000	0.960	0.148	(0.093)	0.966	0.148	(0.087)
l_{21}	0.750	0.813	0.127	(0.094)	0.843	0.122	(0.102)
l_{22}	0.661	0.833	0.185	(0.131)	0.821	0.179	(0.092)
l_{31}	0.000	0.076	0.098	(0.058)	0.000	0.000	(0.000)
l_{32}	− 0.756	− 0.522	0.243	(0.088)	− 0.530	0.229	(0.129)
l_{33}	0.655	0.837	0.183	(0.152)	0.841	0.199	(0.157)
l_{41}	0.000	0.046	0.076	(0.066)	0.000	0.000	(0.000)
l_{42}	0.000	− 0.100	0.124	(0.076)	− 0.042	0.042	(0.091)
l_{43}	0.000	0.025	0.109	(0.068)	0.009	0.029	(0.072)
l_{44}	1.000	1.062	0.129	(0.107)	1.081	0.136	(0.153)
l_{51}	0.000	− 0.023	0.074	(0.052)	− 0.002	0.016	(0.050)
l_{52}	0.000	0.022	0.114	(0.081)	− 0.007	0.022	(0.059)
l_{53}	0.000	0.014	0.087	(0.066)	0.014	0.014	(0.043)
l_{54}	0.500	0.482	0.131	(0.085)	0.500	0.131	(0.082)
l_{55}	0.866	0.946	0.123	(0.103)	0.958	0.126	(0.098)
Mean across Parameters		–	0.136	(0.103)	–	0.113	(0.099)
Mean time		0.147	–	–	39.220	–	–
(SD)		(0.031)			(17.654)		
Sensitivity		–	–	–	0.900	–	–
Specificity					1.000		

parameters are significant (s = 9681, z = 9.871, p-value = 0.000). This does not apply to the non-zero entries (s = 1049, z = 0.987, p-value = 0.324). The mean absolute errors are almost the same here, 0.167 and 0.161 for SJ without and with parameter selection, respectively. No significant differences can be observed in the mean values either (s = 2936, z = 1.413, p-value = 0.158).

The values for sensitivity and specificity are increased compared to before even reach one in case of specificity, which means that no true non-zero parameter has been incorrectly neglected.

Even though the proposed parameter selection approach is computationally intensive compared to the original estimation method, the results presented here show that it

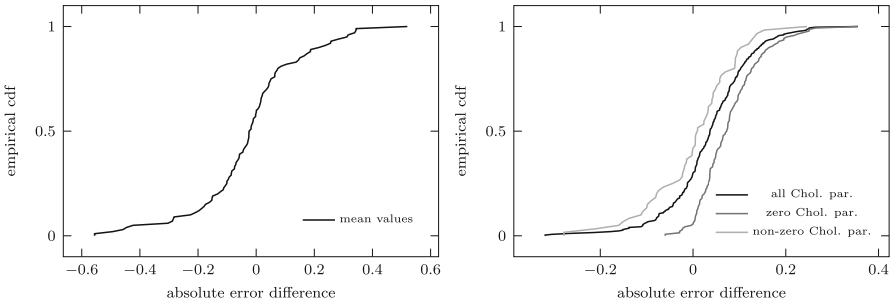

Fig. 2 Empirical cumulative distribution functions of the absolute error differences between SJ estimates without and with parameter selection for L_{Block}, estimated by trust region with cross-validation parameter $c = 5$ for 20 data sets and 2500 observations per data set split up according to mean values (l.) and Cholesky parameters (r.)

seems to work, at least with respect to the intended improved accuracy of the estimation results.

5 Conclusion

Applying analytic approximations for multivariate normal cumulative distribution functions to estimate mixed multinomial probit models has led to a significant improvement in both accuracy of the estimation results and computation time (Bhat and Sidharthan 2011; Patil et al. 2017). The proposed estimation approach of this paper can help to further reduce estimation errors in the estimation of the covariances of the random coefficients of the model.

However, the achieved improvements in the accuracy of the estimation results are accompanied by considerable increases of the computational load, which is mainly due to the underlying discrete optimization problem of the proposed parameter selection approach and the applied meta-heuristic optimization method. Therefore, the proposed procedure is unsuitable for fast repeated calculations, which are necessary in some application areas. Using a different parameter selection approach or another discrete optimization method could possibly reduce computation times and facilitate an extension of the parameter selection procedure to mixed multinomial probit models for panel data. In this regard, it would maybe be beneficial to investigate the impact of memetic algorithms on the computation time in future research.

The enclosed simulation results indicate that more accurate estimation results can already be obtained by the proposed parameter selection approach than with conventional estimation via an analytic approximation, if only a moderate degree of independence between the random coefficients is present. With greater independence, even more accurate estimation results are possible. The online appendix of this study contains an application to a conjoint analysis study of preferences between alternative vehicles, which shows that the proposed parameter selection approach can provide plausible and efficient estimation results. In particular, with respect to real data, a

question for future research may be whether the proposed method is able to reduce parameter identification fragility.

Acknowledgements The author acknowledges the helpful comments of two anonymous reviewers on earlier versions of the paper.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The author states that there is no conflict of interest.

References

Afendras G, Markatou M (2019) Optimality of training/test size and resampling effectiveness in cross-validation. J Stat Plan Inference 199:286–301

Alves MJ, Climaco J (2007) A review of interactive methods for multiobjective integer and mixed-integer programming. Eur J Oper Res 180(1):99–115

Arora JS, Huang MW, Hsieh CC (1994) Methods for optimization of nonlinear problems with discrete variables: a review. Struct Optim 8(2):69–85

Baragatti M (2011) Bayesian variable selection for probit mixed models applied to gene selection. Bayesian Anal 6(2):209–230

Batram M, Bauer D (2019) On consistency of the MACML approach to discrete choice modelling. J Choice Modell 30:1–16

Bhat CR (2011) The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. Transp Res Part B Methodol 45(7):923–939

Bhat CR (2018) New matrix-based methods for the analytic evaluation of the multivariate cumulative normal distribution function. Transp Res Part B Methodol 109:238–256

Bhat CR, Sidharthan R (2011) A simulation evaluation of the maximum approximate composite marginal likelihood (MACML) estimator for mixed multinomial probit models. Transp Res Part B Methodol 45(7):940–953

Dobbin KK, Simon RM (2011) Optimally splitting cases for training and testing high dimensional classifiers. BMC Med Genomics 4(1):1–8

Goldberg DE (1989) Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Boston

Haupt RL, Haupt SE (2004) Practical genetic algorithms. Wiley, Hoboken

Holland JH (1992) Genetic algorithms. Sci Am 267(1):66–73

Joe H (1995) Approximations to multivariate normal rectangle probabilities based on conditional expectations. J Am Stat Assoc 90(431):957–964

Luce RD, Suppes P et al (1965) Preference, utility, and subjective probability. Handb Math Probab 3

Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18(1):50–60

McFadden D (1973) Conditional logit analysis of qualitative choice behavior. Front Economet 105–142

McFadden D (1978) Modeling the choice of residential location. Transp Res Rec 672:72–77

Mendell NR, Elston RC (1974) Multifactorial qualitative traits: genetic analysis and prediction of recurrence risks. Biometrics 30(1):41–57
Patil PN, Dubey SK, Pinjari AR, Cherchi E, Daziano R, Bhat CR (2017) Simulation evaluation of emerging estimation techniques for multinomial probit models. J Choice Modell 23:9–20
Picard RR, Berk KN (1990) Data splitting. Am Stat 44(2):140–147
Savolainen PT, Mannering FL, Lord D, Quddus MA (2011) The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid Anal Prevent 43(5):1666–1676
Solow AR (1990) A method for approximating multivariate normal orthant probabilities. J Stat Comput Simul 37(3–4):225–229
Train KE (2009) Discrete choice methods with simulation. Cambridge University Press, Cambridge
Trinh G, Genz A (2015) Bivariate conditioning approximations for multivariate normal probabilities. Stat Comput 25(5):989–996
Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics 1(6):80–83