Variational Bayesian Inference for Big Data Marketing Models[1]

Asim Ansari Yang Li Jonathan Z. Zhang [2]

August 2014

[1]This is a preliminary version. Please do not cite or circulate.

[2]Asim Ansari is the William T. Dillard Professor of Marketing at Columbia Business School, Yang Li is Assistant Professor of Marketing at Cheung Kong Graduate School of Business in Beijing, and Jonathan Z. Zhang is Assistant Professor of Marketing at University of Washington in Seattle.

Abstract

Bayesian inference plays a central role in empirical marketing. The Bayesian revolution in marketing has resulted in several advances in choice models, targeting, product design, recommendation systems and other areas. Hierarchical Bayesian approaches are popular because they readily yield individual-level parameter estimates that can be used for targeting and personalization decisions. Over the part twenty years, Markov chain Monte Carlo (MCMC) methods have been the methods of choice for estimating hierarchical Bayesian models as they are capable of providing accurate individual-level estimates across a wide range of statistical models. However, MCMC becomes computationally prohibitive and does not scale well when applied to massive data sets that have become common in the current era of "Big Data".

We introduce to the marketing literature a new class of Bayesian estimation techniques known as variational Bayesian (VB) inference. These methods can help marketers in tackling the scalability challenge of big data sets as they use a deterministic optimization approach to approximate the posterior distribution, usually at a fraction of the computational cost associated with simulation-based MCMC methods. We exploit and extend recent developments in variational Bayesian inference and highlight how two VB estimation approaches – Mean-field VB (that is analogous to Gibbs sampling) for conjugate models and Fixed-form VB (which is analogous to Metropolis-Hasting) for nonconjugate models – can be effectively combined for estimating complex marketing models. We also show how recent advances in parallel computing and in stochastic approximation can be used to further enhance the speed of these VB methods. We apply the VB approaches to three commonly used marketing models (e.g. linear model, logit, and ordered logit) to demonstrate how the VB inference is widely applicable for marketing problems.

Keywords: Big Data, Variational Inference, Mean-field Variational Bayes, Fixed-form Variational Bayes, Parallelization, Stochastic Optimization, Adaptive Minibatch.

1 Introduction

The recent advances in information technology, coupled with the rapid decrease in data storage cost, have resulted in the explosive growth in the collection and availability of customer level data. Just in the past decade, we have pushed the limit of data storage from terabytes to petabytes and then to exabytes, and we will soon be staring at the possibility of the zettabyte era. Led by Internet, e-commerce, social media, geographical positioning systems and information-sensing mobile devices, consumers leave a massive trail of data-points, every day, every hour, or even every minute. This has resulted in massive quantities of rich information produced by and about people, things and their interactions, collected at a pace faster than ever before (Boyd and Crawford 2012). As the volume of business data worldwide, across all companies, doubles every 1.2 years (eMarketer 2013), data set sizes in empirical marketing research have also increased tremendously over the years. For example, within the past decade or so, in marketing papers using scanner data, we have seen observations increase from the thousands (Villas-Boas and Winer 1999) to hundreds of thousands (Gordon, Goldfarb and Li 2013).

These novel phenomena, collectively known as "Big Data", provide scholars with exciting opportunities to understand, to explain, and to predict the behavior of consumers in more detailed ways than before. Effective analyses of such large customer databases can yield insights on individual preferences and can help firms increase profit via better targeting, such as in the cases of Harrah's Entertainment, Capital One, and Netflix (Davenport and Harris 2007). Compared to traditional decision-making techniques that rely on managerial experiences and heuristics, data-driven firms can enjoy increased productivity, increased profitability, and an increased market value (Brynjolfsson et al. 2011).

However, this data deluge also challenges firms' abilities to process such data in an effective and managerially timely manner. Big Data is characterized by high volume, high velocity, and high variety that "require new forms of processing to enable enhanced decision making, insight discovery and process optimization" (Gartner 2012). In this research, of the "3Vs of Big Data", we tackle the issues of high data volume and velocity by introducing to the marketing literature a new Bayesian estimation framework, known as Variational Bayesian (VB) inference. This framework is deterministic, scales well in the presence of large number of observations, and allows marketers to expediently estimate individual-level response parameters in order to conduct strategic and tactical marketing actions, when existing statistical methods would take too long to do so.

Understanding heterogeneity in consumer preferences is paramount to many marketing activities. Strategically, an understanding of the distribution of consumers responses to product attributes would guide firms' product design decisions – an insight that would be lost if the preference is examined only at the mean (Allenby and Rossi 1999). Tactically, getting individual-level responses to marketing actions allow firms to adjust allocation of resources across regions, stores, and consumers (Rossi et al. 1996). In the past 20 years, advances in Markov chain Monte Carlo methods have fuelled the Bayesian revolution in marketing, which has allowed researchers to change from a market level macro analysis to study micro consumers. In the hierarchical Bayesian framework, individual response sensitivities are drawn from a distribution, instead of sharing the same average sensitivity. This allows for drawing inferences about individual consumers, and also allows for information pooling across individuals, which attenuates issues of data sparseness.

The marketing literature has used a variety of MCMC methods for estimating hierarchical Bayesian models. For conjugate models, Gibbs sampling (Gelfand and Smith 1990; Gelfand et al. 1990) is the method of choice. In contrast, for non-conjugate models, the full conditional distributions do not have closed-form, and Metropolis-Hasting methods (Chib and Greenberg 1995; Rossi, Allenby and McCulloch 2005) and its extensions such as Hamiltonian MCMC and slice sampling (Neal 2003; Neal 2011), have been used across a wide spectrum of applications. Streams of research that leverage individual-level information have flourished, to address wide topics such coupon (Rossi et al. 1996), recommendation systems (Ansari, Essegaier and Kohli 2000), digital marketing campaigns (Ansari and Mela 2003), B2B communication contacts (Venkatesan and Kumar 2004), pharmaceutical detailing and sampling (Dong et al. 2009; Montoya et al. 2010), targeted promotions (Zhang and Krishnamurthi 2004), price endogeneity (Li and Ansari 2014), and dynamic targeted pricing (Zhang et al. 2014).

Despite their unquestionable usefulness and prevalence in marketing, MCMC methods are constrained by the bottlenecks of speed and scalability, and are, therefore, difficult to apply in large data situations where data sets can often contain hundreds of thousands or even millions of individuals. The situation is further exacerbated by models with complex posteriors such as those involving nonconjugacy, multiple decisions, and dynamics. It is ironical then that while large data sets contain enough information content to support complex models without the risk of overfitting, estimating these models becomes intractable using sampling based methods. MCMC methods suffer from poor mixing, especially in hierarchical models involving a large number of latent variables, and therefore, require a large number of parameter draws for convergence. Mixing also becomes difficult in large data situations because of a concentrated posterior distribution. It

is therefore not uncommon in such situations for an MCMC simulation to run for days or even weeks to ensure convergence and to obtain a reasonable number of effectively uncorrelated samples. In short, in the face of big data, traditional MCMC methods do not scale well and converge too slowly to be useful for making timely managerial decisions such as recalibrating price sensitivities for customers, or making product recommendations (e.g. managerial interests of Netflix and Amazon). To address the scalability issue, it is common to prune the data and only work with a subset of individuals or a subset of observations. This approach, however, discards information that is valuable for estimation and for targeting to every individual, and might lead to poor estimates (Zanutto and Bradlow 2006).

In this paper, we show that Variational Bayesian approaches can be used as an efficient and scalable alternative to MCMC methods in such "Big Data" situations. In contrast to MCMC methods that simulate from the posterior, variational Bayesian methods use an optimization approach to construct an approximation to the posterior. They offer the benefit of significant speedups when compared to MCMC methods and can be made highly accurate. We derive variational algorithms for the conjugate (linear) and nonconjugate (logit and ordered logit) hierarchical models commonly used in marketing, and illustrate how different variational methods that are counterparts to Gibbs and Metropolis-Hasting sampling methods can be used. We also show how recent advances in parallel computing and in stochastic approximation can be used to further enhance the speed of these VB methods. We apply the method to data sets of different sizes and demonstrate that VB methods not only can achieve the same level of accuracy as MCMC, but with speeds that can be up to thousands of times faster than MCMC.

The rest of the paper proceeds as follows. We begin in Section 2 by explaining the basic concept behind variational Bayesian inference. Then in Section 3, we introduce mean-field VB for conjugate models and fixed-form VB for nonconjugate models, and illustrate algorithms for linear-mixed and logit models. Section 4 shows how mean-field and fixed-form can be combined into a "hybrid" VB to efficiently estimate models with both conjugate and nonconjugate components, and offers an algorithm for one such case - a hierarchical ordered logit model. In Section 5, we discuss how VB can be further enhanced via parallelization and stochastic optimization with adaptive minibatch sizes. Section 6 contains applications of these VB methods and extensions to the hierarchical ordered logit model. Lastly in Section 7, we summarize the benefits of VB in marketing and highlight potential avenues for future research. All other details are located in the Appendix. The codes and data sets are available from the authors upon request.

2 Variational Bayesian Inference

Bayesian inference is based on summarizing the posterior distribution of all unknowns. In a generic Bayesian model with observed data $\boldsymbol{y}$ and the unknown parameter vector $\boldsymbol{\theta}$, the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ follows the Bayes' rule,

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}, \tag{1}$$

where, $p(\boldsymbol{\theta})$ is the prior distribution, $p(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood and $p(\boldsymbol{y})$ is the normalizing constant or evidence. For almost all interesting marketing models, the normalizing constant cannot be computed in closed-form. The posterior distribution, therefore, is not available analytically. This necessitates the use of numerical methods to approximately summarize the posterior distribution, via methods such as quadrature, Markov chain Monte Carlo or Variational Bayes.

In contrast to MCMC methods that approximate the posterior by simulating random draws, variational inference seeks to approximate the intractable posterior, $p(\boldsymbol{\theta}|\boldsymbol{y})$, with a simpler distribution, $q(\boldsymbol{\theta})$, called the *variational* distribution (Bishop 2006; Ormerod and Wand 2010). The variational distribution belongs to a family of distributions that is indexed by a set of free parameters. Variational inference searches over the space of the free parameters to find a member of the variational family that is closest to the posterior of interest. In short, variational methods recast the model inference problem as an optimization problem.

VB methods rely on the Kullback-Leibler (KL) divergence (Kullback and Leibler 1951) to measure the dissimilarity (distance) between two probability distributions.[1] The KL divergence between the approximating distribution $q(\boldsymbol{\theta})$ and the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ is defined as

$$\mathrm{KL}\left[q(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta}|\boldsymbol{y})\right] = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{\theta} \geq 0, \tag{2}$$

where, the inequality holds for all densities $q(\boldsymbol{\theta})$, with equality if and only if $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{y})$ almost everywhere. If KL divergence is zero, then $q(\boldsymbol{\theta})$ is identical to $p(\boldsymbol{\theta}|\boldsymbol{y})$. Hence, our goal here is to find a approximating variational distribution $q(\boldsymbol{\theta})$ that minimizes KL as close to 0 as possible. However, because the posterior is unknown to begin with, we need to impose restrictions on the approximating variational distribution for inference to proceed. The machine learning literature has explored a number of different

[1] Note that we cannot judge the dissimilarity of two probability distributions simply by calculating the *Euclidean* distance between distributional parameters. For instance, the two normal distributions $\mathrm{N}(0, 100^2)$ and $\mathrm{N}(10, 100^2)$ are almost indistinguishable, and the Euclidean distance between their parameter vectors is 10. In contrast, the distributions $\mathrm{N}(0, 0.1^2)$ and $\mathrm{N}(0.1, 0.1^2)$ barely overlap, but this is not reflected in the Euclidean distance between their parameter vectors, which is only 0.1. Therefore, Euclidean distance is often a poor measure of closeness between probability distributions (Hoffman et al. 2013).

approaches for structuring the approximating variational distributions. In the current paper, we focus on two approaches that can be effectively applied for marketing models – mean-field approximations for conjugate models and fixed-form approximations for nonconjugate ones. These two approaches can also be used in combination as a hybrid approximation for models involving both conjugate and nonconjugate elements.

Simple manipulation on the Kullback-Leibler divergence gives

$$\begin{aligned} \mathrm{KL}\left[q(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta}|\boldsymbol{y})\right] &= \mathrm{E}_q[\log q(\boldsymbol{\theta})] - \mathrm{E}_q[\log p(\boldsymbol{\theta}|\boldsymbol{y})] \\ &= \mathrm{E}_q[\log q(\boldsymbol{\theta})] - \mathrm{E}_q[\log p(\boldsymbol{y}, \boldsymbol{\theta})] + \log p(\boldsymbol{y}), \end{aligned} \tag{3}$$

where, the last term, $\log p(\boldsymbol{y})$, is a constant. The minimization of the Kullback-Leibler divergence is thus equivalent to maximizing the scalar quantity,

$$\mathcal{L}(q) = \mathrm{E}_q[\log p(\boldsymbol{y}, \boldsymbol{\theta})] - \mathrm{E}_q[\log q(\boldsymbol{\theta})], \tag{4}$$

which is usually referred as the evidence lower bound (ELBO; Bishop 2006; Ormerod and Wand 2010). Compared to the minimization of the KL divergence, the maximization of the ELBO is often a more convenient objective of the optimization over the free distributional parameters.

3 Mean-field and Fixed-form Methods

3.1 Mean-field Variational Bayes (MFVB)

The mean-field[2] approximation can be considered the deterministic counterpart to Gibbs sampling, and as in Gibbs sampling, it is applicable for conjugate or semiconjugate models (Ormerod and Wand 2010; Grimmer 2010). In mean-field inference, the variational distribution $q(\boldsymbol{\theta})$ is restricted to a factorized product form $\prod_{i=1}^{D} q_i(\boldsymbol{\theta}_i)$, over some partition $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D\}$ of $\boldsymbol{\theta}$. Underlying this restriction is the assumption of independence across the different parameter blocks. Such an approximation is nonparametric in spirit. Note that we make no parametric assumptions regarding the function form of $q_i(\boldsymbol{\theta}_i)$, and the only imposed assumption is the factorized product form.

Under the above assumption, setting $\partial\mathcal{L}(q)/\partial q = 0$ leads to the following optimal solution to the minimization of the Kullback-Leibler divergence (Ormerod and Wand 2010)

$$q_i^*(\boldsymbol{\theta}_i) \propto \exp\{\mathrm{E}_{\theta_{-i}}[\log p(\boldsymbol{\theta}_i|\boldsymbol{y}, \boldsymbol{\theta}_{-i})]\}, \tag{5}$$

[2]The name "mean-field" originated from physics.

where $\boldsymbol{\theta}_{-i}$ denotes parameter blocks that exclude $\boldsymbol{\theta}_i$, and $p(\boldsymbol{\theta}_i|\boldsymbol{y}, \boldsymbol{\theta}_{-i})$ is the posterior full conditional distribution for $\boldsymbol{\theta}_i$. When a conjugate prior is used, the optimal variational component, $q_i^*(\boldsymbol{\theta}_i)$, is available in closed-form and belongs to the same distributional family as the prior. Denote $q_i(\boldsymbol{\theta}_i) = q_i(\boldsymbol{\theta}_i|\boldsymbol{\eta}_i)$, and $\boldsymbol{\eta}_i$ is the parameter for the ith variational distribution. Then finding the optimal density merely requires updating the variational parameters $\{\boldsymbol{\eta}_i\}_{\forall i}$. This can be done using simple coordinate ascent optimization in which the different variational parameters are updated sequentially until convergence. We now illustrate the mean-field approximation using a cross-nested mixed linear model.

3.1.1 A Cross-Nested Mixed Linear Model

Here we consider a linear model with cross-nested random coefficients to account for both individual heterogeneity as well as product heterogeneity. Specifically, we simulate panel data sets from the following model,

$$\begin{aligned} y_{ij} &= \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_j'\boldsymbol{\lambda}_i + \boldsymbol{w}_i'\boldsymbol{\gamma}_j + e_{ij}, \\ e_{ij} &\sim \mathrm{N}(0, \sigma^2), \quad \boldsymbol{\lambda}_i \sim \mathrm{N}(0, \boldsymbol{\Lambda}), \quad p(\boldsymbol{\gamma}_j) \sim \mathrm{N}(0, \boldsymbol{\Gamma}), \end{aligned} \tag{6}$$

where, y_{ij} represent the response for person i on item j, where $i = 1, \ldots, I$. Each person is assumed to respond to an idiosyncratic set of $j \in J_i$ items. This yields an unbalanced data set with a total of $\sum_{i=1}^{I} J_i = N$ observations. Such a model arises, for instance, in recommendation systems where in users rate different items (products). The vector $\boldsymbol{x}_{ij}$ contains covariates that characterize the persons and items, $\boldsymbol{z}_j$ contains item-specific covariates and $\boldsymbol{w}_i$ contains person-specific covariates such as demographics. The vector $\boldsymbol{\lambda}_i$ contains the person-specific coefficients and the vector $\boldsymbol{\gamma}_j$ contains item-specific coefficients.

Given the linear setup, one can assume typical semiconjugate priors $p(\boldsymbol{\beta}) = \mathrm{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$, $p(\sigma^2) = \mathrm{IG}(a, b)$, $p(\boldsymbol{\Lambda}) = \mathrm{IW}(\rho_\Lambda, \boldsymbol{R}_\Lambda)$ and $p(\boldsymbol{\Gamma}) = \mathrm{IW}(\rho_\Gamma, \boldsymbol{R}_\Gamma)$. While each of these priors are individually conjugate to the normal likelihood, given the other parameters, the joint prior is not conjugate and thus the resulting posterior distribution for the model is not tractable. However, one can use a mean-field approach and approximate the posterior using a factorization as follows:

$$q(\boldsymbol{\beta}, \{\boldsymbol{\lambda}_i\}_{\forall i}, \{\boldsymbol{\gamma}_j\}_{\forall j}, \boldsymbol{\Lambda}, \boldsymbol{\Gamma}, \sigma^2) = q(\boldsymbol{\beta})q(\boldsymbol{\Lambda})q(\boldsymbol{\Gamma})q(\sigma^2)\prod_{i=1}^{I} q(\boldsymbol{\lambda}_i)\prod_{j=1}^{J_i} q(\boldsymbol{\gamma}_j). \tag{7}$$

Given the semiconjugacy, the resulting full conditional distributions are available in closed form. We can therefore, use (5) to derive the closed form variational distributions from these full conditionals. These

variational distributions are detailed in Appendix A. As can be seen from the appendix, the mean-field assumption leads to explicit solutions such that the functional form of each variational distribution is determined by its full conditionals. We also see that each variational parameter is defined in terms of the others. For instance, applying (5) to the full conditional of $\boldsymbol{\beta}$ results in $q(\boldsymbol{\beta})$ being a normal distribution, $\mathrm{N}(\boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)})$, where the variational parameters are given by,

$$\begin{aligned}
\boldsymbol{\Sigma}_{q(\beta)}^{-1} &= \boldsymbol{\Sigma}_{\beta}^{-1} + \frac{a_{q(\sigma^2)}}{b_{q(\sigma^2)}} \boldsymbol{X}'\boldsymbol{X} \\
\boldsymbol{\mu}_{q(\beta)} &= \boldsymbol{\Sigma}_{q(\beta)} \Big(\boldsymbol{\Sigma}_{\beta}^{-1} \boldsymbol{\mu}_{\beta} + \frac{a_{q(\sigma^2)}}{b_{q(\sigma^2)}} \big(\boldsymbol{X}'\boldsymbol{Y} - \sum_{i=1}^{M} \boldsymbol{X}_i' \boldsymbol{Z}_i \boldsymbol{\mu}_{q(\lambda_i)} - \sum_{j=1}^{\max[J_i]} \boldsymbol{X}_j' \boldsymbol{W}_j \boldsymbol{\mu}_{q(\gamma_j)} \big) \Big),
\end{aligned} \tag{8}$$

where $\boldsymbol{X}$ contains the $\boldsymbol{x}_{ij}'$ vectors for all the N observations, and $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are similarly obtained. We can see that the variational parameters of $\boldsymbol{\beta}$ are dependent on those of σ^2, $\{\boldsymbol{\lambda}_i\}_{\forall i}$ and $\{\boldsymbol{\gamma}_j\}_{\forall j}$. The variational parameters for the other model parameters are similarly dependent on each other. Therefore, we can maximize the evidence lower bound in (4) by iteratively updating each variational parameter, given the current value of other variational parameters. Algorithm 1 shows the iterative procedure for the variational inference scheme.

Algorithm 1. MFVB for the Cross-Nested Mixed Linear Model

1. Initialize variational parameters of the model in (6).

2. Iteratively update each of the following parameters using results in Appendix A:
$\boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)}, \{\boldsymbol{\mu}_{q(\lambda)_i}\}_{\forall i}, \{\boldsymbol{\Sigma}_{q(\lambda)_i}\}_{\forall i}, \{\boldsymbol{\mu}_{q(\gamma)_j}\}_{\forall j}, \{\boldsymbol{\Sigma}_{q(\gamma)_j}\}_{\forall j}, \rho_{q(\Lambda)}, \boldsymbol{R}_{q(\Lambda)}, \rho_{q(\Gamma)}, \boldsymbol{R}_{q(\Gamma)}, a_{q(\sigma^2)}, b_{q(\sigma^2)}$
until the evidence lower bound converges.

3.1.2 Simulated Data Comparison

We compared the mean-field approach with Gibbs sampling on simulated data sets of varying sizes. To ensure a fair comparison, we code both MFVB and MCMC in Mathematica 9 and compile them to C. The codes are run on the same Mac Air with 2G Dual Core i7 and 8G RAM.

Table 1 shows the sizes of the different data sets and the time needed for convergence in MFVB, at a tolerance of 10^{-4} on the changes in ELBO. The last column gives the time for the Gibbs sampling algorithm to finish 5000 iterations, which should be a conservative estimate on the number of MCMC iterations

required for estimating such a model. One can see from the table that MFVB requires very few iterations for convergence. When compared to the time taken by MCMC, it is clear that the variational approach results in a substantial reduction in computational time. Also, the MFVB approach scales much better than MCMC when data size increases, and requires even fewer variational iterations for larger data. In terms of estimation accuracy, MFVB and MCMC produce equally precise parameter estimates. These estimates are available from the authors upon request.

Table 1: Compare MFVB with MCMC on the Mixed Linear Model

			MFVB (Tol $= 10^{-4}$)		MCMC (5000 iter.)
I	J	# Obs.$= I \times J$	# Iter.	Time (sec.)	Time (sec.)
300	50	15,000	5	0.20	127.70
3,000	50	150,000	4	1.31	1,317.86
3,000	500	1,500,000	3	5.61	10,116.50
30,000	500	15,000,000	2	193.30	784,860.00

3.2 Fixed-form Variational Bayes (FFVB)

The fixed-form variational approximation (Honkela et al. 2010; Wang and Blei 2013; Knowles and Minka 2011; Salimans and Knowles 2013) can be used for estimating nonconjugate models. FFVB is analogous to the Metropolis-Hastings algorithm within the umbrella of MCMC in its applicability to a wide variety of nonconjugate models.

In mean-field inference, the full conditional distributions dictate the functional form of the variational distributions. However, in nonconjugate settings, the full conditional distributions are not available in closed-form. Therefore, in FFVB the variational distribution is assumed to belong to a particular family of distributions. This restriction implies that the variational distribution has a fixed functional form. For instance, when restricted to the exponential family, the variational distribution can be written as

$$q(\boldsymbol{\theta}|\boldsymbol{\eta}) = \nu(\boldsymbol{\theta}) \exp\big(S(\boldsymbol{\theta})\boldsymbol{\eta} - Z(\boldsymbol{\eta})\big), \tag{9}$$

where $\boldsymbol{\eta}$ is the vector of natural parameters for the family, the vector $S(\boldsymbol{\theta})$ contains the sufficient statistics of $\boldsymbol{\theta}$, $Z(\boldsymbol{\eta})$ ensures normalization, and $\nu(\boldsymbol{\theta})$ is the base measure. The goal is to find $\boldsymbol{\eta}$ that minimizes the KL divergence in (2):

$$\hat{\boldsymbol{\eta}} = \arg\min_{\eta} \mathrm{E}_q\big[\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\boldsymbol{y})\big]. \tag{10}$$

A number of different approaches have been used to implement fixed-form variational Bayes. Wang and Blei (2013) suggest Laplace variational inference which is based on the Laplace approximation of the posterior. Knowles and Minka (2011) use nonconjugate variational message passing with the delta method (see also Bickel and Doksum 2007; Braun and McAuliffe 2010). Salimans and Knowles (2013) propose stochastic linear regression which we adopt in current paper for fixed-form VB, given its generality and accuracy. Tan (2014) recently compare these approaches in the context of a multinomial logit model.

Salimans and Knowles (2013) and Nott et al. (2013) use the properties of the exponential family to show that the optimization in (10) leads to a fixed point update for the variational parameter

$$\boldsymbol{\eta} = \mathrm{Cov}_q\big[S(\boldsymbol{\theta})^{-1}\big]\mathrm{Cov}_q\big[S(\boldsymbol{\theta}), \log p(\boldsymbol{y}, \boldsymbol{\theta})\big]. \tag{11}$$

Instead of approximating $\mathrm{Cov}_\eta\big[S(\boldsymbol{\theta})^{-1}\big]$ and $\mathrm{Cov}_\eta\big[S(\boldsymbol{\theta}), \log p(\boldsymbol{y}, \boldsymbol{\theta})\big]$ directly, Salimans and Knowles (2013) iteratively evaluate these terms using weighted Monte Carlo with random samples of $\boldsymbol{\theta}^*$ generated from $q(\boldsymbol{\theta}|\boldsymbol{\eta})$. In the case of multivariate normal, i.e., $q(\boldsymbol{\theta}) = \mathrm{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ and $\boldsymbol{\theta} = \{\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q\}$, Minka (2001) and Opper and Archambeau (2009) show that (11) implies

$$\boldsymbol{\Sigma}_q^{-1} = -\mathrm{E}_q\big[\frac{\partial^2 \log p(\boldsymbol{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\big] \quad \text{and} \quad \boldsymbol{\mu}_q = \mathrm{E}_q\big[\boldsymbol{\theta}\big] + \boldsymbol{\Sigma}_q \mathrm{E}_q\big[\frac{\partial \log p(\boldsymbol{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\big], \tag{12}$$

where $\partial/\partial\boldsymbol{\theta}$ and $\partial^2/\partial\boldsymbol{\theta}^2$ denote the gradient vector and Hessian matrix of $\log p(\boldsymbol{y}, \boldsymbol{\theta})$, respectively. As in the general case, one can use weighted Monte Carlo to stochastically approximate the quantities $\boldsymbol{H} = -\mathrm{E}_q[\partial^2 \log p(\boldsymbol{y}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}^2]$, $\boldsymbol{g} = \mathrm{E}_q[\partial \log p(\boldsymbol{y}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}]$, and $\boldsymbol{m} = \mathrm{E}_q[\boldsymbol{\theta}]$. Algorithm 2 details the basic procedure of FFVB using stochastic linear regression. For more details about the algorithm and the theory behind it, we refer the interested readers to Salimans and Knowles (2013) and Salimans (2014).

Note that FFVB requires no explicit expression for the evidence lower bound. We declare convergence when the changes in the Euclidean norm of the parameter estimates is below 10^{-4}. We test different stopping criteria and the tolerance at 10^{-4} leads to good convergence and accurate parameter estimates.

3.2.1 A Logit Model

To illustrate the FFVB method, we conduct a simulation study based on a simple homogeneous logit model. Suppose there are J options being chosen in I occasions. The logit choice probability is given by

$$p(y_{ij} = 1) = \frac{\exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta})}{\sum_{k=1}^{J} \exp(\boldsymbol{x}_{ik}'\boldsymbol{\beta})}, \tag{13}$$

Algorithm 2. FFVB for Gaussian Variational Inference

1. Initialize $\boldsymbol{\mu}_q$, $\boldsymbol{\Sigma}_q$, $\boldsymbol{H}$, $\boldsymbol{g}$ and $\boldsymbol{m}$.
2. Initialize $\bar{\boldsymbol{H}} = \mathbf{0}$, $\bar{\boldsymbol{g}} = \mathbf{0}$ and $\bar{\boldsymbol{m}} = \mathbf{0}$.
3. Set the total number of iterations M and step size $\omega = 1/\sqrt{M}$.

4. At each iteration $n = 1, \ldots, M$:
   (a) Generate a draw $\boldsymbol{\theta}^*$ from $\mathrm{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$.
   (b) Calculate the gradient $\boldsymbol{g}^*$ and the Hessian $\boldsymbol{H}^*$ of $\log p(\boldsymbol{y}, \boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$.
   (c) Set
   $$\boldsymbol{g} = (1-\omega)\boldsymbol{g} + \omega \boldsymbol{g}^*, \boldsymbol{H} = (1-\omega)\boldsymbol{H} - \omega \boldsymbol{H}^* \text{ and } \boldsymbol{m} = (1-\omega)\boldsymbol{m} + \omega \boldsymbol{\theta}^*.$$
   (d) Update $\boldsymbol{\Sigma}_q = \boldsymbol{H}^{-1}$ and $\boldsymbol{\mu}_q = \boldsymbol{\Sigma}_q \boldsymbol{g} + \boldsymbol{m}$.
   (e) If $n > M/2$, then $\bar{\boldsymbol{g}} = \bar{\boldsymbol{g}} + \frac{2}{M}\boldsymbol{g}^*, \bar{\boldsymbol{H}} = \bar{\boldsymbol{H}} - \frac{2}{M}\boldsymbol{H}^*$ and $\bar{\boldsymbol{m}} = \bar{\boldsymbol{m}} + \frac{2}{M}\boldsymbol{\theta}^*$.

5. Set $\boldsymbol{\Sigma}_q = \bar{\boldsymbol{H}}^{-1}$ and $\boldsymbol{\mu}_q = \boldsymbol{\Sigma}_q \bar{\boldsymbol{g}} + \bar{\boldsymbol{m}}$.

where $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

For the parameter $\boldsymbol{\beta}$, we give a normal prior $\mathrm{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$, and assume the variational distribution is fixed as a normal density. Let $\bar{\boldsymbol{x}}_i = \sum_{j=1}^{J} p(y_{ij} = 1) \cdot \boldsymbol{x}_{ij}$. The gradient and Hessian of the logarithm of the unnormalized posterior can be derived as

$$\frac{\partial \log p(\boldsymbol{y}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij} \left(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i\right) - \boldsymbol{\Sigma}_\beta^{-1} \left(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta\right) \tag{14}$$

$$\frac{\partial^2 \log p(\boldsymbol{y}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = -\sum_{i=1}^{I} \sum_{j=1}^{J} p(y_{ij} = 1) \cdot \left(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i\right) \left(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i\right)' - \boldsymbol{\Sigma}_\beta^{-1}. \tag{15}$$

By plugging in these derivatives in Algorithm 2 and setting $\boldsymbol{\theta} = \boldsymbol{\beta}$, we can perform FFVB for the homogeneous logit model to estimate $\boldsymbol{\beta}$.

We simulate multiple data sets with varying sizes according to (13), and compare FFVB and MCMC in Table 2. One can see from the table that the variation approach significantly reduces the computational time when compared with MCMC, and is therefore more suitable for massive choice data sets. Again, the parameter estimates are equally accurate between the two methods.

Table 2: Compare FFVB with MCMC for the Logit Model

	FFVB (Tol $= 10^{-4}$)		MCMC (10000 iter.)
# Obs.	# Iter.	Time (sec.)	Time (sec.)
10,000	66	1.30	37.61
100,000	66	13.23	367.87
1,000,000	67	140.91	3743.86
10,000,000	67	2570.29	40618.89

4 Hybrid Variational Bayes

In marketing, we often face complex panel data situations that require hierarchical models with both conjugate and nonconjugate components. In such models, we can use a hybrid approach that combines both mean-field and fixed-form VB. Consider a generic hierarchical model with data from I individuals, $\{\boldsymbol{y}_i\}_{i=1}^I$. The unknowns are individual-specific parameters $\boldsymbol{\lambda}_i$, and two sets of hyperparameters, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The dependency structure is given by

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \{\boldsymbol{\lambda}_i\}, \{\boldsymbol{y}_i\}) = \prod_{i=1}^{I} p(\boldsymbol{y}_i|\boldsymbol{\lambda}_i, \boldsymbol{\theta}_1) p(\boldsymbol{\lambda}_i|\boldsymbol{\theta}_2) p(\boldsymbol{\theta}_2) p(\boldsymbol{\theta}_1). \tag{16}$$

The resulting augmented posterior distribution is given by

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \{\boldsymbol{\lambda}_i\}, |\{\boldsymbol{y}_i\}) = \frac{p\big(\{\boldsymbol{y}_i\}\big|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \{\boldsymbol{\lambda}_i\}\big) p\big(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \{\boldsymbol{\lambda}_i\}\big)}{\int p\big(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \{\boldsymbol{\lambda}_i\}, \{\boldsymbol{y}_i\}\big) d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2 d\boldsymbol{\lambda}_1 \dots d\boldsymbol{\lambda}_I}. \tag{17}$$

Nonconjugacy arises when the likelihood is not conjugate with the population distribution and therefore the full conditionals for the individual-level coefficients, $\boldsymbol{\lambda}_i$, $\forall i$, are not available in closed form. Similarly, we assume that the prior $p(\boldsymbol{\theta}_1)$ is not conjugate to the likelihood. On the other hand, it is common to use a multivariate normal distribution to model unobserved heterogeneity $p(\boldsymbol{\lambda}_i|\boldsymbol{\theta}_2)$. Given this restriction to the exponential family, the priors $p(\boldsymbol{\theta}_2)$ can be chosen to be conjugate or semiconjugate to the population distribution.

One can then specify the variational distribution in a factorized form for both the conjugate and nonconjugate components as follows

$$q(\{\boldsymbol{\lambda}_i\}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \prod_{i}^{I} q(\boldsymbol{\lambda}_i) q(\boldsymbol{\theta}_1) q(\boldsymbol{\theta}_2), \tag{18}$$

where the nonconjugate factors $q(\boldsymbol{\lambda}_i)$ and $q(\boldsymbol{\theta}_1)$ have fixed-forms (such as multivariate normal), whereas, the conjugate factor $q(\boldsymbol{\theta}_2)$ can be optimally deduced from the known full conditional for $\boldsymbol{\theta}_2$. The resulting estimation process involves using stochastic linear regression (Algorithm 2) for $\left\{\{\boldsymbol{\lambda}_i\}_{\forall i}, \boldsymbol{\theta}_1\right\}$, which serves an inner loop within an outer loop that updates the conjugate parameters $\boldsymbol{\theta}_2$. We now illustrate the hybrid framework within the context of a hierarchical ordered logit model.

4.1 Hierachical Ordered Logit Model

Ordered logit models are used for modeling ordered responses (Train 2009). A typical example involves responses on a rating scale, as in movie ratings within a movie recommendation system. It is common to model the observed ratings using a latent utility, u_{ij}, that represents user i's preference for the jth product. The observed rating changes whenever the underlying utility crosses a threshold. For example, in the movie recommendation context, user i's ratings can be represented as follows

$$r_{ij} = \begin{cases} 1, & u_{ij} < 0 \\ 2, & 0 < u_{ij} < \tau_1 \\ 3, & \tau_1 < u_{ij} < \tau_2 \\ 4, & \tau_2 < u_{ij} < \tau_3 \\ \vdots & \\ R, & u_{ij} > \tau_L, \end{cases}$$

where $\{\tau_1, \ldots, \tau_L\}$ are the utility cutoffs for the $R = L + 2$ rating categories. Suppose the utility of user i on item j is a function of both observed variables and unobserved components, $U_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\lambda}_i + \epsilon_{ij}$, and the distribution of the unobservable ϵ_{ij} is logistic, i.e., $F(\epsilon_{ij}) = \exp(\epsilon_{ij})/(1 + \exp(\epsilon_{ij}))$. The probability of giving a rating l, for instance, is the following

$$\begin{aligned} p(\tau_{l-2} < U_{ij} < \tau_{l-1}) &= p(u_{ij} < \tau_{l-1}) - p(u_{ij} < \tau_{l-2}) \\ &= p(\epsilon_{ij} < \tau_{l-1} - \boldsymbol{x}'_{ij}\boldsymbol{\lambda}_i) - p(\epsilon_{ij} < \tau_{l-2} - \boldsymbol{x}'_{ij}\boldsymbol{\lambda}_i) \\ &= \frac{\exp(\tau_{l-1} - \boldsymbol{x}'_{ij}\boldsymbol{\lambda}_i)}{1 + \exp(\tau_{l-1} - \boldsymbol{x}'_{ij}\boldsymbol{\lambda}_i)} - \frac{\exp(\tau_{l-2} - \boldsymbol{x}'_{ij}\boldsymbol{\lambda}_i)}{1 + \exp(\tau_{l-2} - \boldsymbol{x}'_{ij}\boldsymbol{\lambda}_i)} \quad . \end{aligned} \tag{19}$$

In this model, the individual specific parameters $\boldsymbol{\lambda}_i$'s are assumed to come from a normal population distribution, $\boldsymbol{\lambda}_i \sim \mathrm{N}(\boldsymbol{\beta}, \boldsymbol{\Lambda})$. We use typical priors $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$ and $\boldsymbol{\Lambda} \sim \mathrm{IW}(\rho_\Lambda, \boldsymbol{R}_\Lambda)$. To ensure that the cutoffs are properly ordered we use a reparametrization such that $\tau_l = \sum_{k=1}^{l} \exp(\kappa_k)$. We then impose a multivariate normal prior for the transformed parameter vector $\boldsymbol{\kappa} \sim \mathrm{N}(\boldsymbol{\mu}_\kappa, \boldsymbol{\Sigma}_\kappa)$.

The hierarchical setting results in a model setup that is a specific instantiation of the general hybrid framework outlined earlier. The hyperparameters for the normal population distribution have closed-form full conditionals which can be handled using MFVB (e.g., including the population mean $\boldsymbol{\beta}$, and the population covariance $\boldsymbol{\Lambda}$). The other parameters require FFVB, for example, the individual-level parameters $\boldsymbol{\lambda}_i$, $\forall i$, and the vector of reparametrized cutoffs $\boldsymbol{\kappa}$.

Again, we assume a factorized form for the variational approximation to the true posterior,

$$q(\boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\kappa}, \{\boldsymbol{\lambda}_i\}) = q(\boldsymbol{\beta})q(\boldsymbol{\Lambda})q(\boldsymbol{\kappa}) \prod_{i=1}^{M} q(\boldsymbol{\lambda}_i)$$

Conditional on $\boldsymbol{\lambda}_i$'s, MFVB allows us to deduce from the full conditionals closed-form variational distributions $q(\boldsymbol{\beta}) = \mathrm{N}(\boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)})$ and $q(\boldsymbol{\Lambda}) = \mathrm{IW}(\rho_{q(\Lambda)}, \boldsymbol{R}_{q(\Lambda)})$. On the other hand, FFVB assumes the distributional form is fixed for the nonconjugate parameters, e.g., $q(\boldsymbol{\lambda}_i) = \mathrm{N}(\boldsymbol{\mu}_{q(\lambda_i)}, \boldsymbol{\Sigma}_{q(\lambda_i)})$ and $q(\boldsymbol{\kappa}) = \mathrm{N}(\boldsymbol{\mu}_{q(\kappa)}, \boldsymbol{\Sigma}_{q(\kappa)})$. Following the above outlined hybrid framework, we embed the FFVB updating within the MFVB iterations to estimate the hierarchical ordered logit model. Algorithm 3 provides the detailed procedure. Table 3 describes the simulation results.

5 Speeding Up Variational Bayes

When fitting a complex model with many latent variables (such as the hierarchical ordered logit shown above) to a big data set, the hybrid variational inference procedure can become computationally challenging because the gradient and hessian are needed for every individual. Recent research has explored algorithmic and computational strategies to speed up the variational approximation. Here we propose two methods: (1) the "divide-and-recombine," or parallelization strategy (Huang and Gelman 2005; Nott et al. 2013); and (2) stochastic optimization with adaptive minibatch (Hoffman et al. 2013; Tan 2014). We apply these strategies to the hierarchical ordered logit and show in Section 6 that an integration of these two methods can provide significant improvements to the speed and scalability of variational Bayesian inference. Furthermore, as we will discuss later, parallelization can be used to address the issue of high data velocity.

Algorithm 3. Hybrid VB for the Ordered Logit Model

1. Initialize $\boldsymbol{\mu}_{q(\beta)}$, $\boldsymbol{\Sigma}_{q(\beta)}$, $\boldsymbol{\mu}_{q(\kappa)}$, $\boldsymbol{\Sigma}_{q(\kappa)}$, $\{\boldsymbol{\mu}_{q(\lambda_i)}\}_{\forall i}$, $\{\boldsymbol{\Sigma}_{q(\lambda_i)}\}_{\forall i}$, $\rho_{q(\Lambda)}$ and $\boldsymbol{R}_{q(\Lambda)}$.

2. Set the number of FFVB inner iterations M, and step size ω.

3. Update for $\boldsymbol{\lambda}_i$, $\forall i$, as follows
   (1) Initialize $\boldsymbol{H}_{\lambda_i} = \boldsymbol{\Sigma}_{q(\lambda_i)}^{-1}$, $\boldsymbol{g}_{\lambda_i} = \boldsymbol{\mu}_{q(\lambda_i)}$, $\boldsymbol{m}_{\lambda_i} = \mathbf{0}$.
   (2) Initialize $\bar{\boldsymbol{H}}_{\lambda_i} = \mathbf{0}$, $\bar{\boldsymbol{g}}_{\lambda_i} = \mathbf{0}$, $\bar{\boldsymbol{m}}_{\lambda_i} = \mathbf{0}$.
   (3) At each iteration $n = 1, \ldots, M$:
      (a) Generate a draw $\boldsymbol{\lambda}_i^*$ from $\mathrm{N}(\boldsymbol{\mu}_{q(\lambda_i)}, \boldsymbol{\Sigma}_{q(\lambda_i)})$.
      (b) Calculate the gradient $\boldsymbol{g}_{\lambda_i}^*$ and Hessian $\boldsymbol{H}_{\lambda_i}^*$ of $\log p(\boldsymbol{y}, \{\boldsymbol{\lambda}_i\}_{\forall i}, \boldsymbol{\kappa})$ at $\boldsymbol{\lambda}_i^*$.
      (c) Set
      $$\boldsymbol{g}_{\lambda_i} = (1-\omega)\boldsymbol{g}_{\lambda_i} + \omega \boldsymbol{g}_{\lambda_i}^*, \boldsymbol{H}_{\lambda_i} = (1-\omega)\boldsymbol{H}_{\lambda_i} - \omega \boldsymbol{H}_{\lambda_i}^* \text{ and } \boldsymbol{m}_{\lambda_i} = (1-\omega)\boldsymbol{m}_{\lambda_i} + \omega \boldsymbol{\lambda}_i^*.$$
      (d) Update $\boldsymbol{\Sigma}_{q(\lambda_i)} = \boldsymbol{H}_{\lambda_i}^{-1}$ and $\boldsymbol{\mu}_{q(\lambda_i)} = \boldsymbol{\Sigma}_{q(\lambda_i)} \boldsymbol{g}_{\lambda_i} + \boldsymbol{m}_{\lambda_i}$.
      (e) If $n > M/2$, then
      $$\bar{\boldsymbol{g}}_{\lambda_i} = \bar{\boldsymbol{g}}_{\lambda_i} + \tfrac{2}{M}\boldsymbol{g}_{\lambda_i}^*, \bar{\boldsymbol{H}}_{\lambda_i} = \bar{\boldsymbol{H}}_{\lambda_i} - \tfrac{2}{M}\boldsymbol{H}_{\lambda_i}^* \text{ and } \bar{\boldsymbol{M}}_{\lambda_i} = \bar{\boldsymbol{m}}_{\lambda_i} + \tfrac{2}{M}\boldsymbol{\lambda}_i^*.$$
   (4) Set $\boldsymbol{\Sigma}_{q(\lambda_i)} = \bar{\boldsymbol{H}}_{\lambda_i}^{-1}$ and $\boldsymbol{\mu}_{q(\lambda_i)} = \boldsymbol{\Sigma}_{q(\lambda_i)} \bar{\boldsymbol{g}}_{\lambda_i} + \bar{\boldsymbol{m}}_{\lambda_i}$.

4. Update for κ_l, $\forall l$, as follows
   (1) Initialize $\boldsymbol{H}_{\kappa} = \boldsymbol{\Sigma}_{q(\kappa)}^{-1}$, $\boldsymbol{g}_{\kappa} = \boldsymbol{\mu}_{q(\kappa)}$, $\boldsymbol{m}_{\kappa} = \mathbf{0}$.
   (2) Initialize $\bar{\boldsymbol{H}}_{\kappa} = \mathbf{0}$, $\bar{\boldsymbol{g}}_{\kappa} = \mathbf{0}$, $\bar{\boldsymbol{m}}_{\kappa} = \mathbf{0}$.
   (3) At each iteration $n = 1, \ldots, M$:
      (a) Generate a draw $\boldsymbol{\kappa}^*$ from $\mathrm{N}(\boldsymbol{\mu}_{q(\kappa)}, \boldsymbol{\Sigma}_{q(\kappa)})$.
      (b) Calculate the gradient $\boldsymbol{g}_{\kappa}^*$ and Hessian $\boldsymbol{H}_{\kappa}^*$ of $\log p(\boldsymbol{y}, \{\boldsymbol{\lambda}_i\}_{\forall i}, \boldsymbol{\kappa})$ at $\boldsymbol{\kappa}^*$.
      (c) Set
      $$\boldsymbol{g}_{\kappa} = (1-\omega)\boldsymbol{g}_{\kappa} + \omega \boldsymbol{g}_{\kappa}^*, \boldsymbol{H}_{\kappa} = (1-\omega)\boldsymbol{H}_{\kappa} - \omega \boldsymbol{H}_{\kappa}^* \text{ and } \boldsymbol{m}_{\kappa} = (1-\omega)\boldsymbol{m}_{\kappa} + \omega \boldsymbol{\kappa}^*.$$
      (d) Update $\boldsymbol{\Sigma}_{q(\kappa)} = \boldsymbol{H}_{\kappa}^{-1}$ and $\boldsymbol{\mu}_{q(\kappa)} = \boldsymbol{\Sigma}_{q(\kappa)} \boldsymbol{g}_{\kappa} + \boldsymbol{m}_{\kappa}$.
      (e) If $n > M/2$, then
      $$\bar{\boldsymbol{g}}_{\kappa} = \bar{\boldsymbol{g}}_{\kappa} + \tfrac{2}{M}\boldsymbol{g}_{\kappa}^*, \bar{\boldsymbol{H}}_{\kappa} = \bar{\boldsymbol{H}}_{\kappa} - \tfrac{2}{M}\boldsymbol{H}_{\kappa}^* \text{ and } \bar{\boldsymbol{m}}_{\kappa} = \bar{\boldsymbol{m}}_{\kappa} + \tfrac{2}{M}\boldsymbol{\kappa}^*.$$
   (4) Set $\boldsymbol{\Sigma}_{q(\kappa)} = \bar{\boldsymbol{H}}_{\kappa}^{-1}$ and $\boldsymbol{\mu}_{q(\kappa)} = \boldsymbol{\Sigma}_{q(\kappa)} \bar{\boldsymbol{g}}_{\kappa} + \bar{\boldsymbol{m}}_{\kappa}$.

5. Update
   $$\boldsymbol{\Sigma}_{q(\beta)} = \left(\boldsymbol{\Sigma}_{\beta}^{-1} + I\rho_{q(\Lambda)}\boldsymbol{R}_{q(\Lambda)}^{-1}\right)^{-1} \text{ and } \boldsymbol{\mu}_{q(\beta)} = \boldsymbol{\Sigma}_{q(\beta)}\left(\boldsymbol{\Sigma}_{\beta}^{-1}\boldsymbol{\mu}_{\beta} + \rho_{q(\Lambda)}\boldsymbol{R}_{q(\Lambda)}^{-1}\textstyle\sum_{i=1}^{I}\boldsymbol{\mu}_{q(\lambda_i)}\right)$$

6. Update
   $$\rho_{q(\Lambda)} = \rho_{\Lambda} + I \text{ and } \boldsymbol{R}_{q(\Lambda)} = \boldsymbol{R}_{\Lambda} + I\boldsymbol{\Sigma}_{q(\beta)} + \textstyle\sum_{i=1}^{I}\left((\boldsymbol{\mu}_{q(\lambda_i)} - \boldsymbol{\mu}_{q(\beta)})(\boldsymbol{\mu}_{q(\lambda_i)} - \boldsymbol{\mu}_{q(\beta)})' + \boldsymbol{\Sigma}_{q(\lambda_i)}\right)$$

7. Repeat Steps 3-6 until convergence.

5.1 Parallelization

In the parallelization strategy, the original data is partitioned into independent subsets. When dealing with panel data, it is important to ensure that all observations for a given individual are kept together as part of the same subset. Variational Bayesian inference is performed separately on each subset of the data. The resulting variational Bayesian distributions on the subsets are then combined appropriately to yield overall estimates of model parameters. Huang and Gelman (2005) discuss this strategy in the context of MCMC simulations, whereas, Nott et al. (2013) implement it in the context of variational inference for generalized linear mixed models. We show the gains from such parallelization in the context of the more complex hierarchical ordered logit model in Section 6.

When data $\boldsymbol{y}$ are partitioned into K subsets, $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K\}$, that are conditionally independent given the model parameter $\boldsymbol{\theta}$, the posterior distribution after the "division" becomes

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &\propto p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta}) \\ &= p(\boldsymbol{\theta}) \prod_{k=1}^{K} p(\boldsymbol{y}_k|\boldsymbol{\theta}) \\ &= p(\boldsymbol{\theta})^{1-K} \prod_{k=1}^{K} \{p(\boldsymbol{\theta})p(\boldsymbol{y}_k|\boldsymbol{\theta})\} \\ &\propto p(\boldsymbol{\theta})^{1-K} \prod_{k=1}^{K} p(\boldsymbol{\theta}|\boldsymbol{y}_k) \quad . \end{aligned} \tag{20}$$

That is, the posterior can be obtained as a combination of the separate posteriors from each data subset. This allows us to compute the variational approximations $q(\boldsymbol{\theta}|\boldsymbol{\eta}_k)$ to each subset posterior $p(\boldsymbol{\theta}|\boldsymbol{y}_k)$, $k = 1, \ldots, K$, in *parallel* using the different subsets on different processing units (kernels) simultaneously. Such parallelization can result in considerable savings in computation time because each variational approximation happens on a much smaller subset of the original data. After convergence, the subset variational approximations are recombined to yield the overall approximation to the posterior of interest

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \approx p(\boldsymbol{\theta})^{1-K} \prod_{k=1}^{K} q(\boldsymbol{\theta}|\boldsymbol{\eta}_k). \tag{21}$$

The parallelization strategy is particularly useful for variational distributions from the exponential family, which we already described in (9). Suppose $\boldsymbol{\eta}_0$ is the natural parameter of the prior $p(\boldsymbol{\theta})$, and $\boldsymbol{\eta}_k$ of

the variational distribution on the kth subset of the data, the recombination in (21) implies that the natural parameter of $p(\boldsymbol{\theta}|\boldsymbol{y})$ can be computed as

$$\boldsymbol{\eta} = \sum_{k=1}^{K} \boldsymbol{\eta}_k - (K-1)\boldsymbol{\eta}_0. \tag{22}$$

Now we translate the above general result to the context of an ordered logit model. Note that the variational distribution for $\boldsymbol{\beta}$ is a normal. Suppose that we separately estimate $q(\boldsymbol{\beta})^{(k)} = \mathrm{N}(\boldsymbol{\mu}_{q(\beta)}^{(k)}, \boldsymbol{\Sigma}_{q(\beta)}^{(k)})$ for all the data subsets, where the superscript indicates the kth data subset, then the variational parameters in $q(\boldsymbol{\beta}) = \mathrm{N}(\boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)})$ for the whole data set are

$$\boldsymbol{\Sigma}_{q(\beta)}^{-1} = \sum_{k=1}^{K} (\boldsymbol{\Sigma}_{q(\beta)}^{(k)})^{-1} - (K-1)\boldsymbol{\Sigma}_{\beta}^{-1} \quad \text{and} \quad \boldsymbol{\mu}_{q(\beta)} = \boldsymbol{\Sigma}_{q(\beta)} \cdot \Big(\sum_{k=1}^{K} (\boldsymbol{\Sigma}_{q(\beta)}^{(k)})^{-1} \boldsymbol{\mu}_{q(\beta)}^{(k)} - (K-1)\boldsymbol{\Sigma}_{\beta}^{-1}\boldsymbol{\mu}_{\beta} \Big). \tag{23}$$

The variational distributions for the cutoff vector $\boldsymbol{\kappa}$ can also be combined in a similar fashion.

As for $q(\boldsymbol{\Lambda}) = \mathrm{IW}(\rho_{q(\Lambda)}, \boldsymbol{R}_{q(\Lambda)})$, once we obtain the separate variational estimates, $\mathrm{IW}(\rho_{q(\Lambda)}^{(k)}, \boldsymbol{R}_{q(\Lambda)}^{(k)})$, from the independent runs, the combined variational parameters can be computed as

$$\rho_{q(\Lambda)} = \sum_{k=1}^{K} \rho_{q(\Lambda)}^{(k)} - (K-1)\rho_{\Lambda} \quad \text{and} \quad \boldsymbol{R}_{q(\Lambda)} = \sum_{k=1}^{K} \boldsymbol{R}_{q(\Lambda)}^{(k)} - (K-1)\boldsymbol{R}_{\Lambda}. \tag{24}$$

Next we describe a stochastic optimization technique with adaptive minibatch sizes, and use it to further speed up the variational inference of the ordered logit model.

5.2 Stochastic Optimization with Adaptive Minibatch

We start the discussion by distinguishing between *global* and *local* parameters that are involved in hierarchical models. Global parameters are invariant across observations, whereas local parameters are observation-specific. For instance, in the hierarchical ordered logit model, $\boldsymbol{\beta}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\kappa}$ are global parameters, and $\boldsymbol{\lambda}_i$'s are local because different individuals have different $\boldsymbol{\lambda}_i$. A closer look into Algorithm 3 reveals that every $\boldsymbol{\lambda}_i$ has to be updated before we can update the global parameters just once. This is not efficient because the updates for the local parameters are based upon the unconverged global parameters. The situation worsens if there are a larger number of individuals in the data. However, efficiency can be improved if the estimation does not require a full pass through the entire data in every variational iteration.

To this end, Hoffman et al. (2013) demonstrate a stochastic variational approach based on stochastic gradient descent (Robins and Monro 1951) that only utilizes a *randomly* selected subset of the whole data set – a minibatch, in each variational iteration to update global parameters. Note that these subsets of the same size vary over the iterations due to random sampling, whereas in the parallelization strategy, the subsets after division remain fixed across the iterations. A stochastic gradient is calculated based upon the randomly sampled minibatch to replace the true gradient (based on the entire data) and is used to optimize the variational objective function. Under certain regularity conditions, it has been proved that such stochastic optimization process can probabilistically converge to an optimum (Spall 2003). Hoffman et al. (2013) apply this method to topics models, and demonstrate that the use of minibatches, the size of which is fixed across iterations, can substantially facilitate the convergence of variational inference problems.

Intuitively, however, the inefficiency issue has not completely gone using such a stochastic optimization schedule. At early iterations, the global parameters are far from the optimum and there is a need to sample even fewer observations, i.e., a smaller minibatch at the beginning. Later, when the estimates get closer to the optimum, a more accurate direction for optimization is needed and a larger minibatch with more observations should be provided. Thus, the efficiency of stochastic optimization can further improve if the minibatch size is automatically and adaptively increased during the stochastic optimization process. Tan (2014) explore such an adaptive procedure for the variational inference of a multinomial logit model. We apply this adaptive approach in the context of the ordered logit model, and integrate it with the parallelization strategy discussed above. Next we outline the basics of implementing adaptive minibatch sizes. For more technical details, we refer the readers to Tan (2014).

The intuition behind the adaptive method is that a minibatch size should be increased only if the current data batch can no longer supply adequate information about the direction in which the optimization should move (Tan 2014). With a constant step size on the global parameter, initially the iterates move monotonically towards the optimum. After reaching the neighbourhood of the optimum, they tend to bounce back and forth instead of converging because the step size remains unchanged. We can use this oscillating phenomenon as an indicator that the current data batch cannot contribute to the optimization further, and increase the minibatch size accordingly.

To quantify the bouncing behavior, Tan (2014) consider the "ratio of progress and path" defined by

Gaivoronski (1988). For any univariate model parameter θ, the ratio at the hth iteration is

$$\zeta(h) = \frac{|\theta(h) - \theta(h-V)|}{\sum_{k=h-V}^{h-1} |\theta(k) - \theta(k+1)|}, \tag{25}$$

At any iteration h, the ratio lies between zero and one. It equals one if the path of change in θ during the last V iterations are monotonic, and zero if there is no progress made, i.e., $\theta(h) = \theta(h-V)$. Large value of $\zeta(h)$ implies good progress being made towards one direction (the optimum), whereas small value indicates a lot of bouncing around movement. Therefore, the ratio can be used to judge the need to increase the size of the current minibatch. Once the ratio goes below certain threshold $\bar{\zeta}$, we increase the minibatch size by a factor ψ. Both $\bar{\zeta}$ and ψ are the parameters to set by the researcher for the adaptive procedure.

6 Ordered Logit Model on Simulated Data

We use the hierarchical ordered logit model to show how the adaptive minibatch technique can be used in combination with the parallelization strategy. This involves using the minibatch framework on each subset of the overall data. The resulting variational approximations on each subset can then be combined to yield inferences on the entire data set.

For the simulation, we generate a data set of ordered choices from $I = 10,000$ individuals and $J = 20$ occasions, i.e., a total of $200,000$ observations. In each occasion, the individual chooses one out of five rating categories. This is a massive data set for studies of ordered choices that account for individual heterogeneity. Table 3 shows the results of variational inference and compares the performance across different variational approaches. As we can see in the table, there is strong synergistic effect for combining minibatch and parallelization, and can reduce the estimation time of VB by three-folds.

7 Conclusion

The rapid growth in customer database volume and data collection velocity, which characterize Big Data, have opened up exciting opportunities to understand individual customers and to customize offerings at much finer details and faster pace than ever before. These insights on individual customer behavior are crucial for the strategic and tactical success of firms in this era of big data (Brynjolfsson et al. 2011). Firms therefore need to estimate individual-level response sensitivities accurately and in a timely fashion, and

Table 3: Simulation Results for the Ordered Logit Model

Parameter	Truth	Hybrid VB	Adaptive minibatch	Parallelization	Adaptive minibatch & parallelization
β_1	0.5	0.49	0.52	0.55	0.49
		(0.01)	(0.01)	(0.01)	(0.01)
β_2	0.7	0.68	0.72	0.68	0.67
		(0.01)	(0.01)	(0.01)	(0.01)
β_3	-0.3	-0.27	-0.31	-0.29	-0.33
		(0.01)	(0.01)	(0.01)	(0.01)
β_4	0.6	0.59	0.58	0.59	0.60
		(0.01)	(0.01)	(0.01)	(0.01)
τ_1	0.5	0.50	0.49	0.50	0.49
		(0.00)	(0.00)	(0.00)	(0.00)
τ_2	1.3	1.29	1.28	1.30	1.29
		(0.00)	(0.00)	(0.00)	(0.00)
τ_3	1.8	1.78	1.78	1.79	1.78
		(0.00)	(0.00)	(0.00)	(0.00)
Λ_{11}	0.5	0.49	0.37	0.51	0.44
		(0.00)	(0.00)	(0.00)	(0.00)
Λ_{22}	0.5	0.48	0.36	0.43	0.47
		(0.00)	(0.00)	(0.00)	(0.00)
Λ_{33}	0.5	0.44	0.38	0.49	0.51
		(0.00)	(0.00)	(0.00)	(0.00)
Λ_{44}	0.5	0.42	0.37	0.50	0.48
		(0.00)	(0.00)	(0.00)	(0.00)
# Iterations		446	554	600	600
Time (sec.)		3998.22	2548.27	1966.61	1232.53

Note: the simulated data contains $200,000$ choice observations with five ordered categories. Standard deviations are included in round brackets.

need to be ready to update their estimation as new data points arrive. Traditionally, these issues could be addressed via Bayesian inference with MCMC methods (Rossi et al. 2005). However, as data sets become large, MCMC gets too computationally expensive and managerially infeasible. Moreover, a large number of marketing applications require complex models of customer behavior that are nonconjugate in nature – and this further exacerbates the computational problems faced by MCMC.

Variational Bayesian inference offers a versatile solution to the problems that arise with the high volume of big data. As we have shown via in this paper on several commonly used marketing models, VB methods can recover true parameters with the same accuracy as MCMC, but with a speed that could potentially be thousands of times faster than MCMC. In terms of wall clock time, this means the time required to estimate or re-estimate a large database of individual customer preferences would range from minutes to hours, instead of days to weeks – a crucial difference in the demand for rapid business intelligence that characterizes the current digital economy. We also show that recent innovations in parallelization and adaptive minibatch can be used to further enhance the speed of VB. In particular, using parallelization, managers could estimate parameters for the new customers or those customers with new observations as a subset, without having to run the model for the entire population. This yields considerable time savings and can address the issue of high data velocity.

In summary, the attractive features of VB for marketing are (1) the drastically improved estimation speed compared to MCMC, without a sacrifice on accuracy and without the necessity to throw away information via data-pruning, (2) the superior ability to scale to larger datasets, (3) the flexibility that mean-field and fixed-form methods, analogous to Gibbs and Metropolis-Hasting, offer in handling many conjugate and nonconjugate models widely used in marketing applications.

In this research we have illustrated the key advantages of VB for marketing models. We have addressed the issue of big data volume by demonstrating VB's superior speed and scalability, and have also touched upon the issue of big data velocity by showing how parallelization can be used to estimate only new data as they come in. As we offer an introduction to the VB framework, there are other potential avenues for future research. For example, one can apply VB to even more complex situations such as those with multiple decisions, dynamics, or structural models. For streaming data, such as the browsing behavior and choice behavior of customers using Netflix, Amazon, or any e-commerce websites, one can assess how well VB with parallelization and stochastic optimization performs against the time-stamp, and formally evaluate its ability to deal with big data velocity. From a market research prospective, one can use VB to analyze and

categorize large numbers of social media posts to extract sentiments in real time. Even more ambitiously, one can use VB to address volume, velocity and variety – the 3Vs of big data, all at once.

References

Allenby, G. M. and P. E. Rossi (1999). "Marketing models of consumer heterogeneity," *Journal of Econometrics*, 89, 57-78.

Ansari, A., S. Essegaier and R. Kohli (2000). "Internet Recommendation Systems," *Journal of Marketing Research*, 37(3), 363-375.

Ansari, A. and C. Mela (2003). "E-customization," *Journal of Marketing Research*, 40(2), 131-145.

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

Bishop, C. (2006). *Pattern recognition and machine learning*, New York, Springer.

Boyd, D. and K. Crawford (2012). "Critical questions for Big Data," *Information, Communication & Society*, 15(5), 662-679.

Braun, M. and J. McAuliffe (2010). "Variational inference for large-scale models of discrete choice," *Journal of the American Statistical Association*, 105(489), 324-335.

Broderick, T., N. Boyd, A. Wibisono, A. C. Wilson and M. I. Jordan (2013). "Streaming variational Bayes," *Working Paper*.

Brynjolfsson, E., L. Hitt and H. Kim (2011). "Strength in numbers: how does data-driven decision making affect firm performance?" *Working Paper*.

Chib, S. and E. Greenberg (1995). "Understanding the Metropolis-Hastings algorithm," *The American Statistician*, 49(4), 327-335.

Davenport, T. H. and J. G. Harris (2007). *Competing on Analytics: The New Science of Winning*, Harvard Business Review Press.

Dong, X., P. Manchanda and P. Chintagunta (2009). "Quantifying the benefits of individual level targeting in the presence of firm strategic behavior," *Journal of Marketing Research*, 46(4), 207-221.

eMarketer (2013). "Leading priorities for Big Data for business and IT."

Gaivoronski, A. (1988). "Implementation of stochastic quasigradient methods," *Numerical Techniques for Stochastic Optimization*, 313-352, New York, Springer-Verlag.

Gelfand, A. E. , S. E. Hillsb, A. Racine-Poonc and A. F. M. Smith (1990). "Illustration of Bayesian inference in normal data models using Gibbs sampling," *Journal of the American Statistical Association*, 85(412), 972-985.

Gelfand, A. E. and A. F. M. Smith (1990). "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85(410), 398-409.

Gordon, B. R., A. Goldfarb and Y. Li (2013). "Does price elasticity vary with economic growth? A cross-category analysis," *Journal of Marketing Research*, 50(1), 4-23.

Grimmer, J. (2010). "An introduction to Bayesian inference via variational approximations," *Political Analysis*, xxx.

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

Honkela, A., T. Raiko, M. Kuusela, M. Tornio and J. Karhunen (2010). "Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes," *Journal of Machine Learning Research*, 11, 3235-3268.

Huang, Z. and A. E. Gelman (2005). "Sampling for Bayesian computation with large datasets," *Working Paper*.

Knowles, D. A and T. P. Minka (2011). "Non-conjugate variational message passing for multinomial and binary regression," *Advances in Neural Information Processing Systems*, 24.

Kullback, S. and R. A. Leibler (1951). "On information and sufficiency," *The Annals of Mathematical Statistics*, 22, 79-86.

Li, Y. and A. Ansari (2014). "A Bayesian semiparametric approach for endogeneity and heterogeneity in choice models," *Management Science*, 60(5), 1161-1179.

Luts, J., T. Broderick and M. P. Wand (2013). "Real-time semiparametric regression," *Journal of Computational and Graphical Statistics*, 23(3), 589-615.

Minka, T. P. (2001). "A family of algorithms for approximate Bayesian inference," Ph.D. Thesis, MIT.

Montoya, R., O. Netzer and K. Jedidi (2010). "Dynamic allocation of pharmaceutical detailing and sampling for long-term profitability," *Marketing Science*, 29(5), 909-924.

Neal, R. M. (2003). "Slice sampling," *The Annals of Statistics*, 31(3), 705-767.

Neal, R. M. (2011). "MCMC using Hamiltonian dynamics" *Handbook of Markov Chain Monte Carlo*, 113-162.

Nott, D. J., M. Tran, A. Y. C. Kuk and R. Kohn (2013). "Efficient variational inference for generalized linear mixed models with large data sets," *Working Paper*.

Opper, M. and C. Archambeau (2009). "The variational Gaussian approximation revisited," *Neural Computation*, 21(3), 786-792.

Ormerod, J. T. and M. P. Wand (2010). "Explaining variational approximations," *The American Statistician*, 64(2), 140-153.

Ranganath, R., S. Gerrish and D. M. Blei (2013)."Black box variational inference," *Working Paper*.

Robbins, H. and S. Monro (1951). "A stochastic approximation method," *Annals of Mathematical Statistics*, 22, 400-407.

Rossi, P. E., G. M. Allenby and R. McCulloch (2005). *Bayesian Statistics and Marketing*, Wiley, 1st edition.

Rossi, P. E., R. E. McCulloch and G. M. Allenby (1996). "The value of purchase history data in target marketing," *Marketing Science*, 15(4), 321-340.

Salimans, T. (2014). "Implementing and automating fixed-form variational posterior approximation through stochastic linear regression," *Working Paper*.

Salimans, T. and D. A. Knowles (2013). "Fixed-form variational posterior approximation through stochastic linear regression," *Working Paper*.

Spall, J. (2003). *Introduction to stochastic search and optimization: estimation, simulation, and control*, John Wiley and Sons.

Tan, L. S. L. (2014). "Stochastic variational inference for large-scale discrete choice models using adaptive batch sizes," *Working Paper*.

Tan, L. S. L. and D. J. Nott (2013a). "A stochastic variational framework for fitting and diagnosing generalized linear mixed models," *Working Paper*.

Tan, L. S. L. and D. J. Nott (2013b). "Variational inference for generalized linear mixed models using partially noncentered parametrizations," *Statistical Science*, 28(2), 168-188.

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

Venkatesan, R. and V. Kumar (2004). "A customer lifetime value framework for customer selection and resource allocation strategy," *Journal of Marketing*, 68(4), 106-125.

Villas-Boas, J. M. and R. S. Winer (1999). "Endogeneity in brand choice models," *Management Science*, 45(10), 1324-1338.

Wainwright, M. J. and M. I. Jordan (2008). "Graphic models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, 1, 1-305.

Wang, C. and D. M. Blei (2013). "Variational inference in nonconjugate models," *Journal of Machine Learning Research*, 14(1), 1005-1031.

Zanutto, E. L. and E. T. Bradlow (2006). "Data pruning in consumer choice models," *Quantitative Marketing and Economics*, 5, 267-287.

Zhang, J. and L. Krishnamurthi (2004). "Customizing promotions in online stores," *Marketing Science*, 23(4), 561-578.

Zhang J. Z., O. Netzer and A. Ansari (2014). "Dynamic targeted pricing in B2B relationships," *Marketing Science*, 33(3), 317-337.

Appendices

A Variational Distributions for Linear Models

The Priors are[3]

$$p(\boldsymbol{\beta}) = \mathrm{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta), \;\; p(\boldsymbol{\lambda}_i) = \mathrm{N}(\mathbf{0}, \boldsymbol{\Lambda}), \;\; p(\boldsymbol{\Lambda}) = \mathrm{IW}(\rho_\Lambda, \boldsymbol{R}_\Lambda)$$
$$p(\boldsymbol{\gamma}_j) = \mathrm{N}(\mathbf{0}, \boldsymbol{\Gamma}), \;\; p(\boldsymbol{\Gamma}) = \mathrm{IW}(\rho_\Gamma, \boldsymbol{R}_\Gamma), \;\; p(\sigma^2) = \mathrm{IG}(a, b)$$

The full conditionals are

1) $p(\boldsymbol{\beta}|\text{rest}) = \mathrm{N}(\tilde{\boldsymbol{\mu}}_\beta, \tilde{\boldsymbol{\Sigma}}_\beta)$

$$\tilde{\boldsymbol{\Sigma}}_\beta^{-1} = \boldsymbol{\Sigma}_\beta^{-1} + \sigma^{-2}\boldsymbol{X}'\boldsymbol{X} \qquad \text{and} \qquad \tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\Sigma}}_\beta\big(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \sigma^{-2}(\boldsymbol{X}'\boldsymbol{Y} - \sum_{i=1}^{I}\boldsymbol{X}_i\boldsymbol{Z}_i\boldsymbol{\lambda}_i - \sum_{j=1}^{J}\boldsymbol{X}_j\boldsymbol{W}_j\boldsymbol{\gamma}_j)\big)$$

2) $p(\boldsymbol{\lambda}_i|\text{rest}) = \mathrm{N}(\tilde{\boldsymbol{\mu}}_{\lambda_i}, \tilde{\boldsymbol{\Sigma}}_{\lambda_i})$

$$\tilde{\boldsymbol{\Sigma}}_{\lambda_i}^{-1} = \boldsymbol{\Lambda}^{-1} + \sigma^{-2}\boldsymbol{Z}_i'\boldsymbol{Z}_i \qquad \text{and} \qquad \tilde{\boldsymbol{\mu}}_{\lambda_i} = \tilde{\boldsymbol{\Sigma}}_{\lambda_i}\big(\sigma^{-2}\boldsymbol{Z}_i'(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \sum_{j=1}^{J}\boldsymbol{I}_j\boldsymbol{W}_i\boldsymbol{\gamma}_j)\big)$$

where $\boldsymbol{I}_j$ is a $J \times J$ diagonal matrix in which the jth entry is one, and zeros elsewhere.

3) $p(\boldsymbol{\gamma}_j|\text{rest}) = \mathrm{N}(\tilde{\boldsymbol{\mu}}_{\gamma_j}, \tilde{\boldsymbol{\Sigma}}_{\gamma_j})$

$$\tilde{\boldsymbol{\Sigma}}_{\gamma_j}^{-1} = \boldsymbol{\Gamma}^{-1} + \sigma^{-2}\boldsymbol{W}_j'\boldsymbol{W}_j \qquad \text{and} \qquad \tilde{\boldsymbol{\mu}}_{\gamma_j} = \tilde{\boldsymbol{\Sigma}}_{\gamma_j}\big(\sigma^{-2}\boldsymbol{W}_j'(\boldsymbol{Y}_j - \boldsymbol{X}_j\boldsymbol{\beta} - \sum_{i=1}^{I}\boldsymbol{I}_i\boldsymbol{Z}_j\boldsymbol{\lambda}_i)\big)$$

where $\boldsymbol{I}_i$ is a $I \times I$ diagonal matrix in which the ith entry is one, and zeros elsewhere.

4) $p(\boldsymbol{\Lambda}|\text{rest}) = \mathrm{IW}(\tilde{\rho}_\Lambda, \tilde{\boldsymbol{R}}_\Lambda)$

$$\tilde{\rho}_\Lambda = \rho_\Lambda + I \qquad \text{and} \qquad \tilde{\boldsymbol{R}}_\Lambda = \boldsymbol{R}_\Lambda + \sum_{i=1}^{I}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i'$$

[3] We choose to use the following parameterization for the inverse Wishart and inverse Gamma distributions

$$\mathrm{IW}(\boldsymbol{\Lambda}|\rho_\lambda, \boldsymbol{R}_\lambda) = \frac{|\boldsymbol{R}_\lambda|^{\rho_\lambda/2}}{2^{\rho_\lambda h_\lambda/2}\Gamma_{h_\lambda}(\rho_\lambda/2)}|\boldsymbol{\Lambda}|^{-(\rho_\lambda+h_\lambda+1)/2}e^{-\mathrm{Tr}[\boldsymbol{R}_\lambda\boldsymbol{\Lambda}^{-1}]/2}$$
$$\mathrm{IG}(\sigma^2|a, b) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1}e^{-b/\sigma^2}$$

where $\mathrm{Tr}[\cdot]$ denotes the trace of a matrix.

5) $p(\boldsymbol{\Gamma}|\text{rest}) = \text{IW}(\tilde{\rho}_\Gamma, \tilde{\boldsymbol{R}}_\Gamma)$

$$\tilde{\rho}_\Gamma = \rho_\Gamma + J \qquad \text{and} \qquad \tilde{\boldsymbol{R}}_\Gamma = \boldsymbol{R}_\Gamma + \sum_{j=1}^{J} \boldsymbol{\gamma}_j \boldsymbol{\gamma}_j'$$

6) $p(\sigma^2|\text{rest}) = \text{IG}(\tilde{a}, \tilde{b})$

$$\tilde{a} = a + N/2 \qquad \text{and} \qquad \tilde{b} = b + \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(y_{ij} - \boldsymbol{x}_{ij}'\boldsymbol{\beta} - \boldsymbol{z}_{ij}'\boldsymbol{\lambda}_i - \boldsymbol{w}_{ij}'\boldsymbol{\gamma}_j \right)^2$$

Mean-field approximation assumes

$$q(\boldsymbol{\beta}, \{\boldsymbol{\lambda}_i\}_{\forall i}, \{\boldsymbol{\gamma}_j\}_{\forall j}, \boldsymbol{\Lambda}, \boldsymbol{\Gamma}, \sigma^2) = q(\boldsymbol{\beta}) q(\boldsymbol{\Lambda}) q(\boldsymbol{\Gamma}) q(\sigma^2) \prod_{i=1}^{M} q(\boldsymbol{\lambda}_i) \prod_{j=1}^{J} q(\boldsymbol{\gamma}_j) \tag{26}$$

The variational distributions are

1) $q(\boldsymbol{\beta}) \propto \exp\{\text{E}_{\text{rest}}[\log p(\boldsymbol{\beta}|\text{rest})]\} \Rightarrow q(\boldsymbol{\beta}) = \text{N}(\boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)})$, where

$$\boldsymbol{\Sigma}_{q(\beta)}^{-1} = \boldsymbol{\Sigma}_\beta^{-1} + a_{q(\sigma^2)}/b_{q(\sigma^2)} \boldsymbol{X}'\boldsymbol{X}$$

$$\boldsymbol{\mu}_{q(\beta)} = \boldsymbol{\Sigma}_{q(\beta)} \big(\boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\mu}_\beta + a_{q(\sigma^2)}/b_{q(\sigma^2)} (\boldsymbol{X}'\boldsymbol{Y} - \sum_{i=1}^{I} \boldsymbol{X}_i' \boldsymbol{Z}_i \boldsymbol{\mu}_{q(\lambda_i)} - \sum_{j=1}^{J} \boldsymbol{X}_j' \boldsymbol{W}_j \boldsymbol{\mu}_{q(\gamma_j)}) \big)$$

2) $q(\boldsymbol{\lambda}_i) \propto \exp\{\text{E}_{\text{rest}}[\log p(\boldsymbol{\lambda}_i|\text{rest})]\} \Rightarrow q(\boldsymbol{\lambda}_i) = \text{N}(\boldsymbol{\mu}_{q(\lambda_i)}, \boldsymbol{\Sigma}_{q(\lambda_i)})$, where

$$\boldsymbol{\Sigma}_{q(\lambda_i)}^{-1} = \rho_{q(\Lambda)} \boldsymbol{R}_{q(\Lambda)}^{-1} + a_{q(\sigma^2)}/b_{q(\sigma^2)} \boldsymbol{Z}_i'\boldsymbol{Z}_i$$

$$\boldsymbol{\mu}_{q(\lambda_i)} = \boldsymbol{\Sigma}_{q(\lambda_i)} \big(a_{q(\sigma^2)}/b_{q(\sigma^2)} \boldsymbol{Z}_i' (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\mu}_{q(\beta)} - \sum_{j=1}^{J} \boldsymbol{I}_j \boldsymbol{W}_i \boldsymbol{\mu}_{q(\gamma_j)}) \big)$$

3) $q(\boldsymbol{\gamma}_j) \propto \exp\{\text{E}_{\text{rest}}[\log p(\boldsymbol{\gamma}_j|\text{rest})]\} \Rightarrow q(\boldsymbol{\gamma}_j) = \text{N}(\boldsymbol{\mu}_{q(\gamma_j)}, \boldsymbol{\Sigma}_{q(\gamma_j)})$, where

$$\boldsymbol{\Sigma}_{q(\gamma_j)}^{-1} = \rho_{q(\Gamma)} \boldsymbol{R}_{q(\Gamma)}^{-1} + a_{q(\sigma^2)}/b_{q(\sigma^2)} \boldsymbol{W}_j'\boldsymbol{W}_j$$

$$\boldsymbol{\mu}_{q(\gamma_j)} = \boldsymbol{\Sigma}_{q(\gamma_j)} \big(a_{q(\sigma^2)}/b_{q(\sigma^2)} \boldsymbol{W}_j' (\boldsymbol{Y}_j - \boldsymbol{X}_j \boldsymbol{\mu}_{q(\beta)} - \sum_{i=1}^{I} \boldsymbol{I}_i \boldsymbol{Z}_j \boldsymbol{\mu}_{q(\lambda_i)}) \big)$$

4) $q(\boldsymbol{\Lambda}) \propto \exp\{\text{E}_{\text{rest}}[\log p(\boldsymbol{\Lambda}|\text{rest})]\} \Rightarrow q(\boldsymbol{\Lambda}) = \text{IW}(\rho_{q(\Lambda)}, \boldsymbol{R}_{q(\Lambda)})$, where

$$\rho_{q(\Lambda)} = \rho_\lambda + I \qquad \text{and} \qquad \boldsymbol{R}_{q(\Lambda)} = \boldsymbol{R}_\lambda + \sum_{i=1}^{I} (\boldsymbol{\Sigma}_{q(\lambda_i)} + \boldsymbol{\mu}_{q(\lambda_i)} \boldsymbol{\mu}_{q(\lambda_i)}')$$

5) $q(\boldsymbol{\Gamma}) \propto \exp\{\mathrm{E}_{\text{rest}}[\log p(\boldsymbol{\Gamma}|\text{rest})]\} \Rightarrow q(\boldsymbol{\Gamma}) = \mathrm{IW}(\rho_{q(\Gamma)}, \boldsymbol{R}_{q(\Gamma)})$, where

$$\rho_{q(\Gamma)} = \rho_\Gamma + J \qquad \text{and} \qquad \boldsymbol{R}_{q(\Gamma)} = \boldsymbol{R}_\Gamma + \sum_{j=1}^{J} (\boldsymbol{\Sigma}_{q(\gamma_j)} + \boldsymbol{\mu}_{q(\gamma_j)} \boldsymbol{\mu}'_{q(\gamma_j)})$$

6) $q(\sigma^2) \propto \exp\{\mathrm{E}_{\text{rest}}[\log p(\sigma^2|\text{rest})]\} \Rightarrow q(\sigma^2) = \mathrm{IG}(a_{q(\sigma^2)}, b_{q(\sigma^2)})$, where

$$\begin{aligned}
a_{q(\sigma^2)} &= a + N/2 \\
b_{q(\sigma^2)} &= b + \frac{1}{2} \sum_{i=1}^{I} \Big(||\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\mu}_{q(\beta)} - \boldsymbol{Z}_i \boldsymbol{\mu}_{q(\lambda_i)} - \sum_{j=1}^{J} \boldsymbol{I}_j \boldsymbol{W}_i \boldsymbol{\mu}_{q(\gamma_j)}||^2 \\
&\quad + \mathrm{Tr}[\boldsymbol{X}'_i \boldsymbol{X}_i \boldsymbol{\Sigma}_{q(\beta)}] + \mathrm{Tr}[\boldsymbol{Z}'_i \boldsymbol{Z}_i \boldsymbol{\Sigma}_{q(\lambda_i)}] \Big) + \frac{1}{2} \mathrm{Tr}[\sum_{j=1}^{J} \boldsymbol{W}'_j \boldsymbol{W}_j \Sigma_{q(\gamma_j)}].
\end{aligned}$$

The evidence lower bound is

$$\begin{aligned}
\mathcal{L} = &-(\boldsymbol{\mu}_{q(\beta)} - \boldsymbol{\mu}_\beta)' \boldsymbol{\Sigma}_\beta^{-1} (\boldsymbol{\mu}_{q(\beta)} - \boldsymbol{\mu}_\beta) - \mathrm{Tr}[\boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\Sigma}_{q(\beta)}] \\
&- (\rho_\Lambda + I) \log |\boldsymbol{R}_{q(\Lambda)}| - \rho_{q(\Lambda)} \sum_{i=1}^{I} \boldsymbol{\mu}'_{q(\lambda_i)} \boldsymbol{R}_{q(\Lambda)}^{-1} \boldsymbol{\mu}_{q(\lambda_i)} - \rho_{q(\Lambda)} \mathrm{Tr}[\boldsymbol{R}_{q(\Lambda)}^{-1} (\boldsymbol{R}_\Lambda + \sum_{i=1}^{I} \boldsymbol{\Sigma}_{q(\lambda_i)})] \\
&- (\rho_\Gamma + J) \log |\boldsymbol{R}_{q(\Gamma)}| - \rho_{q(\Gamma)} \sum_{j=1}^{J} \boldsymbol{\mu}'_{q(\gamma_j)} \boldsymbol{R}_{q(\Gamma)}^{-1} \boldsymbol{\mu}_{q(\gamma_j)} - \rho_{q(\Gamma)} \mathrm{Tr}[\boldsymbol{R}_{q(\Gamma)}^{-1} (\boldsymbol{R}_\Gamma + \sum_{j=1}^{J} \boldsymbol{\Sigma}_{q(\gamma_j)})] \\
&+ \log |\boldsymbol{\Sigma}_{q(\beta)}| + \sum_{i=1}^{I} \log |\boldsymbol{\Sigma}_{q(\lambda_i)}| + \sum_{j=1}^{J} \log |\boldsymbol{\Sigma}_{q(\gamma_j)}| - 2 a_{q(\sigma^2)} \log b_{q(\sigma^2)}
\end{aligned}$$

B Derivatives of the Ordered Logit Model

Conditional on the model parameters, ordered choices are independent across observations. For brevity here we show the gradient and Hessian based on a single choice observation, i.e, for a particular i and a particular j. Later we can write down these derivatives for the whole model simply by summing up the results appropriately across observations.

Suppose the $L+2$ ordered categories are $r_0, r_1, \ldots, r_{L+1}$. Denote

$$P_l = \frac{\exp(\tau_l - \boldsymbol{x}'_{ij} \boldsymbol{\lambda}_i)}{1 + \exp(\tau_l - \boldsymbol{x}'_{ij} \boldsymbol{\lambda}_i)} \ , \text{ for } l = 1, 2, \ldots, L,$$

and

$$P_0 = \frac{\exp(-\boldsymbol{x}'_{ij} \boldsymbol{\lambda}_i)}{1 + \exp(-\boldsymbol{x}'_{ij} \boldsymbol{\lambda}_i)} \ , \quad P_{L+1} = 1 \quad \text{and} \quad P_{-1} = 0.$$

Then, on choice occasion j, the probability for individual i choosing a category l can be written as $p(r_l) = P_l - P_{l-1}$, for $l = 0, 1, \ldots, L+1$.

The logrithm of the likelihood function of the single observation is

$$\log p(\boldsymbol{y}_{ij}|\boldsymbol{\lambda}_i, \boldsymbol{\kappa}) = \sum_{h=0}^{L+1} d_{ijh} \log\big(P_l - P_{l-1}\big), \tag{27}$$

where d_{ijh} is a binary indicator on whether individual i chose item h or not on occasion j.

Note the useful facts that

$$\frac{\partial P_l}{\partial \lambda_i} = P_l(P_l - 1)\boldsymbol{x}_{ij} \tag{28}$$

$$\frac{\partial P_l}{\partial \tau_l} = P_l(1 - P_l). \tag{29}$$

The gradient vector with respect to $\boldsymbol{\lambda}_i$ is

$$\frac{\partial \log p(\boldsymbol{y}_{ij}, \lambda_i, \boldsymbol{\kappa})}{\partial \lambda_i} = \sum_{l=0}^{L+1} d_{ijl}(P_{ijl} + P_{ij,l-1} - 1)\boldsymbol{x}_{ij} - \boldsymbol{\Sigma}_{\lambda_i}^{-1}(\boldsymbol{\lambda}_i - \boldsymbol{\mu}_{\lambda_i}) \tag{30}$$

The Hessian matrix with respect to $\boldsymbol{\lambda}_i$ is

$$\frac{\partial^2 \log p(\boldsymbol{y}_{ij}, \lambda_i, \boldsymbol{\kappa})}{\partial {\lambda_i}^2} = \sum_{l=0}^{L+1} d_{ijl}\Big(P_l(P_l - 1) + P_{l-1}(P_{l-1} - 1)\Big)\boldsymbol{x}_{ij}\boldsymbol{x}_{ij} - \boldsymbol{\Sigma}_{\lambda_i}^{-1} \tag{31}$$

The derivatives for cutoff parameters are more complicated than those for the $\boldsymbol{\lambda}_i$'s. Denote two scalar quantities,

$$s_1 = \frac{P_l(1 - P_l)}{P_l - P_{l-1}} \quad \text{and} \quad s_2 = \frac{P_{l-1}(1 - P_{l-1})}{P_l - P_{l-1}}. \tag{32}$$

After the reparameterization from $\boldsymbol{\tau}$ to $\boldsymbol{\kappa}$, we can write down the gradient vector with respect to $\boldsymbol{\kappa}$ as

$$\frac{\partial \log p(\boldsymbol{y}_{ij}, \lambda_i, \boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}} = \sum_{l=1}^{L+1} d_{ijl}\big(s_1 \frac{\partial \tau_l}{\partial \boldsymbol{\kappa}} - s_2 \frac{\partial \tau_{l-1}}{\partial \boldsymbol{\kappa}}\big), \tag{33}$$

where $\partial \tau_l / \partial \boldsymbol{\kappa} = \big(e^{\kappa_1}, \ldots, e^{\kappa_l}, 0, \ldots, 0\big)'$, for $l = 1, \ldots, L$.

The Hessian matrix with respect to $\boldsymbol{\kappa}$ is

$$\begin{aligned}\frac{\partial^2 \log p(\boldsymbol{y}_{ij}, \lambda_i, \boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}^2} = &\sum_{l=1}^{L+1} d_{ijl} \cdot \Big(s_1 \frac{\partial^2 \tau_l}{\partial \boldsymbol{\kappa} \partial \boldsymbol{\kappa}'} - s_2 \frac{\partial^2 \tau_{l-1}}{\partial \boldsymbol{\kappa} \partial \boldsymbol{\kappa}'} + s_1(1 - 2P_l - s_1)\frac{\partial \tau_l}{\partial \boldsymbol{\kappa}}\frac{\partial \tau_l}{\partial \boldsymbol{\kappa}'} \\ &- s_2(1 - 2P_{l-1} + s_2)\frac{\partial \tau_{l-1}}{\partial \boldsymbol{\kappa}}\frac{\partial \tau_{l-1}}{\partial \boldsymbol{\kappa}'} + s_1 s_2(\frac{\partial \tau_{l-1}}{\partial \boldsymbol{\kappa}}\frac{\partial \tau_l}{\partial \boldsymbol{\kappa}'} + \frac{\partial \tau_l}{\partial \boldsymbol{\kappa}}\frac{\partial \tau_{l-1}}{\partial \boldsymbol{\kappa}'})\Big),\end{aligned} \tag{34}$$

where $\partial^2 \tau_l / \partial \boldsymbol{\kappa} \partial \boldsymbol{\kappa}'$ is a diagnal matrix with the vector $\partial \tau_l / \partial \boldsymbol{\kappa}$ on the diagnal.