Large scale panel logistic regressions with unobservable effects: an application to demand analysis for restaurant dining services

November 9, 2016

Tomohiro Ando [1] and Jushan Bai [2]

Abstract

We introduce a logistic panel regression model with interactive fixed effects, which allow multiple individual effects and are capable of capturing high-dimensional cross-sectional/serial dependence. The direct likelihood maximization is challenging because of the nonlinearity in both the functional form and the interactive effects. In this paper, we propose a new data-augmentation strategy for the estimation. Our data-augmentation strategy provides simple inference methods. Theoretical properties are also established for the proposed procedure. Due to the presence of the interactive fixed effects, the diverging dimension of the panel size, and the nonlinear nature of the model, the novel asymptotic theories are developed by handling these technical difficulties.

There is a rich opportunity to apply the proposed method. This paper analyzes a large panel data set that consists of individuals' historical consumptions of restaurant dining services. In restaurant management, it is important to match demand and service capacity, and accurate demand forecasting is crucial in this regard. Our result indicates that the treatment of unobservable heterogeneity is helpful for demand forecasting.

Keywords: Endogeneity, factor analysis, cross-sectional dependence, serial dependence.

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

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

1 Introduction

In restaurant management, one of the key managerial aspects that determines operational efficiency is to match demand and service capacity. This matching process forces restaurant managers to consider, among other factors, how much food inventory should be purchased and what predefined level of staffing will be needed in a given time period. It is obvious that for both decisions, food inventory/staffing levels should be reasonable such that the restaurant can meet the actual demand, which cannot be observed in advance. To approach these problems, it is beneficial to have an accurate forecasting system that allows the manager to efficiently capitalize on both the existing and future demand. This is because the elimination of food waste and the optimization of payroll improves the efficiency of restaurant management. Insufficient food inventory/human resources forces the restaurant to miss an opportunity to serve a menu item. By contrast, the excessive food inventory/human resources create food waste and leaves servers standing idle. An accurate forecasting system helps managers to ensure that waste is kept to a minimum, that inventory is always available, and that an appropriate level of staffing can be brought in at any given time of any given day.

With the advancement of information technology and data-gathering techniques, it has become possible for restaurants to accumulate vast amounts of data on individual consumer choices over time. In this paper, we analyze the daily traffic to the residential dining cafeterias located at Los Angeles, California. Dining offers a multi-tiered, variable prepaid meal plan, and individual customers (in this paper, students living in the campus dorms) sign up for one of of these options if the plan is attractive. The enrolled customer swipes his or her customer ID card every time when eating at a restaurant. Unlike the Point of Sale system employed in a typical restaurant, Dining is able to collect an individual customer's consumption data through the meal plan accounting system. Specifically, for each of the customers, the system records whether the customer ate at a restaurant during five meal periods (Breakfast, Lunch, Dinner, Before Mid., After Mid.) every day. From a managerial perspective, it is crucial to understand customers' eating behaviors for each of 5 periods every day. Demand forecasting plays an important role in maintaining an appropriate level of food inventory, thus allowing the dining service to reduce the spoilage rate and have ingredients regularly available. Dining also needs to make a labor plan one week in advance. Thus, Dining needs to forecast future unseen demand in advance. The substantial statistics literature on demand analysis includes studies by Andrews et al. (2011), Calli and Weverbergh (2009), Fildes and Kumar (2002), Harvey and Koopman (1993), Hanssens (1998), Shively and Sager (2009),

Song and Li (2008), and Syntetos et al. (2009).

To forecast whether a customer will eat at a dining hall, discrete choice models are a natural starting point (McFadden, 1981). Discrete choice models are widely used in social science studies. Under the assumption that Dining has information on all covariates that explain customer behavior and that customers make their decisions independently, this approach seems appropriate, as a customer can decide whether to eat at a Dining facility. However, it is unlikely that customers make their decisions independently, as they often have a meal with their friends. For example, a customer group may have a study plan that allows members to have a dinner in a particular week, where they usually forego eating dinner at the dining hall and eat off campus. This means that a standard panel logistic model that maintains the data independence assumption is unlikely to be satisfied. Moreover, the information acquisition cost faced by Dining in collecting all variables that might explain the customers' eating behavior is extremely high.

To address these problems, this paper introduces panel logistic regression models with interactive fixed effects to analyze choices made by individuals among a set of alternatives, where both the cross-sectional and time-series dimensions of the panel are large. Recently, there is a growing literature on panel data models with unobserved factor structures, where both the cross-sectional and time-series dimensions of the panel are large (Ando and Bai (2015a, 2015b, 2016), Bai and Ng (2002), Bai (2009), Bai and Li (2014), Chudik and Pesaran (2015), Fan et al. (2011), Moon and Weidner (2015), Pesaran (2006), and Song (2013), among others). Although a number of studies exist on linear panel data regression models with interactive effects, studies on nonlinear panel models with unobserved factor structures are scant.

While conducting inference for linear panel data regressions with interactive effects is relatively easy, inference in panel logistic regression with interactive effects is a challenging problem due to the analytically inconvenient form of the model structure. In the linear case, one can employ various estimation algorithms, including Bai (2009) for homogeneous panels, Ando and Bai (2015b) for heterogeneous panels with a large number of explanatory variables, and so forth. These algorithms for linear panel data regression models with interactive effects cannot be applied to the panel data models considered here. This is largely because of the nonlinearity in the interactive effects and the nonlinearity in the probabilistic structure. Thus, in contrast to inference in linear panel data regression, obtaining frequentist inference in the proposed model is quite challenging.

In this paper, we present a new data-augmentation algorithm for panel logistic regression models with interactive fixed effects. In the Bayesian statistics litera-

ture on cross-sectional logistic regression models (not panel data models), a data-augmentation strategy is often employed (e.g., Holmes and Held (2006), Fruhwirth-Schnatter and Fruhwirth (2007), Polson and Scott (2013)). However, previous studies have ignored the issue of "endogeneity", meaning that the set of regressors is correlated with the error terms. We show that our data-augmentation strategy will greatly simplify inference regarding the interactive fixed effects. Although our inference is conducted by Markov chain Monte Carlo, it is easy to implement. This is the first study to investigate a data-augmentation approach to the analysis of panel logistic regression models with endogeneity. As shown in the simulation study, failing to account for endogeneity increases the bias of the estimation. Our simulation study also indicates that our proposed procedure improves the estimation accuracy of the regression coefficients in the presence of interactive fixed effects.

In addition to the new computational approach, this paper also develops the asymptotic theory of our proposed methods. Consistency and convergence rate of the estimator are established. Moreover, a novel approach for selecting the dimension of the interactive fixed effects is established. Previous studies, including Bai and Ng (2002) Hallin and Liška (2007, 2011) can not be applied directly due to the analytically inconvenient form of the model structure. Because both time series dimension and individual dimension are diverging, determining the dimension of the interactive effects is a complicated task. Monte Carlo study indicates that the proposed approach is capable of selecting the true dimension of interactive fixed effects. Furthermore, we show that our method can be easily extended to the inference for panel multinomial logistic regression, panel logistic regression with shrinkage approach, and so forth.

The paper is organized as follows. Section 2 introduces the panel logistic regression model with interactive fixed effects. We then propose the data-augmentation approach for the model inference in Section 3. Section 4 establishes new asymptotic results, including the consistency of the proposed estimator and its asymptotic behaviors and the model selection criterion for determining the dimension of interactive fixed effects. we establish a number of theoretical results. All proofs and additional theoretical results are provided in the Appendix. Appendix also contains Monte Carlo simulations results that demonstrate that the proposed method works well. In, Section 5, we also illustrate the proposed procedure through an empirical analysis of daily meal demand in a dining restaurant chain. The data, background information and the detailed data analysis are given. Section 6 provides further investigations. Section 7 concludes.

2 Panel logistic regression model with interactive fixed effects

2.1 Demand 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, labeled alternative 1 and alternative 0. We consider the random utility (the difference in utilities between alternative 1 and alternative 0) associated with the choice for individual i at time t:

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

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

As one of the novelties of this paper, 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. This is known as the interactive effect in the econometric literature (e.g., Bai, 2009). Note that the interactive effects are more general than conventional additive effects. To see this, suppose that there are two factors ($r = 2$), and consider the special factor $f_t = (1, \delta_t)'$ and the special loading $\lambda_i = (\alpha_i, 1)'$. Then $f_t'\lambda_i = \alpha_i + \delta_t$, reducing to the standard individual effect and time effect model (additive effects). In additive effects models, the influence of individual effects (α_i) is constant over time, and the influence of time effects (δ_t) is identical across individuals. In contrast, the interactive effects allow the unobserved individual characteristics (λ_i) to have time-varying effects (through f_t). Another interpretation of the interactive effects is that they allow a vector of common shocks or social trends (f_t) to impact individuals in a heterogeneous way (through λ_i). If $r = 1$, the model then reduces to the panel logistic choice model studied by Chen et al. (2014). We allow for correlation between the factor structure η_{it} and the regressors $\boldsymbol{x}_{it}$ (endogeneity), while the standard logistic regression model does not permit such a situation. Ignoring endogeneity, if it exists, will lead to bias and inconsistent estimation.

Conventional panel data analysis often assumes 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 in modeling high-dimensional cross-sectional dependence.

Let $y_{it} \in \{0, 1\}$ denote the observed choice outcome, taking value 1 if alternative 1 is chosen, 0 otherwise. Alternative 1 will be chosen if and only of $u_{it} > 0$ (the difference in utility is positive). For the logistic specification of the idiosyncratic shock ε_{it}, the conditional probability of such a choice is given by

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

Assuming that the errors ε_{it} are independently and identically distributed, the log of the joint probability of observing the choices $Y \equiv \{y_{it}|i = 1, ..., N, t = 1, ..., T\}$, $L(Y|X, B, F, \Lambda)$, 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 $F = (\boldsymbol{f}_1, ..., \boldsymbol{f}_T)'$.

An important feature of the model is that correlations between the unobserved factor structure η_{it} and the regressors $\boldsymbol{x}_{it}$ are allowed. This correlation arises 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. The endogeneity will cause inconsistent estimation of the traditional maximum likelihood estimator.

In contrast to a linear panel regression model with interactive fixed effects, the likelihood function is nonlinear in the factor structure. This may be one of the reasons that the panel regression model with interactive fixed effects has not been widely studied and applied. Several related studies include Fernandez-Val and Weidner (2015) and Chen et al (2014). Recently, Chen et al. (2014) also consider inference in panel logistic regression models with predetermined explanatory variables and interactive effects. They consider the interactive fixed effects with a single factor and impose homogeneous regression coefficients. Our paper allows multiple factors and heterogenous regression coefficients. We further consider inference in multinomial panel choice models.

Remark 1 Li and Ansari (2014) also considered latent factor structures for choice modeling. In (1), we model the interactive effects η_{it}, which is unobservable, depends the individuals i and time t. In contract, Li and Ansari (2014) modeled the unobserved structure that depends on unobservable choice attributes and time t.

Letting the interactive effects η_{it} in (1) being common over i, our model then reduces to their unobserved structure. Naik et al. (2010) considered the multi-index binary response (MBR) model. They pointed out that their model is related to approximate factor models (e.g., Bai and Ng 2002) based on the fact that both involve linear combinations of the predictors. As well as their estimation procedure, our model estimation procedure explicitly incorporate the information in a choice data in the process of constructing common factors $\boldsymbol{f}_t$ and loadings $\boldsymbol{\lambda}_i$. In contrast to the MBR model, however, our interactive effects η_{it} is not restricted only on the space of linear combinations of the predictors. Thus, our model includes the model formulations of Li and Ansari (2014) and Naik et al. (2010). In the context of the linear instrumental variable regression where endogeneity are likely to be present, Ebbes et al. (2005) introduced the latent instrumental variables method. Similar to Ebbes et al. (2005), our method does not rely on observable instruments. This can be achieved by taking account for dependencies between the predictors $\boldsymbol{x}_{it}$ and the interactive effects η_{it}.

2.2 Assumption

We first define some notations. Let $\|A\| = [tr(A'A)]^{1/2}$ be the usual norm of the matrix A, where "tr" denotes the trace of a square matrix. The equation $a_n = O(b_n)$ states that the deterministic sequence a_n is at most of order b_n; $c_n = O_p(d_n)$ states that the random variable c_n is at most of order d_n in terms of probability, and $c_n = o_p(d_n)$ is of a smaller order in terms of probability. Because the dimensions of B, Λ and F are diverging, we cannot assume the standard regularity conditions for likelihood functions. The set of regularity conditions that are imposed on the proposed model are as follows:

Assumption A1: Common factors

The common factors satisfy $E\|\boldsymbol{f}_t\|^4 < \infty$. Also $T^{-1}\sum_{t=1}^{T} \boldsymbol{f}_t \boldsymbol{f}_t' \to \Sigma_F$ as $T \to \infty$, where Σ_F is an $r \times r$ positive definite matrix.

Assumption A2: Factor loadings

The factor-loading matrix for the common factors $\Lambda = [\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N]'$ satisfies $E\|\boldsymbol{\lambda}_i^4\| < \infty$ and: $\|N^{-1}\Lambda'\Lambda - \Sigma_\Lambda\| \to \mathbf{0} \quad \text{as} \quad N \to \infty$, where Σ_Λ is an $r \times r$ positive definite matrix.

Assumption A3: Error terms

The ε_{it} are iid (over i and t) standard logistic random variables.

Assumption A4: First order derivative

Let $\boldsymbol{\theta} = (\text{vec}(B)', \text{vec}(F)', \text{vec}(\Lambda)')'$ and F satisfies $F'F/T = I_r$. For each (N, T), there exists a parameter value $\boldsymbol{\theta}_0 = (\text{vec}(B_0)', \text{vec}(F_0)', \text{vec}(\Lambda_0)')'$ that satisfies

$$E[\partial \log L(Y|X, B, F, \Lambda)/\partial \boldsymbol{\theta}|_{\theta=\theta_0}] = 0,$$

where the expectation is taken over the true density of y_{it}. Let $\boldsymbol{\theta}_1 = (\text{vec}(B)', \text{vec}(\Lambda)')'$ and $\boldsymbol{\theta}_2 = \text{vec}(F)$. Then, there exist the parameter values $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^*$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_2^*$ that satisfy

$$\begin{aligned} E[\partial L(Y|X, B, F, \Lambda)/\partial \boldsymbol{\theta}_1|_{\theta_1=\theta_1^*}] = 0, \quad &\text{under a fixed } \boldsymbol{\theta}_2, \\ E[\partial L(Y|X, B, F, \Lambda)/\partial \boldsymbol{\theta}_2|_{\theta_2=\theta_2^*}] = 0, \quad &\text{under a fixed } \boldsymbol{\theta}_1. \end{aligned}$$

Assumption A5: Hessian matrix

For some positive constants C_1 and C_2, the expectation of the second derivative of the log-likelihood function

$$I_{NT}(\boldsymbol{\theta}) = -\frac{1}{NT} E[\partial \log L(Y|X, B, F, \Lambda)/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}']$$

satisfies

$$0 < C_1 < \lambda_{\min}(I_{NT}(\boldsymbol{\theta})) < \lambda_{\max}(I_{NT}(\boldsymbol{\theta})) < C_2 < \infty,$$

where $\lambda_{\min}(I_{NT}(\boldsymbol{\theta}))$ and $\lambda_{\max}(I_{NT}(\boldsymbol{\theta}))$ are the minimum and maximum eigenvalues of $I_{NT}(\boldsymbol{\theta})$. Also, each element of $(NT)^{-1} E[\{\partial^2 \log L(Y|X, B, F, \Lambda)/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'\}^2]$ is bounded by some positive constant C_3 with $0 < C_3 < \infty$.

Assumption A6: Third order derivative

There is a large enough open subset Θ which contains the parameter vector $\boldsymbol{\theta}_0$, such that there exists the third derivative of log density of Y, $\frac{1}{NT} \partial^3 \log L(Y|X, B, F, \Lambda)/\partial \theta_{k_1} \partial \theta_{k_2} \partial \theta_{k_3}$ for $\boldsymbol{\theta} \in \Theta$. Here k_1, k_2, k_3 range from 1 to $\dim(\boldsymbol{\theta})$. Furthermore, its absolute value is bounded by a function of Y, $C(Y)$, which satisfies

$$E[C^2(Y)] < C_4.$$

for some positive constant $C_4 < \infty$.

Remark 2 The full rank assumption of Σ_F and Σ_Λ in Assumptions A1 and A2 is necessary for the number of common factors to be r. Assumption A3 may be relaxed to allows cross-sectional and serial correlations and heteroskedasticities in

the idiosyncratic errors ε_{it}. However, exploring the weaker condition is beyond the scope of this paper. Assumptions A4–A6 impose the moments of the likelihood function. Note that $F\boldsymbol{\lambda}_i = (FH')(H\boldsymbol{\lambda}_i)$ for an orthogonal matrix such that $H'H = I_r$. Thus, Assumption A4 contains the expression F_0H as F_0 and $H\boldsymbol{\lambda}_i^0$ as $\boldsymbol{\lambda}_i^0$. The expectation of the second derivative of the likelihood function is assumed to be positive definite with bounded eigenvalues.

In the next section, we introduce the data-augmentation strategy and propose the new inference algorithm.

3 Data-augmentation approach for parameter inference

Inference on interactive fixed effects is a challenging problem. In this paper, we employ the Markov chain Monte Carlo approach and generate a set of posterior samples. Our computation approach is very attractive because the common factor structure can be easily investigated conditional on the individual effects and vice versa.

For the data-augmentation approach, we need to specify the prior distribution 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 will be

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

which does not provide analytical posterior density forms.

When we use the principal component framework (See, e.g., Bai (2009) and references therein), we usually analyze the unobservable common factor and its factor loadings jointly. Thus, the prior specification will take the form $\pi(B, F, \Lambda) = \pi(B)\pi(\Lambda, F)$. In this paper, in contrast, we analyze the regression coefficients and the factor loadings jointly. This treatment will provide a convenient data augmentation for inference on these unknown parameters. Moreover, one might conjecture that equation (1) allows us to easily derive the conditional posterior distributions of the interactive fixed-effect parameters (F, Λ). However, it does not lead to an easy method for sampling from their posterior distribution because the error term ε_{it} is not normal. In this section, we focus on the data-augmentation strategy in a binary choice setting. Some extensions will be discussed in Section 6.

3.1 Data-augmentation strategy for posterior sampling of the interactive fixed effects

We use the following equality (See Theorem 1 in Polson and Scott (2013))

$$\frac{\exp(\kappa)^a}{(1+\exp(\kappa))^b} = 2^{-b}\exp(\kappa z)\int_0^\infty \exp(-\omega\kappa^2/2)p(\omega)d\omega, \tag{5}$$

where $a \in R$, $b > 0$, $z = a - b/2$, and $p(\omega)$ is the density function of the Polya-Gamma distribution with parameters $(b, 0)$. Using this result, the likelihood contribution of observation y_{it} can be expressed as

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

where $z_{it} = y_{it} - 1/2$, and $p(\omega_{it})$ is the density of a Polya-Gamma random variable with parameters $(1, 0)$. Combining the terms from all observations yields the following expression for the conditional posterior of F, given $\Omega \equiv \{\omega_{it} | i = 1, ..., N, t = 1, ..., T\}$:

$$\begin{aligned}
\pi(F|Y, X, B, \Lambda, \Omega) & \propto \pi(F)\prod_{i=1}^N\prod_{t=1}^T \left[\exp\{z_{it}\{\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i\}\} \times \exp\{-\omega_i\{\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i\}^2/2\}\right] \\
& \propto \pi(F)\prod_{i=1}^N\prod_{t=1}^T \exp\left\{-\frac{\omega_{it}}{2}\{z_{it}/\omega_{it} - \boldsymbol{x}_{it}'\boldsymbol{b}_i - \boldsymbol{f}_t'\boldsymbol{\lambda}_i\}^2\right\} \\
& \propto \pi(F)\exp\left\{-\sum_{i=1}^N(\boldsymbol{z}_i^* - F\boldsymbol{\lambda}_i)'\Omega_i(\boldsymbol{z}_i^* - F\boldsymbol{\lambda}_i)\right\},
\end{aligned} \tag{6}$$

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

We further investigate the form of the conditional posterior of F: In this paper, the common factor F is subject to the normalization condition $F'F/T = I_r$ for identification purposes; see Remark 5 below. From $F'F/T = I_r$, F belongs to a hyperball in T dimensions, and its support is restricted to be the Cartesian product of the T-dimensional hyperball. Furthermore, because of the orthogonality requirement, its support is then reduced to a Stiefel manifold $S_{T,r}$ of radius $\sqrt{T}$ (Khatri and Mardia, 1977; Smidl and Quinn, 2007). The Stiefel manifold is employed in various studies (see, e.g., Strachan and Inderb (2004), Koop, et al. (2006), Hoff (2009), Tsay and Ando (2012)).

Therefore, the prior of F is a flat prior over the Stiefel manifold corresponding to orthogonal transformations and, hence, is invariant with respect to the orthogonal group. Specifically, the prior of F is

$$\pi(F) = \frac{1}{C(T,r)} \cdot 1(F \in S_{T,r}), \tag{7}$$

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 the normalizing constant with $\Gamma(\cdot)$ being the Gamma function.

However, under the prior $\pi(F)$ in (7), the analysis of the conditional posterior of F in (6) is still not straightforward. This is mainly because the diagonal matrix Ω_i prevents the derivation of an analytical conditional posterior of F. We therefore use the Metropolis-Hastings algorithm to generate the posterior sample of F. Given the posterior density $\pi(F|Y, X, B, \Lambda, \Omega)$, known up to a constant, and a proposal conditional density $p(F)$, we can generate the posterior sample of F in the following way.

To generate samples from $\pi(F|Y, X, B, \Lambda, \Omega)$, the Metropolis-Hastings algorithm requires us to specify a proposal density $p(F)$. The Metropolis-Hastings algorithm then first draws a candidate parameter value F^{new} from the proposal density $p(F)$. The generated parameter value F^{new} will be 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.

In the 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 a proposal density

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

where F is on the Stiefel manifold and $Z = (\boldsymbol{z}_1, ..., \boldsymbol{z}_N)$. In our simulation study, this proposal density works well.

3.2 Prior specification and posterior analysis on B and Λ

Here, we specify the prior densities on B and Λ and derive their conditional posterior distributions, given F and Ω. For simplicity of notation, we first express the

likelihood contribution of observation y_{it} as

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

where $\boldsymbol{v}_{it} = (\boldsymbol{x}_{it}', \boldsymbol{f}_t')'$, and $\boldsymbol{\gamma}_i = (\boldsymbol{b}_i', \boldsymbol{\lambda}_i')'$.

In the context of Bayesian (cross-sectional) linear regression, it is common to use the multivariate normal prior. In this case, one considers to use the normal prior on $\boldsymbol{\gamma}_i$ with mean $\mathbf{0}$ and variance covariance matrix A_{γ_i}: $\pi(\boldsymbol{\gamma}_i|A_{\gamma_i}) = N(\mathbf{0}, A_{\gamma_i})$. If one seeks to obtain the maximum likelihood estimator, the diffuse prior can be used for $\boldsymbol{\gamma}_i$. See Remark 3 for further details. Then, the conditional posterior density of $\boldsymbol{\gamma}_i = (\boldsymbol{b}_i', \boldsymbol{\lambda}_i')'$ is

$$\begin{aligned}
\pi(\boldsymbol{\gamma}_i|Y, X, B_{-i}, \Lambda_{-i}, \Omega) &\propto \pi(\boldsymbol{\gamma}_i|A_{\gamma_i}) \exp\{z_{it}\boldsymbol{v}_{it}'\boldsymbol{\gamma}_i - \omega_{it}\{\boldsymbol{v}_{it}'\boldsymbol{\gamma}_i\}^2/2\} \\
&\propto \exp\left\{-\frac{1}{2}\boldsymbol{\gamma}_i' A_{\gamma_i}^{-1}\boldsymbol{\gamma}_i\right\} \exp\left\{-\frac{1}{2}(\boldsymbol{z}_i - W_i\boldsymbol{\gamma}_i)'\Omega_i(\boldsymbol{z}_i - W_i\boldsymbol{\gamma}_i)\right\},
\end{aligned}$$

where $W_i = (X_i, F)$ is the design matrix, $B_{-i} = (\boldsymbol{b}_1, ..., \boldsymbol{b}_{i-1}, \boldsymbol{b}_{i+1}, ..., \boldsymbol{b}_N)'$ and $\Lambda_{-i} = (\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_{i-1}, \boldsymbol{\lambda}_{i+1}, ..., \boldsymbol{\lambda}_N)'$. This implies that the conditional posterior density of $\boldsymbol{\gamma}_i$ is the multivariate normal density with mean $(W_i'\Omega_i W_i + A_{\gamma_i}^{-1})^{-1} W_i'\boldsymbol{z}_i$ and variance-covariance matrix $(W_i'\Omega_i W_i + A_{\gamma_i}^{-1})^{-1}$.

In addition to the conditional posterior density of $\boldsymbol{\gamma}_i$, we can easily obtain the conditional posterior densities of ω_{it}, that is,

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

which is a Polya-Gamma distribution with parameter $(1, \boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)$. Again, we can easily draw a posterior sample of ω_{it} using the Gibbs sampler.

As discussed above, we can analytically obtain the conditional posterior distributions of B, Λ and Ω. Therefore, we easily draw the posterior samples by implementing the Gibbs sampling algorithm. To draw F, we can use the Metropolis-Hastings algorithm described in Section 3.1. We now summarize the posterior sampling procedure as follows.

Posterior sampling algorithm:

Step 1. Initialize the parameters.

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

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

Step 4. Sample ω_{it} from $\pi(\omega_{it}|Y, X, B, \Lambda, \Omega_{-\omega_{it}})$.

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

The outcomes of the above algorithm can be regarded as a random sample from the joint posterior density function after a burn-in period. We then obtain a set of H posterior samples $\{B^{(k)}, F^{(k)}, \Lambda^{(k)}; k = 1, ..., H\}$ for inference.

Remark 3 When we wish to obtain the maximum likelihood estimator, we simply use the diffuse prior $\pi(\boldsymbol{\gamma}_i) \propto$ Const.. Then, the conditional posterior density of $\boldsymbol{\gamma}_i = (\boldsymbol{b}_i', \boldsymbol{\lambda}_i')'$ is

$$\begin{aligned}\pi(\boldsymbol{\gamma}_i|Y, X, B_{-i}, \Lambda_{-i}, \Omega) &\propto \exp\{z_{it}\boldsymbol{v}_{it}'\boldsymbol{\gamma}_i - \omega_{it}\{\boldsymbol{v}_{it}'\boldsymbol{\gamma}_i\}^2/2\} \\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{z}_i - W_i\boldsymbol{\gamma}_i)'\Omega_i(\boldsymbol{z}_i - W_i\boldsymbol{\gamma}_i)\right\},\end{aligned}$$

which implies that the conditional posterior density of $\boldsymbol{\gamma}_i$ is the multivariate normal density with mean $(W_i'\Omega_i W_i)^{-1}W_i'\boldsymbol{z}_i$ and variance-covariance matrix $(W_i'\Omega_i W_i)^{-1}$.

The maximum likelihood estimator can be numerically obtained by the MCMC algorithm. Let $\{B^{(k)}, F^{(k)}, \Lambda^{(k)}; k = 1, ..., H\}$ be the set of H posterior samples. Then, maximum likelihood estimator, $\{\hat{B}, \hat{F}, \hat{\Lambda}\}$, is given as

$$\{\hat{B}, \hat{F}, \hat{\Lambda}\} = \text{argmax}_{\{B^{(k)}, F^{(k)}, \Lambda^{(k)}\}; k=1,...,H} L(Y|X, B^{(k)}, F^{(k)}, \Lambda^{(k)}), \tag{8}$$

where the likelihood function is given in (4). We have established the consistency of the maximum likelihood estimator. Details are given in Section 4.

Remark 4 In Step 1, starting values are needed. The initial parameter values are set as follows. First, we set $B^{initial} = (\boldsymbol{b}_1^{initial}, ..., \boldsymbol{b}_N^{initial})'$ as the maximizer of

$$L(Y|X, B) = \prod_{i=1}^{N}\prod_{t=1}^{T}\left[\frac{\exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i)}{1+\exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i)}\right]^{y_{it}}\left[\frac{1}{1+\exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i)}\right]^{1-y_{it}}, \tag{9}$$

where $B = (\boldsymbol{b}_1, ..., \boldsymbol{b}_N)'$. Then, the initial value of F, $F^{initial}$, is set as the first r eigenvectors of the following matrix: $\sum_{i=1}^{N}(\boldsymbol{z}_i - X_i\boldsymbol{b}_i^{initial})(\boldsymbol{z}_i - X_i\boldsymbol{b}_i^{initial})'$ where $\boldsymbol{z}_i = (z_{i1}, ..., z_{iT})$ with $z_{it} = (y_{it} - 1/2)$. Then, the initial value of Λ is set as $\boldsymbol{\lambda}_i^{initial} = F^{initial\prime}(\boldsymbol{z}_i - X_i\boldsymbol{b}_i^{initial})/T$. Finally, the initial value of ω_{it} is prepared by generating a random variable from a Polya-Gamma distribution with parameter $(1, \boldsymbol{x}_{it}'\boldsymbol{b}_i^{initial} + \boldsymbol{f}_t^{initial\prime}\boldsymbol{\lambda}_i^{initial})$.

Remark 5 If one wishes to obtain the Bayesian version of the principal component estimator (for the principal component estimator for the linear panel data, see Connor and Korajzcyk (1986), Bai and Ng (2002) and Stock and Watson (2002)), we impose a set of conditions under which we can identity the columns of F and the columns of Λ from the product $F\Lambda'$. Because $F\Lambda' = FRR^{-1}\Lambda'$ for any $r \times r$

invertible matrix R and R has r^2 free parameters, we need at least r^2 restrictions to identity F and Λ. Similar to Bai and Ng (2013), we consider the restrictions $F'F/T = I_r$ and $\Lambda' = (\Lambda_1', \Lambda_2')'$, with Λ_1 being an invertible lower triangular matrix. As discussed in Bai and Ng (2013), this restriction will lead to the identification of F and Λ. Given the posterior sample F and Λ, it is easy to obtain estimators satisfying the restriction. Note that the product $F\Lambda'$ remains the same even when F and Λ are rotated (See Bai and Ng (2013)). In Step 1, we obtain a QR decomposition of Λ to yield $\Lambda' = QR$, where R is an upper triangular matrix with positive diagonal elements, and Q is an $r \times r$ orthogonal matrix such that $Q'Q = I_r$. This decomposition is unique for any invertible Λ. In Step 2, we define

$$\tilde{F} = FQ \quad \text{and} \quad \tilde{\Lambda} = \Lambda Q = (R, \Lambda_2')'.$$

Note that $\tilde{F}'\tilde{F}/T = Q'(F'F/T)Q = Q'Q = I_r$.

4 Theoretical Analysis

There is a rich opportunity to apply the proposed method, but theoretical results are lacking in the literature. As a theoretical justification of the proposed method, this section first provides the results of the consistency of the maximum likelihood estimator $\hat{F}$, $\hat{B}$ and $\hat{\Lambda}$. Here the true parameter value $\{F_0, B_0, \Lambda_0\}$ is defined in Assumption A4. This true parameter is equivalent to $\{F, B, \Lambda\}$ that minimizes the Kullback–Leibler distance

$$\int \log\{f(Y|X, B_0, F_0, \Lambda_0)/f(Y|X, B, F, \Lambda)\} f(Y|X, B_0, F_0, \Lambda_0) dY$$

subject to the constraint $F'F/T = I$. We have the following theorem.

Theorem 1 *Under Assumptions A1–A6, $\sqrt{N}/T \to 0$ and $\sqrt{T}/N \to 0$ as $N, T \to \infty$, the maximum likelihood estimators $\hat{F}$, $\hat{B}$ and $\hat{\Lambda}$ are the consistent estimators for their true values in the sense that*

$$T^{-1} \sum_{t=1}^{T} \|\hat{\boldsymbol{f}}_t - \boldsymbol{f}_t^0\|^2 = o_p(1),$$
$$N^{-1} \sum_{i=1}^{N} \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^0\|^2 = o_p(1),$$

where $\boldsymbol{\gamma}_i^0 = (\boldsymbol{b}_i^0, \boldsymbol{\lambda}_i^0)$.

Remark 6 Proof of Theorem 1 is given in the supplementary document. The convergence rate of both $T^{-1}\sum_{t=1}^{T}\|\hat{\boldsymbol{f}}_t-\boldsymbol{f}_t^0\|^2$ and $N^{-1}\sum_{i=1}^{N}\|\hat{\boldsymbol{\gamma}}_i-\boldsymbol{\gamma}_i^0\|^2$ are $O_p(1/\min\{T,N\}) = o_p(1)$. Note that the estimator $\hat{F}$, $\hat{B}$ and $\hat{\Lambda}$ are obtained under the constraint $F'F/T = I$. Moreover, the dimensions of the panel size N and T are diverging. Furthermore, these parameters are connected nonlinearly with each other. Therefore, a novel proof is developed by handling these technical difficulties.

In practice, we have to determine the dimension of the interactive effects, or equivalently, the dimension of F. In the literature, several methods are proposed to select the number of factors, e.g., Bai and Ng (2002), Amengual and Watson (2007), Hallin and Liška (2007), and Lam and Yao (2012). However, these methods are fundamentally constructed for linear factor models, not nonlinear models.

One may consider to apply the cross-validation. However, as described in Ando and Bai (2015a), it is not easy to apply cross-validation given the factor structure. When we consider the leave-one-individual-out cross-validation, we estimate the regression coefficients and the factor structures. However, it is not possible to obtain the factor loadings of the deleted individual as the factor loadings are individual dependent. Instead, consider to estimate the model based on the information observed up to time $t-1$, and then forecast the responses of each unit at time t. To evaluate the predictive ability of the estimated model, we compare the difference between the actual realization of responses at t and their forecasts. However, the factor structure at time t is not available. Thus, the pure cross-validation procedure is not easy to apply directly.

Because both N and T are diverging and nonlinear nature of the model, a direct development of analytical model selection criterion is a challenging task. In this paper, we combine the cross-validation approach and an analytical model selection approach. Let N_v be the number of individuals for the k-th validation panel Y_k, and $N-N_v$ be the number of remaining individuals $Y_{(-k)}$. Thus, Y can be expressed as $Y = (Y_1, Y_{(-1)})$, or, $Y = (Y_1, Y_2, ..., Y_M)$ with $M = N/N_v$. (If N/N_v is not an integer, we can set $M = [N/N_v]+1$ with $[a]$ is the maximum integer that does not exceed a). Also, let $B_{(-k)}$ and $\Lambda_{(-k)}$ be the regression coefficients and factor loading matrices for $N-N_v$ individuals $Y_{(-k)}$. Similarly let B_k and Λ_k be the regression coefficients and factor loading matrices for N_v individuals Y_k. The common factors F are estimated based on the information for $N-N_v$ individuals. We obtain the estimated common factors $\hat{F}_{Y_{(-k)}} = (\hat{\boldsymbol{f}}_{1(-k)}, ..., \hat{\boldsymbol{f}}_{T(-k)})'$ which is given by

$$\{\hat{F}_{Y_{(-k)}}, \hat{B}_{Y_{(-k)}}, \hat{\Lambda}_{Y_{(-k)}}\}$$

$$= \text{argmax}_{F,B_{Y_{(-k)}},\Lambda_{Y_{(-k)}}} \sum_{i\in Y_{(-k)}} \sum_{t=1}^{T} \left[y_{it}\{\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i\} - \log\{1+\exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i)\}\right],$$

where $\sum_{i\in Y_{(-k)}}$ is the sum with respect to the $N - N_v$ individual for the estimation. The estimator $\{\hat{F}_{Y_{(-k)}}, \hat{B}_{Y_{(-k)}}, \hat{\Lambda}_{Y_{(-k)}}\}$ can be obtained by using the data-augmentation strategy given in Section 3. Note that N_v individuals are not used for estimating the common factors F.

Replacing the unknown common factors by the estimated common factors $\hat{F}_{Y_{(-k)}}$, the log-likelihood function for N_v validation individuals Y_k, $\log L(Y_k|X_k, B_k, \hat{F}_{Y_{(-k)}}, \Lambda_k)$ becomes $\log L(Y_k|X_k, B_k, \hat{F}_{Y_{(-k)}}, \Lambda_k) = \sum_{i\in Y_k}\sum_{t=1}^{T}[y_{it}\{\boldsymbol{x}_{it}'\boldsymbol{b}_i + \hat{\boldsymbol{f}}_{t(-k)}'\boldsymbol{\lambda}_i\} - \log\{1 + \exp(\boldsymbol{x}_{it}'\boldsymbol{b}_i + \hat{\boldsymbol{f}}_{t(-k)}'\boldsymbol{\lambda}_i)\}]$, where B_k and Λ_k are the regression coefficients and factor loading matrices for N_v individuals Y_k and thus the sub-matrices of B and Λ. Let

$$\{\hat{B}_k, \hat{\Lambda}_k\} = \text{argmax}_{B_k,\Lambda_k} L(Y_k|X_k, B_k, \hat{F}_{Y_{(-k)}}, \Lambda_k).$$

Then, our model evaluation score based on the N_v validation individuals, under the number of common factors r, is

$$\begin{aligned} IC_k(r) &= -2\log L(Y_k|X_k, \hat{B}_k, \hat{F}_{Y_{(-k)}}, \hat{\Lambda}_k) \\ &\quad + \log\left(\log\left(\frac{N_v T}{N_v + T}\right)\right) \max\{N_v, T/M\} \times r, \end{aligned} \tag{10}$$

where $\hat{B}_k$ and $\hat{\Lambda}_k$ are the maximizers of $L(Y_k|X_k, B_k, \hat{F}_{Y_{(-k)}}, \Lambda_k)$. Repeating this process M times over all validation sample Y_k $(k = 1, ..., M)$, we obtain the the model selection criterion;

$$IC(r) = \sum_{k=1}^{M} IC_k(r). \tag{11}$$

The number of common factors is selected by minimizing the $IC(r)$ score (11). In below, we provide a theoretical justification for this model selection criterion.

Theorem 2 *Suppose that Assumptions A1–A6 hold, and* $\sqrt{N}/T \to 0$ *and* $\sqrt{T}/N \to 0$ *as* $N, T \to \infty$. *If* $M = O(1)$, *the information criterion* $IC(r)$ *provides a consistent model selector of the true dimension of the interactive effects (the true number of common factors)* r_0.

Remark 7 Proof of Theorem 2 is given in the supplementary document, Appendix B. When one considers the one-individual-out cross validation (i.e., $N_v = 1$), then the penalty term on the dimension of the interactive effects in (10) may not be strong enough. Thus, there is a positive probability that $IC(r)$ may select the larger

dimension of interactive effects than r_0. Details are given in the supplementary document, Appendix B (About Remark 7). By choosing $N_v = O(N)$, we proved the consistency of model selection. There is also computational advantage for $N_v = O(N)$.

5 Application

5.1 Data and background information

In this paper, we analyze the daily demand of a dining restaurant at Los Angeles, California. Dining operates several clustered restaurants in the sense that customers can easily walk from one location to another. Dining offers several meal plans and collects data via their meal plan accounting system. Customers often sign up for one of the following meal plans: the premium 19 meal plan, the premium 14 meal plan, the regular 11 meal plan, the regular 14 meal plan, and the regular 19 meal plan. We study the daily demand data of customers living in dorms who enrolled in these plans. Under the regular 19 meal plan, customers can eat up to 19 meals per week by swiping their cards, allowing the dining service to automatically record dining information in the database. Even if 19 meals are not consumed in a particular week, the enrolled customers are not allowed to roll over their unused swipes. In contrast, the premium 19 meal plan allows customers to roll over their unused swipes. Although it is possible to eat more than two meals at a given meal time, this is unlikely and represents a negligible fraction of the total traffic. We therefore focus on the customers' decision of whether to swipe their card (i.e., to eat or not).

There are two major operational problems that Dining faces that it should plan for using a demand forecast: food inventory and labor shifts. Food inventory consists primarily of perishable ingredients. An appropriate level of food inventory will reduce the spoilage rate and allow the dining service to serve meals consistently. Inventories should generally be completed several days in advance. The dining service also needs to make a labor plan in advance. Therefore, we consider a situation in which the dining service makes the plan one week in advance.

By sampling 5000 enrolled customers in the Fall quarters of 2008 and 2009, we create a panel with size $(T, N) = (315, 5000)$. Here $T = 315$ implies that the data period spans 9 weeks (5 meal times/per day $\times$ 63 days) and there are $N = 5000$ individual customers. The start dates of the panel data are October 5, 2008, and October 4, 2009. Thus, both panels start on a Monday morning. Therefore, we analyze more than 3 million transaction records (315×5000 panels for fall 2008 and

fall 2009).

Table 1 provides the two-way tables on the empirically calculated probability of eating at the dining restaurant for both years. We can see that weekdays (Monday $\sim$ Friday) have higher demand than do weekends. Moreover, the Lunch and Dinner periods have higher demand than the remaining periods. Although this information is useful, we would like to explore the demand structure using the new model. In the next section, we introduce the new panel logistic model with interactive effects.

5.2 Model specification and estimation results

Let y_{it} be consumption (Eat; Yes=1 or No=0) with respect to an individual customer i. Then, the utility of choice is assumed to be $u_{it} = \boldsymbol{x}_{it}'\boldsymbol{b}_i + \boldsymbol{f}_t'\boldsymbol{\lambda}_i + \varepsilon_{it}$, where $\boldsymbol{f}_t'\boldsymbol{\lambda}_i$ represents the interactive fixed effects, and ε_{it} is the error term. In our analysis, we use the following information to predict y_{it}: Meal time (Morning, Lunch, Evening, Before Mid., or After Mid.), Day of the week (Monday, Tuesday,..., Sunday), and One week prior visit at the same meal time (Yes, or No). Table 2 lists the names of the variables. Note that price is not included because this is a prepaid plan. We use indicator variables for the predictors in Table 2. To understand the data-generating structure without imposing any prior subjective intuitions, we obtain the maximum likelihood estimator with a factor structure. We generate 2,000 iterations with 1000 burn-in iterations using the proposed posterior sampling algorithms.

In practice, we have to select the number of unobservable factors (the dimension of F) to adequately describe the information contained in the observed panel data. Here, we use the proposed model selection criterion $IC(r)$ in (11). For the 2008 panel, we divide a set of 5000 customers into two groups. One group consists of 2500 customers to estimate the model parameters. Under the given value of the number of factors r, the estimator of the factor structure $\hat{F}_{Y_1}$ is obtained. Using this estimated factor structure, we calculate the value of the $IC_1(r)$ in (10) score based on the remaining 2500 customers. Similarly, $IC_2(r)$ score is obtained. We complete this process under the different numbers of factors and select the best number of factors as the minimizer of the $IC(r)$ score in (11). The same procedure is also applied to the 2009 panel. For both panels, the best number of common factors is 1.

By setting $r = 1$, we compute the maximum likelihood estimators of $\boldsymbol{b}_i$ and factor loadings $\boldsymbol{\lambda}_i$ for each of the individual customers. Figure 1 (a) shows the estimated regression coefficients and the factor loadings for the 2008 panel. To observe the relative contrast among the customers, coefficients on meal time are standardized to have mean zero. Let $\hat{\boldsymbol{b}}_{i,meal} = (\hat{b}_{i,morning}, \hat{b}_{i,lunch}, \hat{b}_{i,dinner}, \hat{b}_{i,before\ mid.}, \hat{b}_{i,after\ mid.})'$ be

the coefficient vector on the meal times for the i-th customer. Then, it is standardized by subtracting $\sum_{i=1}^{N} \hat{\boldsymbol{b}}'_{i,meal}\mathbf{1}/(5N)$, where $\mathbf{1}$ is the vector of ones. Similarly, the coefficients on the day of the week are also standardized.

Each column in Figure 1 (a) corresponds to the individual's sensitivity to the predictors and to the common factors. On the top of the figure, trees based on the hierarchical clustering are presented. We can see that the customers' eating behaviors are categorized into several segments. Figure 1 (a) shows that there exists a customer segment such that their visiting behaviors are positively affected by the meal time rather than by the day of the week. The demand on Saturday is much lower than on weekdays. Moreover, a different customer segment shows that the visiting behaviors are positively affected by the meal time. With respect to the factor loadings in Figure 1 (a), the estimated unobservable factor has relatively larger impacts. Furthermore, some segments have large positive coefficients on the estimated unobservable factor, while some segments have small or negative coefficients.

It is also interesting to observe whether some behavioral changes occurred between the 2008 and 2009 panels. Figure 1 (b) shows the estimated regression coefficients and the factor loadings for the 2009 panel. We observe similar patterns in these estimated coefficients.

5.3 Estimated unobservable structure

The 2008 and 2009 panels contain one factor structure. Here, we attempt to interpret these estimated factors. First, the time periods of these two panels exactly match because the stating date of each panel is the second week of the fall quarter in each year. One simple question is whether these factors exhibit similar structures. We therefore conduct standard correlation analysis, which is used to identify and measure the associations between two sets of variables. We find that the estimated unobservable factor in 2008 and that in 2009 have an association at the 1% significance level.

We next explore the economic meanings of the estimated factors. We simply regress the estimated factors on the set of predictors in Table 2 except for the variable "One week prior visit at the same meal time" because this variable varies over i. We find that the estimated factors are strongly related to the Meal time (Morning, Lunch, Evening, Before Mid., or After Mid.) and Day of the week (Monday, Tuesday,..., Sunday) in the sense that all estimated regression coefficients are statistically significant at the 1% level. This implies that the estimated factors can be regarded as a proxy for customers' weekly activities. This also implies that an endogeneity issue is detected because the set of predictors in Table 2 is correlated

with the unobservable error term η_{it} in (1). Thus, it is important to consider the interactive fixed effects.

5.4 Forecasting the demand

There are two major costs that the dining service should manage using a demand forecast: food inventory and labor shifts. We demonstrate the forecasting performance of the proposed model $h = 7$ days ahead of demand. For the 2008 panel, the forecasting period is fixed from November 17, 2008, Monday morning, until the end of the panel. Similarly, the forecasting period for 2009 begins on November 16, 2009, Monday morning. For a given forecast horizon h, the out-of-sample forecasts $y_{i,t+h}$ are obtained as follows.

Here, we demonstrate that accounting for unobserved heterogeneity is beneficial. We compare the proposed maximum likelihood estimator (MLE with factor structure $\boldsymbol{f}_t'\boldsymbol{\lambda}_i$ in (8)) with the maximum likelihood estimator ignoring the factor structure (MLE without factor structure $\boldsymbol{f}_t'\boldsymbol{\lambda}_i$). This is a fair comparison between the model that accounts for unobserved heterogeneity and the model ignoring this effect.

We estimate the model using the panel with the time series length t, i.e, the first t data points are used for $i = 1, ..., N$. We then compute the h-step ahead out-of-sample forecast $y_{i,t+h}$ as follows:

$$\hat{P}(y_{i,t+h} = 1|\boldsymbol{x}_{i,t+h}, \hat{\boldsymbol{b}}_i, \tilde{\boldsymbol{f}}_{t+h}, \hat{\boldsymbol{\lambda}}_i) = \frac{\exp(\boldsymbol{x}_{i,t+h}'\hat{\boldsymbol{b}}_i + \tilde{\boldsymbol{f}}_{t+h}'\hat{\boldsymbol{\lambda}}_i)}{1 + \exp(\boldsymbol{x}_{i,t+h}'\hat{\boldsymbol{b}}_i + \tilde{\boldsymbol{f}}_{t+h}'\hat{\boldsymbol{\lambda}}_i)}.$$

where $\hat{\boldsymbol{b}}_i$ and $\hat{\boldsymbol{\lambda}}_i$ are the maximum likelihood estimates obtained by the panel with size (t, N) (not the full size (T, N)), and we set the number of factors at one. Because the unobserved factor $\tilde{\boldsymbol{f}}_{t+h}$ at time $t + h$ is not available, we use the most recent factor values corresponding to the same weekday and meal period. It is also possible to use other values such as the weighted sum of the historical factor values corresponding to the same weekday and meal period. However, we use the above approach because it is simple to implement. We also note that all elements of $\boldsymbol{x}_{i,t+h}$ are known at time t. Therefore, we can compute $\hat{P}(y_{i,t+h} = 1|\boldsymbol{x}_{i,t+h}, \hat{\boldsymbol{b}}_i, \tilde{\boldsymbol{f}}_{t+h}, \hat{\boldsymbol{\lambda}}_i)$ at time t. Then, the panel size for the estimation is moved forward by one data point. The estimation and forecasting procedure are then iterated until the forecast of the end of the panel.

Let t_s be the starting period of forecasting and t_e be the end of forecasting. Let $y_{i,t+h}$ be the observed choice and $\hat{P}(y_{i,t+h} = 1|\boldsymbol{x}_{i,t+h}, \hat{\boldsymbol{b}}_i, \tilde{\boldsymbol{f}}_{t+h}, \hat{\boldsymbol{\lambda}}_i)$ be the forecast. Then, in the aggregate, the foretasted demand d_{t+h} and the actual consumption

c_{t+h} are

$$d_{t+h} = \sum_{i=1}^{N} \hat{P}(y_{i,t+h} = 1 | \boldsymbol{x}_{i,t+h}, \hat{\boldsymbol{b}}_i, \tilde{\boldsymbol{f}}_{t+h}, \hat{\boldsymbol{\lambda}}_i), \quad \text{and} \quad c_{t+h} = \sum_{i=1}^{N} y_{i,t+h},$$

respectively. It is ideal if d_{t+h} is close to c_{t+h}. To evaluate the out-of-sample forecasting performance, we use the out-of-sample mean absolute forecast errors (MAFE) and the predictive log-likelihood (PLL).

$$\begin{aligned} \text{MAFE} &= \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} |c_t - d_t|, \\ \text{PLL} &= \frac{1}{N(t_e - t_s + 1)} \sum_{t=t_s}^{t_e} \sum_{i=1}^{N} \left[y_{it} \log \hat{P}_{it} + (1 - y_{it}) \log(1 - \hat{P}_{it}) \right], \end{aligned}$$

where $\hat{P}_{it} \equiv \hat{P}(y_{i,t+h} = 1 | \boldsymbol{x}_{i,t+h}, \hat{\boldsymbol{b}}_i, \tilde{\boldsymbol{f}}_{t+h}, \hat{\boldsymbol{\lambda}}_i)$ is the h-step ahead out-of-sample forecast. To show the effectiveness of the model, we also include a management practice employed by the dining service. The dining service's forecasting method is to use last week's aggregated demand, which is easy to calculate.

In the same manner, we also implement the forecasting. The h-step ahead out-of-sample forecast $y_{i,t+h}$ is given as

$$\begin{aligned} \bar{P}(y_{i,t+h} = 1 | \boldsymbol{x}_{i,t+h}) &= \int \frac{\exp(\boldsymbol{x}'_{i,t+h} \boldsymbol{b}_i + \tilde{\boldsymbol{f}}'_{t+h} \boldsymbol{\lambda}_i)}{1 + \exp(\boldsymbol{x}'_{i,t+h} \boldsymbol{b}_i + \tilde{\boldsymbol{f}}'_{t+h} \boldsymbol{\lambda}_i)} d\pi(B, F, \Lambda | Y_t, X_t) \\ &\approx \frac{1}{H} \sum_{k=1}^{H} \left[\frac{\exp(\boldsymbol{x}'_{i,t+h} \boldsymbol{b}_i^{(k)} + \tilde{\boldsymbol{f}}_{t+h}^{(k)'} \boldsymbol{\lambda}_i^{(k)})}{1 + \exp(\boldsymbol{x}'_{i,t+h} \boldsymbol{b}_i^{(k)} + \tilde{\boldsymbol{f}}_{t+h}^{(k)'} \boldsymbol{\lambda}_i^{(k)})} \right], \end{aligned}$$

where $\{\boldsymbol{b}_i^{(k)}, \tilde{\boldsymbol{f}}_{t+h}^{(k)}, \boldsymbol{\lambda}_i^{(k)}; k = 1, ..., H\}$ are the set of H posterior samples based on the information $\{Y_t, X_t\} \equiv \{y_{it}, \boldsymbol{x}_{it} | i = 1, ..., N, t = 1, ..., t\}$ observed up to the time t. Setting the number of unobservable factors $r = 1$ and the prior variance of the normal prior $\pi(\boldsymbol{\gamma}_i | A_{\gamma_i})$ as $A_{\gamma_i} = 10^5 I$, we generate $H = 1000$ posterior samples. Again, the unobserved factor $\tilde{\boldsymbol{f}}_{t+h}^{(k)}$ at time $t + h$ is not available. Similar to the Non-Bayesian approach, we use the posterior sample of most recent factor values corresponding to the same weekday and meal period.

Table 3 reports the measure of forecasting performance at different forecast horizons. From the table, we make the following observations. Both the Bayesian and MLE with factor structure perform much better than without a factor structure. Bayesian and MLE perform comparably. They have much smaller MAFE and the larger PLL for 2008 and 2009. This implies that our modeling procedure can improve the efficiency of the dining operation. Figure 2 is the boxplot of the predictive log-likelihood at each forecasting point: $\text{PLL}_t = \frac{1}{N} \sum_{i=1}^{N} [y_{it} \log \hat{P}_{it} + (1 - y_{it}) \log(1 - \hat{P}_{it})]$.

We can see the model with endogeneity provides higher PLL values. This also indicates that the proposed procedure represents a useful tool for addressing endogeneity.

Finally, we note that the other variables may help to forecast some customers' behaviors. However, collecting such variables increases the data acquisition cost, while it does not guarantee improved forecasting performance. Moreover, obtaining access to such useful information is sometimes nearly impossible, for example, knowing individual customers' private plans to take trips. A customer is obviously not on campus during such a trip. However, it is even difficult to know whether the customer is taking a trip at a given time. Additionally, some micro-behavioral information may be helpful. For example, customers who enrolled in the regular 19 meal plan may have a substantial number of meals remaining to use at the end of a particular week. This might affect their eating behaviors. Because such information can not be obtained until midnight Thursday, it is impossible to use such information when the dining service creates a demand forecast several days in advance. Therefore, the set of predictors should be selected while considering the data acquisition cost and feasibility.

6 Further extensions of the inference procedure

6.1 Multiple alternatives model

In this section, we discuss an extension of our proposed data-augmentation strategy. In general, the number of alternatives will exceed 2. Each individual makes a single choice among many alternatives, such as transportation modes and occupational fields, selecting one candidate out of many. We extend our data-augmentation strategy to this setting.

Suppose that there are $i = 1, ..., N$ individuals and $J + 1$ alternatives labeled $\{0, 1, ..., J\}$. At a time period t, each individual chooses one of the alternatives. Consider the random utility for the j-th alternative

$$u_{ijt} = \boldsymbol{x}'_{it}\boldsymbol{b}_{ij} + \eta_{ijt} + \varepsilon_{ijt}, \quad i = 1, \ldots, N;\ t = 1, \ldots, T; j = 0, 1, 2, ..., J \tag{12}$$

where η_{ijt} denotes the unobserved structure of individual i's choice j, which can vary across time t, and ε_{ijt} follows a type I Extreme Value distribution. Alternative j $(j = 0, 1, 2, ..., J)$ will be chosen if and only of $u_{ijt} > u_{ikt}$ $(k \neq j)$. Thus, at time t, an individual i chooses alternative j if it offers the highest utility among all alternatives. Similar to the arguments in Section 3, we assume that these unobserved structures vary across time and individuals according to a factor structure:

$$\eta_{ijt} = \sum_{\ell=1}^{r} f_{jt\ell}\lambda_{ij\ell} = \boldsymbol{f}'_{jt}\boldsymbol{\lambda}_{ij} \tag{13}$$

and $\boldsymbol{f}_{jt}$ is an $r_j \times 1$ vector of unobservable factors and $\boldsymbol{\lambda}_{ij}$ represents the factor loadings.

Let $y_{ijt} \in \{0, 1\}$ denote the observed choice outcome, taking value 1 if the corresponding alternative j is chosen and 0 otherwise. Let $\boldsymbol{b}_i = (\boldsymbol{b}'_{i1}, ..., \boldsymbol{b}'_{iJ})'$, $\boldsymbol{f}_t = (\boldsymbol{f}'_{1t}, ..., \boldsymbol{f}'_{Jt})'$, and $\boldsymbol{\lambda}_i = (\boldsymbol{\lambda}'_{i1}, ..., \boldsymbol{\lambda}'_{iJ})'$. After normalizing the coefficients $(\boldsymbol{b}'_{i0}, \boldsymbol{\lambda}'_{i0})'$ for alternative 0 to zero, we obtain the multinomial logit specification (See McFadden (1973)) with the following choice probabilities

$$\begin{aligned} P(y_{ijt} = 1|\boldsymbol{x}_{it}, \boldsymbol{b}_i, \boldsymbol{f}_t, \boldsymbol{\lambda}_i) &= \frac{\exp(\boldsymbol{x}'_{it}\boldsymbol{b}_{ij} + \boldsymbol{f}'_{jt}\boldsymbol{\lambda}_{ij})}{1 + \sum_{k=1}^{J} \exp(\boldsymbol{x}'_{it}\boldsymbol{b}_{ik} + \boldsymbol{f}'_{kt}\boldsymbol{\lambda}_{ik})}, \quad j = 1, ..., J, \\ P(y_{i0t} = 1|\boldsymbol{x}_{it}, \boldsymbol{b}_i, \boldsymbol{f}_t, \boldsymbol{\lambda}_i) &= \frac{1}{1 + \sum_{k=1}^{J} \exp(\boldsymbol{x}'_{it}\boldsymbol{b}_{ik} + \boldsymbol{f}'_{kt}\boldsymbol{\lambda}_{ik})}. \end{aligned} \tag{14}$$

Assuming that the errors ε_{ijt} are independently and identically distributed, the joint probability of observing the complete set of choices $Y \equiv \{y_{ijt}|i = 1, ..., N, t = 1, ..., T, j = 1, ..., J\}$ is

$$\begin{aligned} &L(Y|X, B, F, \Lambda) \\ &= \prod_{i=1}^{N} \prod_{t=1}^{T} \prod_{j=1}^{J} \left[\frac{\exp(\boldsymbol{x}'_{it}\boldsymbol{b}_{ij} + \boldsymbol{f}'_{jt}\boldsymbol{\lambda}_{ij})}{1 + \sum_{k=1}^{J} \exp(\boldsymbol{x}'_{it}\boldsymbol{b}_{ik} + \boldsymbol{f}'_{kt}\boldsymbol{\lambda}_{ik})} \right]^{y_{ijt}} \left[\frac{1}{1 + \sum_{k=1}^{J} \exp(\boldsymbol{x}'_{it}\boldsymbol{b}_{ik} + \boldsymbol{f}'_{kt}\boldsymbol{\lambda}_{ik})} \right]^{1 - \sum_{k'=1}^{J} y_{ik't}}, \end{aligned}$$

where $X \equiv \{\boldsymbol{x}_{it}|i = 1, ..., N, t = 1, ..., T\}$, $\Lambda = (\Lambda_1, ..., \Lambda_J)$ with $\Lambda_j = (\boldsymbol{\lambda}_{j1}, ..., \boldsymbol{\lambda}_{jN})'$, $B = (B_1, ..., B_J)$ with $B_j = (\boldsymbol{b}_{j1}, ..., \boldsymbol{b}_{jN})'$ and $F = (F_1, ..., F_J)$ with $F_j = (\boldsymbol{f}_{j1}, ..., \boldsymbol{f}_{jT})'$ is the common factor.

We first consider the posterior sampling procedure of B and Λ because these parts are nearly identical to those presented in the previous section. Similar to the idea of Holmes and Held (2006) and Polson and Scott (2013), we rewrite each probability (14) as

$$P^*(y_{ijt} = 1|\boldsymbol{x}_{it}, \boldsymbol{b}_i, \boldsymbol{f}_t, \boldsymbol{\lambda}_i) = \frac{\exp(\boldsymbol{v}'_{ijt}\boldsymbol{\gamma}_{ij} - \log\{1 + \sum_{k \neq j} \exp(\boldsymbol{v}'_{ikt}\boldsymbol{\gamma}_{ik})\})}{1 + \exp(\boldsymbol{v}'_{ijt}\boldsymbol{\gamma}_{ij} - \log\{1 + \sum_{k \neq j} \exp(\boldsymbol{v}'_{ikt}\boldsymbol{\gamma}_{ik})\})},$$

for $j = 1, ..., J$. Here,

$$\boldsymbol{v}_{ijt} = (\boldsymbol{x}'_{it}, \boldsymbol{f}'_{jt})', \quad \boldsymbol{\gamma}_{ij} = (\boldsymbol{b}'_{ij}, \boldsymbol{\lambda}'_{ij})',$$

and $P(y_{i0t} = 1|\boldsymbol{x}_{it}, \boldsymbol{b}_{ij}, \boldsymbol{f}_{jt}, \boldsymbol{\lambda}_{ij}) = 1 - \sum_{j=1}^{J} P^*(y_{ijt} = 1|\boldsymbol{x}_{it}, \boldsymbol{b}_i, \boldsymbol{f}_t, \boldsymbol{\lambda}_i)$. Note that this transformation makes use of the normalization on alternative 0 (usually referred to as the outside option), while the normalization is not used in the multinomial choice probability in Holmes and Held (2006) or Polson and Scott (2013). Normalization is useful for parameter identification for choice-dependent coefficients; see Greene (2000, page 860).

The likelihood for $\boldsymbol{\gamma}_{ij} = (\boldsymbol{b}'_{ij}, \boldsymbol{\lambda}'_{ij})'$ conditional upon the matrices Λ and B with column vectors $\boldsymbol{\lambda}_{ij}$ and $\boldsymbol{b}_{ij}$ removed is

$$\begin{aligned} & L(\boldsymbol{\gamma}_{ij}|X, B_{-b_{ij}}, F, \Lambda_{-\lambda_{ij}}) \\ = & \prod_{t=1}^{T} \left(\frac{\exp(\boldsymbol{v}'_{ijt}\boldsymbol{\gamma}_{ij} - \log\{1 + \sum_{k\neq j} \exp(\boldsymbol{v}'_{ikt}\boldsymbol{\gamma}_{ik})\})}{1 + \exp(\boldsymbol{v}'_{ijt}\boldsymbol{\gamma}_{ij} - \log\{1 + \sum_{k\neq j} \exp(\boldsymbol{v}'_{ikt}\boldsymbol{\gamma}_{ik})\})} \right)^{y_{ijt}} \\ & \times \left(\frac{1}{1 + \exp(\boldsymbol{v}'_{ijt}\boldsymbol{\gamma}_{ij} - \log\{1 + \sum_{k\neq j} \exp(\boldsymbol{v}'_{ikt}\boldsymbol{\gamma}_{ik})\})} \right)^{1-y_{ijt}}, \end{aligned}$$

where $\boldsymbol{v}_{ijt} = (\boldsymbol{x}'_{it}, \boldsymbol{f}'_{jt})'$. Incorporating the Polya-Gamma auxiliary variable, the likelihood becomes

$$\begin{aligned} & L(\boldsymbol{\gamma}_{ij}|X, B_{-b_{ij}}, F, \Lambda_{-\lambda_{ij}}) \\ \propto & \prod_{t=1}^{T} \Bigg[\exp\{z_{ijt}(\boldsymbol{v}'_{ijt}\boldsymbol{\gamma}_{ij} - \log\{1 + \sum_{k\neq j} \exp(\boldsymbol{v}'_{ikt}\boldsymbol{\gamma}_{ik})\})\} \\ & \times \exp\{-\omega_{ijt}(\boldsymbol{v}'_{ijt}\boldsymbol{\gamma}_{ij} - \log\{1 + \sum_{k\neq j} \exp(\boldsymbol{v}'_{ikt}\boldsymbol{\gamma}_{ik})\})^2/2\} p(\omega_{ijt}) \Bigg], \end{aligned}$$

where $z_{ijt} = y_{ijt} - 1/2$ and $p(\omega_{ijt})$ is the density of a Polya-Gamma random variable with parameters $(1, 0)$.

To simplify the argument, we assume the normal conjugate prior on $\boldsymbol{\gamma}_{ij} \sim N(\mathbf{0}, A_{ij})$. One can employ the adaptive lasso prior because one can further place a hyperprior on A_{ij} in that case (See Section 6.2). Then, the posterior density of $\boldsymbol{\gamma}_{ij}$ is

$$\pi(\boldsymbol{\gamma}_{ij}|X, B_{-b_{ij}}, F, \Lambda_{-\lambda_{ij}}) = N(\boldsymbol{m}_{ij}, V_{ij}),$$

with

$$\begin{aligned} \boldsymbol{m}_{ij} &= (W'_{ij}\Omega_{ij}W_{ij} + A_{ij}^{-1})^{-1} W'_{ij}(\boldsymbol{z}^*_{ij} - \Omega_{ij}\boldsymbol{c}_{ij}), \\ V_{ij} &= (W'_{ij}\Omega_{ij}W_{ij} + A_{ij}^{-1})^{-1}, \end{aligned}$$

where $W_i = (X_i, F_j)$, $\boldsymbol{z}^*_{ij} = (z^*_{ij1}, ..., z^*_{ijT})'$ with $z^*_{ij} = z_{ijt}/\omega_{ijt}$, $\boldsymbol{c}_{ij} = (c_{ij1}, ..., c_{ijT})'$ with $c_{ijt} = \log\{1 + \sum_{k\neq j} \exp(\boldsymbol{v}'_{ikt}\boldsymbol{\gamma}_{ik})\}$, $\Omega_{ij} = \text{diag}\{\omega_{ij,1}, ..., \omega_{ij,T}\}$. Thus, we can easily draw a posterior sample of $\boldsymbol{\gamma}_{ij}$ using the Gibbs sampler.

A conditional posterior of ω_{it} is

$$\pi(\omega_{ijt}|Y, X, B, \Lambda, \Omega_{-\omega_{ijt}}) \propto \exp\{-\omega_{ijt}\{\boldsymbol{v}'_{ikt}\boldsymbol{\gamma}_{ij} - \log\{1 + \sum_{k\neq j} \exp(\boldsymbol{v}'_{ikt}\boldsymbol{\gamma}_{ik})\}^2/2\} p(\omega_{ijt}),$$

which is a Polya-Gamma distribution with parameter $(1, \boldsymbol{v}'_{ikt}\boldsymbol{\gamma}_{ij} - \log\{1 + \sum_{k\neq j} \exp(\boldsymbol{v}'_{ikt}\boldsymbol{\gamma}_{ik})\})$. Again, we can easily draw a posterior sample of ω_{ijt} using the Gibbs sampler. The conditional posterior of the factor structure F_j is easily obtained by simply extending the analysis in Section 3. Details are given in Appendix D.

6.2 High-dimensional predictors

When the dimension of $\boldsymbol{x}_{it}$ is large, some shrinkage methods are useful. The use of a shrinkage prior on $\boldsymbol{\gamma}_i$ allows us to address high-dimensional predictors. In the context of Bayesian (cross-sectional) linear regression, Park and Casella (2008) study the Bayesian lasso to exploit model inference via posterior distributions. To extend the proposed algorithm by incorporating the shrinkage approach, we can employ the Bayesian adaptive lasso prior (Leng et al. (2014)). It is ideal to place a larger penalty on the coefficients of unimportant predictors. The Bayesian adaptive lasso prior allows for variable selection with more flexible penalties than the Bayesian lasso. The Bayesian adaptive lasso prior on $\boldsymbol{\gamma}_i$ is

$$\pi(\boldsymbol{\gamma}_i) = \prod_{k=1}^{p_i+r} \frac{\kappa_{ik}}{2} \exp\left[-\kappa_{ik}|\gamma_{ik}|\right],$$

where κ_{ik} corresponds to the adaptive weights in the adaptive lasso framework (Zou; 2006). Intuitively, a small penalty will be applied to the set of regressors that are relevant to the choices and a large penalty will be applied to those that are irrelevant.

In Appendix D, we summarize the posterior sampling procedure based on the adaptive lasso prior. This extended algorithm can be applied to the panel logit model with high-dimensional predictors in the presence of endogeneity.

7 Conclusion

In this paper, we introduced a new panel logistic regression model with interactive fixed effects. The proposed model will have a wide range of applications, for example, the inference of market structures in marketing research (Elrod and Keane, 1995), the analysis of partisanship patterns of roll-call votes from the United States Senate (Hahn et al 2012), association studies based on high-throughput single nucleotide polymorphism (SNP) data in genomic DNA study (Lee et al. 2010), and the study of firms' decisions to split their shares (Perez et al. 2015). Although our dataset does not need to consider pricing problems, our methodology also contributes to the price-based revenue management literature (Özer and Phillips (2012) and Talluri and Van Ryzin (2005)). Because our proposed modeling methodology easily incorporates price information, one can use the proposed methods as pricing decision support tools. For example, our proposed method could be applied to the analysis of Maddah and Bish (2007), where static pricing and inventory decisions for multiple competing products are considered.

The estimation of the interactive effects was challenging because the likelihood function has an inconvenient form in terms of model parameters. We proposed an

data-augmented inference procedure and developed a posterior sampling algorithm. Theoretical properties were established for the proposed procedure. We also showed that the Bayesian shrinkage approach can be easily incorporated into our method. Moreover, we showed that the method can be easily extended to multinomial logit models.

As shown in the empirical application, the proposed method provides useful information for demand forecasting. Moreover, we showed that the incorporation of endogeneity improves forecasting performance. By applying our superior forecasting tool, the dining service can manage its operations more efficiently. We would like to apply the proposed procedure to the analysis of other large-scale panel datasets.

References

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. (2015a) "Selecting the regularization parameters in high-dimensional panel data models: consistency and efficiency," *Econometric Reviews*, forthcoming.

Ando, T. and Bai, J. (2015b) "Asset pricing with a general multifactor structure," *Journal of Financial Econometrics*, 13, 556–604.

Ando, T. and Bai, J. (2016) "Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures," *Journal of the American Statistical Association*, forthcoming.

Andrews, R.L. Currim, I.S. and Leeflang, P.S.H. (2011), "A comparison of sales response predictions from demand models applied to store-Level versus panel data," *Journal of Business and Economic Statistics*, 29, 319–326.

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

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.

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

Calli, M. K. and Weverbergh, M. (2009), "Forecasting newspaper demand with censored regression," *Journal of the Operational Research Society*, 60, 944 –951.

Chen, M. Fernández-Val, I. and Weidner, M. (2014) Nonlinear Panel Models with Interactive Effects, Working Paper.

Chudik, A. and Pesaran, M.H. (2015) "Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors," *Journal of Econometrics*, 188, 393–420.

Connor, G. and Korajzcyk, R. (1986) "Performance measurement with the arbitrage pricing theory: a new framework for analysis," *Journal of Financial Economics*, 15, 373–394.

Elrod, T. and Keane, M. P. (1995) "A factor-analytic Probit model for representing the market structure in panel data," *Journal of Marketing Research*, 32, 1–16.

Ebbes, P., Wedel, M., Böckenholt, U. and Steerneman, T. (2005) "Solving and testing for regressor-error (in)dependence when no instrumental variables are available: with new evidence for the effect of education on income," *Quantitative Marketing and Economics*, 3, 365–392.

Fan, J. Liao, Y. and Mincheva, M. (2011) "High-dimensional covariance matrix estimation in approximate factor models," *Annals of Statistics*, 39, 3320–3356.

Fernández-Val, I. and Weidner, M. (2015) "Individual and time effects in nonlinear panel data models with large N, T," *Journal of Econometrics*, forthcoming.

Fildes, R. and Kumar, V. (2002), "Telecommunications demand forecasting – a review," *International Journal of Forecasting*, 18, 489-522.

Fruhwirth-Schnatter, S. and Fruhwirth, R. (2007) "Auxiliary mixture sampling with applications to logistic models," *Computational Statistics and Data Analysis*, 51, 3509–3528.

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

Hahn, P.R. Scott, J. and Carvalho, C.M. (2012) "A Sparse Factor-Analytic Probit Model for Congressional Voting Patterns," *Journal of Royal Statistical Society*, C61, 619–635.

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.

Hallin, M., and R. Liška (2011), "Dynamic factors in the presence of blocks," *Journal of Econometrics*, 163, 29–41.

Harvey, A. and Koopman, S.J. (1993), "Forecasting hourly electricity demand using time-varying splines," *Journal of the American Statistical Association*, 88, 1228–1236.

Hanssens, D.M. (1998), "Order forecasts, retail sales, and the marketing mix for consumer durables," *Journal of Forecasting*, 17, 327–346.

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, 438–456,

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

Koop G., Leon-Gonzalez R. and Strachan R. (2010) "Efficient posterior simulation for cointegrated models with priors on the cointegration space," *Econometric Reviews*, 29, 224–242.

Lam, C. and Yao, Q. (2012) "Factor modeling for high-dimensional time series: inference for the number of factors," *Annals of Statistics*, 40, 694–726.

Lee, S., Huang, J.Z. and Hu, J. (2010) "Sparse principal components analysis for binary data," *Annals of Applied Statistics*, 4, 1579–1601

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

Li, Y. and Ansari, A. (2014) "A Bayesian Semiparametric Approach for Endogeneity and Heterogeneity in Choice Models," *Management Science*, 60, 1161–1179.

Maddah, B., E. Bish. (2007). "Joint Pricing, Assortment, and Inventory Decisions for a Retailers Product Line," *Naval Research Logistics* **54**, 315-330.

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.

Naik, P.A., Wedel, M. and Kamakura, W. (2015) "Multi-index binary response analysis of large data sets," *Journal of Business & Economic Statistics*, 28, 67–81.

Özer, Ö., R. Phillips, eds. (2012). The Oxford Handbook of Pricing Management. Oxford University Press.

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

Perez, M.F. Shkilko, A. and Sokolov, K. (2015) "Factor models for binary financial data," *Journal of Banking and Finance*, forthcoming.

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.

Shively, T.S. and Sager, T.W. (2009), "A Bayesian approach to non-parametric monotone function estimation," *Journal of the Royal Statistical Society*, B71, 159–175.

Song, H. and Li, G. (2008), "Tourism demand modelling and forecasting -? A review of recent research," *Tourism Management*, 29, 203–220

Song, M. (2013) Asymptotic theory for dynamic heterogeneous panels with cross-sectional dependence and its applications. Working paper, Columbia University.

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.

Strachan, R. W. and Inder, B. (2004) "Bayesian analysis of the error correction model," *Journal of Econometrics*, 123, 307–325.

Syntetos, A.A.,Boylan,J.E. and Disney, S.M. (2009), "Forecasting for inventory planning: a 50-year review," *Journal of the Operational Research Society*, 60, 149–160

Talluri, K. T., G. J. Van Ryzin. (2005). The Theory and Practice of Revenue Management. Springer.

Tsay, R. and Ando, T. (2012) "Bayesian panel data analysis for exploring the impact of recent financial crisis on the U.S stock market," *Computational Statistics Data Analysis*, 56, 3345–3365.

Zou, H. (2006) "The adaptive Lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Table 1: Estimated eating probability based on 5000 individual customers.

2008

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Morning	0.317	0.298	0.312	0.304	0.293	0.061	0.051
Lunch	0.652	0.636	0.623	0.634	0.614	0.437	0.445
Dinner	0.699	0.690	0.637	0.679	0.502	0.439	0.562
Before Mid.	0.245	0.244	0.233	0.229	0.170	0.132	0.187
After Mid.	0.072	0.076	0.077	0.076	0.093	0.084	0.077

2009

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Morning	0.311	0.316	0.289	0.296	0.302	0.075	0.058
Lunch	0.657	0.633	0.601	0.635	0.626	0.429	0.456
Dinner	0.705	0.691	0.626	0.685	0.514	0.456	0.575
Before Mid.	0.254	0.236	0.220	0.222	0.180	0.156	0.232
After Mid.	0.061	0.075	0.068	0.063	0.094	0.071	0.072

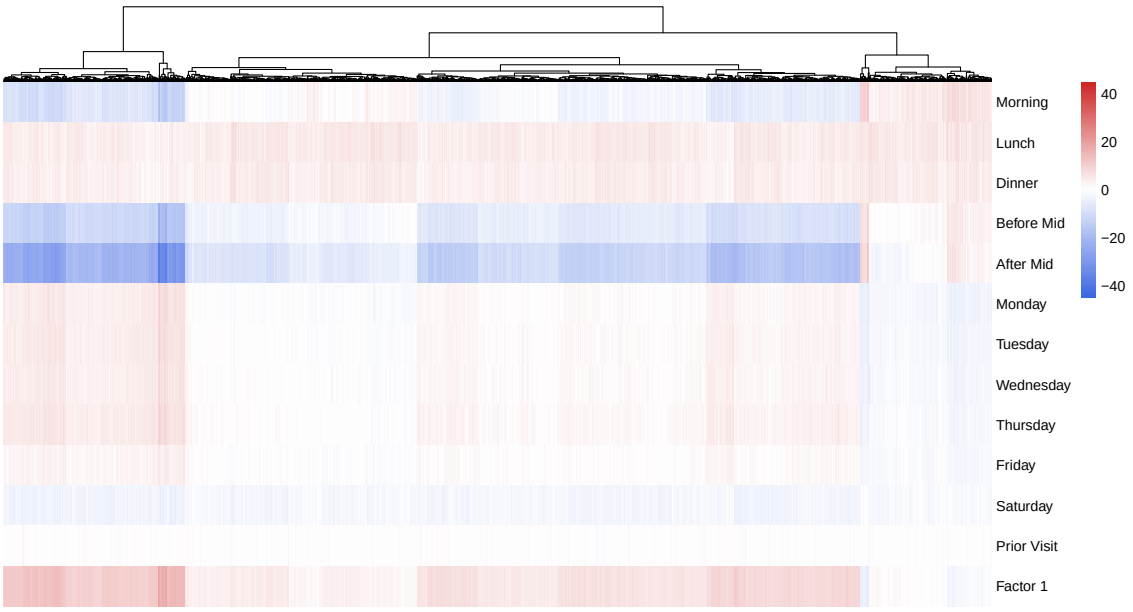
Table 2: The set of 11 variables in the model under study.

Name	Descriptions
Monday	Monday effect (= 1 if t corresponds to Monday, 0 otherwise)
Tuesday	Tuesday effect (= 1 if t corresponds to Tuesday, 0 otherwise)
Wednesday	Wednesday effect (= 1 if t corresponds to Wednesday, 0 otherwise)
Thursday	Thursday effect (= 1 if t corresponds to Thursday, 0 otherwise)
Friday	Friday effect (= 1 if t corresponds to Friday, 0 otherwise)
Saturday	Saturday effect (= 1 if t corresponds to Saturday, 0 otherwise)
Morning	Morning effect (= 1 if t corresponds to Morning, 0 otherwise)
Lunch	Lunch effect (= 1 if t corresponds to Lunch, 0 otherwise)
Dinner	Dinner effect (= 1 if t corresponds to Dinner, 0 otherwise)
Before Mid	Before Mid. effect (= 1 if t corresponds to Before Mid., 0 otherwise)
After Mid	After Mid. effect (= 1 if t corresponds to After Mid., 0 otherwise)
Prior Visit	Past visit effect (= 1 if the customer visited at the corresponding time one week before, 0 otherwise)

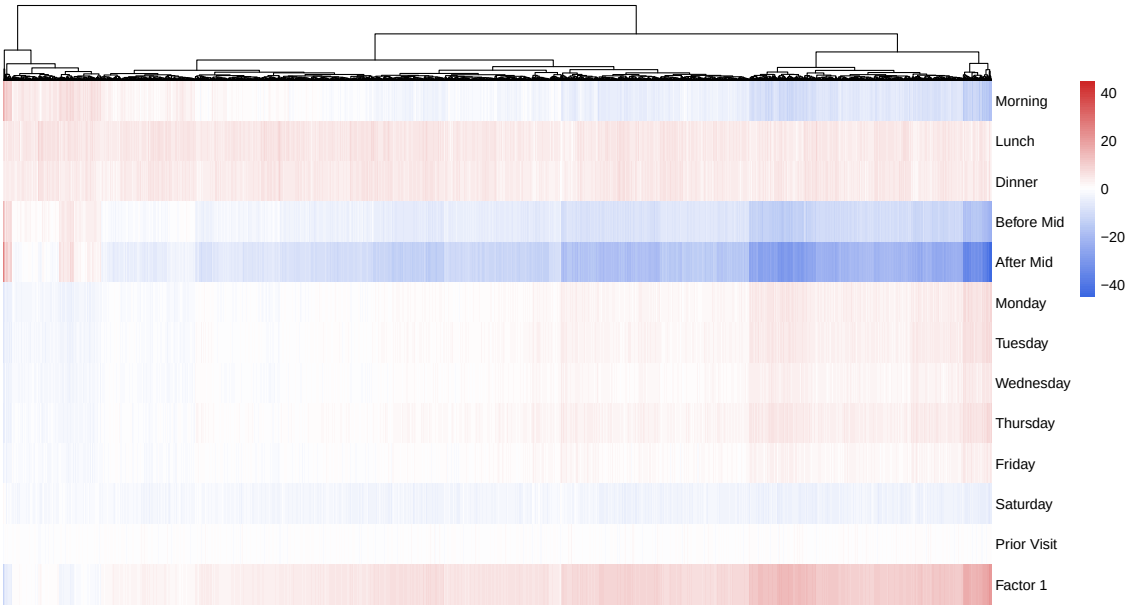
Table 3: Forecasting performance comparison based on MAFE and PLL for the Bayesian estimator with factor structure $\boldsymbol{f}_t'\boldsymbol{\lambda}_i$ (Bayesian with factor structure), the maximum likelihood estimator with factor structure $\boldsymbol{f}_t'\boldsymbol{\lambda}_i$ (MLE with factor structure), the maximum likelihood estimator without factor structure $\boldsymbol{f}_t'\boldsymbol{\lambda}_i$ (MLE without factor structure), and the dining service's approach.

2008	MAFE	PLL
Bayesian with factor structure	276.92	-0.58
MLE with factor structure	284.11	-0.55
MLE without factor structure	1061.92	-0.83
Dinging practice	386.49	———

2009	MAFE	PLL
Bayesian with factor structure	287.68	-0.57
MLE with factor structure	275.46	-0.55
MLE without factor structure	1097.12	-0.92
Dinging practice	370.39	———

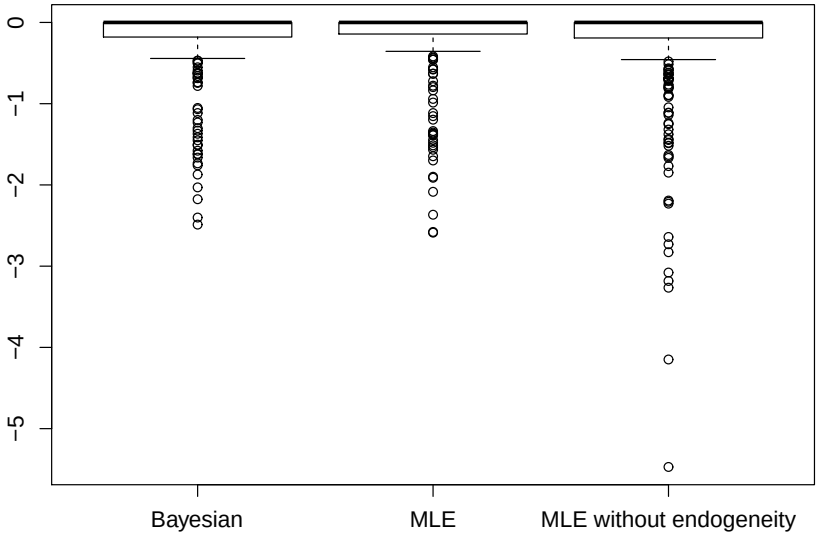


(a) 2008

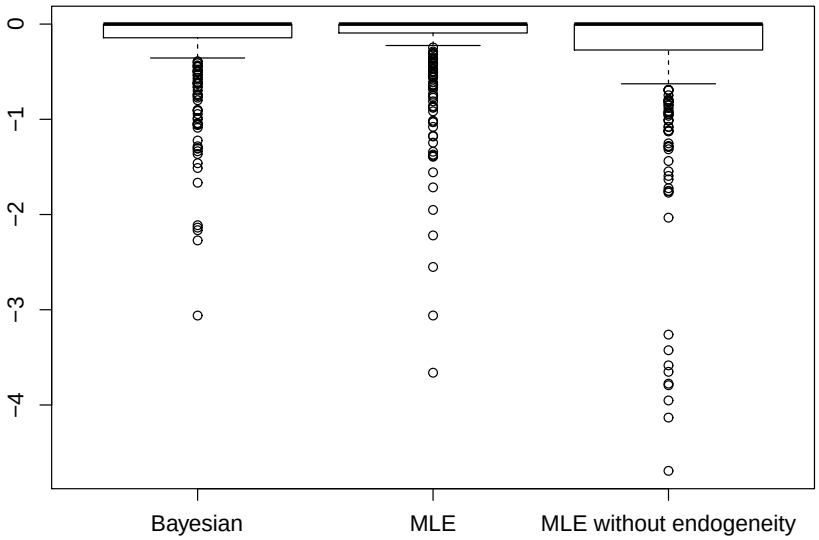


(b) 2009

Figure 1: Maximum likelihood estimates of the estimated regression coefficients and the factor loadings. Each column corresponds to the individual's sensitivity to the predictors and to the common factors. Trees on the top are from the hierarchical clustering.



(a) 2008



(b) 2009

Figure 2: Boxplot of the predictive log-likelihood at each forecasting point. the Bayesian estimator with factor structure $\boldsymbol{f}_t'\boldsymbol{\lambda}_i$ (Bayesian), the maximum likelihood estimator with factor structure $\boldsymbol{f}_t'\boldsymbol{\lambda}_i$ (MLE), and the maximum likelihood estimator without factor structure $\boldsymbol{f}_t'\boldsymbol{\lambda}_i$ (MLE without endogeneity).