Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

A Latent Variable Approach to Learning High-dimensional Multivariate longitudinal Data

Sze Ming Lee, Yunxiao Chen & Tony Sit

To cite this article: Sze Ming Lee, Yunxiao Chen & Tony Sit (07 Jan 2026): A Latent Variable Approach to Learning High-dimensional Multivariate longitudinal Data, Journal of the American Statistical Association, DOI: 10.1080/01621459.2025.2606384

To link to this article: https://doi.org/10.1080/01621459.2025.2606384

© 2026 The Author(s). Published with license by Taylor and Francis Group, LLC

View supplementary material

Accepted author version posted online: 07 Jan 2026.

Submit your article to this journal

Article views: 165

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uasa20

Check for updates

A Latent Variable Approach to Learning High-dimensional Multivariate longitudinal Data

Sze Ming Lee[a], Yunxiao Chen[a,*] and Tony Sit[b]

[a]Department of Statistics, London School of Economics and Political Science

[b]Department of Statistics, The Chinese University of Hong Kong

*Address for correspondence: Yunxiao Chen, Department of Statistics, London School of Economics and Political Science, London, UK. Email: y.chen186@lse.ac.uk

Accepted Manuscript

Abstract

High-dimensional multivariate longitudinal data, which arise when many outcome variables are measured repeatedly over time, are becoming increasingly common in social, behavioral and health sciences. We propose a latent variable model for drawing statistical inferences on covariate effects and predicting future outcomes based on high-dimensional multivariate longitudinal data. This model introduces unobserved factors to account for the between-variable and across-time dependence and assist the prediction. Statistical inference and prediction tools are developed under a general setting that allows outcome variables to be of mixed types and possibly unobserved for certain time points, for example, due to right censoring. A central limit theorem is established for drawing statistical inferences on regression coefficients. Additionally, an information criterion is introduced to choose the number of factors. The proposed model is applied to customer grocery shopping records to predict and understand shopping behavior.

Keywords: factor model; missing data; recurrent event data

1 Introduction

High-dimensional multivariate longitudinal data are becoming increasingly common, especially in social, behavioral and health sciences, where many outcomes are measured repeatedly within individuals. Examples include ecological momentary assessment data collected by smartphones or wearable devices for understanding within-subject social, psychological, and physiological processes in everyday contexts (Bolger and Laurenceau, 2013; Wang et al., 2014), electronic health record data for predicting and understanding health-related conditions (Lian et al., 2015; Zhang et al., 2020), computer logfile data for understanding human-computer interactions from solving complex computer-simulated tasks (Chen et al., 2019; Chen, 2020), and grocery shopping data for market basket analysis (Wan et al., 2017, 2018). These data may involve outcome variables of different types. For example, for ecological momentary assessment, physiological processes are typically measured by continuous variables, such as blood pressure, while psychological processes are recorded by participants' responses to survey items that involve binary or categorical variables. In addition, many multivariate longitudinal data may be derived from multitype recurrent event history (Chapter 2.5, Cook and Lawless, 2007), for which an outcome variable records whether a specific type of event occurs (e.g., purchasing a merchandise item) or the count of its occurrences within a time interval (e.g., the number of purchases).

In this paper, we study high-dimensional multivariate longitudinal data, aiming to (1) infer the effect of covariates on each outcome variable and (2) predict future outcomes based on covariates and historical data. These tasks involve three challenges. First, due to the nature of the data, there is a complex within-individual dependence structure which exists between outcome variables and across time. Valid statistical inference and accurate prediction become a challenge if one fails to account for the dependence properly. Second, the presence of many outcome variables implies a substantial number of item-specific parameters, bringing challenges to the statistical inference. The classical theory for M- or Z-estimators no longer applies, and new asymptotic results concerning the consistency and asymptotic normality under a high-dimensional regime are needed. Third, some observation units may be lost to follow-up or observed only intermittently, resulting in incomplete data. For example, grocery shopping records based on membership may be incomplete if customers occasionally shop without using their membership card.

To tackle these challenges, we propose a high-dimensional generalized latent factor model. In this model, low-dimensional factors are introduced within each observation unit to capture the between-item and across-time dependence that is not attributable to the covariates. The model is very flexible, allowing for many types of outcome variables, including binary, count, and continuous variables. In addition, a computationally efficient joint likelihood estimator is proposed that estimates the unobserved factors, loading parameters, and regression coefficients simultaneously, which treats the factors as fixed parameters. Asymptotic properties of this estimator are established, including a central limit theorem for drawing statistical inferences on regression coefficients and an information criterion for choosing the number of factors. Moreover, we introduce a missing indicator approach (see Chapter 26, Molenberghs and Verbeke, 2005) to account for data missingness. Under a Missing at Random (MAR) assumption, this approach can handle many missingness patterns, including right-censoring that is common to recurrent event data.

Various statistical methods have been proposed for analyzing multivariate longitudinal data. Generalized estimating equation (GEE) methods (e.g., Liang and Zeger, 1986; Prentice,

Accepted Manuscript

1988; Carey et al., 1993; Gray and Brookmeyer, 2000) are widely used for drawing statistical inferences on regression parameters relating the means of outcome variables to a set of covariates and parameters characterizing the marginal association between outcome variables. These methods often provide valid statistical inferences on parameters of interest without a need to specify a full joint distribution for the outcome variables. On the other hand, many joint models have been proposed for multivariate longitudinal data that are better at making predictions while still capable of drawing statistical inferences on parameters of interest, though the latter may be jeopardized by model misspecification. Joint models for multivariate longitudinal data include transition models (Liang and Zeger, 1989; Zeng and Cook, 2007) that are specified through a sequence of conditional probabilities of outcome variables given historical outcome variables and covariates, copula-based models (Lambert and Vandenhende, 2002; Smith et al., 2010; Panagiotelis et al., 2012) that specify a joint distribution via copulas, and latent variable models (Ten Have and Morabia, 1999; Oort, 2001; Liu and Hedeker, 2006; Hsieh et al., 2010; Proust-Lima et al., 2013; Wang et al., 2016; Ounajim et al., 2023; Sørensen et al., 2023) that capture the complex dependence structure by introducing latent variables. Latent variable models are very popular, thanks to their flexibility and interpretability. However, the statistical inference for these traditional latent variable models is carried out based on a marginal likelihood, where the latent variables are treated as random variables and marginalized out. This approach can hardly be extended to the high-dimensional setting with many latent variables due to the high computational cost of optimizing the marginal likelihood. Our method extends the traditional latent variable models to the high-dimensional setting and further overcomes their computational challenge using the proposed joint likelihood estimator.

The proposed method is also related to high-dimensional factor models for multivariate cross-sectional data or panel data that do not directly apply to the current problem. These models are estimated by minimizing a loss function of both unobserved factors and loading parameters. In other words, although unobserved factors may be regarded as random variables in the model specification, they are conditioned upon and treated as unknown parameters at the estimation stage. In this direction, Stock and Watson (2002) and Bai and Li (2012) considered linear factor analysis and proposed estimation methods based on quadratic and likelihood-based loss functions, respectively. Chen et al. (2020) and Liu et al. (2023a) considered generalized latent factor models that allow for various data types and proposed likelihood-based estimation procedures. Moreover, Chen et al. (2021) introduced a quantile factor model for multivariate data and proposed estimators based on the check loss function for quantile estimation. Although Liu et al. (2023a) and Chen et al. (2021) established some asymptotic normality results, they focused on factor models without covariates, and their results are not directly applicable to the current setting.

To summarize, our main contribution is three-fold: First, this study introduces a novel flexible latent variable modeling framework designed to address the analytical challenges inherent in high-dimensional multivariate longitudinal datasets characterized by heterogeneous outcome variables and incomplete observations—a methodological gap that conventional statistical approaches fail to adequately address. This framework accommodates a range of correlation structures by allowing both time-invariant and time-varying regression parameters and factor loadings, as well as structured forms of time-dependent intercepts. Notably, this modeling flexibility is novel compared to the existing high-dimensional factor models that are not tailored to longitudinal settings. Second, we advance a principled approach for statistical inference concerning covariate effects, accompanied by a thorough exposition of the underlying statistical theory. We specifically provide identifiability

Accepted Manuscript

conditions for latent variables that ensure the parametric estimability of covariate effects, while establishing a central limit theorem that demonstrates the asymptotic properties—namely, consistency, validity, and efficiency—of our proposed inference methodology. Third, our theoretical framework facilitates statistical inference procedures for generalized latent factor models with integrated covariate effects, representing a previously unexamined methodological paradigm within the statistical literature. This contribution thus possesses intrinsic theoretical merit and may stimulate further research in the development of covariate-adjusted latent variable modeling approaches.

The rest of the paper is organized as follows. Section 2 introduces a factor model for high-dimensional longitudinal data and proposes a likelihood-based estimator, along with several extensions and variants. Section 3 establishes the theoretical properties of the proposed estimator. Specifically, a central limit theorem is established for statistical inference on regression coefficients, and an information criterion is introduced for choosing the number of factors. The proposed method is evaluated by simulation studies in Section 4 regarding its finite sample performance and is further applied to a grocery shopping dataset in Section 5 for understanding and predicting customers' shopping behavior. The paper is concluded with discussions in Section 6. A software implementation for R is available at https://github.com/Arthurlee51/LVHML. Further details about the computation and proofs of the theoretical results are given in the online supplementary material.

2 Proposed Method

2.1 Setting and Proposed Model

Consider multivariate longitudinal data with *N* individuals and *J* outcome variables observed on discrete time points $t = 1, ..., T$. Let $Y_i = (y_{ijt})_{j=1,...,J,t=1,...,T}$ be a $J \times T$ data matrix for each individual *i*, where y_{ijt} is a random variable indicating the measurement of the *j*th outcome at time *t*. We further use the vector $\mathbf{y}_{it} = (y_{i1t}, \ldots, y_{iJt})^\top$ to denote all of the individual's outcomes at time *t*. Besides the measured outcomes, a set of *p* covariates are collected for each individual *i*, denoted by $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^\top$. To facilitate clarity of presentation and methodological exposition, we restrict our initial analysis to the setting in which covariates are static; the theoretical extension to accommodate time-varying covariate structures will be comprehensively addressed in Section 2.3.1. Furthermore, to account for missing observations, we let r_{it} be a missing indicator for individual *i* at time *t*, where $r_{it} = 1$ if $\mathbf{y}_{it}$ is observed and $r_{it} = 0$ otherwise. We then partition Y_i into Y_i^o and Y_i^m, where Y_i^o contains those $\mathbf{y}_{it}$ for which $r_{it} = 1$ and Y_i^m contains the remaining components. Let $\mathbf{r}_i = (r_{i1}, \ldots, r_{iT})^\top$ denote the vector of the individual's missing indicators. We observe independent and identically distributed (i.i.d.) copies of the triplet Y_i^o, $\mathbf{r}_i$ and $\mathbf{x}_i$, $i = 1, \ldots, N$.

Within this analytical framework, we develop a high-dimensional factor model designed to achieve two primary methodological objectives. Our first objective involves conducting rigorous statistical inference regarding the covariates' effects on individual outcome variables based on the estimated model parameters derived from the training process. The second objective encompasses the development of predictive capabilities, wherein we leverage the up-to-date information to construct a trained model framework for forecasting future

Accepted Manuscript

realizations of outcome variables y_{ijt} at time point $(T+1)$ across all individuals $i=1,...,N$ and outcome dimensions $j=1,...,J$.

To address the complex dependency structure inherent in longitudinal data, we introduce individual-specific latent variables $\boldsymbol{\theta}_i=(\theta_{i1},...,\theta_{iK})^\top$, commonly referred to as latent factors. These latent constructs serve to model the within-individual correlational patterns and systematic variations that remain unexplained by the observable covariate effects, where *K* represents the predetermined dimensionality of the factor space. The latent dimension *K* is assumed to satisfy the condition where *K* is small relative to *N* and *J*, but can still be large in absolute value, ensuring computational tractability.

Suppose that each of $y_{ijt}, i=1,\ldots,N,\ j=1,\ldots,J, t=1,\ldots,T$, follows an exponential family distribution with natural parameter $\gamma_{jt}+\mathbf{a}_j^\top\boldsymbol{\theta}_i+\boldsymbol{\beta}_j^\top\mathbf{x}_i$, and possibly a scale dispersion parameter ϕ_j. Here γ_{jt}, $\mathbf{a}_j=(a_{j1},\ldots,a_{jK})^\top$ and $\boldsymbol{\beta}_j=(\beta_{j1},\ldots,\beta_{jp})^\top$ are item-specific parameters. Specifically, γ_{jt} is a variable- and time-specific intercept capturing the baseline intensity, $\mathbf{a}_j$ is a vector of the loading parameters, and $\boldsymbol{\beta}_j$ contains the regression coefficients. More precisely, the probability density/mass function for y_{ijt} takes the form

$$f(y_{ijt}\mid \gamma_{jt},\mathbf{a}_j,\boldsymbol{\theta}_i,\boldsymbol{\beta}_j,\mathbf{x}_i,\phi_j) = \exp\left(\frac{y_{ijt}(\gamma_{jt}+\mathbf{a}_j^\top\boldsymbol{\theta}_i+\boldsymbol{\beta}_j^\top\mathbf{x}_i)-b_j(\gamma_{jt}+\mathbf{a}_j^\top\boldsymbol{\theta}_i+\boldsymbol{\beta}_j^\top\mathbf{x}_i)}{\phi_j}+c_j(y_{ijt},\phi_j)\right), \quad (1)$$

where $b_j(\cdot)$ and $c_j(\cdot)$ are pre-specified variable-specific functions that are determined by the choice of the exponential family distribution. This model assumption allows us to model outcome variables of mixed types, including binary, count and continuous data. For example, for a binary variable, (2.1) leads to a logistic model where

$$P(y_{ijt}=1\mid \gamma_{jt},\mathbf{a}_j,\boldsymbol{\theta}_i,\boldsymbol{\beta}_j,\mathbf{x}_i)=\frac{\exp(\gamma_{jt}+\mathbf{a}_j^\top\boldsymbol{\theta}_i+\boldsymbol{\beta}_j^\top\mathbf{x}_i)}{1+\exp(\gamma_{jt}+\mathbf{a}_j^\top\boldsymbol{\theta}_i+\boldsymbol{\beta}_j^\top\mathbf{x}_i)}, \quad (2)$$

for which $b_j(\cdot)=\log(1+\exp(\cdot)), \phi_j=1$ and $c_j(\cdot,\cdot)=0$. For count data, (2.1) gives

$$P(y_{ijt}=y\mid \gamma_{jt},\mathbf{a}_j,\boldsymbol{\theta}_i,\boldsymbol{\beta}_j,\mathbf{x}_i)=\frac{\exp(y(\gamma_{jt}+\mathbf{a}_j^\top\boldsymbol{\theta}_i+\boldsymbol{\beta}_j^\top\mathbf{x}_i)-\exp(\gamma_{jt}+\mathbf{a}_j^\top\boldsymbol{\theta}_i+\boldsymbol{\beta}_j^\top\mathbf{x}_i))}{y!}, \quad (3)$$

a Poisson model for which $b_j(\cdot)=\exp(\cdot), \phi_j=1$ and $c_j(y,\phi_j)=-\log(y!)$.

For each individual *i*, we assume that y_{ijt} s are conditionally independent given the latent variables $\boldsymbol{\theta}_i$ and covariates $\mathbf{x}_i$. Furthermore, we assume the missing outcome variables to be MAR, such that the missing indicator $\mathbf{r}_i$ is conditionally independent of the unobserved data

Accepted Manuscript

Y_i^m given the observed data Y_i^o. We denote $A = (a_{jk})_{J\times K}$ as the loading matrix, $\Theta = (\theta_{ik})_{N\times K}$ as the matrix for factor scores, and $X = (x_{il})_{N\times p}$ as the covariate matrix.

To ensure the identifiability of the regression coefficients $\boldsymbol{\beta}_j$, we impose the restriction

$$\Theta^\top X = 0_{K\times p}, \quad (4)$$

where $0_{K\times p}$ is a $K\times p$ matrix with all the entries being zero. This constraint requires the latent factors to be uncorrelated with the observed covariates, coinciding the assumption in traditional random effects models that random effects and regressors are orthogonal.

We present several theoretical observations regarding the proposed modeling framework. First, it is assumed that the within-individual dependence, both between items and across time, is completely captured by covariates $\mathbf{x}_i$ and low-dimensional factors $\boldsymbol{\theta}_i$. Conditioning on these variables, the outcome variables are independent, and the right-hand side of (2.1) does not depend on the outcomes at other time points. Second, one can regard the latent variables $\boldsymbol{\theta}_i$ as random effects capturing unobserved within-individual heterogeneity. Conventional latent variable modeling frameworks for multivariate longitudinal data characteristically require the specification of restrictive parametric distributional assumptions—most commonly normality—concerning the underlying latent variable structure. Statistical inference procedures in these traditional approaches subsequently rely upon marginal likelihood computations wherein the latent variables are analytically or numerically integrated out from the joint distribution, thereby yielding parameter estimates based solely on the observed data likelihood. While this approach works well for low-dimensional latent variable models, it becomes computationally challenging when the latent dimension becomes moderately large (Chapter 6, Skrondal and Rabe-Hesketh, 2004) and unsuitable for the current setting where *K* can be large. In this work, we adopt an approach commonly applied in high-dimensional factor models (Chen et al., 2020, 2021; Liu et al., 2023a), optimizing an objective function involving both the fixed parameters such as γ_{jt}, $\mathbf{a}_j$ and $\boldsymbol{\beta}_j$ and the latent variables $\boldsymbol{\theta}_i$, without requiring distributional assumption on the latent variables. Third, except for $\boldsymbol{\beta}_j$, the remaining unknown parameters in (2.1) are not identifiable without additional constraints. For example, one can add a constant to each entry of Θ and compensate it by adjusting the intercepts γ_{jt}, without changing the density/probability (2.1). It is important to acknowledge that analogous indeterminacy issues constitute a fundamental characteristic of factor analytic methodologies, and the identifiability of these parameters can be rigorously established through the implementation of appropriate normalization constraints analogous to those proposed in, for instance, Bai and Li (2012). This indeterminacy does not affect making predictions but affects interpretation of the factors and inference of the corresponding loading parameters. As we are mainly interested in drawing statistical inferences on the regression coefficients, we do not impose constraints to fix the rotational indeterminacies. As will be rigorously demonstrated in Section 2.2, the proposed estimator for $\boldsymbol{\beta}_j$ exhibits both consistency and asymptotic normality properties, irrespective of the identifiability status of the remaining model parameters.

Accepted Manuscript

Finally, although the covariates, latent variables, and most of the parameters are assumed to be time-independent in (2.1), we can extend our model to allow them to be time-dependent. Some of such extensions are discussed in Sections 2.3 and 6, respectively. However, we should note that these extensions also introduce more model parameters, which may lead to a higher variance in prediction and additional challenges with interpretations.

2.2 Estimation

We consider the estimation of the proposed model based on the joint log-likelihood function

$$l(\boldsymbol{\Xi}) = \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{t=1}^{T} r_{it}\left\{y_{ijt}(\gamma_{jt} + \mathbf{a}_j^\top \boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top \mathbf{x}_i) - b_j(\gamma_{jt} + \mathbf{a}_j^\top \boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top \mathbf{x}_i)\right\}, \quad (5)$$

where $\boldsymbol{\Xi}$ is a vector containing unknown quantities including $\gamma_{jt}, \mathbf{a}_j, \boldsymbol{\beta}_j$ and $\boldsymbol{\theta}_i$, for $i = 1, \ldots, N, j = 1, \ldots, J, t = 1, \ldots, T$. To estimate the regression coefficients $\boldsymbol{\beta}_j$, we maximize $l(\boldsymbol{\Xi})$ with respect to $\boldsymbol{\Xi}$ under certain compactness constraints on the model parameters. More specifically, let $\|\cdot\|$ denote the Euclidean norm, we solve the optimization problem

$$\begin{aligned} &\boldsymbol{\Xi} = \arg\max_{\boldsymbol{\Xi}} l(\boldsymbol{\Xi}) \text{ such that} \\ &\|\boldsymbol{\theta}_i\| \le c_1\sqrt{K}, \text{ and } \|(\mathbf{a}_j^\top, \boldsymbol{\gamma}_j^\top, \boldsymbol{\beta}_j^\top)^\top\| \le c_2\sqrt{T + p + K}, i = 1, \ldots, N, j = 1, \ldots, J, \end{aligned} \quad (6)$$

where c_1 and c_2 are two constraint parameters, and $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jT})^\top$. Numerically, the compactness constraints prevent parameters from taking extreme values, which may happen when observed variables are discrete and certain categories are rarely observed. Theoretically, this constraint plays a crucial role in establishing the estimation consistency; see Section 3 for the details. This optimization problem is solved by a projected gradient descent algorithm (Chapter 4, Bertsekas, 1999) that is guaranteed to converge to a critical point. See Section A in the online supplementary material for the computational details.

We designate the expression in (5) as a joint log-likelihood function to establish a clear distinction from the marginal log-likelihood formulations adopted in conventional latent variable modeling frameworks. This nomenclature reflects the fundamental characteristic that our proposed log-likelihood function explicitly incorporates both the structural model parameters and the latent factor realizations as estimable quantities, contrasting with traditional approaches that marginalize out the latent variables.

We shall acknowledge that, from a theoretical standpoint, the complete joint log-likelihood function assumes a formulation that differs subtly from the simplified representation $l(\boldsymbol{\Xi})$ presented in (5). Specifically, the complete joint log-likelihood takes the form $\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{t=1}^{T} r_{it} \log f(y_{ijt} | \gamma_{jt}, \mathbf{a}_j, \boldsymbol{\theta}_i, \boldsymbol{\beta}_j, \mathbf{x}_i, \phi_j)$. The discrepancy arises because our formulation $l(\boldsymbol{\Xi})$ omits the scale parameters $\{\phi_j\}_{j=1}^{J}$ from the likelihood specification. Notably, these two formulations are asymptotically equivalent (differing only by an additive constant) when the response variables are exclusively binary or count data that adhere to the Bernoulli or Poisson distributional specifications outlined in (2) and (3), respectively.

Accepted Manuscript

The proposed estimator is suitable when all the scale parameters are close to each other. In scenarios characterized by substantial scale parameter heterogeneity, we recommend using the complete joint log-likelihood function to facilitate simultaneous estimation of both $\boldsymbol{\Xi}$ and the entire collection of scale parameters $\{\phi_j\}_{j=1}^{J}$. The asymptotic theoretical properties established in Section 3 can be appropriately extended to accommodate this more comprehensive estimation framework through straightforward analytical adaptations.

2.3 Model Extensions

2.3.1 Extension to Incorporating Time-dependent Covariates

In scenarios where each individual *i* is associated with a time-dependent covariate vector $\mathbf{z}_{it} = (z_{i1t}, z_{i2t}, \ldots, z_{ip_z t})^\top$ at each time point *t*, with corresponding regression parameters $\mathbf{v}_j = (v_{j1}, \ldots, v_{jp_z})^\top$, our model adapts accordingly. The natural parameter of the exponential family distribution can be modeled as $\gamma_{jt} + \mathbf{a}_j^\top \boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top \mathbf{x}_i + \mathbf{v}_j^\top \mathbf{z}_{it}$. The conditional probability density/mass function $f(y_{ijt} \mid \gamma_{jt}, \mathbf{a}_j, \boldsymbol{\theta}_i, \boldsymbol{\beta}_j, \mathbf{x}_i, \mathbf{v}_j, \mathbf{z}_{it}, \phi_j)$ then becomes

$$\exp\Big(\phi_j^{-1}\{y_{ijt}(\gamma_{jt} + \mathbf{a}_j^\top \boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top \mathbf{x}_i + \mathbf{v}_j^\top \mathbf{z}_{it}) - b_j(\gamma_{jt} + \mathbf{a}_j^\top \boldsymbol{\theta}_i + \boldsymbol{\beta}_j^\top \mathbf{x}_i + \mathbf{v}_j^\top \mathbf{z}_{it})\} + c_j(y_{ijt}, \phi_j)\Big).$$

This modification maintains the structure of the likelihood as in equation (5), and, thus, the estimation algorithm described in Section 2.2 can still be applied. By incorporating additional assumptions for time-dependent covariates, we can derive theorems akin to those established under the current model in Section 3. The specific assumptions and proofs of these theorems are elaborated in Sections C and D in the online supplementary material.

2.3.2 Extension for Time-dependent Loadings and Coefficients

To better accommodate the effects of time in complex datasets, the loadings and the coefficients of the covariates may also be made time-dependent by modelling the natural parameter as $\gamma_{jt} + \mathbf{a}_{jt}^\top \boldsymbol{\theta}_i + \boldsymbol{\beta}_{jt}^\top \mathbf{x}_i$. Similar to the extension discussed in Section 2.3.1, we could also include time-dependent covariate Z_t, and the estimation procedure outlined in Section 2.2 can be adapted to incorporate this extension. Theoretical results analogous to those in Section 3 can be established. The required modifications, assumptions and proofs are detailed in Sections A, C and D of the online supplementary material, respectively.

2.3.3 Imposing Dependence Structure on Intercepts

In practical data analysis, it is often desirable to impose structured dependence on $\boldsymbol{\gamma}_j$ to enhance estimation efficiency and predictive accuracy, or reflect prior knowledge. An important example is $\gamma_{jt} = t\gamma_j$, which fits naturally under the framework developed in Section 2.3.1, treating γ_j as the coefficient for the time-dependent covariate *t*. Consequently, the asymptotic results established for that framework continue to hold.

Accepted Manuscript

This structure can also be incorporated into the extension in Section 2.3.2, with additional assumptions required for valid asymptotic theory. These conditions and related proof adjustments are provided in Sections C.2 and D of the online supplementary material.

3 Theoretical Results

3.1 Consistency and Asymptotic Normality

We now establish the asymptotic properties for the estimated regression coefficients $\boldsymbol{\beta}_j$. Let $\mathbf{u}_j^* = \left(\boldsymbol{\gamma}_j^{*\top}, \boldsymbol{\beta}_j^{*\top}, \mathbf{a}_j^{*\top}\right)^\top$ denote the vector of true values of item-specific parameters. Additionally, define $\mathbf{D}_{it} = (D_{it1}, \ldots, D_{itT})^\top$ as a vector of dummy variables indicating the time periods, where $D_{itt} = 1$ and $D_{itt'} = 0$ for $t \neq t'$, $i = 1 \ldots N$. We further define $\mathbf{e}_{it}^* = (\mathbf{D}_{it}^\top, \mathbf{x}_i^\top, \boldsymbol{\theta}_i^{*\top})^\top$ as the vector of true and observed individual-specific quantities. Let K^* denote the true dimension of the latent variables $\boldsymbol{\theta}_i^*$, and $P = T + p + K^*$ denote the dimension of $\mathbf{u}_j^*$. P is assumed to be fixed that does not vary with N and J. Let $\boldsymbol{\Xi}^* = \left(\mathbf{u}_1^{*\top}, \ldots, \mathbf{u}_J^{*\top}, \boldsymbol{\theta}_1^{*\top}, \ldots, \boldsymbol{\theta}_N^{*\top}\right)^\top$ denote the vector of true parameters. Let $\mathcal{U} \subset \mathbb{R}^P$, $\boldsymbol{\Theta} \subset \mathbb{R}^{K^*}$ and define the space of possible parameters $\mathcal{H}^{K^*} = \left\{\boldsymbol{\Xi} \in \mathbb{R}^{NK^*+PJ} : \mathbf{u}_j \in \mathcal{U}, \boldsymbol{\theta}_i \in \boldsymbol{\Theta} \text{ for all } i, j, \ \Theta^\top X = 0_{K^* \times p}\right\}$. For positive sequences a_n and b_n, we write $a_n \lesssim b_n$ if $a_n \leq C b_n$ for some $C > 0$, and $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$. The following regularity conditions ensure consistency of $\boldsymbol{\beta}_j$.

Assumption 1 .

$\mathcal{U}$ and $\boldsymbol{\Theta}$ are compact sets and $\boldsymbol{\Xi}^* \in \mathcal{H}^{K^*}$. Moreover, $\mathbf{x}_i \in \mathcal{X}$ for all i, where $\mathcal{X} \subset \mathbb{R}^p$ is a compact set.

Assumption 2 .

For any compact set $\mathcal{C} \subset \mathbb{R}$, there exists $\bar{b} > \underline{b} > 0$ (depending on $\mathcal{C}$) such that $\bar{b} \geq b_j''(s) \geq \underline{b}$ and $|b_j'''(s)| \leq \bar{b}$ for all $s \in \mathcal{C}, j = 1, \ldots, J$. Moreover, $\{\phi_j\} \lesssim 1$.

Assumption 3 .

$J^{-1} A^{*\top} A^*$ converges to a positive definite matrix as J tends to infinity. Also, $N^{-1} \Theta^{*\top} \Theta^*$ converge to a positive definite matrix as N tends to infinity.

Assumption 4 .

There exists $\kappa_1 > 0$ such that $\inf_{i=1,\ldots,N, t=1,\ldots,T} P(r_{it} = 1) \geq \kappa_1$.

Accepted Manuscript

Assumption 5 .

There exists $\kappa_2 > 0$ *such that* $\liminf_{N\to\infty} \pi_{\min}\left(\left(X, \mathbf{1_N}\right)^\top \left(X, \mathbf{1_N}\right)\right)/N \geq \kappa_2$, *where* $\pi_{\min}(\cdot)$ *is the minimum eigenvalue of a matrix,* $\mathbf{1_N}$ *is length-N vector of ones.*

The following theorem establishes the consistency of $\boldsymbol{\beta}_j$:

Theorem 1 .

Under Assumptions 1 to 5, $\| \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^* \| = o_P(1)$, *as N and J grow to infinity.*

Define $B^* = (\beta_{jl}^*)_{J\times p}$ and $\boldsymbol{\Gamma}_t^* = (\gamma_{1t}^*, \ldots, \gamma_{Jt}^*)^\top$ for $t = 1, \ldots, T$. Furthermore, let $\Theta = (\hat{\theta}_{ik})_{N\times K^*}, A = (\hat{a}_{jk})_{J\times K^*}$, $B = (\hat{\beta}_{jl})_{J\times p}$ and $\boldsymbol{\Gamma}_t = (\hat{\gamma}_{1t}, \ldots, \hat{\gamma}_{Jt})^\top, t = 1, \ldots, T$ be the estimated parameters from $\boldsymbol{\Xi}$. The following theorem provides the average rate of convergence of $\boldsymbol{\Xi}$ and B:

Theorem 2 .

Under Assumptions 1 to 5, we have

$$\max_{t=1,\ldots,T} \frac{\left\| \hat{\Theta}\hat{A}^\top - \Theta^* A^{*\top} + X(\hat{B} - B^*)^\top + \mathbf{1_N}(\boldsymbol{\Gamma}_t - \boldsymbol{\Gamma}_t^*)^\top \right\|_F}{\sqrt{NJ}} = O_P(min\{\sqrt{N}, \sqrt{J}\}^{-1}), \quad (7)$$

$$\frac{1}{\sqrt{J}} \left\| \hat{B}^\top - B^{*\top} \right\|_F = O_P(min\{\sqrt{N}, \sqrt{J}\}^{-1}). \quad (8)$$

We comment on the rate of convergence for $J^{-1/2} \| \hat{B}^\top - B^{*\top} \|_F$. One might expect a $N^{-1/2}$ rate given that the parameter matrix *B* represents the regression coefficients associated with directly observable covariates. However, our analytical framework necessitates the simultaneous estimation of latent variable components, thereby introducing an additional source of statistical uncertainty that can be seen as measurement error. Specifically, the estimated latent component $\hat{\Theta}\hat{A}^\top$ has an estimation error rate of $(NJ)^{-1/2} \left\| \Theta\hat{A}^\top - \Theta^* A^{*\top} \right\|_F = O_P(min\{\sqrt{N}, \sqrt{J}\}^{-1})$. This measurement error dominates the estimation error of the regression component, resulting in the convergence rate stated in Theorem 2. To establish the asymptotic normality for each $\boldsymbol{\beta}_j$, we need two additional assumptions.

Accepted Manuscript

Assumption 6 .

The limits $\Phi_j = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} -\phi_j^{-1} E\left(r_{it}\right) b_j^{''}(\mathbf{u}_{jt}^{*\top}\mathbf{e}_{it}^{*})\mathbf{e}_{it}^{*}\mathbf{e}_{it}^{*\top}$ *and*

$\Psi_i = \lim_{J\to\infty} \frac{1}{J}\sum_{j=1}^{J}\sum_{t=1}^{T} -\phi_j^{-1} E\left(r_{it}\right) b_j^{''}(\mathbf{u}_{jt}^{*\top}\mathbf{e}_{it}^{*})\mathbf{a}_j^{*}\mathbf{a}_j^{*\top}$ *exist for* $i = 1,\ldots,N$ *and* $j = 1,\ldots,J$ *. Moreover, there exists* $\kappa_3 > 0$ *such that* $\pi_{\min}(\Phi_j^{\top}\Phi_j) \geq \kappa_3$ *and* $\pi_{\min}(\Psi_i^{\top}\Psi_i) \geq \kappa_3$.

Assumption 7 .

As $N, J \to \infty$, $N \asymp J$.

Theorem 3 .

Under Assumptions 1 to 7 , we have $\sqrt{N}\left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^{*}\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{E,j}\right)$, *where the asymptotic variance* $\Sigma_{E,j} = (-\Phi_j^{-1})_{(T+1):(T+p),(T+1):(T+p)}$ *is a submatrix of* $-\Phi_j^{-1}$ *that corresponds to its* $T+1$ *to* $(T+p)$*th rows and columns.* $\Sigma_{E,j}$ *is uniquely determined by the true model without being affected by the indeterminacy of* γ_{jt} , $\mathbf{a}_j$, *and* $\boldsymbol{\theta}_i$.

Theorem 3 establishes the asymptotic normality of $\boldsymbol{\beta}_j$, and specifies the form of the asymptotic variance. This shows that $\boldsymbol{\beta}_j$ is efficient, as its asymptotic variance matches the maximum likelihood estimator in generalized linear model regression, where the latent factors $\boldsymbol{\theta}_i^{*}$ are directly observable. Assumptions 1, 3 and 6 are standard in the literature of factor analysis (see e.g., Bai and Ng, 2002 and Bai, 2003). Assumption 2 concerns the regularity conditions of the exponential family. The condition regarding the derivatives of $b_j(\cdot)$ is straightforward to verify and applies to a wide range of commonly used models under the exponential family, including the logistic model for binary data and Poisson model for count data. The condition for the scale parameter is a mild assumption ensuring the variance of y_{ijt} does not explode. Assumptions 4 and 5 are mild conditions that provide lower bounds for the probability of observing the data and guarantee a degree of variability in the values of *X*, respectively. Assumption 7 ensures that *N* and *T* grow at the same rate. Similar assumptions are made when deriving asymptotic normality for high-dimensional factor models; see e.g., Bai and Li (2012), Galvao and Kato (2016) and Chen et al. (2021).

Remark 1 .

In practice, asymptotic variance is unknown and needs to be estimated. Define $\hat{\mathbf{e}}_{it} = (\mathbf{D}_{it}^{\top}, \mathbf{x}_i^{\top}, \boldsymbol{\theta}_i^{\top})^{\top}$. *We can estimate* Φ_j *by* $\hat{\Phi}_j = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} -\hat{\phi}_j^{-1} r_{it} b_j^{''}(\mathbf{u}_j^{\top}\hat{\mathbf{e}}_{it})\hat{\mathbf{e}}_{it}\hat{\mathbf{e}}_{it}^{\top}$, *and estimate* $\Sigma_{E,j}$ *by the corresponding submatrix. We show in Section D.6 in the online supplementary material that* $\Sigma_{E,j}$ *is a consistent estimator for the true asymptotic variance* $\Sigma_{E,j}$, *where* $\hat{\phi}_j$ *is any consistent estimator of the scale parameter* ϕ_j .

Accepted Manuscript

Remark 2 .

Under the conditions of Theorem 3, the established asymptotic normality for each $j \in \{1, \ldots, J\}$ *implies a uniform convergence rate of* $\max_{1 \le j \le J} \| \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^* \| = O_P\left(N^{-1/2}\sqrt{\log J}\right)$. *This result follows by applying a union bound over* $j = 1, \ldots, J$ *to the sub-Gaussian tail probabilities of* $\hat{\beta}_{jl} - \beta_{jl}^*$ *for* $l = 1, \ldots, p$, *noting that p is assumed fixed.*

3.2 Determining the Number of Factors

In real application, the true number of factors K^* is unknown and, thus, needs to be estimated. To do so, we consider a finite set $\mathcal{K}$, containing the candidate numbers of factors. For each value of $K \in \mathcal{K}$, we estimate the proposed model and obtain the estimate $\boldsymbol{\Xi}_K$ and the corresponding log-likelihood function value, $l(\boldsymbol{\Xi}_K)$. We then construct an information criterion taking the form $\mathrm{IC}(K) = -2l(\boldsymbol{\Xi}_K) + K\Lambda_{NJ}$, where Λ_{NJ} is a penalty term to be discussed in the sequel. We then set

$$\hat{K} = \arg\min_{K \in \mathcal{K}} \mathrm{IC}(K). \quad (9)$$

Theorem 4 .

Suppose that Assumptions 1 to 5 hold and $K^* \in \mathcal{K}$. *If the penalty term* Λ_{NJ} *satisfies* $\max\{N, J\} \lesssim \Lambda_{NJ} \lesssim NJ$, *then* $\lim_{N,J \to \infty} P(\hat{K} = K^*) = 1$.

This result is an extension of an information criterion for a generalized latent factor model proposed by Chen and Li (2022) to the current model. Following the choice in Chen and Li (2022). we set $\Lambda_{NJ} = \max\{N, J\} \times \log\left(\max\{N, J\}^{-1} J \sum_{i=1}^{N}\sum_{t=1}^{T} r_{it}\right)$ in implementation, where $J\sum_{i=1}^{N}\sum_{t=1}^{T} r_{it}$ records the total number of data points being observed. It is easy to see that the requirement on Λ_{NJ} is satisfied with this choice.

4 Simulation Study

4.1 Simulation Setting

We assess the finite sample performance of the proposed method via Monte Carlo simulations under a variety of settings. Specifically, we consider $J = 100, 200, 300, 400$ with $N = 5J$ or $10J$, yielding eight combinations. For each setting, we generate 100 replications with the number of time points $T = 4$ and true latent dimensions $K^* = 3$ and 8. Model selection is performed over the candidate set $\mathcal{K} = \{1, 2, \ldots, 10\}$.

We simulate data in a binary response setting to mimic the real data example. The simulated data follows the logistic model given in (2):

Accepted Manuscript

$$P(y_{ijt}=1|\ \gamma_{jt},\mathbf{a}_j,\boldsymbol{\theta}_i,\boldsymbol{\beta}_j,\mathbf{x}_i)=\frac{\exp(\gamma_{jt}+\sum_{k=1}^{K^*}a_{jk}\theta_{ik}+\sum_{l=1}^{5}\beta_{jl}x_{il})}{1+\exp(\gamma_{jt}+\sum_{k=1}^{K^*}a_{jk}\theta_{ik}+\sum_{l=1}^{5}\beta_{jl}x_{il})}. \quad (10)$$

The variables are generated according to the following procedure, with a minor notational inconsistency arising from the use of identical symbols both prior to and following normalization. Intercepts γ_{jt} are sampled from a uniform distribution $U[-1,1]$, and regression coefficients β_{jl} from $U[0.5,1]$. The latent variables θ_{ik} and a_{jk} are sampled from truncated standard normal distributions on $[-3,3]$. The covariates (x_{i1}, x_{i2}) and (x_{i3}, x_{i4}) are two pairs of dummy variables, each derived independently from a binomial distribution $\text{Bin}(2,0.5)$. The last covariate x_{i5} is sampled from $U[-1,1]$. The normalization procedures described in Section B of the online supplementary material is applied to ensure identifiability of regression coefficients. We then independently set half of each normalized coefficients pair (β_{j1},β_{j2}) and (β_{j3},β_{j4}) to zero, as well as half of the β_{j5} coefficients. The missingness indicator $\mathbf{r}_i$ is sampled from all possible binary combinations of 0 and 1 with equal probability, excluding the all-zero case, resulting in approximately 47% of the values in r_{it} being 0.

4.2 Evaluation Criteria

The performance of the proposed estimator is assessed based on several performance metrics, as given in Table 1. Specifically, in each replication, at the true number of factors K^*, we compute the "Loss" metric defined on the left-hand side of (7) to evaluate the convergence of $\boldsymbol{\Xi}$ in finite sample. Additionally, we compute "Bloss" as defined on the left-hand side of (8) to quantify the convergence of $\hat{B}$. The mean "Loss" and "Bloss" across 100 simulations are reported in Table 1.

To further assess the estimator's performance on individual parameters, the mean squared error (MSE) for each β_{jl}, where $j=1,\ldots,J$ and $l=1,\ldots,4$, is computed across all trials. The maximum of these MSE values is reported as "MMSE". Additionally, the proportion of instances where the correct number of factors is accurately identified is denoted by $P(\hat{K}=K^*)$. The asymptotic variance for each simulation is estimated following the methodology proposed in Remark 1, based on which 95% confidence intervals for β_{jl}s are constructed. The empirical coverage probability (ECP) is then determined by aggregating the coverage probabilities across all parameters and simulation repetitions.

In addition, the recovery of the coefficients $\boldsymbol{\beta}_j$s is evaluated against a baseline approach that assumes no factors, that is, $K=0$. The corresponding likelihood is given by $\prod_{i=1}^{N}\prod_{j=1}^{J}\prod_{t=1}^{T}\left\{\exp((\gamma_{jt}+\boldsymbol{\beta}_j^\top\mathbf{x}_i)y_{ijt})(1+\exp(\gamma_{jt}+\boldsymbol{\beta}_j^\top\mathbf{x}_i))^{-1}\right\}^{r_{it}}$. The optimization is carried out using the `glm` function in R, leveraging a logistic regression (LR) approach. Additionally, we compare the proposed method to a logistic regression model incorporating a random intercept

Accepted Manuscript

α_{ij}, where for each *j*, the α_{ij} are assumed to be identically and independently distributed as normal random variables. The likelihood for this model is $\prod_{i=1}^{N}\prod_{j=1}^{J}\prod_{t=1}^{T}\left\{\exp((\gamma_{jt}+\alpha_{ij}+\boldsymbol{\beta}_j^{\top}\mathbf{x}_i)y_{ijt})(1+\exp(\gamma_{jt}+\alpha_{ij}+\boldsymbol{\beta}_j^{\top}\mathbf{x}_i))^{-1}\right\}^{r_{it}}$. This model is optimized using the `glmer` function from the `lme4` package in R. The results, including the metrics "Bloss" and "MMSE" from both approaches, are reported in Table 1.

Furthermore, when drawing statistical inferences on a large set of regression coefficients, it is imperative to account for multiple testing. We present the average false discovery rate (FDR) across 100 simulation runs, computed using the Benjamini–Yekutieli (BY) procedure (Benjamini and Yekutieli, 2001), which maintains validity under arbitrary dependence structures among the hypotheses. Since x_{i1}, x_{i2} are dummy variables for a single covariate, we test the hypotheses $H_{0j}:\beta_{j1}=\beta_{j2}=0$ for $j=1,\ldots,J$. The Wald test is applied using the estimator for asymptotic variance derived in Theorem 3, as detailed in Remark 1. We reject hypotheses at a significance level of 0.05 based on the BY-adjusted *p*-values. The same procedure is applied to the coefficients associated with x_{i3} and x_{i4}. For the continuous covariate x_{i5}, we test $H_{0j}:\beta_{j5}=0$ for each $j=1,\ldots,J$. For each covariate, we compute the mean FDR (MFDR) across the 100 replications and report the maximum as "MMFDR" in Table 1. In addition, we report the maximum of the mean false non-discovery rates (MMFNR), as the proportion of true alternative hypotheses that are incorrectly not rejected, among all hypotheses not rejected. The results in Table 1 are based on setting the constraint parameters in Equation (6) to $c_1=c_2=5$. To assess sensitivity, we repeat the simulations with $c_1=c_2=c$ for $c\in\{3,4,5,6,7\}$. Results are shown in Table S2 of the online supplementary material.

4.3 Results

The simulation results align with the theory for the proposed method. As *N* and *J* increase, the metrics "Loss", "Bloss", and "MMSE" under the proposed method show a decreasing trend, which occurs regardless of the number of factors and the ratio between *N* and *J*. Moreover, the information criterion for determining the true number of factors K^*, introduced in Section 3.2, proves to be effective. This is evidenced by $P(\hat{K}=K^*)$ achieving 1 in every scenario, which means our method always identifies the correct number of factors. Additionally, as *N* and *J* increase, the empirical coverage probability (ECP) approaches the nominal 95% confidence interval level. This validates our asymptotic normality results in Theorem 3 and the asymptotic variance estimator in Remark 1. Furthermore, the metric "MMFDR" consistently remains below the 0.05 significance threshold across all experimental conditions, thereby confirming the efficacy of the Benjamini–Yekutieli (BY) procedure in controlling the false discovery rate. Additionally, the "MMFNR" metric exhibits a decreasing trend with increasing values of *N* and *J*, indicating that the proposed methodology achieves asymptotically high statistical power.

In comparison, the "Bloss" and "MMSE" metrics under the logistic regression (LR) method are consistently higher than those of the proposed method. In addition, they do not further improve as *N* and *J* increase when they are sufficiently large, suggesting that this simplified model suffers from a large bias. In contrast, the logistic regression model with a random

Accepted Manuscript

intercept (LRRI) outperforms the basic LR model by accounting for the effects of unobserved random intercepts. However, our proposed method continues to demonstrate superior performance as *J* and *N* increase, highlighting the importance of considering correlations among different outcomes to achieve optimal results.

Finally, as demonstrated in Table S2, the proposed estimator exhibits consistent performance across varying selections of the constraint parameter *c*. This stability indicates that the method is robust to the specification of constraint values. Accordingly, we adopt $c_1 = c_2 = 5$ in Section 5.

5 Application to Grocery Shopping Data

5.1 Background

We illustrate the proposed method via an application to a grocery shopping dataset. This dataset encompasses household-level transactions over a span of two years from approximately 2,000 frequent shoppers. It includes purchases made by each household, recorded daily, alongside demographic information such as age groups, household sizes, and income levels for around 800 households. We focus on the subset of customers with demographic information to understand how customers' shopping behavior is associated with their demographic variables and evaluate prediction performance based on latent factors and demographic variables.

In this analysis, daily transaction data are aggregated into 25 four-week periods, using the first $T = 24$ intervals for statistical analysis and model training. The 25th interval is reserved for assessing the predictive performance of our proposed model. We focus on the transactions involving the most popular *J* items during the first *T* intervals, with *N* denoting the count of customers who purchased any of the *J* items within these periods. In each time period *t*, let y_{ijt} be a binary indicator of purchase such that $y_{ijt} = 1$ if individual *i* purchased item *j* and $y_{ijt} = 0$ otherwise. The missing indicator, r_{it}, is set to 0 when the *i*th customer did not purchase any item, including those outside the *J* item list.

We introduce a covariate vector $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})^\top$ capturing household sizes and income levels through dummy variables. Here, $x_{i1} = 1$ indicates two-member households, $x_{i2} = 1$ for three or more members, $x_{i3} = 1$ for incomes between \$35,000 and \$74,999, and $x_{i4} = 1$ for incomes above \$75,000. The baseline level with $x_{i1} = x_{i2} = 0$ for size and $x_{i3} = x_{i4} = 0$ for income represents single-member households earning below \$35,000.

5.2 Statistical Inference

We first focus on inferring the effects of covariates on customers' shopping behavior. In this analysis, we focus on the most popular 100 items, i.e., $J = 100$. The number of observations is $N = 800$. We set the candidate set $\mathcal{K} = \{1, 2, ..., 15\}$ when selecting the number of factors. Using the proposed information criterion, we obtain $\hat{K} = 8$. We perform statistical inference under the eight-factor model.

Accepted Manuscript

We start with an overall significance test for all the covariates to see if any of the covariates are associated with customers' shopping behavior. That is, we test the null hypothesis of $B = 0_{J\times 4}$. We use $\| B \|_F$ as the test statistic and perform a permutation test to obtain its reference distribution under the null hypothesis. Specifically, we perform 500 random permutations indexed by $l = 1,...,500$. In each permutation *l*, we randomly shuffle customers' covariates and then estimate the model parameters. Let the estimate of *B* be denoted by $B^{(l)}$. The reference distribution for the test statistic is then obtained by the empirical distribution for $\| B^{(l)} \|_F$, $l = 1,...,500$. This results in a *p*-value < 0.001, suggesting that these covariates are significantly associated with customers' shopping behavior.

We move on to assess the influences of covariates on individual items. For each item *j*, we calculate the *p*-values associated with the null hypotheses of $\beta_{j1} = \beta_{j2} = 0$ and $\beta_{j3} = \beta_{j4} = 0$, respectively, for all *j*. These hypotheses test the effects of household income and size on the likelihood of purchasing item *j*, respectively. *P*-values are derived through Wald tests, utilizing the estimated coefficients and the asymptotic variance $\Sigma_{E,j}$, as elaborated in Remark 1. To account for multiple testing, we adjust the *p*-values using the BY procedure for FDR control, as discussed in Section 4. This adjustment is carried out separately for the covariates of household income and size, enabling the identification of items significantly associated with each at the predetermined FDR threshold of 5%.

Our analysis examines the influence of household income and size on the purchase patterns of grocery items, where the items are categorized into six groups – Vegetables, Dairy and Eggs, Beverages, Fruits, Bakery and Miscellaneous items. The results are reported in Section E.3 of the online supplementary material. Specifically, Table S6 gives the items selected by the BY procedure for the covariate household size, the corresponding regression coefficients, p-value, BY-adjusted p-value (adj p-value), category, subcategory, average price, and package size. Table S7 is similar to Table S6 but gives the results for household income. Item details absent in the original dataset are marked as NA. The subcategory column represents the lowest level classification available within the dataset, and the price column represents the average unit price derived from all recorded transactions.

We examine the results about household size. As presented in Table S6, the coefficients in columns $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$—particularly those corresponding to $\boldsymbol{\beta}_2$, which denote the coefficients for households comprising three or more individuals—suggest an overall increase in the likelihood of purchasing items compared to the baseline scenario of single-person households. This trend aligns with the expectation that households with a greater number of occupants tend to have higher consumption needs.

We then explore the effect of household income on consumer behavior, as revealed in Table S7. Recall that $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$ are the estimated coefficients for the dummy variables of middle and high household incomes, respectively. Notably, most coefficients in the fruits category are positive, suggesting a heightened health consciousness among these households in comparison to their lower-income counterparts. This hypothesis is consistent with the data in the Beverages category, where most soft drinks are associated with negative coefficients. Although there are exceptions, the vegetable category mostly displays positive coefficients, further reinforcing the trend toward healthier dietary preferences. For the Bakery category, a

Accepted Manuscript

consistent negative trend across coefficients suggests that higher-income households are generally accepted to be less inclined to consume breakfast at home. These observations match our knowledge about how income levels are associated with dietary choices and lifestyle habits (see, e.g. French et al., 2019).

On the other hand, divergent preferences across income levels are observed in the Dairy and Egg category. Specifically, we observe two opposite trends for milk: one subset exhibits positive and increasing coefficients across $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$, signifying a preference among higher-income households, while the other shows negative and diminishing coefficients, indicating the contrary. Notably, price and size do not account for these trends, as evidenced by the table. Further investigation is needed to explore the cause, such as brand differentiation, that drives these preferences. These observations may offer insight for further investigations to explain the differences in preferences uncovered by the current exploratory model.

5.3 Model Comparison and Prediction Performance

Beyond inference for coefficients of covariates, a natural application of such models is for predictions and recommendations. In particular, we can estimate the probabilities for the outcome variables at time $T+1$ given $\boldsymbol{\Xi}$, assuming that the model (2) still holds at $t=T+1$. Due to the absence of an estimate for the time-dependent intercept $\gamma_{j,T+1}$, we substitute $\hat{\gamma}_{j,T}$ in practice. More specifically, we predict the occurrence of outcome variable *j* at time $T+1$ based on the predicted probability $1/(1+\exp(-(\hat{\gamma}_{jT}+\hat{\mathbf{a}}_j^\top\boldsymbol{\theta}_i+\boldsymbol{\beta}_j^\top\mathbf{x}_i)))$. To assess the performance of this approach under varied settings, besides the setting where $J=100$, we also consider $J=200,300$ and 400, with all scenarios having $N=800$.

We compare the proposed method with its variants introduced in Sections 2.3.2 (Prop Sect 2.3.2), 2.3.3 (Prop Sect 2.3.3) and the version combining both extensions (Prop Sect 2.3.2 & 2.3.3), in terms of both in-sample $(t=1,\ldots,T)$ and out-of-sample $(t=T+1)$ fittings. Specifically, for $t=1,\ldots,T+1$, we compute the residual deviance defined as $D_t^{\text{res}}=\sum_{j=1}^{J}D_{jt}^{\text{res}}$, where $D_{jt}^{\text{res}}=\sum_{i=1}^{N}-2r_{it}\{y_{ijt}\log(\hat{p}_{ijt})+(1-y_{ijt})\log\{1-\hat{p}_{ijt}\}\}$, with $\hat{p}_{ijt}$ being the estimated/predicted probability of $y_{ijt}=1$ under each model. As shown in Figure 1, the proposed method (Prop), for which $\hat{K}=8$ when $J=100$ and 200 and $\hat{K}=11$ when $J=300$ and 400, consistently achieves the best fit across almost all time points, both in-sample and out-of-sample. Given this superior performance of Prop, we focus on the results from this method in the rest of this section. Additional results and discussions about the model variants and their comparison with Prop are provided in Section E.3 of the online supplementary material.

Given the nature of the dataset, our evaluations focus on recommendation performance. In particular, we compute the sensitivity, namely, the number of actual purchases in the recommendations divided by the total number of actual purchases. We compare four strategies for making recommendations. Prop ranks recommendations based on the sorted predicted probabilities of the *J* items from corresponding model estimates. Hist ranks recommendations by the sorted cumulative purchasing frequency for each individual,

Accepted Manuscript

resorting to random selection when ties occur. Hist-Prop follows the ranking of Hist but employs sorted predicted probabilities from Prop to resolve ties. Lastly, Hist-Hist, like Hist, ranks recommendations but uses the overall cumulative frequency of items across individuals to break ties. Table 2 displays the results for 10, 20, 30, and 40 recommendations across different values of J for all methods.

We observe that Hist generally outperforms Prop. This is not surprising as there is strong tendency for consumers to purchase the same products repeatedly in grocery shopping data (see e.g. Wan et al., 2018), which is captured effectively by the Hist method. Nevertheless, Hist-Prop emerges as the most proficient approach, indicating that our model is beneficial for improving recommendations, especially when customer information is sparse. This shows the capability of our method to borrow information from similar customers and reflect their preference for previously not purchased products. By offering personalized recommendations, this method outperforms Hist-Hist, which merely suggests the most popular items to individuals when there are insufficient individual history data. Finally, we highlight that it is possible to devise more advanced approaches based on our method to further enhance recommendations performances, especially for suggesting relevant new products to customers. For example, instead of using the proposed method to resolve ties only, we could develop more sophisticated criteria to allocate the proportions of recommendations using individual cumulative frequency and sorted predicted probability, respectively.

6 Discussions

This paper concerns the analysis of high-dimensional multivariate longitudinal data. A flexible modeling framework is proposed to account for between-variable and across-time dependence by latent variables. Statistical inference procedures are developed for parameter estimation and model selection, with statistical consistency and asymptotic normality results established. The method's application to customer grocery shopping records demonstrates its ability to identify demographic influences on purchasing patterns and improve recommendation precision, revealing its value for analytical and predictive uses in practical contexts. In particular, we find a positive association between household size and the likelihood of most purchases, whereas income level is positively associated with the consumption probabilities of healthy food and inversely with soft drinks. Moreover, our model's ability to capture information from other customers' purchase behavior allows improved recommendation performance, when combined with the information from one's purchase history.

The current research may be extended in several directions besides the extensions discussed in Section 2.3. First, the current analysis focuses on the regression coefficients. In many applications, especially in applications of social sciences, the substantive interpretation of the factors may be of interest. Section B of the supplementary material presents normalization criteria that allow identification of the latent factors. These results are further supported by additional simulation studies presented in Section E.1 of the online supplementary material. These identification criteria are not unique; rotation techniques (e.g., Liu et al., 2023b and Rohe and Zeng, 2023) and regularized estimation approaches (e.g., Zhu et al., 2016) may be employed to derive more interpretable factors. A theoretical examination of these method-specific criteria within the framework of our model lies beyond the present study's scope and constitutes a valuable avenue for future research.

Accepted Manuscript

Second, as presented in the extension in Section 2.3.2, which permits the factor loadings $\mathbf{a}_j$ to vary temporally, the static factors $\boldsymbol{\theta}_i$ can be generalized to time-varying factors $\boldsymbol{\theta}_{it}$. This alteration would not significantly change the estimation method but would require adjustments to the normalization criteria and assumptions to ensure the identification of the parameters $\boldsymbol{\beta}_j$, as well as $\mathbf{v}_j$ in the model with time-dependent covariates. In this direction, it is of particular interest to consider a change-point setting that assumes the time-dependent factors $\boldsymbol{\theta}_{it}$ to have a piece-wise constant structure, allowing for individual-specific change points. This model allows us to detect structural changes within each individual, based on which adaptive interventions may be made (e.g., individualized marketing strategies). In addition, by controlling for the maximum number of change points, this change point model enables us to find a balance between model flexibility and parsimony, which leads to high prediction accuracy. Finally, the computational cost for the proposed estimator becomes high or even infeasible when some or all of *N*, *J*, *T* and *p* are large. In such cases, stochastic optimization algorithms may be developed to efficiently obtain approximation solutions, and further, central limit theorems may be established for the approximate solutions to facilitate statistical inference.

Data availability

The data that support the findings of this study are openly available at https://www.dunnhumby.com/source-files.

Supplementary material

The supplementary material includes estimation procedure, normalization algorithm, additional conditions and theorems for extension, technical proofs for main theorems and additional simulation and real data analysis results.

Acknowledgements

The authors would like to acknowledge the editor, the associate editor and the anonymous reviewers whose constructive and valuable comments have substantially improved the manuscript.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

The third author acknowledges the financial supports from RGC-14301920 and RGC-14307221.

Accepted Manuscript

References

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40(1):436–465.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Bertsekas, D. (1999). *Nonlinear Programming*. Athena Scientific.

Bolger, N. and Laurenceau, J.-P. (2013). *Intensive longitudinal methods : An introduction to diary and experience sampling research*. Guilford Press, New York, NY.

Carey, V., Zeger, S. L., and Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80(3):517–526.

Chen, L., Dolado, J. J., and Gonzalo, J. (2021). Quantile factor models. *Econometrica*, 89(2):875–910.

Chen, Y. (2020). A continuous-time dynamic choice measurement model for problem-solving process data. *Psychometrika*, 85(4):1052–1075.

Chen, Y. and Li, X. (2022). Determining the number of factors in high-dimensional generalized latent factor models. *Biometrika*, 109(3):769–782.

Chen, Y., Li, X., Liu, J., and Ying, Z. (2019). Statistical analysis of complex problem-solving process data: An event history analysis approach. *Frontiers in Psychology*, 10:486.

Chen, Y., Li, X., and Zhang, S. (2020). Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*, 115(532):1756–1770.

Cook, R. J. and Lawless, J. F. (2007). *The statistical analysis of recurrent events*. Springer, New York, NY.

French, S. A., Tangney, C. C., Crane, M. M., Wang, Y., and Appelhans, B. M. (2019). Nutrition quality of food purchases varies by household income: the shopper study. *BMC Public Health*, 19:1–7.

Galvao, A. F. and Kato, K. (2016). Smoothed quantile regression for panel data. *Journal of Econometrics*, 193(1):92–112.

Accepted Manuscript

Gray, S. M. and Brookmeyer, R. (2000). Multidimensional longitudinal data: estimating a treatment effect from continuous, discrete, or time-to-event response variables. *Journal of the American Statistical Association*, 95(450):396–406.

Hsieh, C.-A., von Eye, A. A., and Maier, K. S. (2010). Using a multivariate multilevel polytomous item response theory model to study parallel processes of change: The dynamic association between adolescents' social isolation and engagement with delinquent peers in the national youth survey. *Multivariate Behavioral Research*, 45(3):508–552.

Lambert, P. and Vandenhende, F. (2002). A copula-based model for multivariate non-normal longitudinal data: analysis of a dose titration safety study on a new antidepressant. *Statistics in Medicine*, 21(21):3197–3217.

Lian, W., Henao, R., Rao, V., Lucas, J., and Carin, L. (2015). A multitask point process predictive model. In *International Conference on Machine Learning*, pages 2030–2038.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Liang, K.-Y. and Zeger, S. L. (1989). A class of logistic regression models for multivariate binary time series. *Journal of the American Statistical Association*, 84(406):447–451.

Liu, L. C. and Hedeker, D. (2006). A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics*, 62(1):261–268.

Liu, W., Lin, H., Zheng, S., and Liu, J. (2023a). Generalized factor model for ultra-high dimensional correlated variables with mixed types. *Journal of the American Statistical Association*, 118(542):1385–1401.

Liu, X., Wallin, G., Chen, Y., and Moustaki, I. (2023b). Rotation to sparse loadings using L^P losses and related inference problems. *Psychometrika*, 88:527–553.

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer, New York, NY.

Oort, F. J. (2001). Three-mode models for multivariate longitudinal data. *British Journal of Mathematical and Statistical Psychology*, 54(1):49–78.

Ounajim, A., Slaoui, Y., Louis, P.-Y., Billot, M., Frasca, D., and Rigoard, P. (2023). Mixture of longitudinal factor analyzers and their application to the assessment of chronic pain. *Statistics in Medicine*, 42(18):3259–3282.

Panagiotelis, A., Czado, C., and Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, 107(499):1063–1072.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, pages 1033–1048.

Accepted Manuscript

Proust-Lima, C., Amieva, H., and Jacqmin-Gadda, H. (2013). Analysis of multivariate mixed longitudinal data: a flexible latent process approach. *British Journal of Mathematical and Statistical Psychology*, 66(3):470–487.

Rohe, K. and Zeng, M. (2023). Vintage factor analysis with varimax performs statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85:1037–1060.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Smith, M., Min, A., Almeida, C., and Czado, C. (2010). Modeling longitudinal data using a pair-copula decomposition of serial dependence. *Journal of the American Statistical Association*, 105(492):1467–1479.

Sørensen, Ø., Fjell, A. M., and Walhovd, K. B. (2023). Longitudinal modeling of age-dependent latent traits with generalized additive latent and mixed models. *Psychometrika*, 88(2):456–486.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Ten Have, T. R. and Morabia, A. (1999). Mixed effects models with bivariate and univariate association parameters for longitudinal bivariate binary response data. *Biometrics*, 55(1):85–93.

Wan, M., Wang, D., Goldman, M., Taddy, M., Rao, J., Liu, J., Lymberopoulos, D., and McAuley, J. (2017). Modeling consumer preferences and price sensitivities from large-scale grocery shopping transaction logs. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1103–1112.

Wan, M., Wang, D., Liu, J., Bennett, P., and McAuley, J. (2018). Representing and recommending shopping baskets with complementarity, compatibility and loyalty. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1133–1142.

Wang, C., Kohli, N., and Henn, L. (2016). A second-order longitudinal model for binary outcomes: Item response theory versus structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(3):455–465.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., and Campbell, A. T. (2014). Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 3–14.

Zeng, L. and Cook, R. J. (2007). Transition models for multivariate longitudinal binary data. *Journal of the American Statistical Association*, 102(477):211–223.

Accepted Manuscript

Zhang, W., Kuang, Z., Peissig, P., and Page, D. (2020). Adverse drug reaction discovery from electronic health records with deep neural networks. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 30–39.

Zhu, Y., Shen, X., and Ye, C. (2016). Personalized prediction and sparsity pursuit in latent factor models. *Journal of the American Statistical Association*, 111(513):241–252.

Accepted Manuscript

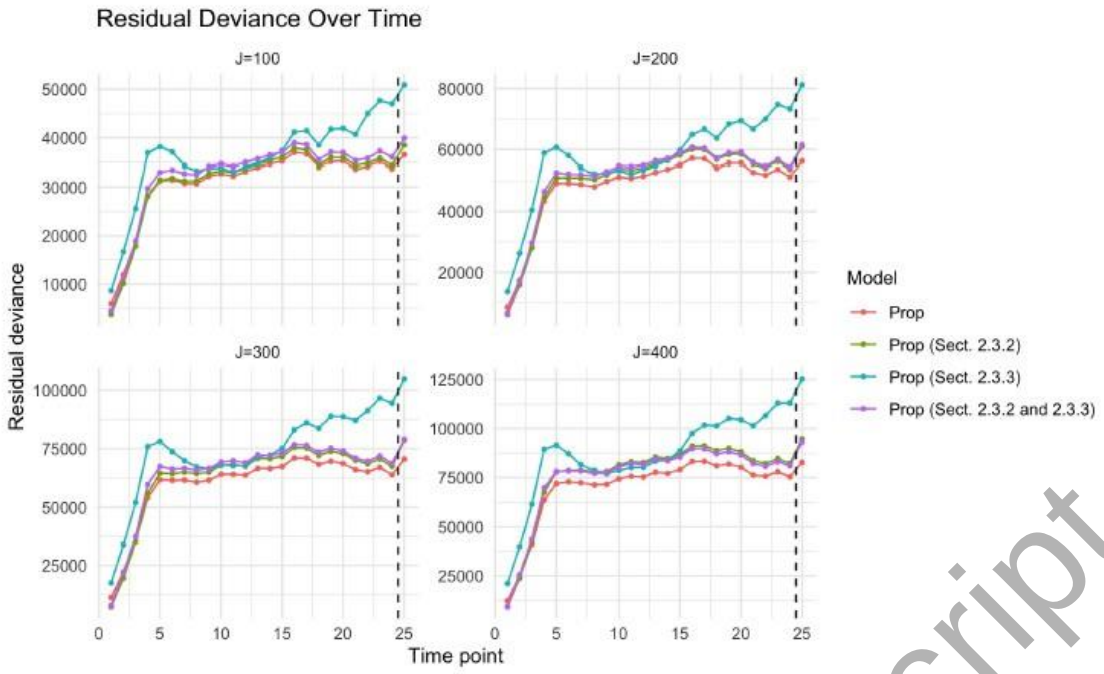

Figure 1: Residual deviances D_t^{res} over time

Accepted Manuscript

Table 1: Summary statistics for the simulation study. The results for the proposed method, logistic regression (LR), and logistic regression with a random intercept (LRRI) across different combinations of N, K^* and J are reported.

		Proposed							**LR**		**LRRI**	
N	J	Loss	$P(\hat{K}=K^*)$	ECP	MMFDR	MMFNR	Bloss	MMSE	Bloss	MMSE	Bloss	MMSE
$K^*=3$												
5J	100	0.55	1	0.94	0.01	0.20	0.49	0.14	0.545	0.512	0.485	0.137
5J	200	0.36	1	0.95	0.01	0.05	0.32	0.05	0.453	0.426	0.338	0.099
5J	300	0.29	1	0.95	0.00	0.00	0.25	0.03	0.427	0.407	0.268	0.104
5J	400	0.24	1	0.95	0.00	0.02	0.22	0.03	0.398	0.328	0.237	0.056
10J	100	0.48	1	0.94	0.01	0.01	0.33	0.06	0.469	0.400	0.333	0.063
10J	200	0.32	1	0.95	0.01	0.00	0.22	0.03	0.409	0.380	0.233	0.047
10J	300	0.26	1	0.95	0.00	0.00	0.18	0.01	0.399	0.388	0.194	0.030
10J	400	0.22	1	0.95	0.00	0.00	0.16	0.01	0.418	0.379	0.185	0.059
$K^*=8$												
5J	100	1.26	1	0.91	0.02	0.24	0.64	0.27	0.700	0.869	0.614	0.371
5J	200	0.68	1	0.94	0.01	0.10	0.39	0.09	0.671	0.787	0.432	0.245
5J	300	0.52	1	0.94	0.01	0.07	0.31	0.06	0.642	1.059	0.366	0.403
5J	400	0.44	1	0.94	0.01	0.02	0.26	0.04	0.629	0.788	0.317	0.434
10J	100	1.12	1	0.91	0.02	0.11	0.45	0.15	0.672	0.955	0.482	1.119
10J	200	0.63	1	0.94	0.01	0.03	0.28	0.05	0.636	0.653	0.319	0.168
10J	300	0.48	1	0.94	0.01	0.00	0.22	0.03	0.613	0.582	0.260	0.113
10J	400	0.41	1	0.94	0.01	0.00	0.18	0.02	0.592	0.492	0.238	0.123

Loss: Frobenius loss measuring the convergence of $\hat{\Xi}$.
P($\hat{K}=K^*$): Proportion of instances where the correct number of factors is identified.
ECP: Empirical coverage probability of the confidence intervals.
MMFDR: Maximum mean false discovery rate across all covariates.
MMFNR: Maximum mean false non-discovery rate across all covariates.
Bloss: Frobenius loss measuring convergence of $\hat{B}$.
MMSE: Maximum mean squared error across all estimated β_{jl} s.

Accepted Manuscript

Table 2: Sensitivity Based on Number of Recommendations

	10 recommendations				20 recommendations			
	J=100	J=200	J=300	J=400	J=100	J=200	J=300	J=400
Hist	0.448	0.348	0.297	0.264	0.652	0.518	0.447	0.404
Prop	0.352	0.256	0.223	0.199	0.533	0.394	0.340	0.304
Hist-Hist	0.451	0.350	0.299	0.266	0.654	0.519	0.450	0.406
Hist-Prop	0.456	0.352	0.301	0.269	0.659	0.525	0.453	0.407
	30 recommendations				40 recommendations			
	J=100	J=200	J=300	J=400	J=100	J=200	J=300	J=400
Hist	0.774	0.627	0.546	0.496	0.848	0.705	0.618	0.565
Prop	0.658	0.490	0.425	0.382	0.752	0.570	0.496	0.447
Hist-Hist	0.775	0.629	0.549	0.499	0.852	0.709	0.621	0.570
Hist-Prop	0.781	0.635	0.555	0.505	0.860	0.714	0.626	0.575

Accepted Manuscript