Spring 2019

A Computationally Efficient Method for Selecting a Split Questionnaire Design

Matthew Stuart

A Computationally Efficient Method for Selecting a Split Questionnaire Design

Matthew Stuart[1], Cindy Yu[1,*]

Department of Statistics
Iowa State University
Ames, IA 50011

Abstract

Split questionnaire design (SQD) is a relatively new survey tool to reduce response burden and increase the quality of responses. Among a set of possible SQD choices, a design is considered as the best if it leads to the least amount of information loss quantified by the Kullback-Leibler divergence (KLD) distance. However, the calculation of the KLD distance requires computation of the distribution function for the observed data after integrating out all the missing variables in a particular SQD. For a typical survey questionnaire with a large number of categorical variables, this computation can become practically infeasible. Motivated by the Horvitz-Thompson estimator, we propose an approach to approximate the distribution function of the observed in much reduced computation time and lose little valuable information when comparing different choices of SQDs. We contrive a thorough simulation study to test if the proposed approximation method can correctly identify the best SQD under several simulation scenarios created to cover different distribution shapes of continuous variables, and different correlation structures in both categorical and continuous variables. Finally, the proposed approach is applied to the 2012 Pet Demographic Survey data. Both of the simulation studies and the empirical study demonstrate that the proposed method is computationally efficient and can accurately select the best SQD design among a set of alternatives.

Keywords: Split Questionnaire Design, Survey Methodology, Survey Sampling

1. Introduction

Universities and corporations continue to use large surveys to make inference of parameters of interest for a population. Advancements in technology have allowed these surveys to become increasingly larger, meaning having more questions to be asked for survey participants, while still maintaining cost effectiveness for survey conductors. However, an increasing concern of these large surveys is that the questionnaire length often leads to response fatigue, hence resulting in a high non-response rate. In addition, those who do complete the survey may provide responses which

*Corresponding author
Email addresses: mstuart@iastate.edu (Matthew Stuart), cindyyu@iastate.edu (Cindy Yu)

are incomprehensible or form a pattern that does not make sense for the question at hand (Adams and Gale [1]).

Split questionnaire designs (SQDs), which have been used to help alleviate this concern, split a long questionnaire into different sets or blocks, and give a subset of these questions to different respondents in the survey. Typically, these designs have a number of core questions that will be answered by all respondents, and the remaining questions are split into blocks. Each block has a certain number of questions, with a particular respondent being asked a set of questions from each block. Thus, for a particular SQD, there are a certain number of patterns of question sets. By administering this type of questionnaire, response burden can be reduced and the quality of the response can improve. After partially answered questionnaires are collected, an imputation method can be applied to fill in the intentionally missing data, which increases the computation time in estimation. Analysis techniques, such as EM algorithms, have been implemented to reduce the computation time in these types of surveys. Raghunathan and Grizzle [16] and Adiguzel and Wedel [2] both proposed a multiple imputation method for the missing values by simulating responses from the posterior predictive distribution via a Gibbs sampler.

This article aims to solve a different problem with SQDs. For a large survey, there are usually several potential ways to split the long questionnaire based on prior knowledge or information. Then, a natural question is which SQD is the best among a set of potential SQDs predetermined by survey conductors? Chipperfield and Steel [6] and Chipperfield and Steel [7] discussed choosing the optimal design based on a function of the design constraints, the variance-covariance matrix of the estimators, and the cost of the complete survey and individual survey questions, assuming univariate and multivariate responses for each respective paper. They proposed choosing the design that either minimizes the variance of the parameter estimates for a fixed cost of the survey or minimizes the cost of the survey for a fixed variance of the parameter estimates. We consider choosing the design that minimizes the amount of information loss defined by Kullback-Leibler divergence (KLD) distance, which is discussed below.

Define the continuous variables as $\boldsymbol{X} = [X_1, \ldots, X_{d_x}]$, where d_x is the number of continuous variables, and the categorical variables as $\boldsymbol{Y} = [Y_1, \ldots, Y_{d_y}]$, where d_y is the number of categorical

Question Patterns	Core		Block 1				Block 2			
p	Y_1	Y_2	Y_3	X_1	Y_4	X_2	Y_5	X_3	Y_6	X_4
1	✓	✓	✓	✓			✓	✓		
2	✓	✓			✓	✓			✓	✓

Table 1: A display of question patterns in a SQD for the toy example of Section 1.

variables. For a particular question pattern in a particular design, let $\boldsymbol{X}_{obs}$ (or $\boldsymbol{Y}_{obs}$) denote the vector of continuous variables $\boldsymbol{X}$ (or categorical variables $\boldsymbol{Y}$) that are planned to be observed, and $\boldsymbol{X}_{mis}$ (or $\boldsymbol{Y}_{mis}$) denote the vector of continuous variable (or categorical variables) that are planned to be skipped. Consider the following toy example with $d_x = 4$ and $d_y = 6$. Based on prior research, either from past data, a pilot study, or literature review, we plan to have Y_1 and Y_2 as the core questions that should be administered to all respondents, and to split the remaining questions into two blocks of 4 questions, with 2 questions from each block assigned in a question pattern p, where $p = 1, \ldots, P$ (here $P = 2$). A visualization of this SQD is in Table 1. Thus $\boldsymbol{X}_{obs} = [X_1, X_3]$ and $\boldsymbol{Y}_{obs} = [Y_1, Y_2, Y_3, Y_5]$ for question pattern $p = 1$, and $\boldsymbol{X}_{obs} = [X_2, X_4]$ and $\boldsymbol{Y}_{obs} = [Y_1, Y_2, Y_4, Y_6]$ for question pattern $p = 2$. $\boldsymbol{X}_{mis}$ (or $\boldsymbol{Y}_{mis}$) is the complement set of $\boldsymbol{X}_{obs}$ (or $\boldsymbol{Y}_{obs}$). Raghunathan and Grizzle [16] theorized via simulation study that loss of efficiency is the least when variables observed in a pattern are not highly correlated. The simulation and empirical studies in this paper use this as motivation to split the questions into different patterns.

Given a set of possible SQDs in which each predetermined question pattern has the same number of questions, one very important question for survey conductors is which SQD is the best in terms of information loss due to reducing the number of questions? We propose to use the KLD distance as described in Kullback and Leibler [11] to quantify this loss. The KLD distance is usually used to measure how one probability distribution is different from a second reference probability distribution. In this article, we specifically consider the KLD distance between the distribution of complete data and the distribution of observed data under a SQD. We attempt to choose a SQD that gives the minimum value of the KLD distance, which is defined as

$$\begin{aligned} KLD =& \int f(\boldsymbol{X}, \boldsymbol{Y}) \ln \left(\frac{f(\boldsymbol{X}, \boldsymbol{Y})}{f(\boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs} \mid D)} \right) d\boldsymbol{X} d\boldsymbol{Y} \\ =& E_{\boldsymbol{X},\boldsymbol{Y}} \left[\ln f(\boldsymbol{X}, \boldsymbol{Y}) \right] - E_{\boldsymbol{X},\boldsymbol{Y}} \left[\ln f(\boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs} \mid D) \right], \end{aligned} \tag{1.1}$$

where $f(\boldsymbol{X}, \boldsymbol{Y})$ is the joint distribution of the complete data and $f(\boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs} \mid D)$ is the distribution of the observed data for a given design D. By comparing the KLD distances across different SQDs, we identify the best design that leads to the minimum amount of "information loss" when choosing to implement a SQD versus a complete survey. Since $E_{\boldsymbol{X},\boldsymbol{Y}}\left[\ln f(\boldsymbol{X}, \boldsymbol{Y})\right]$ is the same across different SQDs, the design that leads to the minimum value of the KLD is the design with the maximal value of $E_{\boldsymbol{X},\boldsymbol{Y}}\left[\ln f(\boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs} \mid D)\right]$, the second term in (1.1).

Denote an estimator for $E_{\boldsymbol{X},\boldsymbol{Y}}\left[\ln f(\boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs} \mid D)\right]$ by

$$\hat{E}_{\boldsymbol{X},\boldsymbol{Y}}\left[\ln f(\boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs} \mid D)\right] = \sum_{p=1}^{P} Prob(p)\hat{E}_{\boldsymbol{X},\boldsymbol{Y}}\left[\ln f(\boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs} \mid p)\right], \tag{1.2}$$

where P is the number of question patterns in a given design D and $Prob(p)$ is the probability of choosing pattern p within design D. For this kind of research, a small set of data from a pilot study or a past dataset from a longitudinal survey is typically assumed to be available to provide information that will guide us in identifying candidates of SQDs. If we assume this dataset has independent respondents and each question pattern has equal chance to be assigned to a respondent, then (1.2) can be approximated by

$$\hat{E}_{\boldsymbol{X},\boldsymbol{Y}}\left[\ln f(\boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs} \mid D)\right] = \frac{1}{n}\frac{1}{P}\sum_{i=1}^{n}\sum_{p=1}^{P}\left[\ln f(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_{i,obs} \mid p)\right], \tag{1.3}$$

where $\boldsymbol{X}_{i,obs}$ (or $\boldsymbol{Y}_{i,obs}$) is the vector containing observed responses for the continuous (or categorical) variables for respondent i in pattern p, and n is the number of respondents in the dataset. If the dataset comes with weights, the equal probability of $\frac{1}{n}$ is replaced by $\frac{w_i}{\sum_{i=1}^{n} w_i}$ inside the summation, where w_i is the weight associated with respondent i. Note that the distribution function $f(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_{i,obs} \mid p)$ in (1.3) is the marginal distribution of $(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_{i,obs})$ after integrating out $(\boldsymbol{X}_{i,mis}, \boldsymbol{Y}_{i,mis})$ over the ranges of $\boldsymbol{X}_{i,mis}$ and all possible cells of $\boldsymbol{Y}_{i,mis}$ in a particular pattern p of design D. In other words,

$$f(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_{i,obs} \mid p) = \int\int f(\boldsymbol{X}_i, \boldsymbol{Y}_i) d\boldsymbol{X}_{i,mis} d\boldsymbol{Y}_{i,mis}. \tag{1.4}$$

The challenge is that, when the number of missing categorical variables $\boldsymbol{Y}_{i,mis}$ increases (even moderately), the number of possible cells that respondent i can fall into becomes enormous. When this occurs, the computation time to calculate $f(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_{i,obs} \mid p)$ in (1.4), and thus the time to compute $\hat{E}_{\boldsymbol{X},\boldsymbol{Y}}\left[\ln f(\boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs} \mid D)\right]$ in (1.3) becomes exhaustive, resulting in valuable time wasted. As an illustration, let us assume that we want to conduct a survey with 45 categorical

variable questions, each of which has 5 levels. Prior experiences suggest us to split the questionnaire into a set of patterns with 25 categorical variable questions planned to be answered in each pattern (i.e. 20 questions are skipped). Then the number of possible cells that $\boldsymbol{Y}_{i,mis}$ can take in this question pattern is $5^{20} \approx 9.5 * 10^{13}$. For such large number of possible cells, it becomes extremely time consuming (if not infeasible) to calculate (1.4) and (1.3). The goal of this paper is to introduce a computationally efficient approach to approximate $f(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_{i,obs} \mid p)$ using the Horvitz-Thompson idea (Horvitz and Thompson [10]), while still accurately identifying the best SQD among a set of alternatives.

The rest of this paper is organized as follows. Section 2 discusses the distribution function of the observed data and introduces our proposed time efficient method to approximate it. Section 3 contrives several simulation studies to evaluate the performance of our proposed approximation method in correctly identifying the best SQD among a set of choices. Section 4 presents an application of our method to a real data from the 2012 Pet Demographic Survey (PDS) conducted by the American Veterinary Medical Association (AVMA). Section 5 concludes with some final remarks.

2. Distribution Function of the Observed and its Approximation in a Split Questionnaire Design

In this section, we will derive the distribution function of the observed data for any question pattern within a SQD in Section 2.1, and then propose a time efficient method to approximate the expected log-distribution function of the observed in Section 2.2.

2.1. Distribution Function of the Observed

We start by writing out the distribution function for the complete data. For any specified density of the continuous variables, the logarithm of the distribution function for the complete data (assuming independence) is

$$\ln f(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{n} \ln f(\boldsymbol{X}_i, \boldsymbol{Y}_i) = \sum_{i=1}^{n} \left[\ln f(\boldsymbol{X}_i \mid \boldsymbol{Y}_i = m) * \pi_{m,i}\right], \tag{2.1}$$

where $f(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ is the joint distribution function of $(\boldsymbol{X}, \boldsymbol{Y})$ for the i^{th} respondent, m is the index for the cell (defined by the categorical variables) that respondent i falls into, and $\pi_{m,i}$ is the

probability that respondent i falls into cell m. Here, $m = 1, \ldots, M$ where $M = \prod_{k=1}^{d_y} I_k$ and I_k is the number of levels that categorical variable Y_k can take. Next, we introduce the modelling set up for the conditional density of the continuous variables $\boldsymbol{X}$ given the categorical variable cell m, i.e. $f(\boldsymbol{X}_i \mid \boldsymbol{Y}_i = m)$ in (2.1), as well as the modelling set up for the cell probability π_m.

We consider two different specifications for $f(\boldsymbol{X}_i \mid \boldsymbol{Y}_i = m)$. First, we assume the continuous variable $\boldsymbol{X}$ given cell m follows a multivariate normal (MVN) distribution specified as follows,

$$\boldsymbol{X}_i \mid (\boldsymbol{Y}_i = m) \sim \text{MVN}(\boldsymbol{\mu}_{m,i} = \mathbf{B}\mathbf{Z}_i, \boldsymbol{\Omega}_i), \tag{2.2}$$

where $\mathbf{Z}_i$ is a $q \times 1$ vector of important categorical variables that determine the mean vector $\boldsymbol{\mu}_{m,i}$, $\mathbf{B}$ is the $d_x \times q$ matrix of coefficients associated with $\mathbf{Z}_i$, and $\boldsymbol{\Omega}_i$ is the variance-covariance matrix for respondent i. We also assume that $\boldsymbol{\Omega}_i = \boldsymbol{\Omega}$ for all $i = 1, \ldots, n$, and $q << M$ to achieve dimension reduction. An example of how to specify $\mathbf{Z}_i$ given important categorical variables is given in the simulation study of Section 3. The multivariate normal distribution is only useful if the continuous variables are not skewed. For skewed distributions, a multivariate gamma (MVG) distribution, defined in Furman [8], is more appropriate to model these types of data, which is our second specification for $f(\boldsymbol{X}_i \mid \boldsymbol{Y}_i = m)$. In the simulation and empirical studies, we explore both MVN and MVG assumptions and compare the results. The details of a MVG distribution can be found in Appendix A.

To model the cell probability π_m, a log-linear model, such as that in Goodman [9] and Bishop et al. [4], is used. For a dataset with d_y categorical variables, where each categorical variable Y_k ($k = 1, \ldots, d_y$) has I_k levels, the cell probability model is written as

$$\begin{aligned} \log\pi_m &= \log\pi_{(l_1,\ldots,l_{d_y})} \\ &= \alpha_0 + \sum_{k=1}^{d_y} \alpha_{k,l_k} + \sum_{k=1}^{d_y} \sum_{k'>k} \alpha_{kk',l_k l_{k'}} + \sum_{k=1}^{d_y} \sum_{k'>k} \sum_{k''>k'} \alpha_{kk'k'',l_k l_{k'} l_{k''}} + \ldots, \end{aligned} \tag{2.3}$$

where α_0 is the intercept, l_k is the level that respondent i takes in variable Y_k, α_{k,l_k} is the main effect for response l_k in variable Y_k, and the remaining $\boldsymbol{\alpha}$ terms are the two- and three-way interaction terms respectively. Constraints on the model coefficients in (2.3) must be set to achieve unique and identifiable results, and to avoid overparameterization of the model. The constraints used in this paper are $\sum_{l_k=1}^{I_k} \alpha_{k,l_k} = 0$ for any k, $\sum_{l_k \text{ or } l_{k'}} \alpha_{kk',l_k l_{k'}} = 0$ for any (k, k'), and $\sum_{l_k \text{ or } l_{k'} \text{ or } l_{k'}} \alpha_{kk'k'',l_k l_{k'} l_{k''}} = 0$ for any (k, k', k''). As an illustration, suppose $d_y = 3$ and π_m

$= \pi_{(l_1,l_2,l_3)}$ is the probability that respondent i has responses $Y_1 = l_1$, $Y_2 = l_2$, $Y_3 = l_3$ where $l_k \in \{1, 2, \ldots, I_k\}$. The log-linear model can be written as $\log\pi_m = \alpha_0 + \alpha_{1,l_1} + \alpha_{2,l_2} + \alpha_{3,l_3} + \alpha_{12,l_1l_2} + \alpha_{13,l_1l_3} + \alpha_{23,l_2l_3} + \alpha_{123,l_1l_2l_3}$.

Instead of modeling the cell probabilities directly, we can set up a model for the cell counts, which is traditionally estimated using a Poisson regression. Define $\lambda_m = \lambda_{(l_1,\ldots,l_{d_y})} = n * \pi_{(l_1,\ldots,l_{d_y})}$ as the count in cell m. The model for the cell counts is

$$\log\lambda_m = \tilde{\alpha}_0 + \sum_{k=1}^{d_y} \alpha_{k,l_k} + \sum_{k=1}^{d_y} \sum_{k'>k} \alpha_{kk',l_kl_{k'}} + \sum_{k=1}^{d_y} \sum_{k'>k} \sum_{k''>k'} \alpha_{kk'k'',l_kl_{k'}l_{k''}} + \ldots. \tag{2.4}$$

Then, based on the relationship between the cell counts and the cell probabilities,

$$\begin{aligned} \log\pi_m &= \log\lambda_m - \log n \\ &= -\log n + \tilde{\alpha}_0 + \sum_{k=1}^{d_y} \alpha_{k,l_k} + \sum_{k=1}^{d_y} \sum_{k'>k} \alpha_{kk',l_kl_{k'}} + \sum_{k=1}^{d_y} \sum_{k'>k} \sum_{k''>k'} \alpha_{kk'k'',l_kl_{k'}l_{k''}} + \ldots. \end{aligned} \tag{2.5}$$

Equation (2.5) is the same as equation (2.3) after defining $\alpha_0 = \tilde{\alpha}_0 - \log n$.

Little and Schluchter [12], along with help from Anderson [3], showed that when the complete data is available, the maximum likelihood estimates of $\mathbf{B}$ and $\mathbf{\Omega}$ for the MVN distribution defined in (2.2), are

$$\hat{\mathbf{B}} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i \mathbf{Z}_i^T\right) \left(\sum_{i=1}^{n} \mathbf{Z}_i \mathbf{Z}_i^T\right)^{-1} \text{ and } \hat{\mathbf{\Omega}} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{X}_i - \hat{\mathbf{B}}\mathbf{Z}_i)(\boldsymbol{X}_i - \hat{\mathbf{B}}\mathbf{Z}_i)^T. \tag{2.6}$$

How to obtain MLE estimates for the MVG case is discussed in Appendix A. The MLE estimates of the $\boldsymbol{\alpha}$ terms are obtained using a Poisson regression and a link function in (2.4). The generalized linear model function `glm` in the Poisson family of R Core Team [15] is used, and the intercept term α_0 is estimated as $\hat{\alpha}_0 = \hat{\tilde{\alpha}}_0 - \log n$. Following Little and Schluchter [12], we assume that three-way and higher interaction terms are unimportant in the simulation and empirical studies. Thus,

$$\hat{\pi}_m = \exp\left[\hat{\alpha}_0 + \sum_{k=1}^{d_y} \hat{\alpha}_{k,l_k} + \sum_{k=1}^{d_y} \sum_{k'>k} \hat{\alpha}_{kk',l_kl_{k'}}\right]. \tag{2.7}$$

Next we define the distribution function of the observed data evaluated at the MLEs for the MVN case. For a question pattern p in a given SQD design D, the distribution function of the observed data is

$$f(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_{i,obs} \mid p, \hat{\mathbf{B}}, \hat{\mathbf{\Omega}}, \hat{\boldsymbol{\alpha}}) = \int f(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_{i,obs}, \boldsymbol{Y}_{i,mis} \mid p, \hat{\mathbf{B}}, \hat{\mathbf{\Omega}}, \hat{\boldsymbol{\alpha}}) d\boldsymbol{Y}_{mis}$$

$$= \sum_{m \in S_{i,mis}} f(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_i = m \mid \hat{\mathbf{B}}, \hat{\boldsymbol{\Omega}}, \hat{\boldsymbol{\alpha}}) = \sum_{m \in S_{i,mis}} f(\boldsymbol{X}_{i,obs} \mid m, \hat{\mathbf{B}}, \hat{\boldsymbol{\Omega}}) \times \hat{\pi}_{m,i}$$
$$= \sum_{m \in S_{i,mis}} (2\pi)^{-d_{x,obs}/2} \mid \hat{\boldsymbol{\Omega}}_{obs} \mid^{-\frac{1}{2}} \exp\left[\frac{-1}{2}\left(\boldsymbol{X}_{i,obs} - \hat{\boldsymbol{\mu}}_{m,i,obs}\right)^T \hat{\boldsymbol{\Omega}}_{obs}^{-1}\left(\boldsymbol{X}_{i,obs} - \hat{\boldsymbol{\mu}}_{m,i,obs}\right)\right] \times \hat{\pi}_{m,i}, \tag{2.8}$$

where $S_{i,mis}$ contains possible cells that respondent i could fall in for $\boldsymbol{Y}_{i,mis}$ given $\boldsymbol{Y}_{i,obs}$, $d_{x,obs}$ is the number of observed continuous variables in pattern p, $\hat{\pi}_{m,i}$ is the estimated probability that respondent i falls in cell m, $\hat{\boldsymbol{\mu}}_{m,i,obs}$ is the estimated mean vector for the observed continuous variables for respondent i who falls in cell m, and $\hat{\boldsymbol{\Omega}}_{obs}$ is the common estimated variance-covariance matrix for the observed continuous variables. Recall that $\hat{\boldsymbol{\mu}}_{m,i} = \hat{\mathbf{B}}\mathbf{Z}_i$ for $i = 1, \ldots, n$. $\hat{\boldsymbol{\mu}}_{m,i,obs}$ (or $\hat{\boldsymbol{\Omega}}_{obs}$) contains the elements of $\hat{\boldsymbol{\mu}}_{m,i}$ (or $\hat{\boldsymbol{\Omega}}$) that correspond to the observed continuous variables $\boldsymbol{X}_{obs}$ in pattern p. Plugging equation (2.8) into equation (1.3), we obtain the estimated expectation of log-distribution function for the observed.

As an illustration of $S_{i,mis}$, we use the toy example of a SQD in Table 1 and assume $p = 1$. Suppose the number of possible levels I_k for the missing categorical variables Y_4 and Y_6 are $I_4 = I_6 = 2$. In this pattern $p = 1$, $S_{i,mis} = \{(l_1, l_2, l_3, Y_4 = 1, l_5, Y_6 = 1)\}, \{(l_1, l_2, l_3, Y_4 = 1, l_5, Y_6 = 2)\}, \{(l_1, l_2, l_3, Y_4 = 2, l_5, Y_6 = 1)\}, \{(l_1, l_2, l_3, Y_4 = 2, l_5, Y_6 = 2)\}$, where l_k $(k = 1, 2, 3, 5)$ is the observed level that categorical variable Y_k takes for respondent i. Then the number of possible cells, denoted as $|S_{i,mis}|$, is equal to 4 in this example. However, when the number of categorical variables planned to be missing is big, $|S_{i,mis}|$ becomes huge. For the example in Section 1 where 20 questions, each of which has 5 levels, are missing in a pattern, the number of possible cells $|S_{i,mis}|$ reaches 9.5×10^{13}. This makes the calculation for the distribution of the observed in (2.8) computationally infeasible. Note that we also need to sum over P and n as shown in (1.3) to obtain the values of $\hat{E}_{\boldsymbol{X},\boldsymbol{Y}}\left[\ln f(\boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs} \mid D)\right]$ for different SQDs, which will be used to compare the KLDs. In the next subsection, we propose a method to approximate the summation $\sum_{m \in S_{i,mis}}$ in (2.8) in tremendously reduced computation time.

2.2. Approximating the Distribution Function of the Observed

In this subsection, we discuss a time efficient approach to approximate $f(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_{i,obs} \mid p)$ in (2.8), thus enabling us to compare the KLD distances across different SQDs. The main motivation

for this approach is through the classic survey sampling estimator, the Horvitz-Thompson (HT) estimator.

Horvitz and Thompson [10] first theorized estimating a population total $T = \sum_{j=1}^{N} U_j$, where U_j is the variable of interest for unit j and N is the population size. N can be very large, just like the norm $|S_{i,mis}|$ defined in the previous subsection. For example, according to Bureau [5], the U.S. population is $N > 308,000,000$ people and continues to grow. This makes it impossible to gather information from every individual j in such large population. Therefore, sampling with sample size of $J << N$ becomes necessary, and an estimator of T using collected information from the sample needs to be developed. Horvitz and Thompson [10] discovered that the best linear unbiased estimator for T is $\hat{T} = \sum_{j=1}^{J} \frac{U_j}{p_j}$, where p_j is the inclusion probability of selecting unit j into the sample. Using this idea, we treat $S_{i,mis}$ from (2.8) as a 'population' and the analogous total that needs to be estimated is $T = \sum_{m_j \in S_{i,mis}} U_j$ where $U_j = f(\boldsymbol{X}_{i,obs}, m_j \mid p, \hat{\mathbf{B}}, \hat{\boldsymbol{\Omega}}, \hat{\boldsymbol{\alpha}})$. We can sample J cells m_j's (for $j = 1, \cdots, J$) according to a proposed sampling design with the inclusion probability $P(m = m_j \mid S_{i,mis})$, which takes the place of p_j in $\hat{T}$ above. Here, J is chosen to be a reasonable number (but much smaller than $|S_{i,mis}|$) to achieve accurate approximation in a timely manner. Thus, the population total can be estimated by $\sum_{j=1}^{J} \frac{U_j}{P(m=m_j|S_{i,mis})}$, i.e. an estimator of the distribution function in (2.8) can be specified as

$$\hat{f}(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_{i,obs} \mid p, \hat{\mathbf{B}}, \hat{\boldsymbol{\Omega}}, \hat{\boldsymbol{\alpha}}) = \sum_{j=1}^{J} f(\boldsymbol{X}_{i,obs} \mid m_j, \hat{\mathbf{B}}, \hat{\boldsymbol{\Omega}}) \times \frac{\hat{\pi}_{m_j,i}}{P(m = m_j \mid S_{i,mis})}$$

$$= \sum_{j=1}^{J} (2\pi)^{-d_{x,obs}/2} \mid \hat{\boldsymbol{\Omega}}_{obs} \mid^{-\frac{1}{2}} \exp\left[\frac{-1}{2}\left(\boldsymbol{X}_{i,obs} - \hat{\boldsymbol{\mu}}_{m_j,i,obs}\right)^T \hat{\boldsymbol{\Omega}}_{obs}^{-1}\left(\boldsymbol{X}_{i,obs} - \hat{\boldsymbol{\mu}}_{m_j,i,obs}\right)\right]$$

$$\times \frac{\hat{\pi}_{m_j,i}}{P(m = m_j \mid S_{i,mis})}, \tag{2.9}$$

where m_j is the index for the j^{th} sampled cell from $S_{i,mis}$ and $P(m = m_j \mid S_{i,mis})$ is the probability that cell m_j is selected into the sample. The approximate expected value of the observed log-distribution in (1.3) is

$$\hat{E}_{\boldsymbol{X},\boldsymbol{Y}}\left[\ln f(\boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs} \mid D)\right] = \frac{1}{nP} \sum_{p=1}^{P} \sum_{i=1}^{n} \left[\ln \hat{f}(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_{i,obs} \mid p, \hat{\mathbf{B}}, \hat{\boldsymbol{\Omega}}, \hat{\boldsymbol{\alpha}})\right], \tag{2.10}$$

where $\hat{f}$ is calculated from (2.9) with a much smaller $J << |S_{i,mis}|$. In both simulation and empirical studies, we examine the effect of different J values on the ability of our approach in identifying the best SQD. The estimated log-distribution of the observed for the MVG case is

obtained in the same way.

We can simply use the multinomial sampling as the proposed sampling design to draw cell m_j from $S_{i,mis}$ with the inclusion probability specified as,

$$P(m = m_j \mid S_{i,mis}) = \prod_{k=1}^{d_{y,mis}} P(Y_k = l_k), \text{ for } j = 1, \cdots, J, \tag{2.11}$$

where $d_{y,mis}$ is the number of missing categorical variables in a pattern p, l_k ($k = 1, \cdots, d_{y,mis}$) is the level for Y_k in cell m_j, and $P(Y_k = l_k)$ is the marginal probability of Y_k taking level l_k. The sampling can be done as follows. For each missing categorical variable Y_k ($k = 1, \cdots, d_{y,mis}$), simulate a $I_k \times 1$ indicator vector $\boldsymbol{\delta}_k \sim Multinomial(1; P(Y_k = 1), \cdots, P(Y_k = I_k))$, where $Multinomial(1; p_1, \cdots, p_{I_k})$ is multinomial distribution with number of trail equal to 1 and category probabilities $p_1, \cdots, p_{I_k}$. The resulting $\boldsymbol{\delta}_k$ contains $I_k - 1$ 0's and only one 1. The position of the 1 in the indicator $\boldsymbol{\delta}_k$ corresponds to the level chosen for Y_k. Repeat this for all missing categorical variables Y_k's ($k = 1, \cdots, d_{y,mis}$) independently to fill in the missing spots in sample cell m_j. We can use historical data, such as that from the U.S. Census Bureau, past survey, pilot study, etc. to estimate the marginal probabilities $P(Y_k = l_k)$ in (2.11). If no historical data is available, equal probability of $P(Y_k = l_k) = \frac{1}{I_k}$ can be used to draw sample m_j. In the simulation and empirical studies, data from a prior survey is used to calculate $P(m = m_j \mid S_{i,mis})$. Other sampling schemes with correctly specified inclusion probabilities can be applied as long as they are easy to be implemented.

3. Simulation Study

The main objective of the simulation study is to evaluate the performance of our proposed approximation method in terms of achieving time efficiency and accurately identifying the rank of SQD designs. The simulation scenarios contrived in this section cover different distributions of continuous variables and different correlation structures.

Following the toy example described in Section 1 and illustrated in Table 1, we conduct a simulation study with 4 continuous variables and 6 categorical variables that have 5 levels, i.e. $I_k = 5$ for $k = 1, \ldots, 6$. Recall, Y_1 and Y_2 are chosen as the core questions with the remaining questions split into two blocks. We simulate population data (for $i = 1, \ldots, N = 50,000$) from 2

conditional distributions of the continuous variables, MVN and MVG, combined with 3 different correlation structures.

First we specify the following 3 cell probability models for the categorical variables,

$$\log \pi_m = \alpha_{1,l_1} + \alpha_{2,l_2} + \alpha_{3,l_3} + \alpha_{4,l_4} + \alpha_{5,l_5} + \alpha_{6,l_6} + \alpha_{34,l_3l_4} + \alpha_{56,l_5l_6} + \alpha_{35,l_3l_5} + \alpha_{46,l_4l_6}, \quad (3.1)$$

$$\log \pi_m = \alpha_{1,l_1} + \alpha_{2,l_2} + \alpha_{3,l_3} + \alpha_{4,l_4} + \alpha_{5,l_5} + \alpha_{6,l_6} + \alpha_{35,l_3l_5} + \alpha_{46,l_4l_6} + \alpha_{36,l_3l_6} + \alpha_{45,l_4l_5}, \text{ and} \quad (3.2)$$

$$\log \pi_m = \alpha_{1,l_1} + \alpha_{2,l_2} + \alpha_{3,l_3} + \alpha_{4,l_4} + \alpha_{5,l_5} + \alpha_{6,l_6} + \alpha_{34,l_3l_4} + \alpha_{56,l_5l_6} + \alpha_{36,l_3l_6} + \alpha_{45,l_4l_5}, \quad (3.3)$$

and then define the following 3 regression models for the mean of the continuous variables,

$$\boldsymbol{X} \sim Y_3 + Y_4 + Y_5 + Y_6 + Y_3 : Y_4 + Y_5 : Y_6 + Y_3 : Y_5 + Y_4 : Y_6, \quad (3.4)$$

$$\boldsymbol{X} \sim Y_3 + Y_4 + Y_5 + Y_6 + Y_3 : Y_5 + Y_4 : Y_6 + Y_3 : Y_6 + Y_4 : Y_5, \text{ and} \quad (3.5)$$

$$\boldsymbol{X} \sim Y_3 + Y_4 + Y_5 + Y_6 + Y_3 : Y_4 + Y_5 : Y_6 + Y_3 : Y_6 + Y_4 : Y_5, \quad (3.6)$$

where $Y_k : Y_{k'}$ represents a two-way interaction term between Y_k and $Y_{k'}$. The regressors on the right hand side determine the vector $\mathbf{Z}$ in (2.2). For example, under the regression model in (3.4), for a respondent with $Y_3 = l_3, Y_4 = l_4, Y_5 = l_5$, and $Y_6 = l_6$, the corresponding $\mathbf{Z} = [\boldsymbol{\delta}_3^T, \boldsymbol{\delta}_4^T, \boldsymbol{\delta}_5^T, \boldsymbol{\delta}_6^T, (\boldsymbol{\delta}_3 \otimes \boldsymbol{\delta}_4)^T, (\boldsymbol{\delta}_5 \otimes \boldsymbol{\delta}_6)^T, (\boldsymbol{\delta}_3 \otimes \boldsymbol{\delta}_5)^T, (\boldsymbol{\delta}_4 \otimes \boldsymbol{\delta}_6)^T]^T$, where $\boldsymbol{\delta}_k$ is the $I_k \times 1$ indicator vector observed for variable Y_k that has 1 in the position of l_k and 0's elsewhere, and $\otimes$ represents a Kronecker product.

The data are simulated following 3 pairs of model specifications, (3.1) with (3.4), (3.2) with (3.5), and (3.3) with (3.6), with appropriately chosen values of $\mathbf{B}$, $\boldsymbol{\Omega}$, and $\boldsymbol{\alpha}$. This leads to the correlation structures displayed in Tables 2a, 2b, and 2c, respectively for these 3 pairs of models. In addition, the coefficients in $\mathbf{B}$ are chosen such that the mean of the continuous variables, $\boldsymbol{\mu}_{m,i}$, for respondent i conditioned on cell m has a large spread based on the different levels of Y_k, making it difficult to estimate the expectation of the log-distribution function for the observed when two highly correlated categorical variables are both missing for a particular pattern. To capture the potential skewness of the data, the MVG case is also considered. Similar correlation structures from Tables 2a, 2b, and 2c are used, and the method to simulate the MVG data can be found in Appendix A. As a result, there are 6 simulation scenarios prescribed by 3 correlation structures and 2 conditional distributions for continuous variables $\boldsymbol{X}$ (MVN and MVG).

	X_1	X_2	X_3	X_4		Y_3	Y_4	Y_5	Y_6
X_1	1	0.9	0.2	0	Y_3	1	0.9	0.2	0
X_2	0.9	1	0	0.2	Y_4	0.9	1	0	0.2
X_3	0.2	0	1	-0.9	Y_5	0.2	0	1	-0.9
X_4	0	0.2	-0.9	1	Y_6	0	0.2	-0.9	1

(a) Correlation structures simulated from the model pair (3.1) and (3.4)

	X_1	X_2	X_3	X_4		Y_3	Y_4	Y_5	Y_6
X_1	1	0	0.9	0.2	Y_3	1	0	0.9	0.2
X_2	0	1	0.2	-0.9	Y_4	0	1	0.2	-0.9
X_3	0.9	0.2	1	0	Y_5	0.9	0.2	1	0
X_4	0.2	-0.9	0	1	Y_6	0.2	-0.9	0	1

(b) Correlation structures simulated from the model pair (3.2) and (3.5)

	X_1	X_2	X_3	X_4		Y_3	Y_4	Y_5	Y_6
X_1	1	0.2	0	0.9	Y_3	1	0.2	0	0.9
X_2	0.2	1	-0.9	0	Y_4	0.2	1	-0.9	0
X_3	0	-0.9	1	0.2	Y_5	0	-0.9	1	0.2
X_4	0.9	0	0.2	1	Y_6	0.9	0	0.2	1

(c) Correlation structures simulated from the model pair (3.3) and (3.6)

Table 2: Description of the three correlation structures for the simulation in Section 3

Next, we define 3 SQDs considered in this simulation. Each SQD is designed so that each pattern within the SQD has two categorical variables and two continuous variables in addition to the 2 core questions. The selection of questions into patterns is motivated from the correlation structures, as described in Tables 2a, 2b, and 2c. For example, SQD 1, described in Table 3a, does not have Y_3 and Y_4, nor X_1 and X_2, together in any particular pattern because both pairs have a correlation of either 0.9 or -0.9 according to Table 2a. SQD 2 (or SQD 3), described in Tables 3b (or Table 3c), is designed following the same reason using the correlation structure in Table 2b (or Table 2c). This is done to further illustrate that, the higher the correlation of variables in different patterns, the less the "information loss" is, as theorized in Raghunathan and Grizzle [16].

From the population of 50,000 observations under each scenario, 100 simple random samples of size $n = 1,000$ are drawn, i.e. $r = 100$ Monte Carlo (MC) samples. For each MC sample, we calculate the approximate expected log-distribution of the observed, $\hat{E}_{\boldsymbol{X},\boldsymbol{Y}}\left[\ln f(\boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs} \mid D)\right]$ in (2.10), for 3 different SQDs. Note that the parameter estimates $\hat{\mathbf{B}}$, $\hat{\boldsymbol{\Omega}}$, and $\hat{\boldsymbol{\alpha}}$ are recalculated under each Monte Carlo sample. The means and standard errors of each of these parameters are reported in the supplemental file. Then we use these values of $\hat{E}_{\boldsymbol{X},\boldsymbol{Y}}\left[\ln f(\boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs} \mid D)\right]$

p	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	X_1	X_2	X_3	X_4
1	✓	✓	✓		✓		✓		✓	
2	✓	✓	✓			✓	✓			✓
3	✓	✓		✓	✓			✓	✓	
4	✓	✓		✓		✓		✓		✓

(a) SQD 1 for the simulation in Section 3

p	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	X_1	X_2	X_3	X_4
1	✓	✓	✓	✓			✓	✓		
2	✓	✓	✓			✓	✓			✓
3	✓	✓			✓	✓			✓	✓
4	✓	✓		✓	✓			✓	✓	

(b) SQD 2 for the simulation in Section 3

p	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	X_1	X_2	X_3	X_4
1	✓	✓	✓	✓			✓	✓		
2	✓	✓	✓		✓		✓		✓	
3	✓	✓			✓	✓			✓	✓
4	✓	✓		✓		✓		✓		✓

(c) SQD 3 for the simulation in Section 3

Table 3: Description of the three choices of SQD for the simulation in Section 3

to rank different SQDs based on the KLD described in (1.1). A design with higher value of $\hat{E}_{\boldsymbol{X},\boldsymbol{Y}}\left[\ln f(\boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs} \mid D)\right]$ leads to smaller amount of "information loss", thus is more preferable. We use $J = \{5, 10\}$ to calculate the approximated density $\hat{f}$ in (2.9) to see the effect of J values on the rank of the approximate expected log-distribution of the observed. These small values of J are chosen because, in each pattern, the number of possible cells is only $|S_{i,mis}| = 5 \times 5 = 25$ in this toy simulation. To see whether our proposed method can capture the "true" ranks of the expected log-distribution of the observed accurately, we need a benchmark. A brute-force sum of the "true" expected log-distribution of the observed is calculated as

$$\tilde{E}_{\boldsymbol{X},\boldsymbol{Y}}\left[\ln f(\boldsymbol{X}_{obs}, \boldsymbol{Y}_{obs} \mid D)\right] = \frac{1}{NP}\sum_{p=1}^{P}\sum_{i=1}^{N} \ln f(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_{i,obs} \mid p), \tag{3.7}$$

where N is the population size and $f(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_{i,obs} \mid p)$ is calculated from (2.8) using the true parameters $\mathbf{B}$, $\boldsymbol{\Omega}$, and $\boldsymbol{\alpha}$ and enumerating all possible cells m's in $S_{i,mis}$.

In addition, we record the average computation times of the true and approximate observed distribution functions calculated in (2.8) and (2.9) per each respondent over 3 SQDs under different scenarios. The goal is to see if the computation time is reduced enough to make the approximation computationally efficient while maintaining accuracy of the rank.

Correlation Structure	Distribution of Continuous Variables	J	"True" Rank of SQDs	Percent of r samples that match "true" rank	Avg. Time of True Density (×1000 sec)	Avg. Time of Approximated Density (×1000 sec)
Table 2a	Normal	5	$2 < 3 < 1$	100%	20.8	3.98
		10	$2 < 3 < 1$	100%		7.94
	Gamma	5	$2 < 3 < 1$	100%	31.5	6.18
		10	$2 < 3 < 1$	100%		12.34
Table 2b	Normal	5	$3 < 1 < 2$	100%	20.8	4.01
		10	$3 < 1 < 2$	100%		7.97
	Gamma	5	$3 < 1 < 2$	100%	31.5	6.22
		10	$3 < 1 < 2$	100%		12.31
Table 2c	Normal	5	$1 < 2 < 3$	100%	20.8	3.95
		10	$1 < 2 < 3$	100%		7.88
	Gamma	5	$1 < 2 < 3$	100%	31.5	6.20
		10	$1 < 2 < 3$	100%		12.37

Table 4: Simulation results from $r = 100$ MC samples for 6 simulation scenarios described in Section 3. The first three columns indicate correlation structures, the assumptions of the conditional distribution of the continuous variables, and the values of J used when calculating (2.9). The next columns provide the "true" rank of the SQDs using the brute force sum, the percentages of the MC samples that give the same rank using the approxmation method, and the computation time for the "true" observed distribution function per respondent and the computation time for the approximate observed distribution functions per respondent.

The simulation results for 6 simulation scenarios and 2 choices of J are shown in Table 4. Table 4 reports the "true" rank of the SQDs using the brute force sum, the percentages of the MC samples that give the same rank using the approximation method, and the average computation time for the true observed distribution function per respondent and the average computation time for the approximate observed distribution functions per respondent. The percent of r samples that match the "true" rank is 100% across all 6 scenarios and 2 choices of J. The "true" rank is different across the correlation structures, suggesting that the correlation structures of the variables has a significant impact on which design is considered the best in terms of "information loss". As expected, SQD 1 is the best for the scenario with Table 2a correlation structure, SQD 2 is the best for Table 2b correlations structure, and SQD 3 is the best for Table 2c correlation structure. In general, the designs with the highest expected log-distribution of the observed (i.e. the best ones) are the ones where the continuous and categorical variables with correlation 0.9 or -0.9 are not in any pattern together, which is consistent with the discussion in Raghunathan and Grizzle [16].

Under the same scenario, the computation time increases linearly as J increases. This makes

sense because the summation in (2.9) is doubled when increasing J from 5 to 10. Because of the structure of the MVG distribution, as discussed in Appendix A, it takes a longer time to compute the true and approximated distribution functions of the observed, compared to the MVN case. In order to see the magnitude of relative difference in time to compute the true and the approximate distribution functions of the observed, we look at the average of ratios of the computation time for the true observed distribution to the computation time for the approximation over the 6 scenarios for $J = 5$. This ratio comes out as $\frac{1}{6}\sum_{j=1}^{6} \frac{\text{Avg. Time of true}}{\text{Avg. Time of approximation}} = 5.15$, meaning that it takes 5.15 times as long to compute the true observed distribution function as compared to the approximated distribution function for the observed. This does not seem to be very computationally efficient, since it only takes about 20.8×10^{-3} seconds to calculate the true distribution function per individual. This is due to the fact that $|S_{i,mis}| = 25$ for all patterns in each of the three possible designs, which is very small. In the more realistic simulation study of the next section, we show that as $|S_{i,mis}|$ increases exponentially, the approximation method becomes very computationally efficient.

4. Application to the 2012 Pet Demographic Survey Data

In this section, we conduct a more realistic simulation study that represents properties of the 2012 Pet Demographic Survey (PDS) in Section 4.1, and apply our proposed approximation method to the real 2012 PDS data in Section 4.2.

4.1. Simulation Study Motivated by the Pet Demographic Survey

In 2016, Iowa State University started an agreement to plan a five-year national survey known as the Pet Demographic Survey (PDS) for the American Veterinary Medical Association (AVMA). This is a survey that collects information about households and their pets, such as whether they own any particular pet (dogs, cats, horses, birds, and fishes), counts of those pets, body types of those pets, amount of money spent on their pets at the veterinarian, etc.. To prepare for the 2017 PDS, data from the previous national survey, the 2012 PDS, was provided as a starting point.

For this data analysis, we choose $d_x = 4$ continuous variables and $d_y = 16$ categorical variables. We choose a much shorter version of the real PDS survey in order to be able to calculate the

brute-force "true" rank, our benchmark. The goal is to see if our proposed approximation method can correctly and quickly identify the best SQD suggested by the "true" rank of expected log-distribution function for the observed in this more realistic simulation set-up. The evaluation is done for 4 model scenarios specified by 2 different correlation structures and 2 distributional forms for the continuous variables (MVN and MVG).

4.1.1. Continuous and Categorical Variables

Table 5 and 6 give detailed descriptions for 4 continuous variables that are related to expenses spent at veterinary visits, and 16 categorical variables that are related to demographic information, household information and number of pets.

4.1.2. Models for π_m and Mean of $\boldsymbol{X}$

We first discuss how to select the model for the cell probability π_m. This selection is done through likelihood ratio tests. Using the notation from (2.3) and motivation from the Neyman-Pearson Lemma as described in Neyman and Pearson [13], we specifically test the hypotheses

$$H_0 : \alpha_{k1} = \cdots = \alpha_{kI_k} = 0, \text{ and}$$

$$H_A : \text{At least one } \alpha_{kj} \neq 0 \ (j = 1, \cdots, I_k),$$

for $k = 1, \ldots, d_y$ where $d_y = 16$. The likelihood ratio test is conducted for each categorical variable, and the variable with the highest p-value is eliminated at each step. The process repeats with the reduced number of categorical variables and the process stops once all p-values are below a significance level of 0.1. Then, the two-way interaction terms for the remaining categorical variables are tested using a chi-squared test backward elimination process, looking at a significance level of 0.1. After all of these elimination procedures, the final cell probability depends on 11 main effects and 5 two-way interaction terms, and is specified as

$$\begin{aligned}\log \pi_m = {} & \alpha_{4,l_4} + \alpha_{5,l_5} + \alpha_{7,l_7} + \alpha_{8,l_8} + \alpha_{9,l_9} + \alpha_{10,l_{10}} + \alpha_{11,l_{11}} + \alpha_{13,l_{13}} + \alpha_{14,l_{14}} \\ & + \alpha_{15,l_{15}} + \alpha_{16,l_{16}} + \alpha_{4,8,l_4 l_8} + \alpha_{8,10,l_8 l_{10}} + \alpha_{8,11,l_8 l_{11}} + \alpha_{8,13,l_8 l_{13}} + \alpha_{8,15,l_8 l_{15}}. \end{aligned} \tag{4.1}$$

Next we discuss how to select categorical variables that are used to determine vector $\mathbf{Z}$ in the conditional mean of $\boldsymbol{X}$. The first step is to run a multiple linear regression model for each of 4 continuous variable individually with all main effects and two-way interactions based on all

Categorical Variable	Definition	Description of Levels
Y_1 ($I_1 = 5$)	Race	White; African-American; Asian or Pacific Islander; American Indian, Aleut, Eskimo; Other
Y_2 ($I_2 = 4$)	Type of Residence	House; Apartment/Condominium; Mobile Home; Other
Y_3 ($I_3 = 5$)	Age of Head of Household	Under 30; 30-39; 40-49; 50-59; 60 and Over
Y_4 ($I_4 = 2$)	Gender	Male,Female
Y_5 ($I_5 = 5$)	Geographic Region	Description in Table 6a
Y_6 ($I_6 = 5$)	Household Income	Under \$20,000; \$20,000 to \$39,999; \$40,000 to \$59,999; \$60,000 to \$99,999; \$100,000 and Over
Y_7 ($I_7 = 4$)	Community Size	Under 100,000; 100,000 to 499,999; 500,000 to 1,999,999; 2,000,000 and Over
Y_8 ($I_8 = 3$)	Marital Status	Married; Never Married/Single; Divorced/Widowed
Y_9 ($I_9 = 4$)	Education Status	High School or Less; Attended College; College Graduate; Advanced Degree
Y_{10} ($I_{10} = 3$)	Home Ownership Status	Own; Rent; Other
Y_{11} ($I_{11} = 3$)	Household Size	1;2;3 or More
Y_{12} ($I_{12} = 7$)	Household Designation	Description in Table 6b
Y_{13} ($I_{13} = 5$)	Employment Status	Full-Time; Part-Time; Retired; Unemployed; Other
Y_{14} ($I_{14} = 2$)	Hispanic Origin	Hispanic; Non-Hispanic
Y_{15} ($I_{15} = 4$)	Number of Dogs Owned in 2011	1; 2; 3; 4 or More
Y_{16} ($I_{16} = 4$)	Number of Cats Owned in 2011	1; 2; 3; 4 or More

(a) Description of Categorical Variables Used in Simulation and Empirical Data in Section 4

Categorical Variable	Definition
X_1	Dollars Spent at Last Veterinary Visit in 2011 for Dogs (\$1000s)
X_2	Dollars Spent at Last Veterinary Visit in 2011 for Cats (\$1000s)
X_3	Dollars Spent in Total at Vet in 2011 for Dogs (\$1000s)
X_4	Dollars Spent in Total at Vet in 2011 for Cats (\$1000s)

(b) Description of Continuous Variables Used in Simulation and Empirical Data in Section 4

Table 5: Description of $\boldsymbol{X}$ and $\boldsymbol{Y}$ Used in Simulation and Empirical Data in Section 4

Geographic Region	States
1	Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont
2	Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, New Jersey, New York, North Dakota, Ohio, Pennsylvania, South Dakota, Wisconsin
3	Alabama, Arkansas, Delaware, District of Columbia, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, West Virginia
4	Alaska, Arizona, Colorado, Hawaii, Idaho, Montana, New Mexico, Oregon, Utah, Washington, Wyoming
5	California, Nevada

(a) Description of Categorical Variable Geographic Region from Section 4

Marriage Designation	States
1	Husband and Wife
2	Male, No Wife, Child, and/or Other Relative Present
3	Female, No Husband, Child, and/or Other Relative Present
4	Male Living Alone
5	Female Living Alone
6	Male Living with Non-Relative
7	Female Living with Non-Relative

(b) Description of Categorical Variable Household Designation from Section 4

Table 6: Description of Additional Categorical Variables from Section 4

categorical variables. A backward elimination regression method is used to further reduce the number of variables in each of these regression models. Once this process is completed, any main effect or two-way interaction that is determined to be significant in at least 3 out of the 4 regression models is selected into the final model. Further reduction of the models is constructed on a case by case basis, resulting in the final regression model as

$$\begin{aligned}\boldsymbol{X} \sim\ & Y_1 + Y_3 + Y_4 + Y_5 + Y_6 + Y_7 + Y_8 + Y_{10} + Y_{14} + Y_{15} \\ & + Y_{16} + Y_1 : Y_3 + Y_1 : Y_{10} + Y_5 : Y_{15} + Y_5 : Y_{16} + Y_{10} : Y_{14}. \end{aligned} \tag{4.2}$$

The important categorical variables affecting the continuous variables $\boldsymbol{X}$ are listed on the right hand side of (4.2), which lead to the definition of the vector $\mathbf{Z}$ in (2.2). The resulting important variables in both (4.1) and (4.2) help in choosing the appropriate SQDs as well.

4.1.3. Choices of SQD Designs

The next decision is how to divide the questionnaire into different patterns to form potential SQDs. Four core questions, which are deemed to be the most important in the survey, are determined in this step and data on these variables should be gathered for every individual being surveyed. The remaining 16 questions are divided into 2 blocks with 8 questions in each. The decisions of the core questions and the formation of the blocks are motivated using a Pearson chi-squared test of independence between the categorical variables (Pearson [14]). Table 7a shows the p-values of the Chi-Squared Independence tests for all 16 categorical variables. If a p-value is lower than 0.001, the symbol $' < 0.001'$ is used to indicate a low p-value. Any pair of categorical variables with a p-value greater than a significance level of 0.1 are considered to be independent of each other and are more likely to be placed in the same block or same pattern of a particular design. Correlation between the continuous variables from the survey is also checked. Table 7b shows the absolute values of the correlation coefficients for all pairs of continuous variables. It can be seen that none of the continuous variables have a strong relationship with each other, so they are not a major factor in the choice of designs and patterns.

Using the information in Tables 7a and 7b for guidance, the core questions are chosen. We first select a set of variables that affect both $\log\pi_m$ and mean of $\boldsymbol{X}$ models in (4.1) and (4.2). We then use Table 7a to see which variables are more correlated with the others and remove one of

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}	Y_{16}
Y_1	<0.001															
Y_2	0.008	<0.001														
Y_3	<0.001	<0.001	<0.001													
Y_4	0.001	0.233	0.001	<0.001												
Y_5	<0.001	<0.001	0.07	0.032	<0.001											
Y_6	0.077	<0.001	<0.001	0.514	<0.001	<0.001										
Y_7	0.055	<0.001	0.009	0.196	<0.001	<0.001	<0.001									
Y_8	<0.001	<0.001	<0.001	<0.001	0.01	<0.001	0.001	<0.001								
Y_9	<0.001	<0.001	<0.001	0.001	<0.001	<0.001	0.043	0.257	<0.001							
Y_{10}	<0.001	<0.001	<0.001	0.753	<0.001	<0.001	0.123	<0.001	<0.001	<0.001						
Y_{11}	0.261	<0.001	<0.001	0.53	0.129	<0.001	0.114	<0.001	<0.001	<0.001	<0.001					
Y_{12}	<0.001	<0.001	<0.001	<0.001	0.614	<0.001	0.124	<0.001	<0.001	<0.001	<0.001	<0.001				
Y_{13}	0.001	<0.001	<0.001	0.022	0.051	<0.001	0.227	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001			
Y_{14}	<0.001	0.014	<0.001	0.045	<0.001	0.618	0.001	<0.001	0.738	<0.001	<0.001	<0.001	0.054	<0.001		
Y_{15}	0.154	<0.001	0.198	0.209	<0.001	0.021	<0.001	0.997	0.205	<0.001	0.025	0.689	0.005	0.089	<0.001	
Y_{16}	0.014	<0.001	0.001	<0.001	0.22	0.004	<0.001	0.621	0.199	0.022	0.348	<0.001	0.253	0.32	<0.001	<0.001

(a) Pearson Chi-Squared Test p-values for Independence of the Categorical Variables in the PDS survey

	X_1	X_2	X_3	X_4
X_1	1.00	0.18	0.51	0.12
X_2	0.18	1.00	0.13	0.50
X_3	0.51	0.13	1.00	0.23
X_4	0.12	0.50	0.23	1.00

(b) Absolute Correlation Coefficients of the Continuous Variables in the PDS survey

Table 7: Correlation structures of the categorical and continuous variables for use of choice of SQD

Block 1	Y_1	Y_5	Y_{11}	X_1	Y_6	Y_{12}	Y_{13}	X_2
Block 2	Y_4	Y_7	Y_{10}	X_3	Y_2	Y_3	Y_9	X_4

Table 8: Final Choices for Blocks for Simulation Study in Section 4

them. Based on these criteria, the four core questions are Y_8, Y_{14}, Y_{15}, and Y_{16}.

Table 7a is also used to determine how the remaining variables are split into two equal blocks, as well as the distribution of the patterns within each block. The blocks are split such that each block has an equal number of categorical variables and continuous variables. The categorical variables are split such that variables that are determined to be independent of each other based on the chi-squared tests are more likely to be placed in the same block. The continuous variables are split such that each block has one question regarding dogs and one question regarding cats. The final split of blocks is described in Table 8.

Now that the blocks have been determined, we define 3 SQDs that need to be compared. The SQDs are created such that they differ by the number of continuous variables between SQDs, the number of continuous variables between patterns in a given SQD, and the number of core questions between SQDs. Again, the choice of variables in a particular pattern in a given SQD is guided by the criteria, whether or not variables are considered to be independent, based on the results in

Tables 7a and 7b.

Table 9a gives the patterns in SQD 1. For SQD 1, we split each block into two subblocks, each containing 3 categorical variables and 1 continuous variable. Categorical variables that are considered to be independent of each other are given first priority of being placed together in the same subblock, meaning some subblocks will have variables that are not considered to be independent of one another. After the categorical variables are determined, a continuous variable is randomly placed into each subblock, creating the four subblocks, $1a = [Y_1, Y_5, Y_{11}, X_1]$, $1b = [Y_6, Y_{12}, Y_{13}, X_2]$, $2a = [Y_4, Y_7, Y_{10}, X_3]$, $2b = [Y_2, Y_3, Y_9, X_4]$, where '1' or '2' is the index for the blocks, and 'a' or 'b' is the index for the subblocks. Each '1' subblock is placed with a '2' subblock, creating a total of $p = 4$ patterns for this design.

Table 9b gives the patterns in SQD 2. For SQD 2, we create a design such that only 1 continuous variable is asked in a particular pattern. We first split the questions into four subblocks using the same methods for SQD 1, making sure the subblocks are different from SQD 1. These choices of subblocks are $1a = [Y_1, Y_6, Y_{12}, X_1]$, $1b = [Y_5, Y_{11}, Y_{13}, X_2]$, $2a = [Y_2, Y_3, Y_{10}, X_3]$, and $2b = [Y_4, Y_7, Y_9, X_4]$. Now, we need to split the categorical variables in each block into additional subblocks in order to have only one continuous variable in a particular pattern. We choose subblocks that are not mutually exclusive, so that each categorical variable is together in at least one pattern with every other categorical variable in this design. A pair of categorical variables that are considered to be independent will be placed in the same subblock together, and placed in a subblock with every other pair of variables. The final choices of these subblocks are $1c = [Y_1, Y_5, Y_{11}, Y_{12}]$, $1d = [Y_1, Y_6, Y_{11}, Y_{13}]$, $1e = [Y_5, Y_6, Y_{12}, Y_{13}]$, $2c = [Y_2, Y_4, Y_7, Y_{10}]$, $2d = [Y_2, Y_3, Y_4, Y_9]$, and $2e = [Y_3, Y_7, Y_9, Y_{10}]$. Each 'a' and 'b' subblock is placed with a 'c', 'd', or 'e' subblock with a different block number, creating a total of $p = 12$ patterns for this design.

Table 9c gives the patterns in SQD 3. For SQD 3, we increase the number of core questions, but the total number of questions in each pattern is still kept the same as other SQDs. Using the same criteria to determine the core questions earlier in this section, we select one categorical variable from each block to be an additional core question. We also choose one continuous variable in each block such that one of the questions relates to dogs and the other related to cats. The

four additional core questions chosen are Y_5, X_1, Y_7, and X_4. The remaining questions are split into subblocks of two questions each, using the same independence criteria as in SQD 1 and SQD 2. The only exception is a categorical variable that is more correlated with the other categorical variables in its block is likely to be placed with the remaining continuous variable in its block. The subblocks in this SQD are $1a = [Y_1, Y_{11}]$, $1b = [Y_{12}, X_2]$, $1c = [Y_6, Y_{13}]$, $2a = [Y_2, Y_4]$, $2b = [Y_{10}, X_3]$, and $2c = [Y_3, Y_9]$. Each '1' subblock is placed with a '2' subblock, creating a total of $p = 9$ patterns for this design.

4.1.4. Data Simulation and Results

The parameters, $\mathbf{B}$, $\mathbf{\Omega}$ and $\boldsymbol{\alpha}$, used for simulation are obtained using the MLE method and the complete data from the 2012 PDS with 3327 respondents who said they owned at least one dog and one cat in the calendar year 2011, the year when the survey was conducted. A population data with size $N = 50,000$ is simulated first. To generate the cell indicator for respondent i, we use a multinomial distribution with cell probabilities specified in (4.1). Once the categorical data is simulated, the continuous variables $\boldsymbol{X}$ conditioning on the cell m are simulated, following MVN and MVG separately. For the MVN case, we use the distribution as outlined in (2.2) where the vector $\mathbf{Z}_i$ is determined by the categorical variables on the right hand side of (4.2). Unlike the MVN, the MVG does not have closed form for its MLEs. So a method of profile likelihood is used to obtain the MLEs. A detailed explanation of the method used to find the MLE for MVG is in Appendix A.

Multiple population datasets are simulated following different correlation structures in the continuous variables and categorical variables. For the MVN case, in addition to using the MLE estimator for $\mathbf{\Omega}$, we also consider $\mathbf{\Omega}$ to be a diagonal matrix to introduce independence between the continuous variables. A discussion about how to manipulate the correlation matrix for the MVG variables is given in Appendix A. For the categorical variables, in addition to using the MLE for $\boldsymbol{\alpha}$, we also set $\alpha_{kk',(l_k,l_{k'})} = 0$ from (4.1) to see the effect of no interaction among the categorical variables. This produces a total of 4 simulated population datasets under each of MVN and MVG assumptions.

From the population of 50,000 observations, 100 MC samples of size $n = 1,000$ are drawn using

Question Patterns	Block 1								Block 2							
p	Y_1	Y_5	Y_{11}	X_1	Y_6	Y_{12}	Y_{13}	X_2	Y_4	Y_7	Y_{10}	X_3	Y_2	Y_3	Y_9	X_4
1	✓	✓	✓	✓					✓	✓	✓	✓				
2	✓	✓	✓	✓									✓	✓	✓	✓
3					✓	✓	✓	✓	✓	✓	✓	✓				
4					✓	✓	✓	✓					✓	✓	✓	✓

(a) Pattern Choices for SQD 1 in Section 4

Question Patterns	Block 1								Block 2							
p	Y_1	Y_5	Y_{11}	X_1	Y_6	Y_{12}	Y_{13}	X_2	Y_4	Y_7	Y_{10}	X_3	Y_2	Y_3	Y_9	X_4
1	✓			✓	✓	✓			✓	✓	✓		✓			
2	✓			✓	✓	✓			✓				✓	✓	✓	
3	✓			✓	✓	✓				✓	✓			✓	✓	
4		✓	✓				✓	✓	✓	✓	✓		✓			
5		✓	✓				✓	✓	✓				✓	✓	✓	
6		✓	✓				✓	✓	✓	✓			✓	✓		
7	✓	✓	✓			✓					✓	✓	✓	✓		
8	✓		✓		✓		✓				✓	✓	✓	✓		
9		✓			✓	✓	✓				✓	✓	✓	✓		
10	✓	✓	✓			✓			✓	✓					✓	✓
11	✓		✓		✓		✓		✓	✓					✓	✓
12		✓			✓	✓	✓		✓	✓					✓	✓

(b) Pattern Choices for SQD 2 in Section 4

Question Patterns	Block 1								Block 2							
p	Y_1	Y_5	Y_{11}	X_1	Y_6	Y_{12}	Y_{13}	X_2	Y_4	Y_7	Y_{10}	X_3	Y_2	Y_3	Y_9	X_4
1	✓	✓	✓	✓					✓	✓			✓			✓
2	✓	✓	✓	✓						✓	✓	✓				✓
3	✓	✓	✓	✓						✓				✓	✓	✓
4		✓		✓		✓		✓	✓	✓			✓			✓
5		✓		✓		✓		✓		✓	✓	✓				✓
6		✓		✓		✓		✓		✓				✓	✓	✓
7		✓		✓	✓		✓		✓	✓			✓			✓
8		✓		✓	✓		✓			✓	✓	✓				✓
9		✓		✓	✓		✓			✓				✓	✓	✓

(c) Pattern Choices for SQD 3 in Section 4

Table 9: Description of the SQD choices in Section 4

Distribution of Continuous Variables	$\log \pi_m$ model	J	"True" Rank of SQDs	Percent of r samples that match "true" rank	Avg. Time of True Density (×1000 sec)	Avg. Time of Approximate Density (×1000 sec)
Correlated	$\alpha_{kk',l_k l_{k'}} \neq 0$	50	$3 < 1 < 2$	100%	13,660.47	84.23
		100	$3 < 1 < 2$	100%		161.91
		1000	$3 < 1 < 2$	100%		1612.73
	$\alpha_{kk',l_k l_{k'}} = 0$	50	$3 < 1 < 2$	100%	13,654.97	83.15
		100	$3 < 1 < 2$	100%		160.14
		1000	$3 < 1 < 2$	100%		1596.41
Uncorrelated	$\alpha_{kk',l_k l_{k'}} \neq 0$	50	$3 < 1 < 2$	100%	13,646.59	82.86
		100	$3 < 1 < 2$	100%		160.93
		1000	$3 < 1 < 2$	100%		1611.55
	$\alpha_{kk',l_k l_{k'}} = 0$	50	$3 < 1 < 2$	100%	13,661.26	83.51
		100	$3 < 1 < 2$	100%		160.32
		1000	$3 < 1 < 2$	100%		1598.50

Table 10: Simulation results from the $r = 100$ MC samples for the 4 simulation scenarios under MVN case in Section 4.1. The first three columns indicate our assumption of the correlation between the continuous variables, whether or not two-way interaction are included in the model for $\log \pi_m$, and the value of J from (2.9) used in the calculation. The next columns provide the "true" rank of SQDs using the brute force sum, the percentages of the MC samples that give the same rank using the approximation method, and the computation time of the "true" distribution function per respondent and the computation time of the approximate distribution function per respondent.

simple random sampling. For each MC sample, we calculate the expected value of log-distribution function in (2.10) using the approximate distribution function of the observed from (2.9). Note, the parameter estimates $\hat{\boldsymbol{\alpha}}$, $\hat{\mathbf{B}}$, and $\hat{\boldsymbol{\Omega}}$ are recalculated under each MC sample. The MC means and standard errors of these parameters are provided in the supplemental file. A brute-force sum of the "true" expected log-distribution of the observed is calculated using (3.7) for each population. In this simulation study, there are totally 16 categorical variables with 5 or 6 missing in each pattern, and the average number of possible cells of $|S_{i,mis}|$ is 8,109. We set $J = \{50, 100, 1000\}$ to see how the the choice of J affects the estimated ranks of expected log-distribution of the observed.

The simulation results for 4 scenarios are shown in Tables 10 and 11 for MVN and MVG respectively. Both tables report the "true" rank of SQDs using the brute force sum, the percentages of the MC samples that give the same rank using the approximation method, and the average computation time of the true distribution function per respondent over 3 SQDs and the average computation time of the approximate distribution function per respondent over 3 SQDs.

Distribution of Continuous Variables	$\log \pi_m$ model	J	"True" Rank of SQDs	Percent of r samples that match "true" rank	Avg. Time of True Density ($\times$1000 sec)	Avg. Time of Approximate Density ($\times$1000 sec)
Correlated	$\alpha_{kk',l_k l_{k'}} \neq 0$	50	$3 < 1 < 2$	93%	27,336.42	180.75
		100	$3 < 1 < 2$	93%		357.21
		1000	$3 < 1 < 2$	93%		3565.307
	$\alpha_{kk',l_k l_{k'}} = 0$	50	$3 < 1 < 2$	94%	27,329.25	176.54
		100	$3 < 1 < 2$	94%		350.44
		1000	$3 < 1 < 2$	94%		3524.11
Uncorrelated	$\alpha_{kk',l_k l_{k'}} \neq 0$	50	$3 < 1 < 2$	100%	13,599.88	77.82
		100	$3 < 1 < 2$	100%		151.06
		1000	$3 < 1 < 2$	100%		1503.15
	$\alpha_{kk',l_k l_{k'}} = 0$	50	$3 < 1 < 2$	100%	13,576.74	74.77
		100	$3 < 1 < 2$	100%		145.37
		1000	$3 < 1 < 2$	100%		1454.72

Table 11: Simulation results from the r MC samples for the 4 simulation scenarios under MVG case in Section 4.1. The first three columns indicate our assumption of the correlation between the continuous variables, whether or not two-way interaction are included in the model for $\log \pi_{m,i}$, and the value of J from (2.9) used in the calculation. The next columns provide the "true" rank from the brute force sum, the percentages of the MC Samples that give the same rank of expected approximate log-distribution of the observed, and the computation time of the "true" log-distribution function per respondent and the computation of the approximate log-distribution function per respondent.

Comparing the ranks suggesetd by the approximate expected log-distribution function of the observed from the $r = 100$ MC samples to the "true" rank from the brute force sums in Table 10 and 11, 100% of the r samples matched the "true" rank for 6 of the 8 datasets. For the case of the correlated MVG, regardless of the categorical variable correlation structure, the percent of r samples that match the "true" rank is between 93% and 94%. When a MVG distribution is used, additional parameters γ_0 and β_0 need to be estimated, as explained in Appendix A. This introduces more estimation errors, thus the variance of the approximate expected log-distribution of the observed will increase and the estimated ranks become less precise.

As discussed in Section 3, computation time increase linearly as J increases and the MVG distribution takes about twice as long as the MVN assumption to compute the distribution function. Similarly, we compute the average ratios of the computation time of the true distribution to the computation time of the approximate distribution over all 4 scenarios under each of MVN and MVG assumption for $J = 50$. The ratio calculated as $\frac{1}{4}\sum_{j=1}^{4} \frac{\text{Avg. Time of true}}{\text{Avg. Time of approximated}}$ is 163.67 for

Distribution of Continuous Variables	J	Rank of Designs	Avg. Computation Time of Approximate Density ($\times$1000 sec)
Normal	50	$3 < 1 < 2$	90.36
	100	$3 < 1 < 2$	179.65
	1000	$3 < 1 < 2$	1789.52
Gamma	50	$3 < 1 < 2$	185.52
	100	$3 < 1 < 2$	369.66
	1000	$3 < 1 < 2$	3700.88

Table 12: Empirical results for the rank of approximate expected log-distribution of the observed for both MVN and MVG assumptions.

MVN case and 165.59 for MVG case. This means that it takes over 160 times as long to compute the true distribution function for the observed from (2.8) as compared to the approximate distribution function of the observed from (2.9). This suggests that our method is computationally efficient.

One last remark is that, in our simulation, we choose to have $d_y = 16$ categorical variables with about 5 or 6 missing questions, in order to make it it feasible to calculate the "true" expected log-distribution function. In large surveys, the number of categorical variables is often greater than 50 and the number of planned missing questions is likely greater than 30, making it more infeasible to calculate the true expected log-distribution of the observed. Our proposed approximation method provides an efficient and accurate way to identify the best SQD among a set of possible choices.

4.2. Empirical Data Results

In this section, we apply the same procedure from Section 4.1 to the real 2012 PDS data. The data comes from 3327 respondents from the 2012 PDS who owned at least one dog or one cat during the calendar year 2011 (i.e., $n = 3327$). Since the underling conditional distribution for continuous variables is unknown, we analyze the data using both MVN and MVG to account for potential skewness of the data. Because the same data set is used to motivate the simulation study in Section 4.1, the same set of MLEs are used in calculation of the observed log-distribution functions. We consider the same 3 design choices of SQDs as described in Tables 3a, 3b, and 3c. We also look at $J = \{50, 100, 1000\}$.

The results of the empirical analysis are found in Table 12. Under each distribution assumption,

the estimated rank stays the same when J increases from 50 to 100, then to 1,000. The results under both distribution assumptions suggest the same rank, which is also in line with the results in Section 4.1. As a conclusion, SQD 2 should be the best SQD among the three choices.

5. Conclusion

SQDs are more commonplace today with advancements in computing technology. Using the information presented in Sections 3 and 4, we conclude that the best choice of SQD design when presented with multiple possibilities can be accurately identified using the proposed approximation method in this article. While the correct order of the SQDs based on the amount of "information loss" is not the same across the simulation scenarios, our approximation method reach the same conclusion of design choice as the "true" expected log-distribution function indicates. The distribution functions discussed in this paper are general, and different probability density functions can also be applied, even though this paper only looks at the results for normal and gamma distributions.

Moreover, this article assumes that each pattern in a particular design has equal chance to be assigned to a respondent, i.e. the weight of $\frac{1}{P}$ is used in (2.10). However, this does not have to be the case. For example, given the demographic information of a respondent, some question patterns can have higher chances to be assigned to this particular respondent than other patterns. So the weight of $\frac{1}{P}$ will change to be a function of auxiliary information available. How to determine the weight function so that the information loss can be further reduced relative to the equal weight case is a research topic in the future.

Acknowledgements

The authors thank Cooperative Agreement (ISU award ID number 015735-00001) between the National Center for Food and Agricultural Policy and the Center for Survey Statistics and Methodology at Iowa State University.

[1] La Mar L. Adams and Darwin Gale. Solving the quandary between questionnaire length and response rate in educational research. *Research in Higher Education*, 17(3):231–240, 1982.

[2] Feray Adiguzel and Michael Wedel. Split questionnaire design for massive surveys. *American Marketing Association*, 45(5):608–617, 2008.

[3] T.W. Anderson. An introduction to multivariate statistical analysis. *New York:Wiley*, 1958.

[4] Y.M. Bishop, S.E. Fienberg, and P.W. Holland. Discrete multivariate analysis: Theory and practice. *M.I.T. Press*, 1974.

[5] United States Census Bureau. Current population survey, 2017. URL `https://www.census.gov/cps/data/cpstablecreator.html`.

[6] James O. Chipperfield and David G. Steel. Design and estimation for split questionnaire surveys. *Journal of Official Statistics*, 25(2):227–244, 2009.

[7] James O. Chipperfield and David G. Steel. Efficiency of split questionnaire surveys. *Journal of Statistical Planning and Inference*, 141:1925–1932, 2011.

[8] Edward Furman. On a multivariate gamma distribution. *Statistics and Probability Letters*, 2008. doi: 10.1016/j.spl.2008.01.012.

[9] L. A. Goodman. The analysis of cross-classified data. independence, quasi-independence, and interaction in contingency tables with or without missing entries. *Journal of the American Statistical Association*, 63:1091–1131, 1968.

[10] D.G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[11] Solomon Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[12] Roderick J. A. Little and Mark D. Schluchter. Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, 72(3):497–512, 1985.

[13] Jerzy Neyman and Egon Pearson. Ix: On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Societies A*, 231:694–706, 1933.

[14] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophy Magazine*, 5:157–175, 1900.

[15] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

[16] Trivellore E. Raghunathan and James E. Grizzle. A split questionnaire survey design. *Journal of the American Statistical Association*, 90(429):54–63, 1995.

Appendix A. Multivariate Gamma Distribution

Furman [8] theorized a method for a multivariate gamma distribution that takes into account the potential skewness of dependent continuous variables. Define a $(d_x + 1) \times 1$ column vector of independent and latent gamma distributed random variables $\boldsymbol{W}_i = [W_{i,0}, W_{i,1}, \ldots, W_{i,d_x}]^T$ for respondent i with shape parameter vector $\boldsymbol{\gamma}_i = [\gamma_{i,0}, \gamma_{i,1}, \ldots, \gamma_{i,d_x}]^T$ and rate parameter vector $\boldsymbol{\beta}_i = [\beta_{i,0}, \beta_{i,1}, \ldots, \beta_{i,d_x}]^T$, i.e. $W_{i,k} \overset{ind}{\sim} Gamma(\gamma_{i,k}, \beta_{i,k})$ for $i = 1, \ldots, n$ and $k = 0, \ldots, d_x$. Now, define a $d_x \times (d_x + 1)$ matrix $\mathbf{A}$ that connects $\boldsymbol{W}_i$ to the observed random variables, $\boldsymbol{X}_i = \mathbf{A}\boldsymbol{W}_i$, as

$$\mathbf{A} = \begin{bmatrix} \frac{\beta_0}{\beta_1} & 1 & 0 & \ldots & 0 \\ \frac{\beta_0}{\beta_2} & \frac{\beta_1}{\beta_2} & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\beta_0}{\beta_{d_x}} & \frac{\beta_1}{\beta_{d_x}} & \frac{\beta_2}{\beta_{d_x}} & \ldots & 1 \end{bmatrix}.$$

Then, we define the distribution of the continuous variables $\boldsymbol{X}_i$ as

$$\boldsymbol{X}_i \sim MVG(\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i), \tag{A.1}$$

where $E(X_{i,k}) = \frac{\sum_{o=1}^{k} \gamma_{i,o}}{\beta_i}, Var(X_{i,k}) = \frac{\sum_{o=1}^{k} \gamma_{i,o}}{\beta_i^2}$, and $Cov(X_{i,k}, X_{i,k'}) = Var(X_{i,k})$ for $k < k'$. The distribution function for the complete data for respondent i evaluated at the true values of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ is

$$g(\boldsymbol{X}_i \mid \boldsymbol{Y}_i, \boldsymbol{\gamma}_i, \boldsymbol{\beta}_i) = e^{-\beta_{i,d_x}\left(x_{i,d_x}\right)} \prod_{k=0}^{d_x} \frac{\beta_{i,k}^{\gamma_k}}{\Gamma(\gamma_k)} \prod_{k=2}^{d_x} \left[\left(x_{i,k} - \frac{\beta_{i,k-1}}{\beta_{i,k}} x_{i,k-1} \right)^{\gamma_k - 1} \right]$$
$$\times \int_0^{x^*} w_{i,0}^{\gamma_0 - 1} \left(x_{i,1} - \frac{\beta_{i,0}}{\beta_{i,1}} w_{i,0} \right)^{\gamma_1 - 1} dw_{i,0}, \tag{A.2}$$

where $w_{i,0}$ is the 1^{st} unobserved variable from $\boldsymbol{W}_i$ and $x^* = \min(\frac{\gamma_1}{\gamma_0} x_{i,1}, \frac{\gamma_2}{\gamma_0} x_{i,2}, \ldots, \frac{\gamma_{d_x}}{\gamma_0} x_{i,d_x})$ based on the relationship between $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$.

Just like the MVN as described in Section 2, we need to express the MVG in terms of the conditional distribution of the continuous variables given cell m. We express this as

$$\boldsymbol{X}_i \mid (\boldsymbol{Y}_i = m) \sim MVG(\boldsymbol{\gamma}_i, \boldsymbol{\beta}_{m,i}), \tag{A.3}$$

and the distribution function is

$$g(\boldsymbol{X}_i \mid \boldsymbol{Y}_i, \boldsymbol{\gamma}_i, \boldsymbol{\beta}_{m,i}) = e^{-\beta_{m,i,d_x}\left(x_{i,d_x}\right)} \prod_{k=0}^{d_x} \frac{\beta_{m,i,k}^{\gamma_k}}{\Gamma(\gamma_k)} \prod_{k=2}^{d_x} \left[\left(x_{i,k} - \frac{\beta_{m,i,k-1}}{\beta_{m,i,k}} x_{i,k-1} \right)^{\gamma_k - 1} \right]$$
$$\times \int_0^{x^*} w_{i,0}^{\gamma_0 - 1} \left(x_{i,1} - \frac{\beta_{m,i,0}}{\beta_{m,i,1}} w_{i,0} \right)^{\gamma_1 - 1} dw_{i,0}, \tag{A.4}$$

where $\boldsymbol{\beta}_{m,i} = [\beta_{m,i,0}, \beta_{m,i,1}, \ldots, \beta_{m,i,d_x}]^T$. We also assume that $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}$ and $\beta_{m,i,0} = \beta_0$ for all $i = 1, \ldots, n$ and $m \in M$.

Though the distribution function is expressed in (A.4), we reparameterize by $\mu_{m,i,k} = \frac{\gamma_k}{\beta_{m,i,k}} = \exp(\mathbf{C}_k \mathbf{Z}_i)$ for $k = 1, \ldots, d_x$ where γ_k and $\beta_{m,i,k}$ are the $(k+1)^{th}$ elements of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_{m,i}$, $\mathbf{Z}_i$ is

the $q \times 1$ vector of important categorical variables that determine the mean vector $\boldsymbol{\mu}_{m,i}$, assuming $q << M$ (an example of which is discussed in Section 3), $\mathbf{C}$ is the $d_x \times q$ matrix of coefficients associated with $\mathbf{Z}_i$ in $\boldsymbol{\mu}_{m,i}$, and $\mathbf{C}_k$ is the k^{th} row of matrix $\mathbf{C}$. Then, we substitute $\beta_{m,i,k} = \frac{\gamma_k}{\exp(\mathbf{C}_k \mathbf{Z}_i)}$ into the distribution function from (A.4) to compute the value of the distribution.

Assuming the continuous variables follow a gamma distribution means the MLE of the parameters $\mathbf{C}$ and $\boldsymbol{\gamma}$ are not known in closed form. Therefore, we use a gamma regression over different values of β_0 and γ_0 using methods of profile likelihood. The parameter estimates that produce the highest value of the log-likelihood are the MLE parameters. Again, we assume $\hat{\beta}_{m,i,k} = \frac{\hat{\gamma}_k}{\exp(\hat{\mathbf{C}}_k \mathbf{Z}_i)}$

Now, define a function g as the distribution function assuming the conditional distribution of the continuous variables follow a multivariate gamma distribution as that in (A.3). Then, using the parameterization as defined earlier, the distribution function of the observed evaluated at the MLE is

$$
\begin{aligned}
&g(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_{i,obs} \mid p, \hat{\boldsymbol{\gamma}}, \hat{\mathbf{C}}, \hat{\boldsymbol{\alpha}}) = \int g(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_{i,obs}, \boldsymbol{Y}_{i,mis} \mid p, \hat{\boldsymbol{\gamma}}, \hat{\mathbf{C}}, \hat{\boldsymbol{\alpha}}) d\boldsymbol{Y}_{mis} \\
&= \sum_{m \in S_{i,mis}} g(\boldsymbol{X}_{i,obs}, \boldsymbol{Y}_i = m \mid \hat{\boldsymbol{\gamma}}, \hat{\mathbf{C}}, \hat{\boldsymbol{\alpha}}) = \sum_{m \in S_{i,mis}} g(\boldsymbol{X}_{i,obs} \mid m, \hat{\boldsymbol{\gamma}}, \hat{\mathbf{C}}) * \hat{\pi}_{m,i} \\
&= \sum_{m \in S_{i,mis}} e^{-\hat{\beta}_{m,i,d_{x,obs}} \left(x_{i,d_{x,obs}}\right)} \prod_{k=0}^{d_{x,obs}} \frac{\hat{\beta}_{m,i,k}^{\hat{\gamma}_k}}{\Gamma(\hat{\gamma}_k)} \prod_{k=2}^{d_{x,obs}} \left[\left(x_{i,k} - \frac{\hat{\beta}_{m,i,k-1}}{\hat{\beta}_{m,i,k}} x_{i,k-1} \right)^{\hat{\gamma}_k - 1} \right] \\
&\times \int_0^{x^*} w_{i,0}^{\hat{\gamma}_0 - 1} \left(x_{i,1} - \frac{\hat{\beta}_0}{\hat{\beta}_{m,i,1}} w_{i,0} \right)^{\hat{\gamma}_1 - 1} dw \times \hat{\pi}_{m,i}, \qquad \text{(A.5)}
\end{aligned}
$$

where $x^* = \min(\frac{\gamma_1}{\gamma_0} x_{i,1}, \frac{\gamma_2}{\gamma_0} x_{i,2}, \ldots, \frac{\gamma_{d_{x_{obs}}}}{\gamma_0} x_{i,d_{x_{obs}}})$, $\hat{\boldsymbol{\gamma}}_{obs}$ is the common estimated vector of shape parameters for the observed continuous variables for respondent i, and $\hat{\boldsymbol{\beta}}_{m,i,obs}$ is the estimated matrix of rate parameters for the observed continuous variables for respondent i in who falls in cell m. Recall that $\hat{\boldsymbol{\beta}}_{m,i} = \frac{\hat{\boldsymbol{\gamma}}}{\exp(\hat{\mathbf{C}} \mathbf{Z}_i)}$ for $i = 1, \ldots, n$. $\hat{\boldsymbol{\beta}}_{m,i,obs}(\hat{\boldsymbol{\gamma}}_{obs})$ contains the elements of $\hat{\boldsymbol{\beta}}_{m,i}(\hat{\boldsymbol{\gamma}})$ that correspond to the observed continuous variables $\boldsymbol{X}_{obs}$ and the latent variable W_0 in pattern p.

The functions g and its approximation ($\hat{g}$) using the same methods as described in Section 2.2 can be substituted for f and $\hat{f}$ in (3.7) and (2.10) respectively for the comparison of the "true" expected log-distribution of the observed and the approximate expected log-distribution of the observed when the conditional distribution of the continuous variables is assumed to follow a gamma distribution in the simulation studies from Sections 3 and 4 and the empirical study from Section 4.2.

For the simulation studies, we first simulate $W_{i,k} \mid m \overset{ind}{\sim} Gamma(\gamma_k, \beta_{m,i,k})$ for $k = 1, \ldots, d_x$ and $W_{i,0} \sim Gamma(\gamma_0, \beta_0)$ indepedent of $(W_{i,k} \mid m)$ for $k = 1, \ldots, d_x$. Then, we obtain $\boldsymbol{X}_{i,k} = \mathbf{A}\boldsymbol{W}_{i,k}$ for each $i = 1, \ldots, N$. The same procedure is used in Section 4 for the cases where the correlation structure exists between the continuous variables. For the cases where the continuous variables are independent of each other, we simply simulate $X_{i,k} \mid m \overset{ind}{\sim} Gamma(\gamma_k, \beta_{m,i,k})$ for $k = 1, \ldots, d_x$ directly. In other words, we assume

$$
\mathbf{A} = \begin{bmatrix} 0 & \mathbf{0}_{1 \times d_x} \\ \mathbf{0}_{d_x \times 1} & \boldsymbol{I}_{d_x} \end{bmatrix},
$$

where $\boldsymbol{I}_{d_x}$ is a $d_x \times d_x$ identity matrix.