Author/s:
Ando, T;Bai, J;Li, K

Title:
Bayesian and maximum likelihood analysis of large-scale panel choice models with unobserved heterogeneity

Date:
2022-09-01

Citation:
Ando, T., Bai, J. & Li, K. (2022). Bayesian and maximum likelihood analysis of large-scale panel choice models with unobserved heterogeneity. Journal of Econometrics, 230 (1), pp.20-38. https://doi.org/10.1016/j.jeconom.2020.11.013.

Persistent Link:
https://hdl.handle.net/11343/258637

Bayesian and maximum likelihood analysis of large-scale panel choice models with unobserved heterogeneity*

Tomohiro Ando†, Jushan Bai‡ and Kunpeng Li§

December 1, 2020

Abstract

This paper considers the estimation and inference procedures for the case of a logistic panel regression model with interactive fixed effects, where multiple individual effects are allowed and the model is capable of capturing high-dimensional cross-section dependence. The proposed model also allows for heterogeneous regression coefficients. New Bayesian and non-Bayesian approaches are introduced to estimate the model parameters. We investigate the asymptotic behaviors of the estimated parameters. We show the consistency and asymptotic normality of the estimated regression coefficients and the estimated interactive fixed effects when both the cross-section and time-series dimensions of the panel go to infinity. We prove that the dimensionality of the interactive effects can be consistently estimated by the proposed information criterion. Monte Carlo simulations demonstrate the satisfactory performance of the proposed method. Finally, the method is applied to study the performance of New York City medallion drivers in terms of efficiency.

JEL Classification: **C11, C33, C35**.
Key Words: Cross-sectional and serial dependence; Endogeneity; Factor analysis; Heterogeneous panel; Nonlinear panel data.

*The authors would like to thank the guest editors and anonymous reviewers for their constructive and helpful comments, which have considerably improved the quality of the paper. The authors sincerely thank Mehmet Caner, Ulrich Müller, Hashem Pesaran, Matthew Shum, Liangjun Su, Michael Wolf and Jun Yu for their constructive comments, especially Liangjun Su for his comment on a panel data model with common regressors. We are grateful for the comments and suggestions from the participants in the 4th annual conference of the International Association for Applied Econometrics 2017 (IAAE 2017), the 14th International Symposium on Econometric Theory and Applications (SETA 2018), the 29th $(\mathrm{EC})^2$ conference "Big Data Econometrics with Applications", and seminars at Monash University, Singapore Management University, the University of Southern California, and the University of Sydney. The authors are listed in alphabetical order and contribute equally to this paper. Kunpeng Li gratefully acknowledges financial support from NSFC No. 71722011, No. 71571122.

†Melbourne Business School, Melbourne University, T.Ando@mbs.edu. 200 Leicester Street, Carlton, Victoria 3053, Australia.

‡Department of Economics, Columbia University, jb3064@columbia.edu. 1019 International Affairs Building 420 West 118 Street New York, NY 10027 USA

§International School of Economics and Management, Capital University of Economics and Business, likp.07@gmail.com. Huaxiang Town, Fengtai District Beijing, China, 100070

1 Introduction

This paper conducts Bayesian and non-Bayesian analyses of a large-scale panel choice model with interactive fixed effects (IFEs), where both the cross-section and time-series dimensions go to infinity. An advantage of our proposed model is that it deals with high-dimensional cross-section dependence and endogeneity in that the regressors are allowed to have arbitrary correlations with the IFEs. Additionally, the regression coefficients are heterogeneous, and thus, the model is very flexible compared with a homogeneous model, where all individual units share the same sensitivity to the regressors.

There are several benefits of the Bayesian approach. First, one can incorporate prior knowledge into the model and model parameters. For example, one can incorporate past information about a parameter to form a prior distribution for future analysis. When new observations arrive, the previous posterior information about the parameter can be employed as a prior distribution. Second, with the help of Bayesian Markov chain Monte Carlo (MCMC) simulations, one can directly estimate arbitrary functions of the parameters together with credible intervals. Third, parameter uncertainty is addressed in the Bayesian approach, whereas it is ignored in the frequentist approach. Fourth, the success of the frequentist approach relies critically on using asymptotic confidence intervals to approximate finite sample intervals. When the sample size is limited, the reliability of this approximation may become questionable. In contrast, Bayesian MCMC simulations directly construct credible intervals based on panel data and prior information without relying on large sample theory. With this advantage, it is thus worthwhile to develop an MCMC algorithm for Bayesian estimation and inference.

In the literature on Bayesian methodologies, studies on logistic regression models usually focus on the analysis of cross-sectional data (e.g., Polson and Scott (2013), Holmes and Held (2006)). Studies on panel data sets are lacking. Bayesian studies on panel logistic models with endogeneity are even rarer, although modeling endogeneity is an important issue in economics. To fill this gap, this paper conducts a Bayesian analysis on the proposed logit model and develops a Bayesian MCMC algorithm for parameter estimation. The proposed Bayesian analysis approach can be easily extended to the probit case.

In addition to the proposed Bayesian framework, another primary contribution of this paper is to develop a frequentist approach to parameter estimation by maximizing the likelihood function. We emphasize that this is a non-trivial task because of the existence of IFEs. The asymptotic properties of maximum likelihood estimators (MLEs), including consistency and asymptotic normality, are studied. Our asymptotic analysis encounters several theoretical difficulties, including the nonlinearity of the logistic objective function, the largeness of the dimensional regression coefficients, and the cross-sectional and temporal incidental parameters present due to the IFEs. To address these difficulties, some new arguments are developed. We establish the average convergence rates of the estimators and show their

asymptotic normality. To the best of our knowledge, this is the first study to rigorously derive these results for panel logistic regression models with IFEs under large panel dimensions. We also prove that the dimensionality of interactive effects can be consistently determined by the information criterion if the penalty satisfies some regularity conditions.

Although there are a rapidly increasing number of studies on linear panel models with unobserved factor structures where both the cross-section and time-series dimensions of the panel are large (Bai (2009), Pesaran (2006), among others), studies on panel choice models with unobserved factor structures are still in their infancy. Several related studies include those of Fernández-Val and Weidner (2016), Sun (2016), Moon, Shum and Weidner (2018), Charbonneau (2017), Boneva and Linton (2017) and Chen, Fernández-Val and Weidner (2018). All of these studies considered only the frequentist approach. Fernández-Val and Weidner (2016) and Sun (2016) studied nonlinear panel data models with individual and time effects. Charbonneau (2017) also studied nonlinear panel data models with additive effects and homogeneous regression coefficients. These studies therefore did not address panel choice models with IFEs. Recently, Chen, Fernández-Val and Weidner (2018) also considered inferences based on a similar model to ours. However, they specified homogeneous regression coefficients and a single factor, while our study allows for heterogeneous regression coefficients and multidimensional IFEs. Because of these two features, our arguments are quite different from theirs. Boneva and Linton (2017) studied a panel probit model with interactive effects. They proposed an estimator belonging to the class of common correlated effects estimators (Pesaran (2006)). While their estimator is simple to compute, as Boneva and Linton (2017) noted, their approach is valid only if the unobserved factors are contained within the space spanned by the observed factors and the cross-sectional averages of the regressors. In contrast to their approach, our method does not require this restrictive assumption and is thus more desirable in applications. Additionally, we show how our idea can be extended to a direct inference approach for a panel probit model with IFEs.

Our contributions to the literature are summarized as follows. First, this paper is the first to investigate estimation and inference for a high-dimensional panel choice model with multidimensional IFEs and heterogeneous slope coefficients. Second, as a Bayesian contribution, we develop a Bayesian data augmentation approach for estimating the parameters of our proposed logit model. Because of the nonlinearity of the objective function and the restrictions on the parameter space, it is not a straightforward task to develop this Bayesian estimation procedure. Additionally, we show that the proposed Bayesian data augmentation method can be extended to a probit model. More specifically, it can be implemented based solely on the Gibb sampling approach. It is well known that the computational efficiency of Gibb sampling is better than that of the Metropolis-Hastings algorithm. Polson and Scott (2013) noted that Bayesian probit models are extensively used, for example, in both political science and market research

(e.g., Rossi, Allenby and McCulloch (2005) and Jackman (2009)). Thus, this paper provides great benefits for a wide audience. Third, we also consider the frequentist estimation method and establish the consistency and asymptotic behavior of the MLE. Some novel arguments have been developed to establish asymptotic properties. The information criterion for selecting the dimensions of the IFEs is proposed. Fourth, we apply the proposed method to a large-scale panel dataset on the performance of New York City medallion drivers in terms of efficiency.

The paper is organized as follows. Section 2 introduces the formulation of a panel logistic regression model with IFEs. In Section 3, we introduce the chosen estimation and model selection methods. Section 4 investigates some asymptotic properties of the proposed method. To save space, all technical proofs are provided in the accompanying online supplementary document. In Section 5, the proposed method is applied to data on taxi capacity utilization. Section 6 extends our proposed procedures to the probit model. Section 7 concludes the paper.

2 Panel logistic regression model with interactive fixed effects

Suppose that there are $i = 1, ..., N$ individuals facing binary choices. At time t, each individual chooses one of the alternatives, which are labeled alternative 1 and alternative 0. We consider the random utility (the difference in utility between alternative 1 and alternative 0) associated with the choice for individual i at time t:

$$u_{it} = \boldsymbol{x}_{it}'\boldsymbol{b}_i + \eta_{it} + \epsilon_{it}, \quad i = 1, \ldots, N, \ t = 1, \ldots, T, \tag{1}$$

where $\boldsymbol{x}_{it}$ is a p-dimensional vector of observed attributes of the alternatives or observed characteristics of the individual; η_{it} denotes the unobserved structure of individual i's utility, which can vary with t; and ϵ_{it} denotes the nonmodeled component of utility (or shocks to preference). Alternative 1 is chosen if and only if $u_{it} > 0$ (i.e., the corresponding utility is higher than that of alternative 0).

We focus on the case in which the unobserved structure η_{it} is modeled with a factor structure:

$$\eta_{it} = \sum_{\ell=1}^{r} f_{\ell t}\lambda_{i\ell} = \boldsymbol{f}_t'\boldsymbol{\lambda}_i, \tag{2}$$

where $\boldsymbol{f}_t$ is an $r \times 1$ vector of unobservable factors, and $\boldsymbol{\lambda}_i$ represents the factor loadings. These are known as interactive effects in the econometric literature (e.g., Bai, 2009). Note that interactive effects are more general than conventional additive effects. To see this, consider the special case in which $f_t = (1, \delta_t)'$ and $\lambda_i = (\alpha_i, 1)'$. It is seen that $f_t'\lambda_i = \alpha_i + \delta_t$, which reduces to the standard model with individual effects and time effects (additive effects). In additive effects models, the influence of the individual effects (α_i) is constant over time, and the influence of the time effects (δ_t) is identical across

individuals. In contrast, a model with interactive effects allows unobserved individual characteristics (λ_i) to have time-varying effects (through f_t). Another interpretation of interactive effects models is that they allow a vector of common shocks or social trends (f_t) to impact individuals in a heterogeneous way (through λ_i).

An important feature of this model is that correlations between the unobserved factor structure η_{it} and the regressors $\boldsymbol{x}_{it}$ (endogeneity) are allowed, while a standard logistic regression model does not permit such correlations. These correlations arise because some of the explanatory variables are themselves decision variables, which are correlated with the unobserved individual effects. This endogeneity problem is common in economics and other social sciences. Ignoring endogeneity, if it exists, results in inconsistent estimations of the parameters of interest. Conventional panel data analyses often assume cross-sectional independence. Interactive-effects models provide a way of modeling cross-sectional dependence because individuals share the same common shocks f_t. These models are effective for modeling high-dimensional cross-sectional dependence.

Discrete choice models are widely used in social science studies. Let $y_{it} \in \{0, 1\}$ denote the observed outcome of a choice, taking value 1 if alternative 1 is chosen and value 0 otherwise. Alternative 1 is chosen if and only if $u_{it} > 0$. For the logistic specification of an idiosyncratic shock ϵ_{it}, the conditional probability of such a choice is given by

$$P(y_{it} = 1|\boldsymbol{x}_{it}, \boldsymbol{b}_i, \boldsymbol{f}_t, \boldsymbol{\lambda}_i) = \frac{\exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)}{1 + \exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)}. \tag{3}$$

Under the assumption that the errors ϵ_{it} are independently and identically distributed, the joint probability of the observed choices $Y \equiv \{y_{it}|i = 1, ..., N; t = 1, ..., T\}$, denoted by $L(Y|X, B, F, \Lambda)$, is

$$L(Y|X, B, F, \Lambda) = \prod_{i=1}^{N}\prod_{t=1}^{T}\left[\frac{\exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)}{1 + \exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)}\right]^{y_{it}}\left[\frac{1}{1 + \exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)}\right]^{1-y_{it}}, \tag{4}$$

where $X \equiv \{\boldsymbol{x}_{it}|i = 1, ..., N, t = 1, ..., T\}$, $\Lambda = (\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_N)'$, $B = (\boldsymbol{b}_1, ..., \boldsymbol{b}_N)'$ and $F = (\boldsymbol{f}_1, ..., \boldsymbol{f}_T)'$.

In the next section, we introduce a new parameter estimation procedure. We also introduce an information criterion for determining the dimensionality of the IFEs.

3 Estimation and model selection

The inference problem for a panel logistic regression model with IFEs is non-trivial.

3.1 Frequentist estimation

It is well known that the factor loadings $\boldsymbol{\lambda}_i$ and the factors $\boldsymbol{f}_t$ suffer the so-called rotational indeterminacy. We refer to Bai and Ng (2013), Bai and Li (2012) and Stock and Watson (2002) for the identifica-

tion issue in factor models. Therefore, we need to impose conditions to achieve a full identification. In this paper, we adopt the following identification conditions.

Identification (normalization) conditions (IC1):

$$\frac{1}{T}F'F = I_r,\ \frac{1}{N}\Lambda'\Lambda = D,\ \text{where } D \text{ is a diagonal matrix whose diagonal elements are distinct.} \quad (5)$$

In some studies, the above identification conditions are also called the normalization conditions. The latter terminology emphasizes that the above restrictions have no effect on the maximum value of the likelihood function, that is, they are loose restrictions. In Appendix A of the online supplementary document, we provide a rigorous proof of the zero values of Lagrange multipliers. With this fact, the analysis can be greatly simplified. We note that the distinct values in the diagonal elements of D are crucial in our proof. The same requirement has been emphasized in other studies to achieve full identification in factor models.

Throughout the paper, we assume that the true factor and true factor loadings also satisfy (5), i.e.,

$$\frac{1}{T}F_0'F_0 = I_r, \quad \frac{1}{N}\Lambda_0'\Lambda_0 = D.$$

Under these two conditions, the rotation matrix H is an identity matrix, as shown in previous studies (Bai and Ng (2013)). We can interpret the true factors and factor loadings that satisfy the above normalization conditions. To see this, let R be an $r \times r$ orthonormal matrix such that $R'R = RR' = I_r$ (to be determined later). Note that

$$\boldsymbol{f}_{t,0}'\boldsymbol{\lambda}_{i,0} = \boldsymbol{f}_{t,0}'(\frac{1}{T}F_0'F_0)^{-1/2}RR'(\frac{1}{T}F_0'F_0)^{1/2}\boldsymbol{\lambda}_{i,0}.$$

Let $\boldsymbol{f}_{t,0}^* = R'(\frac{1}{T}F_0'F_0)^{-1/2}\boldsymbol{f}_{t,0}$ and $\boldsymbol{\lambda}_{i,0}^* = R'(\frac{1}{T}F_0'F_0)^{1/2}\boldsymbol{\lambda}_{i,0}$. It follows that $\boldsymbol{f}_{t,0}'\boldsymbol{\lambda}_{i,0} = \boldsymbol{f}_{t,0}^{*\prime}\boldsymbol{\lambda}_{i,0}^*$. However, we can readily verify that

$$\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_{t,0}^*\boldsymbol{f}_{t,0}^{*\prime} = I_r.$$

and

$$\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\lambda}_{i,0}^*\boldsymbol{\lambda}_{i,0}^{*\prime} = R'(\frac{1}{T}F_0'F_0)^{1/2}(\frac{1}{N}\Lambda_0'\Lambda_0)(\frac{1}{T}F_0'F_0)^{1/2}R.$$

If we choose R to be the matrix consisting of the eigenvectors of the matrix $(\frac{1}{T}F_0'F_0)^{1/2}(\frac{1}{N}\Lambda_0'\Lambda_0)(\frac{1}{T}F_0'F_0)^{1/2}$, then $\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\lambda}_{i,0}^*\boldsymbol{\lambda}_{i,0}^{*\prime}$ is the diagonal matrix consisting of the eigenvalues of the same matrix. We rename $\boldsymbol{f}_{t,0}^*$ as $\boldsymbol{f}_{t,0}$ and rename $\boldsymbol{\lambda}_{i,0}^*$ as $\boldsymbol{\lambda}_{i,0}$.

Remark 1 When y_{it} remains constant over time, it is difficult to estimate the corresponding regression coefficients $\boldsymbol{b}_i$ and factor loadings $\boldsymbol{\lambda}_i$. This difficulty also arises when estimating the parameters of a

standard logistic regression model. In such a case, we need to drop the corresponding individuals and estimate the model based on the remainder of the sample. □

In this section, we develop a new frequentist estimation procedure. The maximum likelihood estimator $\{\hat{B}, \hat{F}, \hat{\Lambda}\}$ is given by

$$\{\hat{B}, \hat{F}, \hat{\Lambda}\} = \text{argmax}_{\{B,F,\Lambda\}} L(Y|X, B, F, \Lambda), \tag{6}$$

subject to the normalization conditions described above (IC1). Here, $\hat{\Lambda} = (\hat{\boldsymbol{\lambda}}_1, ..., \hat{\boldsymbol{\lambda}}_N)'$, $\hat{B} = (\hat{\boldsymbol{b}}_1, ..., \hat{\boldsymbol{b}}_N)'$ $\hat{F} = (\hat{\boldsymbol{f}}_1, ..., \hat{\boldsymbol{f}}_T)'$ and the likelihood function is as given in (4).

We note that the likelihood function $L(Y|X, B, F, \Lambda)$ in (4) is a nonlinear function of F, B and Λ and that the product $\boldsymbol{f}_t'\boldsymbol{\lambda}_i$ appears in this likelihood function. Moreover, the identification restrictions on F and Λ are imposed. Therefore, direct optimization is not a straightforward task. For a given number of common factors r, we introduce the following new algorithm.

Step 1. Given F, update $\boldsymbol{b}_i$ and $\boldsymbol{\lambda}_i$ to maximize

$$L(\boldsymbol{b}_i, \boldsymbol{\lambda}_i) \equiv \prod_{t=1}^{T} \left[\frac{\exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)}{1 + \exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)}\right]^{y_{it}} \left[\frac{1}{1 + \exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)}\right]^{1-y_{it}}$$

for $i = 1, ..., N$.

Step 2. Given B and Λ, update $\boldsymbol{f}_t$ to maximize

$$L(\boldsymbol{f}_t) = \prod_{i=1}^{N} \left[\frac{\exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)}{1 + \exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)}\right]^{y_{it}} \left[\frac{1}{1 + \exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)}\right]^{1-y_{it}}$$

for $t = 1, ..., T$.

Step 3. Repeat Step 1 $\sim$ Step 2 until some tolerance condition is satisfied.

Step 4. Let $\widetilde{F}$ and $\widetilde{\Lambda}$ be the final estimators after the iteration process. First, calculate the matrix $M_{NT} = (\frac{1}{T}\widetilde{F}'\widetilde{F})^{1/2} \times (\frac{1}{N}\widetilde{\Lambda}'\widetilde{\Lambda})(\frac{1}{T}\widetilde{F}'\widetilde{F})^{1/2}$ and its associated diagonalization $M_{NT} = Q_{NT}D_{NT}Q_{NT}'$, where Q_{NT} is an orthogonal matrix and D_{NT} is a diagonal matrix. Then, $\hat{\Lambda} = \widetilde{\Lambda}(\frac{1}{T}\widetilde{F}'\widetilde{F})^{1/2}Q_{NT}$ and $\hat{F} = \widetilde{F}(\frac{1}{T}\widetilde{F}'\widetilde{F})^{-1/2}Q_{NT}$.

Remark 2 Because we update $\boldsymbol{\lambda}_i$ given $\boldsymbol{f}_t$ and update $\boldsymbol{f}_t$ given $\boldsymbol{\lambda}_i$, the same rotation matrix passes down from the current iteration to the next iteration. Therefore, it is unnecessary to normalize the estimators during each iteration. This basic fact is used in the Expectation Maximization (EM) algorithm for maximum likelihood estimators in high-dimensional factor models; see Bai and Li (2012). In our simulation study, we set the tolerance condition as

$$\frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{b}_i^{new} - \boldsymbol{b}_i^{current}\|^2 + \frac{1}{NT}\left\|F^{new}(\Lambda^{new})' - F^{current}(\Lambda^{current})'\right\|^2 < 10^{-3}.$$

Here, $\boldsymbol{b}_i^{new}$ and $\boldsymbol{b}_i^{current}$ are the updated and current values, respectively, of $\boldsymbol{b}_i$, and the other symbols are defined similarly. Our simulation study finds that the above algorithm converges quickly.

To implement the above procedure, we need an initial value of F. First, we estimate the regression coefficients $\boldsymbol{b}_i$ $(i = 1, ..., N)$ by maximizing $L(Y|X, B) = \prod_{i=1}^{N} \prod_{t=1}^{T} \left[\frac{\exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i)}{1+\exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i)}\right]^{y_{it}} \left[\frac{1}{1+\exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i)}\right]^{1-y_{it}}$. Given the initial values $\tilde{\boldsymbol{b}}_i$ $(i = 1, ..., N)$, we define a set of variables $\mathcal{Z} = (\mathcal{Z}_1, \ldots, \mathcal{Z}_N)$ as $\mathcal{Z}_i = \boldsymbol{y}_i - X_i\tilde{\boldsymbol{b}}_i$. We then obtain the principal-component-based estimates of F and Λ as described by Bai (2003), that is, $\sqrt{T}$ times the eigenvectors corresponding to the r largest eigenvalues of the $T \times T$ matrix $\mathcal{Z}'\mathcal{Z}$. □

3.2 Bayesian data augmentation approach for parameter inference

In this section, we develop the MCMC approach and generate a set of posterior samples. As discussed in Section 1, the Bayesian approach provides various benefits. In addition, the results obtained by the Bayesian approach are often more numerically stable than those of the maximum likelihood approach. Under a small sample, the maximum likelihood estimator for the logistic regression model may occasionally become unstable. In such cases, the Bayesian approach with a regularization prior for the slope coefficients can be used to avoid such instability issues. It is therefore worthwhile to conduct a Bayesian analysis of our proposed model and provide an MCMC simulation algorithm. Indeed, our simulation study (see Appendix G.3) reveals that the proposed Bayesian approach achieves better estimation accuracy than that of a corresponding frequentist estimator for a small sample size.

For the data augmentation approach, we need to specify the prior distributions of the parameters. For ease of computation, we assume that the priors of the factors and factor loadings are mutually independent, i.e., $\pi(B, F, \Lambda) = \pi(B, \Lambda)\pi(F)$. Then, the posterior density is

$$\pi(B, F, \Lambda|Y, X) \propto L(Y|X, F, \Lambda, B)\pi(B, \Lambda)\pi(F),$$

which does not yield an analytical form.

Within a principal component framework (see, e.g., Bai (2009) and the references therein), the unobservable common factors and their factor loadings are analyzed jointly such that the specified prior takes the form $\pi(B, F, \Lambda) = \pi(B)\pi(\Lambda, F)$. In contrast, in this paper, we jointly analyze the regression coefficients and factor loadings. This treatment provides a convenient data augmentation approach for inference on these unknown parameters. This approach was also employed by Ando and Bai (2020) for estimating the parameters of panel quantile regression models with interactive effects. Moreover, although one might conjecture that equation (1) should allow us to easily derive the conditional posterior distributions of the interactive fixed-effect parameters (F, Λ), it does not lead to an easy method for sampling from these posterior distributions because the error term ϵ_{it} is not normal. Because there have

been no previous studies on the Bayesian analysis of our model (1), it is important to conduct prior and posterior analyses to derive the conditional posterior distributions.

Similar to the frequentist estimation approach, we need to impose identification (normalization) conditions. In this section, we adopt the following identification conditions to develop the Bayesian estimation procedure.

Identification (normalization) conditions (IC2):

$$\frac{1}{T}F'F = I_r,\ \Lambda' = (\Lambda_1', \Lambda_2')',\ \text{with } \Lambda_1 \text{ being an invertible lower triangular matrix.} \tag{7}$$

We use different identification conditions for frequentist and Bayesian estimations because we want to conform with the traditions of existing studies. For the frequentist estimation of factor models, IC1 is widely adopted. However, for the Bayesian estimation of factor models, IC2 is more popular; see, e.g., Geweke and Zhou (1996). We present prior and posterior analyses in the next section.

3.2.1 Prior specification and posterior analysis for B and Λ

Here, we specify the prior densities for B and Λ and derive their conditional posterior distributions. First, the likelihood contribution of observation y_{it} can be expressed as

$$\begin{aligned}
& \left[\frac{\exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)}{1+\exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)}\right]^{y_{it}} \times \left[\frac{1}{1+\exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)}\right]^{1-y_{it}} \\
= & \frac{[\exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)]^{y_{it}}}{1+\exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)} \\
\propto & \exp\Big[v_{it}(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)\Big] \times \int_0^\infty \exp\Big[-\frac{1}{2}\omega_{it}(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)^2\Big] p(\omega_{it}) d\omega_{it} \\
\equiv & \exp(v_{it}\boldsymbol{z}_{it}'\boldsymbol{\gamma}_i) \times \int_0^\infty \exp\Big[-\frac{1}{2}\omega_{it}(\boldsymbol{z}_{it}'\boldsymbol{\gamma}_i)^2\Big] p(\omega_{it}) d\omega_{it},
\end{aligned}$$

where $v_{it} = y_{it} - 1/2$ and $p(\omega_{it})$ is the density of a Pólya-Gamma random variable with parameters $(1, 0)$. In the cross-sectional context, this expression was also obtained by Polson and Scott (2013). The role of ω_{it} is important because it allows us to obtain the conditional posterior distribution of $\boldsymbol{\gamma}_i$ analytically. For simplicity of notation, we adopt $\boldsymbol{z}_{it} = (\boldsymbol{x}_{it}', \boldsymbol{f}_t')'$ and $\boldsymbol{\gamma}_i = (\boldsymbol{b}_i', \boldsymbol{\lambda}_i')'$. Note that the first r factor loading vectors $\boldsymbol{\lambda}_i$ correspond to the invertible lower triangular matrix Λ_1, which comes from the identification restriction. If some elements of $\boldsymbol{\lambda}_i$ must be zero for identification purposes, we can ignore these elements in the estimation process and denote the nonzero elements of $(\boldsymbol{b}_i', \boldsymbol{\lambda}_i')'$ as $\boldsymbol{\gamma}_i$.

Here, we simply use the diffuse prior $\pi(\boldsymbol{\gamma}_i) \propto \text{Const}$. Then, the conditional posterior density of $\boldsymbol{\gamma}_i = (\boldsymbol{b}_i', \boldsymbol{\lambda}_i')'$ is

$$\pi(\boldsymbol{\gamma}_i | Y, X, F, B_{-i}, \Lambda_{-i}, \Omega) \quad \propto \quad \prod_{t=1}^{T} \exp\Big[v_{it}\boldsymbol{z}_{it}'\boldsymbol{\gamma}_i - \frac{1}{2}\omega_{it}(\boldsymbol{z}_{it}'\boldsymbol{\gamma}_i)^2\Big]$$

$$\propto \quad \exp\Big[-\frac{1}{2}(\Omega_i^{-1}\boldsymbol{v}_i - W_i\boldsymbol{\gamma}_i)'\Omega_i(\Omega_i^{-1}\boldsymbol{v}_i - W_i\boldsymbol{\gamma}_i)\Big],$$

which implies that the conditional posterior density of $\boldsymbol{\gamma}_i$ is a multivariate normal density with a mean of $(W_i'\Omega_i W_i)^{-1}W_i'\boldsymbol{v}_i$ and a variance-covariance matrix of $(W_i'\Omega_i W_i)^{-1}$. Here, $\Omega \equiv \{\omega_{it}|i = 1, ..., N, t = 1, ..., T\}$, $\Omega_i = \text{diag}(\omega_{i1}, ..., \omega_{iT})$, $W_i = (X_i, F)$ is the design matrix, $B_{-i} = (\boldsymbol{b}_1, ..., \boldsymbol{b}_{i-1}, \boldsymbol{b}_{i+1}, ..., \boldsymbol{b}_N)'$, $\Lambda_{-i} = (\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_{i-1}, \boldsymbol{\lambda}_{i+1}, ..., \boldsymbol{\lambda}_N)'$ and $\boldsymbol{v}_i = (v_{i1}, ..., v_{iT})'$.

Remark 3 Different priors imply different prior knowledge about the model. For example, one can impose the multivariate normal prior $N(\boldsymbol{\mu}_{\gamma_i}, V_{\gamma_i})$ on $\boldsymbol{\gamma}_i$. A large variance V_{γ_i} should be set when the confidence of the information is weak, and vice versa. When the dimensionality of $\boldsymbol{x}_{it}$ is high and the true structure is sparse, a Bayesian shrinkage prior can be employed (see, for example, Park and Casella (2008), Carvalho, Polson and Scott (2010) and Leng, Tran and Nott (2014)). □

3.2.2 Conditional posterior density of ω_{it}

We can easily obtain the conditional posterior density of ω_{it}, that is,

$$\pi(\omega_{it}|Y, X, F, B, \Lambda, \Omega_{-\omega_{it}}) \quad \propto \quad \exp\Big[-\frac{1}{2}\omega_{it}(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)^2\Big]p(\omega_{it}),$$

which is a Pólya-Gamma distribution with parameters $(1, \boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)$. Here, $\Omega_{-\omega_{it}} \equiv \{\omega_{js}|j = 1, ..., N, s = 1, ..., T, j \neq i, s \neq t\}$, Again, we can easily draw posterior samples of ω_{it} using the Gibbs sampler.

3.2.3 Prior specification and posterior analysis for F

Combining the terms from all observations yields the following expression for the conditional posterior of F:

$$\begin{aligned}
\pi(F|Y, X, B, \Lambda, \Omega) \quad &\propto \quad \pi(F)\prod_{i=1}^{N}\prod_{t=1}^{T}\Big[\exp\{v_{it}(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)\} \times \exp\{-\omega_i(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)^2/2\}\Big] \\
&\propto \quad \pi(F)\prod_{i=1}^{N}\prod_{t=1}^{T}\exp\Big[-\frac{\omega_{it}}{2}(\frac{v_{it}}{\omega_{it}} - \boldsymbol{x}_{it}'\boldsymbol{b}_i - \boldsymbol{f}_t'\boldsymbol{\lambda}_i)^2\Big] \\
&\propto \quad \pi(F)\exp\Big[-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{v}_i^* - F\boldsymbol{\lambda}_i)'\Omega_i(\boldsymbol{v}_i^* - F\boldsymbol{\lambda}_i)\Big], \qquad (8)
\end{aligned}$$

where $\boldsymbol{v}_i^* = (v_{i1}^*, ..., v_{iT}^*)$, with $v_{it}^* = v_{it}/\omega_{it} - \boldsymbol{x}_{it}'\boldsymbol{b}_i = (y_{it} - 1/2)/\omega_{it} - \boldsymbol{x}_{it}'\boldsymbol{b}_i$.

Now, let us further investigate the form of the conditional posterior of F. In this paper, the common factors F are subject to the normalization condition $F'F/T = I_r$ for identification purposes. Because $F'F/T = I_r$, F belongs to a hyperball in T dimensions, and its support is restricted to be a Cartesian

product of the T-dimensional hyperball. Moreover, because of the orthogonality requirement, its support is further reduced to a Stiefel manifold $S_{T,r}$ of radius $\sqrt{T}$ (Khatri and Mardia (1977)). Therefore, the prior of F is a flat prior over the Stiefel manifold corresponding to the orthogonal transformations and, hence, it is invariant with respect to the orthogonal group. Specifically, the prior of F is

$$\pi(F) = \frac{1}{C(T,r)} \cdot 1(F \in S_{T,r}), \tag{9}$$

where $1(\cdot)$ is the indicator function and

$$C(T,k) = \frac{2^k \pi^{kT/2} T^{k(2T-k-1)/4}}{\pi^{k(k-1)/4} \prod_{j=1}^{k} \Gamma\{(T-j+1)/2\}}$$

is a normalizing constant, with $\Gamma(\cdot)$ being the gamma function.

However, under the prior $\pi(F)$ given in (9), the analysis of the conditional posterior of F in (8) is still not straightforward. This is mainly because the diagonal matrix Ω_i prevents the derivation of an analytical conditional posterior for F. We therefore use the Metropolis-Hastings algorithm to generate posterior samples of F. Given the posterior density $\pi(F|Y, X, B, \Lambda, \Omega)$, which is known up to a constant, and a proposal conditional density $p(F)$, we can generate posterior samples of F in the following way.

To generate samples from $\pi(F|Y, X, B, \Lambda)$, the Metropolis-Hastings algorithm requires us to specify a proposal density $p(F)$. Then, the algorithm draws a candidate parameter value F^{new} from the proposal density $p(F)$. The generated parameter value F^{new} is either accepted or rejected based on the acceptance probability

$$\alpha = \min\left\{1, \frac{L(Y|X, F^{new}, \Lambda, B)\pi(B, F^{new}, \Lambda)/p(F^{new})}{L(Y|X, F^{old}, \Lambda, B)\pi(B, F^{old}, \Lambda)/p(F^{old})}\right\},$$

where F^{old} is the current state of F.

Thus, for a practical implementation of the Metropolis-Hasting algorithm, we need to prepare a proposal density. Here, the random-walk Metropolis-Hastings algorithm is used. We draw a new candidate F^{new} from the proposal density

$$p(F) \propto \exp\left\{-\mathrm{tr}\{(V - F\Lambda)'(V - F\Lambda)\}\right\} \cdot 1(F \in S_{T,r}),$$

where F is on the Stiefel manifold and $V = (\boldsymbol{v}_1, ..., \boldsymbol{v}_N)$. A simulation of the matrix Bingham-von Mises-Fisher distribution has been described by Hoff (2009). In our simulation study, this proposal density works well.

3.2.4 Posterior sampling algorithm

As discussed above, we can analytically obtain the conditional posterior distributions of B, Λ and Ω. Therefore, we can easily draw posterior samples by implementing the Gibbs sampling algorithm. To

sample F, we can use the Metropolis-Hastings algorithm. For a given number of common factors r, our Bayesian MCMC algorithm is summarized as follows.

Posterior sampling algorithm:

Step 1. Initialize the parameters.

Step 2. Sample F from $\pi(F|Y, X, B, \Lambda, \Omega)$.

Step 3. Sample $\boldsymbol{\gamma}_i$ from $\pi(\boldsymbol{\gamma}_i|Y, X, F, B_{-i}, \Lambda_{-i}, \Omega)$ for $i = 1, ..., N$.

Step 4. Sample ω_{it} from $\pi(\omega_{it}|Y, X, F, B, \Lambda, \Omega_{-\omega_{it}})$ for $i = 1, ..., N$ and $t = 1, ..., T$.

Step 5. Repeat Step 2 to Step 4 for a sufficiently large number of iterations.

The outcome of the above algorithm can be regarded as a random sample from the joint posterior density function after a burn-in period. To implement this posterior sampling algorithm, we use the frequentist estimator for parameter initialization. With this starting point, the posterior sampling process quickly converges to the posterior distribution. We thus obtain a posterior sample of size H, denoted by $\{B^{(k)}, F^{(k)}, \Lambda^{(k)}; k = 1, ..., H\}$, for inference purposes. Using the MCMC approach, Jacquier, Johannes and Polson (2005) developed a pure simulation-based method for computing maximum likelihood estimates from latent variable models. Chernozhukov and Hong (2003) studied Laplace-type estimators, which include quantiles of quasi-posterior distributions. In our case, the maximum likelihood estimators $\{\tilde{B}, \tilde{F}, \tilde{\Lambda}\}$ based on MCMC sampling are given by

$$\{\tilde{B}, \tilde{F}, \tilde{\Lambda}\} = \text{argmax}_{\{B^{(k)}, F^{(k)}, \Lambda^{(k)}\}; k=1,...,H} L(Y|X, B^{(k)}, F^{(k)}, \Lambda^{(k)}), \tag{10}$$

where the likelihood function is as given in (4).

For Step 1, we can employ the estimation procedure of the frequentist algorithm. After the process converges to the maximum likelihood estimator like in (6), the initial starting values of F and Λ can be obtained by applying the normalization procedure (IC2) of Bai and Ng (2013).

3.3 Model selection

In practice, we need to determine the dimensionality of the interactive effects. Because of the presence of iterative effects, cross-validation cannot be applied easily. Ando and Bai (2017), Bai and Ng (2002), and Hallin and Liška (2007) proposed several model selection criteria. Unfortunately, these criteria are applicable only for linear panel data models and thus cannot be applied to panel choice models. In this paper, we develop an information criterion exclusively for our panel choice model. More specifically, the dimensionality of the interactive effects is determined by minimizing the following information criterion:

$$IC(r) = -2 \log L(Y|X, \hat{B}^{(r)}, \hat{F}^{(r)}, \hat{\Lambda}^{(r)}) + r \times q(N, T), \tag{11}$$

where $\hat{B}^{(r)}$, $\hat{F}^{(r)}$ and $\hat{\Lambda}^{(r)}$ are the estimated model parameters when the dimensionality of the interactive effects is r. The function $q(N,T)$ is a penalty for model overfitting. In our numerical study, we specify this function as follows:

$$q(N,T) = \left(\frac{N+T}{NT}\right)\log\left(\frac{NT}{N+T}\right). \tag{12}$$

The asymptotic behavior of $IC(r)$ in (11) is investigated in the next section. The penalty function (12) satisfies the conditions given in Theorem 3 in the next section (see, e.g., Bai and Ng (2002)). One can also consider alternative penalty functions. However, this consideration is beyond the scope of this paper.

4 Asymptotic results

There are rich opportunities to apply the proposed method, but theoretical results are lacking in the literature. Therefore, this section presents a theoretical justification of the proposed method. We first state the assumptions needed for the asymptotic analysis. Because the dimensions of B, Λ and F are divergent, we cannot assume that the likelihood functions adhere to the standard regularity conditions.

4.1 Assumptions

We first define some notations. Let $\|A\| = [\text{tr}(A'A)]^{1/2}$ be the usual norm of the matrix A, where "tr" denotes the trace of a square matrix. The equation $a_n = O(b_n)$ indicates that the deterministic sequence a_n is of order b_n at most; $c_n = O_p(d_n)$ indicates that the random variable c_n is of order d_n at most in terms of probability, whereas $c_n = o_p(d_n)$ indicates that c_n is of a smaller order than d_n in terms of probability. We use $\boldsymbol{b}_{i,0}$ to denote the true regression coefficients. Additionally, $\boldsymbol{f}_{t,0}$ and $\boldsymbol{\lambda}_{i,0}$ are the true common factors at time t and the true factor loadings of individual i, respectively. The set of regularity conditions imposed on the proposed model is as follows:

Assumption A: Common factors

Let $\mathcal{F}$ be a set that contains all possible values of the generic factors $\boldsymbol{f}_t$ for $t \leq T$. We assume that $\mathcal{F}$ is a compact subset of $\mathbb{R}^r$. The true factors $\boldsymbol{f}_{t,0}$ are a sequence of fixed constants such that $T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_{t,0}\boldsymbol{f}'_{t,0} = I_r$.

Assumption B: Factor loadings and regression coefficients

Let $\mathcal{B}$ and $\mathcal{L}$ be two sets that contain all possible values of the generic regression coefficients $\boldsymbol{b}_i$ and generic factor loadings $\boldsymbol{\lambda}_i$ for all i, respectively. We assume that $\mathcal{B}$ and $\mathcal{L}$ are compact subsets of $\mathbb{R}^p$ and $\mathbb{R}^r$, respectively. In addition, the true loadings $\boldsymbol{\lambda}_{i,0}$ are a sequence of fixed constants such that

$\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\lambda}_{i,0}\boldsymbol{\lambda}_{i,0}' \to D_0$, where D_0 is a diagonal matrix with its diagonal elements distinct and the smallest diagonal element bounded away from zero.

Assumption C: Idiosyncratic error terms

The idiosyncratic error ϵ_{it} in (1) has a standard logistic distribution, is independent over i and t and is independent of $\boldsymbol{x}_{js}$ for all j and s.

Assumption D: Predictors and design matrix

(D.1): For a positive constant C, the predictors are random and almost surely satisfy $\sup_{it}\|\boldsymbol{x}_{it}\| < C$.

(D.2): Define $A_i = \frac{1}{T}X_i'M_FX_i$, $B_i = (\boldsymbol{\lambda}_{i,0}\boldsymbol{\lambda}_{i,0}')\otimes I_T$, $C_i = \frac{1}{\sqrt{T}}[\boldsymbol{\lambda}_{i,0}\otimes(M_FX_i)]'$, $\boldsymbol{\eta} = \frac{1}{\sqrt{T}}\text{vec}(M_FF_0)$, and $M_F = I - F(F'F)^{-1}F'$. Let $\mathscr{F}$ be the collection of F such that $\mathscr{F} = \{F : F'F/T = I_r\}$. We assume that with a probability approaching one,

$$\inf_{F\in\mathscr{F}}\lambda_{\min}\left[\frac{1}{N}\sum_{i=1}^{N}E_i(F)\right] > 0,$$

where $\lambda_{\min}(Q)$ denotes the minimum eigenvalue of a matrix Q and $E_i(F) = B_i - C_i'A_i^{-1}C_i$.

(D.3): Let $\mathcal{X}(B)$ be an $N\times T$ matrix with its (i,t)th entry equal to $\boldsymbol{x}_{it}'\boldsymbol{b}_i$, where $B = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_N)'$. For all any nonzero B, there exists a positive constant $\breve{c} > 0$ such that with a probability approaching one,

$$\frac{1}{NT}\|M_{\Lambda_0}\mathcal{X}(B)M_{F_0}\|^2 \geq \breve{c}\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{b}_i\|^2.$$

(D.4): For each i, we assume that there exists a constant $c > 0$ such that for each i, with a probability approaching one,

$$\liminf_{T\to\infty}\lambda_{\min}\Big(\frac{1}{T}X_i'M_{F^0}X_i\Big) \geq c.$$

Remark 4 Assumption A treats the factors as parameters. Similar assumptions were made in Bai (2009), Bai and Li (2014), Moon and Weidner (2015), etc. In cases where the factors are random, the analysis in this paper can be viewed as conditioning on a particular realization of factors. This assumption makes the model highly flexible. For example, we can allow one factor to be a normalized time trend (i.e., $f_{tk} = c\frac{t}{T}$ for some constant c) or a constant (i.e., $f_{tk} = 1$ for all t). Assumption A requires that the sample covariance of $\boldsymbol{f}_{t,0}$ equals the identity matrix of the same size. Assumption B requires that the sample covariance of $\boldsymbol{\lambda}_{i,0}$ converges to a diagonal matrix. These two requirements are consistent with our normalization conditions and entail no loss of generality. To see this, for any sets of factors and loadings $(\tilde{F}_0, \tilde{\Lambda}_0)$ such that $\frac{1}{T}\tilde{F}_0'\tilde{F}_0 \to \Sigma_F > 0$ and $\frac{1}{N}\tilde{\Lambda}_0'\tilde{\Lambda}_0 \to \Sigma_\Lambda > 0$, as long as the eigenvalues of

$\Sigma_F\Sigma_\Lambda$ are distinct, we can always perform the manipulations before Remark 1 to obtain the normalized sets (F_0, Λ_0) that satisfy the two requirements. An implicit implication of these two conditions is that the number of common factors is r. Note that we assume that the range of all the parameters, i.e., $\boldsymbol{b}_i, \boldsymbol{\lambda}_i$ and $\boldsymbol{f}_t$ for all i and t, are confined in compact sets. To be consistent with this assumption, the estimators are restricted in compact sets when implementing the estimation procedure. This assumption is used in our theoretical analysis.

Assumption C imposes the independence assumption on the choice error ϵ_{it}. However, we emphasize that this assumption can be relaxed to allow for weak correlations across two dimensions. As noted in Appendix F of the online supplement, the random part of the likelihood function is the same as that of a linear panel model with interactive effects. Therefore, the extension to the general correlation structure follows the standard arguments. This paper does not pursue this work because we want to keep our model simple to highlight the sources of the theoretical difficulties.

Assumption D is used to obtain the asymptotic distributions of the regression coefficients and the structure of the interactive effects. Assumption D.1 is primarily used for the convenience of theoretical analysis and can be relaxed based on the existence of some high moment conditions. With the presence of such high moment conditions and the maximal inequality, we can show that the leading terms of the estimators are unaffected, but the remaining terms become slightly inflated. Therefore, the theoretical results remain the same. Assumption D.2 is the identification assumption for the factor space $P_{F^0} = F_0(F_0'F_0)^{-1}F_0'$. Although the factors suffer from rotational indeterminacy, the factor space is uniquely determined. Assumption D.3 imposes a regularity condition on the design matrix X_i. It is an identification condition for $\boldsymbol{b}_{i,0}$. To see this, we note that Bai (2009) imposed the condition $\inf_{F\in\mathscr{F}} D(F) > 0$ for the case of homogeneous coefficients, where

$$D(F) = \frac{1}{NT}\sum_{i=1}^{N} \widetilde{X}_i'\widetilde{X}_i, \text{ with } \widetilde{X}_i = M_F X_i - \sum_{j=1}^{N} M_F X_j \boldsymbol{\lambda}_{j,0}'(\Lambda_0'\Lambda_0)^{-1}\boldsymbol{\lambda}_{i,0}.$$

If $\boldsymbol{b}_i = \boldsymbol{b}$ for all i, Assumption D.3 reduces to $D(F^0) > 0$ as in Bai (2009) since $\frac{1}{NT}\|M_{\Lambda_0}\mathcal{X}(B)M_{F_0}\|^2 = \boldsymbol{b}'D(F^0)\boldsymbol{b}$. Here, we directly use $D(F^0)$ instead of $\inf_{F\in\mathscr{F}} D(F)$ because $P_{\hat{F}}$ is identified in Assumption D.2. Because of the heterogeneous coefficient specifications, the condition $D(F^0)$ in Bai (2009) is generalized to Assumption D.3. We note that it can be shown that Assumption D.3 holds if $\boldsymbol{x}_{it}$ has an independent and identically distributed component. Both Assumptions D.2 and D.3 preclude some cases of low-rank regressors. In Appendix F of the supplementary document, we provide a detailed discussion on the presence of common regressors (low-rank regressors). Roughly speaking, we need to impose additional assumptions for identification in this case. Assumption D.4 guarantees that the limiting variance of $\hat{\boldsymbol{b}}_i$ is positive definite for each i. □

4.2 Theoretical results

All proofs are given in the online supplementary document. Because the panel size dimensions N and T are divergent, a novel proof is developed. We use $B_0 = (\boldsymbol{b}_{1,0}, ..., \boldsymbol{b}_{N,0})'$, $\Lambda_0 = (\boldsymbol{\lambda}_{1,0}, ..., \boldsymbol{\lambda}_{N,0})'$, and $F_0 = (\boldsymbol{f}_{1,0}, ..., \boldsymbol{f}_{T,0})'$ to denote the true parameter values. The following theorem shows the average convergence rates of $\hat{\boldsymbol{\gamma}}_i \equiv (\hat{\boldsymbol{b}}_i', \hat{\boldsymbol{\lambda}}_i')'$ and $\hat{\boldsymbol{f}}_t$. Because of the high dimensionalities of B, F and Λ, the convergence rate results can only be formulated with some chosen norm. Following Bai (2003), we choose the dimension-adjusted Frobenius norm in this paper.

Theorem 1 *Let* $\delta_{NT} = \max(\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{T}})$. *Under Assumptions A–D, together with the identification conditions, we have*

$$\frac{1}{N}\sum_{i=1}^{N} \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{i,0}\|^2 = O_p(\delta_{NT}^2)$$
$$\frac{1}{T}\sum_{t=1}^{T} \|\hat{\boldsymbol{f}}_t - \boldsymbol{f}_{t,0}\|^2 = O_p(\delta_{NT}^2).$$

Remark 5 Theorem 1 presents the average convergence rates of the estimators, and these are the same as those in high-dimensional factor models (see, e.g., Bai (2003)). Theorem 1 is obtained by working with the likelihood function. As pointed out above, the random component of the likelihood function is the same as that of a linear panel data model. Therefore, this component is relatively easy to analyze. The primary issue comes from the nonrandom part of the likelihood function. Let $L_{NT}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \ell(y_{it}|\boldsymbol{x}_{it}, \theta)$ be the likelihood function, where $\theta = (\boldsymbol{b}_1', \ldots, \boldsymbol{b}_N', \boldsymbol{\lambda}_1', \ldots, \boldsymbol{\lambda}_N', \boldsymbol{f}_1', \ldots, \boldsymbol{f}_T')'$ contains all the parameters to be estimated. Let $E_\vartheta(\cdot)$ denote the expectation of y given all $\boldsymbol{x}_{it}$. Therefore, $E_\vartheta(L_{NT}(\theta))$ denotes the nonrandom part of the objective function. According to classical M-estimation theory on convergence rates (see, e.g., Van Der Vaart and Wellner (1996, Page 289)), we would like to show

$$E_\vartheta(L_{NT}(\theta_0)) - E_\vartheta(L_{NT}(\theta)) \geq c\frac{1}{NT}\|\theta - \theta_0\|^2.$$

We emphasize that without the normalization conditions, such a c does not exist. The proof of Theorem 1 centers on showing the existence of this c with regard to the normalization conditions. Since rotational indeterminacy only occurs in $\boldsymbol{\lambda}_i$ and $\boldsymbol{f}_t$, we need to separate $\boldsymbol{x}_{it}'\boldsymbol{b}_i$ from $\boldsymbol{\lambda}_i'\boldsymbol{f}_t$. In the proof of Theorem 1, we provide some new arguments about this separation. Once the separation is achieved, we apply Lemma 2 of Chen, Dolado and Gonzalo (2019) to show the existence of c. □

With the assumption of average convergence rates, we next show that the asymptotic distributions of the estimated parameters, $\hat{\boldsymbol{\gamma}}_i$, are multivariate normal. Similarly, the asymptotic distributions of the estimated common factors $\hat{\boldsymbol{f}}_t$ are also multivariate normal distributions.

Theorem 2 *Under Assumptions A – D, if $\sqrt{N}/T \to 0$ and $\sqrt{T}/N \to 0$, together with the identification conditions, the asymptotic distribution of $T^{1/2}(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{i,0})$ is a multivariate normal distribution with a mean of* $\mathbf{0}$ *and a covariance matrix of*

$$\overline{\Gamma}_{i,0} \equiv \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \pi_{it}^0 (1 - \pi_{it}^0) \boldsymbol{z}_{it,0} \boldsymbol{z}_{it,0}',$$

where $\boldsymbol{z}_{it,0} = (\boldsymbol{x}_{it}', \boldsymbol{f}_{t,0}')'$ and $\pi_{it}^0 = P(y_{it} = 1 | \boldsymbol{x}_{it}, \boldsymbol{b}_{i,0}, \boldsymbol{f}_{t,0}, \boldsymbol{\lambda}_{i,0})$ is the true choice probability. Moreover, the asymptotic distribution of $N^{1/2}(\hat{\boldsymbol{f}}_t - \boldsymbol{f}_{t,0})$ is a multivariate normal distribution with a mean of $\mathbf{0}$ *and a covariance matrix of*

$$\overline{\Psi}_{t,0} \equiv \operatorname*{plim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \pi_{it}^0 (1 - \pi_{it}^0) \boldsymbol{\lambda}_{i,0} \boldsymbol{\lambda}_{i,0}'.$$

Remark 6 For inference purposes, we need to estimate $\overline{\Gamma}_{i,0}$ and $\overline{\Psi}_{t,0}$. Let $\hat{\pi}_{it} = \exp(\boldsymbol{x}_{it}'\hat{\boldsymbol{b}}_i + \hat{\boldsymbol{f}}_t'\hat{\boldsymbol{\lambda}}_i)/\{1 + \exp(\boldsymbol{x}_{it}'\hat{\boldsymbol{b}}_i + \hat{\boldsymbol{f}}_t'\hat{\boldsymbol{\lambda}}_i)\}$ be the estimated conditional probability. Then, $\overline{\Gamma}_{i,0}$ and $\overline{\Psi}_{t,0}$ are estimated by $\hat{\Gamma}_i = \frac{1}{T}\sum_{t=1}^{T} \hat{\pi}_{it}(1 - \hat{\pi}_{it}) \hat{\boldsymbol{z}}_{it} \hat{\boldsymbol{z}}_{it}'$ and $\hat{\Psi}_t = \frac{1}{N}\sum_{i=1}^{N} \hat{\pi}_{it}(1 - \hat{\pi}_{it}) \hat{\boldsymbol{\lambda}}_i \hat{\boldsymbol{\lambda}}_i'$, respectively, with $\hat{\boldsymbol{z}}_{it} = (\boldsymbol{x}_{it}', \hat{\boldsymbol{f}}_t')'$. □

To consistently determine the number of factors in the panel choice model, we impose the following assumption.

Assumption E: Identification of *b* for an overfitted model

For each k, there exists a constant $\breve{c}_k > 0$ such that, with a probability approaching one,

$$\inf_{F^k, \frac{1}{T}F^{k\prime}F^k = I_k} \frac{1}{NT} \|M_{\Lambda_0} \mathcal{X}(B) M_{F^k}\|^2 \geq \breve{c}_k \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{b}_i\|^2,$$

where $\mathcal{X}(B)$ is defined as in Assumption D.3.

To demonstrate the necessity of Assumption E, consider the linear model

$$Y_i = X_i \boldsymbol{b}_{i,0} + F_0 \boldsymbol{\lambda}_{i,0} + \varepsilon_i, \quad \text{or equivalently} \quad Y = \mathcal{X}(\boldsymbol{b}_0) + \Lambda_0' F_0 + \varepsilon.$$

If there exists an F^k such that $\frac{1}{NT}\|\mathcal{X}(B) M_{F^k}\|^2 = o_p(1)\frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{b}_i\|^2$, we can simply post-multiply by $M_{[F^k, F_0]}$ on both sides, and it is seen that all the useful information (i.e., $\mathcal{X}(B_0) + \Lambda_0' F_0$) is projected out. Therefore, $\boldsymbol{b}_{i,0}$ is not identifiable. Given this, it is necessary to impose $\frac{1}{NT}\|\mathcal{X}(B) M_{F^k}\|^2 \geq c\frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{b}_i\|^2$ for all k. Assumption E is slightly stronger than this requirement because it still requires the preclusion of the effect from the loadings Λ_0, but the theory behind it is the same.

We have the following theorem on the estimation of the number of factors.

Theorem 3 *Under Assumptions A–E, if* $N \to \infty, T \to \infty, N/T \to c \in (0, \infty)$*, then, under the model selection criterion* $IC(r)$ *with a penalty* $q(N,T)$ *satisfying*

$$q(N,T) \to 0 \text{ and } C_{NT} \times q(N,T) \to \infty,$$

where $C_{NT} = \min\{N,T\}$*, the true number of common factors* r_0 *is consistently estimated.*

Remark 7 Under the assumption of homogeneity for the regression coefficients ($\hat{\boldsymbol{b}}_i = \boldsymbol{b}$ for $i = 1, ..., N$), Chen, Fernández-Val and Weidner (2018) studied an average partial effects method to measure the effect of the covariates on the moments of the distribution of the outcome conditional on the covariates and unobserved effects. In the proof of Theorem 2, we obtain the asymptotic expansions of $\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{i,0}$. It would be interesting to explore the asymptotic distribution of the average of the estimated regression coefficients $N^{-1} \sum_{i=1}^{N} \hat{\boldsymbol{\gamma}}_i$. We would like to explore this topic in more detail in a future study. □

5 Application

The concept of operational efficiency has long played an important role in service management. One of the key managerial aspects that determines operational efficiency is the ability to match demand and service capacity. Examples of businesses for which such efficiency is necessary include hair salons, restaurants, and taxi services. This matching process forces a manager to consider, among other factors, what predefined level of capacity/staffing will be needed in a given time period. It is obvious that for a successful decision, the service capacity/staffing levels should be reasonable such that the manager can meet the actual demand, which cannot be observed in advance. To assess performance from this perspective, it is beneficial to have an efficiency measure for capacity utilization. This key performance indicator provides useful information for managers. This section studies efficient capacity utilization in the taxi industry. From a regulatory perspective, the improvement of efficiency measures is important because inefficient taxi service management will lead to idle time and cause passengers to face long wait times.

5.1 Background information and data

We analyze data collected by the New York City Taxi and Limousine Commission (TLC). These data include yellow medallion taxi IDs, masked driver initials, pick-up times, drop-off times, the geographical locations of trip origins and destinations, travel distances and fares. We use a data set made available by Brian and Dan (2016). This data set was obtained through a Freedom of Information Law request from the New York City TLC.

The New York City taxi market is highly regulated in terms of both pricing and entry. First, all medallion taxis use the same pricing scheme. Second, the number of medallions (legal permits to operate a taxi) is capped. Third, yellow medallion taxis are not authorized to conduct prearranged pick-ups. Passengers must pick up yellow medallion taxis from the street. In other words, taxis and passengers need to find one another. From management and regulatory perspectives, it is thus important to understand the efficiency of each yellow medallion taxi in terms of its capacity utilization.

Let y_{it} be an efficiency measure indicating whether medallion taxi i achieves a prespecified capacity utilization rate (Yes=1 or No=0) at time t. More specifically, we set a prespecified level of capacity utilization for every hour. If medallion taxi i drives with passengers for longer than 20 minutes during time period t (in this study, one hour; for example, 10:00 AM–11:00 AM), then, the prespecified capacity utilization rate is successfully achieved. We note that an alternative prespecified rate can be considered depending on management priorities. By sampling data from 3000 medallion taxis between March 1, 2013, and March 31, 2013, we have created a panel of size $(T, N) = (744, 3000)$. Here, $T = 744$ indicates that the data period spans 31 days (24 hours per day $\times$ 31 days), and there are $N = 3000$ individual medallion taxis. Therefore, more than 2 million trip records are considered in this analysis.

Figure 1 provides examples of the measured performances of 3 medallion taxis. The horizontal axis shows the dates, and the vertical axis represents the hours of the studied period. A colored cell indicates that the individual medallion taxi achieved the prespecified capacity utilization rate. A fair amount of heterogeneity can be observed in Figure 1.

Through our empirical analysis, we explore the following questions: Are there any performance variations among the medallion taxis? Are any medallion taxis doing better than the others? If not, how can the capacity utilization rate be improved? Additionally, we address the question of how many common factors underlie a specified explanatory variable. If common factors exist and are correlated with the explanatory variables, then standard maximum likelihood estimation without a factor structure will lead to inconsistent estimation of the regression coefficients. Thus, addressing this issue is important to ensure that the performance measure (see below) is unbiased.

5.2 Model specification and estimation results

In our analysis, we employ the following information for $\boldsymbol{x}_{it}$: Time frame (every hour: MIDNIGHT–1 AM, 1 AM–2 AM,..., 10 PM–11 PM, 11 PM–MIDNIGHT), Weekday (Monday, Tuesday, ..., Friday) and Weekend (Saturday, Sunday, Holiday). By combining Time Frame and Weekday (and Weekend), we can create a set of $48 (= 24 \times 2)$ indicator variables. We also use the one-week-prior performances of the taxis in the same time frame with respect to capacity utilization (Yes or No). For a given week, if a taxi achieved the prespecified capacity utilization rate in the same time frame during the previous week, then

the value of this performance indicator is 1; otherwise, it is 0.

To understand the data generation structure, we obtain an estimator with interactive effects. We generate 2,000 iterations using the proposed posterior sampling algorithms. In practice, we need to select a value for the number of unobservable factors (the dimensionality of F) to adequately describe the information contained in the observed panel data. For this purpose, we use the proposed model selection criterion $IC(r)$ given in (11). We select the optimal number of common factors as the number that minimizes the $IC(r)$ score. In this way, the optimal number of common factors is found to be 2.

After setting the selected number of common factors, we compute the proposed estimators of $\boldsymbol{b}_i$ and the factor loadings $\boldsymbol{\lambda}_i$ for each of the individual customers. Because our purpose is to compare the relative performances of 3000 medallion taxis, the estimated coefficients contain useful information. Figure 2 shows the estimated regression coefficients and the factor loadings[1]. Each column in Figure 2 corresponds to an individual taxi's performance as determined by the predictors and the common factors. At the top of the figure, trees constructed via hierarchical clustering are presented. We can see that the performance of the individual taxis can be categorized into several clusters.

Finally, we check the correlation between the estimated factors (the first factor f_{1t} and the second factor f_{2t}) and the mean of the one-week-prior performances at the same time for all taxis. Note that 'one-week prior performance at the same time' is one of the explanatory variables in the model. The magnitude of the correlation with respect to the first factor is 0.11, and the p-values of the test for zero correlation are less than 1%. Although such a significant correlation is not obtained for the second factor, our findings imply that the unobserved factor structure is correlated with the explanatory variable. As a result, neglecting the unobserved heterogeneity (i.e., endogeneity) leads to inconsistent estimates. Therefore, the term in (2) plays an important role in capturing the unobserved heterogeneity.

5.3 Evaluation of efficient capacity utilization

By using the estimated model, we can evaluate the performances of the 3000 medallion taxis. First, we evaluate the overall performances of the 3000 medallion taxis by comparing their individual regression coefficients. Let $\tilde{B} = (\tilde{\boldsymbol{b}}_1, \tilde{\boldsymbol{b}}_2, ..., \tilde{\boldsymbol{b}}_{3000})'$ be the matrix of the estimated regression coefficients used to create Figure 2. Then, the overall performances can be measured by

$$\boldsymbol{p}_{\text{overall}} \equiv \tilde{B}\mathbf{1},$$

where $\mathbf{1}$ is a vector of ones. The i-th element of the vector $\boldsymbol{p}_{\text{overall}}$ corresponds to the overall performance of the i-th medallion taxi. A large positive value of $p_{i,\text{overall}}$ implies that the i-th medallion taxi has been

[1]To allow the variations to be seen in detail, we multiply the estimated factor loadings by $\sqrt{T}$. Additionally, we truncate the magnitude of the parameter estimates at a certain threshold R. Thus, if the absolute value of an estimate is larger than R, its value is set to either $-R$ or R, depending on the sign.

achieving superior performance to those of the other taxis, and vice versa. By sorting the elements of $\boldsymbol{p}_{\text{overall}}$ in decreasing order, we obtain the bar plot presented in Figure 3. We can see that there are enormous performance variations among the medallion taxis.

Can the taxis in the inferior group improve their performances? To identify a reason for their inferior performance, we analyze the bottom 10% of the medallion taxis. The results of separately applying hierarchical clustering to $\hat{\boldsymbol{b}}_i$ for the top 10% and the bottom 10% of the medallion taxis are shown in Figure 4 (a) and (b), respectively. We can make the following observations. First, the top 10% of the medallion taxis show high homogeneity, as seen in Figure 4 (a). This implies that there seems to be a certain common set of good tactics for achieving high performance. Second, the bottom 10% of the medallion taxis can be roughly separated into two groups, as seen in Figure 4 (b). With respect to daytime performance, one group (on the right in the figure) performs less well than the other. The taxi drivers in this group need to improve their skills in terms of attracting passengers as quickly as possible when vacant. The performance of the other group (in the center and on the left side of the figure) during the daytime on weekday afternoons is only slightly inferior to that of the top 10% of the medallion taxis. However, for some reason, their performance during the morning, during the evening, before midnight, and after midnight is inferior to that of the top 10%. In general, taxi drivers operate in one of two separate shifts (12 hours each, including time for meals and passing the taxi to the next driver). If a manager wants to improve the capacity utilization rate of a taxi, it is recommended to ensure that both shifts are operated properly.

To further explore the performance between 5:00 PM and 8:00 AM the next morning, the pickup locations are explored. Figure 5 and Figure 6 compare the densities of the pick-up locations recorded by the top 10% and the bottom 10% of the medallion taxis, respectively. Darker points correspond to points of higher density. These figures show that the medallion taxis in the bottom 10% tend to be hailed by passengers around Midtown Manhattan. Specifically, they tend to be hailed near Times Square. In contrast, the top 10% are hailed in Lower Manhattan in addition to Midtown Manhattan. Additionally, their pick-up density around Times Square is lower than that of the bottom 10%. Similar patterns are observed for the daytime pick-up densities (between 8:00 AM and 5:00 PM) in Figure 7 and Figure 8. These observations suggest that to improve their performance, the medallion taxi drivers in the bottom 10% should explore matching opportunities around Lower Manhattan.

6 Panel probit regression model with interactive effects

Here, we note that our data augmentation strategy can be extended to the case of a panel probit model with IFEs. Under the assumption of $\varepsilon_{it} \sim N(0,1)$ in (1), we obtain the probit specification of the

conditional choice probability as follows:

$$P(y_{it}=1|\boldsymbol{x}_{it},\boldsymbol{b}_i,\boldsymbol{f}_t,\boldsymbol{\lambda}_i)=\Phi\left(\boldsymbol{x}_{it}'\boldsymbol{b}_i+\boldsymbol{f}_t'\boldsymbol{\lambda}_i\right),$$

where $\Phi(\cdot)$ is the standard normal distribution function. Albert and Chib (1993) developed a data augmentation procedure for a standard probit regression model in the cross-sectional context. We show that we can incorporate their idea to infer a panel probit model with IFEs.

Boneva and Linton (2017) proposed an estimator belonging to the class of common correlated effects estimators. As pointed out in the introduction section, their method depends on some restrictive assumptions. Here, we show that direct inference can be implemented without imposing such assumptions. Because all conditional posterior densities are obtained analytically, one can simply use the Gibbs sampling algorithm.

To develop the data augmentation strategy, we introduce the latent variables z_{it} as follows:

$$w_{it}=\boldsymbol{x}_{it}'\boldsymbol{b}_i+\boldsymbol{f}_t'\boldsymbol{\lambda}_i+\varepsilon_{it},\quad \varepsilon_{it}\sim N(0,1),$$

such that $y_{it}=1(w_{it}\geq 0)$ where $1(\cdot)$ is the indicator function, which takes a value of one if the expression in the brackets is satisfied; otherwise, its value is zero. Noting that $w_{it}=\boldsymbol{x}_{it}'\boldsymbol{b}_i+\boldsymbol{f}_t'\boldsymbol{\lambda}_i+\varepsilon_{it}\geq 0$ implies $-\varepsilon_{it}\leq \boldsymbol{x}_{it}'\boldsymbol{b}_i+\boldsymbol{f}_t'\boldsymbol{\lambda}_i$, we have

$$\begin{aligned}P(y_{it}=1|\boldsymbol{x}_{it},\boldsymbol{b}_i,\boldsymbol{f}_t,\boldsymbol{\lambda}_i)&=P(w_{it}\geq 0|\boldsymbol{x}_{it},\boldsymbol{b}_i,\boldsymbol{f}_t,\boldsymbol{\lambda}_i)=\Phi\left(\boldsymbol{x}_{it}'\boldsymbol{b}_i+\boldsymbol{f}_t'\boldsymbol{\lambda}_i\right)\\&=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(w_{it}-\boldsymbol{x}_{it}'\boldsymbol{b}_i-\boldsymbol{f}_t'\boldsymbol{\lambda}_i)^2\right]I(w_{it}\geq 0)dw_{it}\end{aligned}$$

and

$$P(y_{it}=0|\boldsymbol{x}_{it},\boldsymbol{b}_i,\boldsymbol{f}_t,\boldsymbol{\lambda}_i)=1-P(y_{it}=1|\boldsymbol{x}_{it},\boldsymbol{b}_i,\boldsymbol{f}_t,\boldsymbol{\lambda}_i).$$

Hereafter, we analyze the conditional posterior distributions of F, B and Λ together with the latent variables $\boldsymbol{w}_i=(w_{i1},...,w_{iT})'$ for $i=1,...,N$.

For simplicity, we specify the same prior used in Section 3. We recall the notations $\boldsymbol{z}_{it}=(\boldsymbol{x}_{it}',\boldsymbol{f}_t')'$ and $\boldsymbol{\gamma}_i=(\boldsymbol{b}_i',\boldsymbol{\lambda}_i')'$. By dropping the terms in the joint posterior distribution that are irrelevant with respect to $\boldsymbol{\gamma}_i$, we obtain

$$\pi(\boldsymbol{\gamma}_i|Y,X,B_{-i},\Lambda_{-i},\boldsymbol{w}_i)\propto\exp\Big[-\frac{1}{2}\sum_{t=1}^{T}(w_{it}-\boldsymbol{z}_{it}'\boldsymbol{\gamma}_i)^2\Big].$$

Thus, the conditional posterior distribution of $\boldsymbol{\gamma}_i$ is a normal distribution with a mean of $(\sum_{t=1}^{T}\boldsymbol{z}_{it}\boldsymbol{z}_{it}')^{-1}$ $\sum_{t=1}^{T}\boldsymbol{z}_{it}w_{it}$ and a variance covariance matrix of $(\sum_{t=1}^{T}\boldsymbol{z}_{it}\boldsymbol{z}_{it}')^{-1}$.

Likewise, we obtain

$$\pi(w_{it}|Y,X,F,B,\Lambda) \propto \exp\left[-\frac{(w_{it}-\boldsymbol{z}_{it}'\boldsymbol{\gamma}_i)^2}{2}\right] \times \{y_{it}I(w_{it}\geq 0)+(1-y_{it})I(w_{it}<0)\},$$

which implies that the conditional posterior distribution of w_{it} is

$$\pi(w_{it}|Y,X,F,B,\Lambda) = \begin{cases} TN\left(\boldsymbol{z}_{it}'\boldsymbol{\gamma}_i,1,+\right), & \text{if } y_{it}=1 \\ TN\left(\boldsymbol{z}_{it}'\boldsymbol{\gamma}_i,1,-\right), & \text{if } y_{it}=0 \end{cases},$$

where $TN(\cdot,\cdot,+)$ is a truncated normal distribution with a mean of $\boldsymbol{z}_{it}'\boldsymbol{\gamma}_i$, a variance of 1 and a support of $[0,\infty)$, and $TN(\cdot,\cdot,-)$ is the corresponding truncated normal distribution with a support of $(-\infty,0)$. We can easily generate random samples from such a truncated normal density (see, e.g., Robert (1995)).

With the prior $\pi(F)$ specified in Section 3, the conditional posterior distribution of F is

$$\begin{aligned}
\pi(F|Y,X,B,\Lambda,Z) &\propto \prod_{i=1}^{N}\prod_{t=1}^{T}\Phi(\boldsymbol{x}_{it}'\boldsymbol{b}_i+\boldsymbol{f}_t'\boldsymbol{\lambda}_i)^{y_{it}}\left[1-\Phi(\boldsymbol{x}_{it}'\boldsymbol{b}_i+\boldsymbol{f}_t'\boldsymbol{\lambda}_i)\right]^{1-y_{it}}\Big] \times \pi(F) \\
&\propto \prod_{i=1}^{N}\prod_{t=1}^{T}\exp\Big[-\frac{1}{2}(w_{it}-\boldsymbol{x}_{it}'\boldsymbol{b}_i-\boldsymbol{\lambda}_i'\boldsymbol{f}_t)^2\Big] \times \pi(F) \\
&\propto \exp\Big[-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}_i^*-F\boldsymbol{\lambda}_i)'(\boldsymbol{w}_i^*-F\boldsymbol{\lambda}_i)\Big] \times \pi(F) \\
&\propto \exp\Big[-\frac{1}{2}\mathrm{tr}\Big((W^*-F\Lambda')'(W^*-F\Lambda')\Big)\Big] \cdot 1(F\in S_{T,r}),
\end{aligned}$$

where F is on the Stiefel manifold and $W^*=(\boldsymbol{w}_1^*,...,\boldsymbol{w}_N^*)$ with $\boldsymbol{w}_i^*=(w_{i1}^*,...,w_{iT}^*)'$, where $w_{it}^*=w_{it}-\boldsymbol{x}_{it}'\boldsymbol{b}_i$. Thus, an analytical expression is obtained for the conditional posterior distribution of F, that is, the matrix-variate Bingham-von Mises-Fisher distribution (Hoff (2009)).

Based on the above posterior analysis, the data augmentation strategy for a panel probit model with interactive effects can be summarized as follows. Note that all of the conditional posterior densities are already given above.

Step 1. Initialize the parameters.

Step 2. Sample F from $\pi(F|Y,X,B,\Lambda,Z)$.

Step 3. Sample $\boldsymbol{\gamma}_i$ from $\pi(\boldsymbol{\gamma}_i|Y,X,B_{-i},\Lambda_{-i},\boldsymbol{z}_i)$ for $i=1,...,N$.

Step 4. Sample w_{it} from $\pi(w_{it}|Y,X,F,B,\Lambda)$ for $i=1,...,N$, $t=1,...,T$.

Step 5. Repeat Step 2 to Step 4 for a sufficiently large number of iterations.

Because all conditional posterior densities are obtained analytically, we can simply use the Gibbs sampling algorithm. This is one of the advantages of this data augmentation strategy for the analysis of a panel probit model with interactive effects. Finally, we note that the above idea can be extended to inferences for the multinomial probit case. This can be done by combining our idea with previous

results (see, for example, Geweke, Keane and Runkle (1994), McCulloch, Polson and Rossi (2000) and the references therein).

7 Conclusion

In this paper, we have introduced a new panel logistic regression model with heterogeneous coefficients and IFEs. We have proposed both Bayesian and non-Bayesian parameter estimation procedures, as well as the information criterion required to determine the dimensionality of the IFEs. Numerical results show that the proposed procedure performs well.

This paper has made several theoretical contributions. First, we delivered the average convergence rates of the ML estimates under large cross-section and time-series dimensions. We also studied the asymptotic distributions of the estimated parameters. Moreover, model selection consistency was established for the proposed information criterion. We developed some novel arguments for analyzing these asymptotic results to address the challenging issues related to the nonlinear objective function, heterogeneous coefficients and IFEs.

Here, we point out some interesting extensions. First, individuals generally make a single choice among multiple alternatives, such as transportation modes and occupational fields, by selecting one candidate out of many. Thus, the number of choices exceeds two (see, e.g., McFadden (1973) and Greene (2000, page 860)). Our Bayesian data augmentation strategy should be extended to this more general multiple-choice model with the allowance of IFEs. Second, one may consider a general nonlinear model with the objective function $\ell(y_{it}|\boldsymbol{x}_{it}, \theta_{it})$, which is a concave class of functions in terms of the model parameter θ_{it}. Whether the arguments developed in this paper would work in the general setup is still an open question. Note that the objective function for the logistic model can be represented as a mixture of Gaussians with respect to a Pólya-Gamma distribution. This expression was important for developing our Bayesian data augmentation approach. The Bayesian MCMC procedure for a general nonlinear model may need a different sampling strategy when a mixed expression is not available for $\ell(y_{it}|\boldsymbol{x}_{it}, \theta_{it})$. Finally, as noted in Remark 4, the relaxation of our assumption may be an interesting avenue of investigation. We would like to explore these topics in future studies.

References

Albert, J. and Chib, S. (1993) Bayesian Analysis of Binary and Polychotomous Response Data *Journal of the American Statistical Association*, 88, 669–679.

Amengual, D. and Watson, M. W. (2007) Consistent estimation of the number of dynamic factors in a large N and T panel. *Journal of Business and Economic Statistics*, 25, 91–96.

Ando, T. and Bai, J. (2017) Clustering huge number of time series: A panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association*, 112, 1182–1198.

Ando, T. and Bai, J. (2020) Quantile co-movement in financial markets; A panel quantile model with unobserved heterogeneity. *Journal of the American Statistical Association*, 115, 266–279.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71, 135-171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77, 1229–1279.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *Annals of Statistics*, 40, 436–465

Bai, J. and Li, K. (2014). Theory and methods of panel data models with interactive effects. *Annals of Statistics*, 42, 142–170.

Bai, J. and Ng, S. (2002) Determining the number of factors in approximate factor models. *Econometrica*, 70, 191–221.

Bai, J. and Ng, S. (2013) Principal components estimation and identification of static factors. *Journal of Econometrics*, 176, 18–29.

Boneva, L. and Linton, O. (2017) A discrete choice model for large heterogeneous panels with interactive fixed effects with an application to the determinants of corporate bond issuance. *Journal of Applied Econometrics*, 32, 1226-1243.

Brian, D. and Dan, W. (2016). New York City Taxi Trip Data (2010-2013). University of Illinois at Urbana-Champaign.

Carvalho, C., Polson, N.G. and Scott, J. (2010) The horseshoe estimator of sparse signals, *Biometrika*, **97**, 465–480.

Charbonneau, K. (2017) Multiple fixed effects in binary response panel data models. *Econometrics Journal*, **20**, S1S13.

Chen, M. Fernández-Val, I. and Weidner, M. (2018) Nonlinear factor models for network and panel data. Forthcoming in *Journal of Econometrics*.

Chen, L., Dolado, J., and Gonzalo, J. (2019). Quantile factor models. Unpublished manuscript.

Chernozhukov, V. and Hong, H. (2003) An MCMC approach to classical estimation. *Journal of Econometrics*, 115, 293–346.

Fernández-Val, I. and Weidner, M. (2016) Individual and time effects in nonlinear panel data models with large N, T. *Journal of Econometrics*, 192, 291–312.

Geweke, J., Keane, M. and Runkle, D. (1994) Alternative computational approaches to inference in the multinomial probit model. *Review of Economics and Statistics*, 76, 609–632.

Geweke, J., and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies*, 9, 557-587.

Greene, W. (2000) *Econometric Analysis*, 4th Edition, Prentice Hall.

Hallin, M., and R. Liška (2007) The generalized dynamic factor model: determining the number of factors. *Journal of the American Statistical Association*, 102, 603–617.

Hoff, P. D. (2009). Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2), 438-456.

Holmes, C. and Held, L. (2006) Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1, 145–168.

Jackman, S. (2009). Bayesian analysis for the social sciences (Vol. 846). John Wiley & Sons.

Jacquier, E., Johannes, M. and Polson, N. (2005) MCMC Maximum Likelihood For Latent State Models. *Journal of Econometrics*, 137, 277–728.

Khatri, C. G. and Mardia, K. V. (1977) The von Mises.Fisher distribution in orientation statistics. *Journal of the Royal Statistical Society*, B39, 95–106.

Leng, C., Tran, M.N. and Nott, D. (2014) Bayesian Adaptive Lasso, *Annals of Institute of Statistical Mathematics*, 66, 221–244.

McCulloch, R. E., Polson, N. G. and Rossi. P. E. (2000) A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99, 173–193.

McFadden, D. (1973) Conditional logit analysis qualitative choice behavior. In *Frontiers of Econometrics*, ed. by P. Zarembka, Academic Press, N.Y., pp. 105–42.

Moon, H. and Weidner, M. (2015) Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, 83, 1543–1579.

Moon, H., Shum, M. and Weidner, M. (2018) Estimation of random coefficients logit demand models with interactive fixed effects. *Journal of Econometrics*, 206, 613–644.

Park, T. and Casella, G. (2008) The Bayesian Lasso, *Journal of the American Statistical Association*,103, 681–686.

Pesaran, M. H. (2006) Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*,74, 967–1012.

Polson, N.G. and Scott, J. (2013) Bayesian inference for logistic models using Polya-Gamma latent variables. *Journal of the American Statistical Association*, 108, 1339–1349.

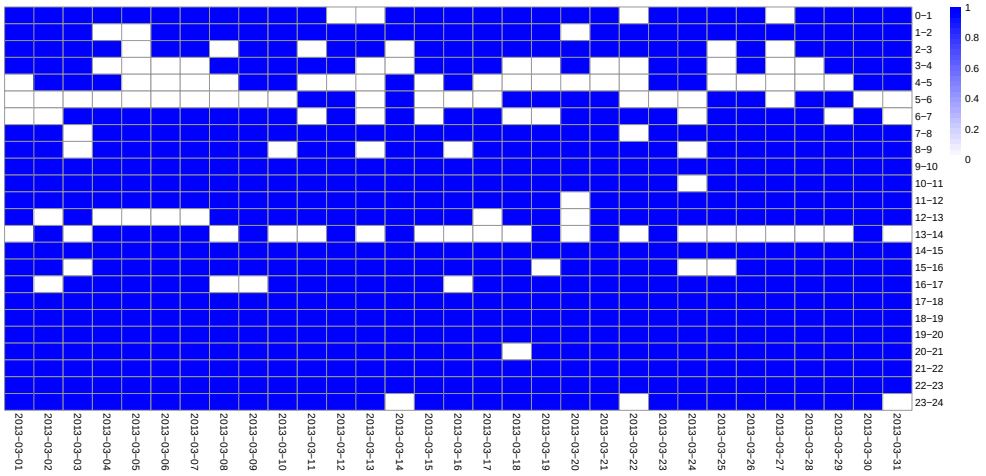
Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing*, 5, 121-125.

Rossi, P.E., Allenby, G., and McCulloch, R., (2005). Bayesian Statistics and Marketing. John Wiley & Sons, NJ.

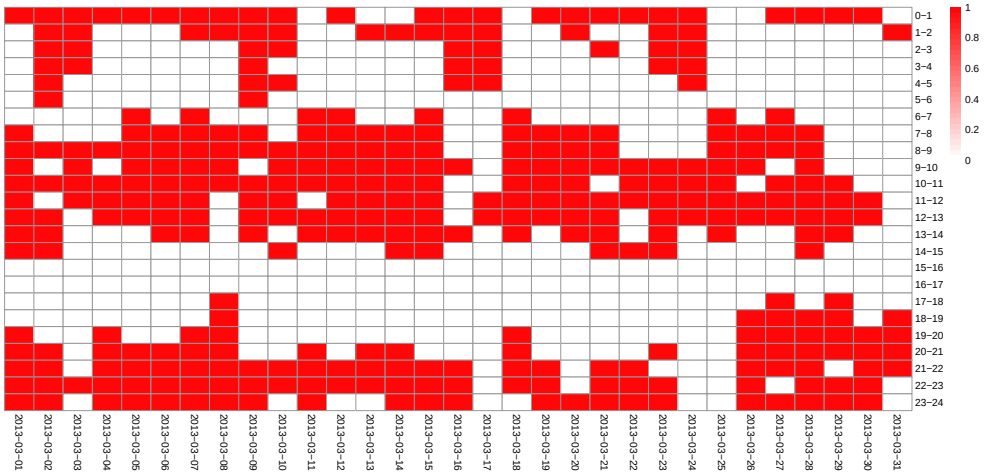
Stock, J. H., and Watson, M. W. (2002) Forecasting using principal components from a large number of observable factors. *Journal of the American Statistical Association*, 97, 1167–1179.

Sun, Y. (2016) Likelihood-based inference for nonlinear models with both individual and time effects. Discussion paper series, KU Leuven Department of Economics.

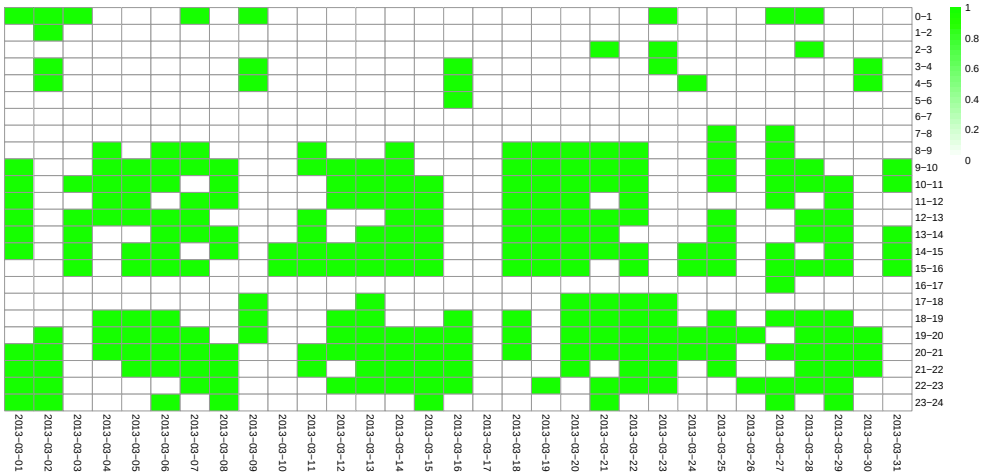
Van Der Vaart A. W., and Wellner J. A. (1996). Weak convergence and empirical processes with applications to statistics. Springer.



Medallion taxi A.



Medallion taxi B.



Medallion taxi C.

Figure 1: Examples of performance achievements from 3 medallion taxis in March 2013. Horizontal axis; date, vertical axis; hours. A colored cell indicates that the corresponding medallion taxi achieved the prespecified level of performance.

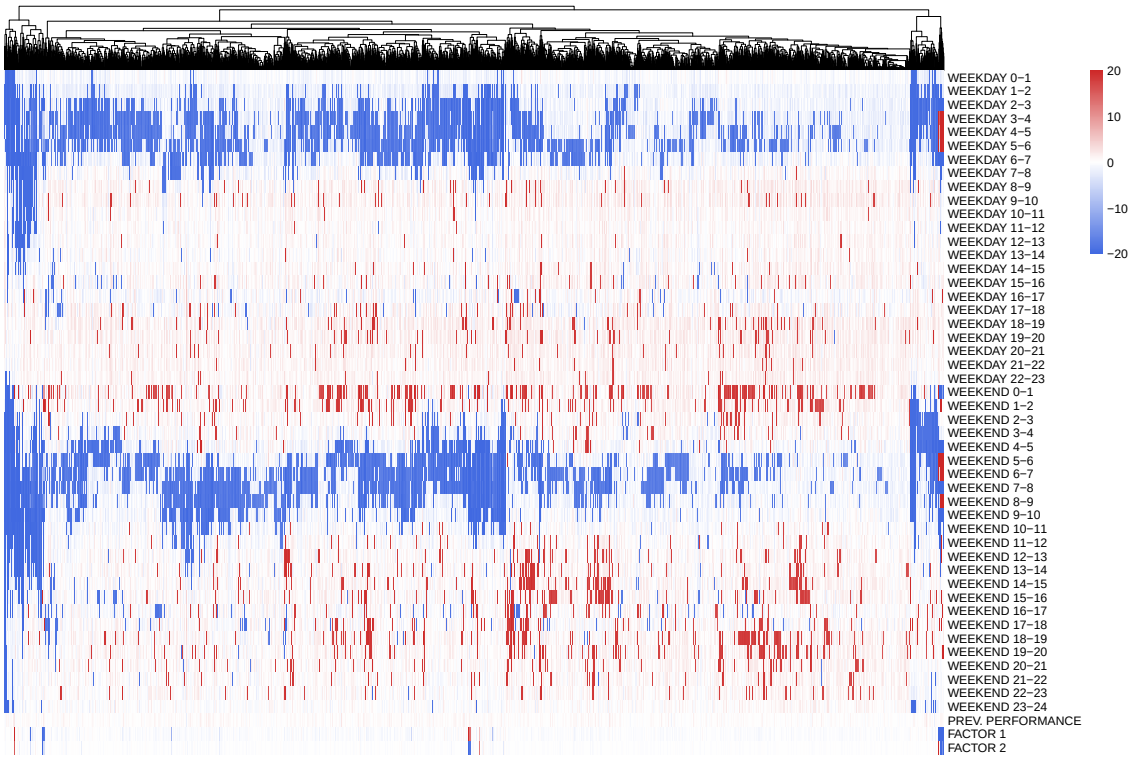


Figure 2: Hierarchical clustering results. Each column corresponds to the individual's sensitivity to the predictors and to the common factors. The trees on the top are from the hierarchical clustering procedure.

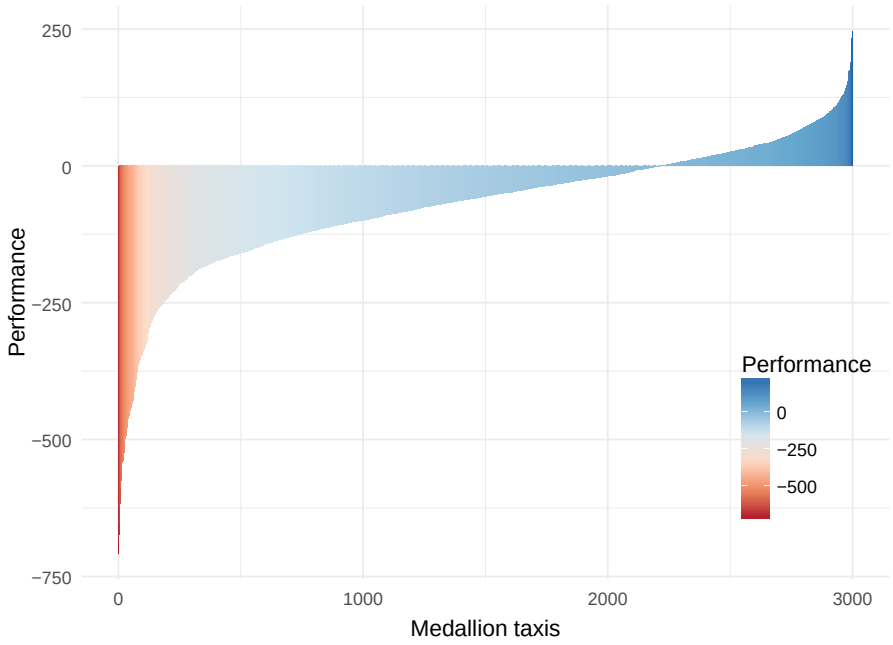
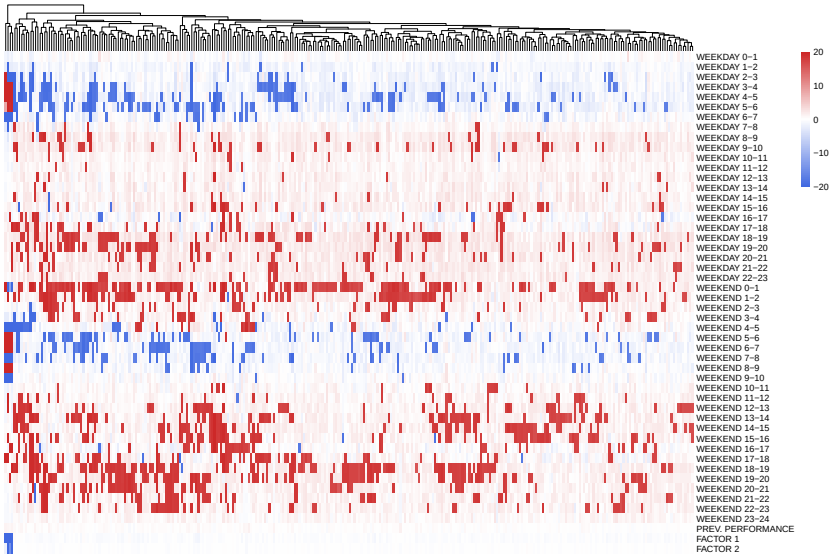
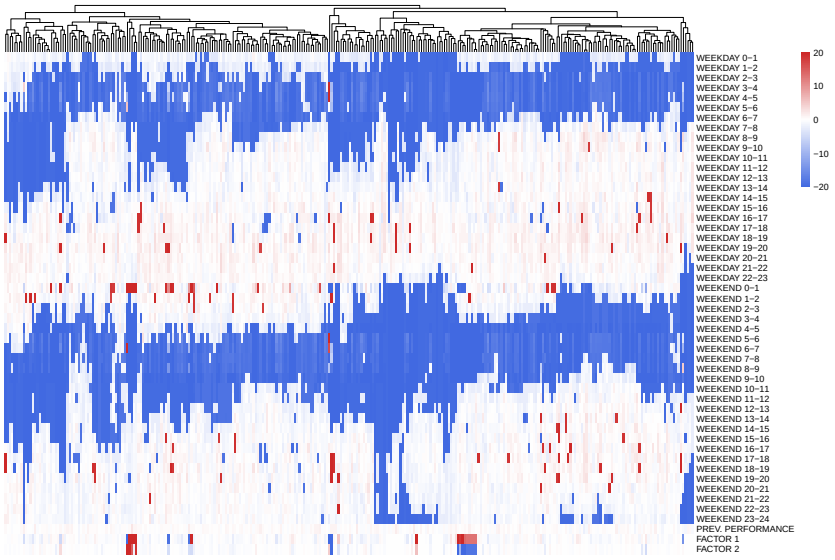


Figure 3: Overall performance of each taxi based on $\boldsymbol{p}_{\text{overall}}$.



(a) Top 10%



(b) Bottom 10%

Figure 4: Hierarchical clustering results of the estimated parameters for the top 10% and the bottom 10% taxis. Each column corresponds to the individual's sensitivity to the predictors and to the common factors. The trees on the top are from the hierarchical clustering procedure.

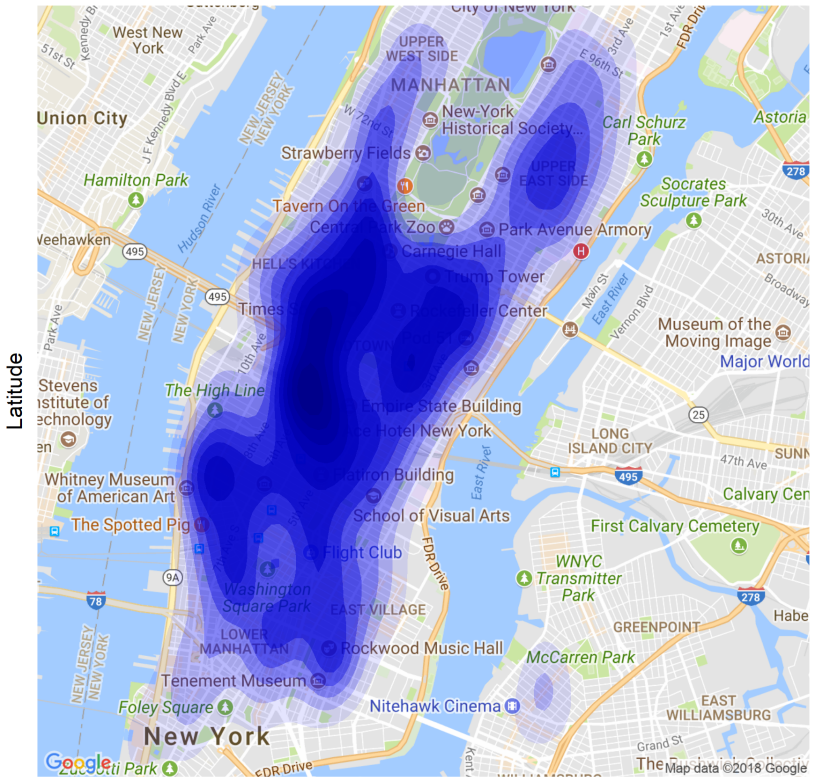


Figure 5: Densities of pickup locations for the top 10% taxis after 5:00 PM until 8:00 AM the next morning. Darker points correspond to points of higher density.

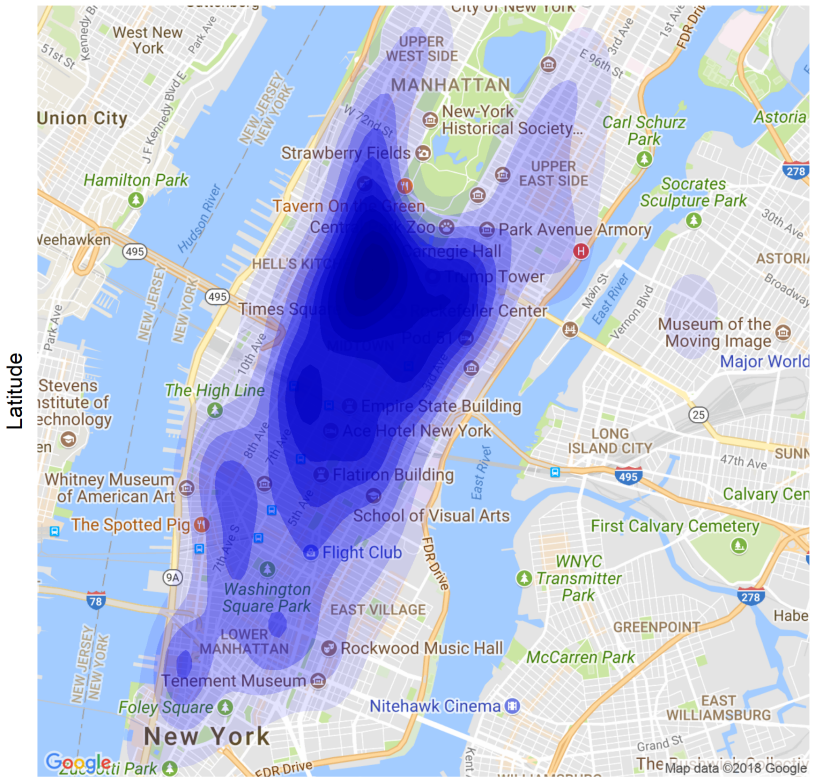


Figure 6: Densities of pickup locations for the bottom 10% taxis after 5:00 PM until 8:00 AM the next morning. Darker points correspond to points of higher density.

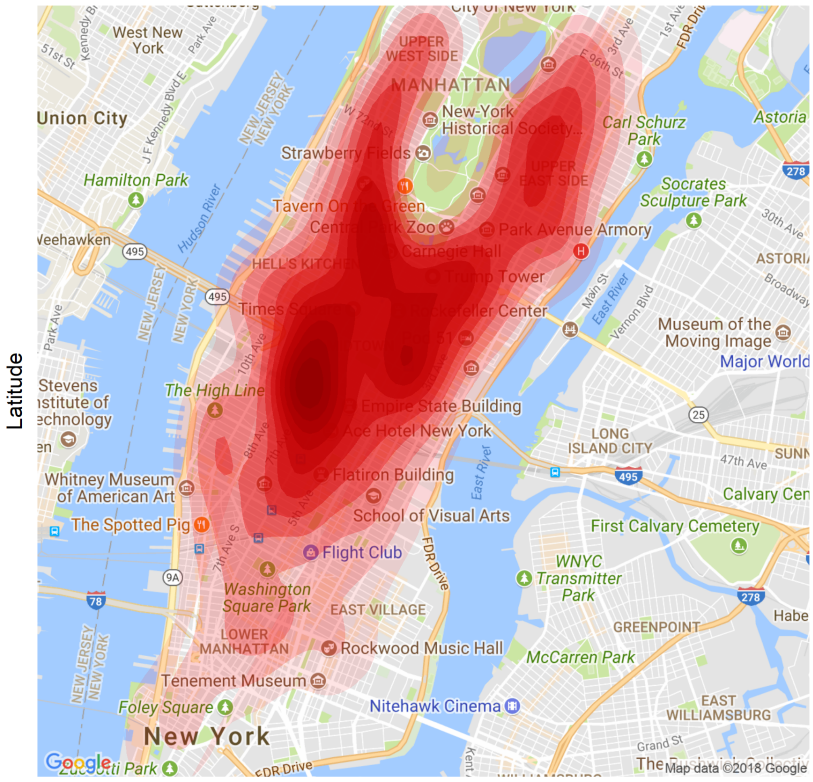


Figure 7: Densities of pickup locations for the top 10% taxis between 8:00 AM and 5:00 PM. Darker points correspond to points of higher density.

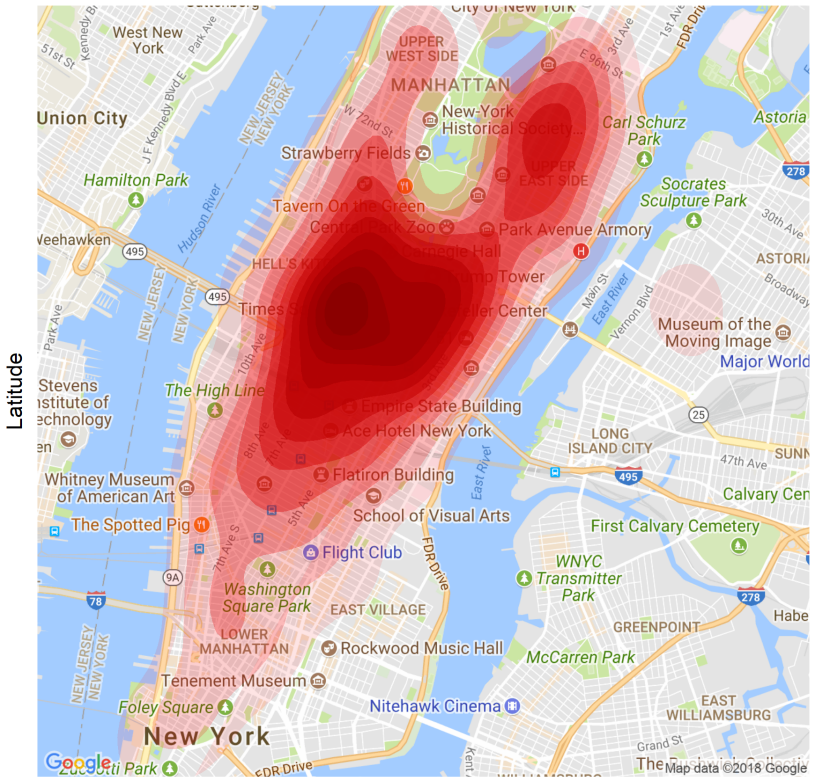


Figure 8: Densities of pickup locations for the bottom 10% taxis between 8:00 AM and 5:00 PM. Darker points correspond to points of higher density.