Variational Bayesian Inference for Mixed Logit Models with Unobserved Inter- and Intra-Individual Heterogeneity

Rico Krueger

Prateek Bansal

Michel Bierlaire

Ricardo A. Daziano

Taha H. Rashidi

UNSW Australia, Cornell University, EPFL **April 2019**

STRC | 19th Swiss Transport Research Conference
Monte Verità / Ascona, May 15 – 17, 2019

UNSW Australia, Cornell University, EPFL

Variational Bayesian Inference for Mixed Logit Models with Unobserved Inter- and Intra-Individual Heterogeneity

Rico Krueger, Taha H. Rashidi
Research Centre for Integrated Transport Innovation, School of Civil and Environmental Engineering
UNSW Australia
Sydney NSW 2052, Australia
phone: NA
fax: NA
r.krueger@student.unsw.edu.au

Prateek Bansal, Ricardo A. Daziano
School of Civil and Environmental Engineering
Cornell University
Ithaca NY, United States
phone: NA
fax: NA
pb422@cornell.edu

Michel Bierlaire
Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
Station 18, Lausanne 1015, Switzerland
phone: NA
fax: NA
michel.bierlaire@epfl.ch

April 2019

Abstract

Variational Bayes (VB) methods have emerged as a fast and computationally-efficient alternative to Markov chain Monte Carlo (MCMC) methods for Bayesian estimation of mixed logit models. In this paper, we derive a VB method for posterior inference in mixed multinomial logit models with unobserved inter- and intra-individual heterogeneity. The proposed VB method is benchmarked against MCMC in a simulation study. The results suggest that VB is substantially faster than MCMC but also noticeably less accurate, because the mean-field assumption of VB is too restrictive. Future research should thus focus on enhancing the expressiveness and flexibility of the variational approximation.

Keywords

Keywords: mixed logit, inter- and intra-individual heterogeneity, Bayesian estimation

1 Introduction

The representation of taste heterogeneity is a principal concern of discrete choice analysis, as information on the distribution of tastes is critical for demand forecasting, welfare analysis and market segmentation. From the analyst's perspective, taste variation is often random, as differences in sensitivities cannot be related to observed or observable characteristics of the decision-maker or features of the choice context.

Mixed random utility models such as mixed logit (McFadden and Train, 2000) provide a powerful framework to account for unobserved taste heterogeneity in discrete choice models. When longitudinal choice data are analysed with the help of random utility models, it is standard practice to assume that tastes vary randomly across decision-makers but not across replications for the same individual (Revelt and Train, 1998). The implicit assumption underlying this treatment of unobserved heterogeneity is that tastes are unique and stable (Stigler and Becker, 1977). Contrasting views of preference formation postulate that preferences are constructed in an ad-hoc manner at the moment of choice (Bettman *et al.*, 1998) or learnt and discovered through experience (Kivetz *et al.*, 2008).

From the perspective of discrete choice analysis, these alternative views of preference formation justify accounting for both inter- and intra- individual heterogeneity (also see Hess and Giergiczny, 2015). A straightforward way to accommodate unobserved inter- and intra-individual heterogeneity in mixed random utility models is to augment a mixed logit model with a multivariate normal mixing distribution in a hierarchical fashion such that case-specific parameters are generated as normal perturbations around the individual-specific parameters (Hess and Rose, 2009, Hess and Train, 2011).

Mixed logit models with unobserved inter- and intra-individual heterogeneity can be estimated with the help of maximum simulated likelihood methods (Hess and Rose, 2009, Hess and Train, 2011). However, this estimation strategy is computationally expensive, as it involves the simulation of iterated integrals. Becker *et al.* (2018) propose a MCMC method, which builds on the Allenby-Train procedure (Train, 2009) for mixed logit models with only inter-individual heterogeneity. While MCMC methods constitute a powerful framework to perform posterior inference in complex probabilistic models (e.g. Gelman *et al.*, 2013), MCMC methods are subject to several bottlenecks, which inhibit their scalability to large datasets, namely i) long computation times, ii) high storage costs for the posterior draws, iii) difficulties in assessing convergence.

Variational Bayes methods (Blei *et al.,* 2017, Jordan *et al.,* 1999, Ormerod and Wand, 2010) have emerged as a fast and computationally-efficient alternative to MCMC methods for posterior inference in discrete choice models. VB addresses the shortcomings of MCMC by re-casting Bayesian inference into an optimisation problem in lieu of a sampling problem. Several studies derive and assess VB methods for mixed logit models with only inter-individual heterogeneity (Bansal *et al.,* 2019, Braun and McAuliffe, 2010, Depraetere and Vandebroek, 2017, Tan, 2017). These studies establish that VB is substantially faster than MCMC at practically no compromises in predictive accuracy.

Motivated by these recent advances in Bayesian estimation of discrete choice models, this current paper has two objectives: First, we derive a VB method for posterior inference in mixed logit models with unobserved inter- and intra-individual heterogeneity. Second, we benchmark the VB method against MCMC in a simulation study.

We organise the remainder of this paper as follows. First, we give the formulation of a mixed logit model with unobserved inter- and intra-individual heterogeneity. Then, we derive the VB method for this model and benchmark the performance of this method against MCMC in a simulation study. Finally, we conclude.

2 Model formulation

The mixed logit (MXL) model with unobserved inter- and intra-individual heterogeneity (Hess and Rose, 2009, Hess and Train, 2011) is established as follows: On choice occasion $t \in \{1, \dots T_n\}$, a decision-maker $n \in \{1, \dots N\}$ derives utility $U_{ntj} = V(\boldsymbol{X}_{ntj}, \boldsymbol{\beta}_{nt}) + \epsilon_{ntj}$ from alternative j in the set C_{nt}. Here, $V()$ denotes the representative utility, $\boldsymbol{X}_{ntj}$ is a row-vector of covariates, $\boldsymbol{\beta}_{nt}$ is a collection of taste parameters, and ϵ_{ntj} is a stochastic disturbance. The assumption $\epsilon_{ntj} \sim$ Gumbel(0, 1) leads to a multinomial logit (MNL) kernel such that the probability that decision-maker n chooses alternative $j \in C_{nt}$ on choice occasion t is

$$P(y_{nt} = j | \boldsymbol{X}_{ntj}, \boldsymbol{\beta}_{nt},) = \frac{\exp\left\{V(\boldsymbol{X}_{ntj}, \boldsymbol{\beta}_{nt})\right\}}{\sum_{k \in C_{nt}} \exp\left\{V(\boldsymbol{X}_{ntk}, \boldsymbol{\beta}_{nt})\right\}}, \tag{1}$$

where $y_{nt} \in C_{nt}$ captures the observed choice.

Note that the taste parameters $\boldsymbol{\beta}_{nt}$ are specified as being observation-specific. To allow for dependence between replications for the same individual and to accomodate inter-individual taste heterogeneity, it has become standard practice to adopt Revelt's and Train's (1998) panel estimator for the mixed logit model. Under this specification, taste homogeneity across replications is assumed such that $\boldsymbol{\beta}_{n,t} = \boldsymbol{\beta}_n \; \forall \; t = 1, \ldots, T_n$. To accomodate intra-individual taste heterogeneity in addition to inter-individual taste heterogeneity, the taste vector $\boldsymbol{\beta}_{n,t}$ can be defined as a normal perturbation around an individual-specific parameter $\boldsymbol{\mu}_n$, i.e. $\boldsymbol{\beta}_{n,t} \sim \mathrm{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W) \; t = 1, \ldots, T_n$, where $\boldsymbol{\Sigma}_B$ is a covariance matrix. The distribution of individual-specific parameters $\boldsymbol{\mu}_{1:N}$ is also assumed to be multivariate normal, i.e. $\boldsymbol{\mu}_n \sim \mathrm{N}(\boldsymbol{\zeta}, \boldsymbol{\Sigma}_B)$ for $n = 1, \ldots, N$, where $\boldsymbol{\zeta}$ is a mean vector and $\boldsymbol{\Sigma}_B$ is a covariance matrix.

In a fully Bayesian setup, the parameters $\boldsymbol{\zeta}$, $\boldsymbol{\Sigma}_B$, $\boldsymbol{\Sigma}_W$ are also considered to be random parameters and are thus given priors. We use a normal prior for mean vector $\boldsymbol{\zeta}$. Following Tan (2017) and Akinc and Vandebroek (2018), we employ Huang's half-t prior (Huang and Wand, 2013) for the covariance matrices $\boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_W$, as this prior specification exhibits superior noninformativity properties compared to other prior specifications for covariance matrices (Huang and Wand, 2013, Akinc and Vandebroek, 2018). In particular, (Akinc and Vandebroek, 2018) show that Huang's half-t prior (Huang and Wand, 2013) outperforms the inverse Wishart prior, which is often employed in fully Bayesian specifications of MMNL models (e.g. Train, 2009), in terms of parameter recovery.

Stated succinctly, the generative process of mixed logit model with inter- and intra-individual heterogeneity is as follows:

$$a_{B,k}|A_{B,k} \sim \mathrm{Gamma}\left(\frac{1}{2}, \frac{1}{A_{B,k}^2}\right), k = 1, \ldots, K, \tag{2}$$

$$a_{W,k}|A_{W,k} \sim \mathrm{Gamma}\left(\frac{1}{2}, \frac{1}{A_{W,k}^2}\right), k = 1, \ldots, K, \tag{3}$$

$$\boldsymbol{\Sigma}_B|\nu_B, \boldsymbol{a}_B \sim \mathrm{IW}\left(\nu_B + K - 1, 2\nu_B \mathrm{diag}(\boldsymbol{a}_B)\right), \quad \boldsymbol{a}_B = \begin{bmatrix} a_{B,1} & \ldots & a_{B,K} \end{bmatrix}^\top \tag{4}$$

$$\boldsymbol{\Sigma}_W|\nu_W, \boldsymbol{a}_W \sim \mathrm{IW}\left(\nu_W + K - 1, 2\nu_W \mathrm{diag}(\boldsymbol{a}_W)\right), \quad \boldsymbol{a}_W = \begin{bmatrix} a_{W,1} & \ldots & a_{W,K} \end{bmatrix}^\top \tag{5}$$

$$\boldsymbol{\zeta}|\boldsymbol{\xi}_0, \boldsymbol{\Xi}_0 \sim \mathrm{N}(\boldsymbol{\xi}_0, \boldsymbol{\Xi}_0) \tag{6}$$

$$\boldsymbol{\mu}_n|\boldsymbol{\zeta}, \boldsymbol{\Sigma}_B \sim \mathrm{N}(\boldsymbol{\zeta}, \boldsymbol{\Sigma}_B), n = 1, \ldots, N, \tag{7}$$

$$\boldsymbol{\beta}_{nt}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W \sim \mathrm{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W), n = 1, \ldots, N, t = 1, \ldots, T_n, \tag{8}$$

$$y_{nt}|\boldsymbol{\beta}_{nt}, \boldsymbol{X}_{nt} \sim \mathrm{MNL}(\boldsymbol{\beta}_{nt}, \boldsymbol{X}_{nt}), n = 1, \ldots, N, \; t = 1, \ldots, T_n, \tag{9}$$

where $\{\boldsymbol{\xi}_0, \boldsymbol{\Xi}_0, \nu_B, \nu_W, A_{B,1:K}, A_{W,1:K}\}$ are known hyper-parameters, and $\boldsymbol{\theta} = \{\boldsymbol{a}_B, \boldsymbol{a}_W, \boldsymbol{\Sigma}_B, \boldsymbol{\Sigma}_W, \boldsymbol{\zeta}, \boldsymbol{\mu}_{1:N}, \boldsymbol{\beta}_{1:N,1:T_n}\}$ is a collection of model parameters whose posterior distribution we wish to estimate.

The generative process implies the following joint distribution of data and model parameters:

$$P(\boldsymbol{y}_{1:N}, \boldsymbol{\theta}) = \prod_{n=1}^{N} \prod_{t=1}^{T_n} P(\boldsymbol{y}_{nt}|\boldsymbol{\beta}_{nt}, \boldsymbol{X}_{nt}) P(\boldsymbol{\beta}_{nt}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W) \prod_{n=1}^{N} P(\boldsymbol{\mu}_n|\boldsymbol{\zeta}, \boldsymbol{\Sigma}_B) \dots$$
$$\dots P(\boldsymbol{\zeta}|\boldsymbol{\xi}_0, \boldsymbol{\Xi}_0) P(\boldsymbol{\Sigma}_B|\omega_B, \boldsymbol{B}_B) P(\boldsymbol{\Sigma}_W|\omega_W, \boldsymbol{B}_W) \prod_{k=1}^{K} P(a_{W,k}|s, r_{W,k}) P(a_{B,k}|s, r_{B,k}) \tag{10}$$

where $\omega_B = \nu_B + K - 1$, $\boldsymbol{B}_B = 2\nu_B \text{diag}(\boldsymbol{a}_B)$, $\omega_W = \nu_W + K - 1$, $\boldsymbol{B}_W = 2\nu_W \text{diag}(\boldsymbol{a}_W)$, $s = \frac{1}{2}$, $r_{B,k} = A_{B,k}^{-2}$ and $r_{W,k} = A_{W,k}^{-2}$.[1] By Bayes' rule, the posterior distribution of interest is then given by

$$P(\boldsymbol{\theta}|\boldsymbol{y}_{1:N}) = \frac{P(\boldsymbol{y}_{1:N}, \boldsymbol{\theta})}{\int P(\boldsymbol{y}_{1:N}, \boldsymbol{\theta}) d\boldsymbol{\theta}} \propto P(\boldsymbol{y}_{1:N}, \boldsymbol{\theta}). \tag{11}$$

Exact inference of this posterior distribution is not possible, because the model evidence $\int P(\boldsymbol{y}_{1:N}, \boldsymbol{\theta}) d\boldsymbol{\theta}$ is not tractable. Becker *et al.* (2018) propose a Gibbs sampler for posterior inference in the described model. While this method has been shown to perform reasonably well, it is subject to the known limitations of MCMC. In the subsequent section, we derive a VB method for scalable inference in mixed logit with unobserved inter- and intra-individual heterogeneity. For completeness, the Gibbs sampler proposed by Becker *et al.* (2018) is given in Appendix A.

3 Variational Bayes estimation

3.1 Background

Variational Bayesian inference (e.g. Blei *et al.*, 2017, Jordan *et al.*, 1999, Ormerod and Wand, 2010) differs from MCMC in that approximate Bayesian inference is viewed as an optimization problem rather than a sampling problem. To describe the fundamental principles of mean-field variational Bayes, we consider a generative model $P(\boldsymbol{y}, \boldsymbol{\theta})$ consisting of observed data $\boldsymbol{y}$ and unknown parameters $\boldsymbol{\theta}$. Our goal is to learn the posterior distribution of $\boldsymbol{\theta}$, i.e. $P(\boldsymbol{\theta}|\boldsymbol{y})$. Variational Bayesian inference aims at finding a variational distribution $q(\boldsymbol{\theta})$ over the unknown parameters

[1] To be clear, the following forms of the Gamma and inverse Wishart distributions are considered:

$$P(a_k|s, r_k) \propto a_k^{s-1} \exp(-r_k a_k),$$
$$P(\boldsymbol{\Omega}|\omega, \boldsymbol{B}) \propto |\boldsymbol{B}|^{\frac{\omega}{2}} |\boldsymbol{\Omega}|^{-\frac{\omega+K+1}{2}} \exp\left(-\frac{1}{2}\text{tr}\left(\boldsymbol{B}\boldsymbol{\Omega}^{-1}\right)\right),$$

whereby $\boldsymbol{\Omega}$ and $\boldsymbol{B}$ are $K \times K$ positive-definite matrices.

that is close to the exact posterior distribution. A computationally-convenient way to measure the distance between two probability distributions is the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). The KL divergence between $q(\boldsymbol{\theta})$ and $P(\boldsymbol{\theta}|\boldsymbol{y})$ is given by

$$\text{KL}\left(q(\boldsymbol{\theta})||P(\boldsymbol{\theta}|\boldsymbol{y})\right) = \int \ln\left(\frac{q(\boldsymbol{\theta})}{P(\boldsymbol{\theta}|\boldsymbol{y})}\right) q(\boldsymbol{\theta}) dx = \mathbb{E}_q\left\{\ln q(\boldsymbol{\theta})\right\} - \mathbb{E}_q\left\{\ln P(\boldsymbol{\theta}|\boldsymbol{y})\right\}. \tag{12}$$

Consequently, the goal of variational inference is to solve

$$q^*(\boldsymbol{\theta}) = \arg\min_q \left\{\text{KL}\left(q(\boldsymbol{\theta})||P(\boldsymbol{\theta}|\boldsymbol{y})\right)\right\}. \tag{13}$$

Note that $P(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{P(\boldsymbol{y},\boldsymbol{\theta})}{P(\boldsymbol{y})}$. Hence,

$$\begin{aligned} \text{KL}\left(q(\boldsymbol{\theta})||P(\boldsymbol{y},\boldsymbol{\theta})\right) &= \text{KL}\left(q(\boldsymbol{\theta})||P(\boldsymbol{\theta}|\boldsymbol{y})\right) - \ln P(\boldsymbol{y}) \\ &= \mathbb{E}_q\left\{\ln q(\boldsymbol{\theta})\right\} - \mathbb{E}_q\left\{\ln P(\boldsymbol{y},\boldsymbol{\theta})\right\} \end{aligned} \tag{14}$$

The term $\mathbb{E}_q\left\{\ln P(\boldsymbol{y},\boldsymbol{\theta})\right\} - \mathbb{E}_q\left\{\ln q(\boldsymbol{\theta})\right\}$ is referred to as the evidence lower bound (ELBO). Thus, minimizing the KL divergence between the approximate variational distribution and the intractable exact posterior distribution is equivalent to maximizing the ELBO. The goal of VB can therefore be re-formulated as follows:

$$\begin{aligned} q^*(\boldsymbol{\theta}) &= \arg\max_q \left\{\text{ELBO}(q)\right\} \\ &= \arg\max_q \left\{\mathbb{E}_q\left\{\ln P(\boldsymbol{y},\boldsymbol{\theta})\right\} - \mathbb{E}_q\left\{\ln q(\boldsymbol{\theta})\right\}\right\}. \end{aligned} \tag{15}$$

The functional form of the variational distribution $q(\boldsymbol{\theta})$ remains to be chosen. We can appeal to the mean-field family of distributions (e.g. Jordan *et al.,* 1999), under which the variational distribution factorizes as $q(\theta_{1:M}) = \prod_{m=1}^{M} q(\theta_m)$, where $m \in \{1,\ldots,M\}$ indexes the model parameters. The mean-field assumption breaks the dependence between the model parameters by imposing mutual independence of the variational factors. It can be shown that the optimal density of each variational factor is given by $q^*(\theta_M) \propto \exp \mathbb{E}_{-\theta_m}\left\{\ln P(\boldsymbol{y},\boldsymbol{\theta})\right\}$, i.e. the optimal density of each variational factor is proportional to the exponentiated expectation of the logarithm of the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{\theta}$, where the expectation is taken with respect to all parameters other than θ_m (Ormerod and Wand, 2010, Blei *et al.,* 2017). Provided that the model of interest is conditionally conjugate, the optimal densities of all variational factors belong to recognizable families of distributions (Blei *et al.,* 2017). Due to the implicit nature of the expectation operator $\mathbb{E}_{-\theta_m}$, the ELBO can then be maximized using an iterative coordinate ascent algorithm (Bishop, 2006), in which the variational factors are updated one at a time conditional on the current estimates of the other variational factors. Iterative updates with respect to each variational factor are performed by equating each of the variational factors to its respective optimal density, i.e. we set $q(\theta_m) = q^*(\theta_m)$ for $m = 1, \ldots, M$.

3.2 Variational Bayes for mixed logit with unobserved inter- and intra-individual heterogeneity

In the present application, we are interested in approximating the posterior distribution of the model parameters $\{\boldsymbol{a}_B, \boldsymbol{a}_W, \boldsymbol{\Sigma}_B, \boldsymbol{\Sigma}_W, \boldsymbol{\zeta}, \boldsymbol{\mu}_{1:N}, \boldsymbol{\beta}_{1:N,1:T_n}\}$ (see expression 11) through a fitted variational distribution. We posit a variational distribution from the mean-field family, i.e. the variational distribution factorises as follows:

$$q(\boldsymbol{\theta}) = \prod_{k=1}^{K} q(a_{B,k}) q(a_{W,k}) q(\boldsymbol{\Sigma}_B) q(\boldsymbol{\Sigma}_W) q(\boldsymbol{\zeta}) \prod_{n=1}^{N} q(\boldsymbol{\mu}_n) \prod_{n=1}^{N} \prod_{t=1}^{T_n} q(\boldsymbol{\beta}_{n,t}). \tag{16}$$

Recall that the optimal densities of the variational factors are given by $q^*(\theta_i) \propto \exp \mathbb{E}_{-\theta_i} \{\ln P(\boldsymbol{y}, \boldsymbol{\theta})\}$. We find that $q^*(a_{B,k}|c_B, d_{B,k})$, $q^*(a_{W,k}|c_W, d_{W,k})$, $q^*(\boldsymbol{\Sigma}_B|w_B, \boldsymbol{\Theta}_B)$, $q^*(\boldsymbol{\Sigma}_W|w_W, \boldsymbol{\Theta}_W)$, $q^*(\boldsymbol{\zeta}|\boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta)$, and $q^*(\boldsymbol{\mu}_n|\boldsymbol{\mu}_{\mu_n}, \boldsymbol{\Sigma}_{\mu_n})$ are common probability distributions (see Appendix C). However, $q^*(\boldsymbol{\beta}_{nt})$ is not a member of recognizable family of distributions, because the MNL kernel does not have a general conjugate prior. For simplicity and computational convenience, we assume that $q(\boldsymbol{\beta}_{nt}) = \text{Normal}(\boldsymbol{\mu}_{\beta_{nt}}, \boldsymbol{\Sigma}_{\beta_{nt}})$ for all $n \in \{1, \ldots, N\}$,$t \in \{1, \ldots, T_n\}$.

The ELBO is maximized using an iterative coordinate ascent algorithm. Iterative updates of $q(a_{B,k})$, $q(a_{W,k})$, $q(\boldsymbol{\Sigma}_B)$, $q(\boldsymbol{\Sigma}_W)$, $q(\boldsymbol{\zeta})$, and $q(\boldsymbol{\mu}_n)$ are performed by equating each variational factor to its respective optimal distribution $q^*(a_{B,k})$, $q^*(a_{W,k})$, $q^*(\boldsymbol{\Sigma}_B)$, $q^*(\boldsymbol{\Sigma}_W)$, $q^*(\boldsymbol{\zeta})$, and $q^*(\boldsymbol{\mu}_n)$, respectively. Then, updates for the nonconjugate variational factor $q(\boldsymbol{\beta}_{nt})$ are performed with the help of either quasi-Newton (QN) methods (e.g. Nocedal and Wright, 2006) or nonconjugate variational message passing (NCVMP; Knowles and Minka, 2011). Whereas updates for nonconjugate variational factors are obtained by maximizing the ELBO over the parameters of the variational factor in QN methods, NCVMP translates this optimization problem into fixed point updates:

$$\begin{aligned} \boldsymbol{\Sigma}_{\beta_{nt}} &= -\left[2\ vec^{-1}\left\{\nabla_{vec(\boldsymbol{\Sigma}_{\beta_{nt}})}\left(\mathbb{E}_q\left(\ln(P(\boldsymbol{y}_{1:N}, \boldsymbol{\theta}))\right)\right)\right\}\right]^{-1} \\ \boldsymbol{\mu}_{\beta_{nt}} &= \boldsymbol{\mu}_{\beta_{nt}} + \boldsymbol{\Sigma}_{\beta_{nt}}\left[\nabla_{\boldsymbol{\mu}_{\beta_{nt}}}\left(\mathbb{E}_q\left(\ln(P(\boldsymbol{y}_{1:N}, \boldsymbol{\theta}))\right)\right)\right] \end{aligned} \tag{17}$$

These updates involve $\mathbb{E}_q\left(\ln(P(\boldsymbol{y}_{1:N}, \boldsymbol{\theta}))\right)$ which does not have a closed-form expression due to intractable expectation of the logsum of exponentials (E-LSE) term $g_{nt} = \ln\left[\sum_{k \in C_{nt}} \exp(\boldsymbol{X}_{ntk}\boldsymbol{\beta}_{nt})\right]$. After approximating, E-LSE using the delta method in Appendix B (Tan, 2017), we derive the required gradients in Appendix C.5. Algorithm 1 succinctly summarises the proposed VB method for posterior inference in MXL models with unobserved inter- and intra-individual heterogeneity.

Initialization:
Set hyper-parameters: $\boldsymbol{\xi}_0$, $\boldsymbol{\Xi}_0$, ν_B, ν_W, $A_{B,1:K}$, $A_{W,1:K}$;
Provide starting values: $\boldsymbol{\mu}_\zeta$, $\boldsymbol{\Sigma}_\zeta$, $\boldsymbol{\mu}_{\beta_{1:N,1:T_n}}$, $\boldsymbol{\Sigma}_{\beta_{1:N,1:T_n}}$, $d_{B,1:K}$, $d_{W,1:K}$, $\boldsymbol{\mu}_{\mu_{1:N}}$, $\boldsymbol{\Sigma}_{\mu_{1:N}}$;
Coordinate ascent:
$c_B = \frac{\nu_B+K}{2}$; $c_W = \frac{\nu_W+K}{2}$; $w_B = \nu_B + N + K - 1$; $w_W = \nu_W + \sum_{n=1}^N T_n + K - 1$;
$\boldsymbol{\Theta}_B = 2\nu_B \text{diag}\left(\frac{c_B}{\boldsymbol{d}_B}\right) + N\boldsymbol{\Sigma}_\zeta + \sum_{n=1}^N \left(\boldsymbol{\Sigma}_{\mu_n} + (\boldsymbol{\mu}_{\mu_n} - \boldsymbol{\mu}_\zeta)(\boldsymbol{\mu}_{\mu_n} - \boldsymbol{\mu}_\zeta)^\top\right)$;
$\boldsymbol{\Theta}_W = 2\nu_W \text{diag}\left(\frac{c_W}{\boldsymbol{d}_W}\right) + \sum_{n=1}^N T_n\boldsymbol{\Sigma}_{\mu_n} + \sum_{n=1}^N \sum_{t=1}^{T_n} \left(\boldsymbol{\Sigma}_{\beta_{nt}} + (\boldsymbol{\mu}_{\beta_{nt}} - \boldsymbol{\mu}_{\mu_n})(\boldsymbol{\mu}_{\beta_{nt}} - \boldsymbol{\mu}_{\mu_n})^\top\right)$;
while *not converged* **do**
Update $\boldsymbol{\mu}_{\beta_{nt}}$, $\boldsymbol{\Sigma}_{\beta_{nt}}$ for $\forall n, \forall t$ using equation 17;
$\boldsymbol{\Sigma}_{\mu_n} = \left(w_B\boldsymbol{\Theta}_B^{-1} + T_n w_W \boldsymbol{\Theta}_W^{-1}\right)^{-1} \forall n$;
$\boldsymbol{\mu}_{\mu_n} = \boldsymbol{\Sigma}_{\mu_n}\left(w_B\boldsymbol{\Theta}_B^{-1}\boldsymbol{\mu}_\zeta + w_W\boldsymbol{\Theta}_W^{-1}\sum_{t=1}^{T_n}\boldsymbol{\mu}_{\beta_{nt}}\right) \forall n$;
$\boldsymbol{\Sigma}_\zeta = \left(\boldsymbol{\Xi}_0^{-1} + Nw_B\boldsymbol{\Theta}_B^{-1}\right)^{-1}$;
$\boldsymbol{\mu}_\zeta = \boldsymbol{\Sigma}_\zeta\left(\boldsymbol{\Xi}_0^{-1}\boldsymbol{\xi}_0 + w_B\boldsymbol{\Theta}_B^{-1}\sum_{n=1}^N\boldsymbol{\mu}_{\mu_n}\right)$;
$\boldsymbol{\Theta}_B = 2\nu_B \text{diag}\left(\frac{c_B}{\boldsymbol{d}_B}\right) + N\boldsymbol{\Sigma}_\zeta + \sum_{n=1}^N \left(\boldsymbol{\Sigma}_{\mu_n} + (\boldsymbol{\mu}_{\mu_n} - \boldsymbol{\mu}_\zeta)(\boldsymbol{\mu}_{\mu_n} - \boldsymbol{\mu}_\zeta)^\top\right)$;
$\boldsymbol{\Theta}_W = 2\nu_W \text{diag}\left(\frac{c_W}{\boldsymbol{d}_W}\right) + \sum_{n=1}^N T_n\boldsymbol{\Sigma}_{\mu_n} + \sum_{n=1}^N \sum_{t=1}^{T_n} \left(\boldsymbol{\Sigma}_{\beta_{nt}} + (\boldsymbol{\mu}_{\beta_{nt}} - \boldsymbol{\mu}_{\mu_n})(\boldsymbol{\mu}_{\beta_{nt}} - \boldsymbol{\mu}_{\mu_n})^\top\right)$;
$d_{B,k} = \frac{1}{A_{B,k}^2} + w_B\nu_B\left(\boldsymbol{\Theta}_B^{-1}\right)_{kk} \forall k$;
$d_{W,k} = \frac{1}{A_{W,k}^2} + w_W\nu_W\left(\boldsymbol{\Theta}_W^{-1}\right)_{kk} \forall k$;
end

Algorithm 1: Pseudo-code representations of variational Bayes method for posterior inference in MXL models with unobserved inter- and intra-individual heterogeneity

4 Simulation study

4.1 Data and experimental setup

For the simulation study, we devise a simple synthetic data generating process (DGP). Decision-makers are assumed to be utility maximisers and to evaluate alternatives based on the utility specification $U_{ntj} = \boldsymbol{X}_{ntj}\boldsymbol{\beta}_{n,t} + \epsilon_{ntj}$. Here, $n \in \{1, \dots, N\}$ indexes decision-makers, $t \in \{1, \dots, T\}$ indexes choice occasions, and $j \in \{1, \dots, 5\}$ indexes alternatives. $\boldsymbol{X}_{ntj}$ is a row-vector of attributes drawn from a standard uniform distribution. ϵ_{ntj} is a stochastic disturbance sampled from Gumbel(0, 1). The DGP of the observation-specific taste parameters $\boldsymbol{\beta}_{n,t}$ is as follows:

$$\boldsymbol{\mu}_n|\boldsymbol{\zeta}, \boldsymbol{\Sigma}_B \sim \text{N}(\boldsymbol{\zeta}, \boldsymbol{\Sigma}_B), n = 1, \ldots, N, \tag{18}$$

$$\boldsymbol{\beta}_{nt}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W \sim \text{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W), n = 1, \ldots, N, t = 1, \ldots, T_n. \tag{19}$$

The assumed values of $\boldsymbol{\zeta}$, $\boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_W$ are enumerated in Appendix D. The scale of the population-level parameters is set such that the error rate is approximately 50%, i.e. in 50% of the cases decision-makers deviate from the deterministically-best alternative due to the stochastic utility component. We set $N = 1,000$ and allow T to take a value in $\{20, 40\}$. For each combination of N and T, we consider ten replications, whereby the data for each replication are generated based on a different random seed.

4.2 Accuracy assessment

We evaluate the performance of the considered estimation approaches in terms of their predictive accuracy, as is common in the context of Bayesian estimation of discrete choice models (see Bansal *et al.,* 2019, Braun and McAuliffe, 2010, Depraetere and Vandebroek, 2017, Tan, 2017). Predictive accuracy accounts for the uncertainty in the estimates and allows for a succinct summary of estimation accuracy, when the number of model parameters is large (Depraetere and Vandebroek, 2017). In the present application, we consider two out-of-sample prediction scenarios. In the first scenario, we predict choice probabilities for a new set of individuals, i.e. we predict *between* individuals. In the second scenario, we predict choice probabilities for new choice sets for individual who are already in the sample, i.e. we predict *within* individuals. For each of these scenarios, we calculate the total variation distance (TVD; Braun and McAuliffe, 2010) between the true and the estimated predictive choice distributions. We proceed as follows:

1. To evaluate the between-individual predictive accuracy, we compute TVD for a validation sample, which we generate along with each training sample. Each validation sample is based on the same DGP as its respective training sample, whereby the number of decision-makers is set to 25 and the number of observations per decision-maker is set to one. The true predictive choice distribution for a choice set C_{nt} with attributes $\boldsymbol{X}^*_{nt}$ from the validation sample is given by

$$P_{\text{true}}(y^*_{nt}|\boldsymbol{X}^*_{nt}) = \int \left(\int P(y^*_{nt} = j|\boldsymbol{X}^*_{nt}, \boldsymbol{\beta}) f(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_W) d\boldsymbol{\beta} \right) f(\boldsymbol{\mu}|\boldsymbol{\zeta}, \boldsymbol{\Sigma}_B) d\boldsymbol{\mu} \tag{20}$$

The corresponding estimated predictive choice distribution is

$$\hat{P}(y^*_{nt}|\boldsymbol{X}^*_{nt}, \boldsymbol{y}) = \int\int\int\left(\int\left(\int P(y^*_{nt}|\boldsymbol{X}^*_{nt}, \boldsymbol{\beta}) f(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_W) d\boldsymbol{\beta}\right) f(\boldsymbol{\mu}, \boldsymbol{\Sigma}_B) d\boldsymbol{\mu}\right) P(\boldsymbol{\zeta}, \boldsymbol{\Sigma}_B, \boldsymbol{\Sigma}_W|\boldsymbol{y}) d\boldsymbol{\zeta} d\boldsymbol{\Sigma}_B d\boldsymbol{\Sigma}_W \tag{21}$$

TVD_B is given by

$$\text{TVD}_B = \frac{1}{2}\sum_{j\in C_{nt}} \left|P_{\text{true}}(y^*_{nt} = j|\boldsymbol{X}^*_{nt}) - \hat{P}(y^*_{nt} = j|\boldsymbol{X}^*_{nt}, \boldsymbol{y})\right|. \tag{22}$$

For succinctness, we calculate averages across decision-makers and choice sets.

2. To evaluate the within-individual predictive accuracy, we compute TVD for another validation sample, which we generate along with each training sample. For 25 individuals from the training sample, we generate one additional choice set. Then, the true predictive choice distribution for a choice set C_{nt} with attributes $\boldsymbol{X}^*_{nt}$ from the validation sample is given by

$$P_{\text{true}}(y^\dagger_{nt}|\boldsymbol{X}^\dagger_{nt}) = \int P(y^\dagger_{nt} = j|\boldsymbol{X}^\dagger_{nt}, \boldsymbol{\beta}) f(\boldsymbol{\beta}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W) d\boldsymbol{\beta} \tag{23}$$

The corresponding estimated predictive choice distribution is

$$\hat{P}(y^\dagger_{nt}|\boldsymbol{X}^\dagger_{nt}, \boldsymbol{y}) = \int\int\left(\int P(y^\dagger_{nt}|\boldsymbol{X}^\dagger_{nt}, \boldsymbol{\beta}) f(\boldsymbol{\beta}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W) d\boldsymbol{\beta}\right) P(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W|\boldsymbol{y}) d\boldsymbol{\mu}_n d\boldsymbol{\Sigma}_W \tag{24}$$

TVD_W is given by

$$\text{TVD}_W = \frac{1}{2}\sum_{j\in C_{nt}} \left|P_{\text{true}}(y^\dagger_{nt} = j|\boldsymbol{X}^\dagger_{nt}) - \hat{P}(y^\dagger_{nt} = j|\boldsymbol{X}^\dagger_{nt}, \boldsymbol{y})\right|. \tag{25}$$

Again, we calculate averages across decision-makers and choice sets for succinctness.

4.3 Implementation details

We implement the MCMC and VB methods by writing our own Python code and make an effort that the implementations of the different estimators are as similar as possible to allow for fair comparisons of estimation times. For MCMC, the sampler is executed with two parallel Markov chains and 200,000 iterations for each chain, whereby the initial 100,000 iterations of each chain are discarded for burn-in. After burn-in, every tenth draws is retained to reduce the amount of autocorrelation in the chains. For VB, we apply the same stopping criterion as Tan (2017): We define $\boldsymbol{\vartheta} = \begin{bmatrix}\boldsymbol{\alpha}^\top & \boldsymbol{\zeta}^\top & \text{diag}(\boldsymbol{\Theta})^\top & \boldsymbol{d}^\top\end{bmatrix}^\top$ and let $\vartheta_i^{(\tau)}$ denote the ith element of $\boldsymbol{\vartheta}$ at iteration τ. We terminate the iterative coordinate ascent algorithm, when $\delta^{(\tau)} = \arg\max_i \frac{|\vartheta_i^{(\tau+1)} - \vartheta_i^{(\tau)}|}{|\vartheta_i^{(\tau)}|} < 0.005$. As $\delta^{(\tau)}$ can fluctuate, $\boldsymbol{\vartheta}^{(\tau)}$ is substituted by its average over the last five iterations. The simulation experiments are conducted on the Katana high performance computing cluster at the Faculty of Science, UNSW Australia.

4.4 Results

Table 1 enumerates the results for the simulation study. We report the means and standard errors of the considered performance metrics for ten replications under different combinations of sample size $N = 1,000$ and choice occasions per decision-maker $T \in \{20, 40\}$. In both experimental conditions, VB is approximately twice as fast as MCMC but noticeably less accurate. In the case of between-individual prediction, TVD is approximately ten times larger for VB than for MCMC. A possible explanation for this discrepancy is the poor recovery of the covariance ($\boldsymbol{\Sigma}_B$) of the individual-specific parameters, which is a consequence of the overly simplistic mean-field assumption. The discrepancy in predictive accuracy between VB and MCMC is less strongly pronounced for the case of within-individual prediction, which suggests that the within-individual covariance matrix ($\boldsymbol{\Sigma}_W$) is recovered reasonably well by VB.

	Estimation time		**TVD_B [10%]**		**TVD_W [10%]**	
	Mean	**Std. err.**	**Mean**	**Std. err.**	**Mean**	**Std. err.**
$N = 1000$; $T = 20$						
MCMC	3049.2	41.2	0.0198	0.0017	0.2028	0.0057
VB	1526.2	14.9	0.1977	0.0061	0.3203	0.0082
$N = 1000$; $T = 40$						
MCMC	5649.5	93.5	0.0182	0.0020	0.1735	0.0047
VB	3199.2	25.7	0.1458	0.0028	0.2543	0.0076

Note: TVD_B: total variation distance for between-individual prediction. TVD_W: total variation distance for within-individual prediction.

Table 1: Results of the simulation study

5 Conclusion

Motivated by recent advances in scalable Bayesian inference for mixed logit models, this current paper derives a mean-field variational Bayes method for the estimation of mixed logit models with unobserved inter- and intra-individual heterogeneity. In a simulation study, we benchmark the performance of the proposed method against MCMC and provide a proof-of-concept of the feasibility of the proposed VB method. We show that VB is substantially faster than MCMC but also find that VB is noticeably less accurate than MCMC. A possible explanation for this discrepancy in predictive accuracy is that the mean-field assumption of VB is too simplistic.

There are several directions in which future research may build on the work presented in the current paper. First, the quality of the variational approximation should be improved by increasing the tightness of the variational lower bound. One possible way to achieve this is to inject structure into the formulation of the variational distribution and to recognise that the model parameters are related in a hierarchical fashion (Ranganath *et al.,* 2016). Alternatively, the expressiveness of the variational distribution could be enhanced by employing more flexible families of distributions such as mixtures or normalising flows (Jaakkola and Jordan, 1998, Rezende and Mohamed, 2015). A second direction for future research is to develop an online inference method which will enable near real-time learning and prediction of individual preferences. Hoffman *et al.* (2013) establish connections between VB and stochastic optimisation and show how VB can be applied to streaming data.

6 References

Akinc, D. and M. Vandebroek (2018) Bayesian estimation of mixed logit models: Selecting an appropriate prior for the covariance matrix, *Journal of choice modelling*, **29**, 133–151.

Bansal, P., R. Krueger, M. Bierlaire, R. A. Daziano and T. H. Rashidi (2019) Bayesian estimation of mixed multinomial logit models: Advances and simulation-based evaluations, *arXiv preprint arXiv:1904.03647*.

Becker, F., M. Danaf, X. Song, B. Atasoy and M. Ben-Akiva (2018) Bayesian estimator for logit mixtures with inter-and intra-consumer heterogeneity, *Transportation Research Part B: Methodological*, **117**, 1–17.

Bettman, J. R., M. F. Luce and J. W. Payne (1998) Constructive consumer choice processes, *Journal of consumer research*, **25** (3) 187–217.

Bishop, C. (2006) *Pattern Recognition and Machine Learning*, Information Science and Statistics, Springer Nature Textbooks ? HE site, Springer-Verlag, New York, ISBN 978-0-387-31073-2.

Blei, D. M., A. Kucukelbir and J. D. McAuliffe (2017) Variational Inference: A Review for Statisticians, *Journal of the American Statistical Association*, **112** (518) 859–877, April 2017, ISSN 0162-1459.

Braun, M. and J. McAuliffe (2010) Variational Inference for Large-Scale Models of Discrete Choice, *Journal of the American Statistical Association*, **105** (489) 324–335, March 2010, ISSN 0162-1459.

Depraetere, N. and M. Vandebroek (2017) A comparison of variational approximations for fast inference in mixed logit models, *Computational Statistics*, **32** (1) 93–125, March 2017, ISSN 0943-4062, 1613-9658.

Gelman, A., H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari and D. B. Rubin (2013) *Bayesian data analysis*, Chapman and Hall/CRC.

Hess, S. and M. Giergiczny (2015) Intra-respondent heterogeneity in a stated choice survey on wetland conservation in belarus: first steps towards creating a link with uncertainty in contingent valuation, *Environmental and Resource Economics*, **60** (3) 327–347.

Hess, S. and J. M. Rose (2009) Allowing for intra-respondent variations in coefficients estimated on repeated choice data, *Transportation Research Part B: Methodological*, **43** (6) 708–719.

Hess, S. and K. E. Train (2011) Recovery of inter-and intra-personal heterogeneity using mixed logit models, *Transportation Research Part B: Methodological*, **45** (7) 973–990.

Hoffman, M. D., D. M. Blei, C. Wang and J. Paisley (2013) Stochastic variational inference, *The Journal of Machine Learning Research*, **14** (1) 1303–1347.

Huang, A. and M. P. Wand (2013) Simple marginally noninformative prior distributions for covariance matrices, *Bayesian Anal.*, **8** (2) 439–452, 06 2013.

Jaakkola, T. S. and M. I. Jordan (1998) Improving the mean field approximation via the use of mixture distributions, in *Learning in graphical models*, 163–173, Springer.

Jordan, M. I., Z. Ghahramani, T. S. Jaakkola and L. K. Saul (1999) An Introduction to Variational Methods for Graphical Models, *Machine Learning*, **37** (2) 183–233, November 1999, ISSN 0885-6125, 1573-0565.

Kivetz, R., O. Netzer and R. Schrift (2008) The synthesis of preference: Bridging behavioral decision research and marketing science, *Journal of Consumer Psychology*, **18** (3) 179–186.

Knowles, D. A. and T. Minka (2011) Non-conjugate variational message passing for multinomial and binary regression, paper presented at the *Advances in Neural Information Processing Systems*, 1701–1709.

Kullback, S. and R. A. Leibler (1951) On Information and Sufficiency, *The Annals of Mathematical Statistics*, **22** (1) 79–86, March 1951, ISSN 0003-4851, 2168-8990.

McFadden, D. and K. Train (2000) Mixed mnl models for discrete response, *Journal of applied Econometrics*, **15** (5) 447–470.

Nocedal, J. and S. Wright (2006) *Numerical optimization*, Springer Science & Business Media.

Ormerod, J. T. and M. P. Wand (2010) Explaining Variational Approximations, *The American Statistician*, **64** (2) 140–153, May 2010, ISSN 0003-1305.

Ranganath, R., D. Tran and D. Blei (2016) Hierarchical variational models, paper presented at the *International Conference on Machine Learning*, 324–333.

Revelt, D. and K. Train (1998) Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level, *The Review of Economics and Statistics*, **80** (4) 647–657, November 1998, ISSN 0034-6535.

Rezende, D. J. and S. Mohamed (2015) Variational inference with normalizing flows, *arXiv preprint arXiv:1505.05770*.

Stigler, G. J. and G. S. Becker (1977) De gustibus non est disputandum, *The american economic review*, **67** (2) 76–90.

Tan, L. S. L. (2017) Stochastic variational inference for large-scale discrete choice models using adaptive batch sizes, *Statistics and Computing*, **27** (1) 237–257, January 2017, ISSN 1573-1375.

Train, K. E. (2009) *Discrete Choice Methods with Simulation*, 2nd edn., Cambridge University Press.

A Gibbs sampler

1. Update $a_{B,k}$ for all $k \in \{1, \dots, K\}$ by sampling $a_{B,k} \sim \text{Gamma}\left(\frac{\nu_B+K}{2}, \frac{1}{A^2_{B,k}} + \nu_B \left(\boldsymbol{\Sigma}_B^{-1}\right)_{kk}\right)$
2. Update $\boldsymbol{\Sigma}_B$ by sampling $\boldsymbol{\Sigma}_B \sim \text{IW}\left(\nu_B + N + K - 1, 2\nu_B \text{diag}(\boldsymbol{a}_B) + \sum_{n=1}^N (\boldsymbol{\mu}_n - \boldsymbol{\zeta})(\boldsymbol{\mu}_n - \boldsymbol{\zeta})^\top\right)$
3. Update $a_{W,k}$ for all $k \in \{1, \dots, K\}$ by sampling $a_{W,k} \sim \text{Gamma}\left(\frac{\nu_W+K}{2}, \frac{1}{A^2_{W,k}} + \nu_W \left(\boldsymbol{\Sigma}_W^{-1}\right)_{kk}\right)$
4. Update $\boldsymbol{\Sigma}_W$ by sampling
   $\boldsymbol{\Sigma}_W \sim \text{IW}\left(\nu_W + \sum_{n=1}^N T_n + K - 1, 2\nu_W \text{diag}(\boldsymbol{a}_W) + \sum_{n=1}^N \sum_{t=1}^{T_n} (\boldsymbol{\beta}_{nt} - \boldsymbol{\mu}_n)(\boldsymbol{\beta}_{nt} - \boldsymbol{\mu}_n)^\top\right)$
5. Update $\boldsymbol{\zeta}$ by sampling $\boldsymbol{\zeta} \sim \text{N}(\boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta)$, where
   $\boldsymbol{\Sigma}_\zeta = \left(\boldsymbol{\Xi}_0^{-1} + N\boldsymbol{\Sigma}_B^{-1}\right)^{-1}$ and $\boldsymbol{\mu}_\zeta = \boldsymbol{\Sigma}_\zeta \left(\boldsymbol{\Xi}_0^{-1}\boldsymbol{\xi}_0 + \boldsymbol{\Sigma}_B^{-1} \sum_{n=1}^N \boldsymbol{\mu}_n\right)$
6. Update $\boldsymbol{\mu}_n$ for all $n \in \{1, \dots, N\}$ by sampling $\boldsymbol{\mu}_n \sim \text{N}(\boldsymbol{\mu}_{\mu_n}, \boldsymbol{\Sigma}_{\mu_n})$, where $\boldsymbol{\Sigma}_{\mu_n} = \left(\boldsymbol{\Sigma}_B^{-1} + T_n \boldsymbol{\Sigma}_W^{-1}\right)^{-1}$ and $\boldsymbol{\mu}_{\mu_n} = \boldsymbol{\Sigma}_{\mu_n} \left(\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\zeta} + \boldsymbol{\Sigma}_W^{-1} \sum_{t=1}^{T_n} \boldsymbol{\beta}_{nt}\right)$
7. Update $\boldsymbol{\beta}_{nt}$ for all $n \in \{1, \dots, N\}$ and $t \in \{1, \dots, T_n\}$:
   a) Propose $\tilde{\boldsymbol{\beta}}_{nt} = \boldsymbol{\beta}_{nt} + \sqrt{\rho}\,\text{chol}(\boldsymbol{\Sigma}_W)\boldsymbol{\eta}$, where $\boldsymbol{\eta} \sim \text{N}(\boldsymbol{0}, \boldsymbol{I}_K)$.
   b) Compute $r = \frac{P(y_{nt}|X_{nt}, \tilde{\beta}_{nt})\phi(\tilde{\beta}_{nt}|\mu_n, \Sigma_W)}{P(y_{nt}|X_{nt}, \beta_{nt})\phi(\beta_{nt}|\mu_n, \Sigma_W)}$.
   c) Draw $u \sim \text{Uniform}(0, 1)$. If $r \le u$, accept the proposal. If $r > u$, reject the proposal.

B E-LSE

We take a second-order Taylor series expansion of $g_{nt} = \left\{\ln \sum_{k \in C_{nt}} \exp(\boldsymbol{X}_{ntk}\boldsymbol{\beta}_{nt})\right\}$ around $\boldsymbol{\mu}_{\beta_{nt}}$:

$$g_{nt}(\boldsymbol{\beta}_{nt}) \approx g_{nt}(\boldsymbol{\mu}_{\beta_{nt}}) + \nabla g_{nt}(\boldsymbol{\mu}_{\beta_{nt}})[\boldsymbol{\beta}_{nt} - \boldsymbol{\mu}_{\beta_{nt}}] + \frac{1}{2}[\boldsymbol{\beta}_{nt} - \boldsymbol{\mu}_{\beta_{nt}}]^\top \nabla^2 g_{nt}(\boldsymbol{\mu}_{\beta_{nt}})[\boldsymbol{\beta}_{nt} - \boldsymbol{\mu}_{\beta_{nt}}] \tag{26}$$

Then,

$$\begin{aligned} \mathbb{E}_q\{g_{nt}(\boldsymbol{\beta}_{nt})\} \approx & g_{nt}(\boldsymbol{\mu}_{\beta_{nt}}) + \frac{1}{2}\text{tr}\left(\nabla^2 g_{nt}(\boldsymbol{\mu}_{\beta_{nt}})\boldsymbol{\Sigma}_{\beta_{nt}}\right) \\ \approx & \ln \sum_{k \in C_{nt}} \exp(\boldsymbol{X}_{ntk}\boldsymbol{\mu}_{\beta_{nt}}) + \frac{1}{2}\text{tr}\left(\left(\boldsymbol{X}_{nt}^\top \left(\text{diag}(\boldsymbol{p}_{nt0}) - \boldsymbol{p}_{nt0}\boldsymbol{p}_{nt0}^\top\right) \boldsymbol{X}_{nt}\right) \boldsymbol{\Sigma}_{\beta_{nt}}\right), \end{aligned} \tag{27}$$

where $p^0_{ntj} = \frac{\exp(\boldsymbol{X}_{ntj}\boldsymbol{\mu}_{\beta_{nt}})}{\sum_{k \in C_{nt}} \exp(\boldsymbol{X}_{ntk}\boldsymbol{\mu}_{\beta_{nt}})}$ and $\boldsymbol{p}_{nt0} = \left\{p^0_{ntm}\right\}_{m \in C_{nt}}$ is a column-stacked vector of all p^0_{ntm} in C_{nt}.

C Optimal densities of conjugate variational factors

C.1 $q^*(a_{B,k})$ and $q^*(a_{W,k})$

$$\begin{aligned} q^*(a_{B,k}) &\propto \exp \mathbb{E}_{-a_{B,k}} \left\{ \ln P\left(a_{B,k}|s, \frac{1}{A^2_{B,k}}\right) + \ln P(\boldsymbol{\Sigma}_B|\omega_B, \boldsymbol{B}_B) \right\} \\ &\propto \exp \mathbb{E}_{-a_{B,k}} \left\{ (s-1)\ln a_{B,k} - \frac{a_{B,k}}{A^2_{B,k}} + \frac{\omega_B}{2}\ln \boldsymbol{B}_{B,kk} - \frac{1}{2}\boldsymbol{B}_{B,kk}\left(\boldsymbol{\Sigma}_B^{-1}\right)_{kk} \right\} \\ &\propto \exp \left\{ \left(\frac{\nu_B + K}{2} - 1\right)\ln a_{B,k} - \left(\frac{1}{A^2_{B,k}} + \nu_B \mathbb{E}_{-a_{B,k}}\left\{\left(\boldsymbol{\Sigma}_B^{-1}\right)_{kk}\right\}\right) a_{B,k} \right\} \\ &\propto \text{Gamma}(c_B, d_{B,k}), \end{aligned} \tag{28}$$

where $c_B = \frac{\nu_B+K}{2}$ and $d_{B,k} = \frac{1}{A^2_{B,k}} + \nu_B \mathbb{E}_{-a_{B,k}}\left\{\left(\boldsymbol{\Sigma}_B^{-1}\right)_{kk}\right\}$. Furthermore, we note that $\mathbb{E}a_{B,k} = \frac{c_B}{d_{B,k}}$. $q^*(a_{W,k})$ can be derived in the same way. We have $q^*(a_{W,k}) \propto \text{Gamma}(c_W, d_{W,k})$ with $c_W = \frac{\nu_W+K}{2}$ and $d_{W,k} = \frac{1}{A^2_{W,k}} + \nu_W \mathbb{E}_{-a_{W,k}}\left\{\left(\boldsymbol{\Sigma}_W^{-1}\right)_{kk}\right\}$. Moreover, $\mathbb{E}a_{W,k} = \frac{c_W}{d_{W,k}}$.

C.2 $q^*(\boldsymbol{\Sigma}_B)$ and $q^*(\boldsymbol{\Sigma}_W)$

$$\begin{aligned} q^*(\boldsymbol{\Sigma}_B) &\propto \exp \mathbb{E}_{-\boldsymbol{\Sigma}_B} \left\{ \ln P(\boldsymbol{\Sigma}_B|\omega_B, \boldsymbol{B}_B) + \sum_{n=1}^{N} \ln P(\boldsymbol{\mu}_n|\boldsymbol{\zeta}, \boldsymbol{\Sigma}_B) \right\} \\ &\propto \exp \mathbb{E}_{-\boldsymbol{\Sigma}_B} \left\{ -\frac{\omega_B + K + 1}{2}\ln|\boldsymbol{\Sigma}_B| - \frac{1}{2}\text{tr}\left(\boldsymbol{B}_B \boldsymbol{\Sigma}_B^{-1}\right) - \frac{N}{2}\ln|\boldsymbol{\Sigma}_B| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{\mu}_n - \boldsymbol{\zeta})^\top \boldsymbol{\Sigma}_B^{-1}(\boldsymbol{\mu}_n - \boldsymbol{\zeta}) \right\} \\ &= \exp \left\{ -\frac{\omega_B + N + K + 1}{2}\ln|\boldsymbol{\Sigma}_B| - \frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}_B^{-1} \mathbb{E}_{-\boldsymbol{\Sigma}_B}\left\{ \boldsymbol{B}_B + \sum_{n=1}^{N}(\boldsymbol{\mu}_n - \boldsymbol{\zeta})(\boldsymbol{\mu}_n - \boldsymbol{\zeta})^\top \right\}\right)\right\} \\ &\propto \text{IW}(w_B, \boldsymbol{\Theta}_B), \end{aligned} \tag{29}$$

where $w_B = \nu_B + N + K - 1$ and $\boldsymbol{\Theta}_B = 2\nu_B \text{diag}\left(\frac{c_B}{\boldsymbol{d}_B}\right) + N\boldsymbol{\Sigma}_\zeta + \sum_{n=1}^{N}\left(\boldsymbol{\Sigma}_{\mu_n} + (\boldsymbol{\mu}_{\mu_n} - \boldsymbol{\mu}_\zeta)(\boldsymbol{\mu}_{\mu_n} - \boldsymbol{\mu}_\zeta)^\top\right)$. We use $\mathbb{E}\left(\boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top\right) = \boldsymbol{\mu}_{\mu_n}\boldsymbol{\mu}_{\mu_n}^\top + \boldsymbol{\Sigma}_{\mu_n}$ and $\mathbb{E}\left(\boldsymbol{\zeta}\boldsymbol{\zeta}^\top\right) = \boldsymbol{\mu}_\zeta\boldsymbol{\mu}_\zeta^\top + \boldsymbol{\Sigma}_\zeta$. Furthermore, we note that $\mathbb{E}\{\boldsymbol{\Sigma}_B^{-1}\} = w_B\boldsymbol{\Theta}_B^{-1}$ and $\mathbb{E}\{\ln|\boldsymbol{\Sigma}_B|\} = \ln|\boldsymbol{\Theta}_B| + C$, where C is a constant. $q^*(\boldsymbol{\Sigma}_W)$ can be derived in the same way. We have $q^*(\boldsymbol{\Sigma}_W) \propto \text{IW}(w_W, \boldsymbol{\Theta}_W)$ with $w_W = \nu_W + \sum_{n=1}^{N} T_n + K - 1$ and $\boldsymbol{\Theta}_W = 2\nu_W \text{diag}\left(\frac{c_W}{\boldsymbol{d}_W}\right) + \sum_{n=1}^{N} T_n \boldsymbol{\Sigma}_{\mu_n} + \sum_{n=1}^{N}\sum_{t=1}^{T_n}\left(\boldsymbol{\Sigma}_{\beta_{nt}} + (\boldsymbol{\mu}_{\beta_{nt}} - \boldsymbol{\mu}_{\mu_n})(\boldsymbol{\mu}_{\beta_{nt}} - \boldsymbol{\mu}_{\mu_n})^\top\right)$. Moreover, $\mathbb{E}\{\boldsymbol{\Sigma}_W^{-1}\} = w_W\boldsymbol{\Theta}_W^{-1}$ and $\mathbb{E}\{\ln|\boldsymbol{\Sigma}_W|\} = \ln|\boldsymbol{\Theta}_W| + C$, where C is a constant.

C.3 $q^*(\zeta)$

$$
\begin{aligned}
q^*(\zeta) &\propto \exp \mathbb{E}_{-\zeta}\left\{\ln P(\zeta|\xi_0, \Xi_0) + \sum_{n=1}^{N} \ln P(\mu_n|\zeta, \Sigma_B)\right\} \\
&\propto \exp \mathbb{E}_{-\zeta}\left\{-\frac{1}{2}\zeta^\top \Xi_0^{-1}\zeta + \zeta^\top \Xi_0^{-1}\xi_0 - \frac{N}{2}\zeta^\top \Sigma_B^{-1}\zeta + \sum_{n=1}^{N} \zeta^\top \Sigma_B^{-1}\mu_n\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\zeta^\top\left(\Xi_0^{-1} + N\mathbb{E}_{-\zeta}\left\{\Sigma_B^{-1}\right\}\right)\zeta - 2\zeta^\top\left(\Xi_0^{-1}\xi_0 + \mathbb{E}_{-\zeta}\left\{\Sigma_B^{-1}\right\}\sum_{n=1}^{N}\mathbb{E}_{-\zeta}\mu_n\right)\right)\right\} \\
&\propto \text{Normal}(\mu_\zeta, \Sigma_\zeta),
\end{aligned}
\tag{30}
$$

where $\Sigma_\zeta = \left(\Xi_0^{-1} + N\mathbb{E}_{-\zeta}\left\{\Sigma_B^{-1}\right\}\right)^{-1}$ and $\mu_\zeta = \Sigma_\zeta\left(\Xi_0^{-1}\xi_0 + \mathbb{E}_{-\zeta}\left\{\Sigma_B^{-1}\right\}\sum_{n=1}^{N}\mathbb{E}_{-\zeta}\mu_n\right)$. Furthermore, we note that $\mathbb{E}\zeta = \mu_\zeta$ and $\mathbb{E}\mu_n = \mu_{\mu_n}$.

C.4 $q^*(\mu_n)$

$$
\begin{aligned}
q^*(\mu_n) &\propto \exp \mathbb{E}_{-\mu_n}\left\{\ln P(\mu_n|\zeta, \Sigma_B) + \sum_{t=1}^{T_n} \ln P(\beta_{nt}|\mu_n, \Sigma_W)\right\} \\
&\propto \exp \mathbb{E}_{-\mu_n}\left\{-\frac{1}{2}\mu_n^\top \Sigma_B^{-1}\mu_n + \mu_n^\top \Sigma_B^{-1}\zeta - \frac{T_n}{2}\mu_n^\top \Sigma_W^{-1}\mu_n + \sum_{t=1}^{T} \mu_n^\top \Sigma_W^{-1}\beta_{nt}\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\mu_n^\top\left(\mathbb{E}_{-\mu_n}\left\{\Sigma_B^{-1}\right\} + T_n\mathbb{E}_{-\mu_n}\left\{\Sigma_W^{-1}\right\}\right)\mu_n - 2\mu_n^\top\left(\mathbb{E}_{-\mu_n}\left\{\Sigma_B^{-1}\right\}\mathbb{E}_{-\mu_n}\{\zeta\} + \mathbb{E}_{-\mu_n}\left\{\Sigma_W^{-1}\right\}\sum_{t=1}^{T_n}\mathbb{E}_{-\beta_{nt}}\beta_{nt}\right)\right)\right\} \\
&\propto \text{Normal}(\mu_{\mu_n}, \Sigma_{\mu_n}),
\end{aligned}
\tag{31}
$$

where $\Sigma_{\mu_n} = \left(\mathbb{E}_{-\mu_n}\left\{\Sigma_B^{-1}\right\} + T_n\mathbb{E}_{-\mu_n}\left\{\Sigma_W^{-1}\right\}\right)^{-1}$ and $\mu_{\mu_n} = \Sigma_{\mu_n}\left(\mathbb{E}_{-\mu_n}\left\{\Sigma_B^{-1}\right\}\mathbb{E}_{-\mu_n}\{\zeta\} + \mathbb{E}_{-\mu_n}\left\{\Sigma_W^{-1}\right\}\sum_{t=1}^{T_n}\mathbb{E}_{-\mu_n}\beta_{nt}\right)$. Furthermore, we note that $\mathbb{E}\mu_n = \mu_{\mu_n}$ and $\mathbb{E}\beta_{nt} = \mu_{\beta_{nt}}$.

C.5 $q^*(\beta_{nt})$

We consider relevant terms of $\ln P(y_{1:N}, \theta)$, which remain non-zero after differentiation:

$$
f(\beta_{nt}) = -\frac{1}{2}(\beta_{nt} - \mu_n)^\top \Sigma_W^{-1}(\beta_{nt} - \mu_n) + \sum_{j \in C_{nt}} y_{ntj} X_{ntj}\beta_{nt} + \ln\left[\sum_{k \in C_{nt}} \exp(X_{ntk}\beta_{nt})\right] \tag{32}
$$

$$\mathbb{E}_q\{f(\boldsymbol{\beta}_{nt})\} = -\frac{w_W}{2}(\boldsymbol{\mu}_{\beta_{nt}} - \boldsymbol{\mu}_{\mu_n})^\top \boldsymbol{\Theta}_W^{-1}(\boldsymbol{\mu}_{\beta_{nt}} - \boldsymbol{\mu}_{\mu_n}) - \frac{w_W}{2}\text{tr}(\boldsymbol{\Sigma}_{\beta_{nt}}\boldsymbol{\Theta}_W^{-1}) - \frac{w_W}{2}\text{tr}(\boldsymbol{\Sigma}_{\mu_n}\boldsymbol{\Theta}_W^{-1}) + \sum_{j \in C_{nt}} y_{ntj}\boldsymbol{X}_{ntj}\boldsymbol{\mu}_{\beta_{nt}} + \mathbb{E}_q\left\{\ln \sum_{k \in C_{nt}} \exp(\boldsymbol{X}_{ntk}\boldsymbol{\beta}_{nt})\right\} \tag{33}$$

where $\mathbb{E}_q\left\{\ln \sum_{k \in C_{nt}} \exp(\boldsymbol{X}_{ntk}\boldsymbol{\beta}_{nt})\right\}$ is obtained using the delta method (see Section B). The required gradients are:

$$\frac{\partial \mathbb{E}_q\{f(\boldsymbol{\beta}_{nt})\}}{\partial \boldsymbol{\mu}_{\beta_{nt}}} = -w_W \boldsymbol{\Theta}_W^{-1}(\boldsymbol{\mu}_{\beta_{nt}} - \boldsymbol{\mu}_{\mu_n}) + \boldsymbol{X}_{nt}^\top(\boldsymbol{y}_{nt} - \boldsymbol{p}_{nt0}) + \boldsymbol{X}_{nt}^\top\Big(\text{diag}(\boldsymbol{p}_{nt0}) - \boldsymbol{p}_{nt0}\boldsymbol{p}_{nt0}^\top\Big)\Big(\boldsymbol{X}_{nt}\boldsymbol{\Sigma}_{\beta_{nt}}\boldsymbol{X}_{nt}^\top\boldsymbol{p}_{nt0} - \frac{1}{2}\text{diag}(\boldsymbol{X}_{nt}\boldsymbol{\Sigma}_{\beta_{nt}}\boldsymbol{X}_{nt}^\top)\Big) \tag{34}$$

$$\frac{\partial \mathbb{E}_q\{f(\boldsymbol{\beta}_{nt})\}}{\partial \text{vec}(\boldsymbol{\Sigma}_{\beta_{nt}})} = -\frac{1}{2}\text{vec}\left(w_W\boldsymbol{\Theta}_W^{-1} + \boldsymbol{X}_{nt}^\top\left(\text{diag}(\boldsymbol{p}_{nt0}) - \boldsymbol{p}_{nt0}\boldsymbol{p}_{nt0}^\top\right)\boldsymbol{X}_{nt}\right) \tag{35}$$

D True population parameters for the simulation study

$$\boldsymbol{\zeta} = \begin{bmatrix} -1.4 & 0.8 & 1.0 & 1.5 \end{bmatrix}^\top, \boldsymbol{\Sigma} = \begin{bmatrix} [r]1.0 & 0.8 & 0.8 & 0.8 \\ 0.8 & 1.0 & 0.8 & 0.8 \\ 0.8 & 0.8 & 1.0 & 0.8 \\ 0.8 & 0.8 & 0.8 & 1.0 \end{bmatrix}, \boldsymbol{\Sigma}_B = \tfrac{3}{2} \cdot \boldsymbol{\Sigma}, \boldsymbol{\Sigma}_W = \tfrac{1}{2} \cdot \boldsymbol{\Sigma}.$$