Transportation Research Part B 45 (2011) 923–939

Contents lists available at ScienceDirect

Transportation Research Part B

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

The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models

Chandra R. Bhat *

The University of Texas at Austin, Department of Civil, Architectural and Environmental Engineering, 1 University Station, C1761, Austin, TX 78712-0278, United States

ARTICLE INFO

Article history:
Received 3 August 2010
Received in revised form 18 April 2011
Accepted 19 April 2011

Keywords:
Multinomial probit
Mixed models
Composite marginal likelihood
Discrete choice models
Spatial econometrics
Panel data

ABSTRACT

The likelihood functions of multinomial probit (MNP)-based choice models entail the evaluation of analytically-intractable integrals. As a result, such models are usually estimated using maximum simulated likelihood (MSL) techniques. Unfortunately, for many practical situations, the computational cost to ensure good asymptotic MSL estimator properties can be prohibitive and practically infeasible as the number of dimensions of integration rises. In this paper, we introduce a maximum approximate composite marginal likelihood (MACML) estimation approach for MNP models that can be applied using simple optimization software for likelihood estimation. It also represents a conceptually and pedagogically simpler procedure relative to simulation techniques, and has the advantage of substantial computational time efficiency relative to the MSL approach. The paper provides a "blueprint" for the MACML estimation for a wide variety of MNP models.

1. Introduction

The "workhorse" multinomial logit model, used extensively in practice for econometric discrete choice analysis, was introduced by Luce and Suppes (1965) and McFadden (1974), and has a very simple and elegant structure. However, it is also saddled with the familiar independence from irrelevant alternatives (IIA) property – that is, the ratio of the choice probabilities of two alternatives is independent of the characteristics of other alternatives in the choice set. This has led to several extensions of the MNL model through the relaxation of the independent and identically distributed (IID) error distribution (across alternatives) assumption. Two common model forms of non-IID error distribution include the generalized extreme-value (GEV) class of models proposed by McFadden (1978) and the multinomial probit (MNP) model that allow relatively flexible error covariance structures (up to certain limits of identifiability; see Train, 2009, Chapter 5). Both of these non-IID kernel structures (or even the IID versions of the GEV and the MNP models, which lead to the MNL and the independent MNP models) can further be combined with (a) continuous mixing factor error structures to accommodate unobserved taste variation across choice occasions through random coefficients (see Bhat and Sardesai, 2006), (b) individual-specific (time-stable or time-dissipating) error terms that generate error dependencies across the choice occasions of the same decision-maker in panel or repeated choice contexts (see Li et al., 2010), (c) spatial/social dependency-based error structure models that generate error dependencies across choice occasions of different decision-makers (see Franzese and Hays, 2008), or (d) some combinations of these model forms. While many different continuous error distributions can be used to accommodate these additional structures, it is most common to adopt a normal distribution for these. For instance, when introducing random coefficients, it is typical to use the multivariate normal distribution for the mixing coefficients, almost to the point that

* Tel.: +1 512 471 4535; fax: +1 512 475 8744.
E-mail address: bhat@mail.utexas.edu

0191-2615/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.trb.2011.04.005

the terms mixed logit or mixed GEV or mixed probit are oftentimes used synonymously with normal mixing (see Fiebig et al., 2010; Dube et al., 2002).[1] Similarly, when introducing panel effects or spatial/social interdependency, the use of normal error structures is ubiquitous, except for some recent developments using general copula forms by Bhat and Sener (2009) and Bhat et al. (2010a).

In the context of the normal error distributions just discussed, the use of a GEV kernel structure leads to a mixing of the normal distribution with a GEV kernel, while the use of an MNP kernel leads once again to an MNP model. Both structures have been widely used in the past, with the choice between a GEV kernel or an MNP kernel really being a matter of "which is easier to use in a given situation" (Ruud, 2007). In recent years, the mixing of the normal with the GEV kernel has been the model form of choice in the economics and transportation fields, mainly due to the relative ease with which the probability expressions in this structure can be simulated (see Bhat et al. (2008) and Train (2009) for detailed discussions). On the other hand, the use of an MNP kernel has not seen as much use in recent years, because the simulation estimation is generally more difficult. In any case, while there have been several approaches proposed to simulate these models with a GEV or an MNP kernel, most of these involve pseudo-Monte Carlo or quasi-Monte Carlo simulations in combination with a quasi-Newton optimization routine in a maximum simulated likelihood (MSL) inference approach (see Bhat, 2001, 2003). In such an inference approach, consistency, efficiency, and asymptotic normality of the estimator is critically predicated on the condition that the number of simulation draws rises faster than the square root of the number of individuals in the estimation sample. This effectively implies that the desirable asymptotic properties of the MSL estimator are obtained at the expense of computational cost. Unfortunately, for many practical situations, the computational cost to ensure good asymptotic estimator properties can be prohibitive and literally infeasible (in the context of the computation resources available and the time available for estimation) as the number of dimensions of integration increases. This is particularly so because the accuracy of simulation techniques is known to degrade rapidly at medium-to-high dimensions, and the simulation noise increases substantially. This leads to convergence problems during estimation, unless a very high number of simulation draws is used. Besides, an issue generally ignored in simulation-based approaches is the accuracy (or lack thereof) of the covariance matrix of the estimator, which is critical for good inference even if the asymptotic properties of the estimator are well established. Specifically, the hessian (or second derivatives) needed with the MSL approach to estimate the asymptotic covariance matrix of the estimator is itself estimated on a highly nonlinear and non-smooth second derivatives surface of the log-simulated likelihood function. This is also usually undertaken numerically because the complex analytic nature of the second derivatives makes them difficult to code. The net result is that Monte Carlo simulation with even three to four decimal places of accuracy in the probabilities embedded in the log-likelihood function can work poorly (see Bhat et al., 2010b), suggesting a critical need to evaluate the likelihood function at a very high level of accuracy and precision. This further increases computational cost. Craig (2008) also alludes to this problem when he states that "(...) the randomness that is inherent in such methods [referring to the Genz-Bretz algorithm (Genz and Bretz, 1999), but applicable in general to MSL methods] is sometimes more than a minor nuisance."

In this paper, we propose a new methodology (which is labeled as the maximum approximate composite marginal likelihood or MACML inference approach) that allows the estimation of models with both GEV and MNP kernels using simple, computationally very efficient, and simulation-free estimation methods. In the MACML inference approach, models with the MNP kernel, when combined with additional normal random components, are much easier to estimate because of the conjugate addition property of the normal distribution (which puts the structure resulting from the addition of normal components to the MNP kernel back into an MNP form). On the other hand, the MACML estimation of models obtained by superimposing normal error components over a GEV kernel requires a normal scale mixture representation for the extreme value error terms, and adds an additional layer of computational effort (see Bhat, 2011). Given that the use of a GEV kernel or an MNP kernel is simply a matter of convenience, we will henceforth focus in this paper on the MNP kernel within the MACML inference approach.

The proposed MACML estimation of the resulting MNP-based models involves only univariate and bivariate cumulative normal distribution function evaluations, regardless of the number of alternatives or the number of choice occasions per individual or the nature of social/spatial dependence structures. This implies substantial computational efficiency relative to the MSL inference approach for models that can be estimated using the MSL approach, and also allows the estimation of several model structures that are literally infeasible to estimate using traditional MSL approaches (with the available computational resources and time available for estimation). For instance, consider the case of dealing with panel data or repeated unordered choice data from each individual in a sample, with individual-specific normally distributed random coefficients. In such a case, the full likelihood function contribution of an individual entails taking the product of the individual's choices across choice occasions conditional on the random coefficients, and then unconditioning out the random coefficients by integration over the multivariate normal domain. The multivariate integration is over the real line with respect to each random coefficient, with the dimensionality being equal to the number of random coefficients. As the number of random coefficients increases, evaluating the likelihood using simulation techniques becomes computationally expensive. Another example where the full likelihood becomes near impractical to work with is when there are individual-specific normal random

[1] To be sure, there have been models with non-normal mixing distributions too, such as the log-normal distribution, the triangular distribution, and the Rayleigh distribution (see, Bhat et al., 2008 for a review). However, it has been well known that using non-normal distributions can lead to convergence/computational problems, and it is not uncommon to see researchers consider non-normal distributions only to eventually revert to the use of a normal distribution (see, for example, Bartels et al., 2006 and Small et al., 2005).

coefficients as well as choice occasion-specific normal random coefficients with panel or repeated unordered choice data. In this case, the result is a double multivariate integral, with the dimensionality of the two multivariate normal integrals being equal to the number of random coefficients in the individual-specific and occasion-specific cases (see Bhat and Castelar, 2002; Bhat and Sardesai, 2006; Hess and Rose, 2009). The explosion of the dimensionality of integration is rapid, making full likelihood evaluation using simulation techniques all but impractical. Finally, in the case of global social interactions or spatial interactions that lead to autoregressive error structures or spatial/social lag effects, the full likelihood is infeasible to estimate using simulation methods in any reasonable time because of the extremely high dimensionality involved (the dimensionality is of the order of the number of decision-makers times the number of alternatives in the multinomial choice situation minus one; for example, with 2000 decision-makers and four alternatives, the dimensionality of integration is 6000). In all these cases and more, the proposed MACML approach offers a computationally convenient inference approach, as we indicate in the rest of this paper. As importantly, the MACML inference approach is simple to code and apply using readily available software for likelihood estimation. It also represents a conceptually and pedagogically simpler procedure relative to simulation techniques.

The paper is structured as follows. The next section presents the two main and fundamental building blocks of the MACML approach. Section 3 presents the MACML inference approach for the cross-sectional MNP model, while Section 4 illustrates the approach for panel and dynamic MNP model structures. Section 5 presents the extension to accommodate spatial/social effects. Section 6 discusses model selection issues in the CML estimation approach. Finally, Section 7 summarizes the contributions of the paper.

2. The basics of the MACML approach to estimate unordered-response models

There are two fundamental concepts in the proposed MACML approach to estimate MNP models. The first is an approximation method to evaluate the multivariate standard normal cumulative distribution (MVNCD) function (discussed in Section 2.1). The second is the composite marginal likelihood (CML) approach to estimation (discussed in Section 2.2).

2.1. Multivariate standard normal cumulative distribution (MVNCD) function

In the general case of an MNP model with I alternatives, the probability expression of an individual choosing a particular alternative involves an $(I-1)$ dimensional MVNCD function (more on this in Section 3). The evaluation of such a function cannot be pursued using quadrature techniques due to the curse of dimensionality when the dimension of integration exceeds two (see Bhat, 2003). Consequently, the probability expression is approximated using simulation techniques in the classical maximum simulated likelihood (MSL) inference approach, usually through the use of the Geweke–Hajivassiliou–Keane (GHK) simulator or the Genz–Bretz (GB) simulator, which are among the most effective simulators for evaluating multivariate normal probabilities (see Bhat et al. (2010b) for a detailed description of these simulators). Some other recent sparse grid-based techniques for simulating the multivariate normal probabilities have also been proposed by Heiss and Winschel (2008), Huguenin et al. (2009), and Heiss (2010). In addition, Bayesian simulation using Markov Chain Monte Carlo (MCMC) techniques (instead of MSL techniques) have been used in the literature (see Albert and Chib, 1993; McCulloch and Rossi, 2000; Train, 2009). However, all these MSL and Bayesian techniques require extensive simulation, are time-consuming, are not very straightforward to implement, and create convergence assessment problems as the number of dimensions of integration increases.

In this paper, we apply an *analytic approximation* method to evaluate the MVNCD function that is quite accurate and very fast even for 20 or more dimensions of integration. Further, unlike Monte Carlo simulation approaches, even two to three decimal places of accuracy in the analytic approximation is generally adequate to accurately and precisely recover the parameters and their covariance matrix estimates because of the smooth nature of the first and second derivatives of the approximated analytic log-likelihood function. While several analytic approximations have been reported in the literature for MVNCD functions (see, for example, Solow, 1990; Joe, 1995, 2008; Gassmann et al., 2002), the one we use here is based on decomposition into a product of conditional probabilities. This approximation appears to have been first proposed by Solow (1990) based on Switzer (1977), and then refined by Joe (1995). However, we are not aware of any earlier research effort that applies this technique for the estimation of parameters in econometric models (such as discrete choice models) involving the evaluation of MVNCD functions. The reason we select this approximation approach is that it is fast and lends itself nicely to combination with the composite marginal likelihood approach of MNP model estimation that we propose in this paper.

To describe the approximation, let $(W_1, W_2, W_3, \ldots, W_I)$ be a multivariate normally distributed random vector with zero means, variances of 1, and a correlation matrix $\boldsymbol{\Sigma}$. Then, interest centers on approximating the following orthant probability:

$$\Pr(\boldsymbol{W} < \boldsymbol{w}) = \Pr(W_1 < w_1, W_2 < w_2, W_3 < w_3, \ldots, W_I < w_I). \tag{1}$$

The above joint probability may be written as the product of a bivariate marginal probability and univariate conditional probabilities as follows ($I \geqslant 3$):

$$\Pr(\boldsymbol{W} < \boldsymbol{w}) = \Pr(W_1 < w_1, W_2 < w_2) \times \prod_{i=3}^{I} \Pr(W_i < w_i | W_1 < w_1, W_2 < w_2, W_3 < w_3, \ldots, W_{i-1} < w_{i-1}). \tag{2}$$

Next, define the binary indicator $\tilde{I}_i$ that takes the value 1 if $W_i < w_i$ and zero otherwise. Then $E(\tilde{I}_i) = \Phi(w_i)$, where $\Phi(.)$ is the univariate normal standard cumulative distribution function. Also, we may write the following:

$$\begin{aligned} &\mathrm{Cov}(\tilde{I}_i, \tilde{I}_j) = E(\tilde{I}_i \tilde{I}_j) - E(\tilde{I}_i)E(\tilde{I}_j) = \Phi_2(w_i, w_j, \rho_{ij}) - \Phi(w_i)\Phi(w_j), i \neq j \\ &Cov(\tilde{I}_i, \tilde{I}_i) = Var(\tilde{I}_i) = \Phi(w_i) - \Phi^2(w_i) = \Phi(w_i)[1 - \Phi(w_i)], \end{aligned} \tag{3}$$

where ρ_{ij} is the ijth element of the correlation matrix $\boldsymbol{\Sigma}$. With the above preliminaries, consider the following conditional probability:

$$\Pr(W_i < w_i | W_1 < w_1, W_2 < w_2, W_3 < w_3, \ldots, W_{i-1} < w_{i-1}) = E(\tilde{I}_i | \tilde{I}_1 = 1, \tilde{I}_2 = 1, \tilde{I}_3 = 1, \ldots, \tilde{I}_{i-1} = 1). \tag{4}$$

The right side of the expression may be approximated by a linear regression model, with $\tilde{I}_i$ being the "dependent" random variable and $\tilde{\boldsymbol{I}}_{<i} = (\tilde{I}_1, \tilde{I}_2, \ldots \tilde{I}_{i-1})$ being the independent random variable vector.[2] In deviation form, the linear regression for approximating Eq. (4) may be written as:

$$\tilde{I}_i - E(\tilde{I}_i) = \boldsymbol{\alpha}'[\tilde{\boldsymbol{I}}_{<i} - E(\tilde{\boldsymbol{I}}_{<i})] + \tilde{\eta}, \tag{5}$$

where $\boldsymbol{\alpha}$ is the least squares coefficient vector and $\tilde{\eta}$ is a mean zero random term. In this form, the usual least squares estimate of $\boldsymbol{\alpha}$ is given by:

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{\Omega}_{<\boldsymbol{i}}^{-\boldsymbol{1}} \cdot \boldsymbol{\Omega}_{\boldsymbol{i},<\boldsymbol{i}},$$

where

$$\boldsymbol{\Omega}_{<\boldsymbol{i}} = \mathrm{Cov}(\boldsymbol{I}_{<\boldsymbol{i}}, \boldsymbol{I}_{<\boldsymbol{i}}) = \begin{bmatrix} \mathrm{Cov}(\tilde{I}_1, \tilde{I}_1) & \mathrm{Cov}(\tilde{I}_1, \tilde{I}_2) & \mathrm{Cov}(\tilde{I}_1, \tilde{I}_3) & \cdots & \mathrm{Cov}(\tilde{I}_1, \tilde{I}_{i-1}) \\ \mathrm{Cov}(\tilde{I}_2, \tilde{I}_1) & \mathrm{Cov}(\tilde{I}_2, \tilde{I}_2) & \mathrm{Cov}(\tilde{I}_2, \tilde{I}_3) & \cdots & \mathrm{Cov}(\tilde{I}_2, \tilde{I}_{i-1}) \\ \mathrm{Cov}(\tilde{I}_3, \tilde{I}_1) & \mathrm{Cov}(I_3, I_2) & \mathrm{Cov}(\tilde{I}_3, \tilde{I}_3) & \cdots & \mathrm{Cov}(\tilde{I}_3, \tilde{I}_{i-1}) \\ \vdots & & & & \\ \mathrm{Cov}(\tilde{I}_{i-1}, \tilde{I}_1) & \mathrm{Cov}(\tilde{I}_{i-1}, \tilde{I}_2) & \mathrm{Cov}(\tilde{I}_{i-1}, \tilde{I}_3) & \cdots & \mathrm{Cov}(\tilde{I}_{i-1}, \tilde{I}_{i-1}) \end{bmatrix}, \tag{6}$$

and

$$\boldsymbol{\Omega}_{\boldsymbol{i},<\boldsymbol{i}} = \mathrm{Cov}(\boldsymbol{I}_{<\boldsymbol{i}}, \boldsymbol{I}_{\boldsymbol{i}}) = \begin{bmatrix} \mathrm{Cov}(\tilde{I}_i, \tilde{I}_1) \\ \mathrm{Cov}(\tilde{I}_i, \tilde{I}_2) \\ \mathrm{Cov}(\tilde{I}_i, \tilde{I}_3) \\ \vdots \\ \mathrm{Cov}(\tilde{I}_i, \tilde{I}_{i-1}) \end{bmatrix}. \tag{7}$$

Finally, putting the estimate of $\hat{\boldsymbol{\alpha}}$ back in Eq. (5), and predicting the expected value of $\tilde{I}_i$ conditional on $\tilde{\boldsymbol{I}}_{<i} = \boldsymbol{1}$ (*i.e.*, $\tilde{I}_1 = 1, \tilde{I}_2 = 1, \tilde{I}_{i-1} = 1$), we get the following approximation for Eq. (4):

$$\Pr(W_i < w_i | W_1 < w_1, W_2 < w_2, \ldots, W_{i-1} < w_{i-1}) \approx \Phi(w_i) + (\boldsymbol{\Omega}_{<\boldsymbol{i}}^{-1} \cdot \boldsymbol{\Omega}_{\boldsymbol{i},<\boldsymbol{i}})'(1 - \Phi(w_1), 1 - \Phi(w_2) \ldots 1 - \Phi(w_{i-1}))' \tag{8}$$

This conditional probability approximation can be plugged into Eq. (2) to approximate the multivariate orthant probability in Eq. (1). The resulting expression for the multivariate orthant probability comprises only univariate and bivariate standard normal cumulative distribution functions.

One remaining issue is that the decomposition of Eq. (1) into conditional probabilities in Eq. (2) is not unique. Further, different permutations (*i.e.*, orderings of the elements of the random vector $\boldsymbol{W} = (W_1, W_2, W_3, \ldots, W_I)$) for the decomposition into the conditional probability expression of Eq. (2) will lead, in general, to different approximations. One approach to resolve this is to average across the $I!/2$ permutation approximations. However, as indicated by Joe (1995), the average over a few randomly selected permutations is typically adequate for the accurate computation of the multivariate orthant probability. In the case when the approximation is used for model estimation (where the integrand in each individual's

[2] Note that, theoretically, this approximation can be viewed as a first-order approximation. The approximation can be continually improved by increasing the order of the approximation. For instance, a second-order approximation would approximate the right side of Eq. (4) by the expectation from a linear regression model that has $\tilde{I}_i$ as the "dependent" random variable and $\tilde{\boldsymbol{I}}_{<i} = (\tilde{I}_1, \tilde{I}_2, \ldots \tilde{I}_{i-1}, \tilde{I}_{12}, \tilde{I}_{13}, \ldots I_{1,i-1}, \tilde{I}_{23}, \tilde{I}_{24}, \ldots \tilde{I}_{2,i-1}, \ldots I_{i-2,i-1})$ as the independent random variable vector, where $\tilde{I}_{i'j'} = \tilde{I}_{i'}\tilde{I}_{j'}$. Essentially this adds second-order interactions in the independent random variable vector (see Joe, 1995). However, doing so entails trivariate and four-variate normal cumulative distribution function (CDF) evaluations (when $I > 4$) as opposed to univariate and bivariate normal CDF evaluations in the first-order approximation, thus increasing computational burden. As we discuss later and show empirically in a companion paper (Bhat and Sidharthan, forthcoming), the first-order approximation is more than adequate (when combined with the CML approach) for estimation of any MNP model. Thus, in the rest of this paper, we will use the term approximation to refer to the first-order approximation evaluation of the MVNCD function.

log-likelihood contribution is a parameterized function of the $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ parameters), even a single permutation of the $\boldsymbol{W}$ vector per choice occasion should typically suffice, as our own experimentations have shown.

2.2. The composite marginal likelihood (CML) estimator

The composite marginal likelihood (CML) estimation approach is a relatively simple approach that can be used when the full likelihood function is practically infeasible to evaluate due to underlying complex dependencies. Unfortunately, in many such cases, the approximation discussed in the previous section for orthant probabilities, by itself, does not work with the full likelihood approach because the dimensionality can be far beyond the accuracy of the MVNCD analytic approximation.

In this paper, we propose the use of the CML approach of estimation, combined with the approximation for orthant probabilities discussed in the previous section. To our knowledge, no earlier study in the literature has considered the CML method in the context of unordered-response models (rather, all earlier studies have used the CML approach for multivariate models such as the multivariate binary probit or the multivariate ordered probit (see, for example, Renard et al., 2004; Zhao and Joe, 2005; Varin and Vidoni, 2006; Feddag and Bacci, 2009; Bhat and Sener, 2009; Varin and Czado, 2010; Bhat et al., 2010a,b). This is because the CML method by itself is not well suited to unordered-response models and does not provide substantial computational benefits in unordered-response models; rather, it is our specific proposal in this paper to combine the CML method with the normal orthant probability approximation method of the previous section that is the key to computational benefits.

The CML approach, which belongs to the more general class of composite likelihood function approaches (see Lindsay, 1988), is based on maximizing a surrogate likelihood function that compounds much easier-to-compute, lower-dimensional, marginal likelihoods (see Varin (2008) and Varin et al. (forthcoming) for recent reviews of the CML method). The CML approach works as follows. Assume that the data originate from a parametric underlying model based on a $D \times 1$ vector random variable $\boldsymbol{Y}$ with density function $f(\boldsymbol{y}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is an unknown $\tilde{K}$-dimensional parameter vector. Suppose that $f(\boldsymbol{y}, \boldsymbol{\theta})$ is difficult or near infeasible to evaluate in reasonable time with the computational resources at hand, so that the corresponding likelihood function from a sampled (observed) vector for $\boldsymbol{Y}$ (say $\boldsymbol{m} = (m_1, m_2, m_3, \ldots m_D)'$) given by $L(\boldsymbol{\theta}; \boldsymbol{m}) = f(\boldsymbol{m}, \boldsymbol{\theta})$ is difficult. However, suppose evaluating the likelihood functions of a set of $\tilde{E}$ observed marginal events (each observed marginal event being a subset of the observed joint event $\boldsymbol{m}$) is easy and/or computationally expedient. Let these observed marginal events be characterized by $(A_1(\boldsymbol{m}), A_2(\boldsymbol{m}), \ldots, A_{\tilde{E}}(\boldsymbol{m}))$. For instance, $A_1(\boldsymbol{m})$ may represent the marginal event that the observed values in the sample for the first two elements of the vector $\boldsymbol{Y}$ are $(\boldsymbol{m}_1, \boldsymbol{m}_2)'$, $A_2(\boldsymbol{m})$ may represent the marginal event that the observed values for the first and third elements of the vector $\boldsymbol{Y}$ are $(m_1, m_3)'$, and so on. Let each event $A_e(\boldsymbol{m})$ be associated with a likelihood object $L_e(\boldsymbol{\theta}; \boldsymbol{m}) = L[\boldsymbol{\theta}; A_e(\boldsymbol{m})]$, which is based on a lower-dimensional marginal joint density function corresponding to the original high-dimensional joint density of $\boldsymbol{Y}$. Then, the general form of the composite marginal likelihood function is as follows:

$$L_{CML}(\boldsymbol{\theta}, \boldsymbol{m}) = \prod_{e=1}^{\tilde{E}} [L_e(\boldsymbol{\theta}; \boldsymbol{m})]^{\omega_e} = \prod_{e=1}^{\tilde{E}} [L(\boldsymbol{\theta}; A_e(\boldsymbol{m})]^{\omega_e}, \tag{9}$$

where ω_e is a power weight to be chosen based on efficiency considerations. If these power weights are the same across events, they may be dropped. The CML estimator is the one that maximizes the above function (or equivalently, its logarithmic transformation).

The CML class of estimators subsumes the usual ordinary full-information likelihood estimator as a special case. For instance, consider the case of repeated unordered choices from a specific individual. Let the individual's choice at time t be denoted by the index C_t, and let this individual be observed to choose alternative m_t at choice occasion t ($t = 1, 2, 3, \ldots, T$). Then, one may define the observed event for this individual as the sequence of observed choices across all the T choice occasions of the individual. Defined this way, the CML function contribution of this individual becomes equivalent to the full-information maximum likelihood function contribution of the individual[3]:

$$L^1_{CML}(\boldsymbol{\theta}, \boldsymbol{m}) = L(\boldsymbol{\theta}, \boldsymbol{m}) = \text{Prob}(C_1 = m_1, C_2 = m_2, C_3 = m_3, \ldots, C_T = m_T). \tag{10}$$

However, one may also define the events as the observed choices at each choice occasion for the individual. Defined this way, the CML function is:

$$L^2_{CML}(\boldsymbol{\theta}, \boldsymbol{m}) = \text{Prob}(C_1 = m_1) \times \text{Prob}(C_2 = m_2) \times \text{Prob}(C_3 = m_3) \times \ldots \times \text{Prob}(C_T = m_t) \tag{11}$$

This CML, of course, corresponds to the independence case between each pair of observations from the same individual. As we will indicate later, the above CML estimator is consistent. However, this approach, in general, does not estimate the parameters representing error correlation effects across choices of the same individual (*i.e.*, only a subset of the vector $\boldsymbol{\theta}$ is estimable). A third approach to estimating the parameter vector $\boldsymbol{\theta}$ in the repeated unordered choice case is to define the events in the CML as the pairwise observations across all or a subset of the choice occasions of the individual. For

[3] In the discussion below, for presentation ease, we will ignore the power weight term ω_e (we revisit this later in Section 4.2, when discussing the case of potentially different numbers of choice occasions across individuals).

presentation ease, assume that all pairs of observations are considered. This leads to a pairwise CML function contribution of individual q as follows:

$$L^3_{CML}(\boldsymbol{\theta},\boldsymbol{y}) = \prod_{t=1}^{T-1}\prod_{w=t+1}^{T} \text{Prob}(C_t = m_t, C_w = m_w) \tag{12}$$

Almost all earlier research efforts employing the CML technique have used the pairwise approach, including Apanasovich et al. (2008), Varin and Vidoni (2009), Engle et al. (2007), Bhat et al. (2010a), and Bhat and Sener (2009). Alternatively, the analyst can also consider larger subsets of observations, such as triplets or quadruplets or even higher dimensional subsets (see Engler et al., 2006; Caragea and Smith, 2007). However, it is generally agreed that the pairwise approach is a good balance between statistical and computational efficiency (besides, in almost all applications, the parameters characterizing error dependency are completely identified based on the pairwise approach). Importantly, the pairwise approach is able to explicitly recognize dependencies across choice occasions in the repeated choice case through the inter-temporal pairwise probabilities. More generally, and as we shall see later, the pairwise approach is adequate to estimate parameters of several MNP-based model structures. The pairwise CML function involves normal orthant probabilities of dimension $(I-1)\times 2$, which can itself be computationally impractical as I (the number of alternatives at each choice occasion in the repeated choice case) becomes large. However, this is where our proposal of combining the pairwise CML with the orthant probability approximation of the previous section comes in.

The properties of the general CML estimator may be derived using the theory of estimating equations (see Cox and Reid, 2004). Under usual regularity conditions (these are the usual conditions needed for likelihood objects to ensure that the logarithm of the CML function can be maximized by solving the corresponding score equations; the conditions are too numerous to mention here, but are listed in Molenberghs and Verbeke, 2005, p. 191), the maximization of the logarithm of the CML function in Eq. (9) is achieved by solving the composite score equations given by $s_{CML}(\boldsymbol{\theta},\boldsymbol{m}) = \nabla \log L_{CML}(\boldsymbol{\theta},\boldsymbol{m}) = \sum_{e=1}^{\tilde{E}} \omega_e s_e(\boldsymbol{\theta},\boldsymbol{m}) = \boldsymbol{0}$, where $s_e(\boldsymbol{\theta},\boldsymbol{m}) = \nabla \log L_e(\boldsymbol{\theta};\boldsymbol{m})$. Since these equations are linear combinations of valid likelihood score functions associated with the event probabilities forming the composite log-likelihood function, they immediately satisfy the requirement of being unbiased. Further, if q independent observations on the vector $\boldsymbol{Y}$ are available (say $\boldsymbol{m}^1, \boldsymbol{m}^2, \boldsymbol{m}^3, \ldots, \boldsymbol{m}^Q$), as would be the case when there are several individuals q (q = 1, 2, 3,. . ., Q) with panel data or repeated choice data, then, in the asymptotic scenario that $Q \to \infty$ with D fixed, a central limit theorem and a first-order Taylor series expansion can be applied in the usual way (see, for example, Godambe, 1960) to the resulting mean composite score function $\left(=\frac{1}{Q}\sum_{q=1}^{Q} s_{CML,q}(\boldsymbol{\theta},\boldsymbol{m^q})\right)$ to obtain consistency and asymptotic normality of the CML estimator (see also Lindsay, 1988; Cox and Reid, 2004; Zhao and Joe, 2005, Theorem 1):

$$\sqrt{Q}(\hat{\boldsymbol{\theta}}_{CML} - \boldsymbol{\theta}) \xrightarrow{d} N_{\tilde{K}}[(0, \boldsymbol{G}^{-1}(\boldsymbol{\theta})], \tag{13}$$

where $\boldsymbol{G}(\boldsymbol{\theta})$ is the Godambe information matrix defined as $\boldsymbol{H}(\boldsymbol{\theta})[\boldsymbol{J}(\boldsymbol{\theta})]^{-1}[\boldsymbol{H}(\boldsymbol{\theta})]$. $\boldsymbol{H}(\boldsymbol{\theta})$ and $\boldsymbol{J}(\boldsymbol{\theta})$ take the following form:

$$\boldsymbol{H}(\boldsymbol{\theta}) = E\left[-\frac{\partial^2 \log L_{CML}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] \quad \text{and} \quad \boldsymbol{J}(\boldsymbol{\theta}) = E\left[\left(\frac{\partial \log L_{CML}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \log L_{CML}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)\right]. \tag{14}$$

These may be estimated in a straightforward manner at the CML estimate $\hat{\boldsymbol{\theta}}_{\boldsymbol{CML}}$ as follows:

$$\hat{\boldsymbol{H}}(\hat{\boldsymbol{\theta}}) = -\left[\sum_{q=1}^{Q} \frac{\partial^2 \log L_{CML,q}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]_{\hat{\boldsymbol{\theta}}_{CML}}, \quad \text{and} \quad \hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}) = \sum_{q=1}^{Q}\left[\left(\frac{\partial \log L_{CML,q}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \log L_{CML,q}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)\right]_{\hat{\boldsymbol{\theta}}_{CML}}. \tag{15}$$

Even in the case where the data include very few or no independent replicates (as would be the case with global social or spatial interactions across all individuals in a cross-sectional data in which D is equal to the number of individuals), the CML estimator will retain the good properties of being consistent and asymptotically normal as long as the data is formed by pseudo-independent and overlapping subsets of observations (such as would be the case when the social interactions taper off relatively quickly with the social separation distance between decision-makers, or when spatial interactions rapidly fade with geographic distance based on an autocorrelation function decaying toward zero; see Cox and Reid (2004) for a technical discussion).[4]

Of course, the Cramer–Rao inequality implies that the CML estimator loses some asymptotic efficiency from a theoretical perspective relative to a full likelihood estimator (Lindsay, 1988; Zhao and Joe, 2005).[5] On the other hand, when simulation methods have to be used to evaluate the likelihood function, there is also a loss in asymptotic efficiency in the maximum simulated likelihood (MSL) estimator relative to a full likelihood estimator (see McFadden and Train, 2000).[6] Consequently,

[4] Otherwise, there may be no real solution to the CML function maximization and the asymptotic results laid out above will not hold.

[5] The theoretical efficiency loss of the CML estimator compared to the full information maximum likelihood (ML) estimator, if such an estimator is feasible, originates from the failure of the information identity (*i.e.*, $[\boldsymbol{H}(\boldsymbol{\theta})]^{-1} \neq \boldsymbol{J}(\boldsymbol{\theta})$). This is also referred to as the failure of the second Bartlett identity. In particular, from a theoretical standpoint, the difference between the asymptotic variances of the CML estimator (*i.e.*, $\boldsymbol{V_{CML}}(\boldsymbol{\theta}) = [\boldsymbol{G}(\boldsymbol{\theta})]^{-1}$) and the ML estimator ($[\boldsymbol{H}(\boldsymbol{\theta})]^{-1}$) is positive semi-definite (see Cox and Reid, 2004; Zhao and Joe, 2005).

[6] Specifically, McFadden and Train (2000) indicate, in their use of independent number of random draws across observations, that the difference between the asymptotic covariance matrix of the MSL estimator obtained as the inverse of the sandwich information matrix and the asymptotic covariance matrix of the ML estimator obtained as the inverse of the cross-product of first derivatives is also theoretically positive semi-definite for finite number of draws per observation.

it is difficult to state from a theoretical standpoint whether the CML estimator efficiency will be higher or lower than the MSL estimator efficiency. However, in a simulation comparison of the CML and MSL methods for multivariate ordered response systems, Bhat et al. (2010b) found that the CML estimator's relative efficiency was almost as good as that of the MSL estimator, but with the benefits of a very substantial reduction in computational time and much superior convergence properties. Even when the log-likelihood function may be maximized without simulations, several studies have found that the efficiency loss of the CML estimator (relative to the maximum likelihood (ML) estimator) appears to be negligible to small from an empirical standpoint (see Zhao and Joe, 2005; Lele, 2006; Joe and Lee, 2009).[7]

In the next few sections, we discuss how the MVNCD function approximation of Section 2.1 and the CML inference approach of Section 2.2 can be gainfully combined in the proposed MACML inference approach for MNP models.

3. Cross-sectional multinomial probit model

In the discussion below, we will assume that the number of choice alternatives in the choice set is the same across all individuals. The case of different numbers of choice alternatives per individual poses no complications whatsoever, since the only change in such a case is that the dimensionality of the multivariate normal cumulative distribution (MVNCD or normal orthant probability) function changes from one individual to the next.

A multinomial probit model may arise from several underlying structural motivations. In this section, we will consider a random-coefficients normal formulation superimposed on an IID normal error structure as the mechanism that leads to the typical cross-sectional MNP model. We do so to streamline the presentation, especially because the panel MNP structure we consider in the next section is a natural extension of the cross-sectional random-coefficients MNP structure. In the random-coefficients formulation, write the utility that an individual q associates with alternative I as follows:

$$U_{qi} = \boldsymbol{\beta}_{\boldsymbol{q}}' \boldsymbol{x}_{\boldsymbol{qi}} + \varepsilon_{qi} \tag{16}$$

where $\boldsymbol{x}_{\boldsymbol{qi}}$ is a $(K \times 1)$-column vector of exogenous attributes, $\boldsymbol{\beta}_{\boldsymbol{q}}$ is an individual-specific $(K \times 1)$-column vector of corresponding coefficients that varies across individuals based on unobserved individual attributes, and ε_{qi} is assumed to be an independently and identically distributed (across alternatives and individuals) normal error term with a variance of one-half. Assume that the $\boldsymbol{\beta}_{\boldsymbol{q}}$ vector in Eq. (16) is a realization from a multivariate normal distribution with a mean vector $\boldsymbol{b}$ and covariance matrix $\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{L}'$, where $\boldsymbol{L}$ is the lower-triangular Cholesky factor of $\boldsymbol{\Omega}$.

The traditional classical MSL approach to estimate the model of Eq. (16) depends upon the dimensionality of $\boldsymbol{\beta}_{\boldsymbol{q}}$ (=K) relative to the dimensionality of I. If $K < I - 2$, then it is convenient to write the likelihood contribution of individual q who chooses alternative m as:

$$L_q(\boldsymbol{b}, \boldsymbol{\Omega}) = \int_{\boldsymbol{\beta}=-\infty}^{\infty} \left\{ \int_{\lambda=-\infty}^{\infty} \left(\prod_{i \neq m} [\Phi\{[-\sqrt{2}(\boldsymbol{\beta}' \boldsymbol{z}_{\boldsymbol{qim}})] + \lambda\}] \right) \phi(\lambda) d\lambda \right\} f(\boldsymbol{\beta} | \boldsymbol{b}, \boldsymbol{\Omega}) d\boldsymbol{\beta} \tag{17}$$

where $\lambda = \sqrt{2} \cdot \varepsilon_{qm}$, $\phi(\cdot)$ is the standard normal density function, and $f(\cdot)$ is the multivariate normal density function with mean $\boldsymbol{b}$ and covariance $\boldsymbol{\Omega}$. The result is a multi-dimensional integral of dimension $K + 1$. On the other hand, if $K > I - 2$, then it is convenient to write the likelihood contribution of individual q by noting that the latent utility differentials $y^*_{qim} (= U_{qi} - U_{qm}, i \neq m)$ have a mean vector of $\boldsymbol{B}_{\boldsymbol{q}} (\boldsymbol{b}' \boldsymbol{z}_{\boldsymbol{q1m}}, \boldsymbol{b}' \boldsymbol{z}_{\boldsymbol{q2m}}, \ldots, \boldsymbol{b}' \boldsymbol{z}_{\boldsymbol{qIm}})'$ and a covariance matrix given by $\boldsymbol{\Sigma}_{\boldsymbol{q}} = \boldsymbol{z}_{\boldsymbol{q}} \boldsymbol{\Omega} \boldsymbol{z}_{\boldsymbol{q}}' + \boldsymbol{I}\boldsymbol{D}_{(\boldsymbol{I-1})}$, where $\boldsymbol{z}_{\boldsymbol{q}}$ is an $(I - 1) \times K$ matrix of independent variables obtained by vertically concatenating the transpose of the $K \times 1$ vectors $\boldsymbol{z}_{\boldsymbol{qim}}$ ($\boldsymbol{z}_{\boldsymbol{qim}} = \boldsymbol{x}_{\boldsymbol{qi}} - \boldsymbol{x}_{\boldsymbol{qm}}$, $i = 1, 2, \ldots, I, i \neq m$), and $\boldsymbol{I}\boldsymbol{D}_{\boldsymbol{I-1}}$ is a matrix of size $(I - 1)$ with ones on the diagonal and values of 0.5 on the off-diagonals. The likelihood contribution of individual q choosing alternative m then takes the multidimensional $(I - 1)$ integral form below:

$$L_q(\boldsymbol{b}, \boldsymbol{\Omega}) = F_{(I-1)}(-\boldsymbol{B}_{\boldsymbol{q}}, \boldsymbol{\Sigma}_{\boldsymbol{q}}) \tag{18}$$

where F_{I-1} is the multivariate cumulative normal distribution of $(I - 1)$ dimensions.

[7] A handful of studies (see Hjort and Varin, 2008; Mardia et al., 2009; Cox and Reid, 2004) have also theoretically examined the limiting normality properties of the CML approach, and compared the asymptotic variance matrices from this approach and the ML approach. However, such a precise theoretical analysis is possible only for extremely simple models.

The MVNCD approximation of Section 2.1 is computationally efficient and straightforward to implement when maximizing the likelihood function of Eq. (18).[8,9] As such, the MVNCD approximation can be used for any value of K and any value of I, as long as there is data support for the estimation of parameters. Of course, parsimonious factor-analytic or other spatial structures may be imposed on the covariance matrix $\boldsymbol{\Omega}$ based on the process under study to reduce the number of parameters to be estimated and increase estimator efficiency.

One final issue in the MACML estimation relates to the procedure to ensure that the symmetric matrix $\boldsymbol{\Omega}$ is positive-definite (that is, all the eigenvalues of the matrix should be positive, or, equivalently, the determinant of the entire matrix and every principal submatrix of $\boldsymbol{\Omega}$ should be positive). To do so, $\boldsymbol{\Omega}$ may be reparameterized through a Cholesky matrix decomposition, and then these Cholesky-decomposed parameters may be estimated.

4. Panel multinomial probit models

In the discussion below, we will assume that the number of choice occasions per individual is the same across all individuals. We discuss the case of different numbers of choice occasions per individual in Section 4.2.

4.1. The panel MNP model

Consider the following model with 't' now being an index for choice occasion:

$$U_{qit} = \boldsymbol{\beta}_{\boldsymbol{q}}' \boldsymbol{x}_{\boldsymbol{qit}} + \varepsilon_{qit}, \boldsymbol{\beta}_{\boldsymbol{q}} \sim MVN(\boldsymbol{b}, \boldsymbol{\Omega}), q = 1, 2, \ldots, Q,\ i = 1, 2, \ldots, I,\ t = 1, 2, \ldots, T \tag{19}$$

Let ε_{qit} be IID normal over individuals, alternatives, and choice occasions with a variance of 0.5. We will assume that the coefficients $\boldsymbol{\beta}_{\boldsymbol{q}}$ are constant over choice situations of a given decision maker.

The traditional simulation procedures are similar to the cross-sectional case. Consider an individual who selects alternative m_t at the tth choice occasion. When the number of random coefficients K (the cardinality of the vector $\boldsymbol{\beta}_{\boldsymbol{q}}$) is less than $[(I-1)*T]-2$, as will mostly be the case in application, it is convenient to write the likelihood contribution of individual q as:

$$L_q(\boldsymbol{b}, \boldsymbol{\Omega}) = \int_{\boldsymbol{\beta}=-\infty}^{\infty} \prod_{t=1}^{T} \left[\int_{\lambda=-\infty}^{\infty} \left(\prod_{i \neq m_t} [\Phi\{[-\sqrt{2}(\boldsymbol{\beta}' \boldsymbol{z}_{\boldsymbol{qim_t t}})] + \lambda\}] \right) \phi(\lambda) d\lambda \right] f(\boldsymbol{\beta} | \boldsymbol{b}, \boldsymbol{\Omega}) \boldsymbol{d\beta} \tag{20}$$

where $\boldsymbol{z}_{\boldsymbol{qim_t t}} = (\boldsymbol{x}_{\boldsymbol{qit}} - \boldsymbol{x}_{\boldsymbol{qm_t t}})$. Another approach is to write the likelihood contribution in terms of the latent utility differentials $y^*_{qim_t t} = \boldsymbol{\beta}_{\boldsymbol{q}}' \boldsymbol{z}_{\boldsymbol{qim_t t}} + \eta_{qim_t t}, \eta_{qim_t t} = \varepsilon_{qit} - \varepsilon_{qm_t t}\ (i \neq m_t)$. These latent utility differentials have an $(I-1)*T$ mean vector $\boldsymbol{B}_{\boldsymbol{q}}(\boldsymbol{b}' \boldsymbol{z}_{\boldsymbol{q1m1}}, \boldsymbol{b}' \boldsymbol{z}_{\boldsymbol{q2m_1 1}} \ldots \boldsymbol{b}' \boldsymbol{z}_{\boldsymbol{qIm_1 1}} (i \neq m_1); \boldsymbol{b}' \boldsymbol{z}_{\boldsymbol{q1m_2 2}}, \boldsymbol{b}' \boldsymbol{z}_{\boldsymbol{q2m_2 2}} \ldots \boldsymbol{b}' \boldsymbol{z}_{\boldsymbol{qIm_2 2}} (i \neq m_2); \ldots, \boldsymbol{b}' \boldsymbol{z}_{\boldsymbol{q1m_T T}}, \boldsymbol{b}' \boldsymbol{z}_{\boldsymbol{q2m_T T}}, \ldots, \boldsymbol{b}' \boldsymbol{z}_{\boldsymbol{qIm_T T}} (i \neq m_T))$ and a covariance matrix given by $\boldsymbol{\Sigma}_{\boldsymbol{q}} = \tilde{\boldsymbol{z}}_{\boldsymbol{q}} \boldsymbol{\Omega} \tilde{\boldsymbol{z}}_{\boldsymbol{q}}' + \boldsymbol{ID}_{\boldsymbol{T \times (I-1)}}$, where $\tilde{\boldsymbol{z}}_{\boldsymbol{q}}$ is a $[T \times (I-1)] \times K$ matrix obtained by vertically concatenating the transpose of the $K \times 1$ vectors $z_{\boldsymbol{qim_t t}}$ $(i = 1, 2, \ldots, I,\ i \neq m_t;\ t = 1, 2, \ldots, T)$ (note that there are $T \times (I-1)$ vectors in $z_{\boldsymbol{qim_t t}}$), and $\boldsymbol{ID}_{\boldsymbol{T \times (I-1)}}$ is a block-diagonal matrix with each block matrix of size $(I-1) \times (I-1)$ with values of one along the diagonal and values of 0.5 on the off-diagonals. The likelihood contribution of individual q then takes the multidimensional $(I-1) \times T$ integral form below:

$$L_q(\boldsymbol{b}, \boldsymbol{\Omega}) = F_{(I-1) \times T}(-\boldsymbol{B}_{\boldsymbol{q}}, \boldsymbol{\Sigma}_{\boldsymbol{q}}), \tag{21}$$

with $F_{(I-1) \times T}$ being the multivariate cumulative normal distribution of $(I-1) \times T$ dimensions.

The simulation approaches for evaluating the panel likelihood function are time-consuming. In our MACML estimation approach, we propose a combination of the approximation method for multivariate normal orthant probabilities and the composite marginal likelihood method. Specifically, based on Eq. (12) and the notation defined there, the analyst may construct the following pairwise CML function across the choice occasions of individual q:

$$L_{CML,q}(\boldsymbol{b}, \boldsymbol{\Omega}) = \prod_{t=1}^{T-1} \prod_{w=t+1}^{T} \text{Prob}(C_{qt} = m_t, C_{qw} = m_w) = \prod_{t=1}^{T-1} \prod_{w=t+1}^{T} \text{Prob}\ [y^*_{qim_t t} < 0 \forall i \neq m_t \text{and} y^*_{qim_w w} < 0 \forall i \neq m_w] \tag{22}$$

[8] As indicated earlier, the CML class of estimators subsumes the usual ordinary full-information likelihood estimator as a special case. It is this characteristic of the CML approach that leads us to the label MACML for the estimation approach proposed here. Specifically, even in cross-sectional mixing distribution contexts, when our approach involves only the approximation of the maximum likelihood function, the MACML label is appropriate since the maximum likelihood function is a special case of the CML function. Of course, in a panel context or in cross-sectional/panel contexts with spatial/social error dependencies, we use a specific pairwise (and non-ML) technique within the CML approach for estimation, as discussed in Section 4 and Section 5.

[9] The use of the MVNCD approximation (as discussed in Section 2.1) has been shown to be accurate in the context of evaluating single multivariate integrals. Joe (1995) indicates that the approximation has an error (even in the worst case of high correlations) in the third decimal place. In a companion paper, we have examined the performance of the MVNCD approximation in the context of estimating parameters in cross-sectional and panel multinomial probit models. The results indicate that the approximation provides parameter values very close to the "true" population parameter values in simulation experiments, with the empirical absolute percentage bias being smaller than that from regular simulation techniques to evaluate the MVNCD function. Thus, the MVNCD-approximated log-likelihood function as proposed here should be close to the log-likelihood function for all parameters in a neighborhood of the "true" parameter values, which implies that the covariance matrix computed using our MACML procedure should also be an accurate approximation to the actual covariance matrix.

The computational effort is reduced in the CML above because only pairwise marginal multivariate probabilities are being considered across choice occasions. However, each multivariate orthant probability above still has a dimension equal to $(I-1)\times 2$. But such an orthant probability is conveniently computed using the approximation of Section 2.1, leading to solely bivariate and univariate cumulative normals. This is a remarkable decomposition and simplification. In this approximation, the bivariate probability expression for the latent variables within the same time period t takes the form shown below ($i, g \neq m_t$, $i \neq g$):

$$\text{Prob}[y^*_{qim_t t} < 0, y_{qgm_t t} < 0] = \Phi_2\left[\frac{-\boldsymbol{b}'\boldsymbol{z}_{\boldsymbol{qim_{bit}t}}}{\sqrt{\text{Var}(y^*_{qim_t t})}}, \frac{-\boldsymbol{b}'\boldsymbol{z}_{\boldsymbol{qgm_t t}}}{\sqrt{\text{Var}(y^*_{qgm_t t})}}, \rho_{qigm_t t}\right], \text{ where} \tag{23}$$

$\rho_{qigm_t t} = \frac{\text{Cov}(y^*_{qim_t t}, y^*_{qgm_t t})}{\sqrt{\text{Var}(y^*_{qim_t t})\text{Var}(y^*_{qgm_t t})}}$, and $\text{Var}(y^*_{qim_t t}), \text{Var}(y^*_{qgm_t t})$, and $\rho_{qigm_t t}$ are obtained from the following (2×2) matrix:

$$\boldsymbol{\Sigma}_{\boldsymbol{qigt}} = \boldsymbol{\Delta}_{\boldsymbol{igt}}(\tilde{\boldsymbol{z}}_{\boldsymbol{q}}\boldsymbol{\Omega}\tilde{\boldsymbol{z}}'_{\boldsymbol{q}})\boldsymbol{\Delta}'_{\boldsymbol{igt}} + \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}. \tag{24}$$

In the equation above, $\boldsymbol{\Delta}_{\boldsymbol{igt}}$ is a selection matrix of size $2\times[(I-1)\times T]$. It has a value of one in the $[(I-1)\times(t-1)+i]$th column in the first row if $i < m_t$ or a value of one in the $[(I-1)\times(t-1)+(i-1)]$th column in the first row if $i > m_t$. Similarly, it has a value of one in the $[(I-1)\times(t-1)+g]$th column in the second row if $i < m_t$ or a value of one in the $[(I-1)\times(t-1)+(g-1)]$th column in the second row if $i > m_t$. The matrix has values of zero everywhere else.The bivariate probability expressions for the latent variables across time periods t and w take the form shown below:

$$\text{Prob}[y^*_{qim_t t} < 0, y_{qgm_w w} < 0] = \Phi_2\left[\frac{-\boldsymbol{b}'\boldsymbol{z}_{\boldsymbol{qim_t t}}}{\sqrt{\text{Var}(y^*_{qim_t t})}}, \frac{-\boldsymbol{b}'\boldsymbol{z}_{\boldsymbol{qgm_w w}}}{\sqrt{\text{Var}(y^*_{qgm_w w})}}, \rho_{qigm_t m_w tw}\right], \tag{25}$$

where $\rho_{qigm_t m_w tw} = \frac{\text{Cov}(y^*_{qim_t t}, y^*_{qgm_w w})}{\sqrt{\text{Var}(y^*_{qim_t t})\text{Var}(y^*_{qgm_w w})}}$, and $\text{Var}(y^*_{qim_t t}), \text{Var}(y^*_{qgm_w w})$, and $\text{Cov}(y^*_{qim_t t}, y^*_{qgm_w w})$ are obtained from the following (2×2) matrix:

$$\boldsymbol{\Sigma}_{\boldsymbol{qigtw}} = \boldsymbol{\Delta}_{\boldsymbol{igtw}}(\tilde{\boldsymbol{z}}_{\boldsymbol{q}}\boldsymbol{\Omega}\tilde{\boldsymbol{z}}'_{\boldsymbol{q}})\boldsymbol{\Delta}'_{\boldsymbol{igtw}} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{26}$$

In the above expression, $\boldsymbol{\Delta}_{\boldsymbol{igtw}}$ is a selection matrix of size $2\times[(I-1)\times T]$. This matrix has a value of one in the $[(I-1)\times(t-1)+i]$th column in the first row if $i < m_t$ or a value of one in the $[(I-1)\times(t-1)+(i-1)]$th column in the first row if $i > m_t$. Similarly, it has a value of one in the $[(I-1)\times(w-1)+g]$th column in the second row if $g < m_w$ or a value of one in the $[(I-1)\times(w-1)+(g-1)]$th column in the second row if $g > m_w$. The matrix has values of zero everywhere else.

The covariance matrix of the estimator may then be obtained using the inverse of the Godambe information matrix $G(=\boldsymbol{H}(\boldsymbol{\theta})[\boldsymbol{J}(\boldsymbol{\theta})]^{-1}\boldsymbol{H}(\boldsymbol{\theta}))$, where $\boldsymbol{H}(\boldsymbol{\theta})$ and $\boldsymbol{J}(\boldsymbol{\theta})$ may be estimated as in Eq. (15) with $\boldsymbol{\theta} = [\boldsymbol{b}', (\text{Vech}(\boldsymbol{\Omega}))']'$, where $\text{Vech}(\boldsymbol{\Omega})$ represents the column vector of upper triangle elements of $\boldsymbol{\Omega}$.

4.2. The case of unequal number of choice occasions

In Section 4.1, we assumed the same number of choice occasions T per individual. Now, consider the case when there are unequal numbers of choice occasions per individual, which is not uncommon with panel data. Let individual q be observed to contribute T_q choice occasions. Then, the unweighted CML functions presented earlier will give more weight to individuals who have more choice occasions than those who have fewer choice occasions. To address this situation, a weighted CML function may be used (see Kuk and Nott, 2000; Renard et al., 2004). Note that, in the context of the discussion in Sections 4.1, and in the absence of any random coefficient effects (that is, when all elements of $\boldsymbol{\Omega}$ are zero), the correlation between all inter-temporal pairs of utility differentials $(y^*_{qim_t t}, y^*_{qjm_w w}), i \neq m_t, g \neq m_w)$ are zero, and Eq. (22) collapses as follows:

$$\begin{aligned} L_{CML,q}(\boldsymbol{b},\boldsymbol{\Omega}) &= \prod_{t=1}^{T_q-1}\prod_{w=t+1}^{T_q}\text{Prob}(C_{qt}=m_t)\times\text{Prob}(C_{qw}=m_w) \\ &= \prod_{t=1}^{T_q-1}\prod_{w=t+1}^{T_q}\text{Prob}[y^*_{qim_t t}<0\forall i\neq m_t]\times\text{Prob}[y^*_{qim_w w}<0\forall i\neq m_w] \end{aligned} \tag{27}$$

On the other hand, when there are random coefficients, the overall correlation between any inter-temporal pair of utility differentials could fall or rise from 0.0, depending upon the nature of the matrix $\boldsymbol{\Omega}$ and the matrix $\boldsymbol{z}_{\boldsymbol{q}}$. However, except in extreme cases, such as when exogenous variables are of opposite signs across time periods for different alternatives or for the same alternative, the net result of the presence of random coefficients will be an increase in the magnitude of correlation from the value of zero.

In the first case discussed above (that is, the absence of random coefficients), Le Cessie and Van Houwelingen (1994) suggest, in their binary data context, that each individual should contribute about equally to the CML function. This may

be achieved by power-weighting each individual's likelihood contribution by a factor that is the inverse of the number of choice occasions minus one (in our context, this is $[T_q - 1]^{-1}$). The net result is that the composite likelihood contribution of individual q collapses to the likelihood contribution of the individual under the case of independence across choice occasions. In a general correlated binary data context, Kuk and Nott (2000) confirmed the above result for efficiently estimating the mean vector $\boldsymbol{b}$ in the case when the correlation is close to zero. However, their analysis suggested that the unweighted CML function remains superior for estimating the correlation parameters (in our context, this corresponds to the elements of the $\boldsymbol{\Omega}$ matrix). In a recent paper, Joe and Lee (2009) theoretically studied the issue of efficiency in the context of a simple random effect binary choice model. They indicate that the weights suggested by Le Cessie and Van Houwelingen (1994) and Kuk and Nott (2000) can provide poor efficiency even for the mean vector $\boldsymbol{b}$ when the correlation between pairs of the underlying latent variables for the repeated binary choices over time is moderate to high. Based on analytic and numeric analyses using a longitudinal binary choice model with an autoregressive correlation structure, they suggest that using a weight of $(T_q - 1)^{-1}[1 + 0.5(T_q - 1)]^{-1}$ for individual q appears to do well in terms of efficiency for all parameters and across varying dependency levels. Thus, our suggested composite likelihood contribution for individual q in the unbalanced panel case is to weight each individual's composite likelihood contribution by a power factor of $(T_q - 1)^{-1}[1 + 0.5(T_q - 1)]^{-1}$. Of course, for MNP contexts where the number of choice occasions across individuals is not substantially different, the difference between the weighted and unweighted CML functions in terms of estimator efficiency may not be substantial. This is an issue that needs to be studied empirically, and is left as a direction for further research.

4.3. Other panel extensions

We now discuss the MACML estimation approach for a few more extensions of the traditional panel probit models. In the following discussion, we assume the same number of choice occasions across individuals. This will help keep the presentation streamlined. Extending the models below to the case of different numbers of choice occasions across individuals poses no substantial difficulties, since weights can be used as just discussed.

4.3.1. Intra- and cross-temporal random coefficients

Consider the following utility function:

$$U_{qit} = \boldsymbol{\beta}'_{\boldsymbol{qt}}\boldsymbol{x}_{\boldsymbol{qit}} + \varepsilon_{qit} \tag{28}$$

where $\boldsymbol{\beta}_{\boldsymbol{qt}} = \boldsymbol{b} + \tilde{\boldsymbol{\beta}}_{\boldsymbol{q}} + \tilde{\boldsymbol{\beta}}_{\boldsymbol{qt}}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{q}} \sim N(0, \boldsymbol{\Omega}), \tilde{\boldsymbol{\beta}}_{\boldsymbol{qt}} \sim N(0, \tilde{\boldsymbol{\Omega}})$, and ε_{qit} is IID normal across individuals, alternatives, and choice occasions with a variance of 0.5. The above form for $\boldsymbol{\beta}_{\boldsymbol{qt}}$ generates covariance across all the choice occasions of individual q and across alternatives i at any time t, due to the $\tilde{\boldsymbol{\beta}}_{\boldsymbol{q}}$ term (this is the same as the usual panel mixed model). However, there is an additional covariance across alternatives i at any time t due to the $\tilde{\boldsymbol{\beta}}_{\boldsymbol{qt}}$ term (see Bhat and Castelar (2002) and Bhat and Sardesai (2006) for such a specification, but with a logit kernel). Using the same notation as earlier, the MACML estimation of this model is straightforward and takes the same form as Eqs. (22), (23), and (25) with

$$\boldsymbol{\Sigma}_{\boldsymbol{qigt}} = \Delta_{\boldsymbol{igt}}[\tilde{\boldsymbol{z}}_{\boldsymbol{q}}(\boldsymbol{\Omega} + \tilde{\boldsymbol{\Omega}})\tilde{\boldsymbol{z}}'_{\boldsymbol{q}}]\Delta'_{\boldsymbol{igt}} + \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \tag{29}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{qigtw}} = \Delta_{\boldsymbol{igtw}}[\tilde{\boldsymbol{z}}_{\boldsymbol{q}}\boldsymbol{\Omega}\tilde{\boldsymbol{z}}'_{\boldsymbol{q}}]\Delta_{\boldsymbol{igtw}} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

4.3.2. Autoregressive random coefficients structure

In this structure, consider an autoregressive structure of order one:

$$\boldsymbol{\beta}_{\boldsymbol{qt}} = \boldsymbol{b} + \tilde{\boldsymbol{\beta}}_{\boldsymbol{qt}}, \tag{30}$$

where $\tilde{\boldsymbol{\beta}}_{\boldsymbol{qt}} = \tilde{\rho}\tilde{\boldsymbol{\beta}}_{\boldsymbol{q,t-1}} + \boldsymbol{\mu}_{\boldsymbol{qt}}, \mu_{qt} \sim N(0, \tilde{\boldsymbol{\Omega}}), \mathrm{Cov}(\boldsymbol{\mu}_{\boldsymbol{qt}}, \boldsymbol{\mu}_{\boldsymbol{qw}}) = 0$ for $t \neq w$ and $\tilde{\rho}$ is an autocorrelation parameter with $|\tilde{\rho}| < 1$ to ensure stationarity (note that $\tilde{\rho}$ itself can be a function of individual-specific attributes and can be written as $\tilde{\rho}_q$; however, this is a trivial extension from a conceptual standpoint and so we will restrict the presentation to a single $\tilde{\rho}$ parameter). The specification above implies that $E(\tilde{\boldsymbol{\beta}}_{\boldsymbol{qt}}) = 0, \mathrm{Var}(\tilde{\boldsymbol{\beta}}_{\boldsymbol{qt}}) = \frac{\tilde{\boldsymbol{\Omega}}}{1-\tilde{\rho}^2}$, and $\mathrm{Cov}(\tilde{\boldsymbol{\beta}}_{\boldsymbol{qt}}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{qw}}) = \frac{\tilde{\rho}^{|t-w|}\tilde{\boldsymbol{\Omega}}}{1-\tilde{\rho}^2}$.

For this model, the CML function is the same again as in Eqs. (22), (23), and (25) with

$$\boldsymbol{\Sigma}_{\boldsymbol{qigt}} = \frac{1}{1-\tilde{\rho}^2}\Delta_{\boldsymbol{igt}}[\tilde{\boldsymbol{z}}_{\boldsymbol{q}}\tilde{\boldsymbol{\Omega}}\tilde{\boldsymbol{z}}'_{\boldsymbol{q}}]\Delta_{\boldsymbol{igt}} + \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \tag{31}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{qigtw}} = \frac{\tilde{\rho}^{|t-w|}}{1-\tilde{\rho}^2}\Delta_{\boldsymbol{igtw}}[\tilde{\boldsymbol{z}}_{\boldsymbol{q}}\tilde{\boldsymbol{\Omega}}\tilde{\boldsymbol{z}}'_{\boldsymbol{q}}]\Delta'_{\boldsymbol{igtw}} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

4.3.3. Individual-constant and autoregressive random coefficients structure

This structure takes the following form:

$$\boldsymbol{\beta_{qt}} = \boldsymbol{b} + \tilde{\boldsymbol{\beta}}_{\boldsymbol{q}} + \tilde{\boldsymbol{\beta}}_{\boldsymbol{qt}}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{q}} \sim N(0, \boldsymbol{\Omega}), \tag{32}$$

and

$$\tilde{\boldsymbol{\beta}}_{\boldsymbol{qt}} = \tilde{\rho}\tilde{\boldsymbol{\beta}}_{\boldsymbol{q},t-1} + \boldsymbol{\mu_{qt}}, \boldsymbol{\mu_{qt}} \sim N(0, \tilde{\boldsymbol{\Omega}}), \mathrm{Cov}(\boldsymbol{\mu_{qt}}, \boldsymbol{\mu_{qw}}) = 0 \quad \text{for} \quad t \neq w, |\tilde{\rho}| < 1.$$

The CML function is as in Eqs. (22), (23), and (25) with

$$\boldsymbol{\Sigma_{qigt}} = \Delta_{\boldsymbol{igt}} \left[\tilde{\boldsymbol{z}}_{\boldsymbol{q}} \left\{ \frac{\tilde{\boldsymbol{\Omega}}}{1-\tilde{\rho}^2} + \boldsymbol{\Omega} \right\} \tilde{\boldsymbol{z}}'_{\boldsymbol{q}} \right] \Delta'_{\boldsymbol{igt}} + \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \tag{33}$$

and

$$\boldsymbol{\Sigma_{qigtw}} = \Delta_{\boldsymbol{igtw}} [\tilde{\boldsymbol{z}}_{\boldsymbol{q}} \left\{ \frac{\tilde{\rho}^{|t-w|}\tilde{\boldsymbol{\Omega}}}{1-\tilde{\rho}^2} + \boldsymbol{\Omega} \right\} \tilde{\boldsymbol{z}}'_{\boldsymbol{q}}] \Delta'_{\boldsymbol{igtw}} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

4.3.4. Dynamic panel structure

Consider a first-order lagged dependence structure that takes the following form:

$$U_{qit} = \boldsymbol{b}'\boldsymbol{x_{qit}} + \tilde{\rho} U_{qit-1} + \varepsilon_{qit}, \quad \text{with} \quad |\tilde{\rho}| < 1. \tag{34}$$

ε_{qit} is IID normal across individuals, alternatives, and time periods with a variance of 0.5. Taking utility differences ($U_{qit} - U_{qm_t t}$) at the tth occasions ($i = 1, 2, \ldots, I, i \neq m_t$), the result is:[10]

$$y^*_{qim_t t} = \boldsymbol{b}'\boldsymbol{z_{qim_t t}} + \tilde{\rho} y^*_{qim_{t-1}t-1} + \eta_{qim_t t} \tag{35}$$

The expected value and the variance of $y^*_{qim_t t}$ based on the above equation may be written as $E(y^*_{qim_t t}) = \sum_{r=0}^{t-1} \tilde{\rho}^r \boldsymbol{b}'\boldsymbol{z_{qim_t t-r}}$, and $\mathrm{Var}(y^*_{qim_t t}) = \frac{1}{1-\tilde{\rho}^2}$. Also, we have the following marginal bivariate correlations ($i \neq m_t$, $g \neq m_w$):

$$\mathrm{Cor}(y^*_{qim_t t}, y^*_{qgm_t t}) = \rho_{qigm_t t} = 0.5(i \neq g, \ g \neq m_t) \tag{36}$$

$$\mathrm{Cor}(y^*_{qim_t t}, y^*_{qgm_w w}) = \rho_{qigm_t m_w tw} = \begin{cases} -\tilde{\rho}^{|t-w|} & \text{if } m_t \neq m_w; i \neq g; i = m_w; g = m_t \\ -0.5\tilde{\rho}^{|t-w|} & \text{if } m_t \neq m_w; i \neq g; i \neq m_w; g = m_t \\ -0.5\tilde{\rho}^{|t-w|} & \text{if } m_t \neq m_w; i \neq g; i = m_w; g \neq m_t \\ 0 & \text{if } m_t \neq m_w; i \neq g; i \neq m_w; g \neq m_t \\ 0.5\tilde{\rho}^{|t-w|} & \text{if } m_t \neq m_w; i = g \\ 0.5\tilde{\rho}^{|t-w|} & \text{if } m_t = m_w; i \neq g \\ \tilde{\rho}^{|t-w|} & \text{if } m_t = m_w; i = g \end{cases}$$

The MACML approach remains the same as for the panel MNP models, with the following bivariate probability expression for the latent variables within the same time period t: ($i, g \neq m_t, i \neq g$):

$$\mathrm{Prob}[y^*_{qim_t t} < 0, y^*_{qgm_t t} < 0] = \Phi_2 \left[-\sqrt{1-\tilde{\rho}^2} \sum_{r=0}^{t-1} \tilde{\rho}^r \boldsymbol{b}'\boldsymbol{z_{qim_t t-r}}, -\sqrt{1-\tilde{\rho}^2} \sum_{r=0}^{t-1} \tilde{\rho}^r \boldsymbol{b}'\boldsymbol{z_{qgm_t t-r}}, \rho_{qigm_t t} \right] \tag{37}$$

The corresponding expression for the bivariate probability expression for the latent variables across time periods is given by ($i \neq m_t, g \neq m_w$):

$$\mathrm{Prob}[y^*_{qim_t t} < 0, y^*_{qgm_w w} < 0] = \Phi_2 [-\sqrt{1-\tilde{\rho}^2} \sum_{r=0}^{t-1} \tilde{\rho}^r \boldsymbol{b}'\boldsymbol{z_{qim_t t-r}}, -\sqrt{1-\tilde{\rho}^2} \sum_{r=0}^{w-1} \tilde{\rho}^r \boldsymbol{b}'\boldsymbol{z_{qgm_w w-r}}, \rho_{qigm_t m_w tw}] \tag{38}$$

5. Extension to incorporate spatial correlation

In this section, we discuss the MACML application to multinomial probit models with spatial correlation, though the same procedures may be adopted to accommodate social interaction effects.

Spatial correlation may exist across discrete choice alternatives (see, for example, Bolduc et al., 1996; Bhat and Guo, 2004; Miyamoto et al., 2004; Sener et al., 2011) or across decision-makers (see, for example, Anselin, 2003; Fleming, 2004;

[10] Individuals are observed making choice from $t = 1$. The lagged utility structure brings up the issue of initial conditions, since we do not observe individual choices before $t = 1$. Here we assume that $U_{qio} = 0$ to resolve the initial conditions issue, which is equivalent to assuming equal probabilities of choice for any alternative i before observation begins. Other ways to accommodate initial conditions are discussed in Heckman and Singer (1986).

Franzese and Hays, 2008; Bhat and Sener, 2009). The focus here will be on spatial correlation across decision makers. Interestingly, in the context of spatial correlation across decision makers, almost all earlier studies have either focused on binary response models or ordered response models. In particular, spatial correlation across individuals has seldom even been discussed (let alone being studied) in the context of unordered-response models. On the other hand, spatial correlation in data may occur in unordered-response models for the same reasons (for example, diffusion effects, social interaction effects, and unobserved location-related effects) they have been studied extensively in binary and ordered-response models.

In terms of estimation of binary and ordered-response discrete choice models with a general spatial correlation structure, the analyst confronts, in the familiar probit model, a multi-dimensional integral over a multivariate normal distribution, which is of the order of the number of observational units in the data. While a number of approaches have been proposed to tackle this situation (see McMillen, 1995; LeSage, 2000; Pinkse and Slade, 1998; Fleming, 2004; Beron et al., 2003; Beron and Vijverberg, 2004), none of these remain practically feasible for moderate-to-large samples. These methods are also quite cumbersome and involved. In the context of unordered-response models, the situation becomes even more difficult – the likelihood function entails a multidimensional integral over a multivariate normal distribution, which is of the order of the number of observational units factored up by the number of alternatives minus one. This situation, however, is relatively easily handled using the MACML method, as we discuss below. As in Section 4, we assume the same number of alternatives across individuals to help keep the presentation streamlined. We also focus on a cross-section spatial formulation. The extension to a panel formulation is similar to the extensions from the cross-sectional to the panel structure in the non-spatial structures discussed earlier.

The next section discusses a spatial error model, while Section 5.2 presents the minor modifications that need to be made to handle the spatial lag model.

5.1. Spatial error model

Consider the following specification of utility for individual q and alternative i:

$$\begin{aligned} U_{qi} &= \boldsymbol{\beta}_{\boldsymbol{q}}' \boldsymbol{x}_{\boldsymbol{qi}} + v_{qi}; \boldsymbol{\beta}_{\boldsymbol{q}} = b + \tilde{\beta}_q, \tilde{\beta}_q \sim N(\boldsymbol{0}, \boldsymbol{\Omega}) \\ v_{qi} &= \rho \sum_{q'} w_{qq'} v_{qi} + \xi_{qi}; \xi_{qi} \sim N(0, \boldsymbol{\Lambda}), |\rho| < 1. \end{aligned} \tag{39}$$

In the above formulation $w_{qq'}$ is the spatial weight corresponding to individuals q and q', with $w_{qq} = 0$ and $\sum_{q'} w_{qq'} = 1$ for each (and all) q. We also assume that ξ_{qi} is independent and identically distributed across q, but allow a general covariance structure across alternatives for individual q. As usual, appropriate scale and level normalization must be imposed on $\boldsymbol{\Lambda}$ or identifiability. Specifically, only utility differentials matter in discrete choice models. Taking the utility differentials with respect to the first alternative, only the elements of the covariance matrix $\boldsymbol{\Lambda}_1$ of $\tilde{\xi}_{qi1} = \xi_{qi} - \xi_{q1} (i \neq 1)$ are estimable. However, the MACML inference approach proposed here, like the traditional GHK simulator, takes the difference in utilities against the chosen alternative during estimation. Thus, if individual q is observed to choose alternative m_q, the covariance matrix $\boldsymbol{\Lambda}_{\boldsymbol{m_q}}$ is desired for the individual. However, even though different differenced covariance matrices are used for different individuals, they must originate in the same original values for $\boldsymbol{\Lambda}$. To achieve this consistency, $\boldsymbol{\Lambda}$ is constructed from $\boldsymbol{\Lambda}_1$ by adding an additional row on top and an additional column to the left. All elements of this additional row and additional column are filled with values of zeros. An additional normalization needs to be imposed on $\boldsymbol{\Lambda}$ because the scale is also not identified. For this, we normalize the element of $\boldsymbol{\Lambda}$ in the second row and second column to the value of one. Note that these normalizations are innocuous and are needed for identification. The $\boldsymbol{\Lambda}$ matrix so constructed is fully general.

Some additional notation now. Define the following: $\boldsymbol{U_q} = (U_{q1}, U_{q2}, ...U_{qI})'$ ($I \times 1$ vector), $\boldsymbol{U} = (\boldsymbol{U_1}, \boldsymbol{U_2}, ...\boldsymbol{U_Q})'$ ($QI \times 1$ vector), $\boldsymbol{\xi_q} = (\xi_{q1}, \xi_{q2}, ...\xi_{qI})'$ ($I \times 1$ vector), $\boldsymbol{\xi} = (\boldsymbol{\xi_1'}, \boldsymbol{\xi_2'}, \ldots \boldsymbol{\xi_Q'})'$ ($QI \times 1$ vector), $\boldsymbol{x_q} = (\boldsymbol{x_{q1}}, \boldsymbol{x_{q2}}, ...\boldsymbol{x_{qI}})'$ ($I \times K$ matrix), $\boldsymbol{x} = (\boldsymbol{x}_1', \boldsymbol{x}_2', \ldots \boldsymbol{x}_Q')'$ ($QI \times K$ matrix), and $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1', \tilde{\boldsymbol{\beta}}_2', \ldots, \tilde{\boldsymbol{\beta}}_Q')'$ ($QK \times 1$ vector). Let $\boldsymbol{IDEN}_E$ be the identity matrix of size E, and $\boldsymbol{1}_E$ be a column vector of size E with all of its elements taking the value of one. Also, define the following matrix:

$$\tilde{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{x}_2 & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{x}_3 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{x}_Q \end{bmatrix} (QI \times QK \text{ matrix}), \tag{40}$$

Then, we can write Eq. (39) in a compact form as:

$$\boldsymbol{U} = \boldsymbol{xb} + \tilde{\boldsymbol{x}}\tilde{\boldsymbol{\beta}} + \boldsymbol{S}\boldsymbol{\xi}, \tag{41}$$

where $\boldsymbol{S} = [\boldsymbol{IDEN_{QI}} - (\boldsymbol{\rho W} \otimes \boldsymbol{IDEN_I})]^{-1}$ ($QI \times QI$ matrix), and $\boldsymbol{W}$ is the $(Q \times Q)$ weight matrix with the weights $w_{qq'}$ as its elements. Let $[\cdot]_e$ indicate the eth element of the column vector $[\cdot]$, and let $d_{qi} = (q-1)I + i$. Eq. (39) can be equivalently written using Eq. (41) as:

$$U_{qi} = [\boldsymbol{xb}]_{d_{qi}} + [\tilde{\boldsymbol{x}}\tilde{\boldsymbol{\beta}} + \boldsymbol{S}\boldsymbol{\xi}]_{d_{qi}}. \tag{42}$$

Define $V_{qi} = [\boldsymbol{xb}]_{d_{qi}}$ and $\varepsilon_{qi} = [\tilde{\boldsymbol{x}}\tilde{\boldsymbol{\beta}} + \boldsymbol{S\xi}]_{d_{qi}}$. Next, as earlier, let individual q be observed to choose alternative m_q. Stack the latent utility differentials $y^*_{qim_q} (= U_{qi} - U_{qm_q}, i \neq m_q)$ as follows: $\boldsymbol{y^*_q} = (y^*_{q1m_q}, y^*_{q2m_q}, \ldots, y^*_{qIm_q})'$, and $\boldsymbol{y^*} = (\boldsymbol{y}^{*\prime}_1, \boldsymbol{y}^{*\prime}_2, \ldots, \boldsymbol{y}^{*\prime}_Q)'$. Thus $\boldsymbol{y^*}$ is an $(I-1)*Q$ vector. Also, let $H_{qim_q} = V_{qi} - V_{qm_q}$. The likelihood of the observed sample (*i.e.*, individual 1 choosing alternative m_1, individual 2 choosing alternative $m_2, \ldots$, individual Q choosing alternative m_Q) may then be written succinctly as Prob[$\boldsymbol{y^* < 0}$]. To write this likelihood function, note that $\boldsymbol{y^*}$ has a mean vector $\boldsymbol{B}$ given by $[H_{11m_1}, H_{12m_1}, \ldots, H_{1Im_1}(i \neq m_1); H_{21m_2}, H_{22m_2}, \ldots, H_{2Im_2}(i \neq m_2); \ldots H_{Q1m_Q}, H_{Q2m_Q}, \ldots, H_{QIm_Q}(i \neq m_Q)]'$. To obtain the covariance matrix of $\boldsymbol{y^*}$, define $\boldsymbol{M}$ as a $[Q \times (I-1)] \times [QI]$ block diagonal matrix, with each block diagonal having $(I-1)$ rows and I columns corresponding to each individual q. This $(I-1) \times I$ matrix for individual q corresponds to an $(I-1)$ identity matrix with an extra column of -1's added as the m_qth column. For instance, consider the case of $I = 4$ and $Q = 2$. Let individual 1 be observed to choose alternative 2 and individual 2 be observed to choose alternative 1. Then $\boldsymbol{M}$ takes the form below.

$$\boldsymbol{M} = \left[\begin{array}{cccc|cccc} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{array}\right] \tag{43}$$

Finally, define the following additional matrices:

$$\begin{aligned} &\tilde{\boldsymbol{\Lambda}} = \boldsymbol{IDEN_Q} \otimes \boldsymbol{\Lambda} (QI \times QI) \text{matrix}, \\ &\tilde{\boldsymbol{\Omega}} = \tilde{\boldsymbol{x}} (\boldsymbol{IDEN_Q} \otimes \boldsymbol{\Omega}) \tilde{\boldsymbol{x}}' (QI \times QI) \text{matrix}, \\ &\tilde{\boldsymbol{F}} = \boldsymbol{S}\tilde{\boldsymbol{\Lambda}}\boldsymbol{S}', \text{a}[(QI) \times (QI)] \text{matrix, and} \\ &\boldsymbol{\Sigma} = \boldsymbol{M}(\tilde{\boldsymbol{F}} + \tilde{\boldsymbol{\Omega}})M', \text{an}[\{Q \times (I-1)\} \times \{Q \times (I-1)\}] \text{matrix} \end{aligned} \tag{44}$$

Then we can write $\boldsymbol{y^*} \sim MVN(\boldsymbol{B}, \boldsymbol{\Sigma})$, and the likelihood function of the sample is:

$$L_{ML}(\boldsymbol{\theta}) = \text{Prob}(\boldsymbol{y^*} < 0) = F_{Q \times (I-1)}(-\boldsymbol{B}, \boldsymbol{\Sigma}), \tag{45}$$

where $\boldsymbol{\theta}$ is the collection of parameters to be estimated: $\boldsymbol{\theta} = [\boldsymbol{b}'; [\text{Vech}(\boldsymbol{\Omega})]'; [\text{Vech}(\boldsymbol{\Lambda})]', \rho]'$, with Vech($\boldsymbol{\Omega}$) representing the column vector of upper triangle elements of $\boldsymbol{\Omega}$. $F_{Q \times (I-1)}$ is the multivariate cumulative normal distribution of $Q \times (I-1)$ dimensions. Of course, maximizing the above likelihood function requires the evaluation of a $Q \times (I-1)$ integral. However, the MACML approach may be again used here. The pairwise likelihood function is as follows:

$$L_{CML}(\theta) = \prod_{q=1}^{Q-1} \prod_{q'=q+1}^{Q} \text{Prob}(C_q = m_q, C_{q'} = m_{q'}) = \prod_{q=1}^{Q-1} \prod_{q'=q+1}^{Q} \text{Prob}[y^*_{qim_q} < 0 \forall i \neq m_q \quad \text{and} \quad y^*_{qim_{q'}} < 0 \forall i \neq m_{q'}] \tag{46}$$

The pairwise likelihood function above is similar to the case of a panel model, except that the pairs correspond to the choices of two different individuals at the same point in time. Each multivariate orthant probability above has a dimension equal to $(I-1) \times 2$, which can be computed using the approximation of Section 2.1. The variances and correlations in the bivariate and univariate cumulative normal distribution expressions in the approximation can be obtained as appropriate sub-matrices of $\boldsymbol{\Sigma}$. The positive-definiteness of $\boldsymbol{\Sigma}$ can be ensured by using a Cholesky-decomposition of the matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\Lambda}$, and estimating these Cholesky-decomposed parameters.

The one important problem with the CML expression in Eq. (46) is that it entails $Q(Q-1)/2$ pairs of multivariate probability computations, which can be very time consuming. Fortunately, in a spatial case where dependency drops quickly with inter-observation distance, the pairs formed from the closest observations provide much more information than pairs that are very far away. In fact, as demonstrated by Varin and Vidoni (2009), Varin and Czado (2010), Bhat et al. (2010a), and Apanasovich et al. (2008) in different empirical contexts, retaining all $Q(Q-1)/2$ pairs not only increases computational costs, but may also reduce estimator efficiency. Typically, in a spatial context, there appears to be an optimal distance for inclusion of observation pairs. This distance threshold may be set based on knowledge about the spatial process or based on testing the efficiency of estimators with varying values of the distance threshold. Using such a distance threshold effectively reduces the number of pairwise terms in the CML function.[11] Let the set of observational units within the threshold distance of unit q be $\tilde{M}_q$. Then, we propose dummy weights to include appropriate pairwise terms in the composite marginal likelihood function of Eq. (45). In particular, $\omega_{qq'} = 1$ if $k \in \tilde{M}_q$ and $\omega_{qq'} = 0$. The CML function may then be refined as follows:

[11] Note also that while we discuss the application of the MACML approach in the context of spatial dependence, the framework is extendable to include social and other forms of dependence too. This is because the weight matrix W that forms the basis for spatial dependence can be the basis for more general forms of dependence. In fact, W itself can be parameterized as a finite mixture of several weight matrices, each weight matrix being related to a specific covariate k: $\boldsymbol{W} = \sum_{k=1}^{K} \varphi_k \boldsymbol{W_k}$, where ϕ_k is the weight on the k^{th} covariate in determining dependency between individuals ($\sum_{k=1}^{K} \varphi_k = 1$), and $\boldsymbol{W_k}$ is a measure of distance between individuals on the k^{th} covariate.

$$L_{CML}(\theta) = \prod_{q=1}^{Q-1} \prod_{q'=q+1}^{Q} [\text{Prob}(C_q = m_q, C_{q'} = m_{q'})]^{\omega_{qq'}}, \tag{47}$$

One final issue. As discussed earlier in Section 1.2, the spatial property that the correlation fades over time implies that the CML estimator will retain the properties of being consistent and asymptotically normal. The "bread" matrix of $\boldsymbol{H}(\theta)$ in Eq. (14) can be estimated in a straightforward manner using the Hessian of the negative of $\log L_{CML}(\theta)$, evaluated at the CML estimate $\hat{\theta}$. This is because the information identity remains valid for each pairwise term forming the composite marginal likelihood. Thus, $\boldsymbol{H}(\theta)$ can be estimated as:

$$\hat{\boldsymbol{H}}(\hat{\theta}) = -\left[\sum_{q=1}^{Q-1} \sum_{q'=q+1}^{Q} \frac{\partial^2 \log L_{CML,qq'}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]_{\hat{\theta}_{CML}}, \tag{48}$$

where $L_{CML,qq'}(\boldsymbol{\theta}) = [P(C_q = m_q, C_{q'} = m_{q'})]^{\omega_{qq'}}$. However, the estimation of the "vegetable" matrix $\boldsymbol{J}(\boldsymbol{\theta})$ is more difficult; it cannot be evaluated at the convergent parameter values because the score function is zero at the convergence point. One cannot empirically estimate $\boldsymbol{J}(\boldsymbol{\theta})$ as the sampling variance of individual contributions to the composite score function (as is possible with panel data) because of the underlying spatial dependence in observations. But, since the spatial dependence fades with distance, we can use a windows resampling procedure (see Heagerty and Lumley, 2000) to estimate $\boldsymbol{J}(\boldsymbol{\theta})$. This procedure entails the construction of suitable overlapping subgroups of the original data that may be viewed as independent replicated observations. Then, $\boldsymbol{J}(\boldsymbol{\theta})$ may be estimated empirically. While there are several ways to implement this, one simple way we suggest is to overlay the spatial region under consideration with a square grid providing a total of $\tilde{Q}$ internal and external nodes. Then, select the observational unit closest to each of the $\tilde{Q}$ grid nodes to obtain $\tilde{Q}$ observational units from the original Q observational units ($\tilde{q} = 1, 2, 3, \ldots, \tilde{Q}$). Next, consider the set $R_{\tilde{q}}$ of observational units h such that $\omega_{\tilde{q}h} = 1$, and include $\tilde{q}$ as an observational unit in this set ($h = 1, 2, \ldots \tilde{H}_{\tilde{q}}$). Let $N_{\tilde{q}} = \sum_{h=1}^{\tilde{H}_{\tilde{q}}-1} \sum_{h'=h+1}^{\tilde{H}_{\tilde{q}}} \omega_{hh'}$ and $\tilde{W} = \sum_{q=1}^{Q-1} \sum_{q'=q+1}^{Q} \omega_{qq'}$. The $\tilde{Q}$ different sets of pairings may be considered as pseudo-independent, given that they are spaced out in a grid and the spatial dependence fades rapidly with distance.[12] An empirical estimate of $\boldsymbol{J}(\boldsymbol{\theta})$ may be obtained as follows:

$$\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}) = \frac{\tilde{W}}{\tilde{Q}} \left[\sum_{\tilde{q}=1}^{\tilde{Q}} \left[\frac{1}{N_{\tilde{Q}}} \left(\sum_{h=1}^{\tilde{H}_{\tilde{q}}-1} \sum_{h'=h+1}^{\tilde{H}_{\tilde{q}}} \omega_{hh'} \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{hh'}\right) \left(\sum_{h=1}^{\tilde{H}_{\tilde{q}}-1} \sum_{h'=h+1}^{\tilde{H}_{\tilde{q}}} \omega_{hh'} \frac{\partial}{\partial \boldsymbol{\theta}'} \log P_{hh'}\right)\right]\right]_{\hat{\theta}_{CML}}, \tag{49}$$

where $P_{hh'} = P(C_h = m_h, C_{h'} = m_{h'})$

5.2. The spatial lag model

The spatial lag model structure, with random coefficients, takes the following form:

$$U_{qi} = \rho \sum_{q'} w_{qq'} U_{q'i} + \boldsymbol{\beta}'_{\boldsymbol{q}} \boldsymbol{x}_{\boldsymbol{qi}} + \xi_{qi}; \xi_{qi} \sim N(0, \boldsymbol{\Lambda}), |\rho| < 1, \tag{50}$$

where ξ_{qi} is independent and identically distributed across q. Using the same notation as earlier, the equivalent of Eq. (41) in the spatial lag model is:

$$\boldsymbol{U} = \boldsymbol{S}[\boldsymbol{x}\boldsymbol{b} + \tilde{\boldsymbol{x}}\tilde{\boldsymbol{\beta}} + \boldsymbol{\xi}], \tag{51}$$

Define $V_{qi} = [\boldsymbol{S}\boldsymbol{x}\boldsymbol{b}]_d$ and $H_{qim_q} = V_{qi} - V_{qm_q}$, where m_q is the alternative chosen by individual q. Then stacking the utility differentials $y^*_{\tilde{q}im_q} (= U_{qi} - U_{qm_q},\ i \neq m_q)$ as in the spatial error model into an $Q \times (I-1)$ vector $\boldsymbol{y}^*$, and using the same notation as earlier, it is easy to see that $\boldsymbol{y}^*$ has a mean vector $\boldsymbol{B}$. The covariance matrix is slightly different from Eq. (44), though the expressions for $\tilde{\boldsymbol{\Lambda}}$ and $\tilde{\boldsymbol{\Omega}}$ are the same as in Eq. (44). Define the following:

$$\begin{aligned} &\tilde{\boldsymbol{F}} = \boldsymbol{S}(\tilde{\boldsymbol{\Lambda}} + \tilde{\boldsymbol{\Omega}})\boldsymbol{S}', \text{a}[(QI) \times (QI)]\text{matrix, and} \\ &\boldsymbol{\Sigma} = \boldsymbol{M}\tilde{\boldsymbol{F}}\boldsymbol{M}', \text{an}[\{Q \times (I-1)\} \times \{Q \times (I-1)\}]\text{matrix} \end{aligned} \tag{52}$$

Thus, $\boldsymbol{y}^* \sim MVN(\boldsymbol{B}, \boldsymbol{\Sigma})$, and the likelihood function is $L_{ML}(\boldsymbol{b}, \boldsymbol{\Omega}, \boldsymbol{\Lambda}, \rho) = \text{Prob}(\boldsymbol{y}^* < 0) = F_{Q\times(I-1)}(-\boldsymbol{B}, \boldsymbol{\Sigma})$. The pairwise likelihood function estimation approach is then similar to the spatial error case.

[12] Obviously, there needs to be a balance here between the number of sets of pairings K and the proximity of points. The smaller the value of K, the less proximal are the sets of observation units and more likely that the sets of observational pairings will be independent. However, at the same time, the value of K needs to be reasonable to obtain a good empirical estimate of J, since this empirical estimate is based on the score functions (computed at the convergent parameter values) across the K sets of observations.

6. Model selection

Procedures similar to those available with the maximum likelihood approach are also available for model selection with the MACML approach, based on results relevant to the composite marginal likelihood (CML) estimation approach. The statistical test for a single parameter may be pursued using the usual t-statistic based on the inverse of the Godambe information matrix. When the statistical test involves multiple parameters between two nested models, an appealing statistic, which is also similar to the likelihood ratio test in ordinary maximum likelihood estimation, is the composite likelihood ratio test (CLRT) statistic. Consider the null hypothesis H_0:$\boldsymbol{\tau} = \boldsymbol{\tau}_0$ against H_1:$\boldsymbol{\tau} \neq \boldsymbol{\tau}_0$, where τ is a subvector of $\boldsymbol{\theta}$ of dimension d; *i.e.*, $\boldsymbol{\theta} = (\boldsymbol{\tau}', \boldsymbol{\alpha}')'$. The statistic takes the familiar form shown below:

$$CLRT = 2[\log L_{CML}(\hat{\boldsymbol{\theta}}) - \log L_{CML}(\hat{\boldsymbol{\theta}}_0)], \tag{53}$$

where $\hat{\boldsymbol{\theta}}_0$ is the composite marginal likelihood estimator under the null hypothesis $(\boldsymbol{\tau}_0', \hat{\boldsymbol{\alpha}}_{\boldsymbol{CML}}'(\boldsymbol{\tau}_0))$. More informally speaking, $\hat{\boldsymbol{\theta}}$ is the CML estimator of the unrestricted model, and $\boldsymbol{\theta_0}$ is the CML estimator for the restricted model. The CLRT statistic does not have a standard chi-squared asymptotic distribution. This is because the CML function that is maximized does not correspond to the parametric model from which the data originates; rather, the CML may be viewed in this regard as a "misspecification" of the true likelihood function because of the independence assumption among the likelihood objects forming the CML function (see Kent, 1982, Section 3). To write the asymptotic distribution of the CLRT statistic, first define $[\boldsymbol{G}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]^{-1}$ and $[\boldsymbol{H}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]^{-1}$ as the $d \times d$ submatrices of $[\boldsymbol{G}(\boldsymbol{\theta})]^{-1}$ and $[\boldsymbol{H}(\boldsymbol{\theta})]^{-1}$, respectively, which correspond to the vector $\boldsymbol{\tau}$. Then, the CLRT has the following asymptotic distribution:

$$CLRT \sim \sum_{i=1}^{d} \lambda_i \tilde{W}_i^2, \tag{54}$$

where $\tilde{W}_i^2$ for $i = 1, 2, \ldots, d$ are independent χ_1^2 variates and $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \lambda_d$ are the eigenvalues of the matrix $[\boldsymbol{H}_{\boldsymbol{\tau}}(\boldsymbol{\theta})][\boldsymbol{G}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]^{-1}$ evaluated under the null hypothesis (this result may be obtained based on the (profile) likelihood ratio test for a mis-specified model; see Kent (1982), Theorem 3.1 and the proof therein). Unfortunately, the departure from the familiar asymptotic chi-squared distribution with d degrees of freedom for the traditional maximum likelihood procedure is annoying. Pace et al. (2011) have recently proposed a way out, indicating that the following adjusted *CLRT* statistic, *ADCLRT*, may be considered to be asymptotically chi-squared distributed with d degrees of freedom:

$$ADCLRT = \frac{[\boldsymbol{S}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]'[\boldsymbol{H}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]^{-1}[\boldsymbol{G}_{\boldsymbol{\tau}}(\boldsymbol{\theta})][\boldsymbol{H}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]^{-1}\boldsymbol{S}_{\boldsymbol{\tau}}(\boldsymbol{\theta})}{[\boldsymbol{S}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]'[\boldsymbol{H}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]^{-1}\boldsymbol{S}_{\boldsymbol{\tau}}(\boldsymbol{\theta})} \times CLRT, \tag{55}$$

where $\boldsymbol{S}_{\boldsymbol{\tau}}(\boldsymbol{\theta})$ is the $d \times 1$ submatrix of $S(\boldsymbol{\theta}) = \left(\frac{\partial \log L_{CML}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)$ corresponding to the vector $\boldsymbol{\tau}$, and all the matrices above are computed at $\hat{\boldsymbol{\theta}}_{\boldsymbol{0}}$. The denominator of the above expression is a quadratic approximation to *CLRT*, while the numerator is a score-type statistic with an asymptotic χ_d^2 null distribution. Thus, *ADCLRT* is also very close to being an asymptotic χ_d^2 distributed under the null. Alternatively, one can resort to parametric bootstrapping to obtain the precise distribution of the *CLRT* statistic for any null hypothesis situation. Such a bootstrapping procedure is rendered simple in the CML approach, and can be used to compute the p-value of the null hypothesis test. The procedure is as follows (see Varin and Czado, 2010):

1. Compute the observed *CLRT* value as in Eq. (53) from the estimation sample. Let the estimation sample be denoted as $\tilde{\boldsymbol{y}}_{\boldsymbol{obs}}$, and the observed *CLRT* value as $CLRT(\tilde{\boldsymbol{y}}_{\boldsymbol{obs}})$.
2. Generate C sample data sets $\tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2, \tilde{\boldsymbol{y}}_3, \ldots, \tilde{\boldsymbol{y}}_C$ using the CML convergent values under the null hypothesis
3. Compute the *CLRT* statistic of Eq. (53) for each generated data set, and label it as $CLRT(\tilde{\boldsymbol{y}}_c)$.
4. Calculate the p-value of the test using the following expression:

$$p = \frac{1 + \sum_{c=1}^{C} I\{CLRT(\tilde{\boldsymbol{y}}_{\boldsymbol{c}}) \geqslant CLRT(\tilde{\boldsymbol{y}}_{\boldsymbol{obs}})\}}{C+1}, \text{where } I\{A\} = 1 \text{ if } A \text{ is true.} \tag{56}$$

The above bootstrapping approach has been used for model testing between nested models in Varin and Czado (2010), Bhat et al. (2010a), and Ferdous et al. (2010).

When the null hypothesis entails model selection between two competing non-nested models, the composite likelihood information criterion (CLIC) introduced by Varin and Vidoni (2005) may be used. The CLIC takes the following form[13]:

$$\log L_{CML}^{*}(\hat{\boldsymbol{\theta}}) = \log \boldsymbol{L}_{CML}(\hat{\boldsymbol{\theta}}) - tr[\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{H}}(\hat{\boldsymbol{\theta}})^{-1}] \tag{57}$$

The model that provides a higher value of CLIC is preferred.

[13] This penalized log-composite likelihood is nothing but the generalization of the usual Akaike's Information Criterion (AIC). In fact, when the candidate model includes the true model in the usual maximum likelihood inference procedure, the information identity holds (*i.e.*, $\boldsymbol{H(\theta) = J(\theta)}$) and the CLIC in this case is exactly the AIC [$= \log L_{ML}(\hat{\boldsymbol{\theta}})$−(# of model parameters)].

7. Conclusions

It is typical to use simulation techniques to estimate multinomial probit-based models arising from general error covariance structures, random coefficients, panel effects, or spatial/social interaction effects. However, the accuracy of simulation techniques is known to degrade rapidly at medium-to-high dimensions, and the simulation noise increases substantially. This leads to convergence problems during estimation. In addition, such simulation-based approaches become impractical in terms of computation time as the number of dimensions of integration grows.

In this paper, we introduce a maximum approximate composite marginal likelihood (MACML) estimation approach for multinomial probit models. The MACML approach introduced here only involves univariate and bivariate cumulative normal distribution function evaluations. As importantly, the approach can be applied using simple optimization software for likelihood estimation.

The author is currently exploring methods to extend the MACML approach to allow a non-normal multivariate mixing distribution (such as non-normal random coefficients) over a multinomial probit kernel. Such non-normal mixing distributions may be accommodated by replacing them with an appropriate finite normal mixture distribution.

Acknowledgements

The author acknowledges the helpful comments of two anonymous reviewers on an earlier version of the paper. The author is grateful to Lisa Macias for her help in typesetting and formatting this document.

References

Albert, J.H., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88 (422), 669–679.
Anselin, L., 2003. Spatial externalities, spatial multipliers and spatial econometrics. International Regional Science Review 26 (2), 153–166.
Apanasovich, T.V., Ruppert, D., Lupton, J.R., Popovic, N., Turner, N.D., Chapkin, R.S., Carroll, R.J., 2008. Aberrant crypt foci and semiparametric modelling of correlated binary data. Biometrics 64 (2), 490–500.
Bartels, R., Fiebig, D.G., van Soest, A., 2006. Consumers and experts: an econometric analysis of the demand for water heaters. Empirical Economics 31 (2), 369–391.
Beron, K.J., Vijverberg, W.P.M., 2004. Probit in a spatial context: a Monte Carlo analysis. In: Anselin, L., Florax, R.J.G.M., Rey, S.J. (Eds.), Advances in Spatial Econometrics: Methodology, Tools and Applications. Springer-Verlag, Berlin, pp. 169–196.
Beron, K.J., Murdoch, J.C., Vijverberg, W.P.M., 2003. Why cooperate? Public goods, economic power, and the Montreal protocol. Review of Economics and Statistics 85 (2), 286–297.
Bhat, C.R., 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transportation Research Part B 35 (7), 677–693.
Bhat, C.R., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. Transportation Research Part B 37 (9), 837–855.
Bhat, C.R., 2011. The MACML estimation of the normally-mixed multinomial logit model. Technical paper, Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin. <http://www.caee.utexas.edu/prof/bhat/ABSTRACTS/MACML_Estim_Norm_MML_Model.pdf>.
Bhat, C.R., Castelar, S., 2002. A unified mixed logit framework for modeling revealed and stated preferences: formulation and application to congestion pricing analysis in the San Francisco Bay area. Transportation Research Part B 36 (7), 593–616.
Bhat, C.R., Guo, J.Y., 2004. A mixed spatially correlated logit model: formulation and application to residential choice modeling. Transportation Research Part B 38 (2), 147–168.
Bhat, C.R., Sardesai, R., 2006. The impact of stop-making and travel time reliability on commute mode choice. Transportation Research Part B 40 (9), 709–730.
Bhat, C.R., Sener, I.N., 2009. A copula-based closed-form binary logit choice model for accommodating spatial correlation across observational units. Journal of Geographical Systems 11 (3), 243–272.
Bhat, C.R., Sidharthan, R., forthcoming. A simulation evaluation of the maximum approximate composite marginal likelihood (MACML) estimator for mixed multinomial probit models. Transportation Research Part B.
Bhat, C.R., Eluru, N., Copperman, R.B., (2008). Flexible model structures for discrete choice analysis. In Handbook of Transport Modelling, 2nd edition, Chapter 5, Hensher, D.A., Button, K.J. (eds.), Elsevier Science, 75-104.
Bhat, C.R., Sener, I.N., Eluru, N., 2010a. A flexible spatially dependent discrete choice model: formulation and application to teenagers' weekday recreational activity participation. Transportation Research Part B 44 (8-9), 903–921.
Bhat, C.R., Varin, C., Ferdous, N., 2010b. A comparison of the maximum simulated likelihood and composite marginal likelihood estimation approaches in the context of the multivariate ordered response model. In: Greene, W.H., Hill, R.C. (Eds.), Advances in Econometrics: Maximum Simulated Likelihood Methods and Applications, vol. 26. Emerald Group Publishing Limited, pp. 65–106.
Bolduc, D., Fortin, B., Fournier, M., 1996. The effect of incentive policies on the practice location of doctors: a multinomial probit analysis. Journal of Labor Economics 14 (4), 703–732.
Caragea, P.C., Smith, R.L., 2007. Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. Journal of Multivariate Analysis 98 (7), 1417–1440.
Cox, D.R., Reid, N., 2004. A note on pseudolikelihood constructed from marginal densities. Biometrika 91 (3), 729–737.
Craig, P., 2008. A new reconstruction of multivariate normal orthant probabilities. Journal of the Royal Statistical Society: Series B 70 (1), 227–243.
Dube, J-P., Chintagunta, P., Petrin, A., Bronnenberg, B., Goettler, R., Seetharam, P.B., Sudhir, K., Tomadsen, R., Zhao, Y., 2002. Structural applications of the discrete choice model. Marketing Letters 13 (3), 207–220.
Engle, R.F., Shephard, N., Sheppard, K., 2007. Fitting and testing vast dimensional time-varying covariance models. Finance Working Papers, FIN-07-046, Stern School of Business, New York University.
Engler, D.A., Mohapatra, G., Louis, D.N., Betensky, R.A., 2006. A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. Biostatistics 7 (3), 399–421.
Feddag, M.-L., Bacci, S., 2009. Pairwise likelihood for the longitudinal mixed Rasch model. Computational Statistics and Data Analysis 53 (4), 1027–1037.
Ferdous, N., Eluru, N., Bhat, C.R., Meloni, I., 2010. A multivariate ordered-response model system for adults' weekday activity episode generation by activity purpose and social context. Transportation Research Part B 44 (8-9), 922–943.
Fiebig, D.C., Keane, M.P., Louviere, J., Wasi, N., 2010. The generalized multinomial logit model: accounting for scale and coefficient heterogeneity. Marketing Science 29 (3), 393–421.

Fleming, M.M., 2004. Techniques for estimating spatially dependent discrete choice models. In: Anselin, L., Florax, R.J.G.M., Rey, S.J. (Eds.), Advances in Spatial Econometrics: Methodology, Tools and Applications. Springer-Verlag, Berlin, pp. 145–168.
Franzese, R.J., Hays, J.C., 2008. Empirical models of spatial interdependence. In: Box-Steffensmeier, J.M., Brady, H.E., Collier, D. (Eds.), The Oxford Handbook of Political Methodology. Oxford University Press, Oxford, pp. 570–604.
Gassmann, H.I., Deák, I., Szántai, T., 2002. Computing multivariate normal probabilities: a new look. Journal of Computational and Graphical Statistics 11 (4), 920–949.
Genz, A., Bretz, F., 1999. Numerical computation of multivariate *t*-probabilities with application to power calculation of multiple contrasts. Journal of Statistical Computation and Simulation 63 (4), 361–378.
Godambe, V.P., 1960. An optimum property of regular maximum likelihood estimation. The Annals of Mathematical Statistics 31 (4), 1208–1211.
Heagerty, P.J., Lumley, T., 2000. Window subsampling of estimating functions with application to regression models. Journal of the American Statistical Association 95 (449), 197–211.
Heckman, J.J., Singer, B., 1986. Econometric analysis of longitudinal data. In: Griliches, Z., Intriligator, M.D. (Eds.), The Handbook of Econometrics, vol. 3. North-Holland, Amsterdam, pp. 1689–1766.
Heiss, F., 2010. The panel probit model: adaptive integration on sparse grids. In: Greene, W.H., Hill, R.C. (Eds.), Advances in Econometrics: Maximum Simulated Likelihood Methods and Applications, vol. 26. Emerald Group Publishing Limited, pp. 41–64.
Heiss, F., Winschel, V., 2008. Likelihood approximation by numerical integration on sparse grids. Journal of Econometrics 144 (1), 62–80.
Hess, S., Rose, J.M., 2009. Allowing for intra-respondent variations in coefficients estimated on repeated choice data. Transportation Research Part B 43 (6), 708–719.
Hjort, N.L., Varin, C., 2008. ML, PL, QL in Markov chain models. Scandinavian Journal of Statistics 35 (1), 64–82.
Huguenin, J., Pelgrin, F., Holly, A., 2009. Estimation of multivariate probit models by exact maximum likelihood. Working Paper 0902, University of Lausanne, Institute of Health Economics and Management (IEMS), Lausanne, Switzerland.
Joe, H., 1995. Approximations to multivariate normal rectangle probabilities based on conditional expectations. Journal of the American Statistical Association 90 (431), 957–964.
Joe, H., 2008. Accuracy of Laplace approximation for discrete response mixed models. Computational Statistics and Data Analysis 52 (12), 5066–5074.
Joe, H., Lee, Y., 2009. On weighting of bivariate margins in pairwise likelihood. Journal of Multivariate Analysis 100 (4), 670–685.
Kent, J.T., 1982. Robust properties of likelihood ratio tests. Biometrika 69 (1), 19–27.
Kuk, A.Y.C., Nott, D.J., 2000. A pairwise likelihood approach to analyzing correlated binary data. Statistics and Probability Letters 47 (4), 329–335.
Le Cessie, S., Van Houwelingen, J.C., 1994. Logistic regression for correlated binary data. Applied Statistics 43 (1), 95–108.
Lele, S.R., 2006. Sampling variability and estimates of density dependence: a composite-likelihood approach. Ecology 87 (1), 189–202.
LeSage, J.P., 2000. Bayesian estimation of limited dependent variable spatial autoregressive models. Geographical Analysis 32 (1), 19–35.
Li, Z., Hensher, D.A., Rose, J.M., 2010. Willingness to pay for reliability in passenger transport: a review and some new empirical evidence. Transportation Research Part E 46 (3), 384–403.
Lindsay, B.G., 1988. Composite likelihood methods. Contemporary Mathematics 80, 221–239.
Luce, R., Suppes, P., 1965. Preference, utility and subjective probability. In: Luce, R., Bush, R., Galanter, E. (Eds.), Handbook of Mathematical Probability, vol. 3. Wiley, New York.
Mardia, K.V., Kent, J.T., Hughes, G., Taylor, C.C., 2009. Maximum likelihood estimation using composite likelihoods for closed exponential families. Biometrika 96 (4), 975–982.
McCulloch, R.E., Rossi, P.E., 2000. Bayesian analysis of the multinomial probit model. In: Mariano, R., Schuermann, T., Weeks, M.J. (Eds.), Simulation-Based Inference in Econometrics. Cambridge University Press, New York, pp. 158–178.
McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.
McFadden, D., 1978. Modeling the choice of residential location. Transportation Research Record 672, 72–77.
McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. Journal of Applied Econometrics 15 (5), 447–470.
McMillen, D.P., 1995. Spatial effects in probit models: a Monte Carlo investigation. In: Anselin, L., Florax, R.J.G.M. (Eds.), New Directions in Spatial Econometrics. Springer-Verlag, Berlin, pp. 189–228.
Miyamoto, K., Vichiensan, V., Shimomura, N., Páez, A., 2004. Discrete choice model with structuralized spatial effects for location analysis. Transportation Research Record 1898, 183–190.
Molenberghs, G., Verbeke, G., 2005. Models for Discrete Longitudinal Data. Springer Series in Statistics. Springer Science + Business Media, Inc., New York.
Pace, L., Salvan, A., Sartori, N., 2011. Adjusting composite likelihood ratio statistics. Statistica Sinica 21 (1), 129–148.
Pinkse, J., Slade, M.E., 1998. Contracting in space: an application of spatial statistics to discrete-choice models. Journal of Econometrics 85 (1), 125–154.
Renard, D., Molenberghs, G., Geys, H., 2004. A pairwise likelihood approach to estimation in multilevel probit models. Computational Statistics & Data Analysis 44 (4), 649–667.
Ruud, P.A., 2007. Estimating mixtures of discrete choice model. Technical Paper, University of California, Berkeley.
Small, K.A., Winston, C., Yan, J., 2005. Uncovering the distribution of motorists' preferences for travel time and reliability. Econometrica 73 (4), 1367–1382.
Sener, I.N., Pendyala, R.M., Bhat, C.R., 2011. Accommodating spatial correlation across choice alternatives in discrete choice models: an application to modeling residential location choice behavior. Journal of Transport Geography 19 (2), 294–303.
Solow, A.R., 1990. A method for approximating multivariate normal orthant probabilities. Journal of Statistical Computation and Simulation 37 (3–4), 225–229.
Switzer, P., 1977. Estimation of spatial distribution from point sources with application to air pollution measurement. Bulletin of the International Statistical Institute 47 (2), 123–137.
Train, K., 2009. Discrete Choice Methods with Simulation, second ed. Cambridge University Press, Cambridge.
Varin, C., 2008. On composite marginal likelihoods. AStA Advances in Statistical Analysis 92 (1), 1–28.
Varin, C., Czado, C., 2010. A mixed autoregressive probit model for ordinal longitudinal data. Biostatistics 11 (1), 127–138.
Varin, C., Vidoni, P., 2005. A note on composite likelihood inference and model selection. Biometrika 92 (3), 519–528.
Varin, C., Vidoni, P., 2006. Pairwise likelihood inference for ordinal categorical time series. Computational Statistics and Data Analysis 51 (4), 2365–2373.
Varin, C., Vidoni, P., 2009. Pairwise likelihood inference for general state space models. Econometric Reviews 28 (1-3), 170–185.
Varin, C., Reid, N., Firth, D., forthcoming. An overview of composite marginal likelihoods. Statistica Sinica.
Zhao, Y., Joe, H., 2005. Composite likelihood estimation in multivariate data analysis. The Canadian Journal of Statistics 33 (3), 335–356.