Approximations of the Information Matrix for a Panel Mixed Logit Model

Wei Zhang [1] **Abhyuday Mandal** [2] **John Stufken** [3]

Abstract

Information matrices play a key role in identifying optimal designs. Panel mixed logit models are more flexible than multinomial logit models for discrete choice experiments. For panel mixed logit models, the information matrix does not have a closed form expression and is difficult to evaluate. We propose three methods to approximate the information matrix, namely importance sampling, Laplace approximation and joint sampling. The three methods are compared through simulations. Since our ultimate goal is to find optimal designs, the three methods are compared on whether they rank designs similarly, not on how accurate the approximations are. Although the Laplace approximation is not as accurate as the other two methods, it can still be used to rank designs accurately and it is much faster than the other two methods. For an optimal design search using an exchange algorithm takes days to run, the Laplace approximation may be the only viable choice to use in practice.

Keywords: Discrete choice experiments, optimal designs, Laplace's method, importance sampling, joint sampling, A-optimality, D-optimality.

1 Introduction

In marketing, transportation and health care, researchers are interested in understanding how people make their choices. Such consumer behaviors can be analyzed with discrete choice models (Train (2009), Rossi, Allenby and McCulloch (2006) and Hensher, Rose and Greene (2005)). One of the most popular discrete choice models is the multinomial logit

[1]Department of Statistics, University of Georgia, Athens, GA 30602, USA. email: zhangwei@uga.edu

[2]Department of Statistics, University of Georgia, Athens, GA 30602, USA. email: amandal@stat.uga.edu

[3]The research of John Stufken was supported in part by NSF Grant DMS-14-06760. School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ 85287, USA. email: jstufken@asu.edu

model, but it has several limitations in representing the choice behaviors (McFadden (1974)). Recently, mixed logit models have become more popular, because they can relax assumptions in the multinomial logit model (McFadden and Train (2000), Bhat (1998), Brownstone and Train (1999), Erdem (1996), Revelt and Train (1998) and Bhat (2000)). However, mixed logit models belong to the class of generalized linear mixed models, for which designing an experiment and analyzing the data are difficult, since the likelihood functions do not have closed-form expressions (McCulloch (1997), Booth and Hobert (1999), Breslow and Clayton (1993), Wand (2007), Moerbeek and Maas (2005) and Waite and Woods (2014)).

When respondents choose from several products, discrete choice models can be used to explore the relationship between their choices and the attributes of the products. The multinomial logit model is popular for its simple analytical form, but it assumes a homogenous population (Train (2009)). Mixed logit models (McFadden and Train (2000)) can account for the heterogeneity in the population. If respondents are asked to choose from more than one choice set, the mixed logit model used is called a panel mixed logit model (Erdem (1996), Revelt and Train (1998) and Bhat (2000)). In a panel mixed logit model, a respondent is assumed to use similar rules to make a sequence of choices, so the choices from the same respondent are correlated.

Unlike multinomial logit models, mixed logit models do not have closed-form likelihood functions, so designing an experiment and analyzing the data are difficult. For the analysis, likelihood functions are simulated by Monte Carlo methods (Revelt and Train (1998)). For the design, information matrices are often used to form criteria that measure qualities of the designs (Atkinson, Donev and Tobias (2007)). Since information matrices also do not have closed-form expressions, we need a method to evaluate information matrices.

For mixed logit models, the expression for the information matrix, which does not have a closed-form expression, is often derived and simplified first, followed by an approximation method based on the simplified expression. For the cross-sectional mixed logit model, Sándor and Wedel (2002) provide an expression for the information matrix that makes the evaluation straightforward using Monte Carlo method. Sándor and Wedel (2002) used cross-sectional mixed logit model for panel data, where responses from the same respondent are assumed to be independent. For the panel mixed logit model, Bliemer and Rose (2010) derive an expression for the information matrix, which is more complex than that for the cross-sectional mixed logit model. Their expression is also too complex to explore the structures in the information matrix. We simplify their expression and make use of the new expression to propose more efficient methods for approximating the information matrix. With respect to a design criterion, the optimal designs are the ones that optimize the criterion and search algorithms can be used to find efficient designs. Since many information matrices are evaluated in search algorithms, efficient methods of approximating the information matrix can reduce the time of the search considerably.

In this paper, we will first derive the simplified expression for the information matrix under a panel mixed logit model. As in Bliemer and Rose (2010), the expression consists of two expectations, but the two expectations involved are different. For the two expectations in our expression, one is with respect to the posterior distribution of the random effects given the responses, the other is with respect to the distribution of the responses. The former is nested within the latter. We can evaluate the expression in two ways − independently or together. If the two expectations are approximated independently, the expectation with respect to the responses is considered first. Then to approximate the expectation with respect to the posterior distribution, we consider techniques from the literature of discrete choice models and generalized linear mixed models: McCulloch (1997) and Rossi, Allenby and McCulloch (2006) use a Metropolis algorithm, Booth and Hobert (1999) use rejection sampling, McCulloch (1997) and Booth and Hobert (1999) use importance sampling, and Tierney and Kadane (1986) and Tierney, Kass and Kadane (1989) apply Laplace's method to approximate the posterior mean. We find that the Metropolis algorithm is too time consuming for approximating the information matrix, rejection sampling is not applicable for the posterior distribution considered here, and importance sampling and the Laplace approximation are viable to use here. If we consider the two expectations together, we propose another method which uses samples from the joint distribution of the responses and the random effects. The three methods, importance sampling, Laplace approximation and joint sampling, are compared in a simulation study. We find that although the Laplace approximation is not as accurate as the other two methods, it can still be used to rank designs and is much faster than the other two methods. Since our ultimate goal is to find efficient designs and not to approximate information matrices, the ranking of the designs is more important than the actual information matrices. We conclude that the Laplace approximation is the most efficient method to use in search algorithms.

The paper is organized as follows. In Section 2, we introduce the panel mixed logit model and give the simplified expression of the information matrix. Methods for approximating the information matrix are discussed in Section 3 and three methods are proposed. In Section 4, we use simulations to compare the three methods. The paper concludes with a discussion in Section 5.

2 Model, Information Matrix and Design Criteria

We start by introducing the formulation of the panel mixed logit model.

2.1 Panel Mixed Logit Model

In a typical choice experiment, there are several questions that ask the respondents to choose one from several alternatives presented to them. The set consisting of the alternatives in each question is called a choice set. From the respondents' choices in the choice sets, we can get information about the preferences of the respondents. The alternatives are identified by the level combinations of the attributes. For example, suppose a beverage has price (low and high) and volume (small and large) as attributes. One beverage with low price and small volume corresponds to a product that is different from another product—a beverage with low price and large volume.

Let S denote the number of choice sets presented to each respondent and J the number of alternatives in each choice set. Let x_{nsj} be the k-dimensional vector containing the coded levels of the q attributes for alternative j in choice set s for respondent n and denote by β_n the corresponding k-dimensional coefficient vector. The details of the coding are given in Section 4. Then, the coded design matrix for respondent n is given by a $SJ \times k$ matrix $X_n = (x_{n11}, x_{n12}, \ldots, x_{nSJ})'$. The corresponding response vector is given by $Y_n = (Y_{n11}, Y_{n12}, \ldots, Y_{nSJ})'$, where $Y_{nsj} = 1$ if respondent n chooses alternative j in choice set s and $Y_{nsj} = 0$ otherwise. In each choice set, $\sum_{j=1}^{J} Y_{nsj} = 1$ where $1 \leq s \leq S$, because the respondent chooses only one alternative in each choice set.

We now introduce the panel mixed logit model. In choice set s, if β_n is given, the probability of respondent n choosing alternative j is

$$P(Y_{nsj} = 1|\beta_n) = \frac{\exp(x'_{nsj}\beta_n)}{\sum_{i=1}^{J} \exp(x'_{nsi}\beta_n)}.$$

In the above formula, β_n is assumed to be constant across the S (> 1) choice sets. Given β_n, the choices made by respondent n are independent and the conditional probability of observing a sequence of choices y_n is

$$P(Y_n = y_n|\beta_n) \;=\; \prod_{s=1}^{S}\prod_{j=1}^{J}\left(\frac{\exp(x'_{nsj}\beta_n)}{\sum_{i=1}^{J} \exp(x'_{nsi}\beta_n)}\right)^{y_{nsj}}.$$

The above expression is the probability of observing y_n in a multinomial logit model where β_n is a fixed parameter vector. In a mixed logit model, β_n is assumed to be a random vector, whose density function is $f_\theta(\beta_n)$ with θ being the vector of unknown parameters.

The unconditional probability of observing y_n is

$$P_\theta(Y_n = y_n) = \int P(Y_n = y_n|\beta_n) f_\theta(\beta_n)\,\mathrm{d}\beta_n = \int \prod_{s=1}^{S}\prod_{j=1}^{J}\left(\frac{\exp(x'_{nsj}\beta_n)}{\sum_{i=1}^{J}\exp(x'_{nsi}\beta_n)}\right)^{y_{nsj}} f_\theta(\beta_n)\,\mathrm{d}\beta_n.$$

The above expression reflects that choices by the same respondent in different choice sets are not independent.

For a sample $y = (y'_1, y'_2, \ldots, y'_N)'$ of N respondents, the likelihood function of θ is

$$L(\theta|Y = y) = \prod_{n=1}^{N} P_\theta(Y_n = y_n).$$

2.2 Information Matrix

The asymptotic variance-covariance matrix of the maximum likelihood estimator of θ is equal to the inverse of the information matrix. The information matrix can be calculated as

$$I(\theta|X) = E_Y\Big((\frac{\partial \log L(\theta|Y)}{\partial\theta})(\frac{\partial \log L(\theta|Y)}{\partial\theta})'\Big),$$

where $X = (X'_1, X'_2, \ldots, X'_N)'$ is the $NSJ \times k$ coded design matrix for the N respondents.

Usually, β_n is assumed to be a random vector from a multivariate normal distribution $N_k(b, \Sigma)$ with $b = (b_1, b_2, \ldots, b_k)'$ and $\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)$. The normal random vector β_n can be written as $\beta_n = b + u_n$ where $u_n \sim N_k(0, \Sigma)$. Let $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_k)'$, then the vector of unknown parameters is $\theta = (b', \sigma')'$. The information matrix for θ is

$$I(\theta|X) = \sum_{n=1}^{N}\begin{pmatrix} E_{Y_n}\Big((\frac{\partial \log L_n}{\partial b})(\frac{\partial \log L_n}{\partial b})'\Big) & E_{Y_n}\Big((\frac{\partial \log L_n}{\partial b})(\frac{\partial \log L_n}{\partial \sigma})'\Big) \\ E_{Y_n}\Big((\frac{\partial \log L_n}{\partial \sigma})(\frac{\partial \log L_n}{\partial b})'\Big) & E_{Y_n}\Big((\frac{\partial \log L_n}{\partial \sigma})(\frac{\partial \log L_n}{\partial \sigma})'\Big) \end{pmatrix},$$

where $L_n = P_\theta(Y_n = y_n)$ is the likelihood function for respondent n and is given by

$$\begin{aligned} & P_\theta(Y_n = y_n) \\ = & \int P_b(Y_n = y_n|u_n) f_\sigma(u_n)\,\mathrm{d}u_n \\ = & \int \prod_{s=1}^{S}\prod_{j=1}^{J}\left(\frac{\exp(x'_{nsj}(b+u_n))}{\sum_{i=1}^{J}\exp(x'_{nsi}(b+u_n))}\right)^{y_{nsj}} (2\pi)^{-k/2}|\Sigma|^{-1/2}\exp(-\frac{1}{2}u'_n\Sigma^{-1}u_n)\,\mathrm{d}u_n. \end{aligned}$$

The score function for respondent n is

$$\frac{\partial \log L_n}{\partial b} = \frac{1}{L_n}\frac{\partial L_n}{\partial b} = X_n'\big(y_n - E_{u_n}(p_n|y_n)\big), \tag{1}$$

where $p_n = (p_{n1}', p_{n2}', \ldots, p_{nS}')'$ with $p_{ns} = (p_{ns1}, p_{ns2}, \ldots, p_{nsJ})'$ and $p_{nsj} = P_b(Y_{nsj} = 1|u_n) = \frac{\exp(x_{nsj}'(b+u_n))}{\sum_{i=1}^{J}\exp(x_{nsi}'(b+u_n))}$; and

$$\frac{\partial \log L_n}{\partial \sigma} = -\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_k}\right)' + E_{u_n}\left[\left(\frac{u_{n1}^2}{\sigma_1^3}, \ldots, \frac{u_{nk}^2}{\sigma_k^3}\right)'|y_n\right], \tag{2}$$

where u_{ni} is the ith element of u_n, $1 \le i \le k$. The above expressions are derived in Appendix 6.1

Then, it can be shown that expressions in the information matrix are given by

$$\begin{aligned}
E_{Y_n}\Big((\frac{\partial \log L_n}{\partial b})(\frac{\partial \log L_n}{\partial b})'\Big) &= X_n'\Big(E_{u_n}(\Delta_n) - E_{u_n}(p_n p_n') + E_{Y_n}\big[E_{u_n}(p_n|Y_n)E_{u_n}(p_n'|Y_n)\big]\Big)X_n,\\
E_{Y_n}\Big((\frac{\partial \log L_n}{\partial b})(\frac{\partial \log L_n}{\partial \sigma})'\Big) &= X_n'\left(E_{u_n}\left[p_n\left(\frac{u_{n1}^2}{\sigma_1^3}, \ldots, \frac{u_{nk}^2}{\sigma_k^3}\right)\right]\right.\\
&\qquad \left. -E_{Y_n}\left[E_{u_n}(p_n|Y_n)E_{u_n}\left(\left(\frac{u_{n1}^2}{\sigma_1^3}, \ldots, \frac{u_{nk}^2}{\sigma_k^3}\right)|Y_n\right)\right]\right),\\
E_{Y_n}\Big((\frac{\partial \log L_n}{\partial \sigma})(\frac{\partial \log L_n}{\partial \sigma})'\Big) &= -\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_k}\right)'\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_k}\right)\\
&\qquad + E_{Y_n}\left[E_{u_n}\left(\left(\frac{u_{n1}^2}{\sigma_1^3}, \ldots, \frac{u_{nk}^2}{\sigma_k^3}\right)'|Y_n\right)E_{u_n}\left(\left(\frac{u_{n1}^2}{\sigma_1^3}, \ldots, \frac{u_{nk}^2}{\sigma_k^3}\right)|Y_n\right)\right],
\end{aligned} \tag{3}$$

where $\Delta_n = \mathrm{diag}(\Delta_{ns})$ with $\Delta_{ns} = \mathrm{diag}(p_{ns}) - p_{ns}p_{ns}'$. These expressions are also derived in Appendix 6.1. They will be used to evaluate the information matrix in order to identify optimal designs, as discussed below.

2.3 Design Criteria

For a univariate estimator, one with a small variance is desirable. For a multivariate estimator, the generalization of variance is the variance-covariance matrix. As mentioned in Subsection 2.2, the asymptotic variance-covariance matrix of the maximum likelihood estimator is equal to the inverse of the information matrix. Hence, a real-valued function of the information matrix is usually used to formulate the design criterion. D-optimality is usually used as the design criterion, which seeks to minimize $\det\big[I(\theta|X)\big]^{-1/2k}$ (often called D-error

in the context of discrete choice experiments) over all possible choices of X, where $2k$ is the number of parameters in θ. A-optimality is another frequently used design criterion, for which the average of the eigenvalues of $I(\theta|X)^{-1}$, i.e., the trace of $I(\theta|X)^{-1}$ divided by $2k$, is minimized.

Note that $I(\theta|X)$ depends on the parameter vector θ, which is unknown prior to the experiment. To overcome this problem, an estimated value of θ from previous studies or an educated guess can be used. Optimal designs found by this method are called locally optimal designs (Chernoff (1953)). Here, locally D-optimal designs are the designs that minimize the D-optimality criterion for a given value of θ. Similarly, the locally A-optimal designs are the designs that minimize the A-optimality criterion for a given value of θ.

3 Approximation of the Information Matrix

The expressions of information matrices for different respondents are the same, but different choices of X_n can be used. Hence, for the demonstration of how to approximate the information matrix, we will use X_1 $(SJ \times k)$ for respondent 1 as an example. Correspondingly, Y_1 $(SJ \times 1)$ and u_1 $(k \times 1)$ are the response and random effect for respondent 1.

The expressions in (3) cannot be evaluated explicitly, because they contain intractable integrals. In (3), the terms $E_{u_1}(\Delta_1)$, $E_{u_1}(p_1 p_1')$ and $E_{u_1}\big[p_1(\frac{u_{11}^2}{\sigma_1^3}, \ldots, \frac{u_{1k}^2}{\sigma_k^3})\big]$ only involve expectations with respect to u_1, so Monte Carlo methods can be applied directly to evaluate these terms.

However, the following terms involve additional expectations with respect to Y_1:

$$E_{Y_1}\big[E_{u_1}(p_1|Y_1)E_{u_1}(p_1'|Y_1)\big], E_{Y_1}\Big[E_{u_1}(p_1|Y_1)E_{u_1}\big((\frac{u_{11}^2}{\sigma_1^3}, \ldots, \frac{u_{1k}^2}{\sigma_k^3})|Y_1\big)\Big],$$
$$\text{and } E_{Y_1}\Big[E_{u_1}\big((\frac{u_{11}^2}{\sigma_1^3}, \ldots, \frac{u_{1k}^2}{\sigma_k^3})'|Y_1\big)E_{u_1}\big((\frac{u_{11}^2}{\sigma_1^3}, \ldots, \frac{u_{1k}^2}{\sigma_k^3})|Y_1\big)\Big]. \quad (4)$$

The two layers of expectations make the approximation of these terms computationally expensive. For simplicity, we denote these terms in a general form as

$$E_{Y_1}\big[E_{u_1}(g(u_1)|Y_1)E_{u_1}(h(u_1)'|Y_1)\big],$$

where both $g(u_1)$ and $h(u_1)$ are vectors of functions of u_1. The approximation methods that we propose for such expressions can be classified into two categories, which are differentiated by whether samples of Y_1 and samples of u_1 are drawn independently or jointly. In Subsection 3.1 , we will discuss different methods for sampling independently, while Subsection 3.2 discusses the method for sampling jointly.

3.1 Approximations Using Samples from Marginal Distributions

For methods in this section, the approximation is done in two steps.

In the first step, a sample is drawn from the marginal distribution of Y_1 to approximate the expectation $E_{Y_1}\left[E_{u_1}(g(u_1)|Y_1)E_{u_1}(h(u_1)'|Y_1)\right]$ with respect to Y_1.

The marginal sample can be easily obtained from a joint sample, so we introduce how to get the joint sample next. The density function for the joint distribution of Y_1 and u_1 is given by $f_\theta(y_1, u_1) = P_b(Y_1 = y_1|u_1)f_\sigma(u_1)$. To get the ith sample point (y_1^i, u_1^i) from the joint distribution, first a u_1^i is drawn from $f_\sigma(u_1)$, then a y_1^i is generated from $P_b(Y_1 = y_1|u_1^i)$ in two steps:

1. In choice set s, given u_1^i the response $(Y_{1s1}, Y_{1s2}, \ldots, Y_{1sJ})'$ follows a multinomial distribution with probabilities $(p_{1s1}^i, p_{1s2}^i, \ldots, p_{1sJ}^i)'$, where $p_{1sj}^i = \frac{\exp(x_{1sj}'(b+u_1^i))}{\sum_{l=1}^J \exp(x_{1sl}'(b+u_1^i))}$. Given u_1^i, a $(y_{1s1}^i, y_{1s2}^i, \ldots, y_{1sJ}^i)'$ is simulated for each choice set s, $1 \le s \le S$.

2. Noting that given u_1^i the responses in different choice sets are independent, the ith sample y_1^i can be obtained by juxtaposing the simulated responses for all choice sets in the previous step.

Suppose the sample size is n_y, then the joint sample is $(y_1^1, u_1^1), \ldots, (y_1^{n_y}, u_1^{n_y})$. Finally, a sample of Y_1 from the marginal distribution can be obtained by using the y part in the joint sample $(y_1^1, u_1^1), \ldots, (y_1^{n_y}, u_1^{n_y})$, which is $y_1^1, \ldots, y_1^{n_y}$.

Now, $E_{Y_1}\left[E_{u_1}(g(u_1)|Y_1)E_{u_1}(h(u_1)'|Y_1)\right]$ is approximated by

$$\frac{1}{n_y}\sum_{i=1}^{n_y} E_{u_1}(g(u_1)|y_1^i)E_{u_1}(h(u_1)'|y_1^i). \tag{5}$$

In the second step, $E_{u_1}(g(u_1)|y_1^i)$, $1 \le i \le n_y$, is considered. Note that $E_{u_1}(g(u_1)|y_1^i)$ is a posterior mean and the posterior density is given by

$$\begin{aligned} f_\theta(u_1|y_1^i) &\propto P_b(Y_1 = y_1^i|u_1) \times f_\sigma(u_1) \\ &\propto \prod_{s=1}^S \prod_{j=1}^J \left(\frac{\exp(x_{1sj}'(b+u_1))}{\sum_{l=1}^J \exp(x_{1sl}'(b+u_1))}\right)^{y_{1sj}^i} \times (2\pi)^{-k/2}|\Sigma|^{-1/2}\exp(-\frac{1}{2}u_1'\Sigma^{-1}u_1). \end{aligned}$$

From the literature, the following methods can be used to approximate $E_{u_1}\big(g(u_1)|y_1^i\big)$.

1. **Metropolis Algorithm:** For generalized linear mixed models, McCulloch (1997) uses a Metropolis algorithm to take samples from the posterior distribution and then form Monte Carlo approximations to the desired posterior means in the Monte Carlo EM algorithm. Rossi, Allenby and McCulloch (2006) consider two Metropolis variants to take samples from the posterior distribution for the multinomial logit model.

   To approximate the information matrix, we need to approximate $E_{u_1}\big(g(u_1)|y_1^i\big)$ where $1 \leq i \leq n_y$, so a sample of $u_1|y_1^i$ is required for every i. Since samples drawn by this method are dependent, a large sample size is usually required for it to work. Additionally, when we search for optimal designs in Section 4, we also need to approximate the information matrices of a large number of designs. Hence, it is not feasible to use the Metropolis algorithm in practice for our problem.

2. **Rejection Sampling:** For generalized linear mixed models, Booth and Hobert (1999) use rejection sampling to take samples from the posterior distribution in the Monte Carlo EM algorithm. The method they use is carried out in two steps. In step 1, a u_1^1 is drawn from $f_\sigma(u_1)$ and a w is drawn from the uniform(0,1) distribution. In step 2, if $w \leq P_b(Y_1 = y_1^i|u_1^1)/\tau$ where $\tau = \sup_{u_1} P_b(Y_1 = y_1^i|u_1)$, then u_1^1 is accepted; otherwise, start from step 1 again. This procedure stops when a desired sample size is attained. In step 2, $P_b(Y_1 = y_1^i|u_1)$ is maximized as a function of u_1.

   Here, since y_1^i is the response vector from respondent 1 and the number of choice sets for a respondent cannot be very large, it is not always possible to find a u_1 that maximizes $P_b(Y_1 = y_1^i|u_1)$. Hence, the previous rejection sampling method is not applicable for the posterior distribution considered here.

3. **Importance sampling:** For generalized linear mixed models, McCulloch (1997) and Booth and Hobert (1999) also use importance sampling, with the former using it to approximate the log-likelihood and the latter for the posterior means in the EM algorithm. To approximate the likelihood function, McCulloch (1997) uses the density function of the random effects as the importance density. Booth and Hobert (1999) use a multivariate t density whose mean and variance match the mode and curvature of the posterior distribution as the importance density.

   For our problem, since the posterior mean can be written as the ratio of two expectations, importance sampling is used to approximate both expectations. Let $u_1^{i1}, u_1^{i2}, \ldots, u_1^{in_u}$ be a set of random samples from the importance density $q(u_1)$ that has the same support as $f_\theta(u_1|y_1^i)$. Then, $E_{u_1}(g(u_1)|y_1^i)$ is approximated by

$$E_{u_1}(g(u_1)|y_1^i) \approx \frac{\sum_{j=1}^{n_u} g(u_1^{ij}) P_b(Y_1 = y_1^i|u_1^{ij}) f_\sigma(u_1^{ij})/q(u_1^{ij})}{\sum_{j=1}^{n_u} P_b(Y_1 = y_1^i|u_1^{ij}) f_\sigma(u_1^j)/q(u_1^{ij})}.$$

   For our problem, we will use the density of the random effects, $f_\sigma(u_1)$, as the importance

density.

As an alternative to (5), $E_{Y_1}\left[E_{u_1}(g(u_1)|Y_1)E_{u_1}(h(u_1)'|Y_1)\right]$ can be calculated directly as

$$\sum_{y_1^i \in A} E_{u_1}(g(u_1)|y_1^i)E_{u_1}(h(u_1)'|y_1^i)P_\theta(Y_1 = y_1^i),$$

where A is the set that contains all possible values for Y_1. In situations where the number of possible values for Y_1 is not very large, we can make use of the above expression. We only need to find a way to approximate $P_\theta(Y_1 = y_1^i)$. Since we have a sample $u_1^{i1}, u_1^{i2}, \ldots, u_1^{in_u}$ from importance density $f_\sigma(u_1)$, we can approximate $P_\theta(Y_1 = y_1^i)$ as $\frac{1}{n_u}\sum_{j=1}^{n_u} P_b(Y_1 = y_1^i|u_1^{ij})$.

4. **Laplace approximation:** Let the lth element of $g(u_1)$ be $g_l(u_1)$. Assuming for now u_1 is univariate and $g_l(u_1)$ is a smooth and positive function of u_1, the posterior mean of $g_l(u_1)$ can be written as

$$E_{u_1}[g_l(u_1)|y_1^i] = \frac{\int e^{\log g_l(u_1)+\log P_b(Y_1=y_1^i|u_1)+\log f_\sigma(u_1)}\,\mathrm{d}u_1}{\int e^{\log P_b(Y_1=y_1^i|u_1)+\log f_\sigma(u_1)}\,\mathrm{d}u_1}.$$

With $Q(u_1) = \log P_b(Y_1 = y_1^i|u_1) + \log f_\sigma(u_1)$ and $q_l(u_1) = \log g_l(u_1) + \log P_b(Y_1 = y_1^i|u_1) + \log f_\sigma(u_1)$, the above expression can be written as

$$E_{u_1}[g_l(u_1)|y_1^i] = \frac{\int e^{q_l(u_1)}\,\mathrm{d}u_1}{\int e^{Q(u_1)}\,\mathrm{d}u_1}.$$

Tierney and Kadane (1986) apply Laplace's method to integrals in the numerator and the denominator and obtain an approximation of the posterior mean. Let $\hat{u}_1$ be the mode of $Q(u_1)$ and $d^2 = -1/Q''(u_1)|_{u_1=\hat{u}_1}$. Then, Laplace's method approximates the integral in the denominator by

$$\int e^{Q(u_1)}\mathrm{d}u_1 \approx \int \exp\left[\frac{Q(\hat{u}_1) - (u_1 - \hat{u}_1)^2}{2d^2}\right]\mathrm{d}u_1 = \sqrt{2\pi}|d|e^{Q(\hat{u}_1)}.$$

Similarly, if $\hat{u}_{1_l}$ is the mode of $q_l(u_1)$ and ${d_l}^2 = -1/(q_l(u_1))''|_{u_1=\hat{u}_{1_l}}$, then Laplace's method approximates integral in the numerator by $\sqrt{2\pi}|d_l|\exp(q_l(\hat{u}_{1_l}))$. Taking the ratio of these two approximations, the Laplace approximation of $E_{u_1}[g_l(u_1)|y_1^i]$ is given by

$$E_{u_1}[g_l(u_1)|y_1^i] \approx \frac{|d_l|}{|d|}\exp\left[q_l(\hat{u}_{1_l}) - Q(\hat{u}_1)\right].$$

If u_1 is multivariate, a similar approximation can be obtained by

$$E_{u_1}[g_l(u_1)|y_1^i] \approx \Big(\frac{|D_l|}{|D|}\Big)^{1/2} \exp\big[q_l(\hat{u}_{1_l}) - Q(\hat{u}_1)\big],$$

where $\hat{u}_{1_l}$ and $\hat{u}_1$ maximize $q_l(u_1)$ and $Q(u_1)$ respectively, D_l is the negative of the inverse of the Hessian of $q_l(u_1)$ evaluated at $\hat{u}_{1_l}$ and D is the negative of the inverse of the Hessian of $Q(u_1)$ evaluated at $\hat{u}_1$.

Applying this approximation to $E_{u_1}(p_{1sj}|y_1^i)$, where $1 \le s \le S$ and $1 \le j \le J$, we have

$$\begin{aligned} E_{u_1}(p_{1sj}|y_1^i) &= \frac{\int p_{1sj} P_b(Y_1 = y_1^i|u_1) f_\sigma(u_1) \mathrm{d}u_1}{\int P_b(Y_1 = y_1^i|u_1) f_\sigma(u_1) \mathrm{d}u_1} \\ &\approx \Big(\frac{|H_{sj}|}{|H|}\Big)^{1/2} \frac{p_{1sj} P_b(Y_1 = y_1^i|u_1) f_\sigma(u_1)|_{u_1=\hat{u}_{1_{sj}}}}{P_b(Y_1 = y_1^i|u_1) f_\sigma(u_1)|_{u_1=\hat{u}_1}}, \end{aligned} \tag{6}$$

where $\hat{u}_{1_{sj}}$ maximizes $\log p_{1sj} + \log P_b(Y_1 = y_1^i|u_1) + \log f_\sigma(u_1)$, $\hat{u}_1$ maximizes $\log P_b(Y_1 = y_1^i|u_1) + \log f_\sigma(u_1)$,

$$\begin{aligned} H_{sj} &= -\Big(\frac{\partial}{\partial u_1}\frac{\partial}{\partial u_1'}\big[\log p_{1sj} + \log P_b(Y_1 = y_1^i|u_1) + \log f_\sigma(u_1)\big]\Big)^{-1}|_{u_1=\hat{u}_{1_{sj}}} \\ &= -(-X_{1s}'\Delta_{1s}X_{1s} - X_1'\Delta_1 X_1 - \Sigma^{-1})^{-1}|_{u_1=\hat{u}_{1_{sj}}}, \end{aligned}$$

where $X_{1s} = (x_{1s1}, x_{1s2}, \ldots, x_{1sJ})'$, and

$$\begin{aligned} H &= -\Big(\frac{\partial}{\partial u_1}\frac{\partial}{\partial u_1'}\big[\log P_b(Y_1 = y_1^i|u_1) + \log f_\sigma(u_1)\big]\Big)^{-1}|_{u_1=\hat{u}_1} \\ &= -(-X_1'\Delta_1 X_1 - \Sigma^{-1})^{-1}|_{u_1=\hat{u}_1}. \end{aligned}$$

The expressions are derived in Appendix 6.2. The previous approximation only applies to a positive function $g_l(u)$, but the elements of $(\frac{u_{11}^2}{\sigma_1^3}, \ldots, \frac{u_{1k}^2}{\sigma_k^3})$ could be zero. Tierney et al. (1989) suggest to add a large constant c to $g_l(u_1)$, so that $g_l(u_1) + c$ is a positive function. Applying this procedure to $E\big((\frac{u_{1j}^2}{\sigma_j^3})|y_1^i\big)$, where $1 \le j \le k$, we get

$$\begin{aligned} E\Big(\frac{u_{1j}^2}{\sigma_j^3}|y_1^i\Big) &= E\Big(\frac{u_{1j}^2}{\sigma_j^3} + c|y_1^i\Big) - c \\ &\approx \Big(\frac{|H_j|}{|H|}\Big)^{1/2} \frac{\frac{u_{1j}^2 + c\sigma_j^3}{\sigma_j^3} P_b(Y_1 = y_1^i|u_1) \log f_\sigma(u_1)|_{u_1=\hat{u}_{1_j}}}{P_b(Y_1 = y_1^i|u_1) f_\sigma(u_1)|_{u_1=\hat{u}_1}} - c, \end{aligned} \tag{7}$$

where $\hat{u}_{1_j}$ maximizes $\log(\frac{u_{1j}^2+c\sigma_j^3}{\sigma_j^3}) + \log P_b(Y_1 = y_1^i|u_1) + \log f_\sigma(u_1)$ and

$$\begin{aligned} H_j &= -\Big(\frac{\partial}{\partial u_1}\frac{\partial}{\partial u_1'}\big[\log(\frac{u_{1j}^2+c\sigma_j^3}{\sigma_j^3}) + \log P_b(Y_1 = y_1^i|u_1) + \log f_\sigma(u_1)\big]\Big)^{-1}|_{u_1=\hat{u}_{1_j}} \\ &= -\left(\frac{2(c\sigma_j^3 - u_{1j}^2)}{(u_{1j}^2+c\sigma_j^3)^2}e_j e_j' - X_1'\Delta_1 X_1 - \Sigma^{-1}\right)^{-1}|_{u_1=\hat{u}_{1_j}}. \end{aligned}$$

The above expressions are also derived in Appendix 6.2.

The Laplace approximation for $E_{u_1}(g(u_1)|y_1^i)$ should run faster than the Monte Carlo method, since optimization usually requires less computation than sampling. In addition, we do not have to decide the sample size of u_1 as in the Monte Carlo method, which is good since we also need to decide the sample size of Y_1.

3.2 Approximation Using Samples from the Joint Distribution

Previously, a sample from the marginal distribution of Y_1 is used and we discuss several methods to approximate posterior means with respect to u_1 given the sample of Y_1. In the second approach, a sample of size n_{yu} from the joint distribution of (Y_1, u_1) is used. The method to take samples from the joint distribution has been described in Subsection 3.1. We denote the joint sample as $(y_1^i, u_1^i),\ 1 \le i \le n_{yu}$.

Suppose there are M unique vectors of y_1 in the joint sample, and denote these by $z_1^1, \ldots, z_1^M$. Then, $E_{u_1}\big(g(u_1)|Y_1 = z_1^m\big)$, $1 \le m \le M$, is approximated by

$$\frac{\sum_{\{i:y_1^i=z_1^m\}} g(u_1^i)}{\#\{i : y_1^i = z_1^m\}},$$

where $\{i : y_1^i = z_1^m\}$ is a set of integers at which y_1^i is equal to z_1^m and $\#\{i : y_1^i = z_1^m\}$ is the number of elements in this set. Next, $E_{Y_1}\big[E_{u_1}(g(u_1)|y_1^i)E_{u_1}(h(u_1)'|y_1^i)\big]$ is approximated by

$$\sum_{j=1}^{M} \frac{\sum_{\{i:y_1^i=z_1^m\}} g(u_1^i)}{\#\{i : y_1^i = z_1^m\}} \frac{\sum_{\{i:y_1^i=z_1^m\}} h(u_1^i)'}{\#\{i : y_1^i = z_1^m\}} \frac{\#\{i : y_1^i = z_1^m\}}{n_{yu}}.$$

In Subsection 3.1, when we use importance sampling, the same sample size of n_u is used for every given y_1^i. Here, when we use the joint sampling, the sample size of u_1 for a given y_1^i is determined from the joint sample. Hence, the sample size of u_1 can be adjusted as needed. Also, we only need to decide the sample size n_{yu} for the joint sample.

4 Simulation

In Section 3, we discuss three methods to approximate the information matrix: importance sampling, Laplace approximation and joint sampling. In this section, we will compare the three methods in simulations.

We consider a case where 2 attributes of 3 levels are of interest and a design with 9 choice sets of size 2 is used for all the respondents. The number of choice sets and the number of alternatives in each choice set cannot be large due to cognitive constraints. We use $3^2/2/9$ to denote this choice design, while other choice designs considered are $3^2/3/6$, $3^2/4/5$ and $3^2/5/4$.

We use effects-type coding for the attributes (Hensher, Rose and Greene (2005)). For example, if the coefficients of the first two levels of an attribute are given by $(\beta_1, \beta_2)'$, where the attribute has 3 levels, then the coefficient of the third level is $-\beta_1-\beta_2$. With effects-type coding, the sum of coefficients for an attribute is zero and the coefficient of each level can be interpreted as its effect relative to the average effect of the attribute, which is zero. Hence, two independent parameters are needed for an attribute of three levels. Here, with effects-type coding, the three levels of an attribute are coded as $(1, 0)$, $(0, 1)$ and $(-1, -1)$. Then, the distribution of random effects is $N_4(b, \Sigma)$, where $b = (b_1, b_2, b_3, b_4)'$ and $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2)$. The unknown parameter vector is $\theta = (b', \sigma')'$, where $\sigma = (\sigma_1, \sigma_2, \sigma_3, \sigma_4)'$. Following Arora and Huber (2001), Toubia et al. (2004) and Yu et al. (2011), values of the parameters are varied in terms of response accuracy and respondent heterogeneity. We take $b = (a, 0, a, 0)'$, where $a = .5$ is used to represent low response accuracy and $a = 3$ is used to represent high response accuracy. With this specification, it is implied that the mean for the third level is $-a$ for each attribute. Arora and Huber (2001) state that it is more meaningful to select the variance relative to the mean. As in Toubia et al. (2004), we take $\sigma = (\sqrt{3a}, \sqrt{3a}, \sqrt{3a}, \sqrt{3a})'$ in the case of high respondent heterogeneity and $\sigma = (\sqrt{0.5a}, \sqrt{0.5a}, \sqrt{0.5a}, \sqrt{0.5a})'$ in the case of low respondent heterogeneity. Thus, the 4 sets of parameter values used in our simulations are (a) high accuracy and high heterogeneity: $b = (3, 0, 3, 0)'$ and $\sigma = (3, 3, 3, 3)'$, (b) high accuracy and low heterogeneity: $b = (3, 0, 3, 0)'$ and $\sigma = (\sqrt{1.5}, \sqrt{1.5}, \sqrt{1.5}, \sqrt{1.5})'$, (c) low accuracy and high heterogeneity: $b = (0.5, 0, 0.5, 0)'$ and $\sigma = (\sqrt{1.5}, \sqrt{1.5}, \sqrt{1.5}, \sqrt{1.5})'$, and (d) low accuracy and low heterogeneity: $b = (0.5, 0, 0.5, 0)'$ and $\sigma = (0.5, 0.5, 0.5, 0.5)'$.

We are only interested in finding good designs, so the (dis)similarities of the three methods are compared on good designs. For a choice design with given values of the parameters, we handpick 100 good designs and approximate the information matrices for these designs using the three methods. The 100 designs are good designs from a computer search (We use a coordinate exchange algorithm with the Laplace approximation, A-optimality, and a sample size of $n_y = 10000$. The setting of the coordinate exchange algorithm is chosen based on preliminary simulation results.).

In the simulation, we use large sample sizes for the three methods so that the approximated values have stabilized and would have very small variation. For importance sampling, if there are 9 choice sets of size 2, there are $2^9 = 512$ possible values for Y. Since 512 is not a large number in this context, instead of taking a sample of Y, we use all possible values of Y with $n_u = 10^6$ in the simulation. We can also use all possible values of Y in the other cases (6 choice sets of size 3, 5 choice sets of size 4 and 4 choice sets of size 5). For joint sampling, we use $n_{yu} = 10^6$. For the Laplace method, we use $n_y = 10^6$. Importance sampling is considered to be the most accurate method because we use all possible values for Y and use 10^6 as the sample size for u.

Importance sampling and joint sampling are Monte Carlo methods, so the simulated information matrices will converge to the information matrices if the corresponding sample sizes (n_y and n_u for importance sampling and n_{yu} for joint sampling) go to infinity. Since the Laplace approximation is a combination of Monte Carlo method and Laplace's method, the simulated information matrices will not converge to the information matrices, but to the approximations of the information matrices, when the sample size (n_y for the Laplace method) goes to infinity. Our eventual goal is to find optimal designs, and not the actual values of the information matrices. Thus, we only want to see whether the three methods can rank the designs similarly.

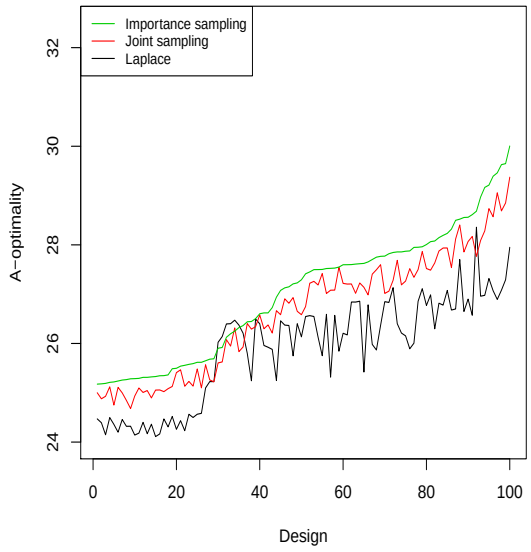
Figures 1 to 4 show the comparisons of the three methods for $3^2/2/9$ and $3^2/5/4$. The figures for $3^2/3/6$ and $3^2/4/5$ are similar, so they are not shown here. The 100 designs are ordered by the values from importance sampling and the x-axis gives the order of the designs. We can see that values from importance sampling and joint sampling are very close. Although values from the Laplace approximation are different from values from the other two methods, the patterns are similar. The three methods largely agree in ordering those 100 good designs.

Another way to assess agreement between the three methods is by studying pairwise correlations of values for a given criterion for the 100 designs. The scatter plot of values from any two of the methods shows that there is a linear pattern. The closer the scatter plot resembles a straight line, the more the two methods would agree in ordering the designs. Correlations depend on the 100 designs used here, since it is more difficult to get high correlations when the designs are similar. Hence, the correlation cannot be used as a useful measure of how the three methods agree. Table 1 shows the correlations between any two of the methods. We see that the correlations between importance sampling and joint sampling are larger than 0.9 in all cases. When the accuracy is high and heterogeneity is high, the correlations between the Laplace method and the other two methods are lower, except for $3^2/5/4$ with A-optimality. When the accuracy is high and the heterogeneity is low, the correlations between the Laplace method and the other two methods are lower, which are around 0.8, in $3^2/2/9$, $3^2/3/6$ and $3^2/4/5$ and all with A-optimality. For these two sets of parameter values, the correlations between the Laplace method and the other two methods are larger in $3^2/5/4$ than in $3^2/2/9$.

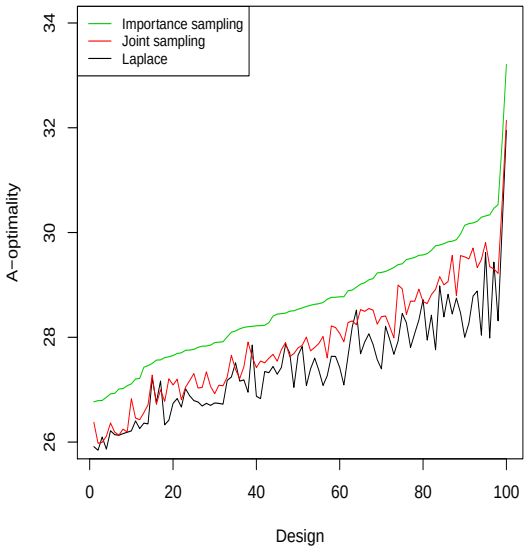
For the other two sets of parameter values, the correlations between the Laplace method and the other two methods are higher than 0.90. This table is consistent with what we observe in Figures 1 to 4.

In order to use the three methods in practice, we need to find appropriate sample sizes for the methods. For each method, relative differences are used to show how values change with sample sizes. We will use the $3^2/5/4$ case with $b = (3, 0, 3, 0)'$ and $\sigma = (3, 3, 3, 3)'$ as an example for illustration. For importance sampling, sample sizes considered are 5000, 10000, ..., 40000. For joint sampling, sample sizes considered are 50000, 100000, ..., 400000. For the Laplace approximation, sample sizes considered are 1000, 2000, ..., 9000. For each method, the relative differences between values from a small sample size and values from the largest sample size (which were also used in the previous simulation, i.e., all possible values of Y with $n_u = 10^6$ for importance sampling, $n_{yu} = 10^6$ for joint sampling and $n_y = 10^6$ for the Laplace method) are calculated. Figure 5 shows the relative differences of values in A-optimality and D-optimality for the three methods for 100 designs. The 100 designs are the same as those used previously for the $3^2/5/4$ case with $b = (3, 0, 3, 0)'$ and $\sigma = (3, 3, 3, 3)'$. We conclude that it suffices to take $n_u = 20000$ for importance sampling, $n_{yu} = 250000$ for joint sampling and $n_y = 3000$ for the Laplace approximation. After these sample sizes, the improvements in the mean and variance of the relative differences become smaller as sample sizes increase. For the other cases, similar conclusions hold. Thus, we can use $n_u = 20000$ for importance sampling, $n_{yu} = 250000$ for joint sampling and $n_y = 3000$ for the Laplace approximation for all the cases considered.

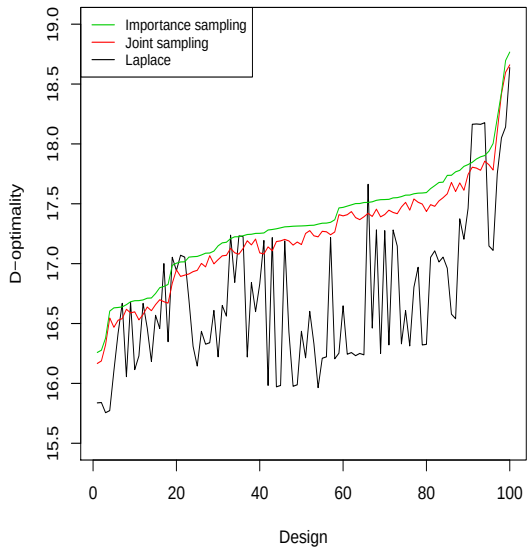
Table 2 shows the running time that the three methods take to approximate the information matrices for 100 designs with the reduced sample sizes. We can see that the Laplace approximation is about 3 times faster than importance sampling and 10 times faster than joint sampling. Note that here all possible values of Y are used for importance sampling. When this is not possible, we need to sample Y, making importance sampling slower, and the advantage of the Laplace approximation in running time will be larger. Another advantage of the Laplace approximation is that only the sample size of Y needs to be decided. For importance sampling with a large number of possible Y values, sample sizes of Y and u are varied simultaneously to find the appropriate ones. For joint sampling, n_{yu} is often much larger than n_y for the Laplace approximation, so it takes more time to find the appropriate sample size. For a given choice experiment, we can see that the time of joint sampling changes with the values of the parameters. The time is shorter for the cases with high response accuracy. In these cases, the mass of Y concentrates on a small proportion of possible values of Y. The algorithm that counts the unique values of Y in the joint sample runs faster when the mass of Y concentrates on a small proportion of possible values of Y than when it is more evenly distributed over possible values of Y.



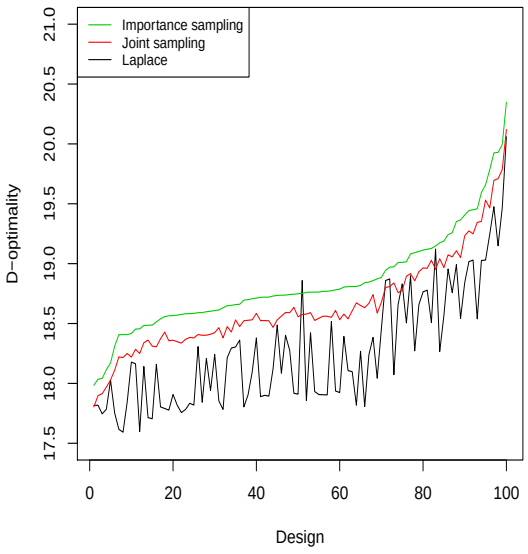
(a) 9 choice sets of size 2



(b) 4 choice sets of size 5

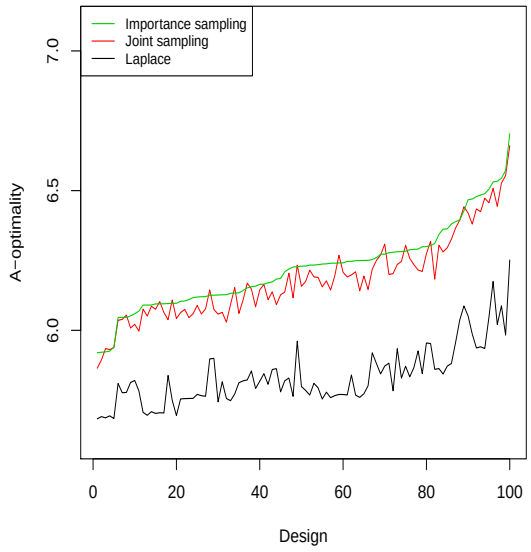


(c) 9 choice sets of size 2

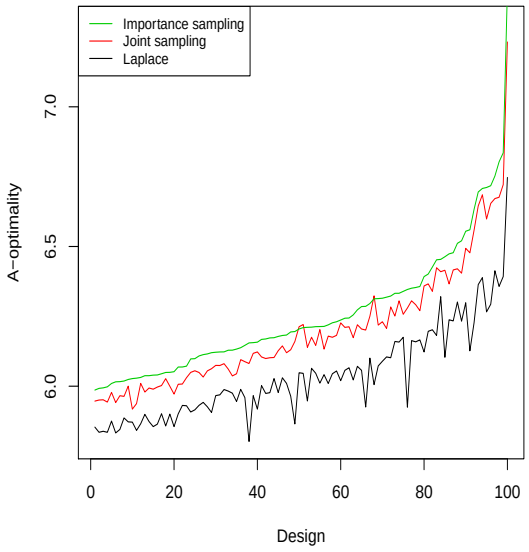


(d) 4 choice sets of size 5

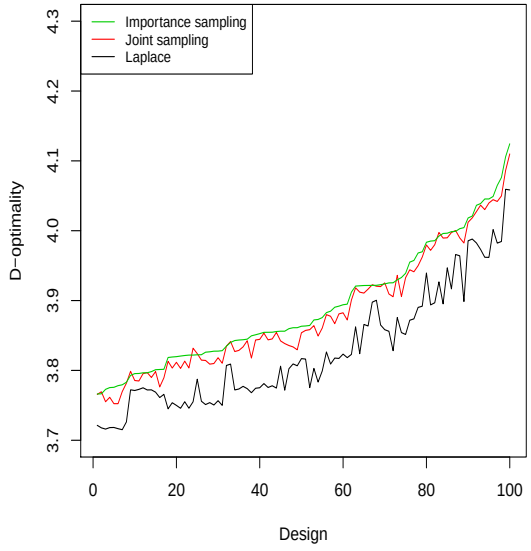
Figure 1: Comparisons of the three methods with A- and D-optimality when the response accuracy is high and the respondent heterogeneity is high.



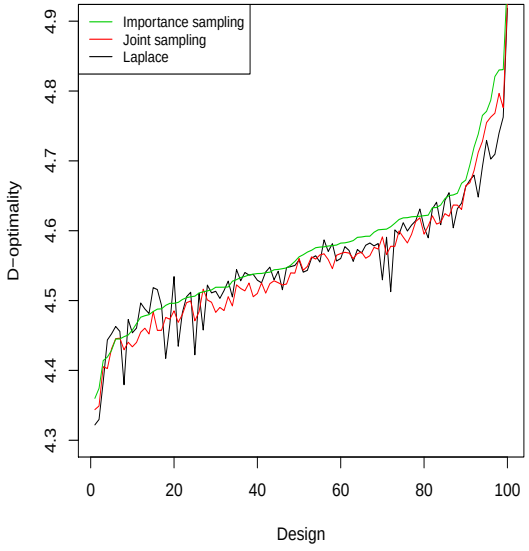
(a) 9 choice sets of size 2



(b) 4 choice sets of size 5

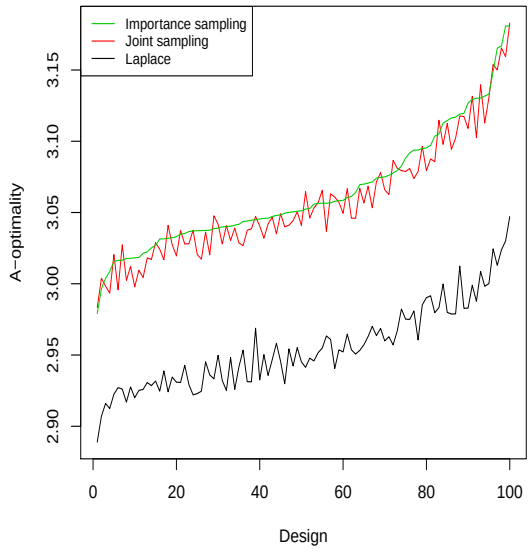


(c) 9 choice sets of size 2

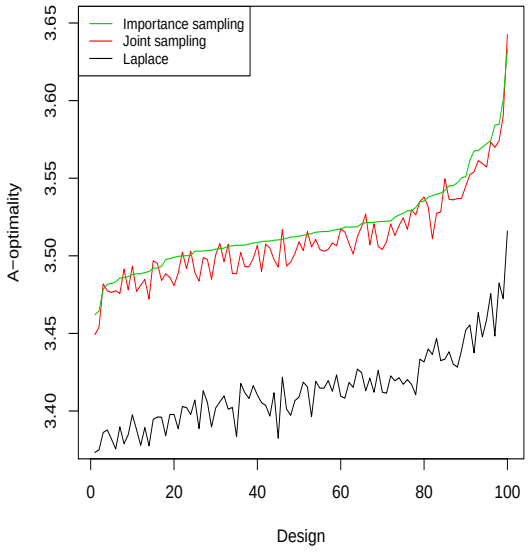


(d) 4 choice sets of size 5

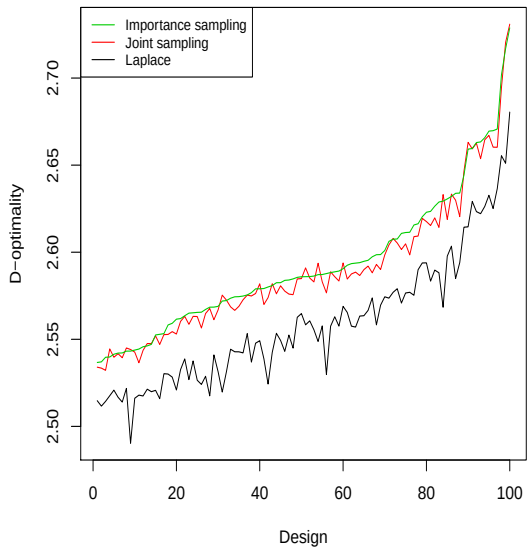
Figure 2: Comparisons of the three methods with A- and D-optimality when the response accuracy is high and the respondent heterogeneity is low.



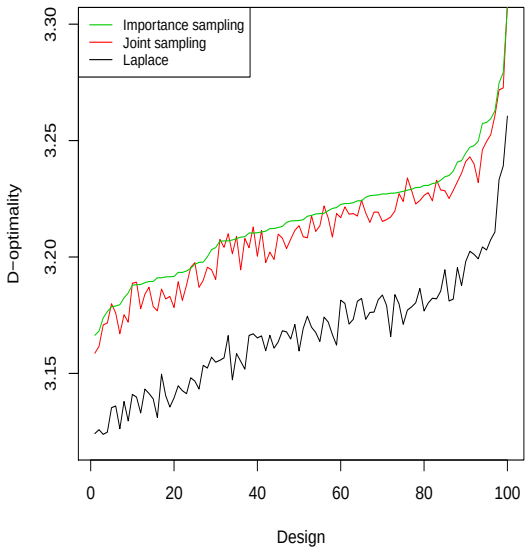
(a) 9 choice sets of size 2



(b) 4 choice sets of size 5

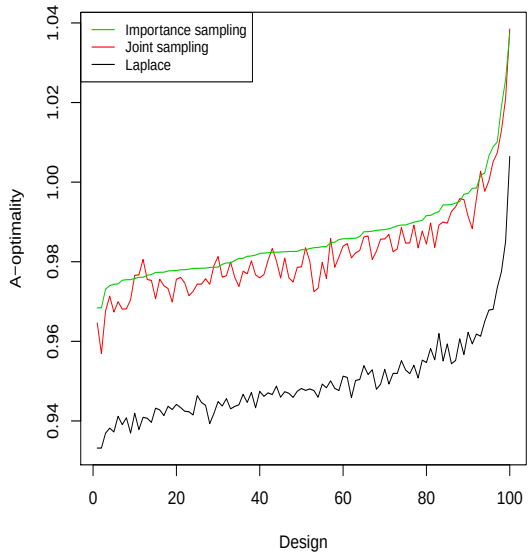


(c) 9 choice sets of size 2

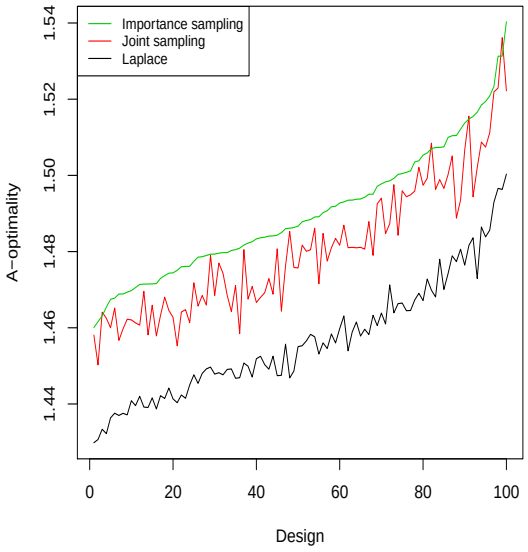


(d) 4 choice sets of size 5

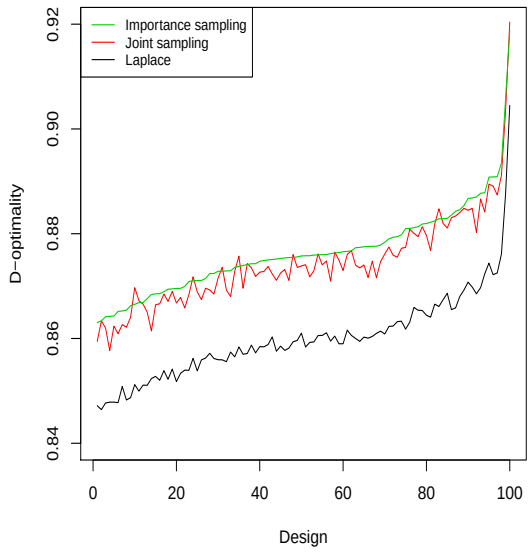
Figure 3: Comparisons of the three methods with A- and D-optimality when the response accuracy is low and the respondent heterogeneity is high.



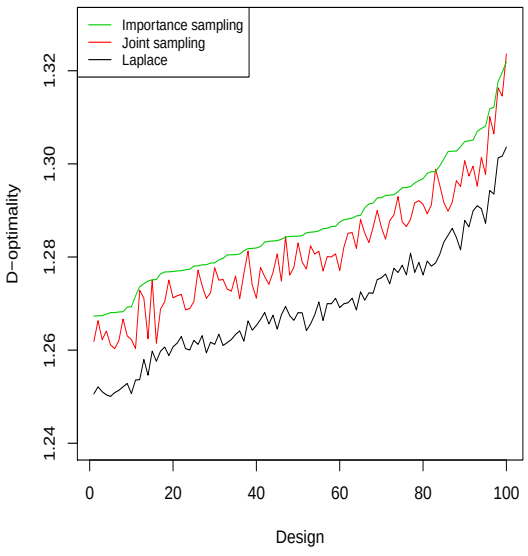
(a) 9 choice sets of size 2



(b) 4 choice sets of size 5



(c) 9 choice sets of size 2



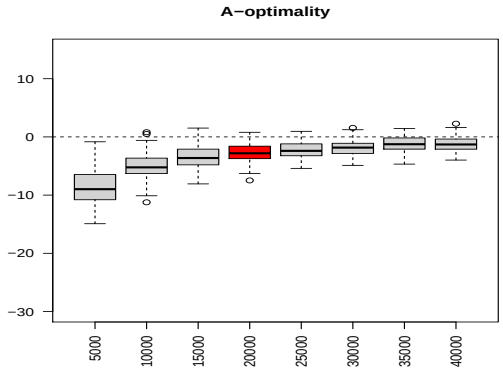
(d) 4 choice sets of size 5

Figure 4: Comparisons of the three methods with A- and D-optimality when the response accuracy is low and the respondent heterogeneity is low.

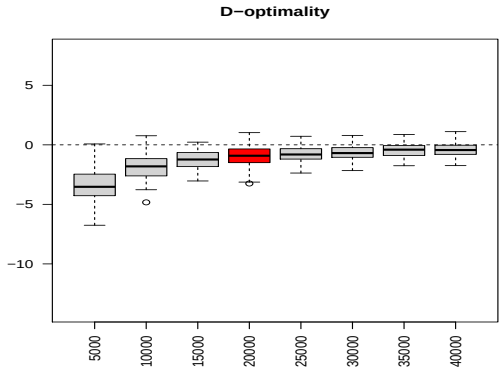
Table 1: Correlations between the three methods

		9 choice sets of size 2				6 choice sets of size 3			
		hh	hl	lh	ll	hh	hl	lh	ll
A-optimality	Importance -Joint	0.98	0.97	0.98	0.97	0.97	0.99	0.98	0.98
	Importance -Laplace	0.88	0.84	0.96	0.98	0.89	0.80	0.98	0.99
	Joint -Laplace	0.88	0.86	0.94	0.96	0.87	0.80	0.97	0.98
D-optimality	Importance -Joint	≈ 1	≈ 1	0.99	0.98	0.99	≈ 1	0.99	0.99
	Importance -Laplace	0.64	0.97	0.98	0.99	0.84	0.96	0.99	≈ 1
	Joint -Laplace	0.64	0.98	0.97	0.98	0.86	0.96	0.99	0.99
		5 choice sets of size 4				4 choice sets of size 5			
		hh	hl	lh	ll	hh	hl	lh	ll
A-optimality	Importance -Joint	0.99	0.99	0.97	0.96	0.98	0.99	0.97	0.95
	Importance -Laplace	0.88	0.78	0.98	0.99	0.94	0.95	0.94	0.99
	Joint -Laplace	0.88	0.80	0.97	0.95	0.94	0.95	0.93	0.94
D-optimality	Importance -Joint	≈ 1	≈ 1	0.99	0.97	≈ 1	≈ 1	0.98	0.97
	Importance -Laplace	0.86	0.94	0.99	0.99	0.86	0.96	0.97	0.99
	Joint -Laplace	0.86	0.95	0.98	0.97	0.86	0.96	0.97	0.97

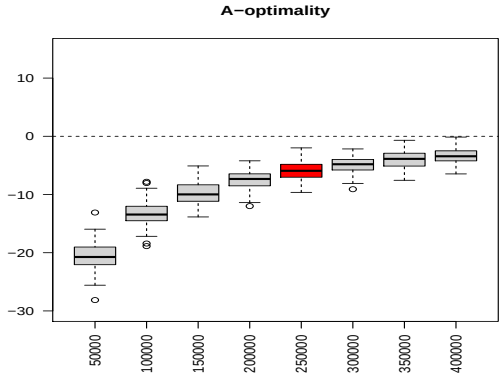
Note: hh represents high accuracy and high heterogeneity ($b = (3, 0, 3, 0)'$ and $\sigma = (3, 3, 3, 3)'$), hl represents high accuracy and low heterogeneity ($b = (3, 0, 3, 0)'$ and $\sigma = (\sqrt{1.5}, \sqrt{1.5}, \sqrt{1.5}, \sqrt{1.5})'$), lh represents low accuracy and high heterogeneity ($b = (0.5, 0, 0.5, 0)'$ and $\sigma = (\sqrt{1.5}, \sqrt{1.5}, \sqrt{1.5}, \sqrt{1.5})'$) and ll represents low accuracy and low heterogeneity ($b = (0.5, 0, 0.5, 0)'$ and $\sigma = (0.5, 0.5, 0.5, 0.5)'$).



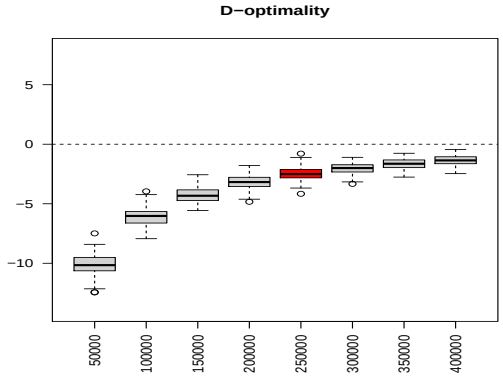
(a) Importance sampling



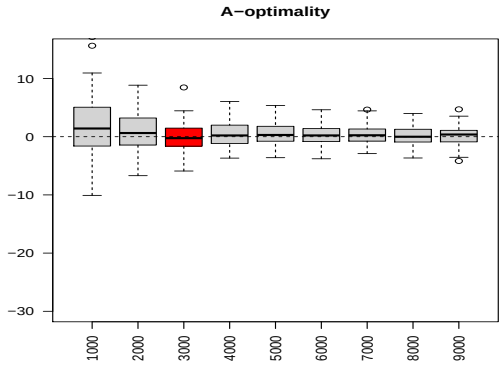
(b) Importance sampling



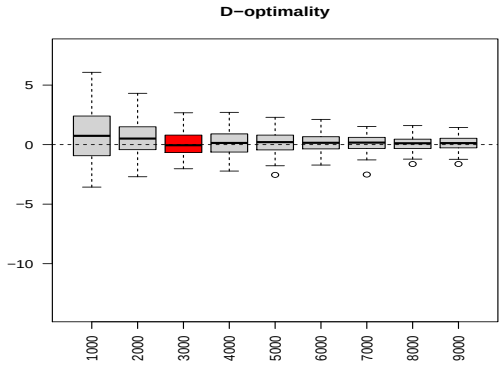
(c) Joint sampling



(d) Joint sampling



(e) The Laplace method



(f) The Laplace method

Figure 5: Relative difference (in %) between values from a sample size on the x-axis and the values from the largest sample size for the $3^2/5/4$ case with $b = (-3, 0, -3, 0)'$ and $\sigma = (3, 3, 3, 3)'$.

Table 2: Time for evaluating 100 designs using the three methods

	9 choice sets of size 2				6 choice sets of size 3			
	hh	hl	lh	ll	hh	hl	lh	ll
Importance, $n_u = 20000$	42m	43m	42m	42m	65m	60m	64m	61m
Joint, $n_{yu} = 250000$	169m	172m	298m	361m	172m	187m	417m	506m
Laplace, $n_y = 3000$	20m	20m	21m	27m	26m	26m	27m	29m
	5 choice sets of size 4				4 choice sets of size 5			
	hh	hl	lh	ll	hh	hl	lh	ll
Importance, $n_u = 20000$	90m	94m	83m	81m	56m	50m	58m	57m
Joint, $n_{yu} = 250000$	209m	198m	499m	630m	173m	166m	405m	487m
Laplace, $n_y = 3000$	30m	32m	30m	35m	33m	35m	30m	32m

Note: hh represents high accuracy and high heterogeneity ($b = (3, 0, 3, 0)'$ and $\sigma = (3, 3, 3, 3)'$), hl represents high accuracy and low heterogeneity ($b = (3, 0, 3, 0)'$ and $\sigma = (\sqrt{1.5}, \sqrt{1.5}, \sqrt{1.5}, \sqrt{1.5})'$), lh represents low accuracy and high heterogeneity ($b = (0.5, 0, 0.5, 0)'$ and $\sigma = (\sqrt{1.5}, \sqrt{1.5}, \sqrt{1.5}, \sqrt{1.5})'$) and ll represents low accuracy and low heterogeneity ($b = (0.5, 0, 0.5, 0)'$ and $\sigma = (0.5, 0.5, 0.5, 0.5)'$).

5 Discussion and Conclusion

For the panel mixed logit model, the information matrix has a complex form and cannot be written in a closed-form expression. We propose three methods to approximate the information matrix: importance sampling, Laplace approximation and joint sampling. For importance sampling, a sample of Y and a sample of u are taken independently, so the sample sizes of the two samples can be changed separately to adjust the precision of the approximation. When the number of possible values for Y is not large, all possible values of Y can be used, which makes the method more efficient. For joint sampling, the sample size for the joint sample is varied to adjust the accuracy of the approximation. From the simulation results, the running time for joint sampling is much longer than for the other two methods. For the Laplace approximation, although it is not as accurate as the other two methods, it ranks designs similarly and is much faster than the other two methods. For finding optimal designs, this ordering is the most important thing. Moreover, when search algorithms are used to find efficient designs, the number of information matrices to be evaluated will be much greater than 100 considered in our simulation and the search algorithm can take days, so using an efficient method to evaluate the information matrix is very important. For larger choice designs, importance sampling and joint sampling may not

be practical and the Laplace approximation may be the only viable method to use. Another advantage of the Laplace approximation is that only the sample size of Y needs to be decided. It is easier and faster to get an appropriate sample size for the Laplace approximation.

6 Appendix

6.1 Information Matrix for Panel Mixed Logit Model

We will show the validity of the expressions for $\frac{\partial \log L_n}{\partial b}$ and $\frac{\partial \log L_n}{\partial \sigma}$ in (1) and (2). First,

$$
\begin{aligned}
\frac{\partial \log L_n}{\partial b} &= \frac{1}{P_\theta(Y_n = y_n)} \frac{\partial P_\theta(Y_n = y_n)}{\partial b} \\
&= \frac{1}{P_\theta(Y_n = y_n)} \frac{\partial \big(\int \prod_{s=1}^{S} \prod_{j=1}^{J} p_{nsj}^{y_{nsj}} f_\sigma(u_n)\, \mathrm{d}u_n \big)}{\partial b} \\
&= \frac{1}{P_\theta(Y_n = y_n)} \int \frac{\partial \big(\prod_{s=1}^{S} \prod_{j=1}^{J} p_{nsj}^{y_{nsj}} \big)}{\partial b} f_\sigma(u_n)\, \mathrm{d}u_n \\
&= \frac{1}{P_\theta(Y_n = y_n)} \int (\prod_{s=1}^{S} \prod_{j=1}^{J} p_{nsj}^{y_{nsj}}) (\sum_s \sum_j y_{nsj} \frac{\partial p_{nsj}}{\partial b}) f_\sigma(u_n)\, \mathrm{d}u_n \\
&= \frac{1}{P_\theta(Y_n = y_n)} \int (\prod_s \prod_j p_{nsj}^{y_{nsj}}) \Big(\sum_s \sum_j y_{nsj} (x_{nsj} - \sum_i p_{nsi} x_{nsi}) \Big) f_\sigma(u_n)\, \mathrm{d}u_n \\
&= \frac{1}{P_\theta(Y_n = y_n)} \int (\prod_s \prod_j p_{nsj}^{y_{nsj}}) \Big(\sum_s \sum_j y_{nsj} x_{nsj} - \sum_s \sum_j y_{nsj} (\sum_i p_{nsi} x_{nsi}) \Big) f_\sigma(u_n)\, \mathrm{d}u_n \\
&= \frac{1}{P_\theta(Y_n = y_n)} \int (\prod_s \prod_j p_{nsj}^{y_{nsj}}) \Big(\sum_s \sum_j y_{nsj} x_{nsj} - \sum_s \sum_j p_{nsj} x_{nsj} \Big) f_\sigma(u_n)\, \mathrm{d}u_n \\
&= \frac{1}{P_\theta(Y_n = y_n)} \int (\prod_s \prod_j p_{nsj}^{y_{nsj}}) (X_n' y_n - X_n' p_n) f_\sigma(u_n)\, \mathrm{d}u_n \\
&= \frac{1}{P_\theta(Y_n = y_n)} X_n' \Big(P_\theta(Y_n = y_n) y_n - \int (\prod_s \prod_j p_{nsj}^{y_{nsj}}) p_n f_\sigma(u_n)\, \mathrm{d}u_n \Big) \\
&= X_n' \Big(y_n - \frac{1}{P_\theta(Y_n = y_n)} \int (\prod_s \prod_j p_{nsj}^{y_{nsj}}) p_n f_\sigma(u_n)\, \mathrm{d}u_n \Big) \\
&= X_n' \big(y_n - E_{u_n}(p_n | y_n) \big),
\end{aligned}
$$

where p_n is defined in (1). For the second expression that is to be evaluated,

$$
\begin{aligned}
&\frac{\partial \log P_\theta(Y_n = y_n)}{\partial \sigma} = \frac{1}{P_\theta(Y_n = y_n)} \frac{\partial P_\theta(Y_n = y_n)}{\partial \sigma} \\
= \;& \frac{1}{P_\theta(Y_n = y_n)} \frac{\partial \big(\int \prod_{s=1}^{S} \prod_{j=1}^{J} p_{nsj}^{y_{nsj}} f_\sigma(u_n)\, \mathrm{d}u_n \big)}{\partial \sigma} \\
= \;& \frac{1}{P_\theta(Y_n = y_n)} \int \big(\prod_{s=1}^{S} \prod_{j=1}^{J} p_{nsj}^{y_{nsj}} \big) \frac{\partial f_\sigma(u_n)}{\sigma}\, \mathrm{d}u_n \\
= \;& \frac{1}{P_\theta(Y_n = y_n)} \int \big(\prod_{s} \prod_{j} p_{nsj}^{y_{nsj}} \big) \Big((2\pi)^{-k/2} (-\frac{1}{2}) |\Sigma|^{-3/2} \exp(-\frac{1}{2} u_n' \Sigma^{-1} u_n) \frac{\partial |\Sigma|}{\partial \sigma} \\
& + (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp(-\frac{1}{2} u_n' \Sigma^{-1} u_n) (-\frac{1}{2}) \frac{\partial (u_n' \Sigma^{-1} u_n)}{\partial \sigma} \Big)\, \mathrm{d}u_n \\
= \;& \frac{1}{P_\theta(Y_n = y_n)} \int \big(\prod_{s} \prod_{j} p_{nsj}^{y_{nsj}} \big) \Big(-\frac{1}{2} f_\sigma(u_n) \big[|\Sigma|^{-1} \frac{\partial |\Sigma|}{\partial \sigma} + \frac{\partial (u_n' \Sigma^{-1} u_n)}{\partial \sigma} \big] \Big)\, \mathrm{d}u_n \\
= \;& \frac{1}{P_\theta(Y_n = y_n)} \int \big(\prod_{s} \prod_{j} p_{nsj}^{y_{nsj}} \big) \Big(f_\sigma(u_n) \big[-(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_k})' + (\frac{u_{n1}^2}{\sigma_1^3}, \ldots, \frac{u_{nk}^2}{\sigma_k^3})' \big] \Big)\, \mathrm{d}u_n \\
= \;& -(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_k})' + E_{u_n} \big((\frac{u_{n1}^2}{\sigma_1^3}, \ldots, \frac{u_{nk}^2}{\sigma_k^3})' | y_n \big),
\end{aligned}
$$

where u_{ni} is the ith element of u_n, $1 \le i \le k$.

Using these partial derivatives, we can now get the expressions for $E_{Y_n}\Big((\frac{\partial \log L_n}{\partial b})(\frac{\partial \log L_n}{\partial b})' \Big)$, $E_{Y_n}\Big((\frac{\partial \log L_n}{\partial b})(\frac{\partial \log L_n}{\partial \sigma})' \Big)$ and $E_{Y_n}\Big((\frac{\partial \log L_n}{\partial \sigma})(\frac{\partial \log L_n}{\partial \sigma})' \Big)$.

First, for $E_{Y_n}\Big((\frac{\partial \log L_n}{\partial b})(\frac{\partial \log L_n}{\partial b})' \Big)$ we have

$$
\begin{aligned}
& E_{Y_n}\Big((\frac{\partial \log L_n}{\partial b})(\frac{\partial \log L_n}{\partial b})' \Big) \\
= \;& E_{Y_n}\Big(X_n' \big[Y_n - E_{u_n}(p_n | Y_n) \big] \big[Y_n' - E_{u_n}(p_n' | Y_n) \big] X_n \Big) \\
= \;& X_n' E_{Y_n}\Big(\big[Y_n - E_{u_n}(p_n | Y_n) \big] \big[Y_n' - E_{u_n}(p_n' | Y_n) \big] \Big) X_n \\
= \;& X_n' \Big(E_{Y_n}(Y_n Y_n') - E_{Y_n}\big[E_{u_n}(p_n | Y_n) Y_n' \big] - E_{Y_n}\big[Y_n E_{u_n}(p_n' | Y_n) \big] \\
& \qquad + E_{Y_n}\big[E_{u_n}(p_n | Y_n) E_{u_n}(p_n' | Y_n) \big] \Big) X_n.
\end{aligned}
$$

Theses expressions are now evaluated separately.

$$
\begin{aligned}
E_{Y_n}(Y_nY_n') &= E_{u_n}(E_{Y_n}(Y_nY_n'|u_n)) \\
&= E_{u_n}\big[E_{Y_n}(\begin{pmatrix} Y_{n1} \\ Y_{n2} \\ \vdots \\ Y_{nS} \end{pmatrix}\begin{pmatrix} Y_{n1}' & Y_{n2}' & \dots & Y_{nS}' \end{pmatrix}|u_n)\big] \\
&= E_{u_n}\begin{pmatrix} E_{Y_n}(Y_{n1}Y_{n1}'|u_n) & E_{Y_n}(Y_{n1}Y_{n2}'|u_n) & \dots & E_{Y_n}(Y_{n1}Y_{nS}'|u_n) \\ E_{Y_n}(Y_{n2}Y_{n1}'|u_n) & E_{Y_n}(Y_{n2}Y_{n2}'|u_n) & \dots & E_{Y_n}(Y_{n2}Y_{nS}'|u_n) \\ \dots & \dots & \dots & \dots \\ E_{Y_n}(Y_{nS}Y_{n1}'|u_n) & E_{Y_n}(Y_{nS}Y_{n2}'|u_n) & \dots & E_{Y_n}(Y_{nS}Y_{nS}'|u_n) \end{pmatrix} \\
&= E_{u_n}\begin{pmatrix} \mathrm{diag}(p_{n1}) & p_{n1}p_{n2}' & \dots & p_{n1}p_{nS}' \\ p_{n2}p_{n1}' & \mathrm{diag}(p_{n2}) & \dots & p_{n2}p_{nS}' \\ \dots & \dots & \dots & \dots \\ p_{nS}p_{n1}' & p_{nS}p_{n2}' & \dots & \mathrm{diag}(p_{nS}) \end{pmatrix},
\end{aligned}
$$

where p_{ns} is defined after (1). Next,

$$
\begin{aligned}
& E_{Y_n}\big[E_{u_n}(p_n|Y_n)Y_n'\big] \\
&= \sum_{y_n}\Big[\Big(\int p_n\frac{\prod_s\prod_j p_{nsj}^{y_{nsj}}}{P_\theta(Y_n=y_n)}f_\sigma(u_n)\,\mathrm{d}u_n\Big)y_n'P_\theta(Y_n=y_n)\Big] \\
&= \int p_n\sum_{y_n}\Big[\prod_s\prod_j p_{nsj}^{y_{nsj}}y_n'\Big]f_\sigma(u_n)\,\mathrm{d}u_n \\
&= \int p_np_n'f_\sigma(u_n)\,\mathrm{d}u_n \\
&= E_{u_n}(p_np_n').
\end{aligned}
$$

Let $\Delta_n = \mathrm{diag}(\Delta_{ns})$ with $\Delta_{ns} = \mathrm{diag}(p_{ns}) - p_{ns}p_{ns}'$. Then

$$
E_{u_n}(\Delta_n) = E_{Y_n}(y_ny_n') - E_{Y_n}\big[E_{u_n}(p_n|y_n)y_n'\big].
$$

Hence, we have

$$
\begin{aligned}
& E_{Y_n}\Big(\Big(\frac{\partial logL_n}{\partial b}\Big)\Big(\frac{\partial logL_n}{\partial b}\Big)'\Big) \\
&= X_n'\Big(E_{u_n}(\Delta_n) - E_{u_n}(p_np_n') + E_{Y_n}\big[E_{u_n}(p_n|Y_n)E_{u_n}(p_n'|Y_n)\big]\Big)X_n.
\end{aligned}
$$

Second, $E_{Y_n}\Big(\Big(\frac{\partial \log L_n}{\partial b}\Big)\Big(\frac{\partial \log L_n}{\partial \sigma}\Big)'\Big)$ can be written as

$$\begin{aligned}
& E_{Y_n}\Big(\Big(\frac{\partial \log L_n}{\partial b}\Big)\Big(\frac{\partial \log L_n}{\partial \sigma}\Big)'\Big) \\
= \;& E_{Y_n}\Big(X_n'\big[Y_n - E_{u_n}(p_n|Y_n)\big]\Big[-\Big(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_k}\Big) + E_{u_n}\Big(\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)|Y_n\Big)\Big]\Big) \\
= \;& -X_n' E_{Y_n}\big[Y_n - E_{u_n}(p_n|Y_n)\big]\Big(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_k}\Big) \\
& \quad + X_n' E_{Y_n}\Big(\big[Y_n - E_{u_n}(p_n|Y_n)\big]E\Big[\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)|Y_n\Big]\Big) \\
= \;& X_n' E_{Y_n}\Big(Y_n E_{u_n}\Big[\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)|Y_n\Big]\Big) \\
& \quad - X_n' E_{Y_n}\Big(E_{u_n}(p_n|Y_n)E_{u_n}\Big[\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)|Y_n\Big]\Big).
\end{aligned}$$

To evaluate the first of these, note that

$$\begin{aligned}
& E_{Y_n}\Big(Y_n E_{u_n}\Big[\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)|Y_n\Big]\Big) \\
= \;& \sum_{y_n}\Big(y_n\Big[\int\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)\frac{\prod_s\prod_j p_{nsj}^{y_{nsj}}}{P_\theta(Y_n=y_n)}f_\sigma(u_n)\,\mathrm{d}u_n\Big]P_\theta(Y_n=y_n)\Big) \\
= \;& \int\Big[\sum_{y_n}\big(\prod_s\prod_j p_{nsj}^{y_{nsj}}\big)y_n\Big]\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)f_\sigma(u_n)\,\mathrm{d}u_n \\
= \;& \int p_n\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)f_\sigma(u_n)\,\mathrm{d}u_n \\
= \;& E_{u_n}\Big[p_n\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)\Big].
\end{aligned}$$

Hence, we have

$$\begin{aligned}
& E_{Y_n}\Big(\Big(\frac{\partial \log L_n}{\partial b}\Big)\Big(\frac{\partial \log L_n}{\partial \sigma}\Big)'\Big) \\
= \;& X_n'\Big(E_{u_n}\Big[p_n\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)\Big] - E_{Y_n}\Big[E_{u_n}(p_n|Y_n)E_{u_n}\Big(\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)|Y_n\Big)\Big]\Big).
\end{aligned}$$

Last, $E_{Y_n}\left(\left(\frac{\partial \log L_n}{\partial \sigma}\right)\left(\frac{\partial \log L_n}{\partial \sigma}\right)'\right)$ can be written as

$$
\begin{aligned}
& E_{Y_n}\left(\left(\frac{\partial \log L_n}{\partial \sigma}\right)\left(\frac{\partial \log L_n}{\partial \sigma}\right)'\right) \\
&= E_{Y_n}\Bigg(\Big[-\Big(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_k}\Big)' + E_{u_n}\Big(\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)'|Y_n\Big)\Big] \\
&\qquad \cdot \Big[-\Big(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_k}\Big) + E_{u_n}\Big(\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)|Y_n\Big)\Big]\Bigg) \\
&= \Big(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_k}\Big)'\Big(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_k}\Big) - \Big(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_k}\Big)' E_{Y_n}\left[E_{u_n}\Big(\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)|Y_n\Big)\right] \\
&\quad - E_{Y_n}\left[E_{u_n}\Big(\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)'|Y_n\Big)\right]\Big(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_k}\Big) \\
&\quad + E_{Y_n}\left[E_{u_n}\Big(\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)'|Y_n\Big)E_{u_n}\Big(\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)|Y_n\Big)\right] \\
&= \Big(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_k}\Big)'\Big(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_k}\Big) - \Big(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_k}\Big)'\Big(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_k}\Big) - \Big(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_k}\Big)'\Big(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_k}\Big) \\
&\quad + E_{Y_n}\left[E_{u_n}\Big(\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)'|Y_n\Big)E_{u_n}\Big(\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)|Y_n\Big)\right] \\
&= -\Big(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_k}\Big)'\Big(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_k}\Big) \\
&\quad + E_{Y_n}\left[E_{u_n}\Big(\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)'|Y_n\Big)E_{u_n}\Big(\Big(\frac{u_{n1}^2}{\sigma_1^3},\ldots,\frac{u_{nk}^2}{\sigma_k^3}\Big)|Y_n\Big)\right].
\end{aligned}
$$

6.1.1 Information Matrix for General Σ

For general Σ, not necessarily a diagonal matrix, a normal random vector β_n can be written as $\beta_n = b + u_n$, where $u_n \sim N_k(0, \Sigma = \Gamma\Gamma')$ with Γ a lower triangular matrix. Let $\gamma = vec(\Gamma')$, the information matrix for $\theta = (b', \gamma')'$ is

$$
I(\theta|X) = \sum_{n=1}^{N} \begin{pmatrix} E_{Y_n}\Big(\big(\frac{\partial \log L_n}{\partial b}\big)\big(\frac{\partial \log L_n}{\partial b}\big)'\Big) & E_{Y_n}\Big(\big(\frac{\partial \log L_n}{\partial b}\big)\big(\frac{\partial \log L_n}{\partial \gamma}\big)'\Big) \\ E_{Y_n}\Big(\big(\frac{\partial \log L_n}{\partial \gamma}\big)\big(\frac{\partial \log L_n}{\partial b}\big)'\Big) & E_{Y_n}\Big(\big(\frac{\partial \log L_n}{\partial \gamma}\big)\big(\frac{\partial \log L_n}{\partial \gamma}\big)'\Big) \end{pmatrix},
$$

where $L_n = P_\theta(Y_n = y_n)$ is the likelihood function for respondent n and is given by

$$\begin{aligned} & P_\theta(Y_n = y_n) \\ = & \int P_b(Y_n = y_n|u_n) f_\gamma(u_n)\,\mathrm{d}u_n \\ = & \int \prod_{s=1}^{S}\prod_{j=1}^{J}\left(\frac{\exp(x'_{nsj}(b+u_n))}{\sum_{i=1}^{J}\exp(x'_{nsi}(b+u_n))}\right)^{y_{nsj}} (2\pi)^{-k/2}|\Sigma|^{-1/2}\exp(-\frac{1}{2}u'_n\Sigma^{-1}u_n)\,\mathrm{d}u_n. \end{aligned}$$

It can be shown that $\frac{\partial \log L_n}{\partial b}$ has the same expression as before. For $\frac{\partial \log L_n}{\partial \gamma}$, the derivation is the same as for $\frac{\partial \log L_n}{\partial \sigma}$ except that we cannot simplify the following expression further,

$$\frac{\partial \log L_n}{\partial \gamma} = \frac{1}{P_\theta(Y_n = y_n)}\int (\prod_{s=1}^{S}\prod_{j=1}^{J} p_{nsj}^{y_{nsj}})\Big(-\frac{1}{2}f_\gamma(u_n)\big[|\Sigma|^{-1}\frac{\partial|\Sigma|}{\partial\gamma} + \frac{\partial(u'_n\Sigma^{-1}u_n)}{\partial\gamma}\big]\Big)\,\mathrm{d}u_n.$$

6.2 Laplace Approximation

In (6), we have

$$\begin{aligned} E_{u_1}(p_{1sj}|y_1^i) &= \frac{\int p_{1sj}P_b(Y_1 = y_1^i|u_1)f_\sigma(u_1)\mathrm{d}u_1}{\int P_b(Y_1 = y_1^i|u_1)f_\sigma(u_1)\mathrm{d}u_1} \\ &= \frac{\int \exp\big[\log p_{1sj} + \log P_b(Y_1 = y_1^i|u_1) + \log f_\sigma(u_1)\big]\mathrm{d}u_1}{\int \exp\big[\log P_b(Y_1 = y_1^i|u_1) + \log f_\sigma(u_1)\big]\mathrm{d}u_1} \\ &\approx \Big(\frac{|H_{sj}|}{|H|}\Big)^{1/2}\frac{p_{1sj}P_b(Y_1 = y_1^i|u_1)f_\sigma(u_1)|_{u_1=\hat{u}_{1_{sj}}}}{P_b(Y_1 = y_1^i|u_1)f_\sigma(u_1)|_{u_1=\hat{u}_1}}, \end{aligned}$$

where $\hat{u}_{1_{sj}}$ maximizes $\log p_{1sj} + \log P_b(Y_1 = y_1^i|u_1) + \log f_\sigma(u_1)$, $\hat{u}_1$ maximizes $\log P_b(Y_1 = y_1^i|u_1) + \log f_\sigma(u_1)$,

$$\begin{aligned} H_{sj} &= -\Big(\frac{\partial}{\partial u_1}\frac{\partial}{\partial u_1'}\big[\log p_{1sj} + \log P_b(Y_1 = y_1^i|u_1) + \log f_\sigma(u_1)\big]\Big)^{-1}|_{u_1=\hat{u}_{1_{sj}}} \\ &= -\Big(\big[\frac{\partial}{\partial u_1}\frac{\partial}{\partial u_1'}\log p_{1sj} + \frac{\partial}{\partial u_1}\frac{\partial}{\partial u_1'}\log P_b(Y_1 = y_1^i|u_1) + \frac{\partial}{\partial u_1}\frac{\partial}{\partial u_1'}\log f_\sigma(u_1)\big]\Big)^{-1}|_{u_1=\hat{u}_{1_{sj}}} \\ &= -(-X'_{1s}\Delta_{1s}X_{1s} - X'_1\Delta_1 X_1 - \Sigma^{-1})^{-1}|_{u_1=\hat{u}_{1_{sj}}}, \end{aligned}$$

and

$$\begin{aligned} H &= -\Big(\frac{\partial}{\partial u_1}\frac{\partial}{\partial u_1'}\big[\log P_b(Y_1 = y_1^i|u_1) + \log f_\sigma(u_1)\big]\Big)^{-1}|_{u_1=\hat{u}_1} \\ &= -\Big(\big[\frac{\partial}{\partial u_1}\frac{\partial}{\partial u_1'}\log P_b(Y_1 = y_1^i|u_1) + \frac{\partial}{\partial u_1}\frac{\partial}{\partial u_1'}\log f_\sigma(u_1)\big]\Big)^{-1}|_{u_1=\hat{u}_1} \\ &= -(-X_1'\Delta_1 X_1 - \Sigma^{-1})^{-1}|_{u_1=\hat{u}_1}. \end{aligned}$$

The validity of these expressions follows because

$$\begin{aligned} \frac{\partial}{\partial u_1}(\log p_{1sj}) &= \frac{1}{p_{1sj}}(\frac{\partial p_{1sj}}{\partial u_1}) \\ &= \frac{1}{p_{1sj}}\frac{\partial}{\partial u_1}(\frac{\exp(x_{1sj}'(b+u_1))}{\sum_{i=1}^J \exp(x_{1si}'(b+u_1))}) \\ &= \frac{1}{p_{1sj}}(p_{1sj}x_{1sj} - p_{1sj}\sum_i p_{1si}x_{1si}) \\ &= x_{1sj} - \sum_i p_{1si}x_{1si}, \end{aligned}$$

so that

$$\begin{aligned} \frac{\partial}{\partial u_1}\frac{\partial}{\partial u_1'}\log p_{1sj} &= \frac{\partial}{\partial u_1}(x_{1sj}' - \sum_i p_{1si}x_{1si}') \\ &= -\sum_i(\frac{\partial}{\partial u_1}p_{1si})x_{1si}' \\ &= -\sum_i(p_{1si}x_{1si} - p_{1si}\sum_l p_{1sl}x_{1sl})x_{1si}' \\ &= -\sum_i p_{1si}x_{1si}x_{1si}' + (\sum_i p_{1si}x_{1si})(\sum_i p_{1si}x_{1si}') \\ &= -X_{1s}'\mathrm{diag}(p_{1s})X_{1s} + X_{1s}'p_{1s}p_{1s}'X_{1s} \\ &= -X_{1s}'\Delta_{1s}X_{1s}. \end{aligned}$$

Further,

$$
\begin{aligned}
\frac{\partial}{\partial u_1}\log P_b(Y_1=y_1^i|u_1) &= \frac{1}{P_b(Y_1=y_1^i|u_1)}\Big(\frac{\partial P_b(Y_1=y_1^i|u_1)}{\partial u_1}\Big)\\
&= \frac{1}{P_b(Y_1=y_1^i|u_1)}\frac{\partial}{\partial u_1}\Big(\prod_s\prod_j p_{1sj}^{y_{1sj}^i}\Big)\\
&= \frac{1}{P_b(Y_1=y_1^i|u_1)}\Big(\prod_s\prod_j p_{1sj}^{y_{1sj}^i}\Big)\Big(\sum_s\sum_j \frac{y_{1sj}^i}{p_{1sj}}\frac{\partial p_{1sj}}{\partial u_1}\Big)\\
&= \sum_s\sum_j \frac{y_{1sj}^i}{p_{1sj}}\Big(p_{1sj}x_{1sj}-p_{1sj}\sum_k p_{1sk}x_{1sk}\Big)\\
&= \sum_s\sum_j (y_{1sj}^i-p_{1sj})x_{1sj},
\end{aligned}
$$

so that

$$
\begin{aligned}
\frac{\partial}{\partial u_1}\frac{\partial}{\partial u_1'}\log P_b(Y_1=y_1^i|u_1) &= \frac{\partial}{\partial u_1}\Big(\sum_s\sum_j (y_{1sj}^i-p_{1sj})x_{1sj}'\Big)\\
&= -\sum_s\sum_j\Big(\frac{\partial}{\partial u_1}p_{1sj}\Big)x_{1sj}'\\
&= -\sum_s\sum_j (p_{1sj}x_{1sj}-p_{1sj}\sum_k p_{1sk}x_{1sk})x_{1sj}'\\
&= -\sum_s (X_{1s}'diag(p_{1s})X_{1s}-X_{1s}'p_{1s}p_{1s}'X_{1s})\\
&= -X_1'\Delta_1 X_1.
\end{aligned}
$$

In (7), we have

$$
\begin{aligned}
E\Big(\frac{u_{1j}^2}{\sigma_j^3}|y_1^i\Big) &= E\Big(\frac{u_{1j}^2}{\sigma_j^3}+c|y_1^i\Big)-c\\
&= \frac{\int \frac{u_{1j}^2+c\sigma_j^3}{\sigma_j^3}P_b(Y_1=y_1^i|u_1)f_\sigma(u_1)\mathrm{d}u_1}{\int P_b(Y_1=y_1^i|u_1)f_\sigma(u_1)\mathrm{d}u_1}-c\\
&= \frac{\int \exp\big[\log(\frac{u_{1j}^2+c\sigma_j^3}{\sigma_j^3})+\log P_b(Y_1=y_1^i|u_1)+\log f_\sigma(u_1)\big]\mathrm{d}u_1}{\int \exp\big[\log P_b(Y_1=y_1^i|u_1)+\log f_\sigma(u_1)\big]\mathrm{d}u_1}-c\\
&\approx \Big(\frac{|H_j|}{|H|}\Big)^{1/2}\frac{\frac{{u_{1j}}^2+c\sigma_j^3}{\sigma_j^3}P_b(Y_1=y_1^i|u_1)\log f_\sigma(u_1)|_{u_1=\hat{u}_{1_j}}}{P_b(Y_1=y_1^i|u_1)f_\sigma(u_1)|_{u_1=\hat{u}_1}}-c,
\end{aligned}
$$

where $\hat{u}_{1_j}$ maximizes $\log(\frac{u_{1j}^2+c\sigma_j^3}{\sigma_j^3}) + \log P_b(Y_1 = y_1^i|u_1) + \log f_\sigma(u_1)$ and

$$\begin{aligned} H_j &= -\Big(\frac{\partial}{\partial u_1}\frac{\partial}{\partial u_1'}\big[\log(\frac{u_{1j}^2+c\sigma_j^3}{\sigma_j^3}) + \log P_b(Y_1 = y_1^i|u_1) + \log f_\sigma(u_1)\big]\Big)^{-1}|_{u_1=\hat{u}_{1_j}} \\ &= -\left(\frac{2(c\sigma_j^3-u_{1j}^2)}{(u_{1j}^2+c\sigma_j^3)^2}e_je_j' - X_1'\Delta_1X_1 - \Sigma^{-1}\right)^{-1}|_{u_1=\hat{u}_{1_j}}. \end{aligned}$$

The validity of this expression follows because

$$\frac{\partial}{\partial u_1}\log\Big(\frac{u_{1j}^2+c\sigma_j^3}{\sigma_j^3}\Big) = \frac{2u_{1j}e_j}{u_{1j}^2+c\sigma_j^3},$$

and

$$\begin{aligned} &\frac{\partial}{\partial u_1}\frac{\partial}{\partial u_1'}\log\Big(\frac{u_{1j}^2+c\sigma_j^3}{\sigma_j^3}\Big) \\ =\ &\frac{\partial}{\partial u_1}\Big(\frac{2u_{1j}e_j'}{u_{1j}^2+c\sigma_j^3}\Big) \\ =\ &\frac{2e_je_j'}{u_{1j}^2+c\sigma_j^3} - \frac{2u_{1j}e_j}{(u_{1j}^2+c\sigma_j^3)^2}(2u_{1j}e_j') \\ =\ &\frac{2e_je_j'(u_{1j}^2+c\sigma_j^3) - 4{u_{1j}}^2e_je_j'}{(u_{1j}^2+c\sigma_j^3)^2} \\ =\ &\frac{2(c\sigma_j^3-u_{1j}^2)}{(u_{1j}^2+c\sigma_j^3)^2}e_je_j'. \end{aligned}$$

7 References

Arora, N., and Huber, J. (2001). Improving parameter estimates and model prediction by aggregate customization in choice experiments. *Journal of Consumer Research*, **28**(2), 273−283.

Atkinson, A. C., Donev, A. N., and Tobias, R. D. (2007). *Optimum experimental designs, with SAS* . New York: Oxford University Press.

Bhat, C. (1998). Accommodating variations in responsiveness to level-of-service variables in travel mode choice models. *Transportation Research A*, **32**, 455−507.

Bhat, C. (2000). Incorporating observed and unobserved heterogeneity in urban work mode choice modeling. *Transportation Science*, **34**, 228−238.

Bliemer, M. C., and Rose, J. M. (2010). Construction of experimental designs for mixed logit models allowing for correlation across choice observations. *Transportation Research Part B: Methodological*, **44**(6), 720−734.

Booth, J. G., and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 265−285.

Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**(421), 9−25.

Brownstone, D. and K. Train (1999). Forecasting new product penetration with flexible substitution patterns. *Journal of Econometrics*, **89**, 109−129.

Chernoff, H. (1953). Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, 586−602.

Erdem, T. (1996). A dynamic analysis of market structure based on panel data. *Marketing Science*, **15**, 359−378.

Hensher, D. A., Rose, J. M., and Greene, W. H. (2005). *Applied choice analysis: a primer*. Cambridge University Press.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**(437), 162−170.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *In: Zarembka P (ed), Frontiers in Econometrics*. Academic Press, New York, pp. 105−142.

McFadden, D. and K. Train (2000). Mixed MNL models of discrete response. *Journal of Applied Econometrics*, **15**, 447−470.

Moerbeek, M., and Maas, C. J. (2005). Optimal experimental designs for multilevel logistic models with two binary predictors. *Communications in Statistics− Theory and Methods*, **34**(5), 1151−1167.

Peter E. Rossi, Greg M. Allenby, and Robert McCulloch (2006). *Bayesian Statistics and Marketing*. John Wiley and Sons, Ltd.

Revelt, D. and K.Train (1998). Mixed logit with repeated choices: households' choices of appliance efficiency level. *Review of Economics and Statistics*, **80**(4), 647−657.

Sándor, Z., and Wedel, M. (2002). Profile construction in experimental choice designs for mixed logit models. *Marketing Science*, **21**(4), 455−475.

Tierney, L., and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**(393), 82−86.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, **84**(407), 710−716.

Toubia, O., Hauser, J. R., and Simester, D. I. (2004). Polyhedral methods for adaptive choice-based conjoint analysis. *Journal of Marketing Research*, **41**(1), 116−131.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Wand, M. P. (2007). Fisher information for generalized linear mixed models. *Journal of Multivariate Analysis*, **98**(7), 1412−1416.

Waite, T.W. and Woods, D.C. (2014) Designs for generalized linear models with random block effects via information matrix approximations. Southampton, GB, Southampton Statistical Sciences Research Institute, 21pp. (Southampton Statistical Sciences Research Institute Methodology Working Papers, M12/01).

Yu, J., Goos, P., and Vandebroek, M. (2011). Individually adapted sequential Bayesian conjoint-choice designs in the presence of consumer heterogeneity. *International Journal of Research in Marketing*, **28**(4), 378−388.