KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

A comparison of variational approximations for fast inference in mixed logit models

Depraetere N, Vandebroek M.

KBI_1503

A comparison of variational approximations for fast inference in mixed logit models

Nicolas Depraetere[1]

KU Leuven, Faculty of Business and Economics, Naamsestraat 69,
3000 Leuven, Belgium.
Email: Nicolas.Depraetere@kuleuven.be
Tel: +32 16 326576

Martina Vandebroek

KU Leuven, Faculty of Business and Economics, Naamsestraat 69,
3000 Leuven, Belgium.
Leuven Statistics Research Center, W. de Croylaan 54, 3001 Leuven-Heverlee, Belgium.
Email: Martina.Vandebroek@kuleuven.be
Tel: +32 16 326975

Nicolas Depraetere was funded by project G.0385.10N of the Flemish Research Foundation (FWO Flanders), Belgium.

[1]Corresponding author

A comparison of variational approximations for fast inference in mixed logit models

Abstract

Variational Bayesian methods aim to address some of the weaknesses (computation time, storage costs and convergence monitoring) of mainstream $MCMC$-based inference at the cost of a biased approximation to the posterior distribution. We investigate the performance of variational approximations in the context of the mixed logit model, which is arguably one of the most used models for discrete choice data. A typical treatment using the variational Bayesian methodology is hindered by the fact that the expectation of the so called log-sum-exponential function has no closed form expression. Therefore, one has to resort to approximating or bounding this term. In this paper we compare seven different possible bounds or approximations. We found that quadratic bounds do not perform particularly well. A recently proposed non-quadratic bound, on the other hand, did perform quite well. We also found that the approximation used in a previous study only performed well for specific settings. Our proposed approximation based on quasi Monte Carlo sampling on the other hand performed consistently well across all simulation settings while remaining computationally tractable.

Keywords **-** Bayesian statistics, Variational Bayes, Discrete choice, Mixed logit models

Acknowledgement

Nicolas Depraetere was funded by project G.0385.10N of the Flemish Research Foundation (FWO Flanders), Belgium.

Introduction

Choice data are often encountered in marketing research. Such data arise when observed subjects (also called persons, households, companies, ...) make choices out of finite sets of mutually exclusive alternatives (revealed preferences) or are asked to make hypothetical choices in hypothetical situations designed by the researcher (stated preferences). Each alternative is characterized by a set of attributes which determine their utility to the subjects. It is generally assumed that subjects will select the alternative with the highest utility as their preferred choice. An example would be the purchase of a car. Important attributes here could be the price, the brand, mileage, size, aesthetic attributes, ... Then, depending on the relative importance of these attributes and the particular values of these attributes, the decision maker chooses the car he prefers. An overview of the rich discrete choice literature can be found in Train (2009). Usually the goal of the researcher is to gage the relative importance of the attributes to the decision makers, based on the observed or stated choices. This is usually done by estimating discrete choice models using logit or probit link functions between a linear function of the attributes and the observed, categorical outcomes. Estimation of non-trivial discrete choice models rapidly becomes difficult and many of such models, like the mixed logit model, are therefore estimated with a hierarchical Bayesian procedure (see for instance, Rossi et al. (2005) and Train (2009)). A Bayesian analysis of discrete choice models, however, is hindered by the fact that there is no naturally intuitive conjugate prior which would allow analytical expressions for the posterior distributions of the model parameters. Hence, a Bayesian analysis will have to take recourse to numerical methods. For relatively low dimensional problems the necessary integrals could be computed using Gaussian quadrature. This, however, becomes quickly infeasible when the dimensionality of the unknown parameter vector increases. Therefore, most researchers have turned to stochastic approximations (this also holds for maximum likelihood where many models are estimated with maximum simulated likelihood (Train 2009)). Typically this is done with a Markov chain Monte Carlo (MCMC) simulation where one draws dependent samples from the posterior distribution of the unknown model parameters and latent variables. Although this approach works well in theory there are some caveats in practice. A properly specified Markov chain is guaranteed to converge to its equilibrium distribution but this may take a long time, especially in large complicated models. Furthermore, the assessment of convergence is non-trivial. Another practical difficulty can be the storage cost. Large models with subject specific parameters require ever more

storage as the number of draws from the posterior increases and the number of subjects increases (Braun and McAuliffe 2010). Nowadays, it is not unheard of to use subsets of the available data for inference in large discrete choice models which can lead to biased estimation (Zanutto and Bradlow 2006). To overcome these practical problems, i.e. computation time, convergence assessment and storage, one can turn to variational approximation methods. Applied in a Bayesian context, variational Bayes (VB) optimizes a well defined functional in order to approximate the posterior distribution. The aim is to find a tractable distribution of the model parameters and latent variables that minimizes some measure of distance between the true posterior distribution (as is sampled from in a properly specified MCMC chain) and the approximate posterior distribution. The distance measure is usually taken to be the Kullback-Leibler divergence. As the inference problem is now transformed into an optimization problem, convergence is generally faster and much easier to assess. Furthermore, the storage requirements are much more modest than for MCMC. For instance, if the posterior of a particular K-dimensional parameter vector is approximated by a multivariate normal distribution one would require the storage of K numbers for the posterior mean and $\frac{K(K+1)}{2}$ numbers for the unique elements of the posterior covariance matrix. In an MCMC scheme however, one requires the storage of $R \times \left(K + \frac{K(K+1)}{2}\right)$ numbers where R represents the number of draws from the posterior. As R is typically in the order of thousands (or more), one can readily see that the difference in storage capacity can be quite large. The downside of the variational approach is that it is necessarily biased as the true posterior is approximated by a (much) simpler, more tractable distribution. The amount of bias can be reduced by allowing more complicated approximate posterior distributions but this more or less defeats the purpose. The most common approach to variational Bayesian approximations is to factorize the posterior distribution in a product of more tractable distributions. This approach is also employed in this paper. As variational approximations are relatively new to the statistical literature, there are not many results yet on their statistical properties. Wang and Titterington (2006) investigated the convergence of factorized, or so called mean-field, variational Bayesian approximations in Gaussian finite mixture models. They showed that asymptotically, as the sample size grows, their estimators converge locally to the maximum likelihood estimator. Wang and Titterington (2005), however, showed that the resulting variance estimates are too narrow compared to maximum likelihood which leads to over-optimistic inference. This phenomenon has been observed by many researchers working with factorized approximations, see for instance Bishop (2006), Consonni and Marin (2007) and Rue et al. (2009). More recently Hall et al. (2011) and Ormerod and Wand (2012) investigated the properties of Gaussian vari-

ational approximations (approximating the distribution of random effects by normal distributions) of maximum likelihood estimation in Poisson mixed models and generalized linear mixed models respectively and proved some consistency results for simple models. In the context of mixed discrete choice models Braun and McAuliffe (2010) used a similar approach in a Bayesian framework. They empirically investigated the performance of these approximations and showed that their approach performed similar to MCMC but their methods were significantly faster and required far less memory.
The goal of this paper is to assess the accuracy of several variational approximations in the context of the very popular mixed logit model. We use several bounds which have never been used for these types of models. Furthermore, we propose a particular approximation based on quasi Monte Carlo sampling which will be shown to work very well. In the following section we will briefly introduce the conditional and the mixed logit model. After that, in the subsequent section we will introduce variational Bayesian approximations for these models. Then, we will present the results of several simulation studies on synthetic data which will be followed by a conclusion.

Logit Models for Discrete Choice

In this section we will briefly introduce the conditional and mixed logit model. In this paper the subjects, also called agents or decision makers, will be denoted by index h going from 1 to H. Each of these subjects is faced with T_h choice sets. This could, for instance, be multiple purchases by the same agent at different time points. Or, in case of stated preferences, this represents the number of hypothetical situations in which the subject is asked to make a choice. Furthermore, we will assume that each choice set is defined by a finite number of J alternatives, indexed by $j = 1, \ldots, J$. Each of these alternatives is characterized by K attributes. The values of these attributes are stored in K-dimensional vectors $\boldsymbol{x}_{htj} = (x_{htj1}, \ldots, x_{htjK})^T$ which contain the K attribute values of alternative j, encountered by subject h at choice situation t. We can collect all J attribute vectors in a J by K-dimensional matrix, denoted by $\boldsymbol{X}_{ht}$ which is called the design matrix of choice set t for subject h. The result of the subjects' decisions is stored in J-dimensional binary vectors $\boldsymbol{y}_{ht} = (y_{ht1}, \ldots, y_{htJ})^T$. In these binary vectors a 1 indicates the chosen alternative and the non-chosen alternatives are indicated by 0s. Hence, each of these vectors contains exactly one 1. As the dependent variable is a binary vector we can adequately model it with a multinomial distribution. Furthermore, we will assume that the choice probabilities are functions of a linear combination of the alter-

natives attributes and the tastes of the subject. A subjects taste is represented by a K-dimensional vector $\boldsymbol{\beta}_h$ which contains the relative importances of each attribute for subject h. All that is required now is a link function to link the probabilities (which must be non-negative and sum to 1) and the linear predictor $\boldsymbol{x}_{htj}^T\boldsymbol{\beta}$. For logit models this is the logit link and this leads to the following expression for the probability that subject h, at choice point t, selects the jth alternative

$$P\left(y_{htj}=1|\boldsymbol{x}_{htj},\boldsymbol{\beta}\right)=p_{htj}=\frac{e^{\boldsymbol{x}_{hyj}^T\boldsymbol{\beta}}}{\sum_{j'=1}^{J}e^{\boldsymbol{x}_{htj'}^T\boldsymbol{\beta}}}. \tag{1}$$

This is the conditional logit model introduced by McFadden (1974). In a full Bayesian analysis we require a prior distribution on the unknown parameter vector $\boldsymbol{\beta}$ which is generally taken to be multivariate normal with mean vector $\boldsymbol{\zeta}$ and covariance matrix $\boldsymbol{\Omega}$. The full specification of a Bayesian conditional logit model is then:

$$\begin{aligned}\boldsymbol{y}_{ht}|\boldsymbol{X}_{ht},\boldsymbol{\beta}&\sim\text{Multinomial}\left(p_{ht1},\ldots,p_{htJ}\right), \qquad h=1,\ldots,H\\ \boldsymbol{\beta}|\boldsymbol{\zeta},\boldsymbol{\Omega}&\sim\mathcal{N}_K\left(\boldsymbol{\zeta},\boldsymbol{\Omega}\right).\end{aligned}$$

Note that in this specification all subjects have the same taste vector. The assumption that the tastes are homogeneous in the population is a fairly restrictive assumption which is usually far from true. Furthermore, when subjects make multiple choices it seems hard to argue that these observations are independent. A popular approach to overcome these shortcomings of the conditional logit model is to allow taste heterogeneity among the subjects, i.e. each subject h has his own personal taste vector $\boldsymbol{\beta}_h$. One could estimate these personal taste vectors by estimating H conditional logit models, one for each subject. This would be a good approach if each subject makes a large number of choices. However, one usually only observes a limited number of choices per subject which would make the resulting inferences very noisy. A way to overcome this is by specifying a distribution of these tastes, usually a multivariate normal distribution. In this scenario, each subject's specific taste parameters are estimated by taking the other tastes into account which is a way of borrowing strength from the other observations. Furthermore, the parameters of the mixing distribution need to be estimated and are usually of prime interest to the researcher. Observations made by the same subjects are now no longer independent which adds to the plausibility of the model. In this paper we will assume a multivariate normal distribution as the mixing distribution of the tastes in the population. As before, in a fully Bayesian analysis we require prior distributions on the mean

vector and the covariance matrix of the mixing distribution. We will assume the typical conjugate priors, i.e. a normal prior for the mean and an inverse-Wishart distribution for the covariance matrix. The full mixed logit model is then [1]:

$$\begin{aligned}
\boldsymbol{y}_{ht}|\boldsymbol{X}_{ht},\boldsymbol{\beta}_h &\sim \text{Multinomial}\left(p_{ht1},\ldots,p_{htJ}\right), \quad h=1,\ldots,H, t=1,\ldots,T_h \\
\boldsymbol{\beta}_h|\boldsymbol{\zeta},\boldsymbol{\Omega} &\sim \mathcal{N}_K\left(\boldsymbol{\zeta},\boldsymbol{\Omega}\right), \quad h=1,\ldots,H \\
\boldsymbol{\zeta}|\boldsymbol{\beta}_0,\boldsymbol{\Omega}_0 &\sim \mathcal{N}_K\left(\boldsymbol{\beta}_0,\boldsymbol{\Omega}_0\right) \\
\boldsymbol{\Omega}|\boldsymbol{S}^{-1},\nu &\sim \mathcal{W}^{-1}\left(\boldsymbol{S}^{-1},\nu\right).
\end{aligned}$$

Now that the model is fully specified, a Bayesian analysis proceeds by obtaining the posterior distribution of the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\zeta}, \boldsymbol{\Omega}, \boldsymbol{\beta}_{1:H})$ which is given by

$$p\left(\boldsymbol{\zeta},\boldsymbol{\Omega},\boldsymbol{\beta}_{1:H}|\mathcal{D}\right) = \frac{p\left(\boldsymbol{\zeta}\right)p\left(\boldsymbol{\Omega}\right)\prod_{h=1}^{H}p\left(\boldsymbol{\beta}_h|\boldsymbol{\zeta},\boldsymbol{\Omega}\right)\prod_{t=1}^{T_h}p\left(\boldsymbol{y}_{ht}|\boldsymbol{X}_{ht}\boldsymbol{\beta}_h\right)}{\int p\left(\boldsymbol{\zeta}\right)p\left(\boldsymbol{\Omega}\right)\prod_{h=1}^{H}p\left(\boldsymbol{\beta}_h|\boldsymbol{\zeta},\boldsymbol{\Omega}\right)\prod_{t=1}^{T_h}p\left(\boldsymbol{y}_{ht}|\boldsymbol{X}_{ht}\boldsymbol{\beta}_h\right)d\boldsymbol{\zeta}d\boldsymbol{\Omega}d\boldsymbol{\beta}_{1:H}}$$

where $\mathcal{D}$ represents the observed data. Even though we use (conditionally) conjugate priors, the denominator is not analytically integrable and we will have to resort to numerical approximations. Furthermore, to obtain marginal posterior distributions of the parameters of the mixing distribution one also requires numerical approximations to integrate the $\boldsymbol{\beta}_h$'s from the numerator.

Variational Bayesian Approximation

Variational Bayes

Variational Bayesian approximations cover a wide set of methods to approximate the posterior distribution. Recent tutorials and literature overviews can be found in Bishop (2006), Ormerod and Wand (2010) and Titterington (2011). The main idea is that one tries to approximate the posterior distribution with a simpler distribution. So, how does one select such a simpler distribution and how does one evaluate how well it approximates the posterior? There are of course several possibilities but most often one tries to minimize the Kullback-Leibler divergence between the approximating distribution $q\left(\boldsymbol{\theta}\right)$, where $\boldsymbol{\theta}$ contains all the unknowns in the model, and the posterior distribution $p\left(\boldsymbol{\theta}|\mathcal{D}\right)$ where $\mathcal{D}$ refers to the observed data. To start we rewrite the Kullback-Leibler divergence

of $p(\boldsymbol{\theta}|\mathcal{D})$ from $q(\boldsymbol{\theta})$ as

$$\begin{aligned} KL(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D})) &= \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})} d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta}) p(\mathcal{D})}{p(\boldsymbol{\theta}, \mathcal{D})} d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}, \mathcal{D})} d\boldsymbol{\theta} + \log p(\mathcal{D}). \end{aligned}$$

The right term on the last line is the natural logarithm of the marginal likelihood of the data under the model, sometimes called the evidence, and it is independent of the model parameters. The left term of the last line is the Kullback-Leibler divergence of the joint distribution of the data and the unknowns from the approximating distribution. We can rewrite this equation and obtain the following decomposition

$$\begin{aligned} \log p(\mathcal{D}) &= \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} + KL(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D})) \\ &\geq \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} = \mathcal{L}(q(\boldsymbol{\theta})) \end{aligned}$$

since $KL(.||.)$ is non-negative. We have now obtained a lower bound $\mathcal{L}(q(\boldsymbol{\theta}))$ on the logarithm of the marginal data likelihood. Making $q(\boldsymbol{\theta})$ as equal (in the Kullback-Leibler divergence sense) to the posterior as possible can now be seen to be equivalent with maximizing this bound with respect to $q(\boldsymbol{\theta})$. Using classical calculus of variations one can show that this optimization results into an optimal $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathcal{D})$. So we have not made much headway yet as we start out from an intractable posterior $p(\boldsymbol{\theta}|\mathcal{D})$. We can, however, put restrictions on $q(\boldsymbol{\theta})$. The most common type of restriction is to factorize the approximating distribution which is known as mean-field theory in physics (Parisi 1988). For the mixed logit model, for instance, we could restrict our approximation, with $\boldsymbol{\theta} = (\boldsymbol{\zeta}, \boldsymbol{\Omega}, \boldsymbol{\beta}_{1:H})$, as $q(\boldsymbol{\zeta}, \boldsymbol{\Omega}, \boldsymbol{\beta}_{1:H}) = q(\boldsymbol{\zeta}) \times q(\boldsymbol{\Omega}) \times \prod_{h=1}^{H} q(\boldsymbol{\beta}_h)$. It can be shown that in fully conjugate exponential family models, the optimal approximate $q(.)$ densities are in the same family as their priors (Winn and Bishop 2005; Ormerod and Wand 2010). From this we can already deduce that $q(\boldsymbol{\zeta}) \sim \mathcal{N}_K(\boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta)$ and $q(\boldsymbol{\Omega}) \sim \mathcal{W}^{-1}(\boldsymbol{\Upsilon}^{-1}, \omega)$. The posterior approximate densities $q(\boldsymbol{\beta}_h), h = 1, \ldots, H$, however, are not conjugate due to the logit link. A natural way to parameterize them is to assume they are independent multivariate normal densities, i.e. $q(\boldsymbol{\beta}_h) \sim \mathcal{N}_K(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h), h = 1, \ldots, H$, which would be conjugate if one approximates the logistic terms by quadratic functions of $\boldsymbol{\beta}_h$ as their prior is a normal distribution, $\mathcal{N}_K(\zeta, \Omega)$. This factorization was also employed in Braun

and McAuliffe (2010). In the terminology of Ormerod and Wand (2010), this is a mix of a product density transform or a mean-field approximation (factorized approximate posterior) and a parametric density transform (assuming a specific parametric posterior). As mentioned before, the drawback of this assumed factorization is that posterior dependencies are lost between the different sets of parameters. This results in posterior second moments being too small and hence posterior credible intervals can be (much) too narrow showing inflated confidence. Further on, this approach will be mixed with yet another approach, the tangent transform, where a tangent lower bound is placed on the variational lower bound.
With factorized approximate posterior distributions one generally maximizes the lower bound by a coordinate ascent algorithm. As will be shown in the next section, each parameter vector depends on (subsets) of the other parameter vectors. Typically, one initializes several sets of parameters and then cycles through the update equations until some measure of convergence is satisfied. Hence, at each step, all other parameter vectors are held constant while one particular parameter vector is found to maximize the resulting lower bound. When there is conjugacy, these updates are usually closed form and can be performed efficiently. When there is no conjugacy on the other hand, it might be necessary to use numerical optimization.

Variational Bayes for the Mixed Logit Model

We have seen before that to obtain a variational approximation one can maximize

$$\mathcal{L}\left(q\left(\boldsymbol{\theta}\right)\right) = \int q\left(\boldsymbol{\zeta}\right) q\left(\boldsymbol{\Omega}\right) \prod_{h=1}^{H} q\left(\boldsymbol{\beta}_h\right) \log \frac{p\left(\boldsymbol{\zeta}, \boldsymbol{\Omega}, \boldsymbol{\beta}_{1:H}, \mathcal{D}\right)}{q\left(\boldsymbol{\zeta}\right) q\left(\boldsymbol{\Omega}\right) \prod_{h=1}^{H} q\left(\boldsymbol{\beta}_h\right)} d\boldsymbol{\zeta} d\boldsymbol{\Omega} \prod_{h=1}^{H} d\boldsymbol{\beta}_h$$

$$= \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log p\left(\boldsymbol{\zeta}, \boldsymbol{\Omega}, \boldsymbol{\beta}_{1:H}, \mathcal{D}\right)\right] + \sum_{h=1}^{H} H\left[q\left(\boldsymbol{\beta}_h\right)\right] + H\left[q\left(\boldsymbol{\zeta}\right)\right] + H\left[q\left(\boldsymbol{\Omega}\right)\right] \quad (2)$$

with respect to the parameters of the variational posterior distribution. The first part of the right hand side is the expectation of the log joint probability of the data and the parameters (log likelihood plus log prior) with respect to the variational posterior distribution and the $H\left[q\left(.\right)\right]$ terms are the differential entropies of the variational posterior[2]. Plugging the known density families into (2) we obtain the following expression which

needs to be maximized with respect to the variational parameters

$$\begin{aligned}\mathcal{L}\left(q\left(\boldsymbol{\theta}\right)\right) = &\sum_{h=1}^{H}\sum_{t=1}^{T_h}\left\{\boldsymbol{y}_{ht}^T\boldsymbol{X}_{ht}\mathbb{E}_{\mathcal{N}_K(\boldsymbol{\beta}_h;\boldsymbol{\zeta},\boldsymbol{\Omega})}\left[\boldsymbol{\beta}_h\right] - \mathbb{E}_{\mathcal{N}_K(\boldsymbol{\beta}_h;\boldsymbol{\zeta},\boldsymbol{\Omega})}\left[\log\left(\sum_{j=1}^{J}e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right)\right]\right\} \\ &+ \sum_{h=1}^{H}\mathbb{E}_{\mathcal{N}_K\left(\boldsymbol{\zeta};\boldsymbol{\mu}_{\boldsymbol{\zeta}},\boldsymbol{\Omega}_{\boldsymbol{\zeta}}\right)\mathcal{W}^{-1}\left(\boldsymbol{\Omega};\boldsymbol{\Upsilon}^{-1},\omega\right)\mathcal{N}_K(\boldsymbol{\beta}_h;\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h)}\left[\log\mathcal{N}_K\left(\boldsymbol{\beta}_h;\boldsymbol{\zeta},\boldsymbol{\Omega}\right)\right] \\ &+ \mathbb{E}_{\mathcal{W}^{-1}\left(\boldsymbol{\Omega};\boldsymbol{\Upsilon}^{-1},\omega\right)}\left[\log\mathcal{W}^{-1}\left(\boldsymbol{\Omega};\boldsymbol{\Upsilon}^{-1},\omega\right)\right] + \mathbb{E}_{\mathcal{N}_K\left(\boldsymbol{\zeta};\boldsymbol{\mu}_{\boldsymbol{\zeta}},\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}\right)}\left[\log\mathcal{N}_K\left(\boldsymbol{\zeta};\boldsymbol{\mu}_{\boldsymbol{\zeta}},\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}\right)\right] \\ &+ \sum_{h=1}^{H}H\left[q\left(\boldsymbol{\beta}_h\right)\right] + H\left[q\left(\boldsymbol{\zeta}\right)\right] + H\left[q\left(\boldsymbol{\Omega}\right)\right]. \end{aligned} \tag{3}$$

Because the variational posterior distribution is factorized all but one of the expectations in this expression are fairly simple to evaluate. Plugging these expectations into (3) one can calculate derivatives of the lower bound with respect to $\boldsymbol{\mu}_{\boldsymbol{\zeta}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}$, ω and $\boldsymbol{\Upsilon}$ and equate them to $\mathbf{0}$. This yields closed form update equations for these sets of parameters. More details on this can be found in appendix A. The only parts of (3) which are troublesome are the updates with respect to $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ for all $h = 1, \ldots, H$. The expected value of the log-sum of exponentials in equation (3) has no analytically closed form. Hence, we have no analytical form in function of the variational parameters over which we can maximize. In the following subsections we will list a number of possible avenues to deal with this problem. Some of these solutions try to approximate the problematic expectation in terms of parameters of the variational posterior. Other solutions bound this expectation in various ways which results into a lower bound on the lower bound. Some of these avenues lead to closed form update equations while others require numeric optimization.

Approximating or Bounding the Log-Sum-Exponential Function

There are $H \times T_h$ terms in equation (3) which have no analytical expectations with respect to a normal distribution:

$$LSE\left(\boldsymbol{X}_{ht}\boldsymbol{\beta}_h\right) = \log\left(\sum_{j=1}^{J}e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right), \qquad \forall h = 1, \ldots, H, t = 1, \ldots, T_h. \tag{4}$$

Therefore, these terms do not allow analytic expressions in terms of the means, $\boldsymbol{\mu}_h$, and covariance matrices, $\boldsymbol{\Sigma}_h$, of the approximate posterior distribution. In order to optimize the lower bound with respect to these parameters one has to replace equation (4) with

an approximation or a bound which can be expressed in terms of the individual agents' parameters. Here we will describe several approaches to do this. Details beyond this exposition can be found in appendix B.

Taylor series. An approach which was proposed by Braun and McAuliffe (2010) is to replace (4) by a second order Taylor series expansion around the current mean $\boldsymbol{\mu}_h$ and taking the expectation which yields

$$\mathbb{E}_{q(\boldsymbol{\beta}_h)}\left[\log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right)\right] \approx \log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\mu}_h}\right) + \frac{1}{2}tr\left[\boldsymbol{\Sigma}_h\boldsymbol{X}_{ht}^T\text{diag}\left(\boldsymbol{p}_{ht}\right)\boldsymbol{X}_{ht}\right] - \frac{1}{2}tr\left[\boldsymbol{\Sigma}_h\boldsymbol{X}_{ht}^T\boldsymbol{p}_{ht}\boldsymbol{p}_{ht}^T\boldsymbol{X}_{ht}\right]$$

where $\boldsymbol{p}_{ht}$ is a J-dimensional vector with entries $\frac{e^{\boldsymbol{x}_{htj}^T\boldsymbol{\mu}_h}}{\sum_{j'=1}^{J} e^{\boldsymbol{x}_{htj'}^T\boldsymbol{\mu}_h}}, \forall j = 1,\ldots,J$ and $\text{diag}\left(\boldsymbol{x}\right)$ is an operator that constructs a diagonal matrix out of the elements in $\boldsymbol{x}$. The resulting expression, plugged into equation (3), can now be optimized with respect to $(\boldsymbol{\mu}_{1:H}, \boldsymbol{\Sigma}_{1:H})$ but does not allow a closed form update equation. Hence, the coordinate ascent algorithm will require a numeric maximization step using any efficient optimization algorithm. Note that using this approach there is no longer a guarantee that the function which is maximized remains a lower bound and the resulting approximate posterior is therefore no longer guaranteed to be the closest approximate posterior (with the chosen factorization and parameterization) to the real posterior distribution with respect to the Kullback-Leibler divergence. Nevertheless, Braun and McAuliffe (2010) empirically showed that the resulting inference is very close to $MCMC$ and is therefore useful. Note also that Braun and McAuliffe (2010) used another simplification in that they restricted the subjects posterior covariances, $\boldsymbol{\Sigma}_h, \forall h = 1,\ldots,H$, to diagonal matrices. In this paper we use both the unrestricted and the restricted approach. The unrestricted approach treats the covariance matrices as dense matrices and is denoted by BM. The restricted approach restricts the subjects covariance matrices to diagonal matrices and is denoted by BM_D.

Quasi Monte Carlo. A different, viable approach would be to approximate the expectation by a Monte Carlo method. Lawrence et al. (2004) used importance sampling to obtain approximations to an intractable expectation within their variational algorithm to improve grid placement for the analysis of DNA microarray data. Girolami and Rogers (2006) also used importance sampling for their variational Bayesian treatment of multinomial probit regression with Gaussian process priors. In this paper, however, we pro-

pose to make us of the considerable research in the field of Quasi Monte Carlo (QMC). Quasi Monte Carlo samples are samples which are constructed to proportionally fill the high density regions of the distribution from which one wishes to sample. The resulting approximate integration can then be performed in a stable manner with fewer samples than with regular Monte Carlo. Inspired by Yu et al. (2010) we use the extensible shifted lattice points (ESLP) algorithm as proposed by Hickernell et al. (2000). This algorithm scales well with the dimensionality of the integral. For details on generating such QMC samples we refer you to the previous two references and appendix C. We thus approximate the expectation of (4) as:

$$\mathbb{E}_{q(\boldsymbol{\beta}_h)}\left[\log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right)\right] \approx \frac{1}{R}\sum_{r=1}^{R}\log\left(\sum_{j=}^{J} e^{\boldsymbol{x}_{htj}^T\left(\boldsymbol{L}_h\boldsymbol{z}^{(r)}+\boldsymbol{\mu}_h\right)}\right)$$

where $\boldsymbol{L}_h$ is the lower triangular Cholesky factor of $\boldsymbol{\Sigma}_h$ such that $\boldsymbol{\Sigma}_h = \boldsymbol{L}_h\boldsymbol{L}_h^T$, R is the number of QMC samples from a standard multivariate normal distribution and each draw is represented by $\boldsymbol{z}^{(r)}$. This expression once again does not allow for analytic update equations and hence requires numeric optimization. For a small number of draws the lower bound can also not be guaranteed but this can be alleviated by increasing the sample size. The number of QMC samples we used is determined by a parameter m which results in a sample size of $R = 2^m$. We used several values for m in the range of $6 - 12$ and found that $R = 2^6 = 64$ worked well enough in our applications. Furthermore, we also considered a restricted version where the subjects covariance matrices were restricted to diagonal matrices. We refer to these respective approaches as QMC and QMC_D.

Jensen's inequality. A different approach to approximate the expectation of (4) was used by Blei and Lafferty (2007), based on Jensen's inequality, in the context of topic models. As log(.) is a concave function one can simply apply Jensen's inequality to obtain:

$$\mathbb{E}_{q(\boldsymbol{\beta}_h)}\left[\log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right)\right] \leq \log\left(\sum_{j=1}^{J}\mathbb{E}_{q(\boldsymbol{\beta}_h)}\left[e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right]\right) = \log\left(\sum_{j=1}^{K} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\mu}_h+\frac{1}{2}\boldsymbol{x}_{htj}^T\boldsymbol{\Sigma}_h\boldsymbol{x}_{htj}}\right).$$

Again we considered an unrestricted and a restricted version and we denote these respective methods as JI and JI_D. Knowles and Minka (2011) improved the flexibility of the former bound by introducing additional parameters $\boldsymbol{a} = (\boldsymbol{a}_{1:H,1:T_h})$ and $\boldsymbol{a}_{ht}$ a J-

dimensional vector. Their bound results in

$$\mathbb{E}_{q(\boldsymbol{\beta}_h)}\left[\log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right)\right] \leq \sum_{j=1}^{J} a_{htj}\boldsymbol{x}_{htj}^T\boldsymbol{\mu}_h + \log\left(\sum_{j=1}^{J} e^{\left(\boldsymbol{x}_{htj}-\sum_{j=1}^{J} a_{htj}\boldsymbol{x}_{htj}\right)^T\boldsymbol{\mu}_h+\frac{1}{2}\left(\boldsymbol{x}_{htj}-\sum_{j=1}^{J} a_{htj}\boldsymbol{x}_{htj}\right)^T\boldsymbol{\Sigma}_h\left(\boldsymbol{x}_{htj}-\sum_{j=1}^{J} a_{htj}\boldsymbol{x}_{htj}\right)}\right).$$

The additional flexibility tightens the inequality at the expense of introducing more parameters which need to be optimized. We again consider an unrestricted and a restricted version of the subjects' covariance matrices. These approaches, denoted here by KM and KM_D respectively, require an additional step in the coordinate ascent algorithm to update the extra HTJ-dimensional parameter vector $\boldsymbol{a}$. Both JI and KM, bound the required expectation and hence they keep the lower bound property of the variational algorithm intact. Both require numeric optimization in each iteration in order to maximize the subjects' variational parameters.

Bï¿½hning-Lindsay. A different approach, which does not require numeric optimization is to introduce a quadratic approximation to (4). The first quadratic approximation we consider is due to a bound on the second order Taylor series expansion of equation (4) around an extra J-dimensional vector of variational parameters $\boldsymbol{\Psi}_{ht}$. The resulting expectation is then

$$\mathbb{E}_{q(\boldsymbol{\beta}_h)}\left[\log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right)\right] \leq \log\left(\sum_{j=1}^{J} e^{\Psi_{htj}}\right) + \left(\boldsymbol{X}_{ht}\boldsymbol{\mu}_h - \boldsymbol{\Psi}_{ht}\right)^T\nabla\left(\boldsymbol{\Psi}_{ht}\right) + \frac{1}{2}\left[\boldsymbol{X}_{ht}^T\boldsymbol{A}\boldsymbol{X}_{ht}\boldsymbol{\Sigma}_h\right] + \frac{1}{2}\left(\boldsymbol{X}_{ht}\boldsymbol{\mu}_h - \boldsymbol{\Psi}_{ht}\right)^T\boldsymbol{A}\left(\boldsymbol{X}_{ht}\boldsymbol{\mu}_h - \boldsymbol{\Psi}_{ht}\right).$$

where $\boldsymbol{A} = \frac{1}{2}\left(\boldsymbol{I}_J - \mathbf{1}_J\mathbf{1}_J^T/J\right)$, $\boldsymbol{I}_J$ is the J-dimensional identity matrix and $\mathbf{1}_J$ is a J-dimensional vector of ones and with $\boldsymbol{\Psi}_{ht}$ a J-dimensional vector of extra variational parameters. Note also that $\nabla\left(\boldsymbol{\Psi}_{ht}\right)$ is the gradient of equation (4) evaluated at $\boldsymbol{\Psi}_{ht}$. This quadratic bound follows from a result from Böhning and Lindsay (1988) and Böhning (1992). We denote this method by BL. This bound has been successfully used by Khan et al. (2010) for a variational treatment of mixed-data factor analysis due to its computational efficiency.

Bouchard. A different quadratic bound we considered is due to Bouchard (2007) which is a multinomial generalization of a quadratic bound developed by Jaakkola and Jordan

(2000). This bound also introduces extra variational parameters $(\alpha_{1:H,1:T_h}, t_{1:H,1:T_h,1:J})$ and is

$$\mathbb{E}_{q(\boldsymbol{\beta}_h)}\left[\log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right)\right] \leq \alpha_{ht} + \sum_{j=1}^{J}\frac{\boldsymbol{x}_{htj}^T\boldsymbol{\mu}_h - \alpha_{ht} - t_{htj}}{2} + \lambda\left(t_{htj}\right)\left[\left(\boldsymbol{x}_{htj}^T\boldsymbol{\mu}_h - \alpha_{ht}\right)^2 - t_{htj}^2 + \boldsymbol{x}_{htj}^T\boldsymbol{\Sigma}_h\boldsymbol{x}_{htj}\right] + \log\left(1+e^{t_{htj}}\right)$$

where $\lambda(t) = \frac{1}{4t}\tanh\left(\frac{t}{2}\right)$. This bound also leads to closed form updates for the subjects' parameters, conditional on the optimal HT-dimensional vector of $\boldsymbol{\alpha} = \alpha_{1:H,1:T_h}$ and HTJ-dimensional vector $\boldsymbol{t} = t_{1:H,1:T_h,1:J}$. We used this bound in several experimental settings but found that it was way too loose and yielded very biased approximations for the posterior mean parameters. Therefore we will not show any results of this bound but it is included here for completeness sake.

Jebara-Choromanska. The final quadratic approach considered is due to Jebara and Choromanska (2012) who developed an algorithm to find a quadratic bound around some $\tilde{\boldsymbol{\beta}}_h$ to equation (4) which tightens BL and generalizes BO. After taking expectations this bound leads to

$$\mathbb{E}_{q(\boldsymbol{\beta}_h)}\left[\log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right)\right] \leq \log z_{ht} + \frac{1}{2}\left(\boldsymbol{\mu}_h - \tilde{\boldsymbol{\beta}}_h\right)^T\boldsymbol{S}_{ht}\left(\boldsymbol{\mu}_h - \tilde{\boldsymbol{\beta}}_h\right) + \frac{1}{2}tr\left[\boldsymbol{\Sigma}_h\boldsymbol{S}_{ht}\right] + \left(\boldsymbol{\mu}_h - \tilde{\boldsymbol{\beta}}_h\right)^T\boldsymbol{m}_{ht}$$

where z_{ht} is a scalar, $\boldsymbol{m}_{ht}$ is a K-dimensional vector and $\boldsymbol{S}_{ht}$ is a $K \times K$-dimensional matrix which are determined by the algorithm of Jebara and Choromanska (2012). The algorithm can be found in appendix B. We will denote this approach as JC. It should be noted that all three quadratic approaches considered here maintain the lower bound property of the variational objective function. Furthermore, due to the fact that the subjects' variational parameters can be updated with closed form updates, these algorithms tend to be computationally very efficient. The requirement that the bounds are quadratic, on the other hand, hinders their flexibility.

Final algorithm and comments. Now that all bounds and approximations have been introduced, we can formulate a typical coordinate ascent algorithm which was used to estimate the models in our simulations. As an example we give the algorithm to obtain the QMC approximation to the posterior distribution (more detail on the derivation of this algorithm is provided in the appendices). These algorithms require several user-specified

input decisions, namely the specification of hyperparameters for the prior distributions,

Algorithm 1 *QMC* Input $\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta}_0, \boldsymbol{\Omega}_0, \boldsymbol{S}^{-1}, \nu, \boldsymbol{Z}$

Initialize: $\boldsymbol{\mu}_{1:H}, \boldsymbol{\Sigma}_{1:H}, \boldsymbol{\mu}_{\boldsymbol{\zeta}}, \boldsymbol{\Sigma}_{\boldsymbol{\zeta}}, \omega = \nu + H$, Convergence = False

$\boldsymbol{\Upsilon} = \left\{ \boldsymbol{S}^{-1} + H\boldsymbol{\Sigma}_{\boldsymbol{\zeta}} + \sum_{h=1}^{H} \left[\boldsymbol{\Sigma}_h + \left(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\boldsymbol{\zeta}} \right) \left(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\boldsymbol{\zeta}} \right)^T \right] \right\}^{-1}$

while Convergence = False **do**

for $h = 1$ to $h = H$ **do**

Obtain $\boldsymbol{L}_h$ from $\boldsymbol{L}_h \boldsymbol{L}_h^T = \boldsymbol{\Sigma}_h$

$\boldsymbol{\beta}_h^{(r)} = \boldsymbol{L}_h \boldsymbol{z}^{(r)} + \boldsymbol{\mu}_h, \qquad \forall r = 1, \ldots, R$

$\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h = \underset{\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h}{\arg\max} \sum_{t=1}^{T_h} \boldsymbol{y}_{ht}^T \boldsymbol{X}_{ht} \boldsymbol{\mu}_h - \frac{1}{R} \sum_{r=1}^{R} \log \left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T \boldsymbol{\beta}_h^{(r)}} \right)$
$-\frac{1}{2} tr \left[\omega \boldsymbol{\Upsilon} \boldsymbol{\Sigma}_h \right] - \frac{\omega}{2} \boldsymbol{\mu}_h \boldsymbol{\Upsilon} \boldsymbol{\mu}_h^T + \omega \boldsymbol{\mu}_h^T \boldsymbol{\Upsilon} \boldsymbol{\mu}_{\zeta} + \frac{1}{2} \log |\boldsymbol{\Sigma}_h|$

end for

$\boldsymbol{\Sigma}_{\boldsymbol{\zeta}} = \left(H\omega \boldsymbol{\Upsilon} + \boldsymbol{\Omega}_0^{-1} \right)^{-1}$

$\boldsymbol{\mu}_{\boldsymbol{\zeta}} = \boldsymbol{\Sigma}_{\boldsymbol{\zeta}} \left(\omega \boldsymbol{\Upsilon} \sum_{h=1}^{H} \boldsymbol{\mu}_h + \boldsymbol{\Omega}_0^{-1} \boldsymbol{\beta}_0 \right)$

$\boldsymbol{\Upsilon} = \left\{ \boldsymbol{S}^{-1} + H\boldsymbol{\Sigma}_{\boldsymbol{\zeta}} + \sum_{h=1}^{H} \left[\boldsymbol{\Sigma}_h + \left(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\boldsymbol{\zeta}} \right) \left(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\boldsymbol{\zeta}} \right)^T \right] \right\}^{-1}$

Test Convergence

end while

Output: $\boldsymbol{\mu}_{1:H}, \boldsymbol{\Sigma}_{1:H}, \boldsymbol{\mu}_{\boldsymbol{\zeta}}, \boldsymbol{\Sigma}_{\boldsymbol{\zeta}}, \boldsymbol{\Upsilon}$

the initialization method, the convergence criterion and tolerance. We used uninformative but proper priors. The prior mean and variance of $\boldsymbol{\zeta}$ were taken as respectively a K-dimensional zero vector, $\boldsymbol{\beta}_0 = \mathbf{0}_K$, and 100 times the K-dimensional identity matrix, $\boldsymbol{\Omega}_0 = 100 \times \boldsymbol{I}_K$. An uninformative prior for an inverse-Wishart distribution is somewhat harder to select. We decided to put the prior degrees of freedom at $\nu = K + 3$ and the prior scale matrix at $\boldsymbol{S}^{-1} = 2 \times \boldsymbol{I}_K$, i.e. two times the K-dimensional identity matrix. As such the prior expected value of $\boldsymbol{\Omega}$ is the K-dimensional identity matrix and the variances of all the elements are infinite. All that is left to specify now is the convergence criterion of the algorithm and a decent method to initialize the parameters. As convergence criterion we took the relative change in the joint Euclidean norm of all the variational parameters. The tolerance level for this convergence criterion was set at 10^{-4}. To initialize $\boldsymbol{\mu}_{\zeta}$ and $\boldsymbol{\Sigma}_{\zeta}$ we used the Laplace approximation to the posterior distribution of a regular conditional logit model based on all the data. As such a model ignores possible (and in our case known) heterogeneity we multiplied the resulting initial posterior variance $\boldsymbol{\Omega}_{\zeta}$ by a factor H. We subsequently initialized the agent specific variational parameters $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ by a Laplace approximation of an individual specific conditional logit model with priors given by $\boldsymbol{\mu}_{\zeta}$ and $\boldsymbol{\Sigma}_{\zeta}$. All computations were done using R (R

Core Team 2012) where we also used the *optim* routine to perform the numerical optimization parts with the BFGS algorithm using analytical gradients for all parameters. We now conclude this section with a brief summary of some published results concerning variational algorithms with respect to multinomial logit models. As far as we are aware the different approaches from the previous subsections have not been compared to each other in the context of discrete choice models. Knowles and Minka (2011) tested several bounds (JI_D, KM_D, BL and BO) and the Taylor series approximation (BM_D here) for their tightness with respect to the expected log-sum-exponential function. They found that BL was very loose and that KM_D dominates JI_D. Furthermore, KM_D performed best of all bounds except when the inputs of the log-sum-exponential function were extremely variable. In that instance, BO performed best but that bound performed much worse in all other cases. They also found that BM_D performed best but did not consider it for further experiments due to the fact that it is not a bound. They also tested BO, KM_D and JI_D on some simulated multinomial logit datasets using a slightly different algorithm (non-conjugate variational message passing) than our variational Bayes method and found that KM_D performed best. With respect to the mixed logit model we are not aware of any results except for Braun and McAuliffe (2010) who used JI and BM_D. They did not report results on JI as they found that in their settings, BM_D was much better. In the following sections all these seven approaches will be compared on their performance in the context of mixed logit models.

Numerical Experiments

Performance Assessment

In order to assess the performance of the various variational approaches in the context of a mixed logit model we performed several simulation experiments which will be detailed in the next subsections. As a benchmark for the variational algorithms we used the function *rhierMnlRwMixture* from the *bayesm* package (Rossi 2012) which uses a Gibbs sampler with a random walk Metropolis step which is explained in Rossi et al. (2005, pg. 136-137). Rather than using this $MCMC$ chain to explore the posterior distribution completely we chose to run this algorithm for as long as it took the variational QMC algorithm to converge. This eliminates the need to and the trouble of checking for convergence of the high dimensional $MCMC$ chain for each experiment, which is a non-trivial problem. Once the chain was stopped we removed the first half of the draws as

burn-in. From the resulting draws we only kept every fifth draw, i.e. thinning, for practical convenience resulting in R draws. So for each dataset we obtain seven variational results which are sets of parameters from the approximate parametric posterior distribution and an $MCMC$ sample of size R from (a part of) the posterior distribution. To measure the accuracy of the different approaches we used the same procedure as Braun and McAuliffe (2010) which compares out-of-sample predictions with the true predictive choice distribution. So, suppose a new choice set $\boldsymbol{X}_{\text{new}}$ is presented to an agent and suppose, for the moment, that we know this agent's tastes, $\boldsymbol{\beta_h}$. We can then calculate the predictive choice distribution for this new choice set based on model (1). This predictive choice distribution yields the probabilities that this respective agent selects the various alternatives specified in the new choice set. Most of the time, however, we are not interested in a specific agent's choices but rather in the choice probabilities of the 'average' agent. Hence, we need to integrate these probabilities over the population heterogeneity distribution which yields

$$p^{\text{true}}\left(\boldsymbol{y}_{\text{new}}|\boldsymbol{X}_{\text{new}},\boldsymbol{\zeta},\boldsymbol{\Omega}\right)=\int p\left(\boldsymbol{y}_{\text{new}}|\boldsymbol{X}_{\text{new}},\boldsymbol{\beta}\right)\mathcal{N}_K\left(\boldsymbol{\beta}|\boldsymbol{\zeta},\boldsymbol{\Omega}\right)d\boldsymbol{\beta}. \tag{5}$$

However, unlike in simulation studies, one generally does not know the heterogeneity distribution. One can estimate this distribution however. As we have posterior distributions over model parameters in a Bayesian setting, this will require another set of integrals to integrate over the posterior distributions of the parameters of the mixing distribution. This results then in an estimated predictive choice distribution given by

$$\hat{p}\left(\boldsymbol{y}_{\text{new}}|\boldsymbol{X}_{\text{new}},\mathcal{D}\right)=\int\int p\left(\boldsymbol{y}_{\text{new}}|\boldsymbol{X}_{\text{new}},\boldsymbol{\beta}\right)\mathcal{N}_K\left(\boldsymbol{\beta}|\boldsymbol{\zeta},\boldsymbol{\Omega}\right)q\left(\boldsymbol{\zeta}|\hat{\boldsymbol{\mu}}_{\zeta},\hat{\boldsymbol{\Sigma}}_{\zeta}\right)q\left(\boldsymbol{\Omega}|\hat{\boldsymbol{\Upsilon}}^{-1},\hat{\omega}\right)d\boldsymbol{\beta}d\boldsymbol{\zeta}d\boldsymbol{\Omega}. \tag{6}$$

In order to calculate the true predictive choice distribution in (5) we averaged the choice probabilities over 1000000 draws of $\boldsymbol{\beta}$ from the known, true heterogeneity distribution, $\mathcal{N}_k\left(\boldsymbol{\zeta},\boldsymbol{\Omega}\right)$. Braun and McAuliffe (2010) used this sample size to ensure that the Monte Carlo error of this estimation is negligible compared to the variability of their results. To calculate the estimated predictive choice distribution for the variational results in (6) we generate 500 samples of $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$ from $q\left(\boldsymbol{\zeta}|\boldsymbol{\mu_\zeta},\boldsymbol{\Sigma_\zeta}\right)q\left(\boldsymbol{\Omega}|\boldsymbol{\Upsilon}^{-1},\omega\right)$. For each of these 500 samples we draw 10000 $\boldsymbol{\beta}$ vectors to evaluate the estimated predictive choice distribution. The average of these 5000000 predictive choice distributions is then the estimated predictive choice distribution. Similarly, for the $MCMC$ results, we use 10000 $\boldsymbol{\beta}$ samples for each of the R draws from the posterior to obtain the estimated predictive choice

distribution. The true predictive choice distribution and the estimated predictive choice distribution were always assessed for 25 new randomly generated choice sets. To compare the true and the estimated predictive choice distributions we used the total variation metric which is a metric that compares probability distributions. The total variation error can then be calculated as(Levin et al. 2009)

$$\begin{aligned}&\mathrm{TV}\left[p^{\mathrm{true}}\left(\boldsymbol{y}_{\mathrm{new}}|\boldsymbol{X}_{\mathrm{new}},\boldsymbol{\zeta},\boldsymbol{\Omega}\right),\hat{p}\left(\boldsymbol{y}_{\mathrm{new}}|\boldsymbol{X}_{\mathrm{new}},\mathcal{D}\right)\right]\\&\qquad=\frac{1}{2}\sum_{j=1}^{J}\left|p_j^{\mathrm{true}}\left(\boldsymbol{y}_{\mathrm{new}}|\boldsymbol{X}_{\mathrm{new}},\boldsymbol{\zeta},\boldsymbol{\Omega}\right)-\hat{p}_j\left(\boldsymbol{y}_{\mathrm{new}}|\boldsymbol{X}_{\mathrm{new}},\mathcal{D}\right)\right|.\end{aligned}$$

This error is contained in the interval $[0,1]$ and obviously smaller errors are preferred. For each simulation scenario this metric was calculated for all the replications and the reported results are based on the median total variation error over the 25 new choice sets.

Uncorrelated Taste Parameters

In this simulation study five experimental factors were varied. The number of decision makers, H, was considered to be 250 or 1000. The number of choice sets per agent were taken as $T_h = 1, 5, 15$ or 25. The number of alternatives J was either 3 or 12 and the number of attributes K was 3 or 10. Finally there was a setting with a relatively high population taste heterogeneity where the true $\boldsymbol{\Omega}$ was set equal to the K-dimensional identity matrix, $\boldsymbol{I}_K$. In the relatively low taste heterogeneity setting the true $\boldsymbol{\Omega}$ was set to 0.25 times the K-dimensional identity matrix, $0.25 \times \boldsymbol{I}_K$. The true mean $\boldsymbol{\zeta}$ was set at K equally spaced values between -2 and 2. Finally, the attribute values were independent identically distributed normal variables, $\mathcal{N}\left(0, 0.5^2\right)$. Each of these simulation settings was replicated 10 times.

-Insert Figure 1 about here-

-Insert Figure 2 about here-

The results of these simulations can be found in figures 1 and 2 which show the average total variation error and average completion time (in minutes) over the 10 replications for all experimental settings. In the cases when the variational algorithm was run with a diagonal version of the decision makers' covariance matrices and with an unrestricted

version, i.e. dense covariance matrices, we only report the restricted results. The reason for this will be explained in a following subsection. We initially included the BO approach in the simulation but stopped this early as this method yielded very biased estimates for the posterior means of the subjects which indicates that the bound is too loose to work properly in these experimental settings. Note also that the figures do not include timing information for the $MCMC$ chains. As these chains were generally not run until convergence, these times are not very interesting here. Just know that these chains were run for as long as the QMC approach was run, which is, generally somewhat slower than the QMC_D version. Looking at figure 1 we can see that there are four distinct clusters with respect to accuracy. We can clearly see that in most settings $MCMC$ performs some orders of magnitude worse than the other algorithms which is an indication that it did not have enough time to fully explore the posterior distribution. The difference is smallest when the number of choice sets per agent is small, i.e. when there is not much information per decision maker in the data. The Taylor series approximation, BM, on the other hand performs very well when the number of choice sets is relatively large but very poorly when the number is low. This shows that this approximation can be very inaccurate when there is not a lot of information per agent. Furthermore, we can discern two other distinct groups. The group with KM and QMC is clearly the most accurate overall which is followed by the group with BL, JC and JI. We can also see that the difference between these two groups increases when the sample size H increases. Finally, it can also be seen that generally the accuracy is higher when there are fewer attributes, i.e. fewer parameters in the model, and when there are more choice sets per agent. Looking at figure 2 many of these patterns reappear. The major difference here is that $MCMC$ does only slightly worse than KM and QMC. Hence, for these settings, the $MCMC$ chains converge quicker than when the heterogeneity is low. This group is again followed by the group of BL, JC and JI, which are very similar when the number of alternatives is small, $J = 3$, and which can be ordered as JI, JC and BL when the number of alternatives is large, $J = 12$. BM, as before, does very well when there are relatively many choice sets per decision maker and is very inaccurate when there are not. Looking at both figures it can clearly be seen that BL is by far the fastest of the algorithms, followed by JC and JI. The former has the edge when there are three alternatives while the latter has the edge when there are twelve alternatives. When there are only three alternatives, we can see that QMC is faster than BM and KM. This distinction however disappears when there are twelve alternatives.

Correlated Taste Parameters

All the specifications of the population variance of the heterogeneous tastes in the previous section were diagonal matrices. This is a highly idealized set-up which represents a small subset of potential heterogeneity distributions. Furthermore, it is doubtful that in reality tastes for different attributes are truly independent. Therefore, we performed a second set of simulations to assess the performance of the various variational approaches in a setting with non-zero covariances between the taste parameters. In order to obtain plausible settings for the population tastes we simulated data based on results reported by Train and Weeks (2005). The original data were obtained by Train and Hudson (2000) and contained stated-preference choices made by 500 households among alternative-fueled vehicles. Each of the respondents considered 15 choice sets with several attributes describing the alternatives. Train and Weeks (2005) estimated several discrete choice models with these data and we will use estimated parameters from a model with $K = 7$ attributes (Train and Weeks 2005, pg.13-14): price, willingness to pay for (WTP) operating cost, WTP range, WTP electric car, WTP hybrid car, WTP High performance, WTP Medium/High performance. Note that several of these coefficients reflect tastes for categorical attributes in the originial data. We will however use these parameters with continuous attribute values. This results in a population mean taste vector $\boldsymbol{\zeta} = (-1.4934, -0.0489, 0.7636, -2.5353, 0.8738, 0.3584, 0.6047)$ and the following covariance matrix $\boldsymbol{\Omega}$ which represents the population taste heterogeneity for these attributes

$$\boldsymbol{\Omega} = \begin{pmatrix} 3.2844 & 0.0532 & 0.6262 & -2.0619 & 1.0965 & 0.4893 & 0.7940 \\ 0.0532 & 0.0028 & 0.0101 & -0.0333 & 0.0179 & 0.0084 & 0.0133 \\ 0.6262 & 0.0101 & 0.1812 & -0.3915 & 0.2091 & 0.0932 & 0.1494 \\ -2.0619 & -0.0333 & -0.3915 & 1.9827 & -0.6851 & -0.3038 & -0.5110 \\ 1.0965 & 0.0179 & 0.2091 & -0.6851 & 2.1182 & 0.1584 & 0.2688 \\ 0.4893 & 0.0084 & 0.0932 & -0.3038 & 0.1584 & 0.5720 & 0.1174 \\ 0.7940 & 0.0133 & 0.1494 & -0.5110 & 0.2688 & 0.1174 & 3.8189 \end{pmatrix}.$$

Based on these taste parameters we simulated data specifying different levels of the number of agents $H = 250$ or $H = 1000$. The number of choice sets per agent were taken as $T_h = 1, 5, 15$ or 25 and the number of alternatives per choice set as $J = 3$ or $J = 12$. Each of these simulation scenarios was once again replicated 10 times and all the attributes were once more independent identically normally distributed variables, $\mathcal{N}(0, 0.5^2)$. The performance of the various variational algorithms was again assessed by the median total variation error of the predictive choice distributions for 25 new, random choice sets

as described in the previous section.

-Insert Figure 3 about here-

The results of these simulations can be found in figure 3 which shows the average total variation error and average completion time (in minutes) over the 10 replications for all experimental settings. In the cases when the variational algorithm was run with a restricted version of the decision makers' covariance matrices and with an unrestricted version, i.e. dense covariance matrices, we again only report the restricted results. Note also that again the figure does not include timing information for the $MCMC$ chains. Looking at figure 3 we can see a similar overall pattern as before. The QMC approach seems to be most accurate overall. The BM approach is very similar to QMC when the number of choice sets is large enough but is extremely inaccurate when this is not the case. QCM and BM (sometimes) are closely followed by KM. The worst accuracy is seen to come from the $MCMC$ approach which indicates it clearly did not have enough time for the chain to converge. The BL, JC and JI methods fall somewhere in between. We can also observe that the accuracy increases when the number of alternatives is lower and when the number of choice sets is larger. We observe again that BL is by far the fastest method followed by JC when there are three alternatives and followed by JI when there are twelve alternatives. We can also see that QMC is slightly faster than KM which in turn is slightly faster than BM when there are three alternatives. When there are twelve alternatives however, we see that KM is slightly faster than QMC and BM.

Diagonal Restriction

In the previous sections the presented results for the BM, JI, KM and QMC methods were based on the diagonally restricted versions.

-Insert Figure 4 about here-

In figure 4 we show the performances of the restricted versions against the unrestricted versions. With respect to time we can see that in most cases the restricted versions converged faster than their unrestricted counterparts, which is to be expected. The case of BM is an outlier here in that in quite a few cases the restricted version converged faster. These, however, are the cases where the algorithm did not perform accurate at all, i.e. diverged. With respect to the accuracy we can see that the restricted and the unrestricted versions are very similar. Considering the significant speed-up of the algorithms, we can see that using restricted decision makers' covariance matrices works very well.

Conclusions and Future Work

We have compared several approaches to approximate the posterior distribution in the framework of mixed logit models. We found that several bounds were too loose to adequately capture the posterior variation which resulted in relatively poor performance. These bounds are the quadratic bounds, BL, BO, JC and the non quadratic bound JI. The proposed bound of Knowles and Minka (2011) did perform well in the context of mixed logit models on the other hand. Furthermore, it appears that the approximations, opposed to bounds, considered here, outperform the bounds. The QMC approach especially performed well in all experimental settings. The BM approximation performed equally well whenever the data contained enough information. In datasets with a small number of agents and/or a small number of choice sets, however, this approximation's bias becomes too large and its performance decreases considerably. All in all, it appears that using an appropriate approximation or bound, the variational approach is viable in the context of mixed logit models. This may indicate an avenue of potential further research, i.e. the development of new, non quadratic bounds which may simplify the algorithms or speed up the optimization. Another potential avenue for further research may be to look for an optimal combination of all useful bounds and/or approximations, i.e. development of some hybrid algorithm. We also did not consider the question of optimal visiting schedules for the various parameter updates. It is very likely that the coordinate ascent algorithm's convergence can be improved by optimizing such a schedule. Finally, as the variational approach seems to work adequately, more complicated models could be considered. For instance, the requirement that the mixing distribution is normal is a suspect assumption which is likely not very realistic. One could improve the flexibility of the model by using a finite mixture of mixed models. Traditional $MCMC$ becomes very burdensome for these types of models due to the multimodality in the posterior and the label switching. A variational approach on the other hand only focuses on one mode and hence there is no need to explore all the equivalent modes due to the label switching. This will speed up the inference considerably at the potential small cost of some approximation bias.

References

Christopher M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer New York, 2006.

David M. Blei and John D. Lafferty. A correlated topic model of Science. *Annals of Applied Statistics*, 1(1):17–35, June 2007.

Dankmar Böhning. Multinomial Logistic Regression Algorithm. *Annals of the Institute of Statistical Mathematics*, 44(1):197–200, 1992.

Dankmar Böhning and Bruce G. Lindsay. Monotonicity of Quadratic-Approximation Algorithms. *Annals of the Institute of Statistical Mathematics*, 40(4):641–663, 1988.

Guillaume Bouchard. Efficient Bounds for the Softmax Function and Applications to Approximate Inference in Hybrid models. In *NIPS 207 Workshop on Approximate Inference in Hybrid Models*, pages 1–9, 2007.

Michael Braun and Jon McAuliffe. Variational Inference for Large-Scale Models of Discrete Choice. *Journal of the American Statistical Association*, 105(489):324–335, March 2010.

Guido Consonni and Jean-Michel Marin. Mean-field variational approximate Bayesian inference for latent variable models. *Computational Statistics & Data Analysis*, 52(2): 790–798, October 2007.

Mark Girolami and Simon Rogers. Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors. *Neural Computation*, 18(8):1790–1817, 2006.

Maya Gupta and Santosh Srivastava. Parametric Bayesian Estimation of Differential Entropy and Relative Entropy. *Entropy*, 12(4):818–843, April 2010.

Peter Hall, John T. Ormerod, and Matt Wand. Theory of Gaussian variational approximation for a Poisson mixed model. *Statistica Sinica*, 21:369–389, 2011.

Fred J. Hickernell, Hee S. Hong, Pierre L'Écuyer, and Christiane Lemieux. Extensible Lattice Sequences for Quasi-Monte Carlo Quadrature. *SIAM Journal on Scientific Computing*, 22(3):1117–1138, January 2000.

Tommi S. Jaakkola and Michael I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37, 2000.

Tony Jebara and Anna Choromanska. Majorization for CRFs and Latent Likelihoods. In *Neural Information Processing Systems (NIPS)*, 2012.

Mohammad E. Khan, Benjamin M. Marlin, Guillaume Bouchard, and Kevin P. Murphy. Variational bounds for mixed-data factor analysis. In *Advances in Neural Information Proceeding Systems*, pages 1–9, 2010.

David A. Knowles and Thomas P. Minka. Non-conjugate variational message passing for multinomial and binary regression. In *Advances in Neural Information Processing Systems*, pages 1701–1709. 2011.

Neil D. Lawrence, Marta Milo, Mahesan Niranjan, Penny Rashbass, and Stephan Soullier. Reducing the variability in cDNA microarray image processing by Bayesian inference. *Bioinformatics (Oxford, England)*, 20(4):518–26, March 2004.

David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

Daniel McFadden. Conditional logit analysis of qualitative choice behaviour. In P. Zarembka, editor, *Frontiers of Econometrics*. Academic Press, New York, 1974.

John T. Ormerod and Matt Wand. Explaining Variational Approximations. *The American Statistician*, 64(2):140–153, May 2010.

John T. Ormerod and Matt Wand. Gaussian Variational Approximate Inference for Generalized Linear Mixed Models. *Journal of Computational and Graphical Statistics*, 21 (1):2–17, January 2012.

Giorgio Parisi. *Statistical Field Theory*. Addison-Wesley, 1988.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`.

Peter Rossi. *bayesm: Bayesian Inference for Marketing/Micro-econometrics*, 2012. URL `http://CRAN.R-project.org/package=bayesm`. R package version 2.2-5.

Peter. Rossi, Greg. Allenby, and Robert McCulloch. *Bayesian Statistics and Marketing.*, volume 169. John Wiley & Sons Ltd, October 2005.

Havard vard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 71:319–392, 2009.

Mike Titterington. The EM algorithm, variational approximations and expectation propagation for mixtures. In K. L. Mengersen, C. P. Robert, and D. M. Titterington, editors, *Mixtures: Estimation and Applications*, pages 1–30. John Wiley & Sons, Ltd, Chichester, UK, 2011.

Kenneth Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, second edition, 2009.

Kenneth Train and Kathleen Hudson. The Impact of Information on Vehicle Choices and the Demand for Electric Vehicles in California. Technical report, Toyota and General Motors, 2000.

Kenneth Train and Melvyn Weeks. Discrete Choice Models in Preference Space and Willingness-to-Pay Space. In Anna Scarpa, Riccardo and Alberini, editor, *Applications of Simulation Methods in Environmental and Resource Economics*, pages 1–16. Springer Netherlands, 2005.

Bo Wang and Mike Titterington. Inadequacy of interval estimates corresponding to variational Bayesian approximations. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, Jan 6-8, 2005, Savannah Hotel, Barbados*, pages 373–380. Society for Artificial Intelligence and Statistics, 2005.

Bo Wang and Mike Titterington. Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3):625–650, 2006.

John Winn and Christopher M. Bishop. Variational Message Passing. *Journal of Machine Learning Research*, 6:661–694, 2005.

Jie Yu, Peter Goos, and Martina Vandebroek. Comparing different sampling schemes for approximating the integrals involved in the efficient design of stated choice experiments. *Transportation Research Part B: Methodological*, 44(10):1268–1289, December 2010.

Elaine L. Zanutto and Eric T. Bradlow. Data pruning in consumer choice models. *Quantitative Marketing and Economics*, 4(3):267–287, August 2006.

Notes

[1]Note that we use the same notation in this paper as Braun and McAuliffe (2010) for consistency.

[2]The differential entropy of a density $f(x)$ is defined as $H[f(x)] = -\int f(x) \log f(x)\, dx$.

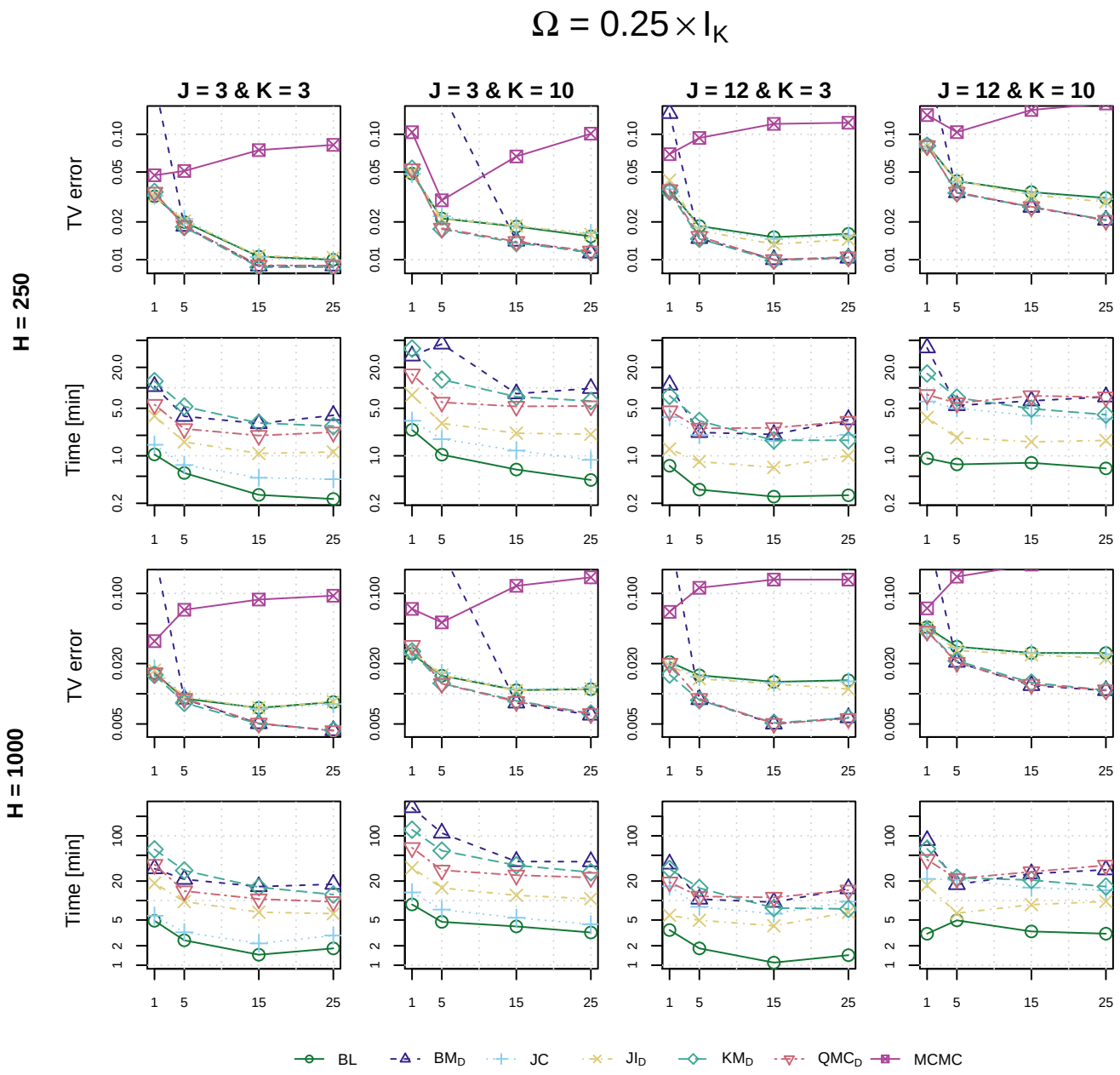

Figure 1: Total variation error and times till convergence (in minutes) for the low heterogeneity setting. Each observation is the average over 10 replications. The x-axes represent the number of choice sets per agent. Note that the y-axes are on a logarithmic scale. Some results are clipped from above to improve the readability.

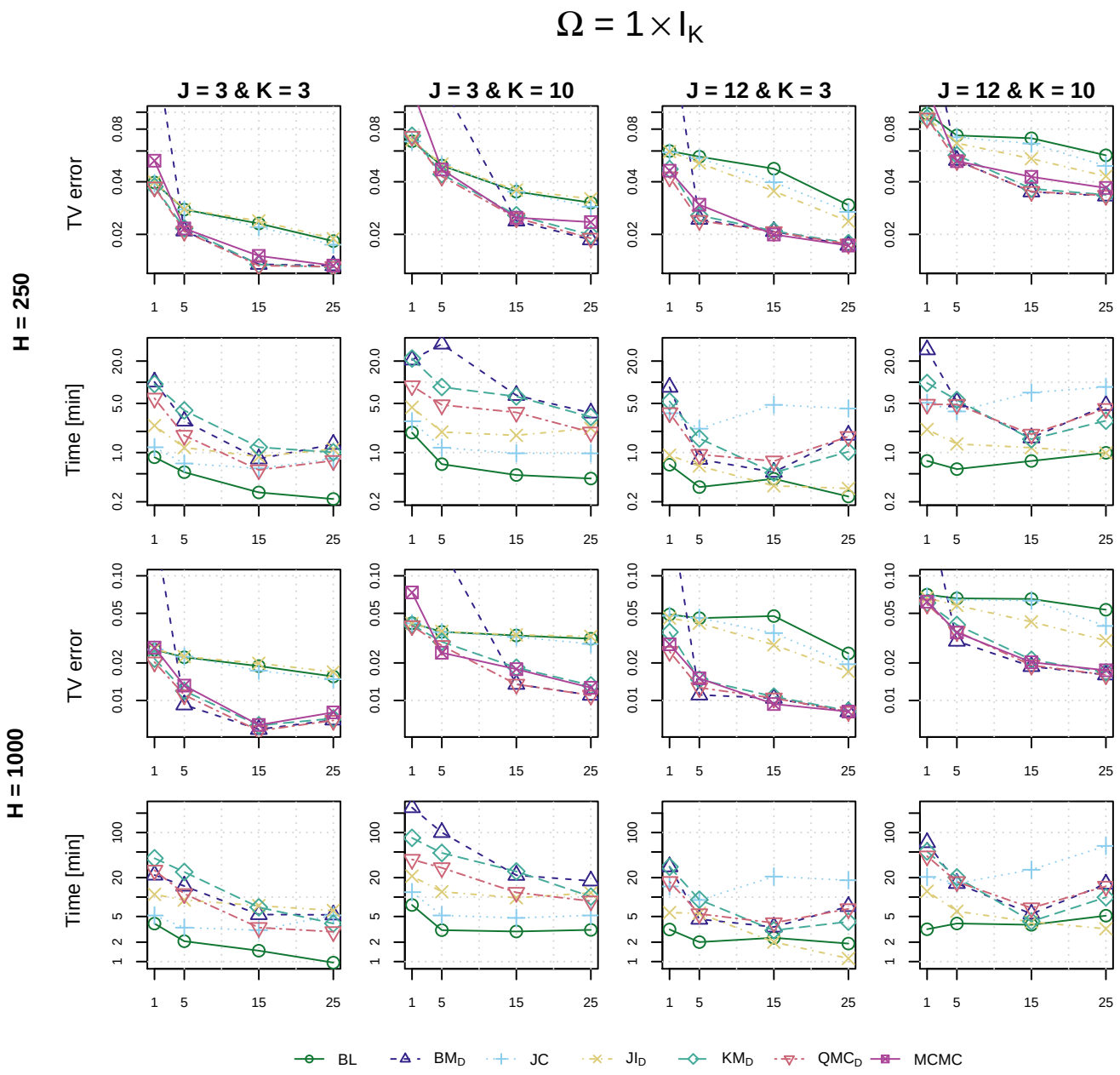

Figure 2: Total variation error and times till convergence (in minutes) for the high heterogeneity setting. Each observation is the average over 10 replications. The x-axes represent the number of choice sets per agent. Note that the y-axes are on a logarithmic scale. Some results are clipped from above to improve the readability.

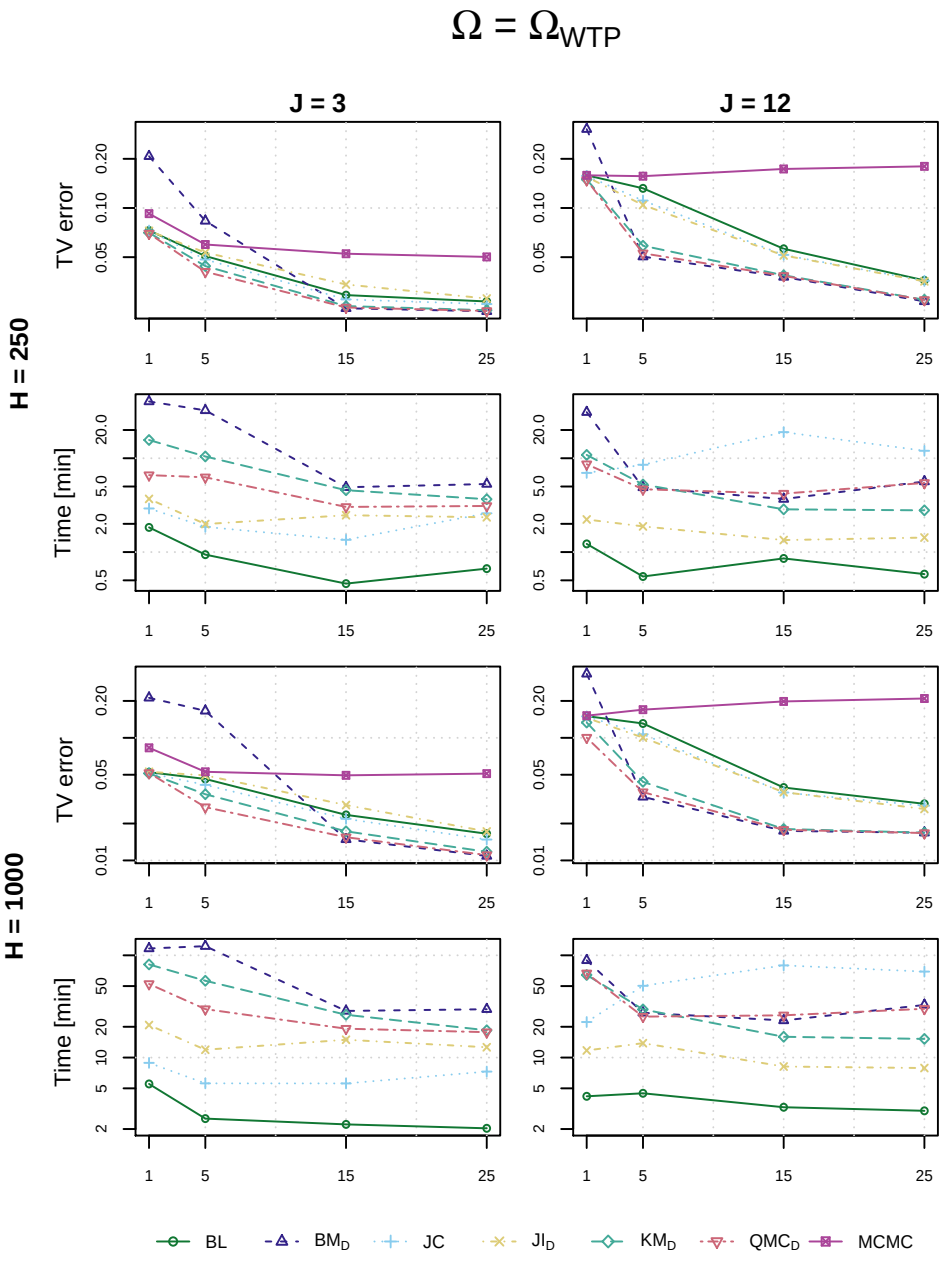

Figure 3: Total variation error and times till convergence (in minutes). Each observation is the average over 10 replications. The x-axes represent the number of choice sets per agent. Note that the y-axes are on a logarithmic scale.

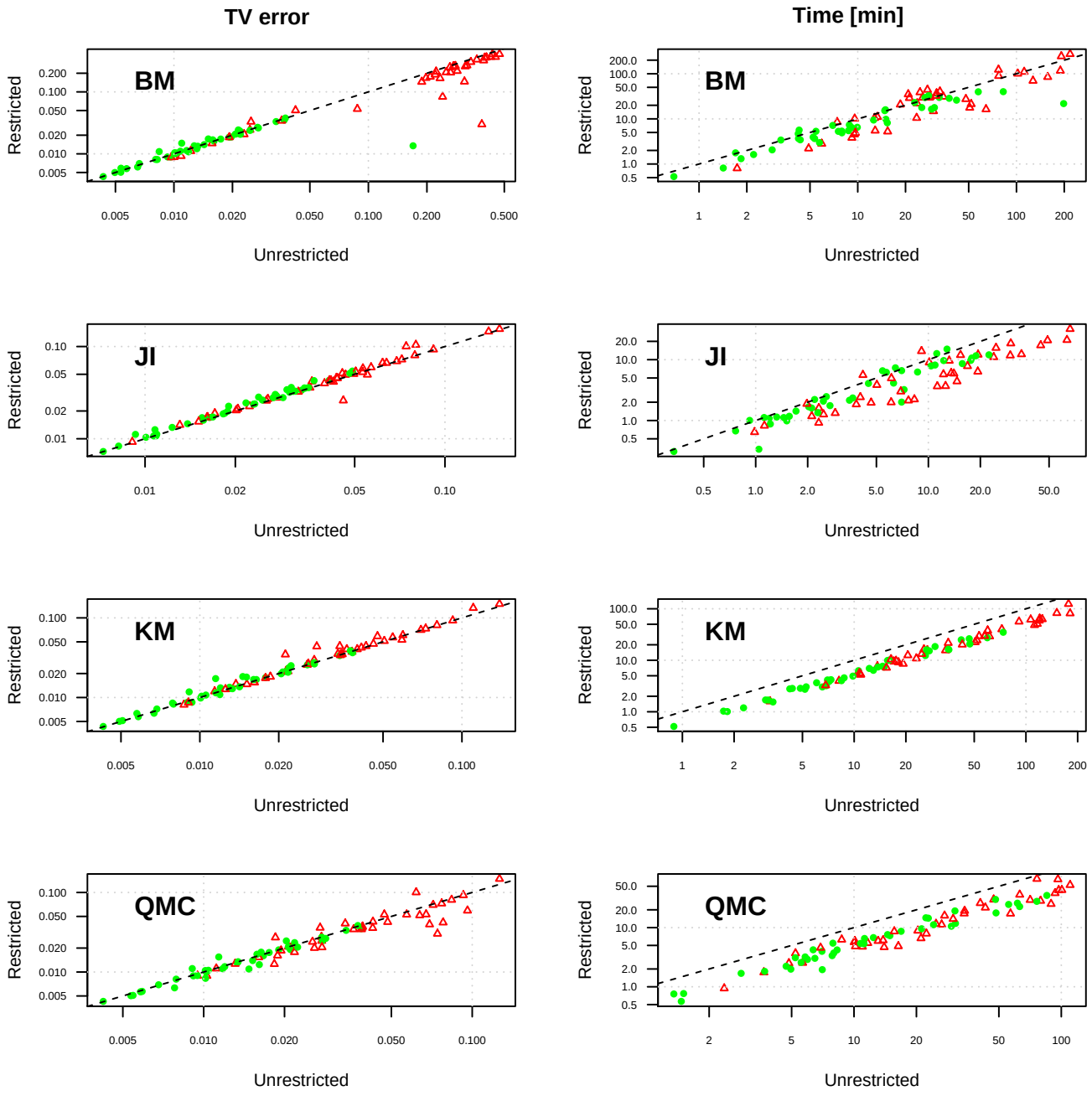

Figure 4: Total variation error and times till convergence (in minutes) for all settings. Each observation is the average over 10 replications. The x-axes represent the performance of the algorithms with unrestricted decision makers' covariance matrices. The y-axes represent the performance of the algorithms with diagonally restricted decision makers' covariance matrices. The green dots represent the cases where the number of choice sets, T_h, was 15 or 25 whereas the cases where the number of choice sets was 1 or 5 are represented by red triangles. The dashed lines are the 45° lines which represent identical performances

A Variational Bayes for the Mixed Logit Model

In this section the development of equation (3), i.e. the expected joint log probability of the data and the priors under the factorized posterior approximation plus the entropy of the variational posterior distribution are shown in more detail. In order to avoid confusion about different parameterizations, we define the following form for the normal and inverse-Wishart densities:

$$p\left(\boldsymbol{\zeta};\boldsymbol{\mu}_{\boldsymbol{\zeta}},\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}\right) \propto \left|\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}\left(\boldsymbol{\zeta}-\boldsymbol{\mu}_{\boldsymbol{\zeta}}\right)^{T}\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{-1}\left(\boldsymbol{\zeta}-\boldsymbol{\mu}_{\boldsymbol{\zeta}}\right)}$$
$$p\left(\boldsymbol{\Omega};\boldsymbol{\Upsilon}^{-1},\omega\right) \propto \left|\boldsymbol{\Omega}\right|^{-\frac{\omega+K+1}{2}} e^{-\frac{1}{2}tr\left(\boldsymbol{\Upsilon}^{-1}\boldsymbol{\Omega}^{-1}\right)}.$$

As the densities of $q\left(\boldsymbol{\beta}_h\right), h=1,\ldots,H$ are equivalent to the density of $\boldsymbol{\zeta}$, only the latter details are shown. The log joint probability of the mixed logit model is, up to a constant (Hyperparameters from priors are set before estimation and are thus constants. Any term that only contains constants required for normalization of the normal and inverse-Wishart distributions is dropped here.):

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{t=1}^{T_h}\left\{\boldsymbol{y}_{ht}^{T}\boldsymbol{X}_{ht}\boldsymbol{\beta}_h - \log\left[\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^{T}\boldsymbol{\beta}_h}\right]\right\} \\
&\qquad + \sum_{h=1}^{H}\left\{-\frac{1}{2}\log\left|\boldsymbol{\Omega}\right| - \frac{1}{2}\boldsymbol{\beta}_h^{T}\boldsymbol{\Omega}^{-1}\boldsymbol{\beta}_h - \frac{1}{2}\boldsymbol{\zeta}^{T}\boldsymbol{\Omega}^{-1}\boldsymbol{\zeta} + \boldsymbol{\beta}_h^{T}\boldsymbol{\Omega}^{-1}\boldsymbol{\zeta}\right\} \\
&\qquad - \frac{1}{2}\boldsymbol{\zeta}^{T}\boldsymbol{\Omega}_0^{-1}\boldsymbol{\zeta} + \boldsymbol{\zeta}^{T}\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 - \frac{\nu+K+1}{2}\log\left|\boldsymbol{\Omega}\right| - \frac{1}{2}tr\left(\boldsymbol{S}^{-1}\boldsymbol{\Omega}^{-1}\right). \qquad (7)
\end{aligned}$$

Because the assumed posterior distribution is factorized, we only require moments of the normal and inverse-Wishart distribution to evaluate the expected value of equation (7), which are fairly easy to derive. In what follows all the expectations are with respect to the approximate posterior densities of the variables over which the expectation is taken. For the normal expectations and entropy for the parameter $\boldsymbol{\zeta}$ we have:

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\zeta}\right] &= \boldsymbol{\mu}_{\boldsymbol{\zeta}} \\
\mathbb{E}\left[\boldsymbol{\zeta}^{T}\boldsymbol{\Omega}_0\boldsymbol{\zeta}\right] &= \mathbb{E}\left[tr\left(\boldsymbol{\Omega}_0\boldsymbol{\zeta}\boldsymbol{\zeta}^{T}\right)\right] = tr\left[\boldsymbol{\Omega}_0\left(\boldsymbol{\Sigma}_{\boldsymbol{\zeta}} + \boldsymbol{\mu}_{\boldsymbol{\zeta}}\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{T}\right)\right] \\
H\left[\boldsymbol{\zeta}\right] &= \frac{K}{2}\log\left(2\pi e\right) + \frac{1}{2}\left|\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}\right| = \frac{1}{2}\left|\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}\right| + \text{Constant}.
\end{aligned}$$

The same expectations are required to evaluate the expectations for the $\boldsymbol{\beta}_{h=1:H}$ parameters. For the inverse-Wishart expectations for the parameter $\boldsymbol{\Omega}$ we have the following

expectations (Gupta and Srivastava 2010):

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\Omega}^{-1}\right] &= \omega\boldsymbol{\Upsilon} \\
\mathbb{E}\left[\log|\boldsymbol{\Omega}|\right] &= -\log|\boldsymbol{\Upsilon}| - \sum_{k=1}^{K}\psi\left(\frac{\omega+1-k}{2}\right) \\
H\left[\boldsymbol{\Omega}\right] &= \sum_{k=1}^{K}\log\Gamma\left(\frac{\omega+1-k}{2}\right) + \frac{\omega K}{2} - \frac{K+1}{2}\log|\boldsymbol{\Upsilon}| \\
&\quad - \frac{\omega+K+1}{2}\sum_{k=1}^{K}\psi\left(\frac{\omega+1-k}{2}\right) + \text{Constant}
\end{aligned}$$

where $\psi(.)$ represents the digamma function, $\psi(x) = \frac{d}{dx}\log\Gamma(x)$, and $\Gamma(x)$ represents the gamma function, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. When we plug these expectations into equation (7) we obtain the following expected joint log probability of the data and the priors, again up to a constant:

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{t=1}^{T_h}\left\{\boldsymbol{y}_{ht}^T\boldsymbol{X}_{ht}\boldsymbol{\mu}_h - \mathbb{E}_{q(\boldsymbol{\beta}_h)}\left[\log\left(\sum_{j=1}^{J}e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right)\right]\right\} \\
&+\sum_{h=1}^{H}\left\{\frac{1}{2}\log|\boldsymbol{\Upsilon}| + \frac{1}{2}\sum_{k=1}^{K}\psi\left(\frac{\omega+1-k}{2}\right) - \frac{1}{2}tr\left[\omega\boldsymbol{\Upsilon}\left(\boldsymbol{\Sigma}_h + \boldsymbol{\Sigma}_{\boldsymbol{\zeta}} + \boldsymbol{\mu}_h\boldsymbol{\mu}_h^T + \boldsymbol{\mu}_{\boldsymbol{\zeta}}\boldsymbol{\mu}_{\boldsymbol{\zeta}}^T\right)\right] + \omega\boldsymbol{\mu}_h^T\boldsymbol{\Upsilon}\boldsymbol{\mu}_{\boldsymbol{\zeta}}\right\} \\
&-\frac{1}{2}tr\left[\boldsymbol{\Omega}_0^{-1}\left(\boldsymbol{\Sigma}_{\boldsymbol{\zeta}} + \boldsymbol{\mu}_{\boldsymbol{\zeta}}\boldsymbol{\mu}_{\boldsymbol{\zeta}}^T\right)\right] + \boldsymbol{\mu}_{\boldsymbol{\zeta}}^T\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 + \frac{\nu+K+1}{2}\log|\boldsymbol{\Upsilon}| + \frac{\nu+K+1}{2}\sum_{k=1}^{K}\psi\left(\frac{\omega+1-k}{2}\right) \\
&-\frac{\omega}{2}tr\left[\boldsymbol{S}^{-1}\boldsymbol{\Upsilon}\right]. \qquad (8)
\end{aligned}$$

The entropy of the variational posterior is up to a constant

$$\begin{aligned}
&\sum_{h=1}^{H}\left\{\frac{1}{2}\log|\boldsymbol{\Sigma}_h|\right\} + \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}| + \sum_{k=1}^{K}\log\Gamma\left(\frac{\omega+1-k}{2}\right) + \frac{\omega K}{2} - \frac{K+1}{2}\log|\boldsymbol{\Upsilon}| \\
&\qquad - \frac{\omega+K+1}{2}\sum_{k=1}^{K}\psi\left(\frac{\omega+1-k}{2}\right). \qquad (9)
\end{aligned}$$

We can now calculate derivatives of the lower bound, i.e. (8) + (9), with respect to $\boldsymbol{\mu}_{\boldsymbol{\zeta}}, \boldsymbol{\Sigma}_{\boldsymbol{\zeta}}, \omega$ and $\boldsymbol{\Upsilon}$ and set them to $\mathbf{0}$.

$$\begin{aligned}
\nabla_{\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}} &= -\frac{1}{2} tr\left[H\omega\boldsymbol{\Upsilon} + \boldsymbol{\Omega}_0^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{-1}\right] \\
\nabla_{\boldsymbol{\mu}_{\boldsymbol{\zeta}}} &= -\left(H\omega\boldsymbol{\Upsilon} + \boldsymbol{\Upsilon}_0^{-1}\right)\boldsymbol{\mu}_{\boldsymbol{\zeta}} + \omega\boldsymbol{\Upsilon}\sum_{h=1}^{H}\boldsymbol{\mu}_h + \boldsymbol{\Upsilon}_0^{-1}\boldsymbol{\beta}_0 \\
\nabla_{\boldsymbol{\Upsilon}} &= \frac{\nu + H}{2}\boldsymbol{\Upsilon}^{-1} - \frac{\omega}{2}\left\{\boldsymbol{S}^{-1} + H\boldsymbol{\Sigma}_{\boldsymbol{\zeta}} + \sum_{h=1}^{H}\left[\boldsymbol{\Sigma}_h + \left(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\boldsymbol{\zeta}}\right)\left(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\boldsymbol{\zeta}}\right)^T\right]\right\} \\
\frac{\partial\left((8)+(9)\right)}{\partial\omega} &= \frac{K}{2} + \frac{H + \nu - \omega}{2}\sum_{k=1}^{K}\frac{\partial\psi\left(\frac{\omega+1-k}{2}\right)}{\partial\omega} \\
&\quad - \frac{1}{2} tr\left\{\boldsymbol{S}^{-1} + H\boldsymbol{\Sigma}_{\boldsymbol{\zeta}} + \sum_{h=1}^{H}\left[\boldsymbol{\Sigma}_h + \left(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\boldsymbol{\zeta}}\right)\left(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\boldsymbol{\zeta}}\right)^T\right]\right\}
\end{aligned}$$

Solving for the variational parameters we get the following closed form update equations:

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{\zeta}} &= \left(H\omega\boldsymbol{\Upsilon} + \boldsymbol{\Omega}_0^{-1}\right)^{-1} \\
\boldsymbol{\mu}_{\boldsymbol{\zeta}} &= \boldsymbol{\Sigma}_{\boldsymbol{\zeta}}\left(\omega\boldsymbol{\Upsilon}\sum_{h=1}^{H}\boldsymbol{\mu}_h + \boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0\right) \\
\omega &= \nu + H \\
\boldsymbol{\Upsilon} &= \left\{\boldsymbol{S}^{-1} + H\boldsymbol{\Sigma}_{\boldsymbol{\zeta}} + \sum_{h=1}^{H}\left[\boldsymbol{\Sigma}_h + \left(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\boldsymbol{\zeta}}\right)\left(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\boldsymbol{\zeta}}\right)^T\right]\right\}^{-1}.
\end{aligned}$$

The degrees of freedom parameter ω of the approximate posterior of $\boldsymbol{\zeta}$ is not data dependent and can be fixed at its optimal value from the start. The only unspecified parts of the estimation algorithm are the updates with respect to $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ for all $h = 1, \ldots, H$.

B Derivation of Bounds and Approximations

B.1 Taylor Series

Consider the second order Taylor series expansion of the function $f(\boldsymbol{\beta}_h) = \log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T \boldsymbol{\beta}_h}\right)$ around the current mean $\boldsymbol{\mu}_h$ which results in

$$f(\boldsymbol{\beta}_h) = \log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T \boldsymbol{\beta}_h}\right) \approx \log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T \boldsymbol{\mu}_h}\right) + (\boldsymbol{\beta}_h - \boldsymbol{\mu}_h)^T \nabla(\boldsymbol{\mu}_h) + \frac{1}{2}(\boldsymbol{\beta}_h - \boldsymbol{\mu}_h)^T H(\boldsymbol{\mu}_h)(\boldsymbol{\beta}_h - \boldsymbol{\mu}_h)$$

where $\nabla(\boldsymbol{\mu}_h) = \boldsymbol{X}_{ht}^T \boldsymbol{p}_{hj}$, $H(\boldsymbol{\mu}_h) = \boldsymbol{X}_{ht}^T \left[\operatorname{diag}(\boldsymbol{p}_{ht}) - \boldsymbol{p}_{ht}\boldsymbol{p}_{ht}^T\right] \boldsymbol{X}_{ht}$ and where $\boldsymbol{p}_{ht}$ is a J-dimensional vector with entries $\frac{e^{\boldsymbol{x}_{htj}^T \boldsymbol{\mu}_h}}{\sum_{j'=1}^{J} e^{\boldsymbol{x}_{htj'}^T \boldsymbol{\mu}_h}}, \forall j = 1, \ldots, J$ and $\operatorname{diag}(\boldsymbol{x})$ is an operator that constructs a diagonal matrix out of the elements in $\boldsymbol{x}$. Taking expectations with respect to $\boldsymbol{\beta}_h$ this leads to

$$\mathbb{E}_{q(\boldsymbol{\beta}_h)}\left[\log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T \boldsymbol{\beta}_h}\right)\right] \approx \log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T \boldsymbol{\mu}_h}\right) + \frac{1}{2} tr\left[\boldsymbol{\Sigma}_h H(\boldsymbol{\mu}_h)\right].$$

If we plug this approximation into equation (7) and collect all terms which only depend on $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ from equations (7) and (9) we obtain the following maximization problem:

$$\arg\max_{\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h} \sum_{t=1}^{T_h} \boldsymbol{y}_{ht}^T \boldsymbol{X}_{ht} \boldsymbol{\mu}_h - \log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T \boldsymbol{\mu}_h}\right) - \frac{1}{2} tr\left[\boldsymbol{\Sigma}_h H(\boldsymbol{\mu}_h)\right] - \frac{\omega}{2} tr\left[\boldsymbol{\Upsilon}\boldsymbol{\Sigma}_h\right] - \frac{\omega}{2}\boldsymbol{\mu}_h^T \boldsymbol{\Upsilon}\boldsymbol{\mu}_h + \omega \boldsymbol{\mu}_h^T \boldsymbol{\Upsilon}\boldsymbol{\mu}_{\boldsymbol{\zeta}} + \frac{1}{2}\log|\boldsymbol{\Sigma}_h|.$$

This approach is the BM and BM_D method where the latter restricts $\boldsymbol{\Sigma}_h$ to a diagonal matrix. Obviously this approach will only work well if the approximation is close enough.

B.2 Quasi Monte Carlo

The maximization function for the QMC approach is

$$\arg\max_{\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h} \sum_{t=1}^{T_h} \boldsymbol{y}_{ht}^T \boldsymbol{X}_{ht} \boldsymbol{\mu}_h - \frac{1}{R}\sum_{r=1}^{R} \log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T \boldsymbol{\beta}_h^{(r)}}\right) - \frac{\omega}{2} tr\left[\boldsymbol{\Upsilon}\boldsymbol{\Sigma}_h\right] - \frac{\omega}{2}\boldsymbol{\mu}_h^T \boldsymbol{\Upsilon}\boldsymbol{\mu}_h + \omega \boldsymbol{\mu}_h^T \boldsymbol{\Upsilon}\boldsymbol{\mu}_{\boldsymbol{\zeta}} + \frac{1}{2}\log|\boldsymbol{\Sigma}_h|.$$

This approach was also considered with unrestricted and a diagonally restricted variance matrices $\boldsymbol{\Sigma}_h$.

B.3 Jensen's Inequality

As $\log(.)$ is a concave function one can apply Jensen's inequality to obtain:

$$\mathbb{E}_{q(\boldsymbol{\beta}_h)}\left[\log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right)\right] \leq \log\left(\sum_{j=1}^{J}\mathbb{E}_{q(\boldsymbol{\beta}_h)}\left[e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right]\right) = \log\left(\sum_{j=1}^{K} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\mu}_h+\frac{1}{2}\boldsymbol{x}_{htj}^T\boldsymbol{\Sigma}_h\boldsymbol{x}_{htj}}\right).$$

The latter expectation is simply the moment generating function of a multivariate normal distribution. If we plug this bound into equation (7) and collect all terms which only depend on $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ from equations (7) and (9) we obtain the following maximization problem:

$$\arg\max_{\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h}\sum_{t=1}^{T_h}\boldsymbol{y}_{ht}^T\boldsymbol{X}_{ht}\boldsymbol{\mu}_h - \log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\mu}_h+\frac{1}{2}\boldsymbol{x}_{htj}^T\boldsymbol{\Sigma}_h\boldsymbol{x}_{htj}}\right) - \frac{\omega}{2}tr\left[\boldsymbol{\Upsilon}\boldsymbol{\Sigma}_h\right] - \frac{\omega}{2}\boldsymbol{\mu}_h^T\boldsymbol{\Upsilon}\boldsymbol{\mu}_h + \omega\boldsymbol{\mu}_h^T\boldsymbol{\Upsilon}\boldsymbol{\mu}_{\boldsymbol{\zeta}} + \frac{1}{2}\log\left|\boldsymbol{\Sigma}_h\right|.$$

This approach is the JI and JI_D method where the latter restricts $\boldsymbol{\Sigma}_h$ to a diagonal matrix. To obtain the KM and KM_D methods we need to introduce additional variational parameters. We start from the identity $\log\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h} = \sum_{j=1}^{J} a_{htj}\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h + \log\sum_{j=1}^{J} e^{\left(\boldsymbol{x}_{htj}-\sum_{j=1}^{J} a_{htj}\boldsymbol{x}_{htj}\right)^T\boldsymbol{\beta}_h}$. Taking expectations and once again applying Jensen's inequality then leads to

$$\begin{aligned}\mathbb{E}_{q(\boldsymbol{\beta}_h)}\left[\log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right)\right] \leq{}& \sum_{j=1}^{J} a_{htj}\boldsymbol{x}_{htj}^T\boldsymbol{\mu}_h \\ &+ \log\left(\sum_{j=1}^{J} e^{\left(\boldsymbol{x}_{htj}-\sum_{j=1}^{J} a_{htj}\boldsymbol{x}_{htj}\right)^T\boldsymbol{\mu}_h+\frac{1}{2}\left(\boldsymbol{x}_{htj}-\sum_{j=1}^{J} a_{htj}\boldsymbol{x}_{htj}\right)^T\boldsymbol{\Sigma}_h\left(\boldsymbol{x}_{htj}-\sum_{j=1}^{J} a_{htj}\boldsymbol{x}_{htj}\right)}\right).\end{aligned}$$

Plugging this into equation (7) we obtain a similar maximization problem as the previous one. However, we have introduced extra variational parameters $\boldsymbol{a} = (a_{1:H,1:T_h,1:J})$ which also need to be updated. Taking derivatives and equating them to 0 results in the following fixed point update equations

$$a_{htj} = \frac{e^{\boldsymbol{x}_{htj}^T\boldsymbol{\mu}_h+\frac{1}{2}\left(\boldsymbol{x}_{htj}-2\sum_{j=1}^{J} a_{htj}\boldsymbol{x}_{htj}\right)^T\boldsymbol{\Sigma}_h\boldsymbol{x}_{htj}}}{\sum_{j'=1}^{J} e^{\boldsymbol{x}_{htj'}^T\boldsymbol{\mu}_h+\frac{1}{2}\left(\boldsymbol{x}_{htj'}-2\sum_{j''=1}^{J} a_{htj''}\boldsymbol{x}_{htj''}\right)^T\boldsymbol{\Sigma}_h\boldsymbol{x}_{htj'}}} \qquad \forall h,t,j.$$

This approach is the KM and KM_D method where the latter restricts $\boldsymbol{\Sigma}_h$ to a diagonal matrix.

B.4 Bï¿½hning-Lindsay

Define $\boldsymbol{A} = \frac{1}{2}\left(\boldsymbol{I}_J - \mathbf{1}_J\mathbf{1}_J^T/J\right)$ where $\boldsymbol{I}_J$ is the J-dimensional identity matrix and $\mathbf{1}_J$ is a J-dimensional vector of ones. Böhning and Lindsay (1988) and Böhning (1992) show that $\boldsymbol{A} \geq \boldsymbol{H}$ with respect to the Loewner ordering ($\boldsymbol{A} \geq \boldsymbol{H}$ with respect to the Loewner ordering if $\boldsymbol{A} - \boldsymbol{H}$ is positive semi-definite.). If we take a second order Taylor series expansion of the function $f\left(\boldsymbol{X}_{ht}\boldsymbol{\beta}_h\right) = \log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right)$ around some parameter vector $\boldsymbol{\Psi}_{ht}$ we know that for some specific vector $\boldsymbol{\Psi}_{ht}^*$ we get the following equality

$$f\left(\boldsymbol{X}_{ht}\boldsymbol{\beta}_h\right) = \log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right) = \log\left(\sum_{j=1}^{J} e^{\Psi_{htj}}\right) + \left(\boldsymbol{X}_{ht}\boldsymbol{\beta}_h - \boldsymbol{\Psi}_{ht}\right)^T \nabla\left(\boldsymbol{\Psi}_{ht}\right) + \frac{1}{2}\left(\boldsymbol{X}_{ht}\boldsymbol{\beta}_h - \boldsymbol{\Psi}_{ht}\right)^T H\left(\boldsymbol{\Psi}_{ht}^*\right)\left(\boldsymbol{X}_{ht}\boldsymbol{\beta}_h - \boldsymbol{\Psi}_{ht}\right)$$

where $\nabla\left(\boldsymbol{\Psi}_{ht}\right)$ and $H\left(\boldsymbol{\Psi}_{ht}^*\right)$ are the gradient of $f\left(\boldsymbol{X}_{ht}\boldsymbol{\beta}_h\right)$ evaluated at $\boldsymbol{\Psi}_{ht}$ and $\boldsymbol{\Psi}_{ht}^*$ respectively. Replacing $H\left(\boldsymbol{\Psi}_{ht}^*\right)$ with $\boldsymbol{A}$ and taking expectations over $\boldsymbol{\beta}_h$ we can obtain the following bound:

$$\mathbb{E}_{q(\boldsymbol{\beta}_h)}\left[\log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right)\right] \leq \log\left(\sum_{j=1}^{J} e^{\Psi_{htj}}\right) + \left(\boldsymbol{X}_{ht}\boldsymbol{\mu}_h - \boldsymbol{\Psi}_{ht}\right)^T \nabla\left(\boldsymbol{\Psi}_{ht}\right) + \frac{1}{2}\left[\boldsymbol{X}_{ht}^T\boldsymbol{A}\boldsymbol{X}_{ht}\boldsymbol{\Sigma}_h\right] + \frac{1}{2}\left(\boldsymbol{X}_{ht}\boldsymbol{\mu}_h - \boldsymbol{\Psi}_{ht}\right)^T \boldsymbol{A}\left(\boldsymbol{X}_{ht}\boldsymbol{\mu}_h - \boldsymbol{\Psi}_{ht}\right).$$

From this bound it is possible to generate analytic update equations for the subject specific parameters by plugging it into equation (7) and equating derivatives with respect to $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ to $\mathbf{0}$ which results in:

$$\boldsymbol{\Sigma}_h = \left(\omega\boldsymbol{\Upsilon} + \sum_{t=1}^{T_h} \boldsymbol{X}_{ht}^T\boldsymbol{A}\boldsymbol{X}_{ht}\right)^{-1}, \quad h = 1, \ldots, H$$

$$\boldsymbol{\mu}_h = \boldsymbol{\Sigma}_h\left\{\omega\boldsymbol{\Upsilon}\boldsymbol{\mu}_{\boldsymbol{\zeta}} + \sum_{t=1}^{T_h} \boldsymbol{X}_{ht}^T\left[\boldsymbol{y}_{ht} - \nabla\left(\boldsymbol{\Psi}_{ht}\right) + \boldsymbol{A}\boldsymbol{\Psi}_{ht}\right]\right\}, \quad h = 1, \ldots, H.$$

Using derivatives again, it can be seen that the update for the extra variational parameters $\boldsymbol{\Psi}_{ht}$, $\forall h, t$, turns out to be $\boldsymbol{\Psi}_{ht} = \boldsymbol{X}_{ht}\boldsymbol{\mu}_h$.

B.5 Bouchard

Bouchard (2007) observed that $\sum_{j=1}^{J} e^{x_j} \leq \prod_{j=1}^{J} (1 + e^{x_j})$. Replacing x_j by $\boldsymbol{x}_{htj}^T \boldsymbol{\beta}_h - \alpha_{ht}$ and taking logarithms we arrive at $\log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T \boldsymbol{\beta}_h}\right) \leq \alpha_{ht} + \sum_{j=1}^{J} \log\left(1 + e^{\boldsymbol{x}_{htj}^T \boldsymbol{\beta}_h - \alpha_{ht}}\right)$. Jaakkola and Jordan (2000) derived the well known tangential bound $\log(1 + e^x) \leq \frac{x-t}{2} + \frac{1}{4t}\tanh\left(\frac{t}{2}\right)(x^2 - t^2) + \log(1 + e^t)$. Combining these two results and taking expectations with respect to $\boldsymbol{\beta}_h$ we obtain the following quadratic lower bound:

$$\mathbb{E}_{q(\boldsymbol{\beta}_h)}\left[\log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T \boldsymbol{\beta}_h}\right)\right] \leq \alpha_{ht} + \sum_{j=1}^{J} \frac{\boldsymbol{x}_{htj}^T \boldsymbol{\mu}_h - \alpha_{ht} - t_{htj}}{2} + \lambda(t_{htj})\left[\left(\boldsymbol{x}_{htj}^T \boldsymbol{\mu}_h - \alpha_{ht}\right)^2 - t_{htj}^2 + \boldsymbol{x}_{htj}^T \boldsymbol{\Sigma}_h \boldsymbol{x}_{htj}\right] + \log\left(1 + e^{t_{htj}}\right)$$

where $\lambda(t) = \frac{1}{4t}\tanh\left(\frac{t}{2}\right)$. From this bound it is possible to generate analytic update equations for the subject specific parameters by plugging it into equation (7) and equating derivatives with respect to $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ to $\mathbf{0}$ which results in:

$$\boldsymbol{\Sigma}_h = \left(\omega \boldsymbol{\Upsilon} + 2\sum_{t=1}^{T_h}\sum_{j=1}^{J} \lambda(t_{htj})\, \boldsymbol{x}_{htj} \boldsymbol{x}_{htj}^T\right)^{-1}, \quad \forall h = 1, \ldots, H$$

$$\boldsymbol{\mu}_h = \boldsymbol{\Sigma}_h\left[\omega \boldsymbol{\Upsilon} \boldsymbol{\mu}_{\boldsymbol{\zeta}} + \sum_{t=1}^{T_h}\sum_{j=1}^{J}\left(\boldsymbol{y}_{htj}^T - \frac{1}{2} + 2\alpha_{ht}\lambda(t_{htj})\right)\boldsymbol{x}_{htj}\right], \quad \forall h = 1, \ldots, H.$$

The extra variational parameters can be updated by fixed point equations which are

$$\alpha_{ht} = \frac{J/2 - 1 + 2\sum_{j=1}^{J} \lambda(t_{htj})\, \boldsymbol{x}_{htj}^T \boldsymbol{\mu}_h}{2\sum_{j=1}^{J} \lambda(t_{htj})} \qquad \forall h, t$$

$$t_{htj} = \sqrt{\left(\boldsymbol{x}_{htj}^T \boldsymbol{\mu}_h - \alpha_{ht}\right)^2 + \boldsymbol{x}_{htj}^T \boldsymbol{\Sigma}_h \boldsymbol{x}_{htj}} \qquad \forall h, t, j.$$

B.6 Jebara-Choromanska

Jebara and Choromanska (2012) developed an algorithm to find a quadratic bound $\log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T \boldsymbol{\beta}_h}\right) \leq \log z_{ht} + \frac{1}{2}\left(\boldsymbol{\beta}_h - \tilde{\boldsymbol{\beta}}_h\right)^T \boldsymbol{S}_{ht}\left(\boldsymbol{\beta}_h - \tilde{\boldsymbol{\beta}}_h\right) + \left(\boldsymbol{\beta}_h - \tilde{\boldsymbol{\beta}}_h\right)^T \boldsymbol{m}_{ht}$ around some $\tilde{\boldsymbol{\beta}}$. The algorithm outputs z_{ht}, $\boldsymbol{m}_{ht}$ and $\boldsymbol{S}_{ht}$ and is:

Algorithm 2 Input $\tilde{\boldsymbol{\beta}}_h$, $\boldsymbol{X}_{ht}$

Initialize: $j = 1, z_{ht} = 0, \boldsymbol{m}_{ht} = \boldsymbol{0}_K, \boldsymbol{S}_{ht} = z_{ht}\boldsymbol{I}_K$
  while $j \leq J$ **do**
    $\alpha = e^{\boldsymbol{x}_{htj}^T \tilde{\boldsymbol{\beta}}_h}$
    $\boldsymbol{l} = \boldsymbol{x}_{htj} - \boldsymbol{m}_{hj}$
    $\boldsymbol{S}_{ht} = \boldsymbol{S}_{ht} + \frac{\tanh\left(\frac{1}{2}\log(a/z_{ht})\right)}{2\log(a/z_{ht})}\boldsymbol{l}\boldsymbol{l}^T$
    $\boldsymbol{m}_{ht} = \boldsymbol{m}_{ht} + \frac{a}{z_{ht}+a}\boldsymbol{l}$
    $z_{ht} = z + a$
    $j = j + 1$
  end while
Output: $z_{ht}, \boldsymbol{m}_{ht}, \boldsymbol{S}_{ht}$

After taking expectations this bound leads to

$$\mathbb{E}_{q(\boldsymbol{\beta}_h)}\left[\log\left(\sum_{j=1}^{J} e^{\boldsymbol{x}_{htj}^T\boldsymbol{\beta}_h}\right)\right] \leq \log z_{ht} + \frac{1}{2}\left(\boldsymbol{\mu}_h - \tilde{\boldsymbol{\beta}}_h\right)^T \boldsymbol{S}_{ht}\left(\boldsymbol{\mu}_h - \tilde{\boldsymbol{\beta}}_h\right) + \frac{1}{2} tr\left[\boldsymbol{\Sigma}_h \boldsymbol{S}_{ht}\right] + \left(\boldsymbol{\mu}_h - \tilde{\boldsymbol{\beta}}_h\right)^T \boldsymbol{m}_{ht}.$$

This quadratic bound again leads to analytic updates of the subjects' variational parameters in the form of

$$\boldsymbol{\Sigma}_h = \left(\omega\boldsymbol{\Upsilon} + \sum_{t=1}^{T_h} \boldsymbol{S}_{ht}\right)^{-1}$$
$$\boldsymbol{\mu}_h = \boldsymbol{\Sigma}_h\left(\omega\boldsymbol{\Upsilon}\sum_{t=1}^{T_h} \boldsymbol{S}_{ht}\tilde{\boldsymbol{\beta}}_h - \boldsymbol{m}_{ht}\right).$$

We chose to update $\tilde{\boldsymbol{\beta}}_h$ as $\boldsymbol{\mu}_{\boldsymbol{\zeta}}$.

C Generating Quasi Monte Carlo Samples

In this section we briefly show how we constructed the QMC samples. We chose to construct the QMC samples according to Hickernell et al. (2000) which are called extensible shifted lattice points (ESLP). For more details on the properties and optimal construction of such samples we refer you to the previously mentioned reference. The goal of QMC samples is to sample from the K-dimensional unit cube $[0,1)^K$ in a way such that the discrepancy between the empirical distribution of the QMC sample and the continuous uniform distribution is small. If this goal is successful, relatively precise high di-

mensional integration can be performed with a relatively small number of samples which benefits the computational efficiency of the algorithm. Say that we require R samples where R is some integer power of an integer base, i.e. $R = b^m$, $b \geq 2$ and b and$m$ are integers. We also require a generating vector $\boldsymbol{h}$ of dimension K. Following Hickernell et al. (2000) we use the generating vector $\boldsymbol{h} = \left(1, \eta, \eta^2, \ldots, \eta^{K-1}\right)^T$. The next step is to write the integers $0, 1, 2, \ldots, b^m - 1$ in base b form. So, for instance, if $b = 2$ and $m = 3$, we have $R = 2^3 = 8$ samples. The integer 0 would be written as $0 \times 2^0 + 0 \times 2^1 + 0 \times 2^2$, 1 would be written as $1 \times 2^0 + 0 \times 2^1 + 0 \times 2^2$ all up to $7 = 1 \times 2^0 + 1 \times 2^1 + 1 \times 2^2$. So now we have for all the integers $0, \ldots, b^m - 1$ the coefficients of its base b representation which can be written as

$$i = \sum_{k=0}^{b^m-1} i_k b^k = i_0 b^0 + i_1 b^1 + \ldots \quad .$$

Define now the function $\phi_b(i)$ as

$$\phi_b(i) = \sum_{k=0}^{b^m-1} i_k b^{-(k+1)} = i_0 b^{-1} + i_1 b^{-2} + \ldots \quad .$$

The final element to generate the QMC sample is to introduce a random shift vector $\boldsymbol{u} = (u_1, \ldots, u_K)^T$ which is an element of the unit cube $[0,1)^K$. The ith QMC sample is now defined as $\left(\{\phi_b(i) h_1 + u_1\}, \ldots, \{\phi_b(i) h_K + u_K\}\right)^T$ where $\{x\}$ is a function which takes the fractional part of x, i.e. $\{x\} = x \pmod 1$. Hickernell et al. (2000) used a periodizing transformation on the final QMC samples as this appeared to increase the accuracy of the method. We also used this transformation which is defined as $x' = |2x - 1|$. Finally, as we are interested in samples from a multivariate normal distribution rather than from a multivariate uniform distribution we apply the inverse normal distribution transformation on all coordinates. This results in a QMC sample from a standard K-dimensional normal distribution. In our algorithms we used base $b = 2$, exponents $m = 6, \ldots, 12$ for the conditional logit model and $m = 6, 7, 8$ for the mixed logit model. Furthermore, we used $\eta = 1571$ from Hickernell et al. (2000, table 4.1) which is appropriate for bases in $6, \ldots, 12$ and up to $K = 33$ dimensions.

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be