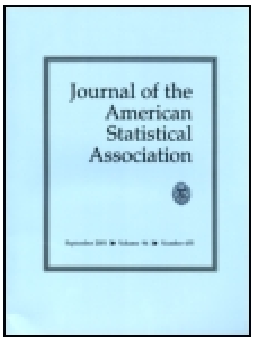

Journal of the
American
Statistical
Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

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

Variational Inference for Large-Scale Models of Discrete Choice

Michael Braun & Jon McAuliffe

To cite this article: Michael Braun & Jon McAuliffe (2010) Variational Inference for Large-Scale Models of Discrete Choice, Journal of the American Statistical Association, 105:489, 324-335, DOI: 10.1198/jasa.2009.tm08030

To link to this article: https://doi.org/10.1198/jasa.2009.tm08030

Variational Inference for Large-Scale Models of Discrete Choice

Michael BRAUN and Jon MCAULIFFE

Discrete choice models are commonly used by applied statisticians in numerous fields, such as marketing, economics, finance, and operations research. When agents in discrete choice models are assumed to have differing preferences, exact inference is often intractable. Markov chain Monte Carlo techniques make approximate inference possible, but the computational cost is prohibitive on the large datasets now becoming routinely available. Variational methods provide a deterministic alternative for approximation of the posterior distribution. We derive variational procedures for empirical Bayes and fully Bayesian inference in the mixed multinomial logit model of discrete choice. The algorithms require only that we solve a sequence of unconstrained optimization problems, which are shown to be convex. One version of the procedures relies on a new approximation to the variational objective function, based on the multivariate delta method. Extensive simulations, along with an analysis of real-world data, demonstrate that variational methods achieve accuracy competitive with Markov chain Monte Carlo at a small fraction of the computational cost. Thus, variational methods permit inference on datasets that otherwise cannot be analyzed without possibly adverse simplifications of the underlying discrete choice model. Appendices C through F are available as online supplemental materials.

KEY WORDS: Bayesian statistics; Convex optimization; Empirical Bayes; Multinomial logit; Random utility model.

1. INTRODUCTION

Discrete choice models have a long history in statistical analysis, appearing in applications as varied as the analysis of consumer choice data (Guadagni and Little 1983; Fader and Hardie 1996), transportation planning (Theil 1969; McFadden 1974; Ben-Akiva and Lerman 1985), economic demand estimation (Train, McFadden, and Ben-Akiva 1987; Revelt and Train 1998), new product development (Moore, Louviere, and Verma 1999), portfolio analysis (Uhler and Cragg 1971), and health services deployment (Hall et al. 2002). They apply to situations where agents (also called choosers or decision makers) select items from a finite collection of alternatives (the choice set), either once or repeatedly over time. For example, in a marketing context, agents are "households"; each household makes a number of "trips" to a store and we observe the items selected for purchase on each trip.

Heterogeneous discrete choice models, which allow preferences to differ across agents, are based on a hierarchical regression formulation. We have agents numbered $h = 1, \dots, H$, each with an unseen parameter vector $\boldsymbol{\theta}_h$ encoding preferences over item attributes. We observe one or more choice events $Y_h \sim p(y_h|\boldsymbol{\theta}_h)$ per agent. The $\boldsymbol{\theta}_h$'s are modeled as independent draws from a prior distribution $p(\boldsymbol{\theta}_h|\boldsymbol{\phi})$, where $\boldsymbol{\phi}$ is a hyperparameter. This prior represents the heterogeneity of preferences in the agent population. Inference in such a hierarchical model allows us to pool information across decision makers. If we use an empirical Bayes point estimate of $\boldsymbol{\phi}$ (Robbins 1955), the posterior distribution of each $\boldsymbol{\theta}_h$ depends on all of $Y_1, \dots, Y_H$, through the common estimate $\hat{\boldsymbol{\phi}}$. In a fully Bayesian setup, integrating out the random variable $\boldsymbol{\phi}$ creates similar dependence.

The marginal likelihood corresponding to one agent in a heterogeneous model is

$$p(y_h|\boldsymbol{\phi}) = \int p(y_h|\boldsymbol{\theta}_h) p(\boldsymbol{\theta}_h|\boldsymbol{\phi})\, d\boldsymbol{\theta}_h. \tag{1.1}$$

In most cases, including the "random utility" discrete choice model we study in this article, Equation (1.1) does not exist in closed form. As a consequence, we must use approximate methods both for empirical Bayes and fully Bayesian inference.

A standard empirical Bayes technique is to approximate Equation (1.1) using Monte Carlo integration. But to match the asymptotics of maximum likelihood, the number of draws per agent must grow faster than the square root of the number of agents (Train 2003), which is infeasible for large-scale problems. The usual approach to the fully Bayesian random utility model is Markov chain Monte Carlo (MCMC) simulation (Albert and Chib 1993; Allenby and Lenk 1994; Allenby and Rossi 1999). MCMC provides approximate draws from the joint posterior distribution on $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_H$ and $\boldsymbol{\phi}$. The draws enable the estimation of Equation (1.1) and related integrals. However, the amount of computation required for MCMC scales up with the number of agents—even if we are only interested in $\boldsymbol{\phi}$, a Gibbs sampling algorithm must repeatedly generate draws for all of $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_H$.

Variational methods (Jordan et al. 1999; Wainwright and Jordan 2003) offer a deterministic alternative for approximate inference. With variational inference, we maximize a data-dependent lower bound on the marginal likelihood found in Equation (1.1), over a set of auxiliary parameters distinct from the model parameters. In the fully Bayesian specification, the end result is an approximate joint posterior distribution for $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_H$ and $\boldsymbol{\phi}$. For empirical Bayes, variational techniques lead to a point estimate $\hat{\boldsymbol{\phi}}$ as well as an approximate posterior distribution for the $\boldsymbol{\theta}_h$'s.

The main advantage of variational methods versus MCMC is computational efficiency. Variational inference algorithms typically converge to their final approximation in far less time than

Michael Braun is the Homer A. Burnell Career Development Professor and Assistant Professor of Management Science, MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02142 (E-mail: *braunm@mit.edu*). Jon McAuliffe is Adjunct Assistant Professor, Department of Statistics, University of California, Berkeley, CA 94720 (E-mail: *jon@stat.berkeley.edu*). The authors gratefully acknowledge research assistance from Alexander Spicer and Liz Theurer, as well as helpful discussions with David Blei. The authors also thank Cathy Trower, Jordan Louviere, and the Center for the Study of Choice at the University of Technology, Sydney for sharing the stated-choice experiment data referenced in this article.

Journal of the American Statistical Association
March 2010, Vol. 105, No. 489, Theory and Methods
DOI: 10.1198/jasa.2009.tm08030

it takes to generate an adequate number of MCMC draws. This advantage comes at the cost of a biased approximation, in contrast to the consistency guarantees that accompany MCMC. We give evidence in Section 4 that, for our random utility discrete choice model, variational bias is minimal and the computational speedup is very large. Variational convergence is also easy to assess, in contrast to MCMC.

Furthermore, the size of the variational representation is fixed, while the size of the MCMC approximation increases with the number of draws. Variational techniques can, therefore, be applied to much larger datasets. For example, with 50,000 agents and 40,000 total MCMC draws, using a 25-dimensional $\boldsymbol{\theta}_h$, the MCMC representation of the posterior for $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_H$ consists of 40,000 × 25 × 50,000 or over 50 billion floating-point values. In fact, many datasets today contain observations on millions of agents, models can contain far more than 25 preference parameters, and MCMC chains may require hundreds of thousands of iterations.

These difficulties are well known. MCMC is rarely applied to large-scale heterogeneous models. Indeed, to address scalability, it is common to work with data from a subset of individuals, or a subset of choice items. However, this approach discards information that is valuable in the inferential process, and it can lead to poor estimates (Zanutto and Bradlow 2006).

In this article, we derive variational algorithms for a common discrete choice model—the mixed multinomial logit (MML) model. We study this model because it is the workhorse of discrete choice theory and is well known in many disciplines, including economics and marketing. There are other popular discrete choice models in the literature, but the mixed multinomial logit has appeal: it is conceptually simple, yet still exhibits the MCMC inference issues just described.

The rest of the article is organized as follows. Section 2 presents the details of the MML model. In Section 3, we describe variational procedures suitable for empirical Bayes and fully Bayesian inference under the MML model. One version of these procedures involves a novel approximation of the variational objective function, based on the multivariate delta method for moments. In Section 4, we compare this version of the variational methods to MCMC for the MML model, using an extensive suite of simulated datasets as well as real-world data from a stated-choice experiment. Section 5 closes with discussion and future directions. Technical arguments, derivations, and MCMC convergence diagnostics are relegated to Appendices A through F.

2. THE MIXED MULTINOMIAL LOGIT MODEL OF DISCRETE CHOICE

Let there be H agents, indexed $h = 1, \dots, H$. We observe a total of T_h choice-event outcomes for agent h. At each choice event, the agent selects from among a fixed set of J items, indexed $j = 1, \dots, J$. The items are differentiated according to K attributes, indexed $k = 1, \dots, K$. The jth item's value for the kth attribute can vary across agents and from one choice event to another. For example, households might shop at different stores charging various prices for the same good, and the price of a good may change over time within a single store. We denote by $\mathbf{x}_{ht}$ the $J \times K$ matrix of attribute values, also called covariates, that agent h encounters at her tth choice event. The jth row of $\mathbf{x}_{ht}$ is denoted $\mathbf{x}_{htj}^\top$. The outcome of this choice event is the categorical random variable $\mathbf{y}_{ht}$, which we represent as a $J \times 1$ indicator vector.

We use the observed $(\mathbf{x}_{ht}, \mathbf{y}_{ht})$ pairs to infer which attributes have the strongest association with item choice. To this end, let U_{ht}^j denote the utility that accrues to agent h if she chooses item j at her tth choice event. This approach, called a "random utility model" (Train 2003), assumes utility is a noisy linear function of the attributes: $U_{ht}^j = \boldsymbol{\beta}_h^\top \mathbf{x}_{htj} + e_{ht}^j$. Here, $\boldsymbol{\beta}_h$ is a $K \times 1$ vector of agent-specific "tastes" or "preference loadings" for the item attributes and e_{ht}^j is a random error term representing unobserved utility.

We assume that each agent, at each choice event, selects the item maximizing her utility. In the mixed multinomial logit model, we further assume the random error terms e_{ht}^j are iid from a Gumbel Type 2 distribution. The implied choice probabilities turn out to be

$$P(y_{ht}^j = 1 | \mathbf{x}_{ht}, \boldsymbol{\beta}_h) = \frac{\exp(\boldsymbol{\beta}_h^\top \mathbf{x}_{htj})}{\sum_{j'} \exp(\boldsymbol{\beta}_h^\top \mathbf{x}_{htj'})}, \qquad j = 1, \dots, J, \tag{2.1}$$

(McFadden 1974). In discrete choice modeling, the right-hand side of Equation (2.1) is called the "multinomial logit" distribution, denoted $\mathrm{MNL}(\mathbf{x}_{ht}, \boldsymbol{\beta}_h)$. It is essentially the same as the multilogistic function used in polychotomous logistic regression, and it is often called the soft-max function in machine learning research.

We further assume that $\boldsymbol{\beta}_{1:H}$ are iid from a K-variate normal distribution with mean vector $\boldsymbol{\zeta}$ and covariance matrix $\boldsymbol{\Omega}$, which we write as $\mathcal{N}_K(\boldsymbol{\zeta}, \boldsymbol{\Omega})$. For empirical Bayes estimation, the model is now completely specified:

$$\mathbf{y}_{ht} | \mathbf{x}_{ht}, \boldsymbol{\beta}_h \overset{\text{ind}}{\sim} \mathrm{MNL}(\mathbf{x}_{ht}, \boldsymbol{\beta}_h), \qquad h = 1, \dots, H, t = 1, \dots, T_h, \tag{2.2}$$

$$\boldsymbol{\beta}_h | \boldsymbol{\zeta}, \boldsymbol{\Omega} \overset{\text{iid}}{\sim} \mathcal{N}_K(\boldsymbol{\zeta}, \boldsymbol{\Omega}), \qquad h = 1, \dots, H. \tag{2.3}$$

The top-level parameters $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$, to be estimated by maximum marginal likelihood, represent the distribution of attribute preferences across the population. In particular, $\boldsymbol{\Omega}$ gives us information about the correlation of preferences between agents.

A fully Bayesian approach requires hyperprior distributions for $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$. As is standard, we use conditionally conjugate distributions:

$$\begin{aligned} \boldsymbol{\zeta} | \boldsymbol{\beta}_0, \boldsymbol{\Omega}_0 &\sim \mathcal{N}_K(\boldsymbol{\beta}_0, \boldsymbol{\Omega}_0), \\ \boldsymbol{\Omega} | \mathbf{S}, \nu &\sim \mathcal{W}^{-1}(\mathbf{S}^{-1}, \nu). \end{aligned} \tag{2.4}$$

In Equation (2.4), $\boldsymbol{\beta}_0$ and $\boldsymbol{\Omega}_0$ are prespecified hyperparameters; $\mathcal{W}^{-1}(\mathbf{S}^{-1}, \nu)$ is the inverse Wishart distribution with scale matrix $\mathbf{S}^{-1}$ and ν degrees of freedom; and $\mathbf{S}$ and ν are hyperparameters fixed in advance. We call the fully Bayesian approach to MML model inference "hierarchical Bayes."

3. VARIATIONAL INFERENCE FOR THE MML MODEL

We presented the component hierarchical distributions in the mixed multinomial logit model. Now we turn to the question of estimation and inference procedures. In the following, variable names inside of $p(\cdot)$ are used to distinguish among densities: we denote the pdf's in Equations (2.2) to (2.4) by $p(\mathbf{y}_{ht}|\mathbf{x}_{ht}, \boldsymbol{\beta}_h)$, $p(\boldsymbol{\beta}_h|\boldsymbol{\zeta}, \boldsymbol{\Omega})$, $p(\boldsymbol{\zeta}|\boldsymbol{\beta}_0, \boldsymbol{\Omega}_0)$, and $p(\boldsymbol{\Omega}|\mathbf{S}, \nu)$, respectively. We let $\mathcal{D} = \{\mathbf{x}_{ht}, \mathbf{y}_{ht}\}$ denote all observed variables, i.e., the data.

For the empirical Bayes version of the MML model, the posterior density of the latent preference vectors, $p(\boldsymbol{\beta}_{1:H}|\mathcal{D}, \boldsymbol{\zeta}, \boldsymbol{\Omega})$, is

$$\prod_{h=1}^{H} \frac{p(\boldsymbol{\beta}_h|\boldsymbol{\zeta}, \boldsymbol{\Omega}) \prod_{t=1}^{T_h} p(\mathbf{y}_{ht}|\mathbf{x}_{ht}, \boldsymbol{\beta}_h)}{\int p(\boldsymbol{\beta}_h|\boldsymbol{\zeta}, \boldsymbol{\Omega}) \prod_{t=1}^{T_h} p(\mathbf{y}_{ht}|\mathbf{x}_{ht}, \boldsymbol{\beta}_h)\, d\boldsymbol{\beta}_h}. \tag{3.1}$$

The joint posterior density for hierarchical Bayes, $p(\boldsymbol{\beta}_{1:H}, \boldsymbol{\zeta}, \boldsymbol{\Omega}|\mathcal{D})$, is

$$p(\boldsymbol{\zeta})p(\boldsymbol{\Omega}) \prod_{h=1}^{H} p(\boldsymbol{\beta}_h|\boldsymbol{\zeta}, \boldsymbol{\Omega}) \prod_{t=1}^{T_h} p(\mathbf{y}_{ht}|\mathbf{x}_{ht}, \boldsymbol{\beta}_h)$$
$$\Bigg/ \Bigg(\int p(\boldsymbol{\zeta})p(\boldsymbol{\Omega}) \prod_{h=1}^{H} \int p(\boldsymbol{\beta}_h|\boldsymbol{\zeta}, \boldsymbol{\Omega})$$
$$\times \prod_{t=1}^{T_h} p(\mathbf{y}_{ht}|\mathbf{x}_{ht}, \boldsymbol{\beta}_h)\, d\boldsymbol{\beta}_h\, d\boldsymbol{\zeta}\, d\boldsymbol{\Omega} \Bigg). \tag{3.2}$$

The numerator in both cases is the joint density of latent and observed variables, computed by multiplying together the densities defined in the model hierarchy in Equations (2.2) to (2.4).

The integrals appearing in these posterior densities have no closed form. As a consequence, exact inference is intractable. Variational inference is a deterministic alternative to the MCMC methods usually applied to this problem. A variational algorithm selects from a prespecified family of distributions $\mathcal{Q}$ the best approximation to the true posterior distribution. We define $\mathcal{Q}$ so that all of its members permit tractable probability calculations. Then, wherever we need the true posterior, such as for an expectation, we use the approximating variational distribution instead. This plug-in idea underlies MCMC methods as well—in place of the true posterior, we substitute the empirical distribution of the MCMC posterior draws.

3.1 Variational Empirical Bayes

In this section we give an overview of a variational algorithm for approximate empirical Bayes estimation in the MML model. Appendix A fills in the details. We first specify a family of approximating distributions $\mathcal{Q} := \{q(\boldsymbol{\beta}_{1:H}|\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \Lambda\}$ for the true posterior distribution in Equation (3.1). Since this posterior factors over h, we take $\mathcal{Q}$ to be a family of factored distributions as well, so that $q(\boldsymbol{\beta}_{1:H}|\boldsymbol{\lambda}) := \prod_h q(\boldsymbol{\beta}_h|\boldsymbol{\lambda}_h)$. In particular, each factor $q(\boldsymbol{\beta}_h|\boldsymbol{\lambda}_h)$ is a K-variate normal density, with mean $\boldsymbol{\mu}_h$ and covariance matrix $\boldsymbol{\Sigma}_h$.

For the particular dataset at hand, we want to find $q(\boldsymbol{\beta}_{1:H}|\boldsymbol{\lambda}^*)$, the best approximation in $\mathcal{Q}$ to the posterior distribution $p(\boldsymbol{\beta}_{1:H}|\mathcal{D}, \boldsymbol{\zeta}, \boldsymbol{\Omega})$. To make the idea of a best approximation precise, we measure discrepancy with the Kullback–Leibler (KL) divergence (also called the relative entropy). Shortening $q(\boldsymbol{\beta}_{1:H}|\boldsymbol{\lambda})$ to $q_{\boldsymbol{\lambda}}$, the optimal variational parameters are given by

$$\boldsymbol{\lambda}^* = \operatorname*{argmin}_{\boldsymbol{\lambda} \in \Lambda} \mathrm{KL}[q_{\boldsymbol{\lambda}} \parallel p]. \tag{3.3}$$

We argue in Appendix D that $\mathrm{KL}[q_{\boldsymbol{\lambda}} \parallel p]$ exists and is finite for all the probability density functions considered in this article, so the optimization problem in Equation (3.3) is well posed.

We can express the KL divergence between $q_{\boldsymbol{\lambda}}$ and p as

$$\mathrm{KL}[q_{\boldsymbol{\lambda}} \parallel p] = \mathbb{E}_{q_{\boldsymbol{\lambda}}} \log\left[\frac{q(\boldsymbol{\beta}_{1:H}|\boldsymbol{\lambda})}{p(\boldsymbol{\beta}_{1:H}|\mathcal{D}, \boldsymbol{\zeta}, \boldsymbol{\Omega})}\right] \tag{3.4}$$
$$= -\mathcal{L}(\boldsymbol{\lambda}; \boldsymbol{\zeta}, \boldsymbol{\Omega}) + \log p(\mathcal{D}|\boldsymbol{\zeta}, \boldsymbol{\Omega}), \tag{3.5}$$

where we define

$$\mathcal{L}(\boldsymbol{\lambda}; \boldsymbol{\zeta}, \boldsymbol{\Omega}) := -\mathbb{E}_{q_{\boldsymbol{\lambda}}} \log\left[\frac{q(\boldsymbol{\beta}_{1:H}|\boldsymbol{\lambda})}{p(\boldsymbol{\beta}_{1:H}, \mathcal{D}|\boldsymbol{\zeta}, \boldsymbol{\Omega})}\right]. \tag{3.6}$$

Here, $\mathbb{E}_{q_{\boldsymbol{\lambda}}}$ denotes an average over $\boldsymbol{\beta}_{1:H}$, using $q(\boldsymbol{\beta}_{1:H}|\boldsymbol{\lambda})$. Because $\mathrm{KL}[q_{\boldsymbol{\lambda}} \parallel p]$ is nonnegative, Equation (3.5) implies that, for all distributions $q_{\boldsymbol{\lambda}}$,

$$\mathcal{L}(\boldsymbol{\lambda}; \boldsymbol{\zeta}, \boldsymbol{\Omega}) \leq \log p(\mathcal{D}|\boldsymbol{\zeta}, \boldsymbol{\Omega}). \tag{3.7}$$

The likelihood is called the "evidence" in some contexts; so we call $\mathcal{L}(\boldsymbol{\lambda}; \boldsymbol{\zeta}, \boldsymbol{\Omega})$ the *evidence lower bound (ELBO)* function.

Using Equation (3.5), we can formulate a maximization problem equivalent to Equation (3.3), having the same optimal point $q(\boldsymbol{\beta}_{1:H}|\boldsymbol{\lambda}^*)$:

$$\boldsymbol{\lambda}^* = \operatorname*{argmin}_{\boldsymbol{\lambda} \in \Lambda} \mathcal{L}(\boldsymbol{\lambda}; \boldsymbol{\zeta}, \boldsymbol{\Omega}). \tag{3.8}$$

The equivalence of Equations (3.3) and (3.8), together with the bound in Equation (3.7), shows that the best approximation in $\mathcal{Q}$ to the posterior also yields the tightest lower bound on the marginal likelihood. We focus on Equation (3.8) rather than Equation (3.3)—to evaluate the KL divergence, we would need to be able to compute the marginal likelihood, bringing us back to the original problem.

As we detail in Appendix A, it turns out that the ELBO objective in Equation (3.8) has no closed form in our discrete choice modeling setup. We consider two surrogate ELBO functions $\tilde{\mathcal{L}}(\boldsymbol{\lambda}; \boldsymbol{\zeta}, \boldsymbol{\Omega})$: one based on Jensen's inequality, which preserves the lower-bounding property in Equation (3.7), and a second based on the multivariate delta method for moments, which does not. The empirical results reported in Section 4 are based on the second surrogate, which yielded better performance in our studies.

Whichever surrogate is used, the optimization problem becomes

$$\boldsymbol{\lambda}^* = \operatorname*{argmin}_{\boldsymbol{\lambda} \in \Lambda} \tilde{\mathcal{L}}(\boldsymbol{\lambda}; \boldsymbol{\zeta}, \boldsymbol{\Omega}). \tag{3.9}$$

The variational parameters to be adjusted are $\boldsymbol{\lambda} = \{\boldsymbol{\mu}_{1:H}, \boldsymbol{\Sigma}_{1:H}\}$. We conduct the variational inference in Equation (3.9) using block coordinate ascent on the coordinate blocks $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_H, \boldsymbol{\Sigma}_H$. The coordinate updates do not have a closed form, but each update solves a smooth, unconstrained convex optimization problem, as we show in Appendix A. There we also derive the gradient and Hessian for the $\boldsymbol{\mu}_h$ update, as well as gradients for the $\boldsymbol{\Sigma}_h$ update under two different parameterizations. Note that, although the coordinate-block optimizations are convex, the variational problem in Equation (3.9) is not convex in the full parameter vector $\boldsymbol{\lambda}$. So, in general, the fitted variational parameters correspond to a local optimum rather than the global optimum.

To finish with empirical Bayes, we explain how to obtain approximate maximum likelihood estimates (MLE's) $\hat{\boldsymbol{\zeta}}$ and $\hat{\boldsymbol{\Omega}}$. Notice the variational inference procedure in Equation (3.9) yields a value $\boldsymbol{\lambda}^*$ for fixed $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$. We alternate this variational inference step with a complementary optimization over $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$. In fact, these optimizations constitute a version of the expectation-maximization (EM) algorithm for computing MLE's. The standard E-step, where we compute the posterior expected complete log likelihood, is replaced with a variational E-step, where the expected complete log likelihood is approximated by finding $q(\boldsymbol{\beta}_{1:H}|\boldsymbol{\lambda}^*)$. The variational EM algorithm alternates between the following steps until $\hat{\boldsymbol{\zeta}}$, $\hat{\boldsymbol{\Omega}}$, and the variational parameters stabilize:

E-step (variational). Using the current $\hat{\boldsymbol{\zeta}}$ and $\hat{\boldsymbol{\Omega}}$, run the block coordinate ascent algorithm as described in Appendix A, yielding new variational parameter values $\{\boldsymbol{\mu}^*_{1:H}, \boldsymbol{\Sigma}^*_{1:H}\}$.

M-step. Using the current variational parameter values, update the empirical Bayes parameter estimates: $(\hat{\boldsymbol{\zeta}}, \hat{\boldsymbol{\Omega}}) \leftarrow \operatorname{argmax}_{\boldsymbol{\zeta},\boldsymbol{\Omega}} \tilde{\mathcal{L}}(\boldsymbol{\mu}^*_{1:H}, \boldsymbol{\Sigma}^*_{1:H}; \boldsymbol{\zeta}, \boldsymbol{\Omega})$.

The M-step maximizes the adjusted $\tilde{\mathcal{L}}$ function from the E-step over $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$, as a surrogate for the true log-likelihood. We then readjust and remaximize until convergence. Appendix A gives details on initialization and the M-step update.

3.2 Variational Hierarchical Bayes

Fully Bayesian inference in the mixed multinomial logit model requires calculations under the posterior $p(\boldsymbol{\beta}_{1:H}, \boldsymbol{\zeta}, \boldsymbol{\Omega}|\mathcal{D})$ given in Equation (3.2). In one sense, the setting is simpler than empirical Bayes: there are no unknown top-level parameters to estimate. All we need is to extend the previous section's variational inference procedure to include $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$. Appendix A.4 reports the details behind the extension; here we summarize the main ideas.

Although the joint posterior in Equation (3.2) is not factorized, we continue to use a family $\mathcal{Q}$ of factorized distributions for the variational approximation:

$$\begin{aligned}\mathcal{Q} &\ni q(\boldsymbol{\beta}_{1:H}, \boldsymbol{\zeta}, \boldsymbol{\Omega}|\boldsymbol{\lambda})\\ &:= q(\boldsymbol{\zeta}|\boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta) q(\boldsymbol{\Omega}|\boldsymbol{\Upsilon}^{-1}, \omega) \prod_{h=1}^{H} q(\boldsymbol{\beta}_h|\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h). \qquad (3.10)\end{aligned}$$

Using a factored family for a nonfactored posterior is commonly called mean-field variational inference (Opper and Saad 2001). In Equation (3.10), $q(\boldsymbol{\zeta}|\boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta)$ is a K-variate normal density; $q(\boldsymbol{\Omega}|\boldsymbol{\Upsilon}^{-1}, \omega)$ is an inverse Wishart density; and the $q(\boldsymbol{\beta}_h)$ factors are K-variate normal densities as before. In the analysis and the algorithm, it is convenient to use a well-known equivalence, treating $q(\boldsymbol{\Omega}|\boldsymbol{\Upsilon}^{-1}, \omega)$ as a Wishart distribution $\mathcal{W}(\boldsymbol{\Upsilon}, \omega)$ on $\boldsymbol{\Omega}^{-1}$. Therefore, we optimize over variational parameters $\boldsymbol{\lambda} := \{\boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta, \boldsymbol{\Upsilon}, \omega, \boldsymbol{\mu}_{1:H}, \boldsymbol{\Sigma}_{1:H}\}$. The variational problem for hierarchical Bayes is to find the best approximating distribution $q(\boldsymbol{\beta}_{1:H}, \boldsymbol{\zeta}, \boldsymbol{\Omega}|\boldsymbol{\lambda}^*)$ in $\mathcal{Q}$ by solving the analog of Equation (3.9):

$$\boldsymbol{\lambda}^* = \operatorname*{argmin}_{\boldsymbol{\lambda}\in\Lambda} \tilde{\mathcal{L}}(\boldsymbol{\lambda}). \qquad (3.11)$$

As with empirical Bayes, we use a block coordinate ascent optimization algorithm to solve Equation (3.11), iterating through the coordinate blocks that define $\boldsymbol{\lambda}$. Here again, all coordinate updates are convex optimizations, while the overall problem is not convex in $\boldsymbol{\lambda}$. The details appear in Appendix A: updates for $\boldsymbol{\mu}_\zeta$, $\boldsymbol{\Sigma}_\zeta$, and $\boldsymbol{\Upsilon}$ all have simple closed forms; ω has a closed form that requires no updating; and the $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ updates are similar to the empirical Bayes case.

4. EMPIRICAL RESULTS

We compared the variational methods described in the previous section to a standard and widely used MCMC approach (Rossi and Allenby 2003), using simulated as well as real data. Specifically, we carried out a suite of simulations, which we will call simulation study A; an analysis of a stated-choice experiment with academic respondents; and additional simulations based on the stated-choice data, called simulation study B. We present each in turn.

4.1 Simulation Study A

This study contains 32 different simulation scenarios, each one based on the discrete choice model given by Equations (2.2) and (2.3). To simulate a dataset with J choice items, K item attributes, and H agents we first fixed values of $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$, the parameters controlling the distribution of preferences in the agent population. We then independently drew a $\boldsymbol{\beta}_h$ vector for each agent, according to Equation (2.3). We also drew for each agent 25 iid $J \times K$ item attribute matrices $\mathbf{x}_{ht}$ consisting of iid $\mathcal{N}(0, 0.5^2)$ entries. Finally, for each agent, we used $\mathbf{x}_{ht}$ and $\boldsymbol{\beta}_h$ to simulate 25 choice events $\mathbf{y}_{ht}$, according to Equation (2.2). Thus, in our datasets, each agent has 25 observed choices.

We created the 32 different scenarios by varying J, K, H, and the selection of $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$. Specifically, each scenario corresponds to a distinct configuration of the following candidate values: 3 or 12 choice items J; 3 or 10 item attributes K; 250, 1000, 5000, or 25,000 agents H; and "low" or "high" heterogeneity of the agent population. In the low-heterogeneity scenario, the $K \times 1$ vector $\boldsymbol{\zeta}$ consists of evenly spaced values from -2 to 2, and the $K \times K$ matrix $\boldsymbol{\Omega}$ is 0.25 times the identity matrix. In the high-heterogeneity scenario, $\boldsymbol{\zeta}$ is the same, but $\boldsymbol{\Omega}$ is the identity matrix. The datasets with high heterogeneity have more diverse collections of preference vectors $\boldsymbol{\beta}_h$.

We simulated each of the scenarios 10 times independently. For each replication of each scenario, we ran variational empirical Bayes (VEB), variational hierarchical Bayes (VB), and the standard MCMC algorithm on the observable data. We used the version of the variational procedures based on approximation D1 (cf. Appendix A). VEB was declared to have converged as soon as an E-step/M-step iteration caused the parameter estimates' joint Euclidean norm to change by less than 10^{-4}, relative to their norm at the beginning of the iteration (here we mean the joint norm of all the variational parameters, together with the model parameter estimates $\hat{\boldsymbol{\zeta}}$ and $\hat{\boldsymbol{\Omega}}$). The convergence criterion for VB was the same, except $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$ do not have point estimates. We looked instead at the change in the variational parameters corresponding to their posterior approximation: $\boldsymbol{\mu}_\zeta$, $\boldsymbol{\Sigma}_\zeta$, and $\boldsymbol{\Upsilon}$.

Choosing MCMC convergence criteria is more delicate. We tried to set the number of burn-in iterations and the thinning ratio algorithmically, using the technique of Raftery and Lewis (1992). But in several of the scenarios, typical control parameter values, such as the default settings for the `raftery.diag` function in the R package `coda` (Plummer et al. 2006) led to a very large number of burn-in iterations. Trace plots of the sampled parameters indicated these large burn-in values were unnecessary, so using them would have been unfair to MCMC in our timing comparisons. Instead, for each scenario we generated 25,000 draws from each of five independent chains, dropped the first half of them as burn-in, and retained every fifth of the remainder. This gave us 12,500 draws per scenario. We confirmed that these draws represent a good approximation of the posterior distributions by computing the Gelman–Rubin scale reduction factors (Gelman and Rubin 1992) and the effective sample sizes (these values are available in Appendix F). However, we did not need this many draws to reach convergence and including all of them would be unfair to MCMC in the timing comparisons. Therefore, timing comparisons are presented using the time to collect 6000 draws per scenario, which appears close to the minimum number required for convergence.

4.1.1 Accuracy. Our measure of accuracy for each inference procedure is based on the *predictive choice distribution*. Informally, this distribution gives the item choice probabilities exhibited by the "average" agent when shown an item attribute matrix $\mathbf{x}_{\text{new}}$. In our simulations, we know the true values of $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$, so we can compute the true predictive choice distribution

$$p(\mathbf{y}_{\text{new}}|\mathbf{x}_{\text{new}}, \boldsymbol{\zeta}, \boldsymbol{\Omega}) = \int p(\mathbf{y}_{\text{new}}|\mathbf{x}_{\text{new}}, \boldsymbol{\beta}) p(\boldsymbol{\beta}|\boldsymbol{\zeta}, \boldsymbol{\Omega})\, d\boldsymbol{\beta} \tag{4.1}$$

$$= \mathbb{E}_{\boldsymbol{\beta}} p(\mathbf{y}_{\text{new}}|\mathbf{x}_{\text{new}}, \boldsymbol{\beta}). \tag{4.2}$$

Equation (4.2) explains the "average agent" interpretation of the predictive choice distribution. A slightly different take on Equations (4.1) and (4.2) is the following: if we want to forecast the item choice probabilities of a previously unobserved decision maker, we can think of her as a random draw from the agent population. Under our model, the choice probabilities for such a randomly drawn agent are precisely Equations (4.1) and (4.2).

Each of the three inference procedures (VEB, VB, and MCMC) furnishes an estimate of the predictive choice distribution. For VEB, the estimate is Equation (4.1), with $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$ replaced by $\hat{\boldsymbol{\zeta}}_{\text{VEB}}$ and $\hat{\boldsymbol{\Omega}}_{\text{VEB}}$. With VB and MCMC, we obtain a posterior distribution over $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$; we take the mean of Equation (4.1) under this posterior as a point estimate of the predictive choice distribution

$$\hat{p}(\mathbf{y}_{\text{new}}|\mathbf{x}_{\text{new}}, \mathcal{D}) = \int \left[\int p(\mathbf{y}_{\text{new}}|\mathbf{x}_{\text{new}}, \boldsymbol{\beta}) p(\boldsymbol{\beta}|\boldsymbol{\zeta}, \boldsymbol{\Omega})\, d\boldsymbol{\beta} \right] \times p(\boldsymbol{\zeta}, \boldsymbol{\Omega}|\mathcal{D})\, d\boldsymbol{\zeta}\, d\boldsymbol{\Omega}. \tag{4.3}$$

For VB, the posterior density $p(\boldsymbol{\zeta}, \boldsymbol{\Omega}|\mathcal{D})$ in Equation (4.3) is approximated by the fitted variational distribution

$$q(\boldsymbol{\zeta}|\boldsymbol{\mu}_{\zeta}, \boldsymbol{\Sigma}_{\zeta}) q(\boldsymbol{\Omega}|\boldsymbol{\Upsilon}^{-1}, \omega). \tag{4.4}$$

To carry out the integral against the VB posterior distribution of $(\boldsymbol{\zeta}, \boldsymbol{\Omega})$ in Equation (4.3), we use a Monte Carlo approximation based on 500 draws from Equation (4.4). For MCMC, the posterior is approximated by the empirical distribution of 12,500 draws from a Markov chain, as discussed. We need fewer draws from the variational posterior because those draws are independent while the MCMC draws have some autocorrelation.

Notice that to compute the true predictive choice distribution and the three estimates thereof, we must evaluate the integral in Equation (4.1) over $\boldsymbol{\beta} \sim \mathcal{N}_K(\boldsymbol{\zeta}, \boldsymbol{\Omega})$. In all cases, we use simple Monte Carlo to approximate this integral. For the estimated predictive choice distributions, the approximation is based on 10,000 iid draws of $\boldsymbol{\beta}$. [When computing Equation (4.3), that means we use a total of 10,000 × C simulated $\boldsymbol{\beta}$'s, where C is 500 for VB and 12,500 for MCMC.] The Monte Carlo variability resulting from this $\boldsymbol{\beta}$ integration contributes to the overall observed variability in the accuracy of these estimators. This variability is reported in the result tables.

However, when we are computing the true predictive choice distribution, we would like to use so many draws of $\boldsymbol{\beta}$ that there is no noticeable effect on the variability of the accuracy measure. It turns out we can guarantee this by using 1,000,000 draws of $\boldsymbol{\beta}$ each time a true predictive choice distribution is calculated. The explanation is in Appendix E, which is best read after finishing the current Section 4.1.1.

We measure the accuracy, or rather the error, of each inference procedure as the distance between its estimated predictive choice distribution and the true predictive choice distribution. We use the total variation (TV) metric on probability distributions (see, for example, Shorack 2000, definition 19.2.3), leading to what we call the "TV error" of each procedure:

$$\begin{aligned} &\mathrm{TV}[p_{\text{true}}(\mathbf{y}_{\text{new}}|\mathbf{x}_{\text{new}}), \hat{p}(\mathbf{y}_{\text{new}}|\mathbf{x}_{\text{new}})] \\ &\quad = \max_{S \subset \{1,\ldots,J\}} \left| \sum_{j \in S} p_{\text{true}}(\mathbf{y}^j_{\text{new}} = 1|\mathbf{x}_{\text{new}}) \right. \\ &\qquad \left. - \sum_{j \in S} \hat{p}(\mathbf{y}^j_{\text{new}} = 1|\mathbf{x}_{\text{new}}) \right|. \end{aligned} \tag{4.5}$$

[Note that Shorack (2000) includes an additional factor of two in his version, which we omit.] In words, TV error equals the maximum, over all choice-item subsets, of the discrepancy between the probability $\hat{p}(\cdot|\mathbf{x}_{\text{new}})$ assigns to the subset versus the probability $p_{\text{true}}(\cdot|\mathbf{x}_{\text{new}})$ assigns to the subset. To compute TV errors, we use the well-known fact (Shorack 2000, proposition 19.2.2) that

$$\begin{aligned} &\mathrm{TV}[p_{\text{true}}(\mathbf{y}_{\text{new}}|\mathbf{x}_{\text{new}}), \hat{p}(\mathbf{y}_{\text{new}}|\mathbf{x}_{\text{new}})] \\ &\quad = \frac{1}{2} \sum_{j=1}^{J} \left| p_{\text{true}}(\mathbf{y}^j_{\text{new}} = 1|\mathbf{x}_{\text{new}}) - \hat{p}(\mathbf{y}^j_{\text{new}} = 1|\mathbf{x}_{\text{new}}) \right|. \end{aligned} \tag{4.6}$$

We need to choose the attribute matrix $\mathbf{x}_{\text{new}}$ at which the true and estimated predictive choice distributions are calculated. In each replication of each simulation scenario, we computed the TV error of VEB on 25 different random draws of $\mathbf{x}_{\text{new}}$. We then compared the three procedures using the $\mathbf{x}_{\text{new}}$, which yielded the median of the 25 TV errors. In this sense our results are representative of accuracy for a "typical" item attribute matrix.

TV error values for the three inference procedures, averaged across the 10 replications of each scenario, are presented in Table 1 for the 3-item simulation scenarios and in Table 2 for the 12-item scenarios. Standard errors computed from the 10 values

Table 1. Total variation error for simulation study A datasets with three choice items

		3 attributes			10 attributes		
Het.	Agents	VEB	VB	MCMC	VEB	VB	MCMC
Low	250	0.67% (0.14)	0.67% (0.13)	0.85% (0.27)	1.06% (0.24)	1.14% (0.27)	1.41% (0.18)
	1000	0.36% (0.08)	0.31% (0.07)	0.47% (0.10)	0.88% (0.18)	0.90% (0.18)	0.72% (0.18)
	5000	0.16% (0.04)	0.15% (0.04)	0.21% (0.03)	0.32% (0.06)	0.30% (0.04)	0.43% (0.08)
	25,000	0.26% (0.05)	0.20% (0.07)	NA	0.40% (0.07)	0.57% (0.13)	NA
High	250	1.47% (0.24)	1.41% (0.23)	1.23% (0.42)	1.71% (0.25)	2.22% (0.27)	1.63% (0.28)
	1000	1.37% (0.29)	1.30% (0.26)	0.47% (0.09)	1.27% (0.18)	1.75% (0.46)	1.03% (0.19)
	5000	0.69% (0.08)	0.69% (0.09)	0.35% (0.10)	0.92% (0.14)	1.20% (0.37)	0.58% (0.12)
	25,000	0.55% (0.12)	0.56% (0.10)	NA	1.15% (0.11)	0.98% (0.11)	NA

NOTE: Reported values are averages across 10 replications of each simulation scenario, with standard errors (based on the 10 values) given in parentheses. MCMC results are unavailable in the 25,000 agent case because the sampler exhausted memory resources before a sufficient number of samples could be drawn.

are given in parentheses. (The NA entries in the tables are explained in the next subsection.) The main conclusion we draw from Tables 1 and 2 is simple: on these datasets, there are no major differences in accuracy among VEB, VB, and MCMC. In most cases, the TV metric between MCMC and the variational methods is less than two standard errors (SE's); in many cases the difference is no more than one SE.

The scale of the TV error for all the procedures is the same; that scale is larger in the 12-item case than the 3-item case, but all three procedures exhibit high accuracy on all datasets. The errors are low simply because there is no model misspecification: by design, the observations follow the MML model exactly. Even the lowest agent count of 250 provides a good deal of information about $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$, considering that we observe 25 choice events per agent.

Another clear pattern in Tables 1 and 2 is that error decreases as agent count grows. This we expect, since each additional agent's observed choices provide further independent information about $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$ (indirectly, through $\boldsymbol{\beta}_h$). In a few cases, the 25,000-agent error is larger than the 5000-agent error, but not significantly so—we attribute this to chance variation.

Figure 1 shows in a different way the comparable accuracy of the three procedures. When there are three choice items, any predictive choice distribution can be plotted as a point in the triangular simplex. The figure shows the close proximity of each procedure's estimated choice distribution to the true distribution in one simulation scenario. Plots for other three-item scenarios are qualitatively the same.

Figure 1 also shows contours of the two approximate posterior distributions for the predictive choice probability vector. Specifically, we view the predictive choice distribution in Equation (4.1) as a function $f(\boldsymbol{\zeta}, \boldsymbol{\Omega})$ of the random pair $(\boldsymbol{\zeta}, \boldsymbol{\Omega})$. This pair has approximate posterior distribution in Equation (4.4) using VB, and it has an MCMC approximate posterior distribution: the empirical distribution of draws from the chain. Each of these two approximate posterior distributions induces a distribution for $f(\boldsymbol{\zeta}, \boldsymbol{\Omega})$ on the simplex. In each case, we smooth the induced distribution using a two-dimensional kernel density estimate (KDE). The KDE uses an axis-aligned bivariate normal kernel with bandwidth chosen by the "normal reference" asymptotic rule (see `R` function `bandwidth.nrd`). Figure 1 then shows contours at heights 80%, 60%, 40%, and 20% of the mode.

We see that VB is producing not only a posterior mean similar to MCMC, but also a similar posterior density in the neighborhood of the mean. However, the variational posterior is more concentrated around the mode than the MCMC posterior. This underdispersion effect is a familiar and general phenomenon when factorized variational families are employed (see, for example, Bishop 2006, section 10.1.2, p. 467).

4.1.2 Speed. For each scenario in simulation study A, we did timing runs of the three procedures in turn on the same unloaded machine, having a 64-bit dual-core 3.2 GHz Intel Xeon processor with 8 GB of main memory. For the MCMC inference, we used the `rhierMnlRwMixture` function in the R

Table 2. Total variation error for simulation study A data sets with 12 choice items

		3 attributes			10 attributes		
Het.	Agents	VEB	VB	MCMC	VEB	VB	MCMC
Low	250	1.41% (0.20)	1.40% (0.20)	1.45% (0.19)	2.02% (0.41)	1.92% (0.38)	2.57% (0.26)
	1000	0.84% (0.14)	0.81% (0.14)	0.73% (0.12)	1.28% (0.23)	1.29% (0.21)	1.52% (0.21)
	5000	0.53% (0.09)	0.49% (0.07)	0.30% (0.04)	0.78% (0.16)	0.77% (0.14)	0.49% (0.07)
	25,000	0.58% (0.09)	0.54% (0.12)	NA	0.48% (0.07)	0.42% (0.06)	NA
High	250	2.31% (0.39)	2.41% (0.39)	2.19% (0.24)	2.83% (0.42)	2.77% (0.37)	3.97% (0.37)
	1000	1.11% (0.16)	1.07% (0.14)	0.81% (0.08)	2.22% (0.15)	2.26% (0.14)	1.92% (0.20)
	5000	0.93% (0.15)	0.86% (0.14)	0.61% (0.07)	1.41% (0.15)	1.29% (0.12)	0.98% (0.11)
	25,000	0.87% (0.12)	0.74% (0.14)	NA	1.29% (0.12)	1.15% (0.19)	NA

NOTE: Reported values are averages across 10 replications of each simulation scenario, with standard errors (based on the 10 values) given in parentheses. MCMC results are unavailable in the 25,000 agent case because the sampler exhausted memory resources before a sufficient number of samples could be drawn.

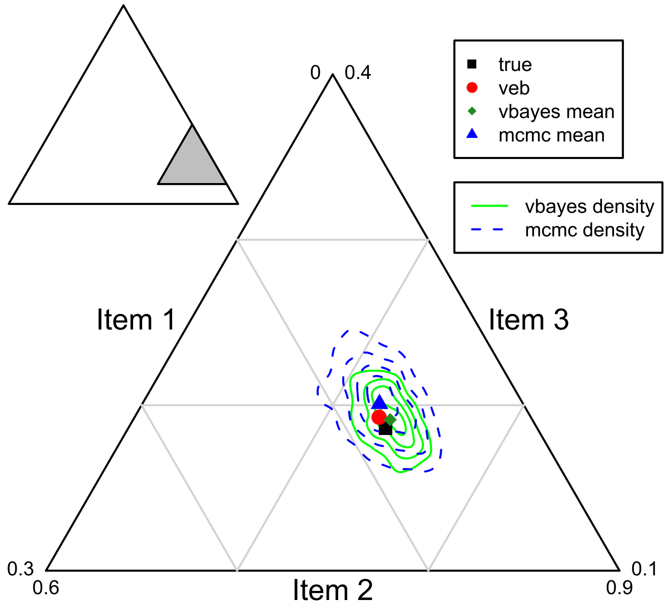

Figure 1. Triangle plot of the true predictive choice distribution and its estimates in the three-item case. Shown are results for the simulation with 250 decision-makers, 10 attributes, and high heterogeneity. Notice that all three estimation methods (variational empirical Bayes, variational hierarchical Bayes, and MCMC) recover the true predictive choice distribution almost exactly. Observe also that the contours of the variational approximate posterior align with the contours of the MCMC approximate posterior. A color version of this figure is available in the electronic version of this article.

package bayesm (Rossi and McCulloch 2007), which has efficient vectorized implementations of the inner sampling routines. This package stores all MCMC draws in memory, however. For our largest datasets, with 25,000 decision makers, the machine's memory was exhausted before MCMC converged. We were able to run MCMC for 1000 iterations in this case. So, for consistency, we timed the MCMC procedure in all cases by running 1000 iterations and using the elapsed time to estimate how long the full 6000 iterations would take. Since the iteration time of MCMC in these simulations is very stable, this approach works well. We implemented the variational algorithms in R, with compiled C components for the numerical optimization routines.

Figure 2 displays time to completion on each dataset for the three procedures, according to the convergence criteria previously described. Within each trellis panel, completion time is plotted as a function of the number of agents for fixed values of the other simulation parameters. Note that the vertical axis shows completion time on a logarithmic scale, to ease comparison of MCMC to the variational methods. All the procedures scale roughly linearly with the number of agents, which leads to the logarithmic curves seen in the figure. The conclusions are the same in all the scenarios we simulated: variational methods complete much more quickly than MCMC and the magnitude of the difference increases with the number of observed agents. MCMC uses two days of computation time for 25,000 agents with 12 choice items, 10 item attributes, and high heterogeneity versus an hour for each of the variational techniques. In the same setting, but with low heterogeneity, MCMC's two-day computation compares with 2 hr for VEB and 6 hr for VB. In other scenarios, variational run times are measured in minutes, as opposed to hours or days for MCMC.

4.2 Stated-Choice Experiment Analysis

We also compared variational inference to MCMC using multinomial choice data from a stated-choice experiment, conducted as part of The Project on Faculty Appointments at the Harvard Graduate School of Education (Trower 2002). In this study, first-year and second-year faculty, as well as graduate students in their last year of study, were asked to state their preferences for hypothetical faculty positions. Each respondent was presented with 16 pairs of jobs and chose among three alternatives for each pair: one of the two positions, or a "neither" option. These positions varied along seven attributes with either two or three levels, for a covariate set of 19 effects-coded indicator variables. There were 1279 respondents to the survey: in our notation, $J = 3$, $K = 19$, $H = 1279$, and $T_h = 16$ for all h.

We ran VEB and VB using the same convergence criteria as in simulation study A. Based on mixing diagnostics, MCMC was run for 10 independent chains of 25,000 post-burn-in iterations, applying an 20 : 1 thinning ratio to yield 12,500 posterior draws as before. Again, convergence diagnostics are available in Appendix F. Since the true predictive choice probabilities are unknown, we instead calculated the TV distance between the Equation (4.3) probabilities produced by VEB on the one hand and MCMC on the other. We did the same calculation a second time, using VB instead of VEB. Thus, we assess agreement between MCMC and each variational method in turn. We computed TV distances using four randomly sampled choice event covariate matrices $\mathbf{x}_{ht}$ from each respondent in the dataset, yielding $4 \times 1{,}279 = 5116$ VEB-to-MCMC distances (and the same number for VB versus MCMC).

Table 3 summarizes the distributions of the two resulting sets of distances. The agreement between MCMC and each variational method is quite reasonable, with average TV distances just over 3%. MCMC appears to agree slightly more closely with VB than VEB, which may be related to the fact that both MCMC and VB are fully Bayesian inference procedures whereas VEB is not. Relative timings for MCMC versus the variational approaches were similar to those reported in the simulation study A scenario with three items, 10 attributes, and 1000 agents (recall that $T_h = 25$ for all h in that study).

4.3 Simulation Study B

We based simulation study B on the data from the stated-choice experiment, as follows. For each of the 1279 respondents, we independently simulated a $\boldsymbol{\beta}_h$ vector according to Equation (2.3), using the empirical Bayes estimates $\hat{\boldsymbol{\zeta}}_{\text{VEB}}$ and $\hat{\boldsymbol{\Omega}}_{\text{VEB}}$ from the analysis of the actual stated-choice data. Then we used Equation (2.2) to simulate 16 choice events per respondent, one for each of the respondent's 16 observed covariate matrices. This results in a dataset having the same agent count, choice event count, and covariate matrices as the real dataset, but with choice event outcomes drawn from the MML model under known parameter values.

We repeated the simulation procedure 10 times independently to obtain 10 such simulated datasets. Then we ran VEB, VB, and MCMC on each. The variational methods used the same convergence criteria as in simulation study A and MCMC

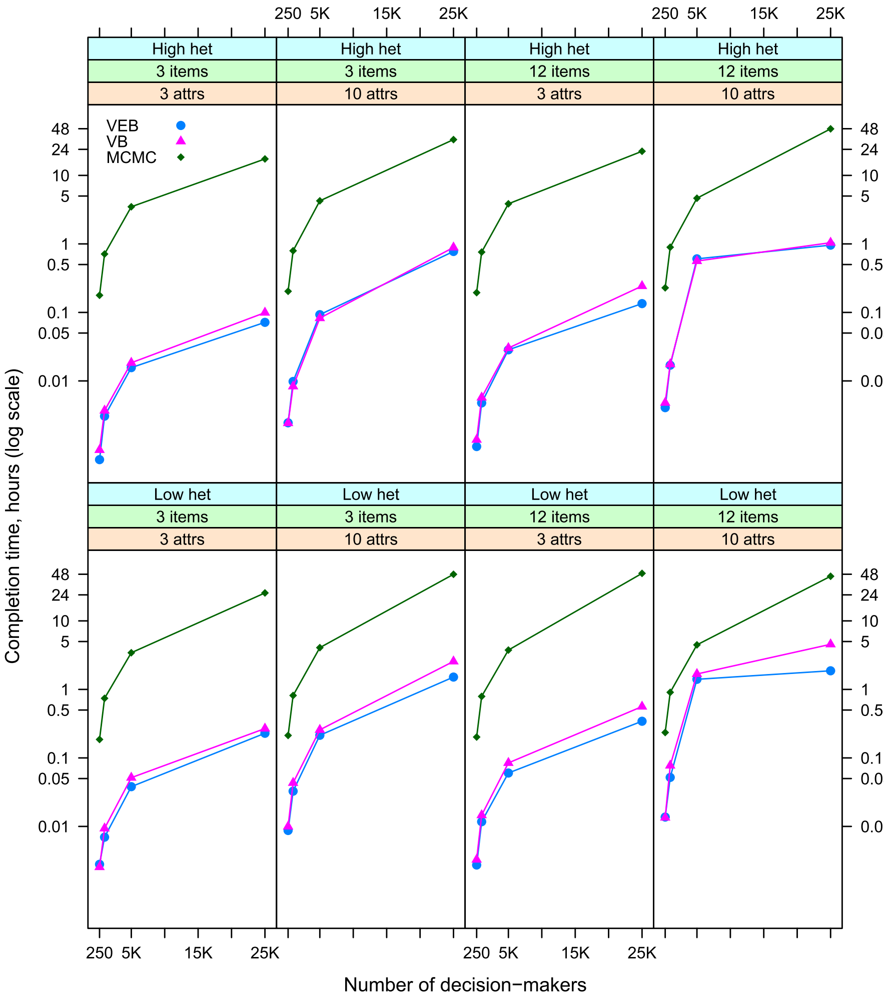

Figure 2. Timing results for Variational Empirical Bayes (VEB), Variational Hierarchical Bayes (VB), and MCMC. Within each panel, completion time is plotted on the log scale as a function of the number of agents, for fixed values of the other simulation parameters (shown at the top of each panel). In all simulated scenarios, variational methods complete more quickly than MCMC. With 25,000 decision-makers, the variational algorithms complete in five minutes to six hours, versus MCMC completion times of one to two days. In the 25,000 agent case, the figure shows the time to generate 6000 MCMC draws, based on a corresponding 1000-draw run (at which point the sampler exhausted memory resources). A color version of this figure is available in the electronic version of this article.

was also run in the same way, but with the number of draws as in the stated-choice experiment. Unlike simulation study A, however, the true population-level covariance matrix $\boldsymbol{\Omega}$ is not diagonal. So the components of $\boldsymbol{\beta}_h$ are correlated and some of the covariate matrices $\mathbf{x}_{ht}$ exhibit multicollinearity.

Table 4 shows average TV error across the 10 datasets for each method, accompanied by standard errors. As in simulation study A, all three methods produced very accurate predictive choice probabilities. In this case, MCMC had a larger average error than the variational approaches. But since the difference amounts to only about 1.5 SE's and the chains satisfied standard convergence diagnostics, it seems plausible to attribute the difference to chance.

5. DISCUSSION

Variational methods allow estimation of hierarchical discrete choice models in a small fraction of the time required for MCMC. They open Bayesian inference to a wide class of problems for which resource constraints make MCMC intractable, such as the MML model with many heterogeneous units. For now, variational methods appear to be the only viable option in these cases.

Table 3. Total variation distances, variational methods vs. MCMC, for the stated-choice experiment data

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	N
VEB vs. MCMC	0.05%	1.95%	3.04%	3.23%	4.39%	8.71%	5116
VB vs. MCMC	0.04%	1.81%	2.82%	3.12%	4.19%	7.19%	5116

NOTE: Each row summarizes the distribution, across respondents, of the TV distance between (1) MCMC predictive choice probabilities and (2) the named variational method's predictive choice probabilities. Of the 16 observed covariate matrices per respondent, a random sample of four was used, so that each respondent contributed four TV distances.

Table 4. Total variation error for simulation study B data sets

	VEB	VB	MCMC
Mean	1.65%	1.53%	2.76%
Std. err.	0.24%	0.25%	0.49%

NOTE: Shown are the average TV errors of the three methods across 10 independently simulated data sets, together with standard errors.

Of course, one can use variational methods to estimate many more types of models than the MML model examined here. Within the MML family, it would be straightforward to add a utility scaling parameter, or allow the heterogeneous coefficients themselves to depend on observed covariates. The value of the variational approach is greatest when subsampling of data is ill-advised. For example, consider a linear model with heterogeneous coefficients on exogenous and endogenous covariates, where the available instrumental variables only weakly explain the endogenous part. To draw inferences about the covariances of the heterogeneous parameters, we may need a large amount of data to achieve reasonable power in hypothesis testing. MCMC is untenable here, but variational methods have promise. When factorized variational distributions are inadequate, alternatives such as mixtures of normals or Dirichlet processes (Blei and Jordan 2006) can be applied.

We emphasize that we do not advocate abandoning MCMC in favor of variational methods. On the contrary, we suggest using MCMC when possible. MCMC offers consistency guarantees with no analog in variational methods and it has withstood decades of scrutiny. But we advise against subsampling the data (i.e., throwing out information) or discarding key modeling elements simply to make the problem fit within the time and resource constraints of MCMC. The possibility of applying variational methods to previously intractable problems makes them an important addition to the statistician's toolkit.

APPENDIX A: VARIATIONAL INFERENCE AND PARAMETER ESTIMATION

In this appendix we describe the variational inference and estimation procedures for the mixed multinomial logit model. A few words on notation: $\mathbf{A} \succ 0$ means the matrix $\mathbf{A}$ is positive definite; $\mathbf{A} \succeq 0$ means positive semidefinite; $|\mathbf{A}|$ is the determinant of $\mathbf{A}$. For a scalar function $s : \mathbb{R} \to \mathbb{R}$ and a vector $\mathbf{v} \in \mathbb{R}^n$, $s(\mathbf{v})$ means the $n \times 1$ vector $(s(v_1), \ldots, s(v_n))^\top$.

A.1 The Empirical Bayes ELBO

Recall that the variational distribution $q(\boldsymbol{\beta}_{1:H}|\boldsymbol{\mu}_{1:H}, \boldsymbol{\Sigma}_{1:H})$ is a product of normal distributions $\mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$. For empirical Bayes in the MML model, the ELBO objective function in Equation (3.6) becomes

$$H(q) + \sum_{h=1}^{H} \mathbb{E}_q \log p(\boldsymbol{\beta}_h|\boldsymbol{\zeta}, \boldsymbol{\Omega}) + \sum_{h=1}^{H}\sum_{t=1}^{T_h} \mathbb{E}_q \log p(\mathbf{y}_{ht}|\mathbf{x}_{ht}, \boldsymbol{\beta}_h). \quad (A.1)$$

The first term in Equation (A.1) is the Shannon entropy of the variational distribution. The second and third terms are (minus) an unnormalized cross entropy—the missing normalization constant is the marginal likelihood.

The first and second terms of Equation (A.1) are straightforward to derive; the third term requires more attention. Using the multinomial logit mass function

$$p(\mathbf{y}_{ht}|\mathbf{x}_{ht}, \boldsymbol{\beta}_h) = \prod_{j=1}^{J} \left[\frac{\exp(\mathbf{x}_{htj}^\top \boldsymbol{\beta}_h)}{\sum_{j'} \exp(\mathbf{x}_{htj'}^\top \boldsymbol{\beta}_h)} \right]^{y_{ht}^j}, \quad (A.2)$$

the third term becomes

$$\sum_{h=1}^{H}\sum_{t=1}^{T_h} \left[\sum_{j=1}^{J} y_{ht}^j (\mathbf{x}_{htj}^\top \boldsymbol{\mu}_h) - \mathbb{E}_{q_\lambda} \log\left(\sum_{j=1}^{J} \exp(\mathbf{x}_{htj}^\top \boldsymbol{\beta}_h) \right) \right]. \quad (A.3)$$

The expected log-sum-exp in Equation (A.3) has no closed form. Therefore, for variational inference, we approximate the ELBO objective function $\mathcal{L}$ using a new objective function $\tilde{\mathcal{L}}$. To construct $\tilde{\mathcal{L}}$, we consider two alternatives: the zeroth-order and first-order delta method for moments (Bickel and Doksum 2007), which we call D0 and D1, respectively. D0 is equivalent to applying Jensen's inequality to the expected log-sum-exp, resulting in the following lower bound to Equation (A.3):

$$[\text{D0}] \quad \sum_{h=1}^{H}\sum_{t=1}^{T_h} \left[\sum_{j=1}^{J} y_{ht}^j (\mathbf{x}_{htj}^\top \boldsymbol{\mu}_h) - \log\left(\sum_{j=1}^{J} \exp(\mathbf{x}_{htj}^\top \boldsymbol{\mu}_h + (1/2)\mathbf{x}_{htj}^\top \boldsymbol{\Sigma}_h \mathbf{x}_{htj}) \right) \right]. \quad (A.4)$$

Here we use the usual formula for the mean of a lognormal random variable. In a different context, Blei and Lafferty (2007) consider an approximation equivalent to D0, but expressed using a redundant variational parameter.

For approximation D1, we restrict $\boldsymbol{\Sigma}_h \succ 0$ to be diagonal, and define $\boldsymbol{\sigma}_h := \log(\mathrm{diag}\{\boldsymbol{\Sigma}_h\}) \in \mathbb{R}^K$. Using results in Appendix B, we obtain the following approximation to Equation (A.3):

$$[\text{D1}] \quad \sum_{h=1}^{H}\sum_{t=1}^{T_h} \left[\sum_{j=1}^{J} y_{ht}^j (\mathbf{x}_{htj}^\top \boldsymbol{\mu}_h) - \log\left(\sum_{j=1}^{J} \exp(\mathbf{x}_{htj}^\top \boldsymbol{\mu}_h) \right) - \frac{1}{2} \exp(\boldsymbol{\sigma}_h)^\top \Theta(\boldsymbol{\mu}_h) \right], \quad (A.5)$$

with $\Theta(\boldsymbol{\mu}_h) \in \mathbb{R}^K$ as defined in Appendix B. Notice that, unlike D0, approximation D1 does not preserve the guarantee that the optimal value of the variational optimization is a lower bound on the marginal likelihood. However, in our simulations, using D1 resulted in more accurate variational approximations to the posterior. Thus, all of our reported empirical results on variational methods employ D1.

In this appendix we give a derivation based on approximation D0. The derivation for D1 is similar, but simpler because $\boldsymbol{\Sigma}_h$ is treated as diagonal. Under D0, the final empirical Bayes objective function is

$$\begin{aligned}
\tilde{\mathcal{L}}&(\boldsymbol{\mu}_{1:H}, \boldsymbol{\Sigma}_{1:H}; \boldsymbol{\zeta}, \boldsymbol{\Omega}) \\
&= \frac{1}{2}\sum_{h=1}^{H} \log[(2\pi e)^K |\boldsymbol{\Sigma}_h|] - \frac{H}{2}\log((2\pi)^K |\boldsymbol{\Omega}|) \\
&\quad - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}^{-1} \sum_{h=1}^{H} \{\boldsymbol{\Sigma}_h + (\boldsymbol{\mu}_h - \boldsymbol{\zeta})(\boldsymbol{\mu}_h - \boldsymbol{\zeta})^\top\} \right] \\
&\quad + \sum_{h=1}^{H}\sum_{t=1}^{T_h} \left[\sum_{j=1}^{J} y_{ht}^j (\mathbf{x}_{htj}^\top \boldsymbol{\mu}_h) - \log\left(\sum_{j=1}^{J} \exp(\mathbf{x}_{htj}^\top \boldsymbol{\mu}_h + (1/2)\mathbf{x}_{htj}^\top \boldsymbol{\Sigma}_h \mathbf{x}_{htj}) \right) \right]. \quad (A.6)
\end{aligned}$$

The first line in Equation (A.6) uses the well-known entropy of the normal distribution. The second line uses the cross-entropy of two normal distributions, also well known. The third line is approximation D0.

A.2 Empirical Bayes Variational E-Step

Here we describe a block coordinate ascent algorithm to maximize Equation (A.6) over the variational parameters $\boldsymbol{\mu}_{1:H}$ and $\boldsymbol{\Sigma}_{1:H}$. Although the problem is not jointly convex in all these parameters, each $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ coordinate update solves a smooth, unconstrained convex optimization problem. The requirement $\boldsymbol{\Sigma}_h \succeq 0$ is satisfied after each update. We initialize the variational parameters at the MLE's from a homogeneous model (in which all agents share a common $\boldsymbol{\beta}$ value).

The concavity of Equation (A.6) in $\boldsymbol{\mu}_h$ follows from the fact that $\boldsymbol{\Omega} \succ 0$ and from the convexity of the log-sum-exp function. We update $\boldsymbol{\mu}_h$ using standard algorithms for unconstrained convex optimization (Boyd and Vandenberghe 2004), supplying an analytic gradient and Hessian as follows. Define the function $\mathbf{w}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{x})$ taking values in $\mathbb{R}^J$, with jth component $\exp(\mathbf{x}_j^\top \boldsymbol{\mu} + (1/2)\mathbf{x}_j^\top \boldsymbol{\Sigma} \mathbf{x}_j)$, and normalized to sum to one across j. The gradient of $\tilde{\mathcal{L}}$ with respect to $\boldsymbol{\mu}_h$ can then be written

$$\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\mu}_h} = -\boldsymbol{\Omega}^{-1}(\boldsymbol{\mu}_h - \boldsymbol{\zeta}) + \sum_{t=1}^{T_h} \sum_{j=1}^{J} [y_{ht}^j - w^j(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h, \mathbf{x}_{ht})]\mathbf{x}_{htj}. \quad (A.7)$$

Note the similarity of this gradient to the gradient from an L_2-regularized multiple logistic regression: it consists of a contribution from the regularizer (the left-hand term), plus a residual-weighted sum of covariate vectors. Abbreviating $\mathbf{w}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h, \mathbf{x}_{ht})$ to $\mathbf{w}_{ht}$, an argument using matrix differentials (Magnus and Neudecker 2007) gives the Hessian

$$\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\mu}_h \, \partial \boldsymbol{\mu}_h^\top}$$

$$= -\boldsymbol{\Omega}^{-1} - \sum_{t=1}^{T_h} [\mathbf{x}_{ht}^\top \operatorname{diag}\{\mathbf{w}_{ht}\}\mathbf{x}_{ht} - (\mathbf{x}_{ht}^\top \mathbf{w}_{ht})(\mathbf{x}_{ht}^\top \mathbf{w}_{ht})^\top]. \quad (A.8)$$

The $\boldsymbol{\Sigma}_h$ coordinate update is harder because we need to insure that $\boldsymbol{\Sigma}_h \succeq 0$. Using a reformulation, we can avoid making the constraint explicit, which would complicate the optimization. Let $\boldsymbol{\Sigma}_h = \mathbf{L}_h \mathbf{L}_h^\top$ for a lower-triangular matrix $\mathbf{L}_h$. Since $\boldsymbol{\Sigma}_h \succeq 0$, one such $\mathbf{L}_h$ always exists—the Cholesky factor. We replace each $\boldsymbol{\Sigma}_h$ in $\tilde{\mathcal{L}}$ with $\boldsymbol{\Sigma}_h(\mathbf{L}_h) := \mathbf{L}_h \mathbf{L}_h^\top$, and optimize over the unconstrained set of lower-triangular matrices $\mathbf{L}_h$.

The objective function in Equation (A.6) remains concave in $\mathbf{L}_h$. To see this, compare the terms depending on $\boldsymbol{\Sigma}_h = \mathbf{L}_h \mathbf{L}_h^\top$ to the function studied in Appendix C. We now give the gradient with respect to $\mathbf{L}_h$. Standard matrix differentiation of Equation (A.6) leads to the $\boldsymbol{\Sigma}_h$ gradient

$$\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\Sigma}_h} = \frac{1}{2}\left[\boldsymbol{\Sigma}_h^{-1} - \boldsymbol{\Omega}^{-1} - \sum_{t=1}^{T_h} \mathbf{x}_{ht}^\top \operatorname{diag}\{\mathbf{w}_{ht}\}\mathbf{x}_{ht}\right]. \quad (A.9)$$

Again using matrix differentials and the Cauchy invariance rule, it is not hard to show that the gradient with respect to $\mathbf{L}_h$ is

$$\frac{\partial \tilde{\mathcal{L}}}{\partial \mathbf{L}_h} = 2\left(\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\Sigma}_h}\right)\mathbf{L}_h$$

$$= \mathbf{L}_h^{-\top} - \left(\boldsymbol{\Omega}^{-1} + \sum_{t=1}^{T_h} \mathbf{x}_{ht}^\top \operatorname{diag}\{\mathbf{w}_{ht}\}\mathbf{x}_{ht}\right)\mathbf{L}_h. \quad (A.10)$$

Note that this is the gradient with respect to a dense matrix $\mathbf{L}_h$. Since we optimize over lower-triangular matrices, i.e., $\operatorname{vech}(\mathbf{L}_h)$, we need only use the lower triangle of the gradient. This is convenient for the term $\mathbf{L}_h^{-\top}$: it is upper-triangular, so its lower triangle is a diagonal matrix. Furthermore, from a standard result of linear algebra, the diagonal entries are simply $1/\ell_{ii}$, where the ℓ_{ii}'s form the diagonal of $\mathbf{L}_h$.

In practice, for a given h, we do the $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ updates in a single step by optimizing jointly over $\boldsymbol{\mu}_h$ and $\mathbf{L}_h$, which remains a convex problem.

A.3 Empirical Bayes M-Step

In the M-step, we maximize Equation (A.6) over $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$. Identifying the terms which depend on $\boldsymbol{\zeta}$, we recognize the usual Gaussian mean estimation problem. Further, Equation (A.6) is easily seen to be concave in $\boldsymbol{\Omega}^{-1}$, with a closed-form solution of the corresponding first-order condition. We obtain the M-step updates

$$\hat{\boldsymbol{\zeta}} \leftarrow \frac{1}{H}\sum_{h=1}^{H} \boldsymbol{\mu}_h, \qquad \hat{\boldsymbol{\Omega}} \leftarrow \frac{1}{H}\sum_{h=1}^{H} \boldsymbol{\Sigma}_h + \widehat{\operatorname{Cov}}(\boldsymbol{\mu}_\cdot). \quad (A.11)$$

Here $\widehat{\operatorname{Cov}}(\boldsymbol{\mu}_\cdot)$ is the empirical covariance of the $\boldsymbol{\mu}_h$ vectors.

A.4 Variational Hierarchical Bayes

In the fully Bayesian MML model, $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$ have prior distributions, with corresponding variational factors given in Equation (3.10). The ELBO in this case has the same form as Equation (A.1), with two differences. First, $H(q)$ contains two new terms

$$H(q(\boldsymbol{\zeta}|\boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta)) + H(q(\boldsymbol{\Omega}|\boldsymbol{\Upsilon}^{-1}, \omega)). \quad (A.12)$$

Second, there are two new cross-entropy terms

$$\mathbb{E}_q \log p(\boldsymbol{\zeta}|\boldsymbol{\beta}_0, \boldsymbol{\Omega}_0) + \mathbb{E}_q \log p(\boldsymbol{\Omega}|\mathbf{S}, \nu). \quad (A.13)$$

Also, the middle term of Equation (A.1) changes in the fully Bayesian case because $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$ are now averaged over rather than treated as constants.

We first write the two new entropy terms explicitly. Let

$$D(\omega, \boldsymbol{\Upsilon}) := \log(2^K|\boldsymbol{\Upsilon}|) + \sum_{i=1}^{K} \Psi\left(\frac{\omega + 1 - i}{2}\right), \quad (A.14)$$

where Ψ is the digamma function. $D(\omega, \boldsymbol{\Upsilon})$ is the expected log determinant of a Wishart random matrix (see, for example, Beal 2003, appendix A). Also let

$$A_\omega(\boldsymbol{\Upsilon}) := \log\left[2^{\omega K/2}\pi^{K(K-1)/4}\prod_{i=1}^{K}\Gamma\left(\frac{\omega+1-i}{2}\right)\right] + \frac{\omega}{2}\log|\boldsymbol{\Upsilon}|, \quad (A.15)$$

which is the log normalization constant of the Wishart distribution (Beal 2003, appendix A). Using known formulas for normal and Wishart entropies (Beal 2003, appendix A), the two new entropy terms are seen to equal

$$\frac{1}{2}\log[(2\pi e)^K|\boldsymbol{\Sigma}_\zeta|] - \frac{\omega - K - 1}{2}D(\omega, \boldsymbol{\Upsilon}) + \frac{\omega K}{2} + A_\omega(\boldsymbol{\Upsilon}). \quad (A.16)$$

The new cross entropy terms for $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$ work out to

$$-\tfrac{1}{2}\big\{\log[(2\pi)^K|\boldsymbol{\Omega}_0|]$$

$$+ \operatorname{tr}\big(\boldsymbol{\Omega}_0^{-1}[\boldsymbol{\Sigma}_\zeta + (\boldsymbol{\mu}_\zeta - \boldsymbol{\beta}_0)(\boldsymbol{\mu}_\zeta - \boldsymbol{\beta}_0)^\top]\big)\big\} \quad (A.17)$$

and

$$-A_\nu(\mathbf{S}^{-1}) + \frac{\nu - K - 1}{2}D(\omega, \boldsymbol{\Upsilon}) - \frac{\omega}{2}\operatorname{tr}(\mathbf{S}^{-1}\boldsymbol{\Upsilon}), \quad (A.18)$$

respectively. The middle term of Equation (A.1) eventually becomes

$$-\frac{H}{2}\{K\log(2\pi) - D(\omega, \boldsymbol{\Upsilon})\} - \frac{\omega}{2}\operatorname{tr}\Bigg[\boldsymbol{\Upsilon}\Bigg(H\boldsymbol{\Sigma}_\zeta$$

$$+ \sum_{h=1}^{H}(\boldsymbol{\Sigma}_h + (\boldsymbol{\mu}_\zeta - \boldsymbol{\mu}_h)(\boldsymbol{\mu}_\zeta - \boldsymbol{\mu}_h)^\top)\Bigg)\Bigg]. \quad (A.19)$$

With these changes, it is not hard to see that $\tilde{\mathcal{L}}$ is concave separately in $\boldsymbol{\mu}_\zeta$ and $\boldsymbol{\Sigma}_\zeta$. The first-order conditions for block coordinate ascent lead to the updates

$$\boldsymbol{\mu}_\zeta \leftarrow (\boldsymbol{\Omega}_0^{-1} + H\omega\boldsymbol{\Upsilon})^{-1}\left(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_0 + \omega\boldsymbol{\Upsilon}\sum_{h=1}^{H}\boldsymbol{\mu}_h\right), \tag{A.20}$$

$$\boldsymbol{\Sigma}_\zeta \leftarrow (\boldsymbol{\Omega}_0^{-1} + H\omega\boldsymbol{\Upsilon})^{-1}. \tag{A.21}$$

By inspection, $\boldsymbol{\Sigma}_\zeta \succeq 0$, so this constraint need not be explicitly enforced. Note the similarity to conjugate posterior updating: on the precision scale, $\boldsymbol{\Sigma}_\zeta$ is the sum of the prior precision matrix $\boldsymbol{\Omega}_0^{-1}$ and H copies of the variational posterior mean $\omega\boldsymbol{\Upsilon}$ for $\boldsymbol{\Omega}^{-1}$. Similarly, $\boldsymbol{\mu}_\zeta$ is a precision-weighted convex combination of the prior vector $\boldsymbol{\beta}_0$ and the empirical average of the variational posterior means $\boldsymbol{\mu}_{1:H}$ for $\boldsymbol{\beta}_{1:H}$.

The updates for $\boldsymbol{\Upsilon}$ and ω are similarly straightforward to derive; we obtain

$$\omega \leftarrow \nu + H, \tag{A.22}$$

$$\boldsymbol{\Upsilon} \leftarrow \left(\mathbf{S}^{-1} + \sum_{h=1}^{H}(\boldsymbol{\Sigma}_h + (\boldsymbol{\mu}_\zeta - \boldsymbol{\mu}_h)(\boldsymbol{\mu}_\zeta - \boldsymbol{\mu}_h)^\top) + H\boldsymbol{\Sigma}_\zeta\right)^{-1}. \tag{A.23}$$

Notice that the solution in Equation (A.22) for ω involves only the constants ν and H. We compute ω once in advance, leaving it unchanged during the variational optimization.

APPENDIX B: AN APPLICATION OF THE DELTA METHOD

Let $f(\mathbf{v})$ be a function from $\mathbb{R}^K$ to $\mathbb{R}$. According to the multivariate delta method for moments (Bickel and Doksum 2007),

$$\mathbb{E}f(\mathbf{V}) \approx f(\mathbb{E}\mathbf{V}) + \frac{1}{2}\operatorname{tr}\left[\left(\frac{\partial f(\mathbb{E}\mathbf{V})}{\partial\mathbf{v}\,\partial\mathbf{v}^\top}\right)\operatorname{Cov}(\mathbf{V})\right], \tag{B.1}$$

provided (i) f has continuous partial derivatives up to order three; (ii) all the order-three partials are bounded in $\mathbf{v}$; and (iii) $\mathbb{E}|V_k|^3 < \infty$, $k = 1, \dots, K$. Consider the case

$$f(\mathbf{v}) = \log(1^\top \exp(\mathbf{x}\mathbf{v})), \tag{B.2}$$

where $\mathbf{x}$ is a $J \times K$ matrix whose rows are the vectors $\mathbf{x}_j^\top$. A routine calculation shows that conditions (i) and (ii) are met. Let $\mathbf{V} \sim \mathcal{N}_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, which clearly satisfies (iii). Restrict $\boldsymbol{\Sigma}$ such that $\boldsymbol{\Sigma} = \operatorname{diag}\{\exp(\boldsymbol{\sigma})\}$ for $\boldsymbol{\sigma} \in \mathbb{R}^K$. We can now rewrite Equation (B.1):

$$\mathbb{E}\log(1^\top\exp(\mathbf{x}\mathbf{V})) \approx \log(1^\top\exp(\mathbf{x}\boldsymbol{\mu})) + \tfrac{1}{2}\Theta(\boldsymbol{\mu})^\top\exp(\boldsymbol{\sigma}), \tag{B.3}$$

where $\Theta(\boldsymbol{\mu})$ is the diagonal of the Hessian of f, evaluated at the point $\boldsymbol{\mu}$. Define $s := 1^\top\exp(\mathbf{x}\boldsymbol{\mu})$. Using matrix differentials, it can be shown that

$$\Theta(\boldsymbol{\mu}) = s^{-1}(\mathbf{x}\odot\mathbf{x})^\top\exp(\mathbf{x}\boldsymbol{\mu}) - s^{-2}(\mathbf{x}^\top\exp(\mathbf{x}\boldsymbol{\mu}))\odot(\mathbf{x}^\top\exp(\mathbf{x}\boldsymbol{\mu})), \tag{B.4}$$

where $\odot$ denotes the Hadamard product.

To use the approximation in Equation (B.3) in an optimization over $\boldsymbol{\mu}$, we need to compute the gradient. The formula for $\Theta(\boldsymbol{\mu})$ makes this a more extensive, but still mechanical exercise in differentials. One obtains

$$\begin{aligned}\frac{\partial\Theta(\boldsymbol{\mu})}{\partial\boldsymbol{\mu}} &= s^{-1}\mathbf{x}^\top e^{\mathbf{x}\boldsymbol{\mu}} \\ &\quad + \frac{1}{2}\mathbf{x}^\top\big[(s^{-1}\operatorname{diag}\{e^{\mathbf{x}\boldsymbol{\mu}}\} - s^{-2}e^{\mathbf{x}\boldsymbol{\mu}}(e^{\mathbf{x}\boldsymbol{\mu}})^\top)(\mathbf{x}\odot\mathbf{x}) \\ &\quad + 2\big(s^{-3}e^{\mathbf{x}\boldsymbol{\mu}}\{(\mathbf{x}^\top e^{\mathbf{x}\boldsymbol{\mu}})\odot(\mathbf{x}^\top e^{\mathbf{x}\boldsymbol{\mu}})\}^\top \\ &\quad - s^{-2}\operatorname{diag}\{e^{\mathbf{x}\boldsymbol{\mu}}\}\mathbf{x}\operatorname{diag}\{\mathbf{x}^\top e^{\mathbf{x}\boldsymbol{\mu}}\}\big)\big]\exp(\boldsymbol{\sigma}).\end{aligned} \tag{B.5}$$

SUPPLEMENTAL MATERIALS

Appendices: Appendices C through F, as referenced in the text. (Braun_McAuliffe_Supplement.pdf)

[Received January 2008. Revised October 2009.]

REFERENCES

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679. [324]

Allenby, G. M., and Lenk, P. J. (1994), "Modeling Household Purchase Behavior With Logistic Normal Regression," *Journal of the American Statistical Association*, 89, 1218–1231. [324]

Allenby, G. M., and Rossi, P. E. (1999), "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89, 57–78. [324]

Beal, M. J. (2003), "Variational Algorithms for Approximate Bayesian Inference," Ph.D. thesis, University College London. [333]

Ben-Akiva, M., and Lerman, S. (1985), *Discrete Choice Analysis: Theory and Application to Travel Demand*, Cambridge, MA: MIT Press. [324]

Bickel, P. J., and Doksum, K. A. (2007), *Mathematical Statistics: Basic Ideas and Selected Topics* (2nd ed.), Vol. 1, Upper Saddle River, NJ: Pearson Prentice Hall. [332,334]

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, New York: Springer. [329]

Blei, D., and Jordan, M. I. (2006), "Variational Inference for Dirichlet Process Mixtures," *Bayesian Analysis*, 1, 121–144. [332]

Blei, D., and Lafferty, J. D. (2007), "A Correlated Topic Model of Science," *The Annals of Applied Statistics*, 1, 17–35. [332]

Boyd, S., and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge, U.K.: Cambridge University Press. [333]

Fader, P. S., and Hardie, B. G. S. (1996), "Modeling Consumer Choice Among SKUs," *Journal of Marketing Research*, 33, 442–452. [324]

Gelman, A., and Rubin, D. B. (1992), "Inference From Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, 457–472. [328]

Guadagni, P. M., and Little, J. D. C. (1983), "A Logit Model of Brand Choice Calibrated on Scanner Data," *Marketing Science*, 2, 203–238. [324]

Hall, J., Kenny, P., King, M., Louviere, J., Viney, R., and Yeoh, A. (2002), "Using Stated Preference Discrete Choice Modelling to Evaluate the Introduction of Varicella Vaccination," *Health Economics*, 11, 457–465. [324]

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. (1999), "An Introduction to Variational Methods for Graphical Models," *Machine Learning*, 37, 183–233. [324]

Magnus, J. R., and Neudecker, H. (2007), *Matrix Differential Calculus With Applications in Statistics and Econometrics* (3rd ed.), New York: Wiley. [333]

McFadden, D. L. (1974), "The Measurement of Urban Travel Demand," *Journal of Public Economics*, 3, 303–328. [324,325]

Moore, W. L., Louviere, J., and Verma, R. (1999), "Using Conjoint Analysis to Help Design Product Platforms," *Journal of Product Innovation Management*, 16, 27–39. [324]

Opper, M., and Saad, D. (eds.) (2001), *Advanced Mean Field Methods: Theory and Practice*, Cambridge, MA: MIT Press. [327]

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006), "CODA: Convergence Diagnosis and Output Analysis for MCMC," *R News*, 6, 7–11. [328]

Raftery, A. E., and Lewis, S. (1992), "How Many Iterations in the Gibbs Sampler?" in *Proceedings of the Fifth Valencia International Conference on Bayesian Statistics*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 763–774. [328]

Revelt, D., and Train, K. E. (1998), "Mixed Logit With Repeated Choices: Households' Choices of Appliance Efficiency Level," *Review of Economics and Statistics*, 80, 647–657. [324]

Robbins, H. (1955), "An Empirical Bayes Approach to Statistics," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA: pp. 131–148. [324]

Rossi, P., and McCulloch, R. (2007), "bayesm: Bayesian Inference for Marketing/Micro-Econometrics," R package version 2.1-3. [330]

Rossi, P. E., and Allenby, G. M. (2003), "Bayesian Statistics and Marketing," *Marketing Science*, 22, 304–328. [327]

Shorack, G. R. (2000), *Probability for Statisticians*, New York: Springer. [328]

Theil, H. (1969), "A Multinomial Extension of the Linear Logit Model," *International Economic Review*, 10, 251–259. [324]

Train, K. E. (2003), *Discrete Choice Methods With Simulation*, Cambridge: Cambridge University Press. [324,325]

Train, K. E., McFadden, D. L., and Ben-Akiva, M. (1987), "The Demand for Local Telephone Service: A Fully Discrete Model of Residential Calling Patterns and Service Choices," *The RAND Journal of Economics*, 18, 109–123. [324]

Trower, C. A. (2002), “Can Colleges Competitively Recruit Faculty Without the Prospect of Tenure?” in *The Questions of Tenure*, ed. R. P. Chait, Cambridge, MA: Harvard University Press, pp. 182–220. [330]

Uhler, R. S., and Cragg, J. G. (1971), “The Structure of Asset Portfolios of Households,” *Review of Economic Studies*, 38, 341–357. [324]

Wainwright, M. J., and Jordan, M. I. (2003), “Graphical Models, Exponential Families, and Variational Inference,” Technical Report 649, UC Berkeley, Dept. Statistics. [324]

Zanutto, E. L., and Bradlow, E. T. (2006), “Data Pruning in Consumer Choice Models,” *Quantitative Marketing and Economics*, 4, 267–287. [325]