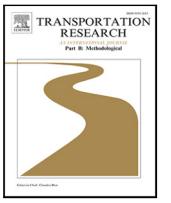
Transportation Research Part B 181 (2024) 102890

Contents lists available at ScienceDirect

Transportation Research Part B

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



Weighting strategies for pairwise composite marginal likelihood estimation in case of unbalanced panels and unaccounted autoregressive structure of the errors

Sebastian Büscher *, Dietmar Bauer

Bielefeld University, Econometrics, Universitätsstraße 25, Bielefeld, 33615, Germany

ARTICLE INFO

MSC:
62J12
62F10
62F12

Keywords:
Probit modelling
Composite marginal likelihood
Weighting
Unbalanced panel data
Unaccounted autoregressive error process
Efficiency

ABSTRACT

Composite Marginal Likelihood (CML) estimation and its advancements are popular ways to reduce the computational burden involved in the estimation of Multinomial Probit (MNP) models. CMLs use the product of marginal likelihoods of decision makers instead of the complete joint likelihood, reducing the numerical load. This allows for the estimation of models for larger and more complex data sets. The definition of the CML involves power weights on the marginal likelihoods that influence the statistical properties of the estimator. In this paper, we discuss how to effectively use the power weights in the cases of (1) unbalanced panel settings, where the weights help to reduce the variance of the estimator, and (2) unaccounted autoregressive structure of the errors, where the weights help to reduce the asymptotic bias of the estimator due to misspecification.

1. Introduction

The mobility behaviour of persons depends heavily on a great number of choices ranging from the choice of the origin and destination as well as the chosen mode for a particular trip to the acquisition of vehicles. In many cases, these choices involve selecting one option from a finite number of alternatives. These decisions depend on the characteristics of the various alternatives as well as the preferences and characteristics of the deciders. In order to learn about these preferences one typically collects data on repeated choice situations.

When modelling (potentially repeated) discrete choices, the two dominant model families are the Multinomial Logit (MNL) and the multinomial probit models (MNP; see, for example, Train, 2009). Both can be formulated as Random Utility Models (RUMs) as

$$y^*_{n,t,j} = X'_{n,t,j}\beta_{n,j} + \varepsilon_{n,t,j},$$

where $y^*_{n,t,j} \in \mathbb{R}$ denotes the unobserved utility the individual $n \in \{1,\dots,N\}$ attributes at time $t \in \{1,\dots,T\}$ to choice alternative $j \in \{1,\dots,J\}$, $X_{n,t,j} \in \mathbb{R}^{R\times 1}$ denotes the set of regressors, and $\beta_{n,j} \in \mathbb{R}^{R\times 1}$ denotes a parameter vector, which can potentially be individual specific. Consequently, $R \in \mathbb{N}$ denotes the number of regressors and, hence, also the number of parameters in $\beta_{n,j}$. Assuming that $\beta_{n,j}$ is drawn from a parametric family of distributions $f_j(\cdot;\theta_j)$ (for parameter vector θ_j) independent of the regressors $X_{n,t,k}$ and the noise terms $\varepsilon_{n,t,k}$, the model allows the representation of unobserved taste heterogeneity (compare Train, 2009, Chapter 6, for the logit case).

* Corresponding author.
E-mail addresses: sebastian.buescher@uni-bielefeld.de (S. Büscher), dietmar.bauer@uni-bielefeld.de (D. Bauer).

https://doi.org/10.1016/j.trb.2024.102890
Received 10 March 2023; Received in revised form 16 October 2023; Accepted 18 January 2024
Available online 30 January 2024

The alternative with the highest random utility is then assumed to be chosen; therefore, for observation $y_{n,t} = k_{n,t}$, we assume $y^*_{n,t,k_{n,t}} \geq y^*_{n,t,j}, j = 1, \dots, J$. By assuming different distributions for the random error terms $\varepsilon_{n,t,j}$, we obtain the different choice model families, both of which are typically estimated using likelihood maximisation. Between these two model families, the MNP offers more modelling flexibility whilst suffering from increased computational costs due to the need to evaluate multivariate normal cumulative distribution functions (MVNCDFs) (see, for example, Train, 2009) to evaluate the choice probabilities.

Moderately sized MNP models were estimated using Maximum Simulated Likelihood (MSL) methods, which show computational difficulties with a growing number of choice alternatives and choice occasions, as demonstrated in Bhat (2014). Whilst developments in Quasi-Monte Carlo (QMC) methods (see Bhat, 2003; Hess et al., 2006; Dick et al., 2017), Sparse Grid Quadrature (SGQ) methods (see Heiss and Winschel, 2008), and, most recently, Designed Quadrature (DQ) methods (see Ryu and Boyd, 2015; Keshavarzzadeh et al., 2018; Bansal et al., 2022) have been shown to reduce computation time significantly in MSL estimation, they continue to be subject to the curse of dimensionality.

To mitigate these difficulties an alternative approach is to use a Composite Marginal Likelihood (CML) (suggested in Varin, 2008). In this approach, the product of the probabilities of subsets of the choices of one individual is used instead of the joint probability. These subsets are called margins and the related probabilities marginal probabilities. In most cases a pairwise CML is used, where the margins consist of pairs of choices. In this approach, the logarithm of the joint probability of the T_n observations $y_{n,1} = k_{n,1}, \dots, y_{n,T_n} = k_{n,T_n}$ from one individual n is replaced in the criterion function by the logarithm of the product of marginal likelihoods of pairs of observations. The difference becomes apparent when comparing the probit log-likelihood function to the resulting CML quasi-log-likelihood function

$$ll(y, X; \theta) = \sum_{n=1}^{N} \log \left(\int \mathbb{P}(y_{n,1} = k_{n,1}, \dots, y_{n,T_n} = k_{n,T_n}, X_n; \tilde{\theta}) f_\theta(\tilde{\theta}) d\tilde{\theta} \right),$$

$$ll_{\text{CML}}(y, X; \theta) = \sum_{n=1}^{N} \sum_{a=1}^{T_n - 1} \sum_{b=a+1}^{T_n} \log \underbrace{\left(\int \mathbb{P}(y_{n,a} = k_{n,a},\ y_{n,b} = k_{n,b}, X_n; \tilde{\theta}) f_\theta(\tilde{\theta}) d\tilde{\theta} \right)}_{\mathbb{P}(y_{n,a}=k_{n,a},\ y_{n,b}=k_{n,b}, X_n; \theta)},$$

where θ denotes the vector of all estimable parameters of the model, considering a mixed effects model with f_θ denoting the Probability Density Function (PDF) of the mixing distribution. In the following, probabilities indexed using θ refer to the mixed choice probabilities (neglecting the dependence on $X_{n,t,j}$ in the notation). Using this quasi-likelihood $ll_{\text{CML}}(y, X; \theta)$ reduces the dimension of the MVNCDFs required for calculating the choice probabilities and can thus significantly reduce the computational burden. Usage of the CMLs and advancements of these, such as Maximum Approximate Composite Marginal Likelihood (MACML) (see Bhat, 2011), have thus become popular (as evidenced by a 'Web of Science' search returning a total of 3717 papers at the time of writing, with an almost steady increase from 34 papers in 1999 to 301 in 2022, with a peak of 354 in 2021, see Clarivate, 2022).

The CML formulation allows for the assignment of different power weights to each bivariate margin. This results in the weighted CML quasi-log-likelihood function

$$ll_{\text{CML}}(y, X; \theta, W) = \sum_{n=1}^{N} \sum_{a=1}^{T_n - 1} \sum_{b=a+1}^{T_n} w_{n,a,b} \log \mathbb{P}(y_{n,a} = k_{n,a}, y_{n,b} = k_{n,b}, X_n; \theta), \tag{1}$$

where W denotes the collection of weights $w_{n,a,b}$, $n = 1, \dots, N$, $a = 1, \dots, T_n - 1$, and $b = a + 1, \dots, T_n$.

The chosen weights affect the statistical properties of the CML estimator (compare Lindsay et al., 2011) and, hence, have the potential to improve the statistical properties of the estimator compared to the unweighted CML estimator with $w_{n,a,b} \equiv 1$. A discussion of the impact of the weights can be found in Section 4.2 of the survey (Varin et al., 2011). More general CML formulations combine marginal and bivariate margins in the estimation, which can be traced back to early papers in psychology such as Muthén (1984). A more recent survey can be found in Cox and Reid (2004). In the transportation literature the pairwise CML formulation as in Eq. (1) is dominant. Beside the unweighted CML estimator also adjacent pairwise formulations are popular, wherein bivariate margins for pairs of adjacent observations are included and all other margins excluded. This reduces the computational cost a lot at the price of reducing the information (and thus increasing the estimation variance) in the CML.

It is not hard to construct synthetic examples (see Appendix A) where the variance can be approximately halved (compared to using the unweighted CML) by including weights.[1] Furthermore, by setting a portion of the weights to zero, the number of bivariate probabilities to be computed can significantly be reduced, which further reduces the computational burden.

Thus, by adopting various weighting strategies, the impact of different pairs of observations on the estimation can be tuned. In this paper, we examine how to optimise some of the most commonly used weighting strategies.

This tuning of the weights first will be investigated under the assumption of correctly specified models. However, in panel data situations, oftentimes some choice occasions arise in spurts, leading to small time differences for some occasions, whereas others arise well separated in time. This is the case, for example, in panel waves of mobility household surveys, such as the German mobility panel (Zumkeller et al., 1999; Zumkeller and Chlond, 2009), wherein the daily mobility of the participants is surveyed for one week every year.

In these cases, it appears plausible that the error terms $\varepsilon_{n,t,j}$ are sampled from an underlying continuous-time stationary process with non-zero autocorrelation being present for short time lags, but correlation diminishing over time. Such dependencies typically

[1] These examples not necessarily are typical for real world applications where the gain can be more modest.

are not modelled explicitly. But by ignoring these correlations, the model is misspecified. Misspecifications typically lead to an asymptotic bias, resulting in an inconsistent estimator. In such cases, we investigate whether weighting schemes can be used to reduce the expected asymptotic bias despite the misspecification, providing a diagnostic tool to check for temporal correlation of error terms.

Whereas there has been a fair amount of research into weighting schemes in the context of CML estimation in the past (see, for example, Bhat (2014) for an overview of this topic, Pedeli and Varin (2020) for the use of weighted pairwise CML in latent autoregressive models, Crastes et al. (2020) for the use on autoregressive ordered probit models), this paper aims to contribute to the existing literature by examining the effects of weighting strategies in two specific cases:

First, in Section 2, we introduce a two-step optimal group-weight CML estimator to optimise (in a specific sense) the estimator variance in the case of unbalanced panel data. This is important because, in contrast to the Maximum Likelihood (ML) probit estimator, the CML estimator is not efficient in general (see Varin, 2008).

Second, in Section 3, we discuss weighting strategies to reduce the asymptotic bias due to misspecification of the model in case of unaccounted autocorrelation of the error terms. In correctly specified models, the use of close-by observations in CML pairs is favoured (see, for example, Crastes et al., 2020; Joe and Lee, 2009; Varin and Vidoni, 2005), whereas the case of misspecification due to unaccounted autocorrelation is, to the best of our knowledge, still to be investigated and benefits from an opposing approach.

In both cases, we first present the theory of the proposed weighting approaches, including asymptotic properties (Sections 2.1 and 3.2), before demonstrating the properties of the proposed weighting schemes in finite-sample simulation studies (Sections 2.2 and 3.3). Proofs of the theorems introduced in the paper can be found in Appendix B.

2. Weighting in case of unbalanced panel data

When using a *full-pairwise* weighting strategy $w_{n,a,b} \equiv 1$, each of the T_n observations of individual n is part of (T_n-1) CML margins, and there is a total of $(T_n-1)T_n/2$ pairs of choice occasions for this individual. Consequently, in an unbalanced panel setting, the number of margins, and thus also the total sum of weights, for one individual scales quadratically with the number of observations. Hence, individuals with more observations have a disproportionately larger influence on the CML function. A priori, it is not clear if this weighting scheme is optimal.

To correct for a possible misrepresentation due to an unequal number of choice occasions, we may group individuals n into clusters according to their number of observations $T_n = s$ and use cluster weights $w_{n,a,b} = w_s$ for $T_n = s$.

Asymptotically, the choice of the weights w_s typically does not affect the consistency of the estimators but will affect the accuracy of the estimator. Cessie and Houwelingen (1994), Joe and Lee (2009) and Kuk and Nott (2000) investigated the effects on the asymptotic variance of the choice of w_s in the context of binary choice occasions and proposed different weighting strategies to account for the imbalance in weights.

Cessie and Houwelingen (1994) used $w_s = (s-1)^{-1}$ for clustered binary data, with $s \in \{2, \dots, \overline{S}\}$, arguing that this weighting strategy results in each cluster's contribution to the likelihood being relative to its size and that for independent observations within the clusters, the CML is equal to the full likelihood.

Kuk and Nott (2000) argue for using the same weights as Cessie and Houwelingen (1994) to estimate the parameters of the utility function but to use the unweighted CML function to estimate the correlation between observations. They argue that, for the estimation of the correlation between observations within one cluster, the number of pairs is relevant instead of the number of observations.

Joe and Lee (2009) derived optimal power weights minimising the variance of the parameter estimators for an exemplary binary choice model $y_n^* = (y_{n,1,1}^*, \dots, y_{n,s,1}^*) \sim \mathcal{N}(\mu, \Sigma_s(\sigma^2, \rho))$, with μ denoting the mean utility of the nonreference alternative, σ^2 the variance of the errors and ρ the correlation between the errors at different choice occasions. For the estimation of μ with known correlation ρ between observations within clusters, they find the optimal weights to be $w_s = (s-1)^{-1}[1+(s-1)\rho]^{-1}$. For uncorrelated observations, this coincides with the weights used by Cessie and Houwelingen (1994) and Kuk and Nott (2000). For both uncorrelated ($\rho = 0$) and perfectly correlated ($\rho = 1$) observations within clusters, these weights result in the CML being equivalent to the full likelihood. They study these weights with different choices for ρ in $w_s = (s-1)^{-1}[1+(s-1)\rho]^{-1}$ in the context of multivariate clustered exchangeable probit models (which lead to a similar correlation structure as the mixed MNP models in the panel case) by performing asymptotic relative efficiency analysis in a simulation study with different correlation coefficients ρ and mixtures of numbers of observations s and find that for unknown correlation ρ using a midway option with $w_s = (s-1)^{-1}[1+1/2(s-1)]^{-1}$, whilst not being the best choice when the correlation is known, performs well over a range of moderate to strong correlation. They, hence, suggest to use $w_s = (s-1)^{-1}[1+1/2(s-1)]^{-1}$ in general for clustered data.

2.1. Two-step optimal group-weight CML Estimator

In this section we will introduce a new two-step optimal group-weight CML estimator, derived from asymptotic properties of the variance of the weighted CML estimator. In order to do so, we will first introduce some assumptions on the initial weights (used in the first step) for the CML estimator (Assumption 1) and on the data generating process (Assumption 2), both of which are reasonable and can be assumed to hold for a wide range of practical model applications. Under these assumptions we introduce and prove two theorems, Theorem 1 on properties of the asymptotic variance of the CML estimator with unbalanced panel data, and Theorem 2 on theoretical optimal group-specific weights for the CML estimator. Based on these theorems, we propose two versions of a new two-step optimal group-weight CML estimator in Algorithm 1 and Algorithm 2.

Following Joe and Lee (2009), also in our setting, the asymptotic variance of the estimator can be calculated as a mixture of variances for deciders with an identical number of choice occasions. We use this to derive variance optimal (in a certain sense) weights depending on the number of choice occasions of an individual.

To calculate the optimal weights, we use a two-step approach commencing from an initial weighting scheme $\mathring{w}_{n,a,b}$ that is subsequently adjusted, using group weights in order to optimise the asymptotic variance.

For the initial weights, several options exist, including those listed in the following assumption:

Assumption 1 (*Initial Weights*)**.** The initial weights for N deciders, where the nth decider faces T_n choice occasions, are chosen according to one of the following five schemes:

(I) $\mathring{w}_{n,a,b} = f(a,b)$ for some bounded positive function $f : \mathbb{N}^2 \to [\underline{w}, \overline{w}], 0 \le \underline{w} \le \overline{w} < \infty$.

(II) $\mathring{w}_{n,a,b} = \mathring{w}_{T_n} \in [\underline{w}, \overline{w}], 0 < \underline{w} \le \overline{w} < \infty$ (*groupwise* CML weights).

(III) $\mathring{w}_{n,a,b} \in \{0,1\}$ chosen randomly independent of all other variables, independent, identically distributed (iid) over deciders, such that

$$\sum_{a=1}^{T_n-1} \sum_{b=a+1}^{T_n} \mathring{w}_{n,a,b} = \mathring{w}_{T_n}$$

(selecting $\mathring{w}_{T_n}$ random pairs from within $T_n(T_n-1)/2$ possible pairs).

(IV) Stratification weights $\mathring{w}_{n,a,b} = \mathring{w}_n$ drawn iid over deciders from some underlying distribution supported on $[\underline{w}, \overline{w}], 0 < \underline{w} \le \overline{w} < \infty$.

(V) A combination of (I)–(IV).

These weights include all commonly used weighting schemes: The full pairwise case $\mathring{w}_{n,a,b} \equiv 1$ can be seen as a special case of any of the schemes above and *adjacent pairwise weighting* is a special case of (I), where $f(a,b) = \mathbb{I}(b = a+1)$ (the indicator function that b equals $a+1$). The proposals of Joe and Lee (2009) correspond to groupwise CML weights (II). (IV) allows for stratification weights, which are typically uniformly bounded from above and below. The various bounds are needed to ensure that (a) each decider has an impact on the criterion function and (b) no decider dominates all others.

In the following, we will use the weights $w_{n,a,b} = w_{T_n} \mathring{w}_{n,a,b}$, where $\mathring{w}_{n,a,b}$ denotes an arbitrary initial weighting scheme from the list above. Subsequently, we use the groupwise weights $w_s, s = 2, \dots, \overline{S}$ to optimise certain aspects of the asymptotic variance.

In the subsequent derivation of these groupwise weights, the total initial weight per decider $C_{n,T_n}(\mathring{W}) = \sum_{a=1}^{T_n} \sum_{b=a+1}^{T_n} \mathring{w}_{n,a,b}$ will come into play. In many examples, this is a function of T_n only. However, for stratification weights, it varies over different deciders facing $T_n = s$ choice occasions. Thus, let $C_{N,s}(\mathring{W}) := {N_s}^{-1} \sum_{n:T_n=s} C_{n,T_n}$ denote the average total weight of all deciders with s choices. In all cases fulfilling Assumption 1, we have $C_{N,s}(\mathring{W}) \to C_s(\mathring{W})$ almost surely.

Standard theory then implies that choosing positive weights under appropriate assumptions (independent of the sample size and not depending on θ; these assumptions are satisfied for all choices discussed above) implies consistent estimators $\hat{\theta}(W) \to \theta_0$ (see Lindsay et al., 2011) and asymptotic normality derived from mean value theorems (see Cox and Reid, 2004; Joe and Lee, 2009), specifically

$$\sqrt{N}(\hat{\theta}(W) - \theta_0) = -(\partial_\theta^2 ll_{CML}(y, X; \bar{\theta}, W)/N)^{-1} (\partial_\theta ll_{CML}(y, X; \theta_0, W)/\sqrt{N}),$$

where $\bar{\theta}$ denotes an intermediate value between $\hat{\theta}(W)$ and θ_0.

Assumption 2 (*Data Generating Process*)**.** The data set $(y_n, X_n), n = 1, \dots, N$ is generated by the following mechanism:

1. A number T_n of choice occasions are drawn from a discrete random distribution supported in $\{2, 3, \dots, \overline{S}\}$.
2. For each decider facing T_n choice occasions, a matrix $X_n = [X_{n,1}, \dots, X_{n,T_n}] \in \mathbb{R}^{JR \times T_n}$ of regressors is chosen iid over deciders such that for each pair of choice occasions (a,b), the distribution of the matrix $[X_{n,a}, X_{n,b}]$ is identical. Furthermore, $\|X_{n,a}\| \le M$ (uniform norm bound).
3. For given T_n and X_n, the vector of choices $y_n = [y_{n,1}, \dots, y_{n,T_n}]' \in \mathcal{J}^{T_n}, \mathcal{J} = \{1, \dots, J\}$ is chosen according to the mixed MNP model corresponding to parameter vector $\theta_0 \in \mathbb{R}^d$ for appropriate integer d.

Note that the assumptions on the regressor variables imply iid sampling over deciders but allow for dependence for the choice occasions faced by one decider. The ordering of the choice occasions must be of no relevance, however, as the distribution of $[X_{n,a}, X_{n,b}]$ for all pairs of choices needs to be identical. This allows for decider-specific regressors that do not vary over choice occasions. Temporal dependence, as would be achieved from systematically posing choices depending on previous answers, is excluded by the assumptions.

A second remark relates to the mechanism for drawing the number of choice occasions: We assume that this choice is performed independently of the regressors or the choice process. This excludes study designs that decide on the number of choice tasks based on either regressors or choices taken so far.

One implication of this data generation is that the number N_s of deciders facing s choice occasions is random, but the fraction N_s/N converges to $f_s \ge 0$, the corresponding probability. The result also holds if $f_s = 0$ for some s. In this case, $\hat{V}_s$ (defined below) is not estimated consistently, whilst $\hat{H}_0$ (see below) still is consistent.

Under this Data Generating Process (DGP), the key to understanding the asymptotic properties then lies in the following representation of the CML, the corresponding score, and the Hessian matrix when using the weighting $w_{n,a,b} = w_{T_n}\mathring{w}_{n,a,b}$:

$$\begin{aligned}
ll_{CML}(y, X; \theta_0, W) &= \sum_{s=2}^{\overline{S}} \sum_{n:T_n=s} ll_{CML}(y_n, X_n; \theta_0, W) \\
&= \sum_{s=2}^{\overline{S}} w_s \sum_{n:T_n=s} \left(\sum_{a=1}^{s-1} \sum_{b=a+1}^{s} \mathring{w}_{n,a,b} \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \theta_0) \right), \\
\partial_\theta ll_{CML}(y, X; \theta_0, W) &= \sum_{s=2}^{\overline{S}} w_s \sum_{n:T_n=s} \underbrace{\left(\sum_{a=1}^{s-1} \sum_{b=a+1}^{s} \mathring{w}_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \theta_0) \right)}_{:=g_n(\mathring{W})} \\
&= \sum_{s=2}^{\overline{S}} w_s \sum_{n:T_n=s} g_n(\mathring{W}), \\
\partial_\theta^2 ll_{CML}(y, X; \theta_0, W) &= \sum_{s=2}^{\overline{S}} w_s \sum_{n:T_n=s} \partial_\theta g_n(\mathring{W}),
\end{aligned}$$

where the next to last equation defines $g_n(\mathring{W})$. Here, we assume that $2 \le T_n \le \overline{S}, n = 1, \dots, N$, such that every decider faces at least two choice occasions and at most $\overline{S}$. Note that

$$\sum_{n:T_n=s} ll_{CML}(y_n, X_n; \theta_0, W) = w_s \sum_{n:T_n=s} ll_{CML}(y_n, X_n; \theta_0, \mathring{W})$$

defines a CML for the balanced subset $\mathcal{S}_s := \{n : T_n = s\}$ whose optimising argument depends on the initial weights $\mathring{W}$ rather than the groupwise weights w_s. Therefore, under standard assumptions (see below), the usual asymptotic properties hold such that $g_n(\mathring{W}), n \in \mathcal{S}_s$ constitutes an iid sequence with $\mathbb{E}g_n(\mathring{W}) = 0$ and variance $V_s(\mathring{W})$. Further independent sampling implies independence for different values of s.

With this notation, we obtain the following result:

Theorem 1 (*Asymptotic Variance*). *Let the data be generated according to Assumption 2 with parameter vector θ_0 and let $\hat{\theta}$ be the CML estimator maximising the weighted CML function (1) using the weights $w_{n,a,b} = w_{T_n}\mathring{w}_{n,a,b}$, where the initial weights $\mathring{w}_{n,a,b}$ adhere to Assumption 1, where $C_s(\mathring{W}) > 0, s = 2, \dots, \overline{S}$.*

Further, let $w_s \ge 0$ denote group-specific weights according to the number of observations $T_n = s,\ s \in \{2, \dots, \overline{S}\}$ of decider n, such that $\sum_{s=2}^{\overline{S}} f_s w_s C_s(\mathring{W}) = 1$. Then the following hold:

(I) $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V_\theta(W_s))$, *where the asymptotic variance–covariance matrix $V_\theta(W_s)$ for a given vector $W_s = (w_2, \dots, w_{\overline{S}})'$ of group-specific weights has the form*

$$V_\theta(W_s) = \sum_{s=2}^{\overline{S}} f_s w_s^2 H_0^{-1} V_s H_0^{-1}, \tag{2}$$

with

$$V_s = \mathbb{E}\ g_n(\mathring{W}) g_n(\mathring{W})', \quad H_0 = \mathbb{E}\ \partial_\theta^2 \log \mathbb{P}(y_{n,1}, y_{n,2}, X_n; \theta_0). \tag{3}$$

(II) H_0 can be estimated consistently as

$$\hat{H}_0 = \frac{1}{\sum_{n,a,b} \mathring{w}_{n,a,b}} \sum_{n,a,b} \mathring{w}_{n,a,b} \partial_\theta^2 \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta}). \tag{4}$$

(III) If $f_s > 0$ and N_s denotes the number of deciders facing s choice occasions, then V_s can be estimated consistently using

$$\hat{V}_s = {N_s}^{-1} \sum_{n:T_n=s} \hat{g}_n(\mathring{W}) \hat{g}_n(\mathring{W})', \tag{5}$$

where $\hat{g}_n(\mathring{W}) := \left(\sum_{a=1}^{s-1} \sum_{b=a+1}^{s} \mathring{w}_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta}) \right)$.

For the proof, see Appendix B.

Since $V_\theta(W)$ is a matrix, it is not obvious that a weighting scheme exists that minimises the variance in the sense of positive definite matrices. Instead, we will investigate optimal choices with respect to linear functions of the form $l : \mathbb{R}^{d \times d} \to \mathbb{R}; V \mapsto \operatorname{tr}(VA)$ (for a positive semidefinite matrix $A \in \mathbb{R}^{d \times d}$), mapping positive definite matrices V to the real line. That is, we want to find weights W that minimise $\operatorname{tr}(V_\theta(W)A)$ for given matrix A.

The form of the linear functional $l(V) = \mathrm{tr}(VA)$ covers, with an appropriate choice of the matrix A, a wide range of possible options. With $A = e_j e_j'$, $e_j \in \mathbb{R}^d$ denoting the jth standard basis vector, we get $l(V_\theta(W_s)) = V_\theta(W_s)_{j,j} = \mathrm{Var}(\theta_j)$, enabling the minimisation of the variance of individual parameters. Kessels et al. (2006) outline four criteria to evaluate choice design efficiency,[2] which are related to A-, D-, G-, and V-optimality. Kessels et al. (2006) uses a Bayesian setting and hence integrate the optimality criteria over the prior distribution. As we work in a frequentist setting, we will neglect this.

All these concepts take estimation accuracy as measured by the variance into account, but differ in the form of dependence. The criterion related to A-optimality, which aims to minimise the average variance of the parameters, is $l(V_\theta(W_s)) = \mathrm{tr}\left(V_\theta(W_s)\right)$, which is obtained by choosing $A = I_d$. The criterion termed V-optimality is defined as $l(V_\theta(W_s)) = \int_X \sum_{y=1}^J c'(y,X) V_\theta(W_s) c(y,X) p_y(X) dF(X) = \mathrm{tr}\left(V_\theta(W_s) \int_X \sum_{y=1}^J c(y,X) c'(y,X) p_y(X) dF(X)\right)$, which is covered by our methodology with $A = \int_X \sum_{y=1}^J c(y,X) c'(y,X) p_y(X) dF(X)$. Here, $c(y,X) = \partial_\theta \mathbb{P}(y,X;\hat{\theta})$ for $X \in \mathbb{R}^{R\times 1}$ and $y \in \{1,\dots,J\}$, $p_y(X) = \mathbb{P}(y,X;\hat{\theta})$, and $F(X)$ denotes the Cumulative Distribution Function (CDF) of X. Using the Delta rule $c'(y,X)V_\theta(W_s)c(y,X)$ equals the variance of the predicted choice probability for choice y and regressor matrix X. V-optimality hence assesses the average variance of the estimation of choice probabilities.

The criterion for D-optimality is $l(V_\theta(W_s)) = \det\left(V_\theta(W_s)\right)$. As the determinant is a non-linear functional, our approach does not cover this particular criterion. G-optimality is defined using $l(V_\theta(W_s)) = \max_{y,X} c'(y,X)V_\theta(W_s)c(y,X)$, which again our methodology does not cover since the maximum function is not linear.

Theorem 2 (*Optimal Group-Specific Weights*). *Let $l : \mathbb{R}^{d\times d} \to \mathbb{R}$ be a linear mapping of the form $l(V) = \mathrm{tr}(VA)$, with $A \in \mathbb{R}^{d\times d}$, $A \neq 0$ symmetric and positive semidefinite, $\hat{\theta}$ be the CML estimator maximising the weighted CML function (1), and let the data be generated subject to Assumption 2. Let $\mathring{w}_{n,a,b}$ be the initial weights fulfilling Assumption 1, where $\sum_{s=2}^{\overline{S}} f_s w_s C_s(\mathring{W}) = 1$.*

Then $l(V_\theta(W_s))$ is minimised over W_s by

$$w_s^* = \left(\sum_{s=2}^{\overline{S}} f_s C_s(\mathring{W})^2 / v_s(\mathring{W})\right)^{-1} C_s(\mathring{W})/v_s(\mathring{W}) \propto C_s(\mathring{W})/v_s(\mathring{W}), \tag{6}$$

with $v_s(\mathring{W}) = l\left(H_0^{-1} V_s(\mathring{W}) H_0^{-1}\right)$.

For the proof, see Appendix B.

These weights W_s^* are, in general, different from $w_s = 1$ or $w_s = C_s(\mathring{W})$. Note that the proportionality constant is not relevant for the estimation but influences the formulas for the asymptotic variance (as the restriction $\sum_{s=2}^{\overline{S}} f_s w_s C_s(\mathring{W}) = 1$ is required in Theorem 1).

Note that the optimal weight may differ for each diagonal entry of V_θ. Since CML is not efficient, we potentially gain from using different functions for each parameter.

The next natural question is how to determine the optimal weights w_s^* in a practical setting, as H_0 and $V_s(\mathring{W})$ need to be estimated. The formula shows that it depends on $C_s(\mathring{W})$, which does not depend on the data, but the CML method we use. Equally weighted full pairwise weighting, for example, implies $C_s(\mathring{W}) = s(s-1)/2$; adjacent pairwise weights lead to $C_s(\mathring{W}) = s - 1$. The second factor is the variance entry $v_s(\mathring{W})$: deciders with choice occasions where we do not learn much from (large $v_s(\mathring{W})$) get a smaller weight, whilst others get a larger weight. This, of course, is reminiscent of Generalised Least Square (GLS) estimation.

From the data, we can estimate V_s as the empirical variance of the derivative of the likelihood contribution of the deciders with a particular number of observations s (compare Theorem 1(III), as well as H_0).

With these estimates, the optimisation above can easily be applied in order to obtain a more efficient second-step estimator. The detailed procedure to calculate such an estimator can be seen in Algorithm 1 in Appendix C.

The estimation accuracy of $\hat{V}_s$ depends on the number of individuals facing s choice occasions. In many data sets, this number will differ heavily between values of $s = 2,\dots,\overline{S}$. It is well known that the estimation of variances is generally noisy. Good estimators, thus, need to rely on sufficient data support. Therefore, we suggest to enhance the estimation accuracy by imposing smooth variation of W_s as a function of s. This is achieved by using a parametric model $w_s = g(s)$ (for details in the simulations, see Section 2.2.2) and leads to a variation of the two-step optimal group-weight CML estimator, as described in Algorithm 2 in Appendix C. The choice of parametric model $g(s)$ is dependent on the available data (it is not useful, for example, if only two or three different values for s occur in a dataset) and is ultimately at the discretion of the practitioner, similar to Feasible Generalised Least Square (FGLS; see Wooldridge, 2015). Weights and estimates calculated using this algorithm will henceforth be denoted with BB_param.

Even though – in contrast to the probit ML estimator – this two-step optimal group-weight CML estimator is not efficient, it is computationally feasible and potentially has lower variance than the standard unweighted CML estimator.

Note also that in the following finite sample simulation example (Section 2.2), we use $l(\cdot) = \mathrm{tr}(\cdot)$, so the weights are calculated to minimise the trace of the variance–covariance matrix of the estimator. Another possible approach could be to optimise differently weighted CMLs for each parameter dimension to obtain the best possible estimator. To do so, one could alternate between the estimation of the parameters of β, Ω, and Σ, with different weighting strategies for each, whilst keeping the other parameters fixed. Joe and Lee (2009) proposed a similar procedure in their paper.

[2] This reference has been pointed out to us by a referee for which we are grateful.

As a further note, observe that calculating the optimal weights w_s^* according to Eq. (6) requires the calculation of the second-order derivative $\partial_\theta^2 \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta})$, which is computationally expensive. An alternative is to use the second Bartlett identity

$$\mathbb{E}\partial_\theta^2 \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta}) = -\mathbb{E}\left(\partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta})\right)\left(\partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta})\right)'.$$

This requires only the calculation of first-order derivatives and leads to a slightly altered algorithm, which has led to similar results as compared to using the exact second derivative.[3]

2.2. Finite-sample simulation study

To investigate the effects of the individual weights on the CML estimator in the MACML setting, we simulate unbalanced panel data sets with an underlying MNP model with mixed effects and estimate the model using MACML with different weighting strategies for individuals with different numbers of observations.

2.2.1. Simulation setup

For $J = 6$ choice alternatives, we simulate 100 panel data sets with 300 individuals, each with the underlying DGP

$$y_{n,t} = \underset{j\in\{1,\dots,6\}}{\arg\max}\ y^*_{n,t,j} \tag{7}$$

$$y^*_{n,t,j} = \beta_{0,j} + \beta_{1,n} x_{1,n,t,j} + \beta_{2,j} x_{2,n,t} + \varepsilon_{n,t,j}, \tag{8}$$

with the parameters

$$\beta_{0,j} = (1 - (j-1)/6)^2 - 1, \quad j = 1, \dots, 6, \tag{9}$$

$$\beta_{1,n} \sim \mathcal{N}(\mu_1 = 1,\ \omega_1 = 0.25) \quad \text{iid drawn for each individual } n, \tag{10}$$

$$\beta_{2,j} = \begin{cases} \sin(30j) & 2 \le j \le 5 \\ 0 & \text{else.} \end{cases} \tag{11}$$

The regressors x were iid drawn for each observation such that

$$x_{1,n,t,j} \sim \begin{cases} \mathcal{N}(0,1) & 1 \le j \le 3 \\ \mathcal{N}(0,0.5) & \text{else} \end{cases} \tag{12}$$

$$x_{2,n,t} \in \{0,1\}, \quad \text{with} \quad \mathbb{P}(x_{2,n,t} = 1) = \mathbb{P}(x_{2,n,t} = 0) = 0.5. \tag{13}$$

The error terms $\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \dots, \varepsilon_{n,t,6})'$ are iid distributed over time and individuals such that $\varepsilon_{n,t} \sim \mathcal{N}(0, \Sigma)$ with Σ as a diagonal matrix with entries $\Sigma_{jj} = |\tilde{\Sigma}_{jj}/\tilde{\Sigma}_{11}|$, $\tilde{\Sigma}_{jj} \sim \mathcal{N}(0,1)$. The entries of Σ are drawn once for each of the 100 data sets and then kept fixed within each data set. To simulate unbalanced data sets, each of the 300 individuals n within each data set gets assigned a random number of choice occasions T_n with $T_n \sim \mathcal{P}_{2\le k\le 20}(\lambda = 10)$, where $\mathcal{P}_{2\le k\le 20}$ denotes a truncated Poisson distribution.

2.2.2. Estimation

We used a version of the MACML estimation procedure with a Solow-Joe (SJ) approximation (Joe, 1995; Solow, 1990) of the MVNCDFs (as proposed by Bhat, 2011) to evaluate the pairwise marginal likelihoods and specified the estimated model according to the DGP. For identification, we fixed $\hat{\beta}_{0,1} = 0$, $\hat{\beta}_{2,1} = 0$, and $\hat{\Sigma}_{1,1} = 1$ at the true values, leaving $\theta = (\beta_{0,2}, \dots, \beta_{0,6}, \mu_1, \omega_1, \beta_{2,2}, \dots, \beta_{2,6}, \Sigma_{22}, \dots, \Sigma_{66})'$ with $3(6-1)+2 = 17$ free parameters to estimate.

Details on the code used for the estimation can be found in Appendix D.

The proposed two-step optimal group-weight CML estimator was estimated according to Algorithm 1 with $\mathring{W} = 1$. The optimal weights and corresponding estimates will be denoted with the abbreviation BB.

To calculate the weights as proposed in Algorithm 2, we used the parametric model

$$1/g(s) = \gamma_0 + \gamma_1 s + \gamma_2 s^2 + \nu_s, \tag{14}$$

with ν_s denoting the random error of the model, and estimated the restricted least squares estimator with the restrictions

$$\gamma_0 + \gamma_1 s + \gamma_2 s^2 \ge \min(\hat{\upsilon}_s / C_s(\mathring{W})), \quad \gamma_1 + 2\gamma_2 s \ge 0, \tag{15}$$

ensuring positive and in s monotonically decreasing weights with

$$\breve{w}_s = (\hat{\gamma}_0 + \hat{\gamma}_1 s + \hat{\gamma}_2 s^2)^{-1}, \tag{16}$$

which has the same form as the optimal weights calculated by Joe and Lee (2009). The estimation was done in R using the COBYLA algorithm as implemented in the *nloptr* package (see Johnson, 2022; Powell, 1994). The optimal weights and the corresponding

[3] Results can be obtained from the authors on request.

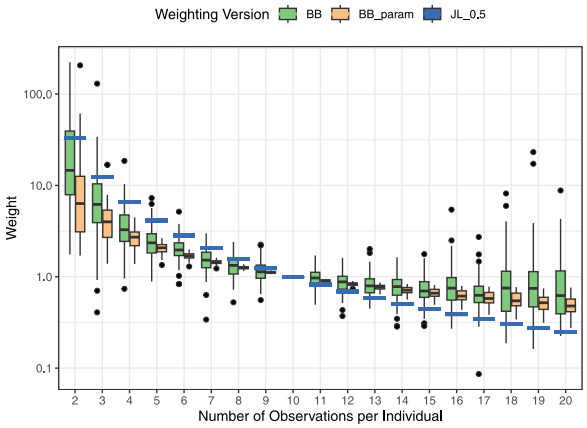


Fig. 1. Boxplots of the distribution of calculated optimal weights for 100 simulated data sets compared to heuristic weights according to Joe and Lee (2009), represented by horizontal lines. Weights are scaled such that individuals with 10 observations get a unit weight. The y-axis has a logarithmic scale.

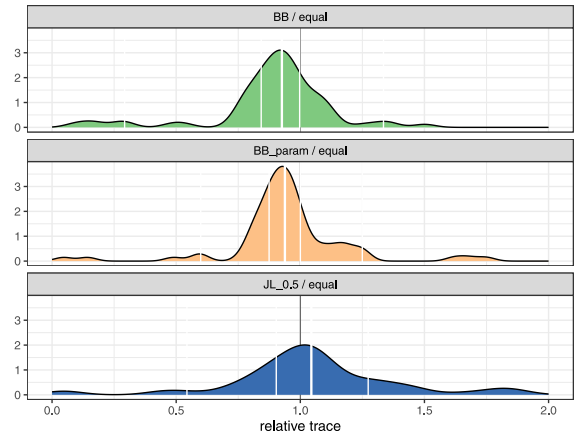


Fig. 2. Distribution of relative trace of variance–covariance matrix $\hat{V}_\theta$ of the weighted models compared to the unweighted model from 77 successful simulations. Vertical breaks in the filling indicate 5%, 25%, 50%, 75%, and 95% percentiles. The x-axis is scaled to $[0, 2]$ for better visibility. 5 of the JL_0.5 weighted and 2 of the BB weighted, as well as the 95% percentile of the JL_0.5 version, are out of scale.

estimates calculated according to Algorithm 2, with the parametric model for the weights as described above in Eqs. (14)–(16), will be abbreviated with BB_param.

For comparison, another weighted model was estimated using the heuristic weights suggested by Joe and Lee (2009) with $\tilde{w}_s = (s-1)^{-1}[1 + 0.5(s-1)]^{-1}$, which will be abbreviated with JL_0.5.

These simulations of the data and estimations of the models were repeated 100 times, of which 77 times all five models were successfully estimated. Successfully estimated means, in this context, that the `nlm()` function from the R package `stats` minimising the CML had as exit code either 1 ("relative gradient is close to zero, current iterate is probably solution".) or 2 ("successive iterates within tolerance, current iterate is probably solution".). In the remaining cases, optimisation ended with code 5 ["maximum step size `stepmax` exceeded five consecutive times. Either the function is unbounded below, becomes asymptotic to a finite value from above in some direction or `stepmax` is too small".], implying that the numerical optimisation algorithm did not successfully converge to a minimum (for details, see R Core Team, 2021). Interestingly, the `nlm()` function was in this study more frequently successful in locating a maximum when using the proposed weighted BB and BB_param procedures, compared to the equally weighted or JL_0.5 versions. Table E.2 in the Appendix provides an overview of the reported `nlm()` exit codes by weighting scheme. To ensure a fair assessment of the weighting methods, the 23 simulations in which at least one method did not lead to an exit code of 1 or 2 were excluded from the subsequent analysis.

An overview of the resulting weights calculated for the successfully estimated models can be seen in Fig. 1.

2.2.3. Results

The following results are calculated for the 77 simulations, in which all five estimators were successfully estimated. Using the optimal BB weighted estimator, as described in Algorithm 1, we can reduce the trace of the covariance matrix V_θ on average by

4.38% over the 77 simulations compared to the unweighted model. In the median over the 77 different simulations, the reduction is, however, by 7.42%, and the BB weighted model has in 58 of the 77 cases a lower trace of the covariance matrix $\hat{V}_\theta$ than the unweighted estimator.

Utilising a parametric function to estimate the weights BB_param, as described in Algorithm 2, leads to more stable weights compared to the BB version, as showcased by the distribution of the weights in Fig. 1. Using this estimator leads to a reduction of the trace of the covariance matrix $\hat{V}_\theta$ on average by 5.25%, in the median by 6.2% and leads to a smaller trace in 72.7% of the cases [56 out of 77], all compared to the trace of the covariance matrix of the equally weighted estimator.

In comparison, the estimator with the JL_0.5 weights has, on average, in the 77 simulations, a 23.02% larger trace of the covariance matrix $\hat{V}_\theta$ compared to the unweighted estimator, in the median a 4.4% larger trace and has in just 41.56% of the cases [32 out of 77] a lower trace than the covariance matrix of the unweighted estimator.

In Fig. 2, the distribution of the quotient between the traces of the covariance matrix of the weighted estimator and that of the unweighted estimator is shown for the different simulated models. An overview of the distributions over the different simulated data sets of the quotients between the traces of differently weighted estimators can be seen in Table E.3 in the Appendix.

Although the optimal weights, as described in Algorithm 1, are asymptotically independent of the group weights for the initial weights $\dot{W}$ (for constant relative weight of the pairs within a group, that is), this may not be the case in finite samples. In addition to the optimal group weights according to Algorithm 1 with a uniformly weighted initial model with $\dot{W} = 1$, a second set of optimal group weights with $\dot{W}$ set to the deterministic group weights JL_0.5 according to Joe and Lee (2009) was computed, which will be denoted BB_fJL. The average ratios between the resulting weights and the ratios between the traces of the variance–covariance matrices were calculated and an overview of the distributions can be seen in Tables E.4 (comparing weights) and E.5 (comparing traces of the variance matrix) in the Appendix. Whilst this example demonstrates that the optimal group weights do depend on initial group weights, the effects are in most cases small.

Beside comparing the trace of the variances one can also measure the precision of the estimated choice probabilities. For this we use the mean squared l_2 distance between the predicted choice probabilities given the model and the choice probabilities according to the true DGP

$$\bar{l}_2^2(p_0, \hat{p}(W)) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{T_n}\sum_{t=1}^{T_n}\sum_{j=1}^{J}\left(\mathbb{P}\left(y_{n,t}=j, X_n; \theta_0\right) - \mathbb{P}(y_{n,t}=j, X_n; \hat{\theta}(W))\right)^2,$$

depending on the weighting schemes W. We calculate the percentual change compared to the measure for the equally weighted model with $W_{\text{equal}} = 1$ as

$$\Delta_{\%}\bar{l}_2^2(W) = \frac{\bar{l}_2^2(p_0, \hat{p}(W_{\text{equal}})) - \bar{l}_2^2(p_0, \hat{p}(W))}{\bar{l}_2^2(p_0, \hat{p}(W_{\text{equal}}))}.$$

An overview of the distribution of the results can be seen in Table E.6 in the Appendix. These indicate that using weights designed to reduce the variance of the estimator also results in an improvement in recovering choice probabilities. Note, however, that this measure does not account for the panel nature of the data.

3. Weighting in case of unaccounted autoregressive error structure

In the previous section, we focused on the relative weighting between observations from deciders with a different number of choice occasions. The weights, however, are also relevant for the relative importance of the different pairs for one decider.

In some situations, it is plausible that the error process follows a continuous time stationary process with nonzero autocorrelation function across time. Such behaviour can be interpreted as taste perseverance if it generates positive correlations.

Typically, for stationary processes, the autocorrelation will decrease with increasing temporal distance between the two observations (compare, for example, Lütkepohl, 2005). An autoregressive process of order 1 (AR(1)), for example, is given as the stationary solution to the difference equation $\varepsilon_t = \rho\varepsilon_{t-1} + \tilde{\varepsilon}_t$ for iid process $\tilde{\varepsilon}_t, t = 1, 2, \ldots$, where $|\rho| < 1$ is assumed (the stability assumption). The correlation between observations at time 1 and at time t is then given as ρ^{t-1} and, hence, decreases with increasing distance t.

The MNP models with random effects, as discussed in the previous section, however, introduce correlation between observations by including a random constant in the error term of the utility: Random effects in the alternative specific constant (ASC), for example, introduce autocorrelation for the various error terms of one individual (see, for example, Train, 2009). These error terms then also form a stationary sequence with random mean vector and constant correlation for nonzero temporal differences, which consequently does not decrease to zero for $t \to \infty$.

This difference in behaviour has consequences in situations where the underlying DGP is not known and a misspecified mixed MNP model is used, neglecting the temporal correlation. Misspecification of the model typically leads to inconsistent estimators. The degree of misspecification depends on the strength of the correlation, which differs between different pairs of observations.

Below, we investigate how the weights for different pairs can be used to reduce the influence of the aforementioned type of misspecification on the estimation of the model. This will first be done in a simplified example where asymptotic properties of the estimators are illustrated, showing the magnitude of asymptotic biases that can occur in conjunction with proposals to limit the corresponding inconsistency. Subsequently, a simulation exercise will show that similar effects can also be observed in finite samples.

3.1. Variance-covariance structure of autocorrelated errors

In this section, we deal with stationary vector processes of the autoregressive type for a given decider n. We will always assume independence and identical distribution across different deciders, and hence consider the properties of the error process only over time for a representative decider n. As we assume a balanced panel data set with independent and identically distributed regressors $x_{n,t}$ and errors $\varepsilon_{n,t}$ over different deciders n, we omit the index n in the remainder of this subsection, as well as in Section 3.2.

Here, to fix notation, the error terms $\varepsilon_t = (\varepsilon_{t,1}, \dots, \varepsilon_{t,J})'$, $t = 1, \dots, T$, constitute a stationary vector autoregressive process of order 1 (VAR(1)) if they are a stationary solution to the difference equation

$$\varepsilon_t = \Psi \varepsilon_{t-1} + \tilde{\varepsilon}_t \tag{17}$$

where $\tilde{\varepsilon}_t = (\tilde{\varepsilon}_{t,1}, \dots, \tilde{\varepsilon}_{t,J})' \sim \mathcal{N}_J(0, \tilde{\Sigma})$ is iid over time $t = 1, \dots, T$ (see, for example, Lütkepohl, 2005).[4]

In this situation, the stationary distribution is a normal distribution with expectation zero and variance Σ solving the Lyapunov equation $\Sigma = \Psi \Sigma \Psi' + \tilde{\Sigma}$ if and only if all eigenvalues of the matrix $\Psi \in \mathbb{R}^{J \times J}$ are inside the unit circle (stability condition).

For a pair of choice observations $p = (a, b)'$ at time points $t_a < t_b$, we can now express the covariance matrix Σ_p of the joint process $(\varepsilon'_{t_a}, \varepsilon'_{t_b})'$ as

$$\Sigma_p = \begin{pmatrix} \Sigma & \Sigma(\Psi^{t_b - t_a})' \\ \Psi^{t_b - t_a} \Sigma & \Sigma \end{pmatrix}.$$

In case of no autocorrelation, where $\Psi = 0$, we have $\Sigma_p = \begin{pmatrix} \tilde{\Sigma} & 0 \\ 0 & \tilde{\Sigma} \end{pmatrix}$. In case of a simple autocorrelation structure, where $\Psi = \rho I_J$, we have the simplified form

$$\Sigma_p = \frac{1}{1 - \rho^2} \begin{pmatrix} \tilde{\Sigma} & \rho^{t_b - t_a} \tilde{\Sigma} \\ \rho^{t_b - t_a} \tilde{\Sigma} & \tilde{\Sigma} \end{pmatrix}.$$

Under the stability assumption $\Psi^k \to 0$ for $k \to \infty$, the covariance $\Psi^{t_b - t_a} \Sigma$ of the errors decreases with increasing temporal distance. This implies that, for distant pairs of observations, the misspecification of the autocorrelation of the error terms (for example by assuming zero correlation) is of less importance.

A DGP with autocorrelated errors and no random effects thus possesses a covariance between two observations equal to Σ_p.

Similarly, a model with randomly mixed ASCs with variance–covariance matrix Ω and iid error terms has the variance matrix

$$\Sigma_p = \begin{pmatrix} \tilde{\Sigma} + \Omega & \Omega \\ \Omega & \tilde{\Sigma} + \Omega \end{pmatrix}$$

for the joint vector of the error terms.

These two matrices coincide when $\Psi^{t_b - t_a} \Sigma = \Omega$ and then $\tilde{\Sigma} = \Sigma - \Omega$. For a given DGP, this is true, for example, for the following special cases:

1. Ψ is such that $\Psi^{t_b - t_a} \Sigma = \Omega$ for some $t_b - t_a$. In this case, symmetry must hold, such that $\Psi^{t_b - t_a} \Sigma = \Sigma(\Psi^{t_b - t_a})'$. This holds, for example, for $\Psi = \rho I_J$. Clearly, stability implies that this may hold for at most one value of $t_b - t_a$ for nonzero Ψ.
2. $\Psi^{t_b - t_a} = 0$ and, hence, $\Omega = 0$.

It follows that the two models lead to different forms of correlation over time. The expressions can only be identical for one value of $t_b - t_a$. Since $\Psi^k \to 0$ for $k \to \infty$, the second case approximately holds for large temporal distances.

In all other cases, misspecifying the model by mistakenly using randomly mixed ASCs when the data generating process uses a stationary but correlated error process will lead to inconsistent estimators.

In the next subsection (Section 3.2), we investigate the corresponding asymptotic bias in a special case, whereas, in the following subsection (Section 3.3), we examine the finite sample properties in a simulation study.

3.2. Deriving the asymptotic bias in misspecified cases

As a starting point, consider the following simple setup: The decision problem involves a number of consecutive binary decisions. This allows us to consider only the difference between the utilities of the two alternatives. Thus, the utility of the first alternative is zero, whilst the utility for the second alternative is assumed to equal $y^*_{t,2} = \beta_1 + x_t + v_t$, with $v_t = \gamma + u_t, u_t = \rho u_{t-1} + \varepsilon_t$, and $\varepsilon_t \sim \mathcal{N}(0, \tilde{\sigma}^2)$. This implies that there exists one alternative specific constant for alternative 2 (the one for alternative 1 is set to zero to fix the level), which contains a random individual specific part γ (with expectation equal to zero and variance ω^2) included in the error term. Additionally, we include a second regressor $x_{n,t}$ drawn iid standard normally distributed over both individuals and choice situations with a corresponding coefficient normalised to 1 (to fix the scale). The error term u_t follows an AR(1) process with autocorrelation coefficient $|\rho| < 1$.

The specification encompasses both the data generating process as well as the model:

[4] The distribution of the error is not important for the stationarity properties, but due to the consideration of probit models.

- For the DGP, the data are considered to be generated with $\omega = 0$, such that there are no random effects, but $\rho \in (-1, 1)$, such that the variance of u_t equals $\sigma^2 = 1$, and the correlation between u_t and u_{t-k} equals $\rho^{|k|}$.
- For the model, we assume that the temporal dependence is neglected, and thus $\rho = 0$ is used. Instead, a random effect in the ASC is postulated. Consequently, three parameters are estimated: β_1, ω^2, and σ^2.

In this setting, we investigate the asymptotic bias of the misspecified estimator, assuming random effects but ignoring the temporal correlation structure.

The calculation of the limiting estimator is achieved using the following setting: We calculate the binary choices observed at days $t \in \mathcal{T} := \{1, 2, 3, 366, 367, 368\}$ corresponding to two waves (in adjacent years) of observations for three days each. The regressors x_t are drawn from an underlying finite set of vectors (we use $M = 100$ in our calculations). For a given vector $x_{:} = [x_t]_{t\in\mathcal{T}}$, we then calculate the choice probabilities for all possible combinations of choices $[y_t]_{t\in\mathcal{T}}$ according to the DGP. These are obtained using the combined random utility vector ($\iota = [1, \dots, 1]'$)

$$U_{:} = \beta_{1,\circ}\iota + x_{:} + v_{:}, \quad \text{with } \mathrm{E}(v_{:}) = 0,$$

$$\mathrm{Var}(v_{:}) = \sigma_\circ^2 \left(\begin{array}{ccc|ccc} 1 & \rho_\circ & \rho_\circ^2 & \rho_\circ^{365} & \rho_\circ^{366} & \rho_\circ^{367} \\ \rho_\circ & 1 & \rho_\circ & \rho_\circ^{364} & \rho_\circ^{365} & \rho_\circ^{366} \\ \rho_\circ^2 & \rho_\circ & 1 & \rho_\circ^{363} & \rho_\circ^{364} & \rho_\circ^{365} \\ \hline \rho_\circ^{365} & \rho_\circ^{364} & \rho_\circ^{363} & 1 & \rho_\circ & \rho_\circ^2 \\ \rho_\circ^{366} & \rho_\circ^{365} & \rho_\circ^{364} & \rho_\circ & 1 & \rho_\circ \\ \rho_\circ^{367} & \rho_\circ^{366} & \rho_\circ^{365} & \rho_\circ^2 & \rho_\circ & 1 \end{array}\right).$$

Note here that, for the off-diagonal blocks, we have $|\rho_\circ^k| \leq |\rho_\circ^{363}| \approx 0$ for practically all values of $|\rho_\circ| < 1$. Using these variances, we calculate $\mathbb{P}(y_{:}; x_{:}, \beta_{1,\circ}, \sigma_\circ^2, \rho_\circ)$, the choice probabilities for all possible combinations according to the data generating process.

In the estimation, we assume for the model random effects as well as $\rho = 0$, which results in $U_{:} = \beta_1\iota + x_{:} + v_{:}$ with $\mathrm{E}(v_{:}) = 0$, $\mathrm{Var}(v_{:}) = \omega^2\iota\iota' + \sigma^2 I_6$, $\iota = [1, \dots, 1]'$.

Using this variance, we obtain for every pair of choices (y_a, y_b) the model choice probabilities $\mathbb{P}(y_a, y_b; x_{:}, \beta_1, \omega^2, \sigma^2)$. Clearly, this is different from the variance due to the DGP, and thus the model is misspecified. As the criterion function, we use the pairwise CML with weights $w_{a,b}$ for the pair (y_a, y_b). For a discrete uniform distribution over $x_{:}$ in the set $\{X_i, i = 1, \dots, 100\}$ this converges to[5]

$$Q_\circ(\beta_1, \omega^2, \sigma^2) = \frac{1}{100}\sum_{i=1}^{100}\sum_{y_{:}} \mathbb{P}(y_{:}; X_i, \beta_{1,\circ}, \sigma_\circ^2, \rho_\circ)\left(\sum_{a=1}^{5}\sum_{b=a+1}^{6} w_{a,b}\log\mathbb{P}(y_a, y_b; X_i, \beta_1, \omega^2, \sigma^2)\right).$$

This limiting asymptotic function is then maximised with respect to the parameters β_1, ω^2, and σ^2 in order to calculate the asymptotic value of the estimators.

In the comparison, we include four different weighting schemes:

FP	The full pairwise CML uses $w_{a,b} \equiv 1$.
growth	The distant pairs, or step growth, CML uses $w_{a,b} = \mathbb{I}(\lvert t_a - t_b\rvert > 7)$. Only pairs measured with more than a week time difference are included.
adj.	The adjacent pairwise CML uses $w_{a,b} = \mathbb{I}(\lvert a - b\rvert = 1)$. Only consecutive observations, irrespective of the distance between them, are used.
decay	The step decay CML uses $w_{a,b} = \mathbb{I}(\lvert t_a - t_b\rvert \leq 7)$. Only pairs of observations less than a week apart are used.

Since the correlation between two pairs is – according to the DGP – given by $\rho^{|t_a - t_b|}$, we see that distant pairs are practically uncorrelated, whilst, for adjacent observations, we obtain a correlation of ρ when $|t_a - t_b| = 1$ and a correlation of almost zero for $|t_a - t_b| = 363$. According to the model, however, independent of the temporal distance, we obtain a variance of $\omega^2 + \sigma^2$ and a covariance of all pairs of ω^2, leading to a correlation of $\omega^2/(\omega^2 + \sigma^2)$.

For the calculations, we use $\beta_{1,\circ} = 1, \sigma_\circ^2 = 1$ and vary $\rho_\circ \in (-1, 1)$. It follows that, in order to mimic the correlation and the variance according to the DGP, for adjacent pairs we must have $\omega^2 + \sigma^2 = 1$ and $\omega^2 = \rho_\circ$, if $\rho_\circ \geq 0$. For negative $\rho_\circ$, the closest ω^2 equals $\omega^2 = 0$.

For distant pairs with one year between observations, the true correlation amounts to almost zero, whilst the model still implies a correlation of $\omega^2/(\omega^2 + \sigma^2)$. Thus, the fit is perfect if $\omega^2 = 0$ and $\sigma^2 = 1$.

Therefore, we expect that the adjacent weighting results in an asymptotic bias for positive values of $\rho_\circ$, where the bias gets larger with larger values of $\rho_\circ$. For the distant weighting approach, we expect that $\omega^2 = 0$ and $\sigma^2 = 1$ throughout. Whilst the temporal correlation is not estimated correctly, it does not lead to an inconsistent estimation for ω if time points are included such that the temporal correlation has vanished. If only close pairs are considered, however, the estimator for ω^2 compensates to match the correlation.

This behaviour can be seen in the three plots in Fig. 3.

[5] The choice of 100 different regressor vectors is arbitrary and only done in order to reduce the impact of the regressors. One way to view this is the approximation of the expectation over continuously uniformly distributed regressors vectors.

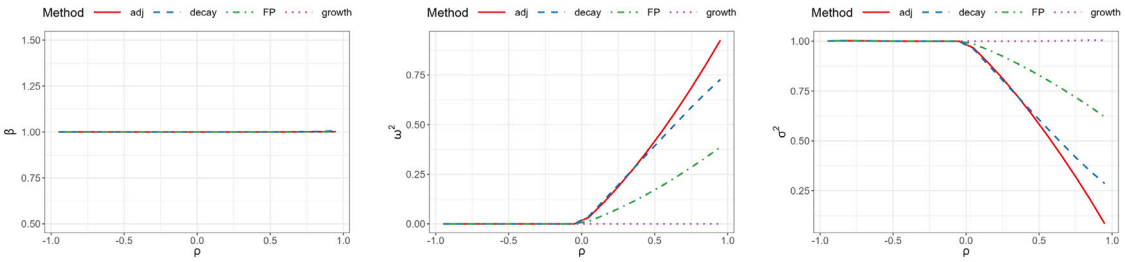


Fig. 3. Visualisation of the asymptotic bias in β, ω^2, and σ^2, respectively, for the four different CML pair structures.

3.3. Finite-sample simulation

In this subsection we demonstrate that the asymptotic effects derived in Section 3.2 in a simple case can also be observed in finite-samples in more complex cases. We use a simulation study with autoregressive VAR(1) errors and no mixed effects in the DGP, but with mixed ASCs and no auto-correlated errors in the estimated model.

3.3.1. Simulation setup

For $J = 6$ choice alternatives, we simulated 100 panel data sets with 300 individuals and 10 observations for each individual at time points $t = 1, 2, 3, 4, 5, 366, 367, 368, 369, 370$, resulting in two waves of five observations each in adjacent years, mimicking the study design of the German Mobility panel; see, for example, Zumkeller and Chlond (2009). For all data sets and individuals, we used the same DGP as described in Section 2.2 in Eqs. (7)–(13), with the modification that $\omega_1 = 0$ instead of $\omega_1 = 1$, such that we have fixed $\beta_{1,n} = 1$.

The error terms $\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \dots, \varepsilon_{n,t,6})'$ are, however, randomly drawn from a VAR(1) process, as described in Eq. (17), and scaled such that $\mathrm{Var}(\varepsilon_{n,t,1}) = 1$. The errors $\varepsilon_{n,1}$ for the first observation of an individual were drawn from the stationary distribution $\mathcal{N}(0, \Sigma)$. The following errors are calculated as

$$\varepsilon_{n,t+1} = \Psi \varepsilon_{n,t} + \tilde{\varepsilon}_{n,t+1},$$

with $\tilde{\varepsilon}_{n,t+1} \sim \mathcal{N}(0, \tilde{\Sigma})$, $\tilde{\Sigma}$ as a diagonal matrix with entries $\tilde{\Sigma}_{jj} = (1 - \rho^2)|\tilde{\sigma}_{jj}|$, $\tilde{\sigma}_{jj} \sim \mathcal{N}(0, 1)$ for $j > 1$ and $\tilde{\Sigma}_{11} = (1 - \rho^2)$. The entries of $\tilde{\Sigma}$ are drawn once for each of the 100 data sets and are then kept fixed for all individuals within one data set.

Five types of errors were simulated for each of the 100 data sets of regressors, resulting in different observations for the choice variable y, leading to a total of 500 distinct data sets. To simulate the errors, five different parameter matrices Ψ were used, with $\Psi = \rho \mathbb{I}_6$, $\rho \in \{-0.95, -0.2, 0, 0.2, 0.95\}$.

3.3.2. Estimation

We used a version of the MACML estimation procedure with an SJ approximation of the MVNCDFs to evaluate the pairwise marginal likelihoods and used a misspecified model, where instead of using autocorrelated errors, we assumed iid errors over individuals and observations. To introduce correlation between the observations, the estimated model assumes mixed effects for the ASCs; thus, $\beta_{0,j} \sim \mathcal{N}(\mu_{0,j}, \omega^2_{0,j})$. For identification, we fixed $\hat{\mu}_{0,1} = 0$, $\hat{\beta}_{2,1} = 0$, and $\hat{\Sigma}_{1,1} = 1$, leaving $\theta = (\mu_{0,2}, \dots, \mu_{0,6}, \omega_{0,1}, \dots, \omega_{0,6}, \beta_1, \beta_{2,2}, \dots, \beta_{2,6}, \Sigma_{22}, \dots, \Sigma_{66})'$ with $3(6-1) + 6 + 1 = 22$ free parameters to estimate.

The model was then estimated using four different CML pair-types, as described in Section 3.2: *step growth*, *full pairwise*, *step decay*, and *adjacent pairwise*. A visualisation of the different pair-types is shown in Fig. 4.

For the MACML estimations the same setup was used as above (see Appendix D).

3.3.3. Results

Possibly owing to the misspecification of the estimated model, our procedures did not converge to a minimum in some cases. This means, in this case, that the `nlm()` function minimising the CML had as exit code neither 1 ("relative gradient is close to zero, current iterate is probably solution".), nor 2 ("successive iterates within tolerance, current iterate is probably solution".). Fig. 5 visualises the relative success rates of the four different models with different pair-types for the five different autocorrelation structures of the errors.

These results indicate that, in case of large positive autocorrelation coefficients in the unaccounted autoregressive process of the error, the inclusion of distant pairs in the CML increases the rate of successfully estimated models substantially.

To evaluate the accuracy of the estimated model further, we calculated the mean absolute deviation of the estimated parameters from the true parameters for each model. To avoid bias in the comparison due to problems with the normalisation implied by fixing the scale via $\Sigma_{1,1} = 1$, we decided to re-scale the estimated models to minimise the sum of squared deviations of the estimated parameters, including the deviation introduced in Σ_{11}, and combine the parameters of Σ and Ω to account for their joint effect in the model. Therefore, we effectively examined the distance within the equivalence class of systems differing only in the scale of the utility.

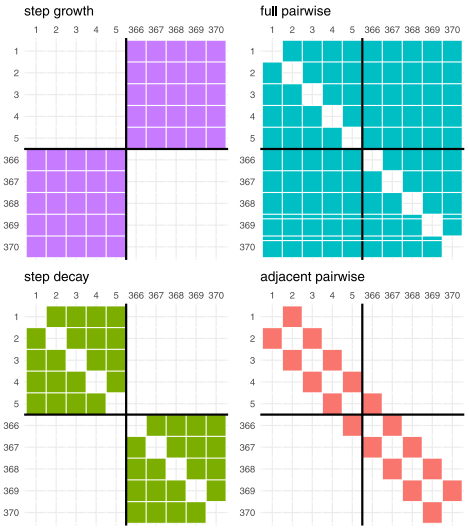


Fig. 4. Visualisation of the four different CML pair structures applied for the estimation process.

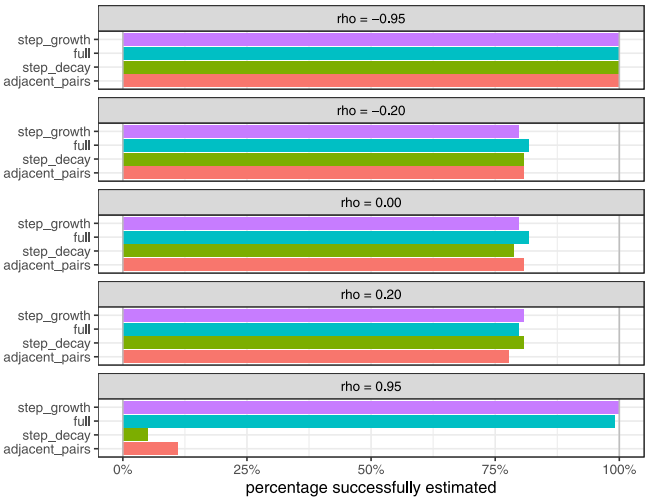


Fig. 5. Percentage of successfully estimated models by autocorrelation coefficient ρ and CML pair-type.

When looking at the mean absolute deviation of the estimated β parameters of the successfully estimated models, we do not see a significant difference in the deviations of the parameters across the different CML pair-types used for the estimation. This aligns with the asymptotic results presented in Section 3.2. An overview of the results is shown in Fig. 6.

3.4. Discussion

When looking at the estimated correlation between the different observations of one individual – in the estimated model introduced by random effects in the ASCs, in the DGP introduced by autocorrelation of the errors – we observe in the finite-sample simulations results similar to those shown in the analysis of the asymptotic behaviour in Section 3.2. This can be observed in Fig. 7, which summarises the distribution of the estimated average correlation $\tilde{\rho} = (6-1)^{-1} \sum_{j=2}^{6} \tilde{\Omega}_{jj} / (\tilde{\Omega}_{jj} + \tilde{\Sigma}_{jj})$.

The estimated correlation between the observations of one individual is due to the mixed ASCs and would, in the standard interpretation, be attributed to individual taste heterogeneity. In the estimated models, this correlation is the same for any pair of observations of one individual, irrespective of the temporal distance between the observations. In the DGP, however, the correlation was introduced due to an autoregressive process of the error terms; hence, the correlation between different observations of one individual decreases as a function of their temporal distance.

In the case of positive autocorrelation of the errors, the models using `full`, `step_decay`, or `adjacent_pairs` CML pair-types in the estimation process estimate significantly larger average correlations between the observations compared to the models using `step_growth` CML pair-types.

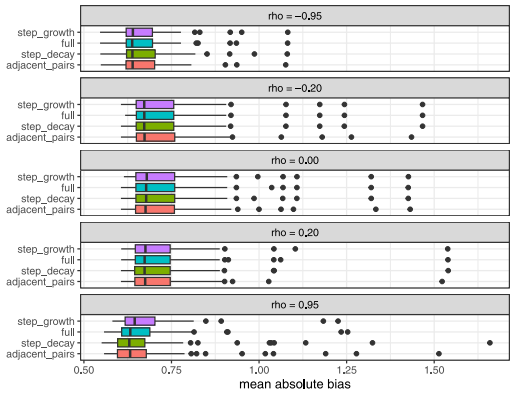


Fig. 6. Distribution of the mean absolute deviation of the estimated β parameters from the true parameters over all estimated models by autocorrelation coefficient ρ and CML pair-type.

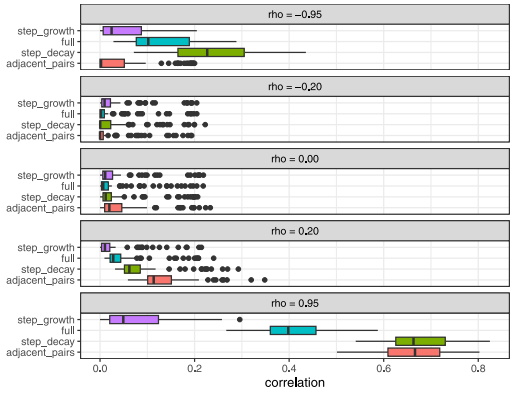


Fig. 7. Distribution of mean estimated correlation $\tilde{\rho}$ over all estimated models by autocorrelation coefficient ρ of the true autoregressive error process and CML pair-type.

In the case of no or small negative autocorrelation, there was no significant difference between the different pair-type structures used in the estimation.

In the case of large negative autocorrelation, both the `full` and `step_decay` models estimate a relatively large positive autocorrelation between the observations. This could be attributed to two factors. First, the models are – due to their structure – incapable of estimating any negative correlation between observations of one individual since $\tilde{\Omega}_{jj}/(\tilde{\Omega}_{jj} + \tilde{\Sigma}_{jj}) \geq 0$. Second, the observations with an even number $2k$ of steps between them have a positive correlation $\rho^{2k} \geq 0$, even for negative autocorrelation of the errors. This is then captured by models that include such pairs in their CML pair structure. The `adjacent_pairs` model has only pairs with temporal distances equal to one or equal to 361 in the CML pair structure, where $\rho < 0$ for $\rho^{361} < 0$; hence, there are no pairs with positive correlation in the estimation process. For this reason, a model with `adjacent_pairs` estimates only a very small correlation between observations in the case of negative autocorrelation of the errors, similar to the `step_growth` models.

Overall the simulation study has shown that when the errors follow an autoregressive process, which is not accounted for in the model, the inclusion of distant observations leads to a reduction of the deviation of the estimated correlation coefficients and makes the estimation process more reliable in terms of estimation success rate. In case of a large positive autocorrelation coefficient in the autoregressive error process, the estimators only relying on distant pairs performed by far the best. Since the direction of the autocorrelation is, however, usually not a priori known, it is, in any case, worth considering using distant observations in the CML when an autoregressive error process in the DGP cannot be ruled out but is also not explicitly modelled.

Note that these findings are due to the misspecification of the model not including temporal autocorrelation in the model, which hence is picked up partially by the random error term. The situation would change if the correct model, including temporal autocorrelation of the error term, were to be used. In that case, Varin and Vidoni (2006) show that pairs with large temporal distances contain less information on the autocorrelation parameter. In the correctly specified case, this is a disadvantage. In the misspecified case, however, it is an advantage.

4. Conclusion

In this paper, we have shown how different weighting strategies for pairwise CML estimation can be used to effectively improve the properties of the estimator in case of (1) unbalanced panel data, and (2) unaccounted autocorrelation of the errors.

In the case of unbalanced panel data, we introduced a new two-step optimal group-weight CML estimator, which exploits the panel structure by grouping individuals with the same number of observations together to assign group weights to effectively reduce the variance of the estimator. The estimator is based on asymptotic theory and has shown to be effective in a finite sample simulation study in which the two-step optimal group-weight CML estimator improved upon the unweighted pairwise CML estimator in 75.3% of the cases when measured in the trace of the variance–covariance matrix of the estimator. Utilising a parametric model to estimate the weights, as described in Algorithm 2, leads in the median to less of a reduction in the trace of V_θ than using the optimal weights directly. However, in the rare cases when the estimated optimal weights lead to an increase of the trace, this effect is less pronounced when using a parametric model for the weights. As a consequence the estimator described in Algorithm 2 has, on average, the lowest trace of the variance–covariance matrix compared to the unweighted estimator out of all tested weighted estimators. Using a parametric model to estimate the optimal weights, hence, seems to lead to more stable results and reduces the dependence on initial group weights $\mathring{W}$.

Although the reduction in estimator variance accomplished by utilising the proposed optimal group weights is less pronounced in this simulation study than the 50% reduction demonstrated in Appendix A, the gains attained are still noteworthy, and the extra computational cost is minor in comparison to the possible benefits.

In comparison, using the weights by Joe and Lee (2009) resulted in most of the cases in larger traces of the variance matrix than the usage of other weights, including the initial weights $w_s = 1$. It resulted, however, in one case in the largest overall improvement compared to the unweighted version. A detailed overview of the results can be seen in Table E.3.

In the case of an unaccounted autoregressive process of the error terms, we were able to reduce the effect of the misspecification of the model by choosing distant observations in the CML pair structure. This was shown in an asymptotic calculation, as well as in a finite sample simulation study. Using distant pairs reduces the effects of autocorrelation of the errors, which decreases over time and which potentially introduces an asymptotic bias in the estimation of the mixed effects, resulting in an inconsistent estimator. This effect is showcased in examples where the data represents a panel data set with two panel waves, in which the differences in covariances between observation pairs is especially apparent. It can be argued that, to distinguish between the covariance induced by individual specific effects and that induced by an autoregressive error process, the inclusion of distant pairs is most effective in case of data with distinct panel waves, such that the known data structure can be used in the estimation process.

Furthermore, the simulation study indicated that the inclusion of distant pairs in the CML results in a substantially higher rate of successfully estimated models when dealing with an unaccounted autoregressive error process with large positive autocorrelation coefficient.

Both results can also be combined using group-specific weights to account for unbalanced panels and stressing distant pairs of observations for a given number of choice occasions to robustify the estimation of the random effects. The latter weighting strategy can be used in order to detect temporal autocorrelation in the error terms, whilst the group-specific weights lead to improved accuracy at almost no numerical costs for the estimation.

CRediT authorship contribution statement

Sebastian Büscher: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Dietmar Bauer:** Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 356500581 which is gratefully acknowledged. The authors also thank Manuel Batram and Lennart Oelschläger who contributed to the codebase used for the calculations.

Table A.1
Ratio of the asymptotic variance for optimal weights versus unweighted case.

s	3	4	5	6	7	8
Variance ratio	0.80	0.66	0.60	0.57	0.55	0.54

Appendix A

In this appendix, we construct an example in which weighting reduces the variance of the estimator by up to 50%. The main idea here is to choose a setting in which pairs of choice decisions are almost perfectly correlated. If this holds for the scores of the pairwise choice probabilities, then for a decider facing s choice occasions, the gradient of the full pairwise CML will be equal to the gradient for one pair times the number $C_s = s(s-1)/2$ of pairs. The Hessian will also be equal to C_s times the Hessian for one pair. Therefore the variance of the score will be equal to $C_s^2 V_1$ (V_1 denoting the variance of the gradient for one pair) and the Hessian equal to $C_s H_1$ (H_1 denoting the Hessian for one pair).

When combining an equal number of deciders with two choice occasions (hence only one pair of choices) and deciders with s choices (where the contribution of these deciders is weighted with a scalar w), we obtain the variance formula

$$V = \frac{1 + w^2 C_s^2}{(1 + wC_s)^2} H_1^{-1} V_1 H_1^{-1}.$$

Without weighting, corresponding to $w = 1$, we obtain a fraction of $(1 + C_s^2)/(1 + C_s)^2$, which tends to 1 for $s \to \infty$ and hence $C_s \to \infty$. Introducing the weight $w = 1/C_s$, however, leads to a fraction $(1 + 1^2)/(1 + 1)^2 = 2/4 = 1/2$. This demonstrates that, for maximal positive correlation, we achieve a variance reduction of 50%.

A situation with a maximal positive correlation is achieved, for example, for a binary decision with no regressors but an ASC for the second choice. The ASC is modelled as $\beta + \gamma, \gamma \sim \mathcal{N}(0, \omega^2)$, where $\beta = 5$ is fixed, and $\omega = 5$ is estimated. Furthermore, we assume that $\varepsilon_{n,t,1} \sim \mathcal{N}(0, 0.01)$. In this model ω is the only parameter estimated.

The noise is negligible compared to the random effect. The choices of a decider almost exclusively are decided via the random ASC. This leads to a model where the random utilities for each decider are almost perfectly positively correlated across choice occasions. It is easy to see that this implies the same for the corresponding gradients.

We computed the ratio of the asymptotic variance of ω with the optimal weighting to the one for the unweighted case for $s = 3, \dots, 8$. Table A.1 shows that the relative variance decreases by almost 50% by using the optimal weights.

Appendix B

In this appendix, we proof the theorems from Section 2. For ease of reading, the theorems are repeated before the respective proofs.

Theorem 1 (*Asymptotic Variance*)**.** Let the data be generated according to Assumption 2 with parameter vector θ_0 and let $\hat{\theta}$ be the CML estimator maximising the weighted CML function (1) using the weights $w_{n,a,b} = w_{T_n} \mathring{w}_{n,a,b}$, where the initial weights $\mathring{w}_{n,a,b}$ adhere to Assumption 1, where $C_s(\mathring{W}) > 0, s = 2, \dots, \overline{S}$.

Further, let $w_s \geq 0$ denote group-specific weights according to the number of observations $T_n = s,\ s \in \{2, \dots, \overline{S}\}$ of decider n, such that $\sum_{s=2}^{\overline{S}} f_s w_s C_s(\mathring{W}) = 1$. Then the following hold:

(I) $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V_\theta(W_s))$, where the asymptotic variance–covariance matrix $V_\theta(W_s)$ for a given vector $W_s = (w_2, \dots, w_{\overline{S}})'$ of group-specific weights has the form

$$V_\theta(W_s) = \sum_{s=2}^{\overline{S}} f_s w_s^2 H_0^{-1} V_s H_0^{-1}, \tag{2}$$

with

$$V_s = \mathbb{E}\ g_n(\mathring{W}) g_n(\mathring{W})', \quad H_0 = \mathbb{E}\ \partial_\theta^2 \log \mathbb{P}(y_{n,1}, y_{n,2}, X_n; \theta_0). \tag{3}$$

(II) H_0 can be estimated consistently as

$$\hat{H}_0 = \frac{1}{\sum_{n,a,b} \mathring{w}_{n,a,b}} \sum_{n,a,b} \mathring{w}_{n,a,b} \partial_\theta^2 \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta}). \tag{4}$$

(III) If $f_s > 0$ and N_s denotes the number of deciders facing s choice occasions, then V_s can be estimated consistently using

$$\hat{V}_s = {N_s}^{-1} \sum_{n: T_n = s} \hat{g}_n(\mathring{W}) \hat{g}_n(\mathring{W})', \tag{5}$$

where $\hat{g}_n(\mathring{W}) := \left(\sum_{a=1}^{s-1} \sum_{b=a+1}^{s} \mathring{w}_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta}) \right)$.

Proof.

(I) The average Hessian $\partial_\theta^2 ll_{CML}(y, X; \bar{\theta}, W_s)/N$ takes the form

$$\partial_\theta^2 ll_{CML}(y, X; \bar{\theta}, W_s)/N = N^{-1} \sum_{n=1}^{N} \left(\sum_{b=a+1}^{T_n} \sum_{a=1}^{T_n-1} w_{n,a,b} \partial_\theta^2 \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \bar{\theta}) \right).$$

Hereby, $\mathbb{E}\left(\partial_\theta^2 \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \theta_0)\right) = H_0$ is independent of a and b if the regressors are drawn from an identical marginal distribution. It follows that

$$\mathbb{E}\partial_\theta^2 ll_{CML}(y, X; \bar{\theta}, W_s)/N = \sum_{s=2}^{\overline{S}} N^{-1} \left(\sum_{n:T_n=s} \left(\sum_{a=1}^{s-1} \sum_{b=a+1}^{s} w_{n,a,b} \right) \right) H_0 \to H(W, f_s),$$

where $f_s = \lim N_s/N, s = 2, \dots, \overline{S}$ denotes the relative frequency of deciders with s choice situations (with a maximum of $\overline{S}$ and a minimum of 2), and $N_s, s = 2, \dots, \overline{S}$ the number of individuals with s choice situations.
Use $w_{n,a,b} = w_s \mathring{w}_{n,a,b}$ as combined weights, and $C_{N,s}(W) = {N_s}^{-1} \sum_{n:T_n=s} \sum_{a=1}^{s-1} \sum_{b=a+1}^{s} w_{n,a,b}$ as the average sum of weights for a pair of observations from an individual with s choice occasions. Then $C_{N,s}(W) \to w_s C_s(\mathring{W})$. This limit for the weights $\mathring{w}_{n,a,b}$ obviously does not depend on n. Otherwise, independent sampling of the stratification weights subject to lower and upper bounds shows that $C_{N,s}(\mathring{W}) \to w_s C_s(\mathring{W})$.
Consequently we obtain

$$H(W, f_s) = \sum_{s=2}^{\overline{S}} f_s w_s C_s(\mathring{W}) H_0 = H_0,$$

by the assumption $\sum_{s=2}^{\overline{S}} f_s w_s C_s(\mathring{W}) = 1$. Furthermore,

$$\partial_\theta^2 ll_{CML}(y, X; \bar{\theta}, W)/N - \mathbb{E}\partial_\theta^2 ll_{CML}(y, X; \bar{\theta}, W)/N \xrightarrow{p} 0$$

follows from independence over deciders and boundedness of the variance implied by the bound on the weights in combination with bounds on the regressors. The choice probabilities depend differentiably to any degree on the underlying parameters and regressors, implying a uniform bound on all moments involved.
With respect to the score $\partial_\theta ll_{CML}(y, X; \theta_0, W)/\sqrt{N}$, note that it is the sum of independent (assuming independent draws over deciders) scores conditional on the number of choice occasions:

$$g_n(W) := \left(\sum_{a=1}^{T_n-1} \sum_{b=a+1}^{T_n} w_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \theta_0) \right),$$

$$\begin{aligned} \sqrt{N} \partial_\theta ll_{CML}(y, X; \theta_0, W) &= \frac{1}{\sqrt{N}} \sum_{s=2}^{\overline{S}} \sum_{n:T_n=s} \left(\sum_{a=1}^{s-1} \sum_{b=a+1}^{s} w_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \theta_0) \right) \\ &= \sum_{s=2}^{\overline{S}} \sqrt{f_s} \left(\frac{1}{\sqrt{f_s N}} \sum_{n:T_n=s} g_n(W) \right). \end{aligned}$$

The inner sum is for each value $s = 2, \dots, \overline{S}$ the sum over independent terms with expectation zero and a variance $V_s(W)$ depending on the number of choice occasions (for iid draws of the regressors over deciders).
Thus the limiting normal distribution of $\sqrt{N} \partial_\theta ll_{CML}(y, X; \theta_0, W)$ has variance (due to the independence of contributions of different deciders)

$$V(W, f_s) = \sum_{s=2}^{\overline{S}} f_s V_s(W).$$

The variance of the estimator then follows the sandwich form

$$V_\theta = H(W, f_s)^{-1} V(W, f_s) H(W, f_s)^{-1} = H_0^{-1} V(W, f_s) H_0^{-1}.$$

Moreover, for n such that $T_n = s$, we have

$$\begin{aligned} g_n(W) &= \sum_{a=1}^{s-1} \sum_{b=a+1}^{s} w_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \theta_0) \\ &= w_s \sum_{a=1}^{s-1} \sum_{b=a+1}^{s} \mathring{w}_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \theta_0) = w_s g_n(\mathring{W}) \end{aligned}$$

and, therefore,

$$V_s(W) = \mathbb{E} g_n(W) g_n(W)' = \mathbb{E} w_s g_n(\mathring{W}) w_s g_n(\mathring{W})' = w_s^2 \mathbb{E} g_n(\mathring{W}) g_n(\mathring{W})' = w_s^2 V_s(\mathring{W}),$$

with $\mathring{W}$ denoting the set of initial weights $\mathring{w}_{n,a,b}$. Then the variance formula simplifies to

$$V_\theta = \sum_{s=2}^{\overline{S}} f_s w_s^2 H_0^{-1} V_s(\mathring{W}) H_0^{-1}.$$

(II) Consistency for the estimator $\hat{H}_0$ follows from the consistency of $\hat{\theta} \to \theta_0$, iid sampling over individuals and the differentiability of the criterion function as a function of the parameter vector. The arguments are standard and, hence, omitted.

(III) The same applies for the estimate $\hat{V}_s$. Here consistency requires $N_s \to \infty$. □

Theorem 2 (*Optimal Group-Specific Weights*)**.** Let $l : \mathbb{R}^{d\times d} \to \mathbb{R}$ be a linear mapping of the form $l(V) = \mathrm{tr}(VA)$, with $A \in \mathbb{R}^{d\times d}$, $A \neq 0$ symmetric and positive semidefinite, $\hat{\theta}$ be the CML estimator maximising the weighted CML function (1), and let the data be generated subject to Assumption 2. Let $\mathring{w}_{n,a,b}$ be the initial weights fulfilling Assumption 1, where $\sum_{s=2}^{\overline{S}} f_s w_s C_s(\mathring{W}) = 1$.

Then $l(V_\theta(W_s))$ is minimised over W_s by

$$w_s^* = \left(\sum_{s=2}^{\overline{S}} f_s C_s(\mathring{W})^2 / v_s(\mathring{W}) \right)^{-1} C_s(\mathring{W})/v_s(\mathring{W}) \propto C_s(\mathring{W})/v_s(\mathring{W}), \tag{6}$$

with $v_s(\mathring{W}) = l\left(H_0^{-1} V_s(\mathring{W}) H_0^{-1}\right)$.

Proof. Due to the linearity of the function $l(\cdot)$, we obtain

$$l\left(V_\theta(W)\right) = l\left(\sum_{s=2}^{\overline{S}} f_s w_s^2 H_0^{-1} V_s H_0^{-1} \right) = \sum_{s=2}^{\overline{S}} f_s w_s^2 l\left(H_0^{-1} V_s H_0^{-1}\right) = \sum_{s=2}^{\overline{S}} f_s w_s^2 v_s(\mathring{W}).$$

It is also ensured that $v_s(\mathring{W}) = l\left(H_0^{-1} V_s(\mathring{W}) H_0^{-1}\right) = \mathrm{tr}\left(H_0^{-1} V_s H_0^{-1} A\right)$ is positive since $H_0^{-1} V_s H_0^{-1}$ is positive definite and A is, by requirement, positive semidefinite.

The Lagrange function for optimising with respect to w_s subject to the restrictions then equals

$$\mathcal{L}(W, \lambda) = \sum_{s=2}^{\overline{S}} f_s w_s^2 v_s(\mathring{W}) - \lambda \left(\sum_{s=2}^{\overline{S}} f_s w_s C_s(\mathring{W}) - 1 \right).$$

The first order condition for w_s evaluated at the optimum then reads:

$$2 f_s w_s^* v_s(\mathring{W}) - \lambda f_s C_s(\mathring{W}) = 0 \quad \Longrightarrow \quad w_s^* = \lambda \frac{f_s C_s(\mathring{W})}{2 f_s v_s(\mathring{W})} = \frac{\lambda}{2} \frac{C_s(\mathring{W})}{v_s(\mathring{W})}.$$

This implies

$$w_s^* = c C_s(\mathring{W})/v_s(\mathring{W}),$$

with the constant $c = \left(\sum_{s=2}^{\overline{S}} f_s C_s(\mathring{W})^2 / v_s(\mathring{W}) \right)^{-1}$. □

Appendix C

In this appendix, algorithms to calculate the different versions of the optimal group weights are shown.

Algorithm 1 Algorithm to calculate two-step optimal group-weight CML estimator

Require: initial weights $\mathring{W} \geq 0$, linear functional $l(\cdot) = \mathrm{tr}(\cdot A)$, $A \in \mathbb{R}^{d\times d}$ symmetric and positive semidefinite

1: Set $w_s \leftarrow 1,\ s = 2, \ldots, \overline{S}$
2: Estimate $\hat{\theta}$, minimising $ll_{\mathrm{CML}}(y, X; \theta, \mathring{W})$
3: Estimate $\hat{H}_0 = \frac{1}{\sum_{n,a,b} \mathring{w}_{n,a,b}} \sum_{n,a,b} \mathring{w}_{n,a,b} \partial_\theta^2 \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta})$
4: Calculate $C_{N_s,s}(\mathring{W}) = N_s^{-1} \sum_{n:T_n=s} \sum_{b=a+1}^{s} \sum_{a=1}^{s-1} \mathring{w}_{n,a,b},\ s = 2, \ldots, \overline{S}$
5: Estimate $\hat{g}_n(\mathring{W}) = \sum_{a=1}^{s-1} \sum_{b=a+1}^{s} \mathring{w}_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta}),\ n = 1, \ldots, N$
6: Estimate $\hat{V}_s = N_s^{-1} \sum_{n:T_n=s} \hat{g}_n(\mathring{W}) \hat{g}_n(\mathring{W})',\ s = 2, \ldots, \overline{S}$
7: Estimate 'optimal' group-weights $\hat{w}_s^* = C_{N_s,s}(\mathring{W})/l(\hat{H}_0^{-1} \hat{V}_s \hat{H}_0^{-1}),\ s = 2, \ldots, \overline{S}$
8: Calculate new weights $\hat{W}^*$ with $\mathring{w}_{n,a,b}^* = \mathring{w}_s^* \mathring{w}_{n,a,b}$ for $T_n = s$
9: Estimate $\tilde{\theta}$, minimising $ll_{\mathrm{CML}}(y, X; \theta, \hat{W}^*)$

Table E.2
Share of `nlm()` exit codes by employed weighting scheme for 100 unbalanced panel data set simulations.

Weighting type	Code 1	Code 2	Code 3	Code 4	Code 5	Code ≤ 2
equal	0.77	0.00	0	0	0.23	0.77
BB	0.91	0.08	0	0	0.01	0.99
BB_param	0.85	0.15	0	0	0.00	1.00
JL_0.5	0.80	0.00	0	0	0.20	0.80

Table E.3
Overview of distribution of quotients between the traces between differently weighted models. The $x\%$ percentile is denoted as Q_x. In each column, the value of the method showing the largest reduction compared to the equally weighted model is highlighted.

Trace quotient	Mean	Min	Q_05	Q_25	Median	Q_75	Q_95	Max	% < 1
BB/equal	0.956	0.090	**0.292**	**0.842**	**0.926**	**0.997**	1.335	4.470	**75.325**
BB/JL_0.5	1.090	0.032	0.348	0.778	0.882	0.975	1.351	13.988	81.818
BB_param/equal	**0.948**	0.046	0.599	0.874	0.938	1.000	**1.250**	**1.749**	72.727
BB_param/JL_0.5	0.909	0.243	0.457	0.806	0.900	0.952	1.135	3.682	83.117
BB_param/BB	1.299	0.060	0.768	0.971	1.020	1.060	1.535	12.250	38.961
JL_0.5/equal	1.230	**0.039**	0.543	0.903	1.044	1.273	2.217	6.909	41.558

Table E.4
Distribution of average ratios between the optimal group weights depending on different initial weights $\mathring{W} = 1$ and $\mathring{W} = \text{JL\_0.5}$ for 100 unbalanced panel simulations.

Weight quotient	Min	Q_05	Q_25	Median	Q_75	Q_95	Max
BB_fJL/BB	0.614	0.938	0.996	1.016	1.045	1.740	4.471
BB_fJL_param/BB_param	0.959	0.976	0.997	1.001	1.006	1.094	4.420

Algorithm 2 Algorithm to calculate two-step optimal group-weight CML estimator with parametric variation

Require: initial weights $\mathring{W} \geq 0$, linear functional $l(\cdot) = \text{tr}(\cdot A)$, $A \in \mathbb{R}^{d\times d}$ symmetric and positive semidefinite

1: Set $w_s \leftarrow 1,\ s = 2, \dots, \overline{S}$
2: Estimate $\hat{\theta}$, minimising $ll_{\text{CML}}(y, X; \theta, \mathring{W})$
3: Estimate $\hat{H}_0 = \frac{1}{\sum_{n,a,b} \mathring{w}_{n,a,b}} \sum_{n,a,b} \mathring{w}_{n,a,b} \partial^2_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta})$
4: Calculate $C_{N_s,s}(\mathring{W}) = {N_s}^{-1} \sum_{n:T_n=s} \sum_{b=a+1}^{s} \sum_{a=1}^{s-1} \mathring{w}_{n,a,b},\ s = 2, \dots, \overline{S}$
5: Estimate $\hat{g}_n(\mathring{W}) = \sum_{a=1}^{s-1} \sum_{b=a+1}^{s} \mathring{w}_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta}),\ n = 1, \dots, N$
6: Estimate $\hat{V}_s = {N_s}^{-1} \sum_{n:T_n=s} \hat{g}_n(\mathring{W}) \hat{g}_n(\mathring{W})',\ s = 2, \dots, \overline{S}$
7: Calculate $\hat{v}_s = l(\hat{H}_0^{-1} \hat{V}_s \hat{H}_0^{-1})$
8: Estimate parametric function $g : \{2, \dots, \overline{S}\} \to \mathbb{R}_+$ to model $C_{N_s,s}(\mathring{W})/\hat{v}_s = g(s)$ and calculate $\breve{w}^*_s = g(s)$
9: Calculate new weights $\breve{W}^*$ with $\breve{w}^*_{n,a,b} = \breve{w}^*_s \mathring{w}_{n,a,b}$ for $T_n = s$
10: Estimate $\tilde{\theta}$, minimising $ll_{\text{CML}}(y, X; \theta, \breve{W}^*)$

Appendix D

For the MACML estimations, the computations were done in the statistical computing language R (Version 4.1.2) (R Core Team, 2021) with the CML function calculations written in C++11, integrated in R via the *Rcpp* package by Eddelbuettel and Francois (2011). The negative CML function is then minimised using the R function `nlm()` with analytic gradients. The variance–covariance matrix $\hat{V}_\theta$ is calculated using the analytic Hessian matrix of the CML function.

The R and C++ code used for the estimation process is bundled into an R-package named *Rprobit* and is available on https://github.com/dbauer72/Rprobit.

Appendix E

See Tables E.2–E.6.

Table E.5
Distribution of ratios between the traces of the variance–covariance matrix of the estimator depending on different initial weights $\hat{W} = 1$ and $\hat{W} = \text{JL\_0.5}$ for 100 unbalanced panel simulations.

Trace quotient	Min	Q_05	Q_25	Median	Q_75	Q_95	Max
BB_fJL/BB	0	0.445	0.988	1.004	1.018	1.327	2.895
BB_fJL_param/BB_param	0	0.836	0.996	1.001	1.013	1.268	239.023

Table E.6
Overview of distribution of relative mean squared l_2 distance of predicted choice probabilities to true choice probabilities, as quotient with mean squared l_2 distance of equally weighted model. The $x\%$ percentile is denoted as Q_x. In each column, the value of the method showing the largest reduction compared to the equally weighted model is highlighted.

Weighting type	Mean	Min	Q_05	Q_25	Median	Q_75	Q_95	Max	% improved
BB/equal	0.981	0.672	0.720	0.841	**0.946**	1.054	1.393	2.193	66.2
BB_param/equal	**0.940**	0.581	0.732	0.865	0.957	**1.014**	**1.161**	**1.329**	**70.1**
JL_0.5/equal	0.971	**0.523**	**0.605**	**0.840**	0.949	1.093	1.366	1.891	57.1

References

Bansal, P., Keshavarzzadeh, V., Guevara, A., Li, S., Daziano, R.A., 2022. Designed quadrature to approximate integrals in maximum simulated likelihood estimation. Econom. J. 25, 301–321. http://dx.doi.org/10.1093/ectj/utab023.

Bhat, C.R., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. Transp. Res. B 37, 837–855. http://dx.doi.org/10.1016/S0191-2615(02)00090-5, URL: https://linkinghub.elsevier.com/retrieve/pii/S0191261502000905.

Bhat, C.R., 2011. The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. Transp. Res. B 45 (7), 923–939. http://dx.doi.org/10.1016/j.trb.2011.04.005, URL: https://www.sciencedirect.com/science/article/pii/S019126151100049X.

Bhat, C.R., 2014. The composite marginal likelihood (CML) inference approach with applications to discrete and mixed dependent variable models. Found. Trends Econom. 7 (1), 1–117. http://dx.doi.org/10.1561/0800000022.

Cessie, S.L., Houwelingen, J.C.V., 1994. Logistic regression for correlated binary data. Appl. Stat. 43 (1), 95. http://dx.doi.org/10.2307/2986114, URL: https://www.jstor.org/stable/2986114?origin=crossref.

Clarivate, 2022. Number of publications on the topic of "composite likelihood" by year. URL: https://www.webofscience.com/wos/woscc/analyze-results/6de73426-9e3f-4ae3-90dd-48c827eee91d-6456e209.

Cox, D.R., Reid, N., 2004. A note on pseudolikelihood constructed from marginal densities. Biometrika 91 (3), 729–737. http://dx.doi.org/10.1093/biomet/91.3.729, URL: http://www.jstor.org/stable/20441134.

Crastes, R., Daly, A., Palma, D., Holz-rau, C., 2020. Weighting strategies for modelling life course history events via pairwise composite marginal likelihood. URL: https://www.stephanehess.me.uk/papers/working_papers/Crastes_et_al_2.pdf.

Dick, J., Gantner, R.N., Gia, Q.T.L., Schwab, C., 2017. Multilevel higher-order quasi-Monte Carlo Bayesian estimation. Math. Models Methods Appl. Sci. 27, 953–995. http://dx.doi.org/10.1142/S021820251750021X.

Eddelbuettel, D., Francois, R., 2011. Rcpp: Seamless R and C++ integration. J. Stat. Softw. 40, http://dx.doi.org/10.18637/jss.v040.i08, URL: https://www.jstatsoft.org/index.php/jss/article/view/v040i08.

Heiss, F., Winschel, V., 2008. Likelihood approximation by numerical integration on sparse grids. J. Econometrics 144, 62–80. http://dx.doi.org/10.1016/j.jeconom.2007.12.004, URL: https://linkinghub.elsevier.com/retrieve/pii/S0304407607002552.

Hess, S., Train, K.E., Polak, J.W., 2006. On the use of a Modified Latin Hypercube Sampling (MLHS) method in the estimation of a Mixed Logit Model for vehicle choice. Transp. Res. B 40 (2), 147–163. http://dx.doi.org/10.1016/j.trb.2004.10.005, URL: https://linkinghub.elsevier.com/retrieve/pii/S0191261505000500.

Joe, H., 1995. Approximations to multivariate normal rectangle probabilities based on conditional expectations. J. Amer. Statist. Assoc. 90 (431), 957–964. http://dx.doi.org/10.2307/2291331.

Joe, H., Lee, Y., 2009. On weighting of bivariate margins in pairwise likelihood. J. Multivariate Anal. 100 (4), 670–685. http://dx.doi.org/10.1016/j.jmva.2008.07.004.

Johnson, S.G., 2022. The NLopt nonlinear-optimization package. URL: https://nlopt.readthedocs.io/en/latest/.

Keshavarzzadeh, V., Kirby, R.M., Narayan, A., 2018. Numerical integration in multiple dimensions with designed quadrature. SIAM J. Sci. Comput. 40, A2033–A2061. http://dx.doi.org/10.1137/17M1137875.

Kessels, R., Goos, P., Vandebroek, M., 2006. A comparison of criteria to design efficient choice experiments. J. Mar. Res. 43, 409–419. http://dx.doi.org/10.1509/jmkr.43.3.409.

Kuk, A.Y., Nott, D.J., 2000. A pairwise likelihood approach to analyzing correlated binary data. Statist. Probab. Lett. 47 (4), 329–335. http://dx.doi.org/10.1016/S0167-7152(99)00174-1, URL: https://linkinghub.elsevier.com/retrieve/pii/S0167715299001741.

Lindsay, B.G., Yi, G.Y., Sun, J., 2011. Issues and strategies in the selection of composite likelihoods. Statist. Sinica 21 (1), 71–105, URL: http://www.jstor.org/stable/24309263.

Lütkepohl, H., 2005. New Introduction To Multiple Time Series Analysis. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–764. http://dx.doi.org/10.1007/978-3-540-27752-1, URL: http://link.springer.com/10.1007/978-3-540-27752-1.

Muthén, B., 1984. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika 49, 115–132. http://dx.doi.org/10.1007/BF02294210, URL: http://link.springer.com/10.1007/BF02294210.

Pedeli, X., Varin, C., 2020. Pairwise likelihood estimation of latent autoregressive count models. Stat. Methods Med. Res. 29 (11), 3278–3293. http://dx.doi.org/10.1177/0962280220924068, arXiv:1805.10865.

Powell, M.J.D., 1994. A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation. Springer Netherlands, Dordrecht, pp. 51–67. http://dx.doi.org/10.1007/978-94-015-8330-5_4, URL: http://link.springer.com/10.1007/978-94-015-8330-5_4.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL: https://www.R-project.org/.

Ryu, E.K., Boyd, S.P., 2015. Extensions of Gauss quadrature via linear programming. Found. Comput. Math. 15, 953–971. http://dx.doi.org/10.1007/s10208-014-9197-9.

Solow, A.R., 1990. A method for approximating multivariate normal orthant probabilities. J. Stat. Comput. Simul. 37 (3–4), 225–229. http://dx.doi.org/10.1080/00949659008811306.

Train, K.E., 2009. Discrete Choice Methods with Simulation, second ed. Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511805271.

Varin, C., 2008. On composite marginal likelihoods. AStA Adv. Stat. Anal. 92 (1), 1–28. http://dx.doi.org/10.1007/s10182-008-0060-7, URL: http://link.springer.com/10.1007/s10182-008-0060-7.

Varin, C., Reid, N., Firth, D., 2011. An overview of composite likelihood methods. Statist. Sinica 21, 5–42, URL: http://www.jstor.org/stable/24309261.

Varin, C., Vidoni, P., 2005. A note on composite likelihood inference and model selection. Biometrika 92 (3), 519–528. http://dx.doi.org/10.1093/biomet/92.3.519, URL: http://www.jstor.org/stable/20441211.

Varin, C., Vidoni, P., 2006. Pairwise likelihood inference for ordinal categorical time series. Comput. Statist. Data Anal. 51, 2365–2373. http://dx.doi.org/10.1016/j.csda.2006.09.009.

Wooldridge, J.M., 2015. Introductory Econometrics: A Modern Approach, fifth ed. Nelson Education.

Zumkeller, D., Chlond, B., 2009. Dynamics of change: Fifteen-Year German mobility panel. In: Transportation Research Board 88th Annual Meeting Compendium of Papers. Transportation Research Board, Washington DC, United States, URL: https://trid.trb.org/view/880678.

Zumkeller, D., Chlond, B., Lipps, O., 1999. Das mobilitäts-panel (MOP) - konzept und realisierung einer bundesweiten längsschnittbeobachtung. In: Hautzinger, H. (Ed.), Schriftenreihe der Deutschen Verkehrswissenschaftlichen Gesellschaft / B. Vol. 217, DVWG, Bergisch Gladbach, pp. 33–72, URL: https://madoc.bib.uni-mannheim.de/12149/.