

Tuzcuoglu, Kerem

**Working Paper**

## Composite likelihood estimation of an autoregressive panel probit model with random effects

Bank of Canada Staff Working Paper, No. 2019-16

**Provided in Cooperation with:**

Bank of Canada, Ottawa

*Suggested Citation:* Tuzcuoglu, Kerem (2019) : Composite likelihood estimation of an autoregressive panel probit model with random effects, Bank of Canada Staff Working Paper, No. 2019-16, Bank of Canada, Ottawa,  
<https://doi.org/10.34989/swp-2019-16>

This Version is available at:

<https://hdl.handle.net/10419/197930>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*



Staff Working Paper/Document de travail du personnel 2019-16

# Composite Likelihood Estimation of an Autoregressive Panel Probit Model with Random Effects



by Kerem Tuzcuoglu

Bank of Canada staff working papers provide a forum for staff to publish work-in-progress research independently from the Bank's Governing Council. This research may support or challenge prevailing policy orthodoxy. Therefore, the views expressed in this paper are solely those of the authors and may differ from official Bank of Canada views. No responsibility for them should be attributed to the Bank.

Bank of Canada Staff Working Paper 2019-16

May 2019

# **Composite Likelihood Estimation of an Autoregressive Panel Probit Model with Random Effects**

by

**Kerem Tuzcuoglu**

Financial Stability Department  
Bank of Canada  
Ottawa, Ontario, Canada K1A 0G9  
[ktuzcuoglu@bankofcanada.ca](mailto:ktuzcuoglu@bankofcanada.ca)

## **Acknowledgements**

This article was one of the chapters of my doctoral dissertation titled “Three Essays in Econometrics” (see Tuzcuoglu (2017)). I would like to thank Serena Ng, Jushan Bai, Bernard Salanié, Aysun Alp, Victor Aguirregabiria, Sokbae Lee, Jean-Jacques Forneron, Christoph Rothe, Jeffrey Wooldridge, Fabio Canova and Cristiano Varin for their helpful comments and suggestions. This work further benefited from discussions and comments by participants at the NBER-NSF Time Series conference in San Diego, International Panel Data conference in Seoul, Asian Meeting of Econometric Society in Seoul, International Association of Applied Econometrics in Montreal, Canadian Econometric Study Group in Toronto, and seminars at Columbia University and the Bank of Canada. All errors are, of course, my own. The views expressed in this paper are mine and do not necessarily reflect the position of the Bank of Canada.

## Abstract

Modeling and estimating persistent discrete data can be challenging. In this paper, we use an autoregressive panel probit model where the autocorrelation in the discrete variable is driven by the autocorrelation in the latent variable. In such a non-linear model, the autocorrelation in an unobserved variable results in an intractable likelihood containing high-dimensional integrals. To tackle this problem, we use composite likelihoods that involve much lower order of integration. However, parameter identification becomes problematic since the information employed in lower dimensional distributions may not be rich enough for identification. Therefore, we characterize types of composite likelihoods that are valid for this model and study conditions under which the parameters can be identified. Moreover, we provide consistency and asymptotic normality results of the pairwise composite likelihood estimator and conduct Monte Carlo simulations to assess its finite-sample performances. Finally, we apply our method to analyze credit ratings. The results indicate a significant improvement in the estimated transition probabilities between rating classes compared with static models.

*Bank topics: Econometric and statistical methods; Economic models; Credit risk management*

*JEL codes: C23, C25, C58, G24*

## Résumé

La modélisation et l'estimation de données discrètes persistantes peuvent s'avérer difficiles. Dans cette étude, nous utilisons un modèle probit autorégressif avec données de panel où l'autocorrélation de la variable discrète dépend de l'autocorrélation de la variable latente. Dans ce type de modèle non linéaire, l'autocorrélation d'une variable non observée entraîne une vraisemblance incalculable contenant des intégrales de haute dimension. Pour résoudre ce problème, nous utilisons des vraisemblances composites où l'ordre d'intégration est nettement inférieur. Cependant, l'identification des paramètres devient problématique, car l'information employée dans les distributions de dimension plus faible peut ne pas être suffisante. Nous caractérisons donc les types de vraisemblances composites valides pour ce modèle et étudions les conditions dans lesquelles les paramètres peuvent être identifiés. De plus, nous démontrons la convergence et la normalité asymptotique de l'estimateur de la vraisemblance composite par paire, et effectuons des simulations de Monte-Carlo pour évaluer les résultats obtenus avec des échantillons finis. Enfin, nous appliquons notre méthode à l'analyse de notes de crédit. Les résultats indiquent une nette amélioration de l'exactitude des probabilités de transition estimées entre les catégories de notes, par rapport aux modèles statiques.

*Sujets : Méthodes économétriques et statistiques; Modèles économiques; Gestion du risque de crédit*

*Codes JEL : C23, C25, C58, G24*

## Non-technical Summary

We frequently encounter discrete variables both in our daily lives and in the economy. For instance, the choices we make in supermarkets, our school grades, our marital status, as well as some economic and financial variables such as our employment status, the state of the economy (recession or expansion), credit ratings, investment decisions, and changes in the target interest rate are all in discrete terms. Some of these discrete variables are very repetitive and do not change frequently. We tend to buy the same brand of milk, our marital or employment status changes rarely, a good (bad) student consistently gets high (low) grades. Similarly, if an economy starts expanding then it keeps expanding for some years, and if an economy enters into a recession then it tends to stay in the recession for some time, and credit ratings of firms and countries do not change frequently, etc.

Probit models are commonly used in the literature to model and predict discrete variables. These models estimate the probability of a discrete variable being in a specific category (as well as the transition probabilities to other categories). However, most of these models in the literature ignore the persistence in the discrete variable and treat the observations in each time period as if they were independent from each other.

In this paper, we focus on a particular type of probit model that takes into account the persistence of the discrete variable. We call this model *autoregressive probit* model, where the autoregressive part refers to the time dependence. However, the non-linear nature of discrete variables as well as their dependence over time render the estimation of this model extremely complex. For this reason, we borrow composite likelihood estimation methods from the statistics literature to facilitate the complex estimation. Then, we provide theoretical sufficient conditions for when the composite likelihood estimation method can be used for this model. Moreover, we prove certain theoretical properties of the estimator and assess its performances via Monte Carlo simulations. The simulation results show that the proposed estimator is not only extremely fast but also accurate and robust. Finally, we apply our methodology to analyze corporate bond ratings and empirically show that taking the persistence into account yields significant improvements in the estimated transition probabilities between rating classes.

# 1 Introduction

Persistent discrete variables are extensively used in both the economics and finance literature. Credit ratings, changes in the federal funds target rate, binary indicators representing recessionary periods, unemployment status, and school grades are just a few important examples. These variables have a fair amount of persistence in them: credit ratings of companies or countries change slowly (Altman and Kao (1992) and Nickell et al. (2000)); the policy rate is usually adjusted gradually by central banks, which is known as monetary policy inertia (Woodford (1999)); an economy in a recession (expansion) tends to stay in the same state (Bernanke et al. (1999) and Kiyotaki and Moore (1997)); unemployment status changes infrequently (Blanchard and Summers (1986)); a good (bad) student consistently gets high (low) grades. To understand the dynamic nature of these binary or categorical variables, one needs to take both persistence and discreteness into consideration.

Modeling and estimating persistent discrete data can be challenging. Incorporating time series concepts – to capture the persistence – into the non-linear nature of discrete data might need complex models that are hard to estimate. An interesting model that embodies both the discreteness and persistence is an autoregressive (AR) panel probit model.<sup>1</sup> This model is introduced by Heckman (1981) as a special case of a general discrete choice model that contains a variety of other probit models.<sup>2</sup> Mathematically, we present the main model of this paper in the following way. For  $i = 1, \dots, N$  and  $t = 1, \dots, T$ , we have

$$y_{it} = s \quad \text{if} \quad \tau_{s-1} \leq y_{it}^* < \tau_s, \quad (1)$$

$$y_{it}^* = \rho y_{i,t-1}^* + \beta' x_{it} + \alpha_i + \varepsilon_{it}, \quad (2)$$

where  $y_{it}$  is the observed discrete variable taking values from the set  $\{1, \dots, S\}$ ,  $y_{it}^*$  is the autoregressive latent variable that governs the underlying continuous process which depends on observable covariates  $x_{it}$ , unobservable individual-specific effects  $\alpha_i$  as well as idiosyncratic

---

<sup>1</sup>In fact, the main model of the paper is ordered probit. But to keep the name of the model simple, we omit the word *ordered*.

<sup>2</sup>The general framework introduced by Heckman (1981) is as follows:  $y_{it} = \mathbb{1}(y_{it}^* > 0)$  where  $y_{it}^* = \beta' x_{it} + \sum_{l=1}^{\infty} \gamma_l y_{i,t-l} + \lambda \sum_{s=1}^{\infty} \sum_{l=1}^s y_{i,t-l} + \sum_{l=1}^{\infty} \delta_l y_{i,t-l}^* + v_{it}$ . The first term captures the effect of the covariates  $x_{it}$ ; the second and third terms capture the effect of the history of the discrete outcomes; the fourth term represents the effect of the history of the underlying latent process.

errors  $\varepsilon_{it}$ , and unknown parameters  $\{\rho, \beta, \tau_0, \dots, \tau_S\}$ . The autocorrelation of the latent variable drives non-linearly that of the discrete variable. Note that in this model, lagged latent variable  $y_{i,t-1}^*$  generates the persistence as opposed to the lagged observed variable  $y_{i,t-1}$ . The former is an autoregressive process that is more akin to typical time series models whereas the latter generates discrete jumps in the underlying process. The former model is named as *habit persistence* and the latter as *state dependence* by Heckman (1981).

One caveat of the AR panel probit model is that it has a complex likelihood function. The model generates non-Markovian discrete outcome  $y_{it}$  even though the underlying latent variable  $y_{it}^*$  is Markovian. Hence, the likelihood is intractable due to  $T$ -dimensional integrals, which arise from integrating the autocorrelated latent variable out. To deal with such a challenging likelihood, we borrow composite likelihood (CL) estimation methods from statistics literature (proposed by Lindsay (1988)). CL methods have been widely used in the statistics field and have started to gain substantial attention in economics.<sup>3</sup> CL estimation is a likelihood-based method that uses partial specifications of full likelihood (see Varin et al. (2011) and Cox and Reid (2004) for a review).

The most common CL functions are the marginal composite likelihood  $L_{MCL}(\mathbf{y}_i|\mathbf{x}_i, \alpha_i; \theta) = \prod_{t=1}^T l(y_{it}|\mathbf{x}_i, \alpha_i; \theta)$  that contains a single integral, and the pairwise composite likelihood of adjacent observations  $L_{PCL}(\mathbf{y}_i|\mathbf{x}_i, \alpha_i; \theta) = \prod_{t=1}^{T-J} \prod_{j=1}^J l(y_{it}, y_{i,t+j}|\mathbf{x}_i, \alpha_i; \theta)$  that contains two-dimensional integrals, where  $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})'$  and  $\mathbf{x}_i = (x'_{i1}, \dots, x'_{iT})'$ . Hence, composite likelihood reduces the complexity of the full likelihood by focusing on sub-likelihoods. In this sense, CL is similar to pseudo likelihoods, quasi likelihoods, or partial likelihoods proposed by several authors, such as Besag (1974), Cox (1975), and Gouriéroux et al. (1984). CL becomes very useful especially in cases where writing or computing the full likelihood is infeasible, yet marginal or conditional likelihoods are easier to formulate. Compared with the traditional maximum likelihood estimator, the CL method is statistically less efficient. However, consistency, asymptotic normality, and significantly faster computation are among

---

<sup>3</sup>Some examples among a few in the economics literature that use CL estimation are Pakel, Shephard, and Sheppard (2011) and Pakel, Shephard, Sheppard, and Engle (2017) for multivariate GARCH models; Qu (2018) and Canova and Matthes (2016) for DSGE models; Varin and Vidoni (2008) for general state space models; Oh and Patton (2016) for high-dimensional copulas; Chu (2017) for panel error-correction models with spatial dependence; Bel et al. (2018) for a multivariate logit model; Chan et al. (2018) for large dimensional VAR models; and Gouriéroux and Monfort (2018) for composite indirect inference methods.



the appealing properties of the CL estimator. Moreover, it can be more robust to model misspecification compared with ML estimation or simulation methods, since one needs only correctly specified sub-models in CL approach whereas MLE requires a fully correct model (Varin and Vidoni (2008)).

Even though a CL is not the correct likelihood, it can still provide consistent estimation since each sub-likelihood used to form the CL is correctly specified. The intuition is similar to the consistency of the method of moments estimation. Similar to the fact that correctly specified moments can yield consistent method of moments estimation, correctly specified sub-likelihoods can yield consistent composite likelihood estimation. But what is not immediately clear is which types of composite likelihoods provide sufficient information for the identification of parameters.

A simple example can show the importance of the choice of composite likelihoods for parameter identification. Consider a linear AR(1) model  $z_t = \delta z_{t-1} + \sigma e_t$  for an observed continuous variable  $z_t$  and an unobserved error  $e_t \sim \mathcal{N}(0, 1)$ . If we build a composite likelihood based on marginal distributions  $f(z_t|\theta) = \mathcal{N}(0, \sigma^2/(1 - \delta^2))$ , then we cannot identify  $(\delta, \sigma)$  separately. However, using conditional distributions  $f(z_t|z_{t-1}; \theta) = \mathcal{N}(\delta z_{t-1}, \sigma^2)$  enables us to identify the parameters. Even in such a simple example, the choice of the composite likelihood matters dramatically in terms of identification. In more complex models, it is not clear, in general, which sub-likelihoods should be included in the CL so that one can identify all of the parameters.

A contribution of this paper is to show that there exist such composite likelihood functions, involving very low order of integration, in which one can identify all parameters of the AR panel probit with correlated random effects model given in (1) and (2). To identify the parameters, we borrow ideas from the average derivative literature instead of showing the negative definiteness of the Hessian matrix because this is an intricate task in dynamic non-linear models (Honoré and Tamer (2006)). Due to the complexity of these models, the identification is implicitly assumed in many non-linear panel data models both in the economics and composite likelihood literature. In this paper, we provide sufficient conditions for parameter identification and characterize composite likelihoods that can consistently estimate the parameters. Subsequently, we prove the consistency and asymptotic normality

for the (pairwise) composite likelihood estimator in this specific model.

For the asymptotic results, we do not need stronger moment conditions than required in a static panel probit model. The only stronger condition required in the AR panel probit compared with the static model is the strict exogeneity of the covariates – which is a typical assumption in non-linear dynamic latent models (see Section 2 for details). After the asymptotic results, we conduct several Monte Carlo simulations to assess the performance of the composite likelihood estimator. We show that the CL estimator has satisfactory finite-sample properties. Moreover, even though an increase in  $T$  could be bad news for simulation-based estimators of the full likelihood – due to the increase in the dimension of integrals – the performance of CLE improves as  $T$  increases. The increased time-dimension does not increase the complexity of the composite likelihoods while it increases the data points (and pairs) to infer from.

This paper also contributes to the finance literature. One branch of this literature analyzes corporate bond ratings based on firm characteristics by using static panel (probit/logit) models (see, for instance, [Altman \(1968\)](#), [Blume et al. \(1998\)](#), [Amato and Furfine \(2004\)](#), and [Alp \(2013\)](#)). Another branch uses dynamic and more sophisticated models (such as panel probit models with latent dynamic factors) at the expense of not using firm-level data, i.e. not controlling for heterogeneity in firms. In these models, all the firms are treated as identical as long as they belong to the same rating class. Therefore, they are called portfolio models where the dynamics of the rating cohorts are modeled. The contribution of this paper to the finance literature is to offer an easy-to-estimate dynamic model while controlling for firm-level observed and unobserved heterogeneity.

In the empirical part, the credit ratings of individual firms are explained by their balance sheet ratios as well as unobserved firm characteristics and macroeconomic business cycle conditions. The results indicate that the AR panel probit model yields a better fit to the data compared with static probit models, which are the common choice of the literature. More importantly, controlling for the autocorrelation, the AR panel probit model provides significantly improved rating transition dynamics than its static counterpart.

CL estimation is not the only estimation method that can be used for the AR panel probit model with random effects. The method of moments, simulation, and Bayesian techniques

have also been used for this model (see [Grether and Maddala \(1982\)](#), [Lechner \(1993\)](#), [Lee \(1997\)](#), [Pudney \(2008\)](#), and [Stegmueller \(2013\)](#)). The first two papers, [Grether and Maddala \(1982\)](#) and [Lechner \(1993\)](#), attempt to estimate the model by treating the panel data model as if it was  $T$ -many separate cross-sectional probit specifications, where each model, for a given  $t = 1, \dots, T$ , includes all the lags of  $x_{it}$  that are considered as different covariates. This results in  $(K + 1)T(T + 1)/2$  estimated reduced-form parameters to recover the underlying  $K + 2$  parameters, where  $K$  is the number of covariates. Finally, the estimation is conducted via minimum distance estimator – a procedure proposed by [Chamberlain \(1982\)](#) and [Chamberlain \(1984\)](#) for panel data models. In terms of simulation-based techniques, [Lee \(1997\)](#) and [Pudney \(2008\)](#) utilize Simulated Maximum Likelihood (SML) and [Stegmueller \(2013\)](#) uses Monte Carlo Markov Chain (MCMC) methods to estimate the model. However, CL methods have speed and robustness advantages over these estimation techniques (see [Varin and Vidoni \(2006\)](#), [Bhat et al. \(2010\)](#), and [Patil et al. \(2017\)](#)).

The estimation procedure of [Grether and Maddala \(1982\)](#) and [Lechner \(1993\)](#) can be fit under a generalized method of moments (GMM) framework ([Hansen \(1982\)](#)) where one can indeed estimate the panel data model as a whole instead of relying on  $T$ -many separate estimations. However, as in many dynamic panel data settings, the moment conditions can be of order  $T^2$  (see, for instance, [Arellano and Bond \(1991\)](#), [Blundell and Bond \(1998\)](#), [Bertschek and Lechner \(1998\)](#), [Alvarez and Arellano \(2003\)](#), [Greene \(2004b\)](#), and [Semykina and Wooldridge \(2013\)](#)).<sup>4</sup> Even for moderate levels of  $T$ , one needs to invert at least a few hundred dimensional matrices to obtain the GMM estimates (e.g. for  $T = 10$  and  $K = 5$ , the number of moments in the AR panel probit model is around 300). This makes the GMM estimation unstable and even renders it impractical in many cases (see [Roodman \(2009\)](#) for further discussion).

Even though choosing orthogonality conditions for GMM is intuitively similar to choosing sub-likelihoods for CLE, their performances in the AR panel probit model differ significantly. To the best of our knowledge, this is the first paper that shows how GMM estimation can be used to estimate the AR panel probit model and compares its performance to the CLE.

---

<sup>4</sup>One can reduce the number of moment conditions to the order of  $T$  by including only contemporaneous orthogonality conditions. However, selecting the most informative moments, in general, is not an easy task (see [Andrews \(1999\)](#) and [Arellano \(2016\)](#) for some optimality conditions).

Based on the simulation results, we find that GMM and CLE yield comparable finite sample performance for very short panels (e.g. when  $T \leq 5$ ). However, for larger  $T$ , composite likelihood methods clearly outperform GMM.<sup>5</sup>

There are also simulation-based methods that can be used to estimate models with complex likelihood functions containing integrals.<sup>6</sup> One of the most frequently used simulators in Limited Dependent Variable (LDV) models is the GHK algorithm – an importance sampling method for evaluating multivariate normal probabilities – based on the work of Geweke (1989, 1991), Hajivassiliou (1990), Hajivassiliou and McFadden (1998), and Keane (1994).<sup>7</sup> Lee (1997, 1999) employs SML estimation via GHK simulator for various discrete dynamic panel models introduced by Heckman (1981), including the habit persistence model.

To tackle multidimensional integral problems in LDV models, simulation methods have been intensively used in empirical work in both economics and finance (see Hyslop (1999), Gerfin and Lechner (2002), Koopman et al. (2012), Gagliardini and Gouriéroux (2005), Feng et al. (2008), and Koopman et al. (2009)). However, the computational difficulty of the simulation approach has been emphasized by many papers such as Hajivassiliou et al. (1996), McFadden and Ruud (1994), and Feng et al. (2008).<sup>8</sup> In comparison to other estimation methods, several studies (Geweke et al. (1994), Geweke et al. (1997), and Patil et al. (2017)) demonstrate via Monte Carlo simulations that the SML (with the GHK simulator) in a multinomial probit model is outperformed by Gibbs sampling and CLE. Moreover, in a panel ordered probit with autocorrelated error terms, where the likelihood contains 6 dimensional integrals, Bhat et al. (2010) find that the CLE approach is 40 times faster than SML and performs similarly to SML in terms of efficiency and recovering the true parameters.<sup>9</sup>

---

<sup>5</sup>See Bertschek and Lechner (1998) for various GMM-based estimators in panel probit models with unknown error structure. But, the problem of large number of moment conditions as  $T$  increases is still present.

<sup>6</sup>Simulation methods have been proposed by Lerman and Manski (1993), McFadden (1989), and Pakes and Pollard (1989), among others. Simulations have been proven to be useful in limited dependent variable models to approximate their complex likelihoods (Börsch-Supan and Hajivassiliou (1993), Lee (1997), Laroque and Salanié (1993), Fermanian and Salanié (2004), and Kristensen and Shin (2012)), moment conditions (McFadden (1989), Duffie and Singleton (1993), and Creel and Kristensen (2012)), or score functions (Hajivassiliou and Ruud (1994), and Hajivassiliou and McFadden (1998)).

<sup>7</sup>See Train (2009) for details, and Börsch-Supan and Hajivassiliou (1993) for its performance.

<sup>8</sup>Hajivassiliou et al. (1996) states that “The problem is computationally difficult except in very special cases.” Similarly, Feng et al. (2008) says “Although the SML estimators are consistent and efficient for a large number of simulations, practitioners may find the procedure quite difficult and time-consuming.”

<sup>9</sup>Bhat et al. (2010)’s result is interesting since CL is supposed to be less efficient than the full likelihood approach. However, SML is efficient when the number of draws tends to infinity; otherwise, the simulation

In the Bayesian literature, MCMC methods are commonly used to deal with the intractable likelihoods of LDV models (see [Albert and Chib \(1993\)](#), [Koop and Poirier \(1993\)](#), [McCulloch and Rossi \(1994\)](#), and [Wei \(1999\)](#)). In particular, [Chauvet and Potter \(2005\)](#), [Dueker \(2005\)](#), and [Stegmueller \(2013\)](#) utilize Gibbs sampling for the AR probit model. However, [Müller and Czado \(2005\)](#) show that the Gibbs sampler exhibits bad convergence properties in this model. Nevertheless, as a solution, they suggest a more sophisticated grouped move multi-grid Monte Carlo Gibbs sampler. Yet, this proposed technique was criticized by [Varin and Vidoni \(2006\)](#) and [Bhat et al. \(2010\)](#) for increasing the computational complexity by adding to the already cumbersome nature of the simulation approach. Finally, Monte Carlo simulations of [Varin and Vidoni \(2006\)](#) and [Patil et al. \(2017\)](#) show that CLE yields a comparable performance to MCMC techniques in a time-series AR probit and multinomial panel probit models, respectively.

Autoregressive panel probit model has been underutilized in the literature since its estimation is not as easy as other dynamic models, such as dynamic probit model where state dependence is modeled by letting  $y_{i,t-1}$  driving the dynamics instead of  $y_{i,t-1}^*$ . However, with the lagged latent variable, AR probit is more akin to standard time series models. While state dependence is an important feature of certain types of discrete data, there are cases where habit persistence can be a better representation of the true data-generating process. For instance, an economy in a recession today is more likely to be in a recession tomorrow not because it is in a recession today but because the underlying bad economic conditions will persist tomorrow and keep the economy in the state of recession. Another example could be unemployment dynamics: there is still an ongoing debate to explain why some people stay in unemployment longer than others (see [Arulampalam et al. \(2000\)](#) and [Heckman and Borjas \(1980\)](#) for examples). Is this a causal effect of being unemployed or is this due to individual traits such as being lazy or underqualified so that these individuals cannot find jobs? Even though this paper does not take any side in the *habit persistence* versus *state dependence* debate, it gives a chance for complex habit persistence models to be more frequently used in

---

error becomes non-negligible. Hence, if one cannot simulate a large number of times – due to computational restrictions – SML also ends up being inefficient. Therefore, CL and SML provide comparable estimation results in terms of root mean square errors, but in terms of computation times, CL has an unquestionable advantage.

the literature by providing an easy, fast, and robust estimation technique with asymptotically normal estimates.

The plan of the rest of the paper is as follows. In Section 2, we formally introduce the model and its marginal and pairwise composite likelihoods. In Section 3, we focus on the identification of the parameters. The asymptotic results are given in Section 4. In Section 5, we report Monte Carlo simulations to study the finite-sample behavior of the composite likelihood estimator and compare it to MLE and GMM. Section 6 contains the empirical results on credit ratings. Finally, Section 7 concludes. The proofs of the results from Section 4 and some mathematical derivations are left to Appendix A – B. Remaining proofs, extra mathematical details and Monte Carlo simulations are in the supplementary material.

A final note on the notation used throughout this paper. For any scalar, vector, or matrix  $v$ , the Euclidean norm is denoted by  $|v| = \sqrt{\text{tr}(vv')}$ . For  $p > 0$ , the  $L_p$ -norm of  $v$  is denoted by  $\|v\|_p = (\text{E}|v|^p)^{1/p}$  for a random variable  $v$  (scalar, vector, or matrix).

## 2 AR Panel Probit Model with Random Effects

In this section, we introduce the autoregressive (AR) panel probit model with time-invariant individual random effects and construct its marginal and pairwise composite likelihoods. Moreover, distributional and exogeneity assumptions are stated.

We consider the following AR panel probit model for the observable discrete outcome variable  $y_{it}$  that can take values in the ordered category set  $\{1, \dots, S\}$ . For  $i = 1, \dots, N$ ,  $t = 1, \dots, T$ , and  $T \geq 3$ , we have

$$y_{it} = s \quad \text{if} \quad \tau_{s-1} \leq y_{it}^* < \tau_s, \quad (3)$$

$$y_{it}^* = \rho y_{i,t-1}^* + \beta' x_{it} + \alpha_i + \varepsilon_{it}, \quad (4)$$

where  $y_{it}^*$  is the autoregressive latent variable that governs the underlying continuous process,  $x_{it}$  is a  $K$ -dimensional vector of observable covariates,  $\alpha_i$  is individual-specific unobservable effects,  $\varepsilon_{it}$  is unobservable error term,  $\rho$  is the autocorrelation coefficient of the latent process,  $\beta$  is a  $K$ -dimensional coefficient vector, and  $\{\tau_0, \dots, \tau_S\}$  are the threshold coefficients such

that  $\tau_0 = -\infty < \tau_1 = 0 < \tau_2 < \dots < \tau_{S-1} < \tau_S = \infty$ .

Since  $y_{i,t-1}^*$  is not observed, we use backward-substitution on the latent process. That is, the current latent variable becomes a weighted sum of the past observations and innovations, where the weights are decreasing at an exponential rate.

$$\begin{aligned} y_{it}^* &= \rho y_{i,t-1}^* + \beta' x_{it} + \alpha_i + \varepsilon_{it} \\ &= \rho^t y_{i0}^* + \sum_{k=1}^t \rho^{t-k} \beta' x_{ik} + \frac{1-\rho^t}{1-\rho} \alpha_i + \sum_{k=1}^t \rho^{t-k} \varepsilon_{ik}. \end{aligned}$$

Then, an initial value problem arises since  $y_{i0}^*$  is now present in the equation. There are several solutions proposed for this matter (see [Wooldridge \(2005\)](#)), including assuming a nonrandom  $y_{i0}^*$ , or drawing it from a stationary distribution, or conditioning on the initial value. A nonrandom initial value in the presence of random effects is not desirable since it would mean that  $y_{i0}^*$  is not affected by the individual heterogeneity  $\alpha_i$ . Conditioning on the initial value is also not possible here since  $y_{i0}^*$  is not observed. Hence, we assume that  $y_{i0}^* = \alpha_i / (1 - \rho) + \varepsilon_{i0} / \sqrt{1 - \rho^2}$ , where the initial value is modeled in a *stationary* way such that the distribution of  $y_{it}^*$  conditional on the covariates is the same for every  $t$  (beyond the dependence of the covariates on  $t$ ). This assumption yields

$$y_{it}^* = \sum_{k=1}^t \rho^{t-k} \beta' x_{ik} + \frac{\alpha_i}{1-\rho} + \frac{\rho^t}{\sqrt{1-\rho^2}} \varepsilon_{i0} + \sum_{k=1}^t \rho^{t-k} \varepsilon_{ik}. \quad (5)$$

Under some distributional and independence assumptions, which will be stated below, it is easy to show that  $E(y_{it}^* | \mathbf{x}_i)$  and  $\text{Var}(y_{it}^* | \mathbf{x}_i)$  do not vary over time, where  $\mathbf{x}_i = (x'_{i1}, \dots, x'_{iT})'$ . In forming the composite likelihoods, we use the functional form (5) instead of (4) for  $y_{it}^*$ .

s In the panel data literature, the individual effects  $\alpha_i$  is treated either as fixed or random. In non-linear panel data models, it is not possible, in general, to “difference away” the individual heterogeneity. Therefore, in non-linear panels with large  $N$  and fixed  $T$ , it is well known that fixed effects estimation suffers from an incidental parameter problem (see [Neyman and Scott \(1948\)](#)). In this case, the fixed effects estimator has a bias of order  $1/T$  regardless of how large  $N$  is.<sup>10</sup> Note that, under large  $N$  and large  $T$  cases, the incidental parameter

---

<sup>10</sup>However, for the fixed effects estimation of panel probit with binary data, [Greene \(2004a\)](#) finds that the

problem can be solved in static and dynamic panel probit models (Arellano and Bonhomme (2009), Hahn and Kuersteiner (2011), and Dhaene and Jochmans (2015)). However, none of the existing methods for bias correction (or bias reduction) allows for lagged latent variable on the right-hand side. Moreover, the proposed techniques assume that the maximum likelihood estimator is feasible, which is challenging in the case of AR panel probit due to a large number of integrals. For all these reasons, we will assume that the individual heterogeneity is a random variable instead of a fixed parameter.

In the random effects literature, it is a common practice to use the correlated random effects approach à la Mundlak (1978) and Chamberlain (1984) where the individual random effects conditional on the covariates are assumed to belong to a specific distribution with finite number of parameters. In this paper, we also take this approach and assume that  $\alpha_i$  conditional on  $\mathbf{x}_i$  is normally distributed. Specific assumptions regarding the distributions of the unobserved variables of the model are discussed in the following assumption.

**Assumption 1.** (*Normality*) Assume that (i)  $\varepsilon_{it}$  is independent and identically distributed over  $i = 1, \dots, N$  and  $t = 1, \dots, T$  with the distribution  $\mathcal{N}(0, (1 - \rho^2)\sigma_\varepsilon^2)$ , (ii)  $\alpha_i$  is independent and identically distributed over  $i = 1, \dots, N$  conditional on the observed covariates such that  $\alpha_i | \mathbf{x}_i \stackrel{iid}{\sim} (1 - \rho)\mathcal{N}(\mu + \gamma'\bar{\mathbf{x}}_i, \sigma_\alpha^2)$  where  $\bar{\mathbf{x}}_i = T^{-1} \sum_{t=1}^T x_{it}$ , and (iii) the normalization  $\sigma_\varepsilon^2 + \sigma_\alpha^2 = 1$  holds.

There are two important implications of this assumption. First, the distributional assumption on the composite error term  $\alpha_i + \varepsilon_{it}$  determines the distribution of the latent variable conditional on the covariates. Second, the normalization assumption  $\sigma_\varepsilon^2 + \sigma_\alpha^2 = 1$  helps us identify the parameters since it is well known that in probit models the parameters are identified up to a scale parameter where the scaling factor is the standard deviation of the error term. That is, only  $\beta/\sqrt{\sigma_\varepsilon^2 + \sigma_\alpha^2}$  could be identified in this model. The multiplication of the distributions of  $\varepsilon_{it}$  and  $\alpha_i | \mathbf{x}_i$  by  $\sqrt{1 - \rho^2}$  and  $(1 - \rho)$ , respectively, is just a reparametrization of the distributions which facilitates the mathematical terms in the distribution of  $y_{it}^*$  in (5). Even though, in this paper, we do keep the independence assumption in the innovations, one can allow for serial correlation. For instance, one can introduce an

---

order of bias is actually larger than  $1/T$  in the simulations. It is quoted that "... a widely accepted result that suggests that the probit estimator is actually relatively well behaved appears to be incorrect."



autoregressive random time effects,  $\varepsilon_{it} = \lambda_t + u_{it}$  with  $\lambda_t = \delta\lambda_{t-1} + \nu_t$ , or model  $\varepsilon_{it}$  as a moving average  $\varepsilon_{it} = e_{it} + \delta e_{i,t-1}$  or as a stationary autoregressive process  $\varepsilon_{it} = \delta\varepsilon_{i,t-1} + e_{it}$ . One can also allow for time-varying heteroskedasticity in  $\varepsilon_{it}$ . But, in all these extensions, one needs to carefully model the normalization and the variance of the composite error term  $\alpha_i + \varepsilon_{it}$  since the identification of the parameters relies on it.

With the correlated random effects assumption, we can now define the vector of parameters as  $\theta = (\rho, \mu, \beta', \gamma', \sigma_\alpha^2, \tau')$  where  $\tau = (\tau_2, \dots, \tau_{S-1})'$ . The length of the parameter vector is  $1 + 1 + K + K_\gamma + 1 + (S - 2)$ , where  $S$  is the number of distinct discrete outcomes,  $K$  is the number of observed covariates, and  $K_\gamma$  is the dimension of observed variables in the correlated random effects. Note that  $K$  and  $K_\gamma$  are not necessarily equal to each other; one may include other variables (such as interaction terms of the covariates or extra control variables) in the random effects. Hence, even though we represent the conditional mean of the correlated random effects as  $\mu + \gamma'\bar{\mathbf{x}}_i$ , one can easily generalize it to  $\mu + \gamma'z_i$  where  $z_i$  is a  $(K_\gamma \times 1)$  vector containing deterministic functions of the covariates  $\mathbf{x}_i$  as well as any other time-invariant observed variables.

In parametric discrete outcome models, normal and logistic distributions are the most common choices for the assumed distribution of the errors. However, in the autoregressive latent variable model, the distribution of  $y_{it}^*$  is not determined by that of the error term only, but instead, by the distribution of the weighted summation of the error terms due to the unobserved right-hand-side variable  $y_{i,t-1}^*$  (see the next subsection for details). Since we need to rely on the distribution of the summation of the error terms, their distribution should belong to a stable distribution family.<sup>11</sup> In other words, the error distribution should be closed under convolution. While a normal distribution is a stable distribution, a logistic distribution is not (see [George and Mudholkar \(1983\)](#) and [Ojo \(2003\)](#) for a mathematical expression of convoluted logistic distributions). Therefore, a normality assumption is needed for a parametric AR panel probit model.

The next assumption is on the strict exogeneity of the covariates and the independence

---

<sup>11</sup>([Feller \(1971, p. 169\)](#)) Let  $X, X_1, X_2, \dots$  be independent and identically distributed. The distribution is called *stable* if  $\forall n \exists c_n > 0$  and  $\gamma \in \mathbb{R}$  such that  $(X_1 + \dots + X_n)$  has the same distribution as  $c_n X + \gamma$ . Stable distributions do not have closed-form formulae densities except for the following ones: Gaussian, Cauchy, and Lévy distributions.

between the innovations and the random effects conditional on the covariates.

**Assumption 2.** (*Exogeneity*) Assume that the covariates are strictly exogenous in the sense that  $f(\varepsilon_{iT}, \dots, \varepsilon_{i0} | \mathbf{x}_i, \alpha_i) = f(\varepsilon_{iT}, \dots, \varepsilon_{i0})$ .

Even though strict exogeneity is a restrictive assumption, it is common in non-linear dynamic panel models with observed covariates (see, for example, [Honoré and Kyriazidou \(2000\)](#), [Wooldridge \(2005\)](#), and [Semykina and Wooldridge \(2013\)](#)). One way to relax this assumption is to allow for endogeneity of the covariates by jointly modeling  $(y_{it}^*, x'_{it})$  in a vector autoregressive panel probit model (see [Dueker \(2005\)](#) for a time series version). This is also an interesting model, but we leave it as a future work for now and retain the strict exogeneity assumption.

## 2.1 Marginal and Pairwise Composite Likelihoods

In this subsection, we compute the univariate and bivariate probabilities of discrete outcomes that form the marginal and pairwise composite likelihoods. First, we compute the conditional distribution of the latent state variable  $y_{it}^*$ . Using [Assumption 1](#) and following the approach [Mundlak \(1978\)](#), we can rewrite [\(5\)](#) as follows:

$$y_{it}^* = \mu + \gamma' \bar{\mathbf{x}}_i + \sum_{k=1}^t \rho^{t-k} \beta' x_{ik} + \eta_i + \frac{\rho^t}{\sqrt{1-\rho^2}} \varepsilon_{i0} + \sum_{k=1}^t \rho^{t-k} \varepsilon_{ik},$$

where  $\eta_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_\alpha^2)$  is independent from  $\mathbf{x}_i$  and  $\varepsilon_i$ . Note that the conditional mean is  $E(y_{it}^* | \mathbf{x}_i) = \mu + \gamma' \bar{\mathbf{x}}_i + \sum_{k=1}^t \rho^{t-k} \beta' x_{ik}$  and the conditional variance is  $\text{Var}(y_{it}^* | \mathbf{x}_i) = \text{Var}(\eta_i + \rho^t / \sqrt{1-\rho^2} \varepsilon_{i0} + \sum_{k=1}^t \rho^{t-k} \varepsilon_{ik}) = \sigma_\alpha^2 + \sigma_\varepsilon^2 = 1$ . Finally, the conditional correlation of the latent variable is  $\text{corr}(y_{it}^*, y_{i,t+j}^* | \mathbf{x}_i) = \sigma_\alpha^2 + \rho^j \sigma_\varepsilon^2$ . This means that the sources of the autocorrelation in the latent variable are the random effects and the lagged latent variable. While the contribution of the former to the autocorrelation is constant, that of the latter is decreasing at an exponential rate.

The conditional distribution  $f(\mathbf{y}_i^* | \mathbf{x}_i, \theta)$  of  $\mathbf{y}_i^* = (y_{i1}^*, \dots, y_{iT}^*)$  conditional on all of the covariates  $\mathbf{x}_i$  is a  $T$ -dimensional normal distribution with a covariance matrix  $\mathbf{V}$  in the

following form:

$$\mathbf{V} = \sigma_\alpha^2 \mathbf{1}\mathbf{1}' + \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \vdots & & \ddots & & \vdots \\ \rho^{T-1} & \rho^{T-2} & \cdots & & 1 \end{bmatrix},$$

where  $\mathbf{1}$  is a  $T \times 1$  vector of ones. Note that the diagonal elements of  $\mathbf{V}$  are all equal to 1 since  $\sigma_\alpha^2 + \sigma_\varepsilon^2 = 1$ . However, since  $y_{it}^*$  is not independent over time,  $\mathbf{V}$  is a non-diagonal matrix – this is true even when there are no random effects. Hence, the (full) likelihood function of an individual  $i$  can be written with a  $T$ -dimensional integral as

$$L(\mathbf{y}_i | \mathbf{x}_i) = \int_{a_{i1}}^{b_{i1}} \int_{a_{i2}}^{b_{i2}} \cdots \int_{a_{iT}}^{b_{iT}} f(y_{i1}^*, \dots, y_{iT}^* | \mathbf{x}_i; \theta) dy_{i1}^* \cdots dy_{iT}^*, \quad (6)$$

where  $\{(a_{it}, b_{it})\}_{t=1}^T$  are truncation points such that  $(a_{it}, b_{it}) = (\tau_{s-1}, \tau_s)$  whenever  $y_{it} = s$ , which results in  $S^T$  possible integration. For moderate or large  $T$ , simulating these integrals is computationally demanding. Composite likelihood reduces the number of integrals by *erroneously* assuming independence between certain subsets of  $(y_{i1}^*, \dots, y_{iT}^*)$ . The simplest composite likelihood arises when conditional independence of  $y_{it}^*$  is assumed for all  $t$ . This is called *Marginal Composite Likelihood* (MCL), which reduces the  $T$ -dimensional integrals to a single-dimensional one. In contrast, a *Pairwise Composite Likelihood* (PCL) takes the dependency between pairs of  $(y_{i1}^*, \dots, y_{iT}^*)$  into account but ignores any higher dimensional dependencies. This results in a double integral which is still easy to compute. In this paper we focus on the bivariate distribution between  $y_{it}^*$  and  $y_{i,t+j}^*$  for  $1 \leq j \leq J < T$ .

$$\begin{bmatrix} y_{it}^* \\ y_{i,t+j}^* \end{bmatrix} \Big| \mathbf{x}_i \sim \mathcal{N} \left( \begin{bmatrix} m_t(\mathbf{x}_i; \theta) \\ m_{t+j}(\mathbf{x}_i; \theta) \end{bmatrix}, \begin{bmatrix} 1 & r_j(\theta) \\ r_j(\theta) & 1 \end{bmatrix} \right), \quad (7)$$

where  $m_t(\mathbf{x}_i; \theta) = \mu + \gamma' \bar{\mathbf{x}}_i + \sum_{k=1}^t \rho^{t-k} \beta' x_{ik}$  is the conditional mean and  $r_j(\theta) = \rho^j (1 - \sigma_\alpha^2) + \sigma_\alpha^2$  is the conditional correlation. Instead of the bivariate distribution between  $y_{it}^*$  and  $y_{i,t+j}^*$ , one can consider that between  $y_{it}^*$  and  $y_{is}^*$  for all  $s \neq t$ . This is called All-Pairs Composite Likelihood. But in a time series setting with an autocorrelation, it is natural to focus on only

$J$ -period distant pairs since most of the information on correlation is contained in the recent pairs. Moreover, far-apart observations might bring almost no information but, instead, end up bringing more noise to the estimation (see [Joe and Lee \(2009\)](#) and [Varin and Vidoni \(2006\)](#)). Model selection, e.g. choosing optimal  $J$ , can be done based on the composite likelihood information criteria  $\text{CLIC}(\hat{\theta}) = \mathcal{L}_{CLE}(\hat{\theta}) - \text{tr}\{\mathcal{I}(\hat{\theta})\mathcal{H}(\hat{\theta})^{-1}\}$  where  $\mathcal{L}_{CLE}(\hat{\theta})$  is the composite log-likelihood,  $\mathcal{I}(\hat{\theta})$  is the variance of the score, and  $\mathcal{H}(\hat{\theta})$  is the Hessian matrix all evaluated at  $\hat{\theta}$  (see Section 4 for more details, and also [Varin and Vidoni \(2005\)](#) and [Lindsay et al. \(2011\)](#)). The CLIC is the counterpart, based on composite likelihood approach, of TIC proposed by [Takeuchi \(1976\)](#), which is a modified version of AIC of [Akaike \(1973\)](#) under misspecification.

### *Marginal Composite Likelihood*

The probability of  $y_{it}$  taking the value  $s$  is

$$\mathbb{P}(y_{it} = s | \mathbf{x}_i; \theta) = \mathbb{P}(\tau_{s-1} \leq y_{it}^* < \tau_s | \mathbf{x}_i; \theta) = \Phi(\tau_s - m_t(\mathbf{x}_i; \theta)) - \Phi(\tau_{s-1} - m_t(\mathbf{x}_i; \theta)), \quad (8)$$

where  $\mathbb{P}$  denotes the probability function. An associated composite log-likelihood based on the univariate probabilities given in (8) is

$$\begin{aligned} \mathcal{L}_{MCL}(\theta | \mathbf{y}, \mathbf{x}) &= \frac{1}{N} \sum_{i=1}^N \ell_i(\theta | \mathbf{y}_i, \mathbf{x}_i) \\ &= \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \ln f(y_{it} | \mathbf{x}_i; \theta) \\ &= \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \sum_{s=1}^S \mathbb{1}(y_{it} = s) \ln \mathbb{P}(y_{it} = s | \mathbf{x}_i; \theta), \end{aligned} \quad (9)$$

where  $\mathbb{1}(\cdot)$  denotes the indicator function,  $\ell_i$  denotes the composite log-likelihood of an individual  $i$  and  $f(y_{it} | \mathbf{x}_i; \theta)$  denotes the likelihood of an observation conditional on the covariates.

Even though the complete independence assumption in MCL seems like an extreme simplification, the MCL estimator maximizing  $\mathcal{L}_{MCL}(\theta | \mathbf{y}, \mathbf{x})$  is still consistent as  $N \rightarrow \infty$ . Moreover, the Monte Carlo simulation results presented in Section 5 show that the MCL estimator's performance is comparable to PCL's. However, one should note that not all

parameters can be identified and estimated in the MCL. For instance,  $\sigma_\alpha^2$  appears only in the correlation between  $y_{it}^*$  and  $y_{i,t+j}^*$ , which is ignored in the MCL (see Section 3 for more details on identification). Thus, we define the marginal composite likelihood estimator as  $\hat{\theta}_{MCL} = \arg \max_{\theta \in \Theta_{/\{\sigma_\alpha^2\}}} \mathcal{L}_{MCL}(\theta|\mathbf{y}, \mathbf{x})$ , where  $\mathcal{L}_{MCL}$  is given in (9) and  $\Theta_{/\{\sigma_\alpha^2\}}$  is the parameter space excluding the dimension along  $\sigma_\alpha^2$ , since it is not included in the MCL.

### Pairwise Composite Likelihood

To facilitate the notation for the bivariate probability, let's denote  $q_{its} \equiv \tau_s - m_t(\mathbf{x}_i; \theta)$ . The bivariate probability can be computed as follows:<sup>12</sup>

$$\begin{aligned} \mathbb{P}(y_{it} = s_1, y_{i,t+j} = s_2 | \mathbf{x}_i; \theta) &= \mathbb{P}(\tau_{s_1-1} \leq y_{it}^* < \tau_{s_1}, \tau_{s_2-1} \leq y_{i,t+j}^* < \tau_{s_2} | \mathbf{x}_i; \theta) \\ &= \Phi_2(q_{i,t,s_1}, q_{i,t+j,s_2} | r_j) + \Phi_2(q_{i,t,s_1-1}, q_{i,t+j,s_2-1} | r_j) \\ &\quad - \Phi_2(q_{i,t,s_1}, q_{i,t+j,s_2-1} | r_j) - \Phi_2(q_{i,t,s_1-1}, q_{i,t+j,s_2} | r_j), \end{aligned} \quad (10)$$

where  $\Phi_2(\cdot, \cdot | r)$  denotes the bivariate standard normal distribution function with the correlation coefficient  $r$ . A pairwise composite log-likelihood based on the bivariate probabilities given in (10) is

$$\begin{aligned} \mathcal{L}_{PCL}(\theta|\mathbf{y}, \mathbf{x}) &= \frac{1}{N} \sum_{i=1}^N \ell_i(\theta|\mathbf{y}_i, \mathbf{x}_i) \\ &= \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta) \\ &= \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^{T-J} \sum_{j=1}^J \sum_{s_1=1}^S \sum_{s_2=1}^S \mathbb{1}(y_{it} = s_1, y_{i,t+j} = s_2) \ln \mathbb{P}(y_{it} = s_1, y_{i,t+j} = s_2 | \mathbf{x}_i; \theta), \end{aligned} \quad (11)$$

where  $f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta)$  denotes the likelihood of a pair of observations conditional on the covariates. Finally, we define the PCL estimator as  $\hat{\theta}_{PCL} = \arg \max_{\theta \in \Theta} \mathcal{L}_{PCL}(\theta|\mathbf{y}, \mathbf{x})$ . Since the main focus of this paper is the pairwise composite likelihood estimation, we will drop the subscript and denote the PCL as  $\mathcal{L}(\theta)$  and the associated estimator as  $\hat{\theta}$ .

<sup>12</sup>Note that for any two random variables  $X$  and  $Y$  with the bivariate cumulative distribution function  $G$ , one can write  $\mathbb{P}(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = G(x_2, y_2) + G(x_1, y_1) - G(x_1, y_2) - G(x_2, y_1)$ .

### 3 Identification of the Parameters

Parameter identification requires special attention in composite likelihood estimations. It might be the case that a parameter does not appear in a chosen composite likelihood or, even if it appears, it may not be identified under certain conditions. For instance,  $\sigma_\alpha^2$  is not included in the marginal composite likelihood given in (9); hence,  $\sigma_\alpha^2$  cannot be estimated in the MCL. Moreover, if all of the covariates are irrelevant, that is, if  $\beta = 0$ , then the parameter  $\rho$  disappears from the marginal likelihood. Hence,  $\rho$  cannot be identified in the MCL if  $\beta = 0$ . On the other hand,  $\rho$  can be identified in the pairwise composite likelihood even if  $\beta = 0$ , given that there are no random effects. The reason is that  $\rho$  appears not only in the conditional mean but also in the correlation coefficient. Hence, when there are no random effects and  $\beta = 0$ , pairwise composite likelihood will include bivariate probabilities of the form  $\Phi_2(0, 0|\rho^j)$ , from which one can easily identify  $\rho$ . However, if the random effects are present, then the bivariate probabilities become  $\Phi_2(\cdot, \cdot|\sigma_\alpha^2 + \rho^j(1 - \sigma_\alpha^2))$ . In this case,  $\rho$  and  $\sigma_\alpha^2$  cannot be separately identified when  $J = 1$ . All these cases show how tricky the identification of parameters in composite likelihood methods can be.

The marginal and pairwise composite likelihoods proposed in (9) and (11), respectively, are highly non-linear in the parameters. One source of the nonlinearity is the normal cumulative distribution function and the other source is the nonlinearities in  $m_i(\mathbf{x}_i; \theta)$  and  $r_j(\theta)$  due to the autocorrelation. The former nonlinearity is also present in regular static probit estimation whereas the latter is not. For the identification, it is possible to show negative definiteness of the Hessian matrix  $\mathcal{H}(\theta^*)$  in a static probit model whereas it is extremely cumbersome to follow this path in autoregressive probit model due to *double* nonlinearity. For this reason, we follow a different strategy to identify the parameters.

We borrow our identification ideas from the estimation techniques of the *average derivative estimation* literature. In this literature, average derivatives are utilized to estimate parameters of interest in semi-parametric models (see [Stoker \(1986\)](#), [Powell et al. \(1989\)](#), [Stoker \(1991\)](#), and [Newey and Stoker \(1993\)](#)).<sup>13</sup> In this paper, we use certain partial effects

---

<sup>13</sup>If  $E(y|x) = F(\alpha + x'\beta)$ , then the average derivative of  $y$  on  $x$ ,  $E[\partial E(y|x)/\partial x]$ , will be proportional to  $\beta$ . Since our model is fully parametric, the function  $F$ , thus, the proportionality constants are known. Hence, we can eliminate the proportionality to isolate the coefficient of interest.

to isolate the parameters to show that, under mild assumptions, the composite likelihoods proposed in this paper contain enough information to identify the parameters in the AR panel probit model. We first show the identification of all the parameters except  $\sigma_\alpha^2$  in the MCL, since it is easier to understand the identification strategy in an easier composite likelihood. Then, we show the identification results in the PCL together with the identification of  $\sigma_\alpha^2$ .

### Identification of $\rho$

Assume that there is at least one continuous and significant covariate in  $x_{it}$ . Without loss of generality, assume that it is the first covariate, denoted by  $x_{it}^{(1)}$  with  $\beta_1 \neq 0$ . Consider  $\mathbb{P}(y_{it} = 1 | \mathbf{x}_i) = \mathbb{P}(-\infty < y_{it}^* < 0 | \mathbf{x}_i) = \Phi(-m_t(\mathbf{x}_i; \theta))$ , where  $m_t(\mathbf{x}_i; \theta) = \mu + \gamma' \bar{\mathbf{x}}_i + \beta' x_{it} + \sum_{k=1}^{t-1} \rho^{t-k} \beta' x_{ik}$ . We will analyze the cases when  $\gamma = 0$  and  $\gamma \neq 0$  separately.

For now, assume that  $\gamma = 0$ , i.e. we have uncorrelated random effects. Then the ratio of the partial effects of  $x_{i,t-1}^{(1)}$  and  $x_{it}^{(1)}$  gives

$$\left. \begin{aligned} \frac{\partial \mathbb{P}(y_{it}=1 | \mathbf{x}_i)}{\partial x_{it}^{(1)}} &= -\frac{\partial m_t(\mathbf{x}_i; \theta)}{\partial x_{it}^{(1)}} \phi(-m_t(\mathbf{x}_i; \theta)) = -\beta_1 \phi(-m_t(\mathbf{x}_i; \theta)) \\ \frac{\partial \mathbb{P}(y_{it}=1 | \mathbf{x}_i)}{\partial x_{i,t-1}^{(1)}} &= -\frac{\partial m_t(\mathbf{x}_i; \theta)}{\partial x_{i,t-1}^{(1)}} \phi(-m_t(\mathbf{x}_i; \theta)) = -\rho \beta_1 \phi(-m_t(\mathbf{x}_i; \theta)) \end{aligned} \right\} \frac{\partial \mathbb{P}(y_{it} = 1 | \mathbf{x}_i) / \partial x_{i,t-1}^{(1)}}{\partial \mathbb{P}(y_{it} = 1 | \mathbf{x}_i) / \partial x_{it}^{(1)}} = \rho.$$

Since  $\mathbb{E}(y_{it} | \mathbf{x}_i) = \sum_{s=1}^S s \mathbb{P}(y_{it} = s | \mathbf{x}_i) = \sum_{s=1}^S s [\Phi(\tau_s - m_t(\mathbf{x}_i; \theta)) - \Phi(\tau_{s-1} - m_t(\mathbf{x}_i; \theta))]$ , the same result can also be found by  $[\partial \mathbb{E}(y_{it} | \mathbf{x}_i) / \partial x_{i,t-1}^{(1)}] / [\partial \mathbb{E}(y_{it} | \mathbf{x}_i) / \partial x_{it}^{(1)}] = \rho$ . This result is simple and intuitive: as  $k$  increases, the partial effect of  $x_{i,t-k}$  on  $y_{it}$  decreases at an exponential rate driven by the autocorrelation coefficient  $\rho$ .

For the case when  $\gamma \neq 0$ , i.e. for the correlated random effects, the contribution of  $x_{it}$  on the probability distribution of  $y_{it}$  is not only driven by  $\rho$  but also by  $\gamma/T$  (due to  $\gamma' \bar{\mathbf{x}}_i$ ). In this case, to isolate  $\rho$ , we need to eliminate the contribution of the random effects. Note that

$$\begin{aligned} \frac{\partial \mathbb{P}(y_{it} = 1 | \mathbf{x}_i)}{\partial x_{i,t-1}^{(1)}} - \frac{\partial \mathbb{P}(y_{it} = 1 | \mathbf{x}_i)}{\partial x_{i,t-2}^{(1)}} &= -(\rho \beta_1 + \frac{\gamma_1}{T}) \phi(-m_t(\mathbf{x}_i; \theta)) + (\rho^2 \beta_1 + \frac{\gamma_1}{T}) \phi(-m_t(\mathbf{x}_i; \theta)) \\ &= \rho(\rho \beta_1 - \beta_1) \phi(-m_t(\mathbf{x}_i; \theta)), \end{aligned} \tag{12}$$

which implies that

$$\frac{\partial \mathbb{P}(y_{it} = 1 | \mathbf{x}_i) / \partial x_{i,t-1}^{(1)} - \partial \mathbb{P}(y_{it} = 1 | \mathbf{x}_i) / \partial x_{i,t-2}^{(1)}}{\partial \mathbb{P}(y_{it} = 1 | \mathbf{x}_i) / \partial x_{it}^{(1)} - \partial \mathbb{P}(y_{it} = 1 | \mathbf{x}_i) / \partial x_{i,t-1}^{(1)}} = \rho. \quad (13)$$

Hence, in the case of correlated random effects, the interpretation of this result is as follows: the partial effect (net of the individual heterogeneity's effect) of  $x_{it}$  on  $y_{it}$  is, on average, decreasing at an exponential rate driven by the autocorrelation coefficient  $\rho$ .

Finally, we can focus on the identification of  $\rho$  in the PCL. We use similar ideas and make the same assumptions regarding  $x_{it}$ . For any given  $j = 1, \dots, J$ , consider  $\mathbb{P}(y_{it} = 1, y_{i,t+j} = 1 | \mathbf{x}_i) = \mathbb{P}(-\infty < y_{it}^* < 0, -\infty < y_{i,t+j}^* < 0 | \mathbf{x}_i) = \Phi(-m_t(\mathbf{x}_i; \theta), -m_{t+j}(\mathbf{x}_i; \theta) | r_j(\theta))$ . Note that  $\partial m_{t+q}(\mathbf{x}_i; \theta) / \partial x_{i,t+k}^{(1)} = \rho^{q-k} \beta_1 + \gamma_1 / T$  for  $q = 0, \dots, J$  and  $-t+1 \leq k \leq q$ . Let's assume again, for now, that  $\gamma = 0$ .

$$\begin{aligned} \frac{\partial \mathbb{P}(y_{it} = 1, y_{i,t+j} = 1 | \mathbf{x}_i)}{\partial x_{it}^{(1)}} &= -\frac{\partial m_t(\mathbf{x}_i; \theta)}{\partial x_{it}^{(1)}} \phi(-m_t(\mathbf{x}_i; \theta)) \Phi\left(\frac{r_j(\theta)m_t(\mathbf{x}_i; \theta) - m_{t+j}(\mathbf{x}_i; \theta)}{\sqrt{1 - r_j(\theta)^2}}\right) \\ &\quad - \frac{\partial m_{t+j}(\mathbf{x}_i; \theta)}{\partial x_{it}^{(1)}} \phi(-m_{t+j}(\mathbf{x}_i; \theta)) \Phi\left(\frac{-m_t(\mathbf{x}_i; \theta) + r_j(\theta)m_{t+j}(\mathbf{x}_i; \theta)}{\sqrt{1 - r_j(\theta)^2}}\right), \end{aligned}$$

$$\begin{aligned} \frac{\partial \mathbb{P}(y_{it} = 1, y_{i,t+j} = 1 | \mathbf{x}_i)}{\partial x_{i,t-1}^{(1)}} &= -\frac{\partial m_t(\mathbf{x}_i; \theta)}{\partial x_{i,t-1}^{(1)}} \phi(-m_t(\mathbf{x}_i; \theta)) \Phi\left(\frac{r_j(\theta)m_t(\mathbf{x}_i; \theta) - m_{t+j}(\mathbf{x}_i; \theta)}{\sqrt{1 - r_j(\theta)^2}}\right) \\ &\quad - \frac{\partial m_{t+j}(\mathbf{x}_i; \theta)}{\partial x_{i,t-1}^{(1)}} \phi(-m_{t+j}(\mathbf{x}_i; \theta)) \Phi\left(\frac{-m_t(\mathbf{x}_i; \theta) + r_j(\theta)m_{t+j}(\mathbf{x}_i; \theta)}{\sqrt{1 - r_j(\theta)^2}}\right). \end{aligned}$$

Taking their ratio yields

$$\frac{\partial \mathbb{P}(y_{it} = 1, y_{i,t+j} = 1 | \mathbf{x}_i) / \partial x_{i,t-1}^{(1)}}{\partial \mathbb{P}(y_{it} = 1, y_{i,t+j} = 1 | \mathbf{x}_i) / \partial x_{it}^{(1)}} = \rho.$$

For the case when  $\gamma \neq 0$ , we again need to subtract the contribution of the random effects to bivariate probability distribution of  $(y_{it}, y_{i,t+j})$ . As in the MCL case in (13),

$$\frac{\partial \mathbb{P}(y_{it} = 1, y_{i,t+j} = 1 | \mathbf{x}_i) / \partial x_{i,t-1}^{(1)} - \partial \mathbb{P}(y_{it} = 1, y_{i,t+j} = 1 | \mathbf{x}_i) / \partial x_{i,t-2}^{(1)}}{\partial \mathbb{P}(y_{it} = 1, y_{i,t+j} = 1 | \mathbf{x}_i) / \partial x_{it}^{(1)} - \partial \mathbb{P}(y_{it} = 1, y_{i,t+j} = 1 | \mathbf{x}_i) / \partial x_{i,t-1}^{(1)}} = \rho. \quad (14)$$



### Identification of $\sigma_\alpha^2$

To identify the parameter  $\rho$ , it was enough to analyze first-order partial derivatives since  $\rho$  appears in the first moment of  $y_{it}$ . However, the parameter  $\sigma_\alpha^2$  appears only in the second moments, in particular in the correlation between  $y_{it}^*$  and  $y_{i,t+j}^*$ . For this reason, first-order partial derivatives are not enough to isolate  $\sigma_\alpha^2$ ; we need to use second-order partials. In addition to the average derivate estimation idea, we use one of the results of [Plackett \(1954\)](#) regarding the correlation coefficient and mixed partial derivatives: for a vector of random variables  $\mathbf{z} = (z_1, \dots, z_n)'$  that has a multivariate standard normal distribution with correlation coefficients  $r_{ii'}$ , the following partial differential equations hold  $\partial\phi(\mathbf{z})/\partial r_{ii'} = \partial^2\phi(\mathbf{z})/\partial z_i\partial z_{i'}$ , for  $i, i' = 1, \dots, n$ . As in the identification of the parameter  $\rho$ , we will analyze cases with  $\gamma = 0$  and  $\gamma \neq 0$  separately. Let's assume that  $\gamma = 0$ . Then, the following ratio of the differentials yields

$$\begin{aligned} \frac{\partial\mathbb{P}(y_{it} = 1, y_{i,t+j} = 1|\mathbf{x}_i)/\partial x_{i,t+1}^{(1)}}{\partial\mathbb{P}(y_{i,t+j} = 1|\mathbf{x}_i)/\partial x_{i,t+1}^{(1)}} &= \frac{-\frac{\partial m_{t+j}(\mathbf{x}_i;\theta)}{\partial x_{i,t+1}^{(1)}}\phi(m_{t+j}(\mathbf{x}_i;\theta))\Phi\left(\frac{m_t(\mathbf{x}_i;\theta)-r_j(\theta)m_{t+j}(\mathbf{x}_i;\theta)}{\sqrt{1-r_j(\theta)^2}}\right)}{-\frac{\partial m_{t+j}(\mathbf{x}_i;\theta)}{\partial x_{i,t+1}^{(1)}}\phi(m_{t+j}(\mathbf{x}_i;\theta))} \\ &= \Phi\left(\frac{m_t(\mathbf{x}_i;\theta) - r_j(\theta)m_{t+j}(\mathbf{x}_i;\theta)}{\sqrt{1-r_j(\theta)^2}}\right). \end{aligned} \quad (15)$$

Next, we will differentiate the probability in (15) with respect to  $x_{it}^{(1)}$  and  $x_{i,t+1}^{(1)}$ , and then take their ratio.

$$\begin{aligned} \frac{\frac{\partial}{\partial x_{it}^{(1)}}\Phi\left(\frac{m_t(\mathbf{x}_i;\theta)-r_j(\theta)m_{t+j}(\mathbf{x}_i;\theta)}{\sqrt{1-r_j(\theta)^2}}\right)}{\frac{\partial}{\partial x_{i,t+1}^{(1)}}\Phi\left(\frac{m_t(\mathbf{x}_i;\theta)-r_j(\theta)m_{t+j}(\mathbf{x}_i;\theta)}{\sqrt{1-r_j(\theta)^2}}\right)} &= \frac{\partial m_t(\mathbf{x}_i;\theta)/\partial x_{it}^{(1)} - r_j(\theta)\partial m_{t+j}(\mathbf{x}_i;\theta)/\partial x_{it}^{(1)}}{-r_j(\theta)\partial m_{t+j}(\mathbf{x}_i;\theta)/\partial x_{i,t+1}^{(1)}} \\ &= \frac{\beta_1 - r_j(\theta)\rho^j\beta_1}{-r_j(\theta)\rho^{j-1}\beta_1}, \end{aligned} \quad (16)$$

where  $r_j(\theta) = \sigma_\alpha^2 + \rho^j(1 - \sigma_\alpha^2)$ . Hence, based on (15) and (16), the ratio of second-order partial derivatives below yields a result that is a function of only  $\rho$  and  $\sigma_\alpha^2$ . Having already

identified the parameter  $\rho$ , the ratio below can identify the parameter  $\sigma_\alpha^2$ .

$$\frac{\frac{\partial}{\partial x_{it}^{(1)}} \left[ \frac{\partial \mathbb{P}(y_{it}=1, y_{i,t+j}=1 | \mathbf{x}_i) / \partial x_{i,t+1}^{(1)}}{\partial \mathbb{P}(y_{i,t+j}=1 | \mathbf{x}_i) / \partial x_{i,t+1}^{(1)}} \right]}{\frac{\partial}{\partial x_{i,t+1}^{(1)}} \left[ \frac{\partial \mathbb{P}(y_{it}=1, y_{i,t+j}=1 | \mathbf{x}_i) / \partial x_{i,t+1}^{(1)}}{\partial \mathbb{P}(y_{i,t+j}=1 | \mathbf{x}_i) / \partial x_{i,t+1}^{(1)}} \right]} = \rho - \frac{1}{\rho^{j-1} r_j(\theta)}.$$

As in the MCL case, when  $\gamma \neq 0$ , we need to subtract the term  $\gamma_1/T$  in the derivatives of  $m_t(\mathbf{x}_i; \theta)$ . But, to implement the subtraction, we need to assume that  $J \geq 2$ . We skip the mathematical derivation since it is similar to (12) derived for the MCL case.

$$\frac{\frac{\partial}{\partial x_{it}^{(1)}} \left[ \frac{\partial \mathbb{P}(y_{it}=1, y_{i,t+j}=1 | \mathbf{x}_i) / \partial x_{i,t+1}^{(1)}}{\partial \mathbb{P}(y_{i,t+j}=1 | \mathbf{x}_i) / \partial x_{i,t+1}^{(1)}} \right] - \frac{\partial}{\partial x_{i,t-1}^{(1)}} \left[ \frac{\partial \mathbb{P}(y_{it}=1, y_{i,t+j}=1 | \mathbf{x}_i) / \partial x_{i,t+1}^{(1)}}{\partial \mathbb{P}(y_{i,t+j}=1 | \mathbf{x}_i) / \partial x_{i,t+1}^{(1)}} \right]}{\frac{\partial}{\partial x_{i,t+2}^{(1)}} \left[ \frac{\partial \mathbb{P}(y_{it}=1, y_{i,t+j}=1 | \mathbf{x}_i) / \partial x_{i,t+1}^{(1)}}{\partial \mathbb{P}(y_{i,t+j}=1 | \mathbf{x}_i) / \partial x_{i,t+1}^{(1)}} \right] - \frac{\partial}{\partial x_{i,t+1}^{(1)}} \left[ \frac{\partial \mathbb{P}(y_{it}=1, y_{i,t+j}=1 | \mathbf{x}_i) / \partial x_{i,t+1}^{(1)}}{\partial \mathbb{P}(y_{i,t+j}=1 | \mathbf{x}_i) / \partial x_{i,t+1}^{(1)}} \right]} = \rho^2 + \frac{1}{r_1(\theta)},$$

where  $r_1(\theta) = \rho + (1 - \rho)\sigma_\alpha^2$ . Hence, the ratio above can single out the parameter  $\sigma_\alpha^2$ .

#### Identification of $(\mu, \beta, \gamma, \tau)$

Having identified  $\rho$ , it is easy to identify the other conditional mean parameters  $(\mu, \beta, \gamma, \tau)$  by a standard identification approach. Assume that  $\tilde{\mu} \neq \mu^*$ ,  $\tilde{\beta} \neq \beta^*$ ,  $\tilde{\gamma} \neq \gamma^*$ . Then, we have  $\tilde{\mu} + \tilde{\gamma}' \bar{\mathbf{x}}_i + \sum_{k=1}^t \rho^{*t-k} \tilde{\beta}' x_{ik} \neq \mu^* + \gamma^{*'} \bar{\mathbf{x}}_i + \sum_{k=1}^t \rho^{*t-k} \beta^{*'} x_{ik}$ , with positive probability. Provided that  $E[\mathbf{x}_i \mathbf{x}_i']$  is positive definite, we have  $(\tilde{\mu} - \mu^*) + (\tilde{\gamma} - \gamma^*)' \bar{\mathbf{x}}_i + \sum_{k=1}^t \rho^{*t-k} (\tilde{\beta} - \beta^*)' x_{ik} \neq 0$ , with positive probability (note that  $\mathbf{x}_i$  is a  $TK \times 1$  dimensional vector whereas  $\bar{\mathbf{x}}_i$  is  $K \times 1$ ). Hence,  $\mathbb{P}(y_{it} = 1 | \mathbf{x}_i; \rho^*, \tilde{\mu}, \tilde{\beta}, \tilde{\gamma}) \neq \mathbb{P}(y_{it} = 1 | \mathbf{x}_i; \rho^*, \mu^*, \beta^*, \gamma^*)$  with positive probability. Finally, after  $\mu$  is identified, it is trivial to identify  $\tau = (\tau_2, \dots, \tau_{S-1})$  in  $\mathbb{P}(y_{it} = s | \mathbf{x}_i)$ , for  $s > 1$ . This concludes the identification of parameters in the MCL estimation. The identification of  $(\mu, \beta, \gamma, \tau)$  in the PCL estimator is very similar, hence we skip it here.

We show that for a pairwise composite likelihood it is sufficient to have  $J \geq 2$  to identify all of the parameters, where  $T \geq 3$ . This condition will be the basis for sufficiency in other types of composite likelihoods. Next, consider a trivariate distribution  $f(y_{it}, y_{i,t+j_1}, y_{i,t+j_1+j_2} | \mathbf{x}_i; \theta)$  with  $j_1 = 1, \dots, J_1$ ,  $j_2 = 1, \dots, J_2$ . The covariance of this distribution will contain terms

involving  $(\sigma_\alpha^2 + \rho^{j_1}\sigma_\varepsilon^2)$ ,  $(\sigma_\alpha^2 + \rho^{j_2}\sigma_\varepsilon^2)$ , and  $(\sigma_\alpha^2 + \rho^{j_1+j_2}\sigma_\varepsilon^2)$ . Note that this distribution already contains pairwise distributions of at least two distinct pairs; thus, it satisfies the sufficient conditions for identification. As a result, we show that a sufficient condition for parameter identification in the AR panel probit with correlated random effects model is to include distributions of at least two distinct pairs of observations into the composite likelihood.

## 4 Asymptotic Distribution of the CL Estimator

In this section, we state the asymptotic properties of the (pairwise) composite likelihood estimator. First, we state our last two assumptions.

**Assumption 3.** (*Moments*) Assume that the following moment conditions are satisfied by the covariates: i)  $E[\mathbf{x}_i\mathbf{x}_i']$  is non-singular and ii)  $\|\mathbf{x}_i\|_4 < \infty$ .

**Assumption 4.** (*Parameters*) Assume that the true parameter value  $\theta^*$  belongs to a compact set  $\Theta$ . Moreover,  $\theta^*$  is an interior point of  $\Theta$ , i.e.  $\theta^* \in \overset{\circ}{\Theta}$ .

These assumptions are standard in the asymptotic theory literature (see Amemiya (1985) and Newey and McFadden (1994)). Therefore, we do not state the assumptions needed for the consistency and asymptotic normality separately (for instance, finite second moments would be enough for consistency whereas finite fourth moments are needed for the asymptotic normality). Note that the moment conditions needed for the CLE in the AR panel probit model are not stronger than those needed for the MLE in the static probit model.

**Theorem 1.** Suppose that Assumptions 1 – 4 hold. Define the pairwise composite likelihood estimator as  $\hat{\theta} = \arg \max_{\theta \in \Theta} \mathcal{L}(\theta|\mathbf{y}, \mathbf{x})$ , where  $\mathcal{L}(\theta|\mathbf{y}, \mathbf{x}) = N^{-1} \sum_{i=1}^N \ell_i(\theta|\mathbf{y}_i, \mathbf{x}_i)$  is given in (11). Then, for  $T < \infty$  and  $N \rightarrow \infty$ , the pairwise composite likelihood estimator is consistent for the true parameter  $\theta^*$  and is asymptotically normal such that  $\sqrt{N}(\hat{\theta} - \theta^*) \rightarrow_d \mathcal{N}(0, \mathcal{H}(\theta^*)^{-1}\mathcal{I}(\theta^*)\mathcal{H}(\theta^*)^{-1})$ , where

$$\mathcal{H}(\theta) = E \left[ \frac{\partial^2 \ell_i(\theta)}{\partial \theta \partial \theta'} \right] \quad \text{and} \quad \mathcal{I}(\theta) = E \left[ \frac{\partial \ell_i(\theta)}{\partial \theta} \frac{\partial \ell_i(\theta)}{\partial \theta'} \right].$$

Note that  $\mathcal{H}(\theta^*) \neq -\mathcal{I}(\theta^*)$  here since not all of the information is used in forming the composite likelihood. Hence, we have a sandwich-form asymptotic variance. Even though

the asymptotic results provided above are about the pairwise composite likelihood estimator, similar results are valid for the marginal composite likelihood estimator  $\hat{\theta}_{MCL}$  as well.

To prove Theorem 1, we will closely follow the results of [Newey and McFadden \(1994\)](#). The asymptotic theory on the PCL estimator in the AR panel probit with random effects model is conceptually not different than the asymptotic theory on maximum or pseudo-likelihood estimators. However, a difficulty arises here due to the nonlinearities in the parameters in the bivariate function  $\Phi_2(q_{its_1}(\theta), q_{i,t+j,s_2}(\theta)|r_j(\theta))$ . The cumulative distribution function  $\Phi_2$  is not the only source of the nonlinearity (as is the case in a static probit model); the function  $q_{its}(\theta)$  is also non-linear in parameters. These ‘double’ nonlinearities result in complicated derivative functions of the pairwise composite likelihood. Hence, identification, computing the derivatives, and finding bounds for them become non-trivial. Despite this extra nonlinearity, the moment conditions on the process  $\mathbf{x}_i$  are not different than those in the static model. For instance, the finiteness of  $\sup_{\theta \in \Theta} |q_{its}(\theta)| = \sup_{\theta \in \Theta} |\tau_s - \mu - \gamma' \bar{\mathbf{x}}_i - \sum_{k=1}^t \rho^{t-k} \beta' x_{ik}| \leq c(1 + |\mathbf{x}_i|)$  is simply implied by the finiteness of  $|\mathbf{x}_i|$  (see the section [B.2](#) of the Appendix for more details). Most of the complications disappear when  $\rho = 0$ ; thus, at any point in the proof, one can recover the conditions for static probit by imposing  $\rho = 0$ .

The continuity and the measurability of the objective function are easy to prove since bivariate Gaussian cumulative distribution function  $\Phi_2$  and  $q_{its}(\theta)$  are all continuous and measurable functions. Thus,  $\ln f(y_{it}, y_{i,t+j}|\mathbf{x}_i; \theta)$  is continuous in  $\theta$  for given  $(y_{it}, y_{i,t+j}, \mathbf{x}_i)$ , and is a measurable function of  $(y_{it}, y_{i,t+j}, \mathbf{x}_i)$  for given  $\theta$ . Moreover,  $\ln f(y_{it}, y_{i,t+j}|\mathbf{x}_i; \theta)$  is twice continuously differentiable. Since each piece of the likelihood (marginals and conditionals) satisfies the Kullback-Leibler inequality, so will the chosen pieces in the composite likelihood. This property helps the estimation procedure to discriminate the true parameter value from other possible ones. Note that  $E[\ln f(y_{it}, y_{i,t+j}|\mathbf{x}_i; \theta^*)] \geq E[\ln f(y_{it}, y_{i,t+j}|\mathbf{x}_i; \theta)]$  for all  $\theta \in \Theta$  since  $E[\ln(f(y_{it}, y_{i,t+j}|\mathbf{x}_i; \theta)/f(y_{it}, y_{i,t+j}|\mathbf{x}_i; \theta^*))] \leq \ln E[f(y_{it}, y_{i,t+j}|\mathbf{x}_i; \theta)/f(y_{it}, y_{i,t+j}|\mathbf{x}_i; \theta^*)] = 0$ . The discussions in [Section 3](#) proves the uniqueness of  $\theta^*$  a maximizer.

Finally, in order to compute the asymptotic covariance matrix, we introduce consistent

estimators for  $\mathcal{H}(\theta^*)$  and  $\mathcal{I}(\theta^*)$ :

$$\widehat{\mathbf{H}}(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^N \frac{1}{T} \sum_{t=1}^{T-J} \sum_{j=1}^J \frac{\partial^2 \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \hat{\theta})}{\partial \theta \partial \theta'} \quad (17)$$

$$\widehat{\mathbf{I}}(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^N \left( \frac{1}{T} \sum_{t=1}^{T-J} \sum_{j=1}^J \frac{\partial \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \hat{\theta})}{\partial \theta} \right) \left( \frac{1}{T} \sum_{t=1}^{T-J} \sum_{j=1}^J \frac{\partial \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \hat{\theta})}{\partial \theta} \right)', \quad (18)$$

where the derivatives of the bivariate likelihood function are given as

$$\begin{aligned} \frac{\partial \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \hat{\theta}_N)}{\partial \theta} &= \sum_{s_1=1}^S \sum_{s_2=1}^S \mathbb{1}_{i,t,t+j,s_1,s_2} \frac{\frac{\partial}{\partial \theta} \mathbb{P}_{i,t,t+j,s_1,s_2}}{\mathbb{P}_{i,t,t+j,s_1,s_2}} \\ \frac{\partial^2 \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \hat{\theta}_N)}{\partial \theta \partial \theta'} &= \sum_{s_1=1}^S \sum_{s_2=1}^S \mathbb{1}_{i,t,t+j,s_1,s_2} \left[ \frac{\frac{\partial^2}{\partial \theta \partial \theta'} \mathbb{P}_{i,t,t+j,s_1,s_2}}{\mathbb{P}_{i,t,t+j,s_1,s_2}} - \frac{\frac{\partial}{\partial \theta} \mathbb{P}_{i,t,t+j,s_1,s_2} \frac{\partial}{\partial \theta'} \mathbb{P}_{i,t,t+j,s_1,s_2}}{\mathbb{P}_{i,t,t+j,s_1,s_2}^2} \right], \end{aligned}$$

where  $\mathbb{1}_{i,t,t+j,s_1,s_2} \equiv \mathbb{1}(y_{it} = s_1, y_{i,t+j} = s_2)$  and  $\mathbb{P}_{i,t,t+j,s_1,s_2} \equiv \mathbb{P}(y_{it} = s_1, y_{i,t+j} = s_2 | \mathbf{x}_i; \theta)$ .

The mathematical details on the derivatives are given in the section [B.1](#).

## 5 Monte Carlo Simulations

In this section, we present Monte Carlo simulation results to illustrate the finite sample performance of the marginal and pairwise composite likelihood estimators. All the simulation results rely on 1000 replications of the following model. For  $i = 1, \dots, N$  and  $t = 1, \dots, T$ ,

$$\begin{aligned} x_{it} &= 0.1 + 0.2x_{i,t-1} + \sqrt{2}e_{it} \\ \alpha_i | \mathbf{x}_i &\stackrel{iid}{\sim} (1 - \rho) \mathcal{N}(\mu + \gamma' \bar{\mathbf{x}}_i, \sigma_\alpha^2) \\ \varepsilon_{it} &\stackrel{iid}{\sim} \sqrt{1 - \rho^2} \mathcal{N}(0, 1 - \sigma_\alpha^2) \\ y_{i0}^* &= \alpha_i / (1 - \rho) + \varepsilon_{i0} / \sqrt{1 - \rho^2} \\ y_{it}^* &= \rho y_{i,t-1}^* + \beta' x_{it} + \alpha_i + \varepsilon_{it} \\ y_{it} &= s \text{ if } \tau_{s-1} \leq y_{it}^* < \tau_s, \end{aligned}$$

where  $e_{it} \stackrel{iid}{\sim} \mathcal{N}(0, 1)$ . The model is simulated for  $T + 1000$  time periods, but only the last  $T$  periods are used for estimation. We consider different sample sizes for the cross-section ( $N = 500, 1000$ ) and for the length of the panel data ( $T = 5, 10, 30$ ). Many other simulation setups, such as DGPs with different true parameters, different values of  $N$ , or different types of covariates (such as discrete variables), can be found in the supplementary material.

We first focus on the finite sample performances of the Marginal Composite Likelihood (MCL) estimator and Pairwise Composite Likelihood (PCL) estimator. We consider two different DGPs:  $(\rho, \mu, \beta, \tau) = (0.5, 0.5, 0.2, 1)$  corresponding to an autoregressive panel probit without random effects, and  $(\rho, \mu, \beta, \gamma, \sigma_\alpha^2, \tau) = (0.3, 0.5, 0.2, 0.5, 0.3, 1)$  corresponding to an autoregressive panel probit with random effects. Both DGPs exhibit a moderate level of persistence with a first autocorrelation around 0.5.<sup>14</sup> The parameters are chosen so that the signal-to-noise ratios are close to 1 and the frequency of the discrete outcomes are even, i.e. close to 33%, in the simulated data sets.

Table 1 presents summary statistics of the Monte Carlo simulations, in particular, the average bias (*bias*), the root mean squared errors (*rmse*), and the average estimated standard deviation of the estimators (*std*) – based on (17) and (18). To assess the finite sample properties of the asymptotic distribution, we calculate the proportion of the replications in which the true parameter is not contained in the 95% symmetric confidence intervals. It is given in the variable *size* representing the 5% two-sided rejection probabilities at the true values. We expect the empirical size of a test for a true hypothesis to be close to the nominal size of 5%. For readability, all of the summary statistics in Table 1 are multiplied by 100.

In each Monte Carlo replication, the MCL estimation algorithm is started at random initial points whereas, for the PCL, it is started in a random neighborhood of the MCL estimates. When, in these two benchmark DGPs, the initial points for the PCL estimator are randomly chosen, then in approximately 15% of the replications the algorithm does not converge and the estimator ends up with a slightly larger *rmse* (see the supplementary material for these simulation results).<sup>15</sup> Even though the results are robust to the choice of

<sup>14</sup>Note that the  $j^{\text{th}}$  autocorrelation in the latent variable  $y_{it}^*$  is given by  $\sigma_\alpha^2 + \rho^j(1 - \sigma_\alpha^2)$ .

<sup>15</sup>If an estimation algorithm results in an estimate that is outside of a 100% error band, then it is considered as non-convergent and started at a new initial point. For instance, if the true value of a parameter is 0.5, then any estimate that is outside of  $(0, 1)$  interval is regarded as non-converged estimate. If the true value is 0, then an estimate outside of  $(-1, 1)$  is regarded similarly. Then, we resimulate until convergence occurs.

the initial points, we start the estimation algorithms in each replication around the MCL estimates to facilitate the convergence of the PCL estimator. In this case, the PCL estimator converges in almost all simulations without any problem. Note that the composite likelihood is not necessarily strictly concave. Therefore, we suggest utilizing MCL estimates, which are already close to the true values, as an initial point for more stable and faster PCL estimation.

The simulation results in Table 1 suggest that both MCL and PCL estimators can uncover the true parameters in each DGP successfully in terms of *bias* and *rmse*. Especially the autocorrelation parameter  $\rho$  can be estimated with a small bias and with a high precision. In theory, the PCL estimator, which is using more information than the MCL estimator, should be more efficient. In the simulations, we can verify this for the parameter  $\rho$ , however, we see opposite results for the other parameters. In the AR panel probit with random effects model, this situation might be due to the estimation of an extra parameter in the PCL, namely  $\sigma_\alpha^2$ . Moreover, the estimation of  $\sigma_\alpha^2$  is not very precise for small  $T$  since it appears only in the correlation coefficient  $r_j(\theta)$  and it is hard to separate the effect of random effects from that of  $\rho y_{i,t-1}^*$ . Hence, this might contaminate the precision of the other parameter estimates in the random effects model. Yet, even in the AR panel probit without random effects model, we still see equal or larger *rmse* for the PCL estimates of the parameters  $(\mu, \beta, \tau)$  – with a slightly smaller *bias* – compared with those for the MCL estimates. Finally, regardless of the DGP and the estimator, doubling the sample size from 500 to 1000 makes the estimates more precise and decreases the estimated standard errors and *rmse*, on average, by a factor of  $\sqrt{2}$ , as the theory suggests.

In terms of the rejection probabilities of true hypotheses, the empirical sizes for both the MCL and PCL estimators are close to 5% in the AR Panel Probit model (both with and without random effects). For the MCL, the average *size* over all estimates is 5.45% where the maximum overrejection and the minimum underrejection occur at 7.2% and 4%, respectively. For the PCL, these statistics are 5.28%, 6.8% and 3.9%, respectively. Hence, the asymptotic distribution of CL estimators can be well approximated in the AR panel probit model. On the other hand, we do not see any significant improvement in the *size* values as  $N$  gets larger. Finally, although not reported here, the median estimates are very close to the mean estimates, suggesting that there are no significant outlier effects in the simulations.

TABLE 1: Summary Statistics of Monte Carlo Simulations for MCL & PCL Estimators

Autoregressive Panel Probit																									
$N = 500$																									
$T$	$J$	$\rho = 0.5$			$\mu = 0.5$			$\beta = 0.2$			$\tau = 1.0$														
		bias	rmse	std	size	bias	rmse	std	size	bias	rmse	std	size												
MCL	5	-0.26	6.79	6.88	5.40	0.78	3.61	3.55	5.10	0.49	1.68	1.66	5.90	-0.25	2.99	3.05	4.60								
MCL	10	0.04	4.57	4.45	6.70	0.24	2.76	2.67	5.90	0.18	1.21	1.18	5.80	-0.10	2.19	2.18	5.80								
MCL	30	0.03	2.46	2.42	5.50	0.13	1.59	1.61	4.90	0.08	0.68	0.68	5.10	-0.06	1.24	1.27	4.00								
PCL	5	0.06	2.62	2.60	5.60	0.86	3.90	3.79	5.60	0.41	1.72	1.68	5.90	-0.26	3.22	3.29	5.30								
PCL	10	0.03	2.17	2.18	4.50	0.28	3.03	2.94	5.70	0.17	1.28	1.24	5.90	-0.06	2.44	2.40	5.90								
PCL	30	0.04	1.70	1.69	5.70	0.12	1.75	1.79	4.90	0.08	0.73	0.73	4.90	-0.05	1.41	1.40	4.70								
$N = 1000$																									
MCL	5	-0.06	5.02	4.87	6.10	0.71	2.54	2.52	4.50	0.39	1.24	1.17	5.60	-0.38	2.19	2.16	5.30								
MCL	10	-0.11	3.02	3.15	4.50	0.43	1.99	1.89	6.90	0.22	0.85	0.83	6.50	-0.18	1.52	1.54	5.00								
MCL	30	-0.14	1.80	1.72	5.80	0.07	1.12	1.14	4.00	0.08	0.49	0.48	5.80	-0.06	0.88	0.90	5.10								
PCL	5	0.19	1.84	1.84	5.60	0.77	2.73	2.68	5.20	0.32	1.22	1.19	5.30	-0.38	2.34	2.32	4.80								
PCL	10	0.00	1.51	1.54	5.10	0.44	2.16	2.08	6.80	0.16	0.89	0.88	4.00	-0.16	1.68	1.70	5.40								
PCL	30	0.00	1.24	1.20	5.70	0.04	1.27	1.26	4.80	0.05	0.52	0.51	6.60	-0.04	0.99	0.99	5.40								
Autoregressive Panel Probit with Random Effects																									
$N = 500$																									
$T$	$J$	$\rho = 0.3$			$\mu = 0.5$			$\beta = 0.2$			$\gamma = 0.5$			$\sigma_\alpha^2 = 0.3$			$\tau = 1.0$								
		bias	rmse	std	size	bias	rmse	std	size	bias	rmse	std	size	bias	rmse	std	size	bias	rmse	std	size				
MCL	5	-0.35	9.71	9.39	7.20	0.45	4.03	3.92	6.20	0.33	1.78	1.74	5.60	-0.15	5.96	5.93	5.50	0.24	3.30	3.30	9.39	7.20			
MCL	10	0.18	5.24	5.42	4.60	0.13	3.39	3.34	4.50	0.21	1.17	1.12	6.30	0.00	6.14	6.07	5.50	0.18	2.50	2.50	5.42	4.60			
MCL	30	-0.06	2.59	2.69	4.60	0.22	3.01	2.97	4.80	0.05	0.64	0.64	5.60	-0.21	8.40	8.45	4.80	0.02	1.61	1.61	2.69	4.60			
PCL	5	-0.14	7.50	8.85	5.90	0.45	4.20	4.09	5.20	0.28	1.91	1.90	4.50	0.03	5.84	6.21	4.60	-0.68	6.66	8.11	5.80	0.28	3.51	3.56	4.20
PCL	10	0.17	4.16	4.31	4.20	0.10	3.55	3.49	4.50	0.22	1.27	1.26	4.80	0.04	6.42	6.25	5.10	-0.49	3.02	3.08	4.10	0.15	2.71	2.62	6.10
PCL	30	0.00	2.31	2.39	4.40	0.23	3.05	3.03	5.30	0.04	0.71	0.70	5.80	-0.19	8.56	8.60	5.00	-0.08	1.87	1.78	5.80	0.00	1.71	1.68	5.30
$N = 1000$																									
MCL	5	0.14	6.78	6.70	6.10	0.27	2.72	2.77	4.10	0.21	1.19	1.23	4.30	0.06	4.10	4.19	4.90	0.04	2.32	2.32	6.70	6.10			
MCL	10	-0.11	3.90	3.83	6.20	0.15	2.34	2.36	4.80	0.14	0.85	0.79	6.80	0.04	4.31	4.29	5.30	0.04	1.77	1.77	3.83	6.20			
MCL	30	-0.04	1.86	1.89	5.60	0.05	2.14	2.10	5.20	0.05	0.47	0.45	5.60	0.07	6.30	5.97	5.90	0.06	1.11	1.11	1.89	5.60			
PCL	5	0.12	5.21	5.66	5.00	0.32	2.91	2.89	5.20	0.20	1.31	1.34	3.90	0.14	4.15	4.24	4.60	-0.53	4.75	4.95	5.20	0.06	2.52	2.52	4.90
PCL	10	0.08	3.03	3.03	4.80	0.16	2.47	2.47	5.80	0.12	0.92	0.89	5.90	0.02	4.41	4.42	5.20	-0.12	2.18	2.15	6.40	-0.08	1.93	1.85	6.50
PCL	30	0.00	1.69	1.69	5.60	0.03	2.18	2.14	5.10	0.06	0.50	0.49	5.00	0.07	6.44	6.08	6.40	-0.03	1.27	1.26	5.60	0.07	1.18	1.20	5.40

Notes: Each DGP is simulated 1000 times and each statistic is multiplied by 100 for readability. The terms *bias*, *rmse*, and *std* stand for the average bias, average root mean squared error, and average estimated standard error of the associated estimator, where the average is taken over 1000 Monte Carlo replications. The term *size* represents the 5% two-sided rejection probabilities for a hypothesis at the true value.



It is known that CLE is not efficient since it does not use all the available information. However, it is difficult to analytically compute the loss of efficiency in nonlinear complex models. Therefore, we present some experimental comparisons via Monte Carlo simulations. Whenever possible, we compare the finite sample performance (in terms of *rmse*) of the CLE to the Maximum Likelihood Estimator (MLE). Furthermore, since choosing orthogonality conditions is similar to choosing composite likelihoods, we include the Generalized Method of Moments estimator in the Monte Carlo simulations for further comparison. To the best of our knowledge, the performances of GMM and CLE have not been compared in the literature. The abbreviation GMM represents the Generalized Method of Moments Estimator with the identity weighting matrix whereas EGMM represents that with the efficient (optimal) weighting matrix, which is based on the inverse of the variance of the estimated moment conditions in GMM.

We consider four different DGPs:  $(\rho, \mu, \beta, \tau) = (0, 0.5, 0.5, 1)$  corresponding to a static panel probit;  $(\rho, \mu, \beta, \gamma, \sigma_\alpha^2, \tau) = (0, 0.5, 0.5, -0.5, 0.5, 1)$  corresponding to a static panel probit with random effects;  $(\rho, \mu, \beta, \tau) = (0.5, 0.5, 0.5, 1)$  corresponding to an autoregressive panel probit; and  $(\rho, \mu, \beta, \gamma, \sigma_\alpha^2, \tau) = (0.5, 0.5, 0.5, -0.5, 0.5, 1)$  corresponding to an autoregressive panel probit with random effects. The maximum likelihood estimation is performed in the static panel probit models (both with and without random effects). These are the DGPs where we can empirically measure the loss of the efficiency of the CLE. On the other hand, GMM is used in all of the DGPs.

In the static panel probit without random effects model, we obtain the MLE by maximizing the pooled log-likelihood  $\mathcal{L}_{pooled}(\theta) = N^{-1} \sum_{i=1}^N \sum_{t=1}^T \sum_{s=1}^S \mathbb{1}(y_{it} = s) \ln \mathbb{P}(y_{it} = s | x_{it})$  where  $\mathbb{P}(y_{it} = s | x_{it}) = \Phi(\tau_s - \mu - \beta'x_{it}) - \Phi(\tau_{s-1} - \mu - \beta'x_{it})$ . In the static panel probit with random effects model, we integrate out the random effects and obtain  $\mathcal{L}_{RE}(\theta) = N^{-1} \sum_{i=1}^N \ln \int f(y_{it} | \mathbf{x}_i, \alpha_i) \phi(\alpha_i) d\alpha_i$  where  $f(y_{it} | \mathbf{x}_i, \alpha_i) = \prod_{t=1}^T \prod_{s=1}^S \mathbb{P}(y_{it} = s | \mathbf{x}_i, \alpha_i)^{\mathbb{1}(y_{it}=s)}$ ,  $\mathbb{P}(y_{it} = s | \mathbf{x}_i, \alpha_i) = \Phi((\tau_s - \alpha_i - \beta'x_{it})/\sigma_\varepsilon) - \Phi((\tau_{s-1} - \alpha_i - \beta'x_{it})/\sigma_\varepsilon)$ , and  $\alpha_i | \mathbf{x}_i \sim \mathcal{N}(\mu + \gamma'\bar{\mathbf{x}}_i, \sigma_\alpha^2)$  such that  $\sigma_\alpha^2 + \sigma_\varepsilon^2 = 1$  (see [Hsiao \(2014\)](#) for further details). Then, we approximate the one-dimensional integral by Gauss-Hermite quadrature as suggested by [Butler and Moffitt \(1982\)](#) and [Guilkey and Murphy \(1993\)](#).<sup>16</sup>

---

<sup>16</sup>Gauss-Hermite quadrature is used for numerical integration. It approximates a specific type of integral as

For method of moments estimations, the first and second moments of  $y_{it}$  are computed by  $M_{1,i,t}(\theta) \equiv E[y_{it}|\mathbf{x}_i] = \sum_{s=1}^S s\mathbb{P}(y_{it} = s|\mathbf{x}_i; \theta)$ ,  $M_{2,i,t}(\theta) \equiv E[y_{it}^2|\mathbf{x}_i] = \sum_{s=1}^S s^2\mathbb{P}(y_{it} = s|\mathbf{x}_i; \theta)$ , and  $M_{3,i,t,t+j}(\theta) \equiv E[y_{it}y_{i,t+j}|\mathbf{x}_i] = \sum_{s_1=1}^S \sum_{s_2=1}^S s_1s_2\mathbb{P}(y_{it} = s_1, y_{i,t+j} = s_2|\mathbf{x}_i; \theta)$ . The associated probabilities are given in (8) and (10). Based on these three conditional moments, we generate the following orthogonality conditions. For a function  $h(\cdot)$  of  $\mathbf{x}_i$ ,

$$\begin{aligned} E[\{y_{it} - M_{1,i,t}(\theta)\}h(\mathbf{x}_i)] &= \mathbf{0} \\ E[\{(y_{it} - M_{1,i,t}(\theta))^2 - M_{2,i,t}(\theta)\}h(\mathbf{x}_i)] &= \mathbf{0} \\ E[\{[y_{it} - M_{1,i,t}(\theta)][y_{i,t+j} - M_{1,i,t+j}(\theta)] - [M_{3,i,t,t+j}(\theta) - M_{1,i,t}(\theta)M_{1,i,t+j}(\theta)]\}h(\mathbf{x}_i)] &= \mathbf{0}. \end{aligned}$$

Note that these orthogonality conditions hold for each  $t = 1, \dots, T$ . Furthermore, the last one holds also for each  $j = 1, \dots, J$ . Since  $\mathbf{x}_i$  is strictly exogenous,  $x_{ik}$  for  $k = 1, \dots, T$  satisfies all orthogonality conditions for each  $t = 1, \dots, T$ . Hence, one can potentially use each time dimension of the covariates as instruments for all time periods. That is, one can choose  $h(\mathbf{x}_i) = (1, \text{vec}(\mathbf{x}_i)')$  as a  $(TK + 1)$  dimensional vector of instruments. Hence, in total there are  $M = T(1 + TK) + T(1 + TK) + \sum_{j=1}^J (T - j)(1 + TK) = (T(J + 2) - J(J + 1)/2)(1 + TK)$  many orthogonality conditions. As a result, the number of the orthogonalities is of order  $T^2KJ$ .

We can reduce this number to an order of  $TKJ$  by using only the contemporaneous instruments such as  $h(\mathbf{x}_i) = (1, x'_{it})'$ , whose dimension is  $(1 + K) \times 1$ . The number of moments is very important for the *efficient* GMM estimation since one needs to invert a matrix with a dimension that is equal to the number of moments. For short panels with a single covariate, such that  $(T, J, K) = (5, 2, 1)$  or  $(T, J, K) = (10, 4, 1)$ , the total number of moments is  $M = 32$  and  $M = 88$ , respectively. If we increase the number of covariates to 5, then the number of moments becomes  $M = 96$  and  $M = 264$ , respectively. Taking the inverse of a 264-dimensional matrix results in unstable and imprecise estimation so that the *Efficient* GMM estimator starts being less efficient than the GMM with simply the identity matrix as its weighting matrix. Therefore, for the EGMM in the Monte Carlo simulations, whenever

---

follows:  $\int_{-\infty}^{\infty} h(z) \exp(-z^2) dz \cong \sum_{k=1}^K w_k h(z_k)$ , where nodes  $z_k$  are the zeros of  $k^{\text{th}}$  order Hermite polynomial and  $w_k$  are the corresponding weights. A table for the nodes and weights can be found in [Abramowitz and Stegun \(1965\)](#).

the number of moments exceeds 200, we take only the inverse of the diagonal elements of the weighting matrix and ignore the off-diagonals.<sup>17</sup>

Table 2 presents the Monte Carlo simulation results for the estimators MLE, MCL, PCL, GMM, and EGMM in static panel probit models (without and with random effects). In addition to *bias* and *rmse*, we present the empirical relative efficiency *ere* and the average estimation time per replication (CPU time). The relative efficiency of an estimator is calculated by dividing the *rmse* of the MLE by that of the estimator. Hence, a number smaller than 1 shows how much efficiency will be lost for a given estimator compared with the MLE.

In the static panel probit without random effects model, the median loss of efficiency of the MCL is negligible (0.15%) since the MCL and the MLE are basically maximizing the same pooled log-likelihood – except that the MCL additionally estimates the autocorrelation parameter  $\rho$ . The *rmse* of the PCL estimator is slightly larger than that of the MLE, resulting in an 8.8% median loss of efficiency. On the other hand, the median loss of efficiencies of the GMM and EGMM are around 30% and 20%, respectively. Note that, even in this simple example, we occasionally get smaller *rmse* for some GMM estimates compared with EGMM estimates, in particular when the number of moments is large. For instance, when there are 520 moments, the *rmse* of  $\hat{\rho}$  in the GMM is 2.06, whereas that in the EGMM is 2.16.

For the static panel probit with random effects model, the MCL and PCL estimators have a median loss of efficiency of 1% and 9%, respectively (similar to without random effects case). For the ML estimation with  $T = 5, 10,$  and  $30$ , we use 10, 20, and 60 Hermite points, respectively, for the Gauss-Hermite quadrature to approximate the single-dimensional integral due to the random effects. It turns out that, for the empirical efficiency of the MLE, it is crucial to have a large number of Hermite points as  $T$  increases (as suggested also in [Guilkey and Murphy \(1993\)](#)). For instance, when  $T = 30$ , if 40 Hermite points are used (instead of 60), then CL estimators turn out to have a smaller *rmse* than the MLE for a couple of parameters, where the efficiency gain of the CL estimators becomes around 4%. If, however, 20 Hermite points are employed, then the CL estimators can have up to 32% more efficiency compared with the MLE (see the results in the supplementary material).

---

<sup>17</sup>For larger panels, the number of moments might become extremely large. For instance, in our application, the average time dimension is equal to 48,  $J = 8$ , and  $K = 11$ . This results in 12,312 orthogonality conditions, even by using only contemporaneous orthogonalities.

TABLE 2: Monte Carlo Simulations to Compare Finite Sample Performances and Relative Efficiencies

<i>Static Panel Probit</i>			$\rho = 0.0$			$\mu = 0.5$			$\beta = 0.5$			$\tau = 1.0$			CPU time (in seconds)		
<i>T</i>	<i>J</i>	<i>M</i>	bias	rmse	ere	bias	rmse	ere	bias	rmse	ere	bias	rmse	ere	bias	rmse	ere
MLE	5	5	—	—	—	0.07	2.83	—	0.03	1.84	—	0.22	3.08	—	0.03	—	—
MCL	5	5	0.13	3.79	0.07	2.83	1.00	0.03	1.87	0.99	0.25	3.09	1.00	0.07	0.07	—	—
PCL	5	2	0.06	3.07	0.11	3.06	0.92	0.08	2.00	0.92	0.24	3.36	0.91	0.51	0.51	—	—
GMM	5	2	0.70	5.87	-0.43	4.28	0.66	-0.19	3.03	0.61	0.18	4.42	0.70	2.41	2.41	—	—
EGMM	5	2	0.49	5.17	0.03	4.12	0.69	0.03	2.01	0.91	0.80	3.36	0.92	3.83	3.83	—	—
MLE	10	10	—	—	—	0.00	2.03	—	0.03	1.35	—	0.07	2.20	—	0.04	—	—
MCL	10	4	-0.13	2.55	0.01	2.03	1.00	0.05	1.37	0.99	0.09	2.20	1.00	0.09	0.09	—	—
PCL	10	4	-0.09	2.33	0.00	2.22	0.91	0.07	1.51	0.89	0.11	2.46	0.90	1.48	1.48	—	—
GMM	10	4	0.08	3.89	-0.08	2.91	0.69	-0.03	2.06	0.65	0.04	3.05	0.72	8.59	8.59	—	—
EGMM	10	4	0.05	3.61	0.25	2.89	0.70	0.11	1.57	0.86	0.73	2.55	0.86	12.04	12.04	—	—
MLE	30	30	—	—	—	0.01	1.12	—	0.05	0.78	—	0.04	1.25	—	0.06	—	—
MCL	30	10	0.03	1.33	0.01	1.12	1.00	0.05	0.79	0.99	0.05	1.25	1.00	0.16	0.16	—	—
PCL	30	10	0.05	1.39	0.00	1.23	0.91	0.06	0.88	0.88	0.05	1.37	0.91	8.93	8.93	—	—
GMM	30	10	0.15	2.06	-0.08	1.62	0.69	0.03	1.09	0.71	0.06	1.55	0.81	74.64	74.64	—	—
EGMM	30	10	0.21	2.16	0.13	1.66	0.68	0.18	1.00	0.78	0.71	1.57	0.80	73.23	73.23	—	—

  

<i>Static Panel Probit with Random Effects</i>			$\rho = 0.0$			$\mu = 0.5$			$\beta = 0.5$			$\gamma = -0.5$			$\sigma_\alpha^2 = 0.5$			$\tau = 1.0$			CPU time (in seconds)			
<i>T</i>	<i>J</i>	<i>M</i>	bias	rmse	ere	bias	rmse	ere	bias	rmse	ere	bias	rmse	ere	bias	rmse	ere	bias	rmse	ere	bias	rmse	ere	
MLE	5	5	—	—	—	0.24	4.25	—	0.09	1.82	—	-0.34	5.06	—	-0.32	2.52	—	0.17	3.30	—	0.17	3.30	—	0.57
MCL	5	5	0.31	3.44	0.22	4.27	1.00	0.12	1.86	0.98	-0.53	5.40	0.94	—	—	—	—	0.14	3.32	0.99	0.14	3.32	0.99	0.13
PCL	5	2	0.19	3.41	0.21	4.37	0.97	0.18	2.00	0.91	-0.51	5.85	0.87	—	—	—	—	0.23	4.79	0.69	0.23	4.79	0.69	1.99
GMM	5	2	0.85	6.49	-0.67	5.74	0.74	-0.65	5.69	0.32	0.62	13.26	0.38	—	—	—	—	-0.61	6.87	0.48	-0.61	6.87	0.48	7.07
EGMM	5	2	0.00	5.71	0.37	5.38	0.79	-0.93	3.82	0.48	2.27	11.94	0.42	—	—	—	—	0.77	4.80	0.69	0.77	4.80	0.69	15.01
MLE	10	10	—	—	—	0.18	3.63	—	0.09	1.39	—	0.17	6.29	—	-0.16	2.02	—	0.26	2.58	—	0.26	2.58	—	1.97
MCL	10	4	-0.03	2.12	0.18	3.63	1.00	0.10	1.42	0.98	0.15	6.38	0.99	—	—	—	—	0.25	2.58	1.00	0.25	2.58	1.00	0.16
PCL	10	4	-0.05	2.25	0.19	3.82	0.95	0.14	1.76	0.79	0.11	6.49	0.97	—	—	—	—	0.20	2.79	0.92	0.20	2.79	0.92	4.06
GMM	10	4	0.34	3.94	-1.41	4.96	0.73	-0.06	6.20	0.22	0.20	21.40	0.29	—	—	—	—	-0.86	9.43	0.21	-0.86	9.43	0.21	25.35
EGMM	10	4	0.08	3.57	0.20	4.17	0.87	-0.40	2.55	0.54	2.33	12.84	0.49	—	—	—	—	-1.12	6.30	0.32	-1.12	6.30	0.32	37.40
MLE	30	30	—	—	—	-0.17	3.53	—	0.03	1.04	—	-0.25	10.36	—	-0.14	1.81	—	0.10	2.06	—	0.10	2.06	—	41.46
MCL	30	10	0.00	1.15	-0.19	3.54	1.00	0.02	1.08	0.97	-0.21	10.36	1.00	—	—	—	—	0.08	2.10	0.98	0.08	2.10	0.98	0.31
PCL	30	10	-0.02	1.31	-0.17	3.57	0.99	0.01	1.30	0.81	-0.29	10.42	0.99	—	—	—	—	0.09	2.16	0.95	0.09	2.16	0.95	18.62
GMM	30	10	0.67	6.80	-2.57	12.20	0.29	-0.68	7.04	0.15	4.64	45.86	0.23	—	—	—	—	-3.75	10.23	0.20	-3.75	10.23	0.20	250.23
EGMM	30	10	0.09	2.12	-0.58	4.10	0.86	-0.37	1.96	0.53	5.23	21.93	0.47	—	—	—	—	0.51	3.06	0.68	0.51	3.06	0.68	243.48

Notes: Each DGP is simulated 1000 times and the estimations are based on  $N = 500$ . The *bias* and *rmse* are multiplied by 100. The term *ere* represents empirical relative efficiency of estimators where the *rmse* of the MLE is taken as the benchmark efficiency. The number *M* represents the number of moments employed in the GMM and EGMM estimators. The CPU time represents average duration of an estimation per replication run in the MATLAB software performed on a personal laptop computer with a 2.6 GHz Intel Core i7 processor.

Even a 40-point approximation for a univariate normal distribution, i.e. for a single integral, can be inferior compared with the composite likelihood’s approximation for the full likelihood. As a result, if integrals in a full likelihood are not well approximated, then the efficiency loss of MLE can be substantial. This partially explains the poor performance of Simulated Maximum Likelihood (SML) estimator in similar complex models with high-dimensional integrals (see [Geweke et al. \(1994\)](#), [Geweke et al. \(1997\)](#), and [Patil et al. \(2017\)](#) for instance).

Regarding the GMM and EGMM, the median loss of efficiency increases substantially to 73% and 47%, respectively, making them an unreliable estimator for this model. In terms of the CPU times, the MCL estimator converges in a fraction of a second. It is even faster than the MLE in the model with random effects. The PCL estimator is slightly slower than the MCL one, yet it only takes a few seconds. However, method of moments estimations might take a significant amount of time.

Finally, the results for the autoregressive panel probit (without and with random effects) models are presented in [Table 3](#). The CL estimators have significantly smaller *bias* and *rmse* in most of the cases compared with the GMM and EGMM estimators. Note also that, in these models, we cannot employ MLE anymore without relying on simulation methods, such as SML or Bayesian techniques. The literature has already compared finite sample performances of these estimators with CLE in models similar to AR panel probit and found that CLE has a big computational advantage over both SML and MCMC techniques. Moreover, it even has smaller *rmse* in many settings.<sup>18</sup>

Overall, the Monte Carlo simulation results in [Tables 1, 2 and 3](#) show the MCL and PCL estimators’ satisfactory finite sample performances as well as their computational attractiveness. In relatively simpler models such as static panel probit, where efficient estimators are also available, the CL estimators’ loss of efficiency is at acceptable levels (overall, the MCL and PCL estimators’ median loss of efficiencies are at 1% and 9%, respectively). Moreover, if we move from a static probit to a more complex model such as the AR probit, where

---

<sup>18</sup>[Varin and Vidoni \(2006\)](#), [Bhat et al. \(2010\)](#), [Bhat \(2011\)](#), and [Patil et al. \(2017\)](#) compare SML and MCMC methods with CLE in various settings such as univariate AR probit model, or a static panel probit with autoregressive errors, or a multinomial probit with correlated random effects. All these models have similar likelihoods to the one analyzed here in the sense that it contains high-dimensional integral.

TABLE 3: Monte Carlo Simulations to compare GMM with CLE

<i>Autoregressive Panel Probit</i>															
	$T$	$J$	$M$	$\rho = 0.5$		$\mu = 0.5$		$\beta = 0.5$		$\tau = 1.0$					
				bias	rmse	bias	rmse	bias	rmse	bias	rmse				
MCL	5			0.79	3.13	1.13	3.75	-0.14	2.15	-2.33	4.10				
PCL	5	2		1.11	2.52	1.32	4.05	-0.24	2.28	-2.20	4.22				
GMM	5	2	32	-2.33	6.78	-23.21	23.60	1.11	3.15	-4.35	6.19				
EGMM	5	2	32	0.54	4.20	-24.47	24.70	0.09	2.21	-1.98	4.29				
MCL	10			0.21	1.94	0.42	2.93	0.02	1.52	-1.42	2.80				
PCL	10	4		0.35	1.77	0.52	3.25	-0.03	1.64	-1.24	2.95				
GMM	10	4	88	-2.59	5.63	-23.56	23.78	0.96	2.53	-3.34	4.84				
EGMM	10	4	88	-0.77	3.24	-24.04	24.18	0.39	1.63	-0.68	2.96				
MCL	30			0.01	1.04	0.18	1.73	0.02	0.83	-0.45	1.46				
PCL	30	10		0.01	1.06	0.15	1.87	0.05	0.91	-0.35	1.59				
GMM	30	10	520	-0.32	2.09	-24.90	24.94	0.02	1.24	-1.07	2.53				
EGMM	30	10	520	-0.45	1.87	-24.45	24.50	0.31	1.06	0.46	1.97				
<i>Autoregressive Panel Probit with Random Effects</i>															
	$T$	$J$	$M$	$\rho = 0.5$		$\mu = 0.5$		$\beta = 0.5$		$\gamma = -0.5$		$\sigma_\alpha^2 = 0.5$		$\tau = 1.0$	
				bias	rmse	bias	rmse	bias	rmse	bias	rmse	bias	rmse	bias	rmse
MCL	5			-0.05	3.16	1.02	4.59	-0.25	2.08	1.66	6.91	—	—	-2.40	4.42
PCL	5	2		0.24	3.07	1.14	4.76	-0.43	2.23	1.78	6.88	-0.44	5.57	-2.25	4.54
GMM	5	2	32	-7.38	13.04	-21.93	24.46	0.07	7.62	7.89	27.75	-0.26	15.98	-7.75	10.70
EGMM	5	2	32	-6.44	9.84	-21.42	22.38	-4.89	7.31	20.48	27.88	-3.29	11.76	-1.98	5.15
MCL	10			-0.02	1.92	0.58	4.05	0.03	1.59	0.85	7.51	—	—	-1.11	3.19
PCL	10	4		0.10	2.04	0.72	4.28	0.01	1.71	0.82	7.81	-0.26	6.36	-0.95	3.41
GMM	10	4	88	-2.49	6.77	-24.25	24.93	1.79	4.82	-3.75	22.10	-1.45	15.52	-5.99	9.21
EGMM	10	4	88	-2.58	4.68	-23.42	23.64	-0.93	2.91	10.04	20.09	-3.23	9.42	-0.49	3.97
MCL	30			0.04	0.89	0.09	3.69	0.10	1.12	0.34	10.58	—	—	-0.30	2.19
PCL	30	10		0.11	2.47	0.08	4.44	0.20	2.17	0.37	11.23	-0.40	3.56	0.24	6.29
GMM	30	10	520	-0.09	5.28	-25.13	29.82	0.02	8.33	-5.33	34.58	-0.75	10.99	-4.08	8.85
EGMM	30	10	520	-0.44	2.59	-24.56	24.73	0.18	2.66	1.92	22.13	-2.27	6.12	-0.20	4.22

*Notes:* Each DGP is simulated 1000 times and the estimations are based on  $N = 500$ . The *bias* and *rmse* are multiplied by 100. The number  $M$  represents the number of moments employed in the GMM and EGMM estimators.

efficient estimators are not feasible, there is no significant performance deterioration in the CL estimators, whereas we see considerable decline in that of the GMM.

## 6 A Large $N$ Moderate $T$ Application: Credit Ratings

In this section, we present an empirical illustration of the attractiveness of both the CL estimation technique and the AR panel probit model. We analyze credit ratings (long-term corporate bond ratings, specifically) controlling for firm-specific balance sheet ratios and business cycle variables in a large  $N$  and moderate  $T$  panel. The “moderate  $T$ ” aspect of the panel is important since there are, on average, 48 time periods in our unbalanced panel data (time series dimension of the firms ranges from 20 to 109 periods with an average of 48 periods). This means that it is impractical to estimate this model by simulation-based methods since its full likelihood contains, on average, 48-dimensional integral.

Credit ratings reflect a borrower’s creditworthiness and ability to meet its financial obligations. A branch of this literature utilizes balance sheet data, especially financial ratios, to explain bond ratings of corporates.<sup>19</sup> However, the static nature of these firm-level models has shifted the interest of the literature into dynamic and more sophisticated portfolio-level models, at a cost of not using firm-level data.<sup>20</sup> The AR panel probit model with random effects fits into the gap between static firm-level models and dynamic portfolio-level models.

Two important aspects of credit ratings are stability and non-Markovian behavior. [Cantor \(2001\)](#), [Standard and Poor’s \(2002\)](#), [Altman and Rijken \(2006\)](#), and [Cantor and Mann \(2007\)](#) explain why the stability of ratings is desired by investors and regulators, and thereby, implemented by the credit rating agencies. The rating stability is achieved by not changing the ratings immediately after seeing a temporary change in the financial situation of a firm; instead, the rating agencies prefer waiting until they think that the changes in the financials of a firm are permanent.<sup>21</sup>

The non-Markovian property of the ratings is well documented by many papers, such

---

<sup>19</sup>See [Horrigan \(1966\)](#), [Altman \(1968\)](#), [Blume et al. \(1998\)](#), [Amato and Furfine \(2004\)](#), [Pagratis and Stringa \(2009\)](#), [Alp \(2013\)](#), and [Baghai et al. \(2014\)](#)

<sup>20</sup>The relevant literature contain the following papers and references therein: [Lando and Skødeberg \(2002\)](#), [Wei \(2003\)](#), [Koopman and Lucas \(2005\)](#), [McNeil and Wendin \(2007\)](#), [Koopman and Lucas \(2008\)](#), [Feng et al. \(2008\)](#), [Koopman et al. \(2009\)](#), [Stefanescu et al. \(2009\)](#), and [Gagliardini and Gouriéroux \(2014\)](#). See also [Alam et al. \(2010\)](#) for a review of different credit risk models.

<sup>21</sup>[Standard and Poor’s \(2002, p. 41\)](#): The ideal is to rate through the cycle. There is no point in assigning high ratings to a company enjoying peak prosperity if that performance level is expected to be only temporary. Similarly, there is no need to lower ratings to reflect poor performance as long as one can reliably anticipate that better times are just around the corner.

as Lando and Skødeberg (2002), Christensen et al. (2004), Altman and Rijken (2006), and Frydman and Schuermann (2008). The credit rating agencies claim that firms within the same rating class are homogeneous in the sense that they have the same rating transition probabilities with respect to future rating movements. However, the aforementioned papers show that two firms with identical current credit ratings can have substantially different transition probabilities. This means that future rating transitions do not depend only on the current ratings but also on the whole rating history, which results in a non-Markovian process. Moreover, there is an empirical phenomenon called *rating momentum* which contributes to the non-Markovian behavior of the ratings.<sup>22</sup>

Based on the stability and non-Markovian characteristics of the credit ratings, the AR panel probit model will be a good candidate to model them since it allows for persistence and non-Markovian behavior in the ratings. In this model, the latent autoregressive variable  $y_{it}^*$  will represent the credit rating agency's view on the creditworthiness of a firm that depend on the entire history of the financial situation of a firm with an exponentially decreasing importance of the past.

Note that the rating agencies first assign a continuous internal rating to each firm, then put them into discrete rating brackets (see Van Gestel et al. (2007) and Standard and Poor's (2018) for more details on the rating process). In calculating a continuous rating, the rating agencies take their past views as well as the current financial situation of the firm and the economy into account. In the model, these are captured by the lagged latent variable  $\rho y_{i,t-1}^*$  and  $\beta' x_{it}$ , respectively. Here,  $\rho$  represents the persistence or 'the weight' of the past and  $\beta$  is the weight of the current variables. The higher the stability or persistence of the ratings the higher will be the coefficient  $\rho$ . The autoregressive property provides both the stability and non-Markovian behavior of credit ratings. Moreover, unobserved firm heterogeneities and idiosyncratic innovations are also taken into account by the model.

---

<sup>22</sup>*Rating momentum* represents the empirical finding that a downgrade is more likely to be followed by another downgrade, and an upgrade is more likely to be followed by another upgrade. Altman and Rijken (2006) explains this fact as partial adjustment of the ratings instead of immediately jumping to a rating that represents the actual credit quality of a firm.



## 6.1 Data

We use the Wharton Research Data Services and the Compustat database to obtain corporate balance sheet data and the S&P Long-Term Issuer Level ratings.<sup>23</sup> The corporate bond ratings are available at the monthly frequency, but we transform them into quarterly frequency by taking the last rating within each quarter to match the frequency of the accounting data. Moreover, we convert the letter ratings {CCC, B, BB, BBB, A, AA, AAA} into ordinal numbers {1, . . . , 7} where CCC=1, B=2, BB=3 and so on.<sup>24</sup>

Based on the literature and *Key Financial Ratios* published by the credit rating agencies (see [Standard and Poor's \(2013\)](#)), we use the following firm-level financial variables with their abbreviations given in parentheses: net income/total assets (*roa*), retained earnings/total assets (*reta*), relative market value (*size*), binary variable for paying dividends (*dividend*), market-to-book ratio (*m/b*), cash flow/total assets (*cash*), capital expenditures/total assets (*capex*), total debt/total assets (*debt*). These ratios capture, in general, the profitability, leverage, financial soundness, liquidity, market power, and valuation of a firm.<sup>25</sup>

We use an unbalanced panel data set for the period 1989Q4–2016Q4. A firm is allowed to enter the data set after the initial date 1989Q4 or leave it before 2016Q4. But note that the parameter estimation in an autoregressive model relies heavily on the dependency between time periods and information accumulation over time. Therefore, we only include firms that do not have any missing data once they entered the data set until they leave. Since the firms with a short span of data are not representative and exhibit large variations, we exclude firms that have less than 5 years of quarterly data at least.<sup>26</sup> Eventually, the data set comprises 88,232 observations containing 1852 firms with an average of 48 quarters. Table 4 shows the total number of observations of each rating class and their associated frequencies in the data

---

<sup>23</sup>Throughout the paper, the term “credit rating of a firm” is frequently used; nevertheless, the corporate bond that is issued by the obligor receives a rating, rather than the obligor itself. An obligor can issue several bonds, but each issue might have a different rating. However, senior unsecured long-term bond ratings, which are considered in this paper, are close to the issuer rating since the debt defaults only when the issuer defaults. Therefore, Long-Term Issuer Level ratings reflect de facto the creditworthiness of the obligor.

<sup>24</sup>Note that we group the ratings without considering the notches +/-. For instance, {AA-, AA, AA+} all belong to a single category denoted by AA. Observations with D (Default), SD (Suspended), or NM (Not meaningful) ratings are excluded.

<sup>25</sup>For more details on the financial ratios, please see the data appendix in the supplementary material.

<sup>26</sup>The results are robust to the choice of minimum number of uninterrupted observations. For example, requiring a minimum of 1 year of observations yields similar results.

set. While around one-third of the ratings are BBB, firms with AAA or CCC constitute only around 1% of the entire data set.

TABLE 4: The Distribution of the Ratings

Ratings	# of Obs.	Percentage
CCC	1258	1.43%
B	14,223	16.12%
BB	21,735	24.63%
BBB	27,916	31.64%
A	18,244	20.68%
AA	4059	4.60%
AAA	797	0.90%

*Notes:* This table reports the distribution of ratings in the unbalanced panel data set. The data contain 1852 firms with at least 20 continuous quarters of observations between 1989Q4–2016Q4. Hence, the time dimension of a firm is minimum 20 quarters and maximum 109 quarters, with an average of 48 quarters. In total, there are 88,232 firm-quarter observations.

During the rating assignment process, credit rating agencies take also the current macro-financial conditions of the economy into account ([Standard and Poor’s \(2002\)](#)). The literature uses various choices of business cycle variables to control for the state of the economy. The NBER recession indicator (denoted by *recession*) seems to be the most common choice; however, the choice of other aggregate macroeconomic variables differs from paper to paper. While some papers use GDP growth rate ([Feng et al. \(2008\)](#), [Koopman et al. \(2009\)](#), and [Alp \(2013\)](#)), others create their own business cycle indicator ([Amato and Furfine \(2004\)](#) and [Freitag \(2015\)](#)). Hence, it is not certain which business cycle variable should be included in the analysis. For this reason, we prefer extracting principal components from a large macro-financial data set and use them as the business cycle variables.

The first two principal components (denoted by  $f_{1t}$  and  $f_{2t}$ ) of the quarterly data set of [McCracken and Ng \(2016\)](#) explain more than 20% of the total variation in 218 business cycle variables over the period 1971Q2–2016Q4. They are especially related to the real economy sector. For instance, they explain around 70% of the variation in real variables such as output, exports, imports, personal income, private investment, and housing starts. After estimating the business cycle factors, we took their corresponding dates that match with the data range of the credit ratings, i.e. 1989Q4–2016Q4. Note that, due to the *generated*

*regressor* problem, using estimated variables might result in larger standard errors or even inconsistent estimates in non-linear models. However, satisfying the conditions in [Bai and Ng \(2008\)](#), the estimated factors in our case do not create any problem since they can be treated as if they are the true factors.<sup>27</sup>

## 6.2 Estimation Results

In this subsection, we present the MCL and PCL estimation results of the AR panel probit model. Just to show potential gains in moving from a static probit model to an autoregressive one, we also present the maximum likelihood estimation results of a static panel probit model, which is the working-horse model in the credit rating literature (for instance, see [Blume et al. \(1998\)](#), [Amato and Furfine \(2004\)](#), [Alp \(2013\)](#), [Baghai et al. \(2014\)](#), and [Dimitrov et al. \(2015\)](#)).<sup>28</sup>

Table 6 presents the ML, MCL, and PCL(4) estimates and some associated summary statistics. PCL(4) stands for the PCL estimation with  $J = 4$ . The estimation results are robust to the choice of  $J$ .<sup>29</sup> The underscored numbers denote the insignificant estimates at the 5% significance level (the estimated standard errors are left to the supplementary material). In the models with random effects, the results for the correlated random effects' conditional mean parameter  $\gamma$  are not presented. Note that to deal with local maxima problems, we start estimation algorithms at several different initial points. Finally, due to computational time constraints, we end up using 41 Hermite points for the ML estimation in the static model with random effects.<sup>30</sup>

---

<sup>27</sup>[Bai and Ng \(2008\)](#) state that an estimated factor  $\hat{f}_t$  can be treated as the true factor  $f_t$  such that  $\tilde{T}^{-1} \sum_{t=1}^{\tilde{T}} \|\hat{f}_t - f_t\|^2 = O_p(1/\min[\tilde{N}, \tilde{T}])$  provided that  $\hat{f}_t$  is extracted from a large panel satisfying  $\tilde{N} \rightarrow \infty$ ,  $\tilde{T} \rightarrow \infty$ , and  $\tilde{T}^{5/8}/\tilde{N} \rightarrow 0$ . In our case, the principal components are extracted from a panel with  $\tilde{N} = 218$  and  $\tilde{T} = 183$ , which satisfy the sufficient conditions.

<sup>28</sup>Note that, in this literature, static panel probit without random effects is the most common model since the random effects model is computationally challenging given the large and extremely unbalanced nature of the data sets in this literature. Moreover, some of these aforementioned papers employ a static panel probit with fixed effects model, whose estimates are in fact inconsistent for fixed  $T$ .

<sup>29</sup>The estimations with  $J \in \{1, 2, 4, 6, 8\}$  yield very similar estimates and they are provided in the supplementary material.

<sup>30</sup>For such a large panel data, fitting a static probit with random effects model is extremely time-consuming. We encountered several problems in estimating the random effects model in the software STATA. But we were able to estimate it in MATLAB. Yet, we could not use a large number of Hermite points; otherwise the estimation would take tens of hours. Based on the Monte Carlo simulation results presented in Table 2, we know that using 41 Hermite points is unlikely to deliver an efficient ML estimation in a model with

The most important difference between the static and autoregressive models is the significant and economically large persistence parameter  $\hat{\rho}$ . It indicates that there is, in fact, a need for a time series component in analyzing credit ratings. In the autoregressive model without random effects, the estimated autocorrelation coefficient  $\hat{\rho}$  is 0.592 and 0.967 in the MCL and PCL estimations, respectively. After controlling for the random effects, the estimates become 0.681 and 0.718, respectively.

Given that both MCL and PCL estimators are consistent, why do we obtain such a different estimate like 0.967 versus the estimates around 0.70? The estimate 0.967 is not an error, but instead, it shows how utilizing lower dimensional distributions provides robustness under misspecification. In the model without random effects,  $\rho$  is the only source of the autocorrelation, and the PCL estimation indicates that ratings have extremely high persistence. However, in the model with the random effects, there are two sources for the autocorrelation:  $\rho$  and  $\sigma_\alpha^2$ . Hence, after controlling for the random effects, the estimated  $\rho$  returns to 0.718.

Note that the implied first autocorrelation coefficient in the random effects model is  $\sigma_\alpha^2 + \rho(1 - \sigma_\alpha^2) = 0.905 + 0.718(1 - 0.905) = 0.973$ , which is not far from the estimated autocorrelation coefficient 0.967 in the model without random effects. Hence, in this model, we can see the composition of the autocorrelation present in the latent process  $y_{it}^*$ . The random effects  $\alpha_i$ , whose conditional variance is estimated to be  $(1 - \hat{\rho})^2 \hat{\sigma}_\alpha^2 = 0.072$ , constitute a large portion of the autocorrelation, while  $\hat{\rho}$  also significantly contributes to the persistence.

A significant and economically large  $\rho$  shows that the credit rating agency's view on the creditworthiness of a company exhibits significant persistence. Any shock to the creditworthiness of a firm is estimated to have a half-life around 2 quarters. Even after a year,  $\hat{\rho}^4 \approx 25\%$  of a shock is affecting the underlying financial soundness of a firm. In the static model, on the contrary, there is no difference between short-term and long-term effects. The effects of any shock disappear within one time period. This is why the estimates of the financial ratios in the static model are larger than those in the corresponding autoregressive model. For instance, the  $\beta$  coefficients of the financial ratios in the MLE without random effects are

---

$T = 48$  on average. Thus, the CL estimation is possibly more efficient than the MLE in this case. This computational difficulty can explain partly why the credit ratings literature prefers the static model without random effects.

TABLE 5: Estimation Results

	Static Probit		Autoregressive Probit			
	No RE	With RE	No RE		With RE	
	MLE	MLE	MCL	PCL(4)	MCL	PCL(4)
$\rho$	—	—	0.592	0.967	0.681	0.718
<i>roa</i>	0.063	0.025	0.101	0.062	0.068	0.031
<i>reta</i>	0.254	0.229	0.106	<u>0.002</u>	0.065	0.036
<i>size</i>	0.615	0.370	0.253	0.022	0.100	0.072
<i>dividend</i>	1.160	0.528	0.508	0.045	0.145	0.149
<i>m/b</i>	0.162	0.082	0.066	<u>0.001</u>	0.031	0.071
<i>cash</i>	-0.115	-0.116	-0.050	-0.005	-0.032	-0.045
<i>capex</i>	-0.072	0.068	-0.041	-0.007	0.059	0.045
<i>debt</i>	-0.416	-0.241	-0.177	-0.026	-0.085	-0.083
<i>recession</i>	-0.073	0.000	-0.168	-0.170	-0.157	-0.030
$\hat{f}_1$	-0.057	-0.018	-0.080	-0.052	-0.060	-0.013
$\hat{f}_2$	<u>-0.012</u>	-0.037	<u>-0.007</u>	<u>0.003</u>	<u>-0.009</u>	0.013
$\sigma_\alpha^2$	—	0.673	—	—	—	0.905
<i>constant</i>	2.773	2.796	2.854	3.047	2.926	2.463
$\tau_2$	1.935	1.750	1.986	1.826	2.080	1.613
$\tau_3$	3.144	3.205	3.225	3.005	3.397	2.973
$\tau_4$	4.431	4.892	4.536	4.305	4.779	4.410
$\tau_5$	5.929	6.670	6.049	5.783	6.349	5.903
$\tau_6$	7.593	8.427	7.703	7.179	8.084	7.452
Correct Predictions						
Total	0.224	0.231	0.229	0.211	0.236	0.230
CCC	0.102	0.199	0.112	0.078	0.123	0.149
B	0.378	0.395	0.405	0.227	0.456	0.370
BB	0.560	0.581	0.568	0.549	0.567	0.581
BBB	0.732	0.793	0.731	0.755	0.727	0.763
A	0.334	0.267	0.356	0.308	0.393	0.326
AA	0.175	0.155	0.178	0.100	0.194	0.176
AAA	0.442	0.508	0.458	0.336	0.483	0.450
Transition Probabilities						
$ \mathbf{M}_{data} - \widehat{\mathbf{M}} $	0.419	0.173	0.214	0.087	0.106	0.072

*Notes:* This table presents the estimation results from an unbalanced panel with  $N = 1852$  and average  $T = 48$ . The total number of observations is 88,232. In the models with random effects, the estimates for the correlated random effects' conditional mean parameter  $\gamma$  are not presented. The estimates that are insignificant at the 5% level are underscored, where the estimated standard errors are given in the supplementary material. The matrix  $\mathbf{M}$  denotes the rating transition probability matrix ( $\mathbf{M}_{data}$  is the observed transition probability matrix, whereas  $\widehat{\mathbf{M}}$  is the estimated counterpart by the associated estimator).

around 2.3 times larger than the corresponding MCL estimates. Note that this magnitude is similar to the 6-quarter cumulative effect implied by the MCL estimates calculated by  $(1 + 0.592 + \dots + 0.592^5) = 2.34$ . In other words, we can conjecture that the coefficients in the static model capture medium-term effects.

In terms of in-sample prediction accuracy, the static and autoregressive models have comparable performances. Around 23% of the credit ratings are correctly predicted in each model. While there are some differences, the correct prediction within each rating class does not, overall, differ significantly between the static and autoregressive model. Hence, taking the autocorrelation into account does not increase the predictive power of the model.

However, it remarkably improves the accuracy of the implied rating transition probabilities. Note that the thresholds are the cut-off points that determine the switch between rating classes. But the estimated thresholds (and the constant term) are of similar magnitudes in static and autoregressive models. Therefore, a rating transition after a given shock is determined by only contemporaneous effects in the static model but by cumulative effects in the autoregressive one. The latter model, which allows for sluggish rating migration, provides superior rating transition estimates compared with the static models since it captures slow and partial rating adjustments that are observed in the data (Altman and Rijken (2006)).

Transition probabilities can be represented in rating transition matrices, which are useful in credit risk models to measure future credit loss (Nickell et al. (2000) and Bangia et al. (2002)). The summary of a performance comparison in terms of producing a rating transition matrix that is closer to the observed one is presented in the last line of Table 6. The matrix  $\mathbf{M}_{data}$  is the observed transition probability matrix throughout the entire data set, whereas the matrix  $\widehat{\mathbf{M}}$  is the estimated counterpart by the associated estimator in a given model. In computing  $\widehat{\mathbf{M}}$ , we utilize the implied transition probabilities computed from the estimates of each estimator in a given model. Based on the Euclidean normed distances  $|\mathbf{M}_{data} - \widehat{\mathbf{M}}|$ , we see that the autoregressive models predict rating migrations significantly better than the static ones.<sup>31</sup>

In Table 7, we present more details on the estimated transition probability matrices. In

---

<sup>31</sup>The Euclidean norm is used also in Bangia et al. (2002) for rating transition matrix comparison. However, since matrix norms are equivalent, the ordering of the normed differences is robust to the choice of norms.

TABLE 6: Rating Transition Probabilities

		CCC	B	BB	BBB	A	AA	AAA
Observed Data $\mathbf{M}_{data}$	CCC	92.6%	7.3%	0.1%	0	0	0	0
	B	1.6%	96.1%	2.4%	0	0	0	0
	BB	0	2.0%	96.8%	1.1%	0	0	0
	BBB	0	0.1%	1.1%	98.2%	0.6%	0	0
	A	0	0	0.1%	1.7%	98.1%	0.2%	0
	AA	0	0	0	0.1%	2.5%	97.4%	0
	AAA	0	0	0	0	0.1%	2.1%	97.8%
		CCC	B	BB	BBB	A	AA	AAA
Static Probit without Random Effects $ \mathbf{M}_{data} - \widehat{\mathbf{M}}_{MLE}  = 0.419$	CCC	87.3%	12.7%	0	0	0	0	0
	B	1.0%	90.9%	8.1%	0	0	0	0
	BB	0	3.9%	90.5%	5.7%	0	0	0
	BBB	0	0	2.9%	94.0%	3.1%	0	0
	A	0	0	0	8.7%	90.1%	1.2%	0
	AA	0	0	0	0	7.1%	89.6%	3.3%
	AAA	0	0	0	0	0	6.7%	93.3%
		CCC	B	BB	BBB	A	AA	AAA
Static Probit with Random Effects $ \mathbf{M}_{data} - \widehat{\mathbf{M}}_{MLE-RE}  = 0.173$	CCC	91.4%	8.6%	0	0	0	0	0
	B	1.4%	92.3%	6.2%	0	0	0	0
	BB	0	2.3%	93.4%	4.4%	0	0	0
	BBB	0	0	2.6%	96.0%	1.5%	0	0
	A	0	0	0	7.9%	90.9%	1.2%	0
	AA	0	0	0	0	7.1%	88.8%	4.1%
	AAA	0	0	0	0	0	6.4%	93.6%
		CCC	B	BB	BBB	A	AA	AAA
Autoregressive Probit with Random Effects $ \mathbf{M}_{data} - \widehat{\mathbf{M}}_{PCL(4)}  = 0.072$	CCC	95.7%	4.3%	0	0	0	0	0
	B	0.8%	95.6%	3.6%	0	0	0	0
	BB	0	1.8%	96.2%	2.0%	0	0	0
	BBB	0	0	1.1%	97.6%	1.3%	0	0
	A	0	0	0	3.8%	95.5%	0.7%	0
	AA	0	0	0	0	3.4%	94.8%	1.9%
	AAA	0	0	0	0	0	3.8%	96.2%

*Notes:* The matrix  $\mathbf{M}_{data}$  is the observed transition probability matrix throughout the entire data set, whereas the matrix  $\widehat{\mathbf{M}}$  is the estimated counterpart by the associated estimator in a given model. The matrices present rating migrations from the ratings listed in the rows towards those in the columns. For instance, in 1.6% of the cases in the observed data, a B-rated bond in a given quarter has become a CCC-rated bond in the following quarter.

particular, we present the data transition matrix  $\mathbf{M}_{data}$ , the results from the static probit model without random effects ( $\widehat{\mathbf{M}}_{MLE}$ ) and with random effects ( $\widehat{\mathbf{M}}_{MLE-RE}$ ) and the PCL(4) estimation results in the autoregressive model with random effects ( $\widehat{\mathbf{M}}_{PCL(4)}$ ). The rating transition is presented from the ratings in the rows towards those in the columns.

We see that, at the quarterly frequency, the ratings are highly sticky. For instance, in the

entire data set, a CCC rating in a given quarter stays in the same rating class in the following quarter in 92.6% of the times, migrates to the rating class B in 7.3% of the times, and to BB in only 0.1% of the times. The results in Table 7 indicate that the static probit without random effects model has a very poor prediction accuracy in terms of rating transitions. Even though controlling for random effects results in twice as good predictions (see the third transition matrix), it still cannot predict the persistence in the ratings successfully. Almost all of the diagonal estimates, which represents the probability of staying within the same rating class, are underestimated by the static model. On the other hand, the autoregressive model, which takes the rating persistence into account, produces substantially more accurate rating migration probabilities (more than twice as close to the observed rating migrations as those produced by static models).

Overall, the autoregressive probit model successfully captures the time series properties of the data, in particular the rating persistence and rating transitions. Even though the largest portion of the rating persistence arises from the random effects, taking the autocorrelation coefficient into account improves significantly the rating migration process.

## 7 Conclusion

In this paper, we are interested in modeling and estimating persistent discrete data. We focus on the autoregressive (AR) panel probit model with correlated random effects, where the discrete outcome variable is a non-linear function of an autocorrelated latent process. In this model, the persistence of the outcome is driven by the persistence of the underlying latent variable. The likelihood function of this model contains  $T$ -dimensional integrals, where  $T$  is the length of the panel data set.

The maximum likelihood estimation (MLE) of this intractable likelihood would be challenging or even infeasible. Hence, to solve this complex problem, we employ composite likelihood (CL) techniques borrowed from the statistics literature. In forming a composite likelihood, one utilizes lower dimensional distributions of the data, such as that of  $(y_{it}, y_{i,t+j})$  for  $j = 1, \dots, J$  rather than that of  $(y_{i1}, \dots, y_{iT})$ . Hence, the CL estimation loses some statistical efficiency but gains significant computational speed (even when the maximum likelihood



estimation is not feasible) and more robustness to misspecification.

Most common composite likelihoods are the *Marginal Composite Likelihood* (MCL), where univariate distributions of the data are modeled, and the *Pairwise Composite Likelihood* (PCL), where bivariate distributions are modeled. Utilizing only such sub-likelihoods and not using all the available information might, however, result in identification problems. Therefore, we especially focus on the identification issue of the CL estimators in the AR panel probit model and provide conditions under which the identification of the parameters can be achieved in this model. Our approach for identification is borrowed from the *average derivative estimation* literature. In particular, we show that certain derivatives of the associated sub-likelihoods can isolate parameters separately so that one can say that the composite likelihood contains enough information to identify the parameters. Moreover, we prove the consistency and asymptotic normality of the PCL estimator in a large  $N$  and fixed  $T$  setting.

To see the finite sample performance of the CL estimators, we conduct Monte Carlo simulations. Moreover, whenever possible, we compare the CLE with the MLE and the method of moments estimator (GMM). The simulations indicate that the composite likelihood estimation offers a fast, reliable, and accurate estimation with a small loss of efficiency. The GMM estimation suffers from the large number of moment conditions as  $T$  increases, and becomes significantly inferior compared with the CLE.

In the empirical part, we apply the AR probit model and the CL estimation techniques to analyze the corporate bond ratings. We use firm balance sheet data as well as some business cycle variables to explain the corporate bond ratings in a quarterly unbalanced panel data between 1989Q4–2016Q4 for 1852 firms. First we estimate a static panel probit (with and without random effects) by the MLE, which is the working-horse model of the firm-level credit ratings literature. Then we use the autoregressive panel probit (with and without random effects) estimated by the CLE to fit the data.

The estimated parameters show that there is significant persistence in the ratings that is captured both by random effects and autocorrelation property of the underlying process. Specifically, an autocorrelation parameter is estimated to be around 0.70 after controlling for random effects indicating a high degree of persistence overall in the discrete ratings. The AR probit model which takes the autocorrelation into account yields slightly better fitted values

but provides significantly better transition probabilities compared with static models.

## References

- Abramowitz, M. and I. A. Stegun (1965). Handbook of mathematical functions: with formulas, graphs, and mathematical tables. *Courier Corporation* 55.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Caski (Eds.), *Proc. Second International Symposium on Information Theory*, pp. 267–281. Budapest: Akademiai Kiado.
- Alam, M., C. Hao, and K. Carling (2010). Review of the literature on credit risk modeling: Development of the past 10 years. *Banks and Bank Systems* 5(3), 43–60.
- Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88(422), 669–679.
- Alp, A. (2013). Structural shifts in credit rating standards. *The Journal of Finance* 68(6), 2435–2470.
- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance* 23(4), 589–609.
- Altman, E. I. and D. L. Kao (1992). Rating drift in high-yield bonds. *The Journal of Fixed Income* 1(4), 15–20.
- Altman, E. I. and H. A. Rijken (2006). A point-in-time perspective on through-the-cycle ratings. *Financial Analysts Journal* 62(1), 54–70.
- Alvarez, J. and M. Arellano (2003). The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica* 71(4), 1121–1159.
- Amato, J. D. and C. H. Furfine (2004). Are credit ratings procyclical? *Journal of Banking & Finance* 28(11), 2641–2677.
- Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.
- Andrews, D. W. (1999). Consistent moment selection procedures for generalized method of moments estimation. *Econometrica* 67(3), 543–563.
- Arellano, M. (2016). Modelling optimal instrumental variables for dynamic panel data models. *Research in Economics* 70(2), 238–261.
- Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies* 58(2), 277–297.
- Arellano, M. and S. Bonhomme (2009). Robust priors in nonlinear panel data models. *Econometrica* 77(2), 489–536.
- Arulampalam, W., A. L. Booth, and M. P. Taylor (2000). Unemployment persistence. *Oxford Economic Papers* 52(1), 24–50.
- Baghai, R. P., H. Servaes, and A. Tamayo (2014). Have rating agencies become more conservative? Implications for capital structure and debt pricing. *The Journal of Finance* 69(5), 1961–2005.
- Bai, J. and S. Ng (2008). Extremum estimation when the predictors are estimated from large panels. *Annals of Economics and Finance* 9(2), 201–222.
- Bangia, A., F. X. Diebold, A. Kronimus, C. Schagen, and T. Schuermann (2002). Ratings

- migration and the business cycle, with application to credit portfolio stress testing. *Journal of Banking & Finance* 26(2), 445–474.
- Bel, K., D. Fok, and R. Paap (2018). Parameter estimation in multivariate logit models with many binary choices. *Econometric Reviews* 37(5), 534–550.
- Bernanke, B. S., M. Gertler, and S. Gilchrist (1999). The financial accelerator in a quantitative business cycle framework. *Handbook of Macroeconomics* 1, 1341–1393.
- Bertschek, I. and M. Lechner (1998). Convenient estimators for the panel probit model. *Journal of Econometrics* 87(2), 329–371.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 192–236.
- Bhat, C. R. (2011). The maximum approximate composite marginal likelihood (macml) estimation of multinomial probit-based unordered response choice models. *Transportation Research Part B: Methodological* 45(7), 923–939.
- Bhat, C. R., C. Varin, and N. Ferdous (2010). A comparison of the maximum simulated likelihood and composite marginal likelihood estimation approaches in the context of the multivariate ordered-response model. *Advances in Econometrics* 26, 65.
- Blanchard, O. J. and L. H. Summers (1986). Hysteresis in unemployment. *NBER Working Paper No. 2035*.
- Blume, M. E., F. Lim, and A. C. MacKinlay (1998). The declining credit quality of US corporate debt: Myth or reality? *The Journal of Finance* 53(4), 1389–1413.
- Blundell, R. and S. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87(1), 115–143.
- Börsch-Supan, A. and V. A. Hajivassiliou (1993). Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. *Journal of Econometrics* 58(3), 347–368.
- Butler, J. S. and R. Moffitt (1982). A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica* 50(3), 761–764.
- Canova, F. and C. Matthes (2016). A composite likelihood approach for dynamic structural models. Technical report.
- Cantor, R. (2001). Moodys investors service response to the consultative paper issued by the Basel Committee on Bank Supervision a new capital adequacy framework. *Journal of Banking & Finance* 25(1), 171–185.
- Cantor, R. and C. Mann (2007). Analyzing the tradeoff between ratings accuracy and stability. *The Journal of Fixed Income* 16(4), 60–68.
- Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics* 18(1), 5–46.
- Chamberlain, G. (1984). Panel data. *Handbook of econometrics* 2, 1247–1318.
- Chan, J. C., E. Eisenstat, C. Hou, and G. Koop (2018). Composite likelihood methods for large Bayesian VARs with stochastic volatility. *CAMA Working Paper No. 26/2018*.
- Chauvet, M. and S. Potter (2005). Forecasting recessions using the yield curve. *Journal of Forecasting* 24(2), 77–103.
- Christensen, J. H., E. Hansen, and D. Lando (2004). Confidence sets for continuous-time rating transition probabilities. *Journal of Banking & Finance* 28(11), 2575–2602.
- Chu, B. (2017). Composite quasi-likelihood estimation of dynamic panels with group-specific heterogeneity and spatially dependent errors. *arXiv preprint arXiv:1704.06613*.

- Cox, D. R. (1975). Partial likelihood. *Biometrika* 62(2), 269–276.
- Cox, D. R. and N. Reid (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* 91(3), 729–737.
- Creel, M. and D. Kristensen (2012). Estimation of dynamic latent variable models using simulated non-parametric moments. *The Econometrics Journal* 15(3), 490–515.
- Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press.
- Dhaene, G. and K. Jochmans (2015). Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies* 82(3), 991–1030.
- Dimitrov, V., D. Palia, and L. Tang (2015). Impact of the Dodd-Frank Act on credit ratings. *Journal of Financial Economics* 115(3), 505–520.
- Dueker, M. (2005). Dynamic forecasts of qualitative variables. *Journal of Business & Economic Statistics* 23(1), 96–104.
- Duffie, D. and K. J. Singleton (1993). Simulated moments estimation of Markov models of asset prices. *Econometrica* 61(4), 929–952.
- Feller, W. (1971). *An Introduction to Probability and Its Applications*, Volume II. Wiley, New York.
- Feng, D., C. Gouriéroux, and J. Jasiak (2008). The ordered qualitative model for credit rating transitions. *Journal of Empirical Finance* 15(1), 111–130.
- Fermanian, J.-D. and B. Salanié (2004). A nonparametric simulated maximum likelihood estimation method. *Econometric Theory* 20(04), 701–734.
- Freitag, L. (2015). Procyclicality and path dependence of sovereign credit ratings: The example of Europe. *Economic Notes* 44(2), 309–332.
- Frydman, H. and T. Schuermann (2008). Credit rating dynamics and Markov mixture models. *Journal of Banking & Finance* 32(6), 1062–1075.
- Gagliardini, P. and C. Gouriéroux (2005). Stochastic migration models with application to corporate risk. *Journal of Financial Econometrics* 3(2), 188–226.
- Gagliardini, P. and C. Gouriéroux (2014). Efficiency in large dynamic panel models with common factors. *Econometric Theory* 30(05), 961–1020.
- George, E. and G. Mudholkar (1983). On the convolution of logistic random variables. *Metrika* 30(1), 1–13.
- Gerfin, M. and M. Lechner (2002). A microeconomic evaluation of the active labour market policy in Switzerland. *The Economic Journal* 112(482), 854–893.
- Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 57(6), 1317–1339.
- Geweke, J. (1991). Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints. In *Computer Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, pp. 571–578.
- Geweke, J., M. Keane, and D. Runkle (1994). Alternative computational approaches to inference in the multinomial probit model. *The Review of Economics and Statistics*, 609–632.
- Geweke, J. F., M. P. Keane, and D. E. Runkle (1997). Statistical inference in the multinomial multiperiod probit model. *Journal of Econometrics* 80(1), 125–165.
- Gouriéroux, C. and A. Monfort (2018). Composite indirect inference with application to corporate risks. *Econometrics and Statistics* 7, 30–45.

- Gouriéroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood methods: Theory. *Econometrica* 52(3), 681–700.
- Greene, W. (2004a). The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *The Econometrics Journal* 7(1), 98–119.
- Greene, W. (2004b). Convenient estimators for the panel probit model: Further results. *Empirical Economics* 29(1), 21–47.
- Grether, D. M. and G. Maddala (1982). A time series model with qualitative variables. In *Games, Economic Dynamics, and Time Series Analysis*, pp. 291–305. Springer.
- Guilkey, D. K. and J. L. Murphy (1993). Estimation and testing in the random effects probit model. *Journal of Econometrics* 59(3), 301–317.
- Hahn, J. and G. Kuersteiner (2011). Bias reduction for dynamic nonlinear panel models with fixed effects. *Econometric Theory* 27(6), 1152–1191.
- Hajivassiliou, V. (1990). Smooth simulation estimation of panel data LDV models. Discussion Paper, Yale University.
- Hajivassiliou, V., D. McFadden, and P. Ruud (1996). Simulation of multivariate normal rect-angle probabilities and their derivatives: Theoretical and computational results. *Journal of Econometrics* 72(1–2), 85–134.
- Hajivassiliou, V. A. and D. L. McFadden (1998). The method of simulated scores for the estimation of LDV models. *Econometrica* 66(4), 863–896.
- Hajivassiliou, V. A. and P. A. Ruud (1994). Classical estimation methods for LDV models using simulation. *Handbook of Econometrics* 4, 2383–2441.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* 50(4), 1029–1054.
- Heckman, J. J. (1981). Statistical models for discrete panel data. In C. F. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 114–178. MIT Press, Cambridge.
- Heckman, J. J. and G. J. Borjas (1980). Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence. *Economica* 47(187), 247–283.
- Honoré, B. E. and E. Kyriazidou (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica* 68(4), 839–874.
- Honoré, B. E. and E. Tamer (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica* 74(3), 611–629.
- Horrigan, J. O. (1966). The determination of long-term credit standing with financial ratios. *Journal of Accounting Research*, 44–62.
- Hsiao, C. (2014). *Analysis of Panel Data*. Econometric Society Monographs, no. 54. Cambridge University Press.
- Hyslop, D. R. (1999). State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. *Econometrica* 67(6), 1255–1294.
- Joe, H. and Y. Lee (2009). On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis* 100(4), 670–685.
- Keane, M. P. (1994). A computationally practical simulation estimator for panel data. *Econometrica* 62(1), 95–116.
- Kiyotaki, N. and J. Moore (1997). Credit cycles. *Journal of Political Economy* 105(2),

211–248.

- Koop, G. and D. J. Poirier (1993). Bayesian analysis of logit models using natural conjugate priors. *Journal of Econometrics* 56(3), 323–340.
- Koopman, S. J., R. Kräussl, A. Lucas, and A. B. Monteiro (2009). Credit cycles and macro fundamentals. *Journal of Empirical Finance* 16(1), 42–54.
- Koopman, S. J. and A. Lucas (2005). Business and default cycles for credit risk. *Journal of Applied Econometrics* 20(2), 311–323.
- Koopman, S. J. and A. Lucas (2008). A non-Gaussian panel time series model for estimating and decomposing default risk. *Journal of Business & Economic Statistics* 26(4), 510–525.
- Koopman, S. J., A. Lucas, and B. Schwaab (2012). Dynamic factor models with macro, frailty, and industry effects for US default counts: The credit crisis of 2008. *Journal of Business & Economic Statistics* 30(4), 521–532.
- Kristensen, D. and Y. Shin (2012). Estimation of dynamic models with nonparametric simulated maximum likelihood. *Journal of Econometrics* 167(1), 76–94.
- Lando, D. and T. M. Skødeberg (2002). Analyzing rating transitions and rating drift with continuous observations. *Journal of Banking & Finance* 26(2), 423–444.
- Laroque, G. and B. Salanié (1993). Simulation-based estimation of models with lagged latent variables. *Journal of Applied Econometrics* 8(S1), S119–S133.
- Lechner, M. (1993). Estimation of limited dependent variable habit persistence models on panel data with an application to the dynamics of self-employment in the former East Germany. In H. Bunzel, P. Jensen, and N. Westergård-Nielson (Eds.), *Panel Data and Labour Market Dynamics*, pp. 263–283. Amsterdam: North-Holland.
- Lee, L.-F. (1997). Simulated maximum likelihood estimation of dynamic discrete choice statistical models: Some Monte Carlo results. *Journal of Econometrics* 82(1), 1–35.
- Lee, L.-F. (1999). Statistical inference with simulated likelihood functions. *Econometric Theory* 15(3), 337–360.
- Lerman, S. and C. Manski (1993). On the use of simulated frequencies to approximate choice probabilities. In C. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 305–319. MIT Press Cambridge, MA.
- Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* 80(1), 221–239.
- Lindsay, B. G., G. Y. Yi, and J. Sun (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica* 21, 71–105.
- McCracken, M. W. and S. Ng (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics* 34(4), 574–589.
- McCulloch, R. and P. E. Rossi (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics* 64(1-2), 207–240.
- McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*: 57(5), 995–1026.
- McFadden, D. and P. A. Ruud (1994). Estimation by simulation. *The Review of Economics and Statistics* 76(4), 591–608.
- McNeil, A. J. and J. P. Wendin (2007). Bayesian inference for generalized linear mixed models of portfolio credit risk. *Journal of Empirical Finance* 14(2), 131–149.
- Müller, G. and C. Czado (2005). An autoregressive ordered probit model with application to high-frequency financial data. *Journal of Computational and Graphical Statistics* 14(2),

- 320–338.
- Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica* 46(1), 69–85.
- Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics* 4, 2111–2245.
- Newey, W. K. and T. M. Stoker (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica* 61(5), 1199–1223.
- Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica* 16(1), 1–32.
- Nickell, P., W. Perraudin, and S. Varotto (2000). Stability of rating transitions. *Journal of Banking & Finance* 24(1), 203–227.
- Oh, D. H. and A. J. Patton (2016). High-dimensional copula-based distributions with mixed frequency data. *Journal of Econometrics* 193(2), 349–366.
- Ojo, M. O. (2003). A remark on the convolution of the generalized logistic random variables. *ASSET serves A* 1(2).
- Pagratis, S. and M. Stringa (2009). Modeling bank senior unsecured ratings: A reasoned structured approach to bank credit assessment. *International Journal of Central Banking* 5(2), 1–39.
- Pakel, C., N. Shephard, and K. Sheppard (2011). Nuisance parameters, composite likelihoods and a panel of GARCH models. *Statistica Sinica* 21, 307–329.
- Pakel, C., N. Shephard, K. Sheppard, and R. F. Engle (2017). Fitting vast dimensional time-varying covariance models. *Working Paper*.
- Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* 57(5), 1027–1057.
- Patil, P. N., S. K. Dubey, A. R. Pinjari, E. Cherchi, R. Daziano, and C. R. Bhat (2017). Simulation evaluation of emerging estimation techniques for multinomial probit models. *Journal of Choice Modelling* 23, 9–20.
- Plackett, R. L. (1954). A reduction formula for normal multivariate integrals. *Biometrika* 41(3/4), 351–360.
- Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica* 57(6), 1403–1430.
- Pudney, S. (2008). The dynamics of perception: Modelling subjective wellbeing in a short panel. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(1), 21–40.
- Qu, Z. (2018). A composite likelihood framework for analyzing singular DSGE models. *Review of Economics and Statistics* 100(5), 916–932.
- Roodman, D. (2009). A note on the theme of too many instruments. *Oxford Bulletin of Economics and Statistics* 71(1), 135–158.
- Semykina, A. and J. M. Wooldridge (2013). Estimation of dynamic panel data models with sample selection. *Journal of Applied Econometrics* 28(1), 47–61.
- Standard and Poor’s (2002). Corporate ratings criteria. *Ratings Direct*.
- Standard and Poor’s (2013). Corporate methodology: Ratios and adjustments. *Ratings Direct*.
- Standard and Poor’s (2018). Guide to credit rating essentials. *S&P Global Ratings*.
- Stefanescu, C., R. Tunaru, and S. Turnbull (2009). The credit rating process and estimation

- of transition probabilities: A Bayesian approach. *Journal of Empirical Finance* 16(2), 216–234.
- Stegmueller, D. (2013). Modeling dynamic preferences: A Bayesian robust dynamic latent ordered probit model. *Political Analysis* 21(3), 314–333.
- Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica* 54(6), 1461–1481.
- Stoker, T. M. (1991). Equivalence of direct, indirect, and slope estimators of average derivatives. In W. A. Barnett, J. L. Powell, and G. Tauchen (Eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, pp. 99–118. Cambridge University Press.
- Takeuchi, K. (1976). The distribution of information statistics and the criterion of goodness of fit of models. *Mathematical Science* 153, 12–18.
- Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press.
- Tuzcuoglu, K. (2017). *Three Essays in Econometrics*. Ph. D. thesis, Columbia University.
- Van Gestel, T., D. Martens, B. Baesens, D. Feremans, J. Huysmans, and J. Vanthienen (2007). Forecasting and analyzing insurance companies’ ratings. *International Journal of Forecasting* 23(3), 513–529.
- Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica* 21, 5–42.
- Varin, C. and P. Vidoni (2005). A note on composite likelihood inference and model selection. *Biometrika* 92(3), 519–528.
- Varin, C. and P. Vidoni (2006). Pairwise likelihood inference for ordinal categorical time series. *Computational Statistics & Data Analysis* 51(4), 2365–2373.
- Varin, C. and P. Vidoni (2008). Pairwise likelihood inference for general state space models. *Econometric Reviews* 28(1–3), 170–185.
- Wei, J. Z. (2003). A multi-factor, credit migration model for sovereign and corporate debts. *Journal of International Money and Finance* 22(5), 709–735.
- Wei, S. X. (1999). A Bayesian approach to dynamic Tobit models. *Econometric Reviews* 18(4), 417–439.
- Woodford, M. (1999). Optimal monetary policy inertia. *The Manchester School* 67, 1–35.
- Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* 20(1), 39–54.



# A Technical Appendix

First, let's introduce some notation to facilitate the readability. We denote the conditional mean of the latent process  $y_{it}^*$  as  $m_{it} \equiv m_{it}(\theta) = m_t(\mathbf{x}_i; \theta) = \mu + \gamma' \bar{\mathbf{x}}_i + \sum_{k=1}^t \rho^{t-k} \beta' x_{ik}$  such that  $y_{it}^* | \mathbf{x}_i \sim \mathcal{N}(m_{it}, 1)$ . Let  $q_{its} \equiv \tau_s - m_{it}$  denote the threshold net of the conditional mean. The first and second subscripts of bivariate functions will be associated with time periods  $t$  and  $t + j$ , respectively. Moreover, for simplicity, we will abuse the notation and suppress the subscripts  $i$ ,  $t$ , and  $t + j$  whenever it does not create ambiguity:  $\mathbb{P}_{s_1 s_2} \equiv \mathbb{P}(y_{it} = s_1, y_{i,t+j} = s_2 | \mathbf{x}_i; \theta)$  and  $\Phi_{s_1 s_2} \equiv \Phi_2(q_{ts_1}, q_{t+j, s_2} | r_j) \equiv \Phi_2(\tau_{s_1} - m_{it}, \tau_{s_2} - m_{i,t+j} | r_j)$ .

Next, let's rewrite the pairwise composite log-likelihood and the associated probabilities with the recently introduced notation.

$$\begin{aligned} \mathcal{L}_{PCL}(\theta | \mathbf{y}, \mathbf{x}) &= \frac{1}{N} \sum_{i=1}^N \ell_i(\theta), \\ \ell_i(\theta) \equiv \ell(\theta | \mathbf{y}_i, \mathbf{x}_i) &= \frac{1}{T} \sum_{t=1}^{T-J} \sum_{j=1}^J \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta), \\ \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta) &= \sum_{s_1=1}^S \sum_{s_2=1}^S \mathbb{1}_{s_1 s_2} \ln \mathbb{P}_{s_1 s_2}(\theta), \\ \mathbb{P}_{s_1 s_2}(\theta) &= \Phi_2(q_{ts_1}, q_{t+j, s_2} | r_j) - \Phi_2(q_{t, s_1-1}, q_{t+j, s_2} | r_j) \\ &\quad - \Phi_2(q_{ts_1}, q_{t+j, s_2-1} | r_j) + \Phi_2(q_{t+j, s_2-1}, q_{t+j, s_2-1} | r_j), \\ &= \Phi_{s_1, s_2} - \Phi_{s_1-1, s_2} - \Phi_{s_1, s_2-1} + \Phi_{s_1-1, s_2-1}. \end{aligned}$$

## A.1 The Proof of Theorem 1

We first prove the consistency and then the asymptotic normality of the PCL estimator. For consistency, we prove the uniform boundedness of the composite likelihood by finding the bounds of bivariate probabilities. For the asymptotic normality, we compute the score, its asymptotic distribution, the Hessian and its uniform bounds. The flow of the proof is standard in the sense that it is similar to the asymptotic results for quasi- or pseudo-maximum likelihood estimators. Hence, we leave many details to the supplementary material and show here the most important points and differences compared with static probit models.

### A.1.1 The Consistency of the PCLE

The lemma below proves the uniform boundedness of the composite likelihood.

**Lemma 1.** *Under the assumptions of Theorem 1, the logarithm of the bivariate distribution function is uniformly bounded. That is,  $E[\sup_{\theta \in \Theta} |\ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta)|] < \infty$ .*

**Proof of Lemma 1.** *Let's analyze the bivariate probability  $\mathbb{P}_{11}(\theta) = \Phi_2(q_{t1}, q_{t+j,1} | r_j)$  and use the mean value expansion, where  $q_{t1} = -(\mu + \gamma' \bar{\mathbf{x}}_i + \sum_{k=1}^t \rho^{t-k} \beta' x_{ik})$  since  $\tau_1 = 0$ . For a  $\bar{\theta}$  between  $\theta$  and 0, we can write*

$$|\ln \mathbb{P}_{11}(\theta)| = \left| \ln \mathbb{P}_{11}(0) + \theta' \frac{\partial \ln \mathbb{P}_{11}(\bar{\theta})}{\partial \theta} \right| = \left| \ln \Phi_2(0, 0 | 0) + \theta' \frac{\frac{\partial \mathbb{P}_{11}(\bar{\theta})}{\partial \theta}}{\mathbb{P}_{11}(\bar{\theta})} \right| \leq \ln(1/4) + |\theta| \left| \frac{\frac{\partial \mathbb{P}_{11}(\bar{\theta})}{\partial \theta}}{\mathbb{P}_{11}(\bar{\theta})} \right|$$

*Let's focus on the last norm. Using the equation (24) in the section B, which provides the first derivative of  $P_{11}$ , yields*

$$\begin{aligned} \left| \frac{\frac{\partial}{\partial \theta} \mathbb{P}_{11}}{\mathbb{P}_{11}} \right| &= \left| \frac{\frac{\partial}{\partial \theta} \Phi(q_{t1}, q_{t+j,1} | r_j)}{\Phi(q_{t1}, q_{t+j,1} | r_j)} \right| \\ &\leq |q'_{t1}| \left| \frac{\phi(q_{t1}) \Phi\left(\frac{-r_j q_{t1} + q_{t+j,1}}{\sqrt{1-r_j^2}}\right)}{\Phi(q_{t1}, q_{t+j,1} | r_j)} \right| + |q'_{t+j,1}| \left| \frac{\phi(q_{t+j,1}) \Phi\left(\frac{q_{t1} - r_j q_{t+j,1}}{\sqrt{1-r_j^2}}\right)}{\Phi(q_{t1}, q_{t+j,1} | r_j)} \right| + |r'_j| \left| \frac{\phi_2(q_{t1}, q_{t+j,1} | r_j)}{\Phi(q_{t1}, q_{t+j,1} | r_j)} \right|, \end{aligned} \quad (19)$$

*where  $q'$  denotes the derivative with respect to  $\theta$ , i.e.  $q'_{t1} = \partial q_{t1}(\theta) / \partial \theta$  and  $\phi_2(\cdot, \cdot | r_j)$  denotes the bivariate standard normal density with the correlation coefficient  $r_j$ . We need to find upperbounds for each term in (19). The idea of these upperbounds can be understood when one considers the special case of no autocorrelation. If  $r_j = 0$ , then bounds for the ratios in (19) become  $(\phi(q_{t1}) / \Phi(q_{t1})) \leq c(1 + |q_{t1}|)$ ,  $(\phi(q_{t+j,1}) / \Phi(q_{t+j,1})) \leq c(1 + |q_{t+j,1}|)$ , and  $(\phi(q_{t1}) \phi(q_{t+j,1}) / \Phi(q_{t1}) \Phi(q_{t+j,1})) \leq c(1 + \max\{q_{t1}^2, q_{t+j,1}^2\})$ , respectively, for some positive constant  $c$ . These are the inverse Mills ratios that appear in a static probit model. A non-zero  $r_j$  does not change the limiting behavior of these ratios, but it makes the proof more complex. The details for the case when  $r_j \neq 0$  are given in the supplementary material.*

*The same upperbounds are found for the generic term involving  $\mathbb{P}_{s_1 s_2}$  (see the supplementary material for details). Without loss of generality, let's assume that  $|q_{t+j,s_2}| \leq |q_{t,s_1}|$  and*

$|q'_{t+j,s_2}| \leq |q'_{t,s_1}|$ , for all  $s_1$  and  $s_2$ . Then,  $|\mathbb{P}'_{s_1 s_2} / \mathbb{P}_{s_1 s_2}| \leq C_{s_1, s_2} (|q'_{t,s_1}|(1+|q_{t,s_1}|) + |r'_j|(1+q_{t,s_1}^2))$ , for some constant  $C_{s_1, s_2} > 0$ . Thus, the limiting behavior of the ratios involving normal pdf and cdf is common for each probability.

To find an upperbound for  $\mathbb{E} [\sup_{\theta \in \Theta} |\ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta)|]$ , we use the upperbounds of functions of  $q_{its}(\theta)$  derived in B.2, and utilize Minkowski's inequality, Hölder inequality, and Loève's  $c_r$ -Inequality.<sup>32</sup> We obtain

$$\begin{aligned}
\mathbb{E} \left[ \sup_{\theta \in \Theta} |\ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta)| \right] &= \left\| \sup_{\theta \in \Theta} \left| \sum_{s_1=1}^S \sum_{s_2=1}^S \mathbb{1}_{s_1 s_2} \ln \mathbb{P}_{s_1 s_2}(\theta) \right| \right\|_1 \leq \sum_{s_1=1}^S \sum_{s_2=1}^S \left\| \sup_{\theta \in \Theta} |\ln \mathbb{P}_{s_1 s_2}(\theta)| \right\|_1 \\
&\leq c + \sum_{s_1=1}^S \sum_{s_2=1}^S C_{s_1, s_2} \left\| \sup_{\theta \in \Theta} |q'_{ts_1}(\theta)|(1 + |q_{ts_1}(\theta)|) + |r'_j(\theta)|(1 + q_{ts_1}^2(\theta)) \right\|_1 \\
&\leq c + C \sum_{s_1=1}^S \sum_{s_2=1}^S \left\| \sup_{\theta \in \Theta} |q'_{ts_1}(\theta)|(1 + |q_{ts_1}(\theta)|) \right\|_1 + \left\| \sup_{\theta \in \Theta} |r'_j(\theta)|(1 + q_{ts_1}^2(\theta)) \right\|_1 \\
&\leq c + C \sum_{s_1=1}^S \sum_{s_2=1}^S \Delta_{2,2} \left\| \sup_{\theta \in \Theta} |q'_{ts_1}(\theta)| \right\|_2 \left\| \sup_{\theta \in \Theta} |q_{ts_1}(\theta)| + 1 \right\|_2 + \left\| J \sup_{\theta \in \Theta} (1 + q_{ts_1}^2(\theta)) \right\|_1 \\
&\leq c + C \sum_{s_1=1}^S \sum_{s_2=1}^S \Delta_{2,2} \left\| c'(1 + |\mathbf{x}_i|) \right\|_2 \left\| c'(1 + |\mathbf{x}_i|) \right\|_2 + \left\| c'(1 + |\mathbf{x}_i|)^2 \right\|_1 \\
&\leq C' (1 + \|\mathbf{x}_i\|_2^2 + \|\mathbf{x}_i\|_1),
\end{aligned}$$

where  $\Delta_{2,2} > 0$  is the constant term given in Loève's  $c_r$ -Inequality and  $c, c', C, C'$  are some positive constants. Hence,  $\mathbb{E} [\sup_{\theta \in \Theta} |\ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta)|] < \infty$  as long as  $\mathbb{E} |\mathbf{x}_i|^2 < \infty$ . This condition is satisfied by Assumption 3, i.e. by  $\mathbb{E} [\mathbf{x}_i \mathbf{x}_i'] < \infty$ .

As a result, Lemma 1 and the analysis in the identification section 3, which proves the uniqueness of the true parameter as the maximizer, i.e.  $\mathbb{E} [\ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta^*)] > \mathbb{E} [\ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta)]$  for all  $\theta \in \Theta$  with  $\theta \neq \theta^*$ , constitute the consistency proof for  $\hat{\theta}$ .

### A.1.2 The Asymptotic Normality of the PCLE

In this section, we analyze the asymptotic distribution of  $\hat{\theta}$ . Let  $s_i(\theta) \equiv \partial \ell_i(\theta) / \partial \theta$ ,  $h_i(\theta) \equiv \partial^2 \ell_i(\theta) / \partial \theta \partial \theta'$ , and the true value be  $\theta^*$ . Then,  $\sqrt{N}(\hat{\theta} - \theta^*) \rightarrow_d \mathcal{N}(0, \mathbf{H}(\theta^*)^{-1} \mathbf{I}(\theta^*) \mathbf{H}(\theta^*)^{-1})$

<sup>32</sup>Loève's  $c_r$ -Inequality: for any  $r > 0$ , we have  $\|\sum_{k=1}^t v_k\|_r \leq \Delta_{r,k} \sum_{k=1}^t \|v_k\|_r$ , where  $\Delta_{r,t} = \max\{1, t^{(1-r)/r}\}$ . See Davidson (1994, p. 139–140) for more details on the mentioned inequalities.

where  $\mathcal{H}(\theta)$  denotes the Hessian matrix  $E(h_i(\theta)|\mathbf{x}_i)$  and  $\mathcal{I}(\theta)$  denotes the information matrix  $E(s_i(\theta)s_i(\theta)'|\mathbf{x}_i)$ . Particularly, we have

$$\begin{aligned}\mathcal{H}(\theta) &= E \left[ \frac{1}{T} \sum_{t=1}^{T-J} \sum_{j=1}^J \frac{\partial^2 \ln f(y_{it}, y_{i,t+j}|x_i; \theta)}{\partial \theta \partial \theta'} \right] \\ \mathcal{I}(\theta) &= E \left[ \left( \frac{1}{T} \sum_{t=1}^{T-J} \sum_{j=1}^J \frac{\partial \ln f(y_{it}, y_{i,t+j}|x_i; \theta)}{\partial \theta} \right) \left( \frac{1}{T} \sum_{t=1}^{T-J} \sum_{j=1}^J \frac{\partial \ln f(y_{it}, y_{i,t+j}|x_i; \theta)}{\partial \theta} \right)' \right].\end{aligned}$$

Note that since composite likelihood does not use the full information as the full likelihood does, we have  $\mathcal{I}(\theta^*) \neq \mathcal{H}(\theta^*)$  in this case. Further details are given in the following subsections. As it is typical in asymptotic normality proofs, we utilize the mean value expansion of the composite likelihood evaluated at the CLE around the true parameter. That is, for a mean value  $\tilde{\theta}$  that lies between  $\hat{\theta}$  and  $\theta^*$ , we have

$$0 = \frac{\partial \mathcal{L}(\hat{\theta})}{\partial \theta} = \frac{\partial \mathcal{L}(\theta^*)}{\partial \theta} + \frac{\partial^2 \mathcal{L}(\tilde{\theta})}{\partial \theta \partial \theta'} (\hat{\theta} - \theta^*) = \frac{1}{N} \sum_{i=1}^N s(\theta^*|\mathbf{y}_i, \mathbf{x}_i) + \left[ \frac{1}{N} \sum_{i=1}^N h(\tilde{\theta}|\mathbf{y}_i, \mathbf{x}_i) \right] (\hat{\theta} - \theta^*),$$

where  $s(\cdot)$  and  $h(\cdot)$  denote the score and Hessian functions. Arranging the terms in the above equation, using the uniform convergence property of the Hessian and the asymptotic normality of the score function, we obtain the desired asymptotic normality result. Further details are provided in the following subsections.

$$\sqrt{N}(\hat{\theta} - \theta^*) = \left[ \frac{1}{N} \sum_{i=1}^N h_i(\tilde{\theta}) \right]^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N s_i(\theta^*) \rightarrow_d \mathcal{N} \left( 0, E[h_i(\theta^*)]^{-1} E[s_i(\theta^*)s_i(\theta^*)'] E[h_i(\theta^*)]^{-1} \right)$$

## The Score

The score of an individual composite likelihood is  $s_i(\theta) = s(\theta|\mathbf{y}_i, \mathbf{x}_i) = \partial \ell_i(\theta)/\partial \theta$ , where

$$\begin{aligned}s(\theta|\mathbf{y}_i, \mathbf{x}_i) &= \frac{1}{T} \sum_{t=1}^{T-J} \sum_{j=1}^J \frac{\partial \ln f(y_{it}, y_{i,t+j}|\mathbf{x}_i; \theta)}{\partial \theta} \\ &= \frac{1}{T} \sum_{t=1}^{T-J} \sum_{j=1}^J \sum_{s_1=1}^S \sum_{s_2=1}^S \frac{\mathbb{1}_{i,t,t+j,s_1,s_2}}{\mathbb{P}_{i,t,t+j,s_1,s_2}} \frac{\partial \mathbb{P}_{i,t,t+j,s_1,s_2}}{\partial \theta}.\end{aligned}$$

Next, we compute the first derivative of the bivariate probabilities. Note that  $\mathbb{P}_{i,t,t+j,s_1,s_2}(\theta) = \Phi_{s_1,s_2} - \Phi_{s_1-1,s_2} - \Phi_{s_1,s_2-1} + \Phi_{s_1-1,s_2-1}$ .

$$\begin{aligned} \frac{\partial \mathbb{P}_{i,t,t+j,s_1,s_2}}{\partial \theta} &= \frac{\partial r_j}{\partial \theta} [\phi_{s_1,s_2} - \phi_{s_1-1,s_2} - \phi_{s_1,s_2-1} + \phi_{s_1-1,s_2-1}] \\ &+ \frac{\partial q_{t,s_1}}{\partial \theta} \phi(q_{t,s_1}) \left[ \Phi \left( \frac{-r_j q_{t,s_1} + q_{t+j,s_2}}{\sqrt{1-r_j^2}} \right) - \Phi \left( \frac{-r_j q_{t,s_1} + q_{t+j,s_2-1}}{\sqrt{1-r_j^2}} \right) \right] \\ &+ \frac{\partial q_{t+j,s_2}}{\partial \theta} \phi(q_{t+j,s_2}) \left[ \Phi \left( \frac{q_{t,s_1} - r_j q_{t+j,s_2}}{\sqrt{1-r_j^2}} \right) - \Phi \left( \frac{q_{t,s_1-1} - r_j q_{t+j,s_2}}{\sqrt{1-r_j^2}} \right) \right] \\ &- \frac{\partial q_{t,s_1-1}}{\partial \theta} \phi(q_{t,s_1-1}) \left[ \Phi \left( \frac{-r_j q_{t,s_1-1} + q_{t+j,s_2}}{\sqrt{1-r_j^2}} \right) - \Phi \left( \frac{-r_j q_{t,s_1-1} + q_{t+j,s_2-1}}{\sqrt{1-r_j^2}} \right) \right] \\ &- \frac{\partial q_{t+j,s_2-1}}{\partial \theta} \phi(q_{t+j,s_2-1}) \left[ \Phi \left( \frac{q_{t,s_1} - r_j q_{t+j,s_2-1}}{\sqrt{1-r_j^2}} \right) - \Phi \left( \frac{q_{t,s_1-1} - r_j q_{t+j,s_2-1}}{\sqrt{1-r_j^2}} \right) \right] \end{aligned}$$

Note that, depending on the values  $s_1$  and  $s_2$  take, some of the terms in the above derivative might drop. For instance,  $q_{t,s_1-1} = -\infty$  for  $s_1 = 1$  or  $q_{t,s_1} = \infty$  for  $s_1 = S$ . These associated terms will drop from the equation; but, to keep the notation simple, we keep the general formula.

Even though the pairwise composite likelihood is a misspecified likelihood as a whole, each piece of it actually belongs to the correctly specified likelihood. Therefore, the score of the composite likelihood is unbiased. In other words, the derivative of the log-pairwise likelihoods at the true value is zero:  $E[\partial \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta^*) / \partial \theta] = 0$ . Hence, we have  $E[s_i(\theta^*) | \mathbf{x}_i] = 0$ , which implies  $E[s_i(\theta^*)] = 0$ . Moreover, conditional on  $\mathbf{x}_i$ ,  $s_i(\theta)$  is independent from  $s_{i'}(\theta)$  for any  $i' \neq i$ . The reason is the independence of  $(\varepsilon_i, \alpha_i)$  from  $(\varepsilon_{i'}, \alpha_{i'})$ , conditional on  $\mathbf{x}_i$ , and that  $s_i(\theta)$  is a measurable function of  $(\varepsilon_i, \alpha_i)$ . Hence, since  $s_i$  is independent and identically distributed with mean 0 and variance  $\mathbf{I}(\theta^*)$ , we can use Lindeberg-Lévy central limit theorem to obtain  $N^{-1/2} \sum_{i=1}^N s_i(\theta^*) \rightarrow_d \mathcal{N}(0, \mathbf{I}(\theta^*))$ , where  $\mathbf{I}(\theta^*) = E[s_i(\theta^*) s_i(\theta^*)']$  and

$$E[s_i(\theta^*) s_i(\theta^*)'] = E \left[ \left( \frac{1}{T} \sum_{t=1}^{T-J} \sum_{j=1}^J \frac{\partial \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta)}{\partial \theta} \right) \left( \frac{1}{T} \sum_{t=1}^{T-J} \sum_{j=1}^J \frac{\partial \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta)}{\partial \theta} \right)' \right].$$

The variance  $\mathcal{I}(\theta^*)$  is finite if  $T^{-1} \sum_{t=1}^{T-J} \sum_{j=1}^J \mathbb{E} \left[ \frac{\partial \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta)}{\partial \theta} \frac{\partial \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta)}{\partial \theta'} \right]$  is finite – due to Cauchy-Schwarz inequality. The finiteness of the expected cross-product is shown in the next subsection where the Hessian is analyzed.

## The Hessian

In this subsection, we compute the Hessian of the composite likelihood function and show that it is uniformly bounded. The Hessian  $h(\theta | \mathbf{y}_i, \mathbf{x}_i) = \partial^2 \ell_i(\theta | \mathbf{y}_i, \mathbf{x}_i) / \partial \theta \partial \theta'$  is found to be

$$\begin{aligned} h(\theta | \mathbf{y}_i, \mathbf{x}_i) &= \\ \frac{1}{T} \sum_{t=1}^{T-J} \sum_{j=1}^J \frac{\partial^2 \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta)}{\partial \theta \partial \theta'} &= \\ \frac{1}{T} \sum_{t=1}^{T-J} \sum_{j=1}^J \sum_{s_1=1}^S \sum_{s_2=1}^S \mathbb{1}_{i,t,t+j,s_1,s_2} &\left[ \mathbb{P}_{i,t,t+j,s_1,s_2}^{-1} \frac{\partial^2 \mathbb{P}_{i,t,t+j,s_1,s_2}}{\partial \theta \partial \theta'} - \mathbb{P}_{i,t,t+j,s_1,s_2}^{-2} \frac{\partial \mathbb{P}_{i,t,t+j,s_1,s_2}}{\partial \theta} \frac{\partial \mathbb{P}_{i,t,t+j,s_1,s_2}}{\partial \theta'} \right]. \end{aligned}$$

The following lemma shows the finiteness of the Hessian.

**Lemma 2.** *The Hessian is uniformly bounded, i.e.  $\mathbb{E}[\sup_{\theta \in \Theta} \left| \frac{\partial^2}{\partial \theta \partial \theta'} \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta) \right|] < \infty$ .*

**Proof of Lemma 2.** *Let's find a uniform upperbound for the generic term of the Hessian.*

*Suppressing the subscripts for  $i$  and  $t$ , and for a positive constant  $c$ , we have*

$$\begin{aligned} &\sup_{\theta \in \Theta} \left| \frac{1}{\mathbb{P}_{s_1 s_2}} \frac{\partial^2 \mathbb{P}_{s_1 s_2}}{\partial \theta \partial \theta'} - \frac{1}{\mathbb{P}_{s_1 s_2}^2} \frac{\partial \mathbb{P}_{s_1 s_2}}{\partial \theta} \frac{\partial \mathbb{P}_{s_1 s_2}}{\partial \theta'} \right| \\ &\leq c(1 + |\mathbf{x}_i|) \left[ \sup_{\theta \in \Theta} \left| \frac{\partial^2 q_{s_1}}{\partial \theta \partial \theta'} \right| + \sup_{\theta \in \Theta} \left| \frac{\partial^2 q_{s_2}}{\partial \theta \partial \theta'} \right| + \sup_{\theta \in \Theta} \left| \frac{\partial^2 q_{s_1-1}}{\partial \theta \partial \theta'} \right| + \sup_{\theta \in \Theta} \left| \frac{\partial^2 q_{s_2-1}}{\partial \theta \partial \theta'} \right| \right] \end{aligned} \quad (20)$$

$$+ c(1 + |\mathbf{x}_i|^2) \left[ \sup_{\theta \in \Theta} \left| \frac{\partial^2 r_j}{\partial \theta \partial \theta'} \right| \right] + c(1 + |\mathbf{x}_i|^4) \left[ \sup_{\theta \in \Theta} \left| \frac{\partial r_j}{\partial \theta} \frac{\partial r_j}{\partial \theta'} \right| \right] \quad (21)$$

$$+ c(1 + |\mathbf{x}_i|^2) \left[ \sup_{\theta \in \Theta} \left| \frac{\partial q_{s_1}}{\partial \theta} \frac{\partial q_{s_1}}{\partial \theta'} \right| + \dots + \sup_{\theta \in \Theta} \left| \frac{\partial q_{s_2-1}}{\partial \theta} \frac{\partial q_{s_2-1}}{\partial \theta'} \right| \right] \quad (22)$$

$$+ c(1 + |\mathbf{x}_i|^3) \left[ \sup_{\theta \in \Theta} \left| \frac{\partial q_{s_1}}{\partial \theta} \frac{\partial r_j}{\partial \theta'} \right| + \sup_{\theta \in \Theta} \left| \frac{\partial q_{s_2}}{\partial \theta} \frac{\partial r_j}{\partial \theta'} \right| + \sup_{\theta \in \Theta} \left| \frac{\partial q_{s_1-1}}{\partial \theta} \frac{\partial r_j}{\partial \theta'} \right| + \sup_{\theta \in \Theta} \left| \frac{\partial q_{s_2-1}}{\partial \theta} \frac{\partial r_j}{\partial \theta'} \right| \right] \quad (23)$$

*The second order derivative terms in (20) are bounded by a linear function of  $|\mathbf{x}_i|$ ; thus, (20) is bounded by  $c_1(1 + |\mathbf{x}_i|)^2$ , for some constant  $c_1 > 0$ . The derivatives of  $r_j$  are bounded; thus, (21) is bounded by  $c_2(1 + |\mathbf{x}_i|^4)$ , for some constant  $c_2 > 0$ . The cross derivative terms in (22)*

are bounded by a quadratic function of  $|\mathbf{x}_i|$ ; thus, (22) is bounded by  $c_3(1 + |\mathbf{x}_i|)^4$ , for some constant  $c_3 > 0$ . Finally, the cross derivative terms with  $r_j$  in (23) are bounded linearly by  $|\mathbf{x}_i|$ ; thus, (23) is bounded by  $c_4(1 + |\mathbf{x}_i|)^4$ , for some constant  $c_4 > 0$ . The details of these derivations can be found in sections B.1 and B.2, and in the supplementary material. As a result, the Hessian term is uniformly bounded by the fourth moment of  $\mathbf{x}_i$ . For some  $C > 0$ ,

$$\mathbb{E} \left[ \sup_{\theta \in \Theta} \left| \frac{\partial^2}{\partial \theta \partial \theta'} \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta) \right| \right] \leq \sum_{s_1}^S \sum_{s_2}^S \left\| \sup_{\theta \in \Theta} \left| \frac{\partial^2 \mathbb{P}_{s_1 s_2}}{\partial \theta \partial \theta'} - \frac{\partial \mathbb{P}_{s_1 s_2}}{\partial \theta} \frac{\partial \mathbb{P}_{s_1 s_2}}{\partial \theta'} \right| \right\|_1 \leq C \|(1 + |\mathbf{x}_i|)^4\|_1.$$

Hence, the Hessian is uniformly bounded by Assumption 3, i.e.  $\|\mathbf{x}_i\|_4 < \infty$ .

Lemma 2 yields the uniform convergence of  $N^{-1} \sum_{i=1}^N h(\tilde{\theta} | \mathbf{y}_i, \mathbf{x}_i)$  for any consistent estimator  $\tilde{\theta}$  such that  $1/N \sum_{i=1}^N h(\tilde{\theta} | \mathbf{y}_i, \mathbf{x}_i) \rightarrow_p \mathcal{H}(\theta^*)$ , where the Hessian matrix is  $\mathcal{H}(\theta^*) = \mathbb{E}[h_i(\theta^*) | \mathbf{x}_i] = 1/T \sum_{t=1}^{T-J} \sum_{j=1}^J \mathbb{E} \left[ \frac{\partial^2 \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta^*)}{\partial \theta \partial \theta'} \right]$ . We need  $\mathcal{H}(\theta^*)$  to be nonsingular. It is usually hard to prove negative definiteness of the Hessian matrix in non-linear models. However, with composite likelihood we can utilize its nice features that it borrows from the full likelihood. Particularly, the Bartlett equality holds for each piece of the composite likelihood even though it does not hold, in general, for the composite likelihood itself. Mathematically,  $\mathbb{E} \left[ \frac{\partial \ell_i(\theta^*)}{\partial \theta} \frac{\partial \ell_i(\theta^*)}{\partial \theta'} \right] \neq -\mathbb{E} \left[ \frac{\partial^2 \ell_i(\theta^*)}{\partial \theta \partial \theta'} \right]$ , yet,

$$\mathbb{E} \left[ \frac{\partial \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta^*)}{\partial \theta} \frac{\partial \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta^*)}{\partial \theta'} \right] = -\mathbb{E} \left[ \frac{\partial^2 \ln f(y_{it}, y_{i,t+j} | \mathbf{x}_i; \theta^*)}{\partial \theta \partial \theta'} \right] < 0.$$

Hence,  $\mathcal{H}(\theta^*)$  is invertible. Therefore, we can conclude that  $[1/N \sum_{i=1}^N h(\tilde{\theta} | \mathbf{y}_i, \mathbf{x}_i)]^{-1} \rightarrow_p \mathcal{H}(\theta^*)^{-1}$ , for any consistent estimator  $\tilde{\theta}$ .

Consequently, by Slutsky's Theorem, we conclude that

$$\sqrt{N}(\hat{\theta} - \theta^*) = \left[ \frac{1}{N} \sum_{i=1}^N h_i(\tilde{\theta}) \right]^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N s_i(\theta^*) \rightarrow_d \mathcal{N}(0, \mathcal{H}(\theta^*)^{-1} \mathcal{I}(\theta^*) \mathcal{H}(\theta^*)^{-1}).$$

## B Mathematical Details

This section analyzes mathematical properties of functions of normal density and normal cumulative distribution, especially the ones that are needed throughout the analysis in this paper. First, the derivative of the bivariate standard normal distribution is

$$\begin{aligned}
\frac{\partial \Phi_2(m_t(\theta), m_{t+j}(\theta) | r(\theta))}{\partial \theta} &= m'_t(\theta) \phi(m_t(\theta)) \Phi\left(\frac{-r(\theta)m_t(\theta) + m_{t+j}(\theta)}{\sqrt{1-r(\theta)^2}}\right) \\
&+ m'_{t+j}(\theta) \phi(m_{t+j}(\theta)) \Phi\left(\frac{m_t(\theta) - r(\theta)m_{t+j}(\theta)}{\sqrt{1-r(\theta)^2}}\right) \\
&+ r'(\theta) \phi_2(m_t(\theta), m_{t+j}(\theta) | r). \tag{24}
\end{aligned}$$

Further details of this derivation are given in the supplementary material.

### B.1 Derivatives of Bivariate Probabilities

The subscripts  $(s_1, s_1 - 1)$  and  $(s_2, s_2 - 1)$  are always associated with  $t$  and  $t + j$ , respectively. Hence, whenever it is clear, we drop the time subscripts  $t$  and  $t + j$ . First, let  $\phi_{s_1, s_2} \equiv \phi_2(q_{t, s_1}, q_{t+j, s_2} | r_j)$ . Next, let  $\xi_{(r)s_1, s_2} \equiv (1 - r_j^2)^{-1/2}(-r_j q_{t, s_1} + q_{t+j, s_2})$ , and  $\xi_{s_1, (r)s_2} \equiv (1 - r_j^2)^{-1/2}(q_{t, s_1} - r_j q_{t+j, s_2})$ . The subscripts  $(r)s_1$  and  $(r)s_2$  indicate whether the term with  $t$  or  $t + j$  is multiplied by  $-r_j$ . The  $\xi_{(r)s_1, s_2}$  variable appears if the derivatives of the bivariate probabilities are taken with respect to  $q_{s_1}$ . Similarly,  $\xi_{s_1, (r)s_2}$  variable appears if the derivatives are taken with respect to  $q_{s_2}$ . The first derivative of the bivariate probability with the newly introduced notation is as follows.

$$\begin{aligned}
\frac{\partial \mathbb{P}_{s_1 s_2}}{\partial \theta} &= \\
&\frac{\partial q_{s_1}}{\partial \theta} \phi_{s_1} [\Phi(\zeta_{(r)s_1, s_2}) - \Phi(\zeta_{(r)s_1, s_2-1})] - \frac{\partial q_{s_1-1}}{\partial \theta} \phi_{s_1-1} [\Phi(\zeta_{(r)s_1-1, s_2}) - \Phi(\zeta_{(r)s_1-1, s_2-1})] + \\
&\frac{\partial q_{s_2}}{\partial \theta} \phi_{s_2} [\Phi(\zeta_{s_1, (r)s_2}) - \Phi(\zeta_{s_1-1, (r)s_2})] - \frac{\partial q_{s_2-1}}{\partial \theta} \phi_{s_2-1} [\Phi(\zeta_{s_1, (r)s_2-1}) - \Phi(\zeta_{s_1-1, (r)s_2-1})] + \\
&\frac{\partial r_j}{\partial \theta} [\phi_{s_1, s_2} + \phi_{s_1-1, s_2-1} - \phi_{s_1-1, s_2} - \phi_{s_1, s_2-1}]
\end{aligned}$$



The cross-product of the first derivative is as follows.

$$\begin{aligned}
& \frac{\partial \mathbb{P}_{s_1 s_2}}{\partial \theta} \frac{\partial \mathbb{P}_{s_1 s_2}}{\partial \theta'} = \frac{\partial r_j}{\partial \theta} \frac{\partial r_j}{\partial \theta'} [\phi_{s_1, s_2} + \phi_{s_1-1, s_2-1} - \phi_{s_1-1, s_2} - \phi_{s_1, s_2-1}]^2 + \\
& \frac{\partial q_{s_1}}{\partial \theta} \frac{\partial q_{s_1}}{\partial \theta'} \phi_{s_1}^2 [\Phi(\zeta_{(r) s_1, s_2}) - \Phi(\zeta_{(r) s_1, s_2-1})]^2 + \\
& \frac{\partial q_{s_1-1}}{\partial \theta} \frac{\partial q_{s_1-1}}{\partial \theta'} \phi_{s_1-1}^2 [\Phi(\zeta_{(r) s_1-1, s_2}) - \Phi(\zeta_{(r) s_1-1, s_2-1})]^2 + \\
& \frac{\partial q_{s_2}}{\partial \theta} \frac{\partial q_{s_2}}{\partial \theta'} \phi_{s_2}^2 [\Phi(\zeta_{s_1, (r) s_2}) - \Phi(\zeta_{s_1-1, (r) s_2})]^2 + \\
& \frac{\partial q_{s_2-1}}{\partial \theta} \frac{\partial q_{s_2-1}}{\partial \theta'} \phi_{s_2-1}^2 [\Phi(\zeta_{s_1, (r) s_2-1}) - \Phi(\zeta_{s_1-1, (r) s_2-1})]^2 + \\
& \left[ \frac{\partial q_{s_1}}{\partial \theta} \frac{\partial q_{s_2}}{\partial \theta'} + \frac{\partial q_{s_2}}{\partial \theta} \frac{\partial q_{s_1}}{\partial \theta'} \right] \phi_{s_1} \phi_{s_2} [\Phi(\zeta_{(r) s_1, s_2}) - \Phi(\zeta_{(r) s_1, s_2-1})] [\Phi(\zeta_{s_1, (r) s_2}) - \Phi(\zeta_{s_1-1, (r) s_2})] + \\
& \left[ \frac{\partial q_{s_1-1}}{\partial \theta} \frac{\partial q_{s_2-1}}{\partial \theta'} + \frac{\partial q_{s_2-1}}{\partial \theta} \frac{\partial q_{s_1-1}}{\partial \theta'} \right] \\
& \quad \times \phi_{s_1-1} \phi_{s_2-1} [\Phi(\zeta_{(r) s_1-1, s_2}) - \Phi(\zeta_{(r) s_1-1, s_2-1})] [\Phi(\zeta_{s_1, (r) s_2-1}) - \Phi(\zeta_{s_1-1, (r) s_2-1})] - \\
& \left[ \frac{\partial q_{s_1-1}}{\partial \theta} \frac{\partial q_{s_2}}{\partial \theta'} + \frac{\partial q_{s_2}}{\partial \theta} \frac{\partial q_{s_1-1}}{\partial \theta'} \right] \\
& \quad \times \phi_{s_1-1} \phi_{s_2} [\Phi(\zeta_{(r) s_1-1, s_2}) - \Phi(\zeta_{(r) s_1-1, s_2-1})] [\Phi(\zeta_{s_1, (r) s_2}) - \Phi(\zeta_{s_1-1, (r) s_2})] - \\
& \left[ \frac{\partial q_{s_1}}{\partial \theta} \frac{\partial q_{s_2-1}}{\partial \theta'} + \frac{\partial q_{s_2-1}}{\partial \theta} \frac{\partial q_{s_1}}{\partial \theta'} \right] \\
& \quad \times \phi_{s_1} \phi_{s_2-1} [\Phi(\zeta_{(r) s_1, s_2}) - \Phi(\zeta_{(r) s_1, s_2-1})] [\Phi(\zeta_{s_1, (r) s_2-1}) - \Phi(\zeta_{s_1-1, (r) s_2-1})] + \\
& \left[ \frac{\partial q_{s_1}}{\partial \theta} \frac{\partial r_j}{\partial \theta'} + \frac{\partial r_j}{\partial \theta} \frac{\partial q_{s_1}}{\partial \theta'} \right] \\
& \quad \times \phi_{s_1} [\Phi(\zeta_{(r) s_1, s_2}) - \Phi(\zeta_{(r) s_1, s_2-1})] [\phi_{s_1, s_2} + \phi_{s_1-1, s_2-1} - \phi_{s_1-1, s_2} - \phi_{s_1, s_2-1}] - \\
& \left[ \frac{\partial q_{s_1-1}}{\partial \theta} \frac{\partial r_j}{\partial \theta'} + \frac{\partial r_j}{\partial \theta} \frac{\partial q_{s_1-1}}{\partial \theta'} \right] \\
& \quad \times \phi_{s_1-1} [\Phi(\zeta_{(r) s_1-1, s_2}) - \Phi(\zeta_{(r) s_1-1, s_2-1})] [\phi_{s_1, s_2} + \phi_{s_1-1, s_2-1} - \phi_{s_1-1, s_2} - \phi_{s_1, s_2-1}] + \\
& \left[ \frac{\partial q_{s_2}}{\partial \theta} \frac{\partial r_j}{\partial \theta'} + \frac{\partial r_j}{\partial \theta} \frac{\partial q_{s_2}}{\partial \theta'} \right] \\
& \quad \times \phi_{s_2} [\Phi(\zeta_{s_1, (r) s_2}) - \Phi(\zeta_{s_1-1, (r) s_2})] [\phi_{s_1, s_2} + \phi_{s_1-1, s_2-1} - \phi_{s_1-1, s_2} - \phi_{s_1, s_2-1}] - \\
& \left[ \frac{\partial q_{s_2-1}}{\partial \theta} \frac{\partial r_j}{\partial \theta'} + \frac{\partial r_j}{\partial \theta} \frac{\partial q_{s_2-1}}{\partial \theta'} \right] \\
& \quad \times \phi_{s_2-1} [\Phi(\zeta_{s_1, (r) s_2-1}) - \Phi(\zeta_{s_1-1, (r) s_2-1})] [\phi_{s_1, s_2} + \phi_{s_1-1, s_2-1} - \phi_{s_1-1, s_2} - \phi_{s_1, s_2-1}]
\end{aligned}$$

The second derivative of the bivariate probability is as follows.

$$\begin{aligned}
& \frac{\partial^2 \mathbb{P}_{s_1 s_2}}{\partial \theta \partial \theta'} \\
&= \frac{\partial^2 q_{s_1}}{\partial \theta \partial \theta'} \phi_{s_1} [\Phi(\zeta_{(r)s_1, s_2}) - \Phi(\zeta_{(r)s_1, s_2-1})] - \frac{\partial^2 q_{s_1-1}}{\partial \theta \partial \theta'} \phi_{s_1-1} [\Phi(\zeta_{(r)s_1-1, s_2}) - \Phi(\zeta_{(r)s_1-1, s_2-1})] \\
&- \frac{\partial^2 q_{s_2-1}}{\partial \theta \partial \theta'} \phi_{s_2-1} [\Phi(\zeta_{s_1, (r)s_2-1}) - \Phi(\zeta_{s_1-1, (r)s_2-1})] + \frac{\partial^2 q_{s_2}}{\partial \theta \partial \theta'} \phi_{s_2} [\Phi(\zeta_{s_1, (r)s_2}) - \Phi(\zeta_{s_1-1, (r)s_2})] \\
&+ \frac{\partial^2 r_j}{\partial \theta \partial \theta'} [\phi_{s_1, s_2} - \phi_{s_1-1, s_2} - \phi_{s_1, s_2-1} + \phi_{s_1-1, s_2-1}] \\
&- \frac{\partial q_{s_1}}{\partial \theta} \frac{\partial q_{s_1}}{\partial \theta'} \{q_{s_1} \phi_{s_1} [\Phi(\zeta_{(r)s_1, s_2}) - \Phi(\zeta_{(r)s_1, s_2-1})] + r_j [\phi_{s_1, s_2} - \phi_{s_1, s_2-1}]\} \\
&+ \frac{\partial q_{s_1-1}}{\partial \theta} \frac{\partial q_{s_1-1}}{\partial \theta'} \{q_{s_1-1} \phi_{s_1-1} [\Phi(\zeta_{(r)s_1-1, s_2}) - \Phi(\zeta_{(r)s_1-1, s_2-1})] + r_j [\phi_{s_1-1, s_2} - \phi_{s_1-1, s_2-1}]\} \\
&+ \frac{\partial q_{s_2-1}}{\partial \theta} \frac{\partial q_{s_2-1}}{\partial \theta'} \{q_{s_2-1} \phi_{s_2-1} [\Phi(\zeta_{s_1, (r)s_2-1}) - \Phi(\zeta_{s_1-1, (r)s_2-1})] + r_j [\phi_{s_1, s_2-1} - \phi_{s_1-1, s_2-1}]\} \\
&- \frac{\partial q_{s_2}}{\partial \theta} \frac{\partial q_{s_2}}{\partial \theta'} \{q_{s_2} \phi_{s_2} [\Phi(\zeta_{s_1, (r)s_2}) - \Phi(\zeta_{s_1-1, (r)s_2})] + r_j [\phi_{s_1, s_2} - \phi_{s_1-1, s_2}]\} \\
&+ \frac{\partial r_j}{\partial \theta} \frac{\partial r_j}{\partial \theta'} \left\{ \phi_{s_1, s_2} \left[ \frac{r_j}{1-r_j^2} + \zeta_{(r)s_1, s_2} \zeta_{s_1, (r)s_2} \right] + \phi_{s_1-1, s_2-1} \left[ \frac{r_j}{1-r_j^2} + \zeta_{(r)s_1-1, s_2-1} \zeta_{s_1-1, (r)s_2-1} \right] \right. \\
&\quad \left. - \phi_{s_1, s_2-1} \left[ \frac{r_j}{1-r_j^2} + \zeta_{(r)s_1, s_2-1} \zeta_{s_1, (r)s_2-1} \right] - \phi_{s_1-1, s_2} \left[ \frac{r_j}{1-r_j^2} + \zeta_{(r)s_1-1, s_2} \zeta_{s_1-1, (r)s_2} \right] \right\} \\
&+ \left( \frac{\partial q_{s_1}}{\partial \theta} \frac{\partial q_{s_2}}{\partial \theta'} + \frac{\partial q_{s_2}}{\partial \theta} \frac{\partial q_{s_1}}{\partial \theta'} \right) \phi_{s_1, s_2} + \left( \frac{\partial q_{s_1-1}}{\partial \theta} \frac{\partial q_{s_2-1}}{\partial \theta'} + \frac{\partial q_{s_2-1}}{\partial \theta} \frac{\partial q_{s_1-1}}{\partial \theta'} \right) \phi_{s_1-1, s_2-1} \\
&- \left( \frac{\partial q_{s_1}}{\partial \theta} \frac{\partial q_{s_2-1}}{\partial \theta'} + \frac{\partial q_{s_2-1}}{\partial \theta} \frac{\partial q_{s_1}}{\partial \theta'} \right) \phi_{s_1, s_2-1} - \left( \frac{\partial q_{s_1-1}}{\partial \theta} \frac{\partial q_{s_2}}{\partial \theta'} + \frac{\partial q_{s_2}}{\partial \theta} \frac{\partial q_{s_1-1}}{\partial \theta'} \right) \phi_{s_1-1, s_2} \\
&- \left( \frac{\partial q_{s_1}}{\partial \theta} \frac{\partial r_j}{\partial \theta'} + \frac{\partial r_j}{\partial \theta} \frac{\partial q_{s_1}}{\partial \theta'} \right) \frac{1}{\sqrt{1-r_j^2}} [\zeta_{s_1, (r)s_2} \phi_{s_1, s_2} - \zeta_{s_1, (r)s_2-1} \phi_{s_1, s_2-1}] \\
&+ \left( \frac{\partial q_{s_1-1}}{\partial \theta} \frac{\partial r_j}{\partial \theta'} + \frac{\partial r_j}{\partial \theta} \frac{\partial q_{s_1-1}}{\partial \theta'} \right) \frac{1}{\sqrt{1-r_j^2}} [\zeta_{s_1-1, (r)s_2} \phi_{s_1-1, s_2} - \zeta_{s_1-1, (r)s_2-1} \phi_{s_1-1, s_2-1}] \\
&+ \left( \frac{\partial q_{s_2-1}}{\partial \theta} \frac{\partial r_j}{\partial \theta'} + \frac{\partial r_j}{\partial \theta} \frac{\partial q_{s_2-1}}{\partial \theta'} \right) \frac{1}{\sqrt{1-r_j^2}} [\zeta_{(r)s_1, s_2-1} \phi_{s_1, s_2-1} - \zeta_{(r)s_1-1, s_2-1} \phi_{s_1-1, s_2-1}] \\
&- \left( \frac{\partial q_{s_2}}{\partial \theta} \frac{\partial r_j}{\partial \theta'} + \frac{\partial r_j}{\partial \theta} \frac{\partial q_{s_2}}{\partial \theta'} \right) \frac{1}{\sqrt{1-r_j^2}} [\zeta_{(r)s_1, s_2} \phi_{s_1, s_2} - \zeta_{(r)s_1-1, s_2} \phi_{s_1-1, s_2}]
\end{aligned}$$

The Hessian term of the pairwise composite likelihood is as follows.

$$\begin{aligned}
& \frac{1}{\mathbb{P}_{s_1 s_2}} \frac{\partial^2 \mathbb{P}_{s_1 s_2}}{\partial \theta \partial \theta'} - \frac{1}{\mathbb{P}_{s_1 s_2}^2} \frac{\partial \mathbb{P}_{s_1 s_2}}{\partial \theta} \frac{\partial \mathbb{P}_{s_1 s_2}}{\partial \theta'} = \frac{\partial^2 r_j}{\partial \theta \partial \theta'} [\phi_{s_1, s_2} - \phi_{s_1-1, s_2} - \phi_{s_1, s_2-1} + \phi_{s_1-1, s_2-1}] \mathbb{P}_{s_1 s_2}^{-1} + \\
& \frac{\partial^2 q_{s_1}}{\partial \theta \partial \theta'} \phi_{s_1} [\Phi(\zeta_{(r) s_1, s_2}) - \Phi(\zeta_{(r) s_1, s_2-1})] \mathbb{P}_{s_1 s_2}^{-1} - \frac{\partial^2 q_{s_1-1}}{\partial \theta \partial \theta'} \phi_{s_1-1} [\Phi(\zeta_{(r) s_1-1, s_2}) - \Phi(\zeta_{(r) s_1-1, s_2-1})] \mathbb{P}_{s_1 s_2}^{-1} + \\
& \frac{\partial^2 q_{s_2}}{\partial \theta \partial \theta'} \phi_{s_2} [\Phi(\zeta_{s_1, (r) s_2}) - \Phi(\zeta_{s_1-1, (r) s_2})] \mathbb{P}_{s_1 s_2}^{-1} - \frac{\partial^2 q_{s_2-1}}{\partial \theta \partial \theta'} \phi_{s_2-1} [\Phi(\zeta_{s_1, (r) s_2-1}) - \Phi(\zeta_{s_1-1, (r) s_2-1})] \mathbb{P}_{s_1 s_2}^{-1} - \\
& \frac{\partial q_{s_1}}{\partial \theta} \frac{\partial q_{s_1}}{\partial \theta'} \left\{ \frac{q_{s_1} \phi_{s_1} [\Phi(\zeta_{(r) s_1, s_2}) - \Phi(\zeta_{(r) s_1, s_2-1})]}{\mathbb{P}_{s_1 s_2}} + \frac{r_j [\phi_{s_1, s_2} - \phi_{s_1, s_2-1}]}{\mathbb{P}_{s_1 s_2}} + \frac{\phi_{s_1}^2 [\Phi(\zeta_{(r) s_1, s_2}) - \Phi(\zeta_{(r) s_1, s_2-1})]^2}{\mathbb{P}_{s_1 s_2}^2} \right\} - \\
& \frac{\partial q_{s_2}}{\partial \theta} \frac{\partial q_{s_2}}{\partial \theta'} \left\{ \frac{q_{s_2} \phi_{s_2} [\Phi(\zeta_{s_1, (r) s_2}) - \Phi(\zeta_{s_1-1, (r) s_2})]}{\mathbb{P}_{s_1 s_2}} + \frac{r_j [\phi_{s_1, s_2} - \phi_{s_1-1, s_2}]}{\mathbb{P}_{s_1 s_2}} + \frac{\phi_{s_2}^2 [\Phi(\zeta_{s_1, (r) s_2}) - \Phi(\zeta_{s_1-1, (r) s_2})]^2}{\mathbb{P}_{s_1 s_2}^2} \right\} + \\
& \frac{\partial q_{s_1-1}}{\partial \theta} \frac{\partial q_{s_1-1}}{\partial \theta'} \left\{ q_{s_1-1} \phi_{s_1-1} [\Phi(\zeta_{(r) s_1-1, s_2}) - \Phi(\zeta_{(r) s_1-1, s_2-1})] \mathbb{P}_{s_1 s_2}^{-1} + r_j [\phi_{s_1-1, s_2} - \phi_{s_1-1, s_2-1}] \mathbb{P}_{s_1 s_2}^{-1} \right. \\
& \left. - \phi_{s_1-1}^2 [\Phi(\zeta_{(r) s_1-1, s_2}) - \Phi(\zeta_{(r) s_1-1, s_2-1})]^2 \mathbb{P}_{s_1 s_2}^{-2} \right\} + \frac{\partial q_{s_2-1}}{\partial \theta} \frac{\partial q_{s_2-1}}{\partial \theta'} \left\{ q_{s_2-1} \phi_{s_2-1} [\Phi(\zeta_{s_1, (r) s_2-1}) \right. \\
& \left. - \Phi(\zeta_{s_1-1, (r) s_2-1})] \mathbb{P}_{s_1 s_2}^{-1} + r_j [\phi_{s_1, s_2-1} - \phi_{s_1-1, s_2-1}] \mathbb{P}_{s_1 s_2}^{-1} - \phi_{s_2-1}^2 [\Phi(\zeta_{s_1, (r) s_2-1}) - \Phi(\zeta_{s_1-1, (r) s_2-1})]^2 \mathbb{P}_{s_1 s_2}^{-2} \right\} + \\
& \frac{\partial r_j}{\partial \theta} \frac{\partial r_j}{\partial \theta'} \mathbb{P}_{s_1 s_2}^{-1} \left\{ - [\phi_{s_1, s_2} - \phi_{s_1-1, s_2} - \phi_{s_1, s_2-1} + \phi_{s_1-1, s_2-1}]^2 \mathbb{P}_{s_1 s_2}^{-1} + \phi_{s_1-1, s_2-1} \left[ \frac{r_j}{1-r_j^2} + \zeta_{(r) s_1-1, s_2-1} \zeta_{s_1-1, (r) s_2-1} \right] \right. \\
& \left. + \phi_{s_1, s_2} \left[ \frac{r_j}{1-r_j^2} + \zeta_{(r) s_1, s_2} \zeta_{s_1, (r) s_2} \right] - \phi_{s_1, s_2-1} \left[ \frac{r_j}{1-r_j^2} + \zeta_{(r) s_1, s_2-1} \zeta_{s_1, (r) s_2-1} \right] - \phi_{s_1-1, s_2} \left[ \frac{r_j}{1-r_j^2} + \zeta_{(r) s_1-1, s_2} \zeta_{s_1-1, (r) s_2} \right] \right\} + \\
& \left[ \frac{\partial q_{s_1}}{\partial \theta} \frac{\partial q_{s_2}}{\partial \theta'} + \frac{\partial q_{s_2}}{\partial \theta} \frac{\partial q_{s_1}}{\partial \theta'} \right] \left\{ \frac{\phi_{s_1, s_2}}{\mathbb{P}_{s_1 s_2}} - \frac{\phi_{s_1} \phi_{s_2} [\Phi(\zeta_{(r) s_1, s_2}) - \Phi(\zeta_{(r) s_1, s_2-1})] [\Phi(\zeta_{s_1, (r) s_2}) - \Phi(\zeta_{s_1-1, (r) s_2})]}{\mathbb{P}_{s_1 s_2}^2} \right\} + \\
& \left[ \frac{\partial q_{s_1-1}}{\partial \theta} \frac{\partial q_{s_2-1}}{\partial \theta'} + \frac{\partial q_{s_2-1}}{\partial \theta} \frac{\partial q_{s_1-1}}{\partial \theta'} \right] \left\{ \phi_{s_1-1, s_2-1} \mathbb{P}_{s_1 s_2}^{-1} - \phi_{s_1-1} \phi_{s_2-1} [\Phi(\zeta_{(r) s_1-1, s_2}) - \Phi(\zeta_{(r) s_1-1, s_2-1})] \right. \\
& \times [\Phi(\zeta_{s_1, (r) s_2-1}) - \Phi(\zeta_{s_1-1, (r) s_2-1})] \mathbb{P}_{s_1 s_2}^{-2} \left. \right\} - \left[ \frac{\partial q_{s_1}}{\partial \theta} \frac{\partial q_{s_2-1}}{\partial \theta'} + \frac{\partial q_{s_2-1}}{\partial \theta} \frac{\partial q_{s_1}}{\partial \theta'} \right] \left\{ \phi_{s_1, s_2-1} \mathbb{P}_{s_1 s_2}^{-1} - \phi_{s_1} \phi_{s_2-1} \right. \\
& \times [\Phi(\zeta_{(r) s_1, s_2}) - \Phi(\zeta_{(r) s_1, s_2-1})] [\Phi(\zeta_{s_1, (r) s_2-1}) - \Phi(\zeta_{s_1-1, (r) s_2-1})] \mathbb{P}_{s_1 s_2}^{-2} \left. \right\} - \left[ \frac{\partial q_{s_1-1}}{\partial \theta} \frac{\partial q_{s_2}}{\partial \theta'} + \frac{\partial q_{s_2}}{\partial \theta} \frac{\partial q_{s_1-1}}{\partial \theta'} \right] \\
& \left\{ \phi_{s_1-1, s_2} \mathbb{P}_{s_1 s_2}^{-1} - \phi_{s_1-1} \phi_{s_2} [\Phi(\zeta_{(r) s_1-1, s_2}) - \Phi(\zeta_{(r) s_1-1, s_2-1})] [\Phi(\zeta_{s_1, (r) s_2}) - \Phi(\zeta_{s_1-1, (r) s_2})] \mathbb{P}_{s_1 s_2}^{-2} \right\} - \\
& \left[ \frac{\partial q_{s_1}}{\partial \theta} \frac{\partial r_j}{\partial \theta'} + \frac{\partial r_j}{\partial \theta} \frac{\partial q_{s_1}}{\partial \theta'} \right] \left\{ [\zeta_{s_1, (r) s_2} \phi_{s_1, s_2} - \zeta_{s_1, (r) s_2-1} \phi_{s_1, s_2-1}] (1-r_j^2)^{-1/2} \mathbb{P}_{s_1 s_2}^{-1} \right. \\
& \left. + \phi_{s_1} [\Phi(\zeta_{(r) s_1, s_2}) - \Phi(\zeta_{(r) s_1, s_2-1})] [\phi_{s_1, s_2} - \phi_{s_1-1, s_2} - \phi_{s_1, s_2-1} + \phi_{s_1-1, s_2-1}] \mathbb{P}_{s_1 s_2}^{-2} \right\} - \\
& \left[ \frac{\partial q_{s_2}}{\partial \theta} \frac{\partial r_j}{\partial \theta'} + \frac{\partial r_j}{\partial \theta} \frac{\partial q_{s_2}}{\partial \theta'} \right] \left\{ [\zeta_{(r) s_1, s_2} \phi_{s_1, s_2} - \zeta_{(r) s_1-1, s_2} \phi_{s_1-1, s_2}] (1-r_j^2)^{-1/2} \mathbb{P}_{s_1 s_2}^{-1} \right. \\
& \left. + \phi_{s_2} [\Phi(\zeta_{s_1, (r) s_2}) - \Phi(\zeta_{s_1-1, (r) s_2})] [\phi_{s_1, s_2} - \phi_{s_1-1, s_2} - \phi_{s_1, s_2-1} + \phi_{s_1-1, s_2-1}] \mathbb{P}_{s_1 s_2}^{-2} \right\} + \\
& \left[ \frac{\partial q_{s_1-1}}{\partial \theta} \frac{\partial r_j}{\partial \theta'} + \frac{\partial r_j}{\partial \theta} \frac{\partial q_{s_1-1}}{\partial \theta'} \right] \left\{ [\zeta_{s_1-1, (r) s_2} \phi_{s_1-1, s_2} - \zeta_{s_1-1, (r) s_2-1} \phi_{s_1-1, s_2-1}] (1-r_j^2)^{-1/2} \mathbb{P}_{s_1 s_2}^{-1} \right. \\
& \left. + \phi_{s_1-1} [\Phi(\zeta_{(r) s_1-1, s_2}) - \Phi(\zeta_{(r) s_1-1, s_2-1})] [\phi_{s_1, s_2} - \phi_{s_1-1, s_2} - \phi_{s_1, s_2-1} + \phi_{s_1-1, s_2-1}] \mathbb{P}_{s_1 s_2}^{-2} \right\} + \\
& \left[ \frac{\partial q_{s_2-1}}{\partial \theta} \frac{\partial r_j}{\partial \theta'} + \frac{\partial r_j}{\partial \theta} \frac{\partial q_{s_2-1}}{\partial \theta'} \right] \left\{ [\zeta_{(r) s_1, s_2-1} \phi_{s_1, s_2-1} - \zeta_{(r) s_1-1, s_2-1} \phi_{s_1-1, s_2-1}] (1-r_j^2)^{-1/2} \mathbb{P}_{s_1 s_2}^{-1} \right. \\
& \left. + \phi_{s_2-1} [\Phi(\zeta_{s_1, (r) s_2-1}) - \Phi(\zeta_{s_1-1, (r) s_2-1})] [\phi_{s_1, s_2} - \phi_{s_1-1, s_2} - \phi_{s_1, s_2-1} + \phi_{s_1-1, s_2-1}] \mathbb{P}_{s_1 s_2}^{-2} \right\}
\end{aligned}$$

## B.2 Bounds on functions of $q_{its}(\theta)$

In this subsection, we analyze the upperbounds for functions of  $q_{its}(\theta)$ , in particular the bounds for  $q_{its}$ ,  $q_{its}^2$ , and  $q'_{its}$ . First, remember that  $q_{its}(\theta) = \tau_s - (\mu + \gamma' \bar{\mathbf{x}}_i + \sum_{k=1}^t \rho^{t-k} \beta' x_{ik})$ . A uniform bound for  $q_{its}(\theta)$  can be calculated as

$$\begin{aligned} & |q_{its}(\theta)| \\ &= |\tau_s - \mu - \gamma' \bar{\mathbf{x}}_i - \sum_{k=1}^t \rho^{t-k} \beta' x_{ik}| \leq \tau_s + |\mu| + |\gamma| |\bar{\mathbf{x}}_i| + \sum_{k=1}^t |\rho|^{t-k} |\beta| |x_{ik}| \\ &\leq \tau_s + |\mu| + \sum_{t=1}^T \left( |\rho|^{T-t} |\beta| + \frac{|\gamma|}{T} \right) |x_{it}| \leq \tau_s + |\mu| + \max_{1 \leq t \leq T} |x_{it}| \sum_{t=1}^T \left( |\rho|^{T-k} |\beta| + \frac{|\gamma|}{T} \right). \end{aligned}$$

Hence, taking supremum over the compact parameter space  $\Theta$  yields

$$\begin{aligned} \sup_{\theta \in \Theta} |q_{its}(\theta)| &\leq \sup_{\theta \in \Theta} \left\{ \tau_s + |\mu| + \max_{1 \leq t \leq T} |x_{it}| \sum_{t=1}^T \left( |\rho|^{T-k} |\beta| + \frac{|\gamma|}{T} \right) \right\} \\ &\leq c_\tau + \bar{\mu} + \left( c_\gamma + c_\beta \frac{1 - \bar{\rho}^T}{1 - \bar{\rho}} \right) \max_{1 \leq t \leq T} |x_{it}| \\ &\leq c(1 + |\mathbf{x}_i|), \end{aligned} \tag{25}$$

where  $\bar{\rho}, \bar{\mu}, c_\beta, c_\gamma, c_\tau > 0$  are upperbounds for their associated parameters,  $\bar{\rho} < 1$ , and  $c > \max\{c_\tau + \bar{\mu}, c_\gamma + c_\beta/(1 - \bar{\rho})\}$ . Let's calculate a uniform bound for  $q_{its}^2(\theta)$ .

$$\begin{aligned} q_{its}^2(\theta) &= \left( \tau_s - \mu - \gamma' \bar{\mathbf{x}}_i - \sum_{k=1}^t \rho^{t-k} \beta' x_{ik} \right)^2 \leq \left( |\tau_s - \mu| + |\gamma' \bar{\mathbf{x}}_i - \sum_{k=1}^t \rho^{t-k} \beta' x_{ik}| \right)^2 \\ &\leq \left( |\tau_s - \mu| + (|\gamma| + |\beta|/(1 - |\rho|)) \max_{1 \leq t \leq T} |x_{it}| \right)^2 \end{aligned}$$

This implies that  $\sup_{\theta \in \Theta} q_{its}^2(\theta) \leq c^2(1 + |\mathbf{x}_i|)^2$  where  $c$  is defined above. Next, let's find the first derivative of  $q_{its}(\theta)$  and then calculate a uniform bound for it.

$$\begin{aligned} \frac{\partial q_{its}(\theta)}{\partial \theta} &= \left( \frac{\partial q_{its}(\theta)}{\partial \rho}, \frac{\partial q_{its}(\theta)}{\partial \mu}, \frac{\partial q_{its}(\theta)}{\partial \beta'}, \frac{\partial q_{its}(\theta)}{\partial \gamma'}, \frac{\partial q_{its}(\theta)}{\partial \sigma_\alpha^2}, \frac{\partial q_{its}(\theta)}{\partial \tau'} \right)' \\ &= \left( -\sum_{k=1}^{t-1} (t-k) \rho^{t-1-k} \beta' x_{ik}, -1, -\sum_{k=1}^t \rho^{t-k} x'_{ik}, -\bar{\mathbf{x}}'_i, 0, e'_{s-1} \right)', \end{aligned}$$

where  $e_s = (0, \dots, 0, 1, 0, \dots, 0)'$  is the basis vector where the 1 appears in the  $s^{\text{th}}$  position. Note that, in the derivative we have  $e_{s-1}$  since  $\tau = (\tau_2, \dots, \tau_{S-1})$ . Based on the

above derivations, we can conclude that  $\sup_{\theta \in \Theta} |\partial q_{its}(\theta) / \partial \theta| \leq c_1(1 + |\mathbf{x}_i|)$  for some constant  $c_1 > \max\{1, c_\beta / (1 - \bar{\rho})^2, 1 / (1 - \bar{\rho})\}$ . The second derivative of  $q_{its}(\theta)$  contains all zeros except for  $\partial^2 q_{its}(\theta) / \partial \rho^2 = -\sum_{k=1}^{t-2} (t-k)(t-k-1) \rho^{t-2-k} \beta' x_{ik}$  and  $\partial^2 q_{its}(\theta) / \partial \rho \partial \beta = (\partial^2 q_{its}(\theta) / \partial \beta' \partial \rho)' = -\sum_{k=1}^{t-1} (t-k) \rho^{t-1-k} x_{ik}$ . Therefore,  $\sup_{\theta \in \Theta} |\partial^2 q_{its}(\theta) / \partial \theta \partial \theta'| \leq c_2(1 + |\mathbf{x}_i|)$  for some constant  $c_2 > \max\{1, 2 / (1 - \bar{\rho})^3\}$ . This is another difference between the static probit and autoregressive probit. In the former model where  $\rho = 0$ , we have  $\partial^2 q_{its}(\theta) / \partial \theta \partial \theta' = 0$ , which facilitates the calculations tremendously and renders the full likelihood function to stay concave.