MARKETING SCIENCE
CELEBRATING 30 YEARS
Vol. 30, No. 4, July–August 2011, pp. 717–736
ISSN 0732-2399 | EISSN 1526-548X | 11 | 3004 | 0717

informs®
DOI 10.1287/mksc.1110.0648
© 2011 INFORMS

No Customer Left Behind: A Distribution-Free Bayesian Approach to Accounting for Missing *X*s in Marketing Models

Yi Qian
Department of Marketing, Kellogg School of Management, Northwestern University, Evanston, Illinois 60208,
yiqian@northwestern.edu

Hui Xie
Division of Epidemiology and Biostatistics, University of Illinois, Chicago, Illinois 60612,
huixie@uic.edu

In marketing applications, it is common that some key covariates in a regression model, such as marketing mix variables or consumer profiles, are subject to missingness. The convenient method that excludes the consumers with missingness in *any* covariate can result in a substantial loss of efficiency and may lead to strong selection bias in the estimation of consumer preferences and sensitivities. To solve these problems, we propose a new Bayesian distribution-free approach, which can ensure that no customer is left behind in the analysis as a result of missing covariates. In this way, all customers are being considered in devising managerial policies. The proposed approach allows for flexible modeling of a joint distribution of multidimensional interrelated covariates that can contain both continuous and discrete variables. At the same time, it minimizes the impact of distributional assumptions involved in covariate modeling because the method does not require researchers to specify parametric distributions for covariates and can automatically generate suitable distributions for missing covariates. We have developed an efficient Markov chain Monte Carlo algorithm for inference. Besides robustness and flexibility, the proposed approach reduces modeling and computational efforts associated with missing covariates and therefore makes the missing covariate problems easier to handle. We evaluate the performance of the proposed method using extensive simulation studies. We then illustrate the method in two real data examples in which missing covariates occur: a mixed multinomial logit discrete-choice model in a ketchup data set and a hierarchical probit purchase incidence model in a retail store data set. These analyses demonstrate that the proposed method overcomes several important limitations of existing approaches for solving missing covariate problems and offers opportunities to make better managerial decisions with the current available marketing databases. Although our applications focus on consumer-level data, the proposed method is general and can be applied to other marketing applications where other types of marketing players are the units of analysis.

Key words: CRM; hierarchical Bayesian; individual marketing; marketing mix variable; MCMC; missing covariates
History: Received: December 14, 2009; accepted: March 7, 2011; Eric Bradlow and then Preyas Desai served as the editor-in-chief and Fred Feinberg served as associate editor for this article. Published online in *Articles in Advance* June 6, 2011.

1. Introduction

Regression models are a main class of econometric tools for empirical marketing and economics studies. For example, a vast marketing literature studies brand choices and purchasing behaviors of individual consumers. In these studies, discrete-choice regression models have become workhorses in assessing determinants of consumer choices among differentiated products and stores. Purchase incidence models are frequently used to evaluate what drives household purchasing/shopping decisions. Regression models for conjoint analysis are essential in studying the importance of product attributes and the design of new products. To create the most effective marketing strategies, it is crucial to obtain valid and precise estimates of consumer preferences and responsiveness to marketing mix strategies.

In practice, missing data issues often arise in the applications of regression models (Little and Rubin 2002, Qian 2007). As noted in Blattberg et al. (2008, p. 301), “Missing variables is a fact of life for DBM [Database Marketing] applications.” The focus of this paper is on how to obtain valid and efficient estimates of regression relationships when some key covariates in a regression model are subject to missingness. To describe the issue more precisely, let Y denote the regression outcome, and let $X = (X_1, \dots, X_K)$ denote K covariates. We are interested in estimating a parametric regression model, $f_\theta(Y \mid X)$, where θ is a vector of parameters of interest. In marketing research, it is not uncommon that some variables in X are subject to missingness. For example, scanner panel data frequently have been used to calibrate discrete-choice models. Although scanner panel data have the advantage of reflecting consumer behaviors in real life, missing data issues can be severe. Missing data can occur in this situation because here, unlike experimental studies, typically only the values of marketing mix variables (e.g., price and coupon values) for the purchased products are recorded in the collection of scanner data (Erdem et al. 1999, henceforth referred as EKS). Other situations in which missing data could occur include item and unit nonresponse in surveys (Bradlow and Zaslavsky 1999), conjoint analysis (Bradlow et al. 2004), attrition and intermittent missingness in panel data (Qian and Xie 2010), data combination from different sources (Kamakura and Wedel 1997, 2000; Gilula et al. 2006; Feit et al. 2010), and one-to-one marketing (Khan et al. 2009). In sum, missing data are ubiquitous in marketing research.

A simple method for avoiding missing covariate problems is the complete-case analysis, where the standard regression analysis is applied to the subset of units with complete data. The complete-case analysis, although convenient, is inefficient, because it does not exploit the available information in those excluded units. Furthermore, when the probability of missingness depends on the regression outcome, exclusion of units with incomplete data leads to inconsistent estimation of population parameters because the resulting subsample is nonrepresentative. For example, in the above scanner panel data example, the missingness of those important marketing mix variables depends on the observed choice outcome Y. A complete-case analysis or an ad hoc method to fill in the missing values without accounting for this dependence is subject to self-selection bias in the estimation of price sensitivity and promotion effect (EKS 1999).

A general approach that avoids the drawbacks of the complete-case analysis is to posit a model for the covariates in X and then estimate a joint model for the outcome Y and the covariates in X. That is, one bases the inference on the following likelihood:

$$L(\theta, \phi; Y, X^{\text{obs}}) \propto \int f_\theta(Y \mid X^{\text{obs}}, X^{\text{mis}}) f_\phi(X^{\text{obs}}, X^{\text{mis}})\, dX^{\text{mis}}, \quad (1)$$

where X^{obs} and X^{mis} denote the observed and missing components of X, respectively; ϕ is a vector of parameters in the density function for the covariates. The above likelihood-based inference is valid when missingness is ignorable (Rubin 1976). Missingness is ignorable if the missing data mechanism is missing at random (MAR) and if the model governing the missing data mechanism has parameters distinct from the parameters θ and ϕ (i.e., parameter distinctness). The MAR assumption is satisfied if missingness is conditionally independent of the unobserved items in the data matrix, given the observed items in the data matrix. It is important to note that MAR is much less restrictive than missing completely at random (MCAR), which implies that missingness is independent of both unobserved and observed data values. The MAR is known to hold in the above scanner panel example, where the missingness of marketing mix variables depends only on the observed choice outcomes.[1]

In the above likelihood, a multidimensional covariate matrix must be carefully modeled, which can be a challenging task. One approach is to posit parametric distributions for X. One limitation of the parametric covariate modeling approach is that misspecification of the covariate distributions can result in a significant estimation bias and misleading inference. Therefore, care must be taken in modeling covariates. On the other hand, as shown in our empirical marketing applications, the multidimensional interrelated covariates usually contain a mixture of continuous, semicontinuous, and discrete variables that often exhibit features such as skewness, multimodality, discreteness, and zero-inflation. It is difficult to specify a joint parametric covariate model to account simultaneously for all the features in these variables. The problem is further exacerbated because, unlike the case in which all the data are observed, it is much harder, if not impossible, to verify whether distributional assumptions in a parametric covariate model are satisfied simultaneously for all the missing covariates. Furthermore, the computation burden can be heavy because one needs to evaluate multiple integration with respect to those missing covariates in the above likelihood. What is needed, then, is a method that minimizes the impact of covariate distributional assumptions and also reduces the extra modeling and computational burden involved in covariate modeling.

To address these challenges, we propose a new distribution-free Bayesian approach to estimating marketing models with multiple missing covariates. Our approach builds on a novel odds ratio modeling framework, first proposed by Chen (2004). The proposed method is robust in that no distributional

[1] The MAR assumption could be questionable in some marketing applications. For example, if the missingness of a consumer profiling variable (e.g., income) relates to the unobserved value of this variable, even after controlling for all the observed information including the observed consumer behaviors (e.g., purchase incidence outcome) and those observed profiling variables of the same consumer, the missingness then becomes missing not at random.

assumptions are required for modeling covariates. It can automatically generate suitable distributions for missing covariates and account for important distributional features including the ones mentioned above. Despite its full freedom from distributional assumptions, the method is flexible enough to allow for dependence among the covariates. Our analyses in both the real data and the simulation studies demonstrate that the proposed method improves the estimation of the marketing model parameters (e.g., consumer preference and marketing mix sensitivities) in the presence of missing covariates. Consequently, the proposed method offers the opportunity for better managerial decisions, such as optimal pricing and more accurate targeting. Furthermore, the proposed approach possesses modeling and computational simplicity, thereby rendering missing covariate problems easier to handle than they are under parametric covariate modeling approaches. As a Bayesian approach, the proposed method is ideal for individualized marketing when individual-level covariates are subject to missingness, and it ensures that no consumer is left behind in managerial considerations.[2] We hope this paper will contribute to expanding the set of tools researchers need to deal with missing covariate problems.

Our approach for missing covariate problems is general and can be applied to a wide range of marketing applications, including the following examples.

- *Market data.* As discussed above, in consumer databases, important variables often are missing (e.g., marketing mix variables, consumer profiles). The proposed method allows researchers to more efficiently and robustly estimate consumer preferences and sensitivity to marketing mix variables.
- *Survey data.* Survey studies are widely used and act as essential tools with which marketing researchers can answer important questions, particularly when market data are not available. Item and unit nonresponse are common in marketing survey data and can threaten effective analyses of survey data. The proposed method can be applied to this data type to improve estimation and inference.
- *Combining data from different sources.* It is becoming increasingly popular in marketing to combine data from different sources to overcome the limitations of a data set from a single source (e.g., Kamakura and Wedel 1997, 2000; Gilula et al. 2006; Feit et al. 2010). The missing covariate problem often occurs when data from different sources are combined (Feit et al. 2010). Our proposed method can be applied to address the problem.
- *Beyond customer-level data.* Marketing research is multifaceted. Regression models are often applied to study the behaviors of marketing players other than consumers, such as firms (manufacturers, retailers), organizations, and countries. Although our applications in this paper focus on the consumer-level data, the method can be applied to address missing covariate problems in empirical applications in which these other marketing agents are the units of analysis.

The rest of this paper is organized as follows. In §2 we review prior literature for missing covariate problems and describe our contributions to the field. In §3 we describe the model and estimation. In §4 we summarize the features and benefits of the proposed method. In §5 we apply the proposed method to two marketing applications with missing covariate problems. We conclude with a discussion in §6.

2. Literature Review and Contributions

Missing data are ubiquitous and problematic not only in marketing contexts but throughout empirical analysis in social sciences. Consequently, the subject has received an enormous amount of attention in the literature (e.g., Little and Rubin 2002, Schafer and Graham 2002, Daniels and Hogan 2008, Tsiatis 2006). A key message from the literature is that methods based on the probability models are preferred for dealing with missing data issues because these methods are based on the established statistical principles with known properties. Furthermore, as the assumptions in the analyses are made explicit, the methods can be evaluated clearly. Our review therefore focuses on model-based methods.

Little (1992) and Ibrahim et al. (2005) review various methods for dealing with missing covariates. Two main approaches for solving missing covariate problems are popular. The first one is the multiple imputation (MI) method (Little and Rubin 2002, Schafer 1997). MI imputes the missing values multiple times using draws from the predictive distributions of missing values. Each imputed data set is analyzed using standard complete-data methods. The resulting multiple estimates and inferences are combined to form one pooled inference using Rubin's combination rule. The second approach is the direct estimation method, in which a joint model for the regression outcome and the covariates is directly estimated using the likelihood or the posterior distribution under the model. Compared with MI, the direct estimation method does not require separate steps to create multiple data sets nor to pool estimates over these data sets.

In marketing literature, the missing covariate issue has also received much attention. Bradlow et al. (2004)

[2] An example is for online purchases, where ongoing predictions need to be made based on sparse individual-level data. We thank the associate editor for suggesting this.

develop an imputation learning model for the missing attributes in a conjoint model that improves model estimation. EKS (1999) and Feit et al. (2010) study missing covariate problems in discrete-choice models. EKS (1999) focus more on the robustness of covariate modeling and use a polynomial probability function to model the nonnormal feature of price and coupon values. One limitation of the approach is its inflexibility to model the potentially strong dependence among the covariates. Feit et al. (2010) use a multivariate normal (MVN) covariate model, which allows for correlated covariates. The MVN model is frequently used in missing data analysis because of its unique mathematical and computational properties.[3] However, there are also significant limitations in using a MVN model to handle missing covariate problems. First, the regression parameter estimators can be substantially biased if the parametric distributional assumption is incorrect. Second, the MVN covariate model does not allow for nonlinear relationships (e.g., a quadratic relationship or a relationship with interaction) among covariates. As shown in §4.5, this inflexibility in modeling covariates can cause bias in outcome regression estimates. Third, except for some limited types of regression models, the computational cost generally is high. For example, when the outcome is nonnormal or when the regression model contains nonlinear or interaction terms of missing covariates, the likelihood of the resulting joint model can involve intractable integrals with respect to missing data.

There are other important works in marketing literature related to missing data problems. Kamakura and Wedel (1997, 2000) develop MI methods to solve data fusion problems. Their novel idea is to use a finite mixture model to identify underlying homogeneous groups; the missing data are then stochastically imputed using observations from the same group. Gilula et al. (2006) propose a direct approach to data fusion, which directly estimates the joint distribution of the variables of interest. Although these methods have been highly successful in addressing the problems for which they were designed, they address issues that differ from missing covariate problems, which are the focus of this paper.[4] Another stream of research studies the missing outcome issues in regression models (Bradlow and Zaslavsky 1999, Ying et al. 2006, Qian and Xie 2010, Yang et al. 2010). In these studies, however, the missingness occurs in the outcome instead of in the covariates, thus obviating the need to model covariates.

There is emerging literature in statistics on weighting methods for missing data (Tsiatis 2006). This class of methods can be considered as a direct estimation approach in which the estimation is based on a set of inversely weighted estimating equations. A major motivation of the weighting methods is the robustness to model misspecifications. In addition to specifying a covariate model, a weighting method requires modeling how covariates are missing even if they are missing at random. Modeling how multiple covariates are missing may not be easy, and thus this approach may require much more modeling work. This is in stark contrast to the likelihood-based approaches in which there is no need to model missing data mechanisms when data are MAR. The benefit of the additional modeling in the weighting method is its property of *double robustness*. In missing covariate problems, this implies that as long as either the missing data model or the covariate model is correctly specified, the resulting inference is consistent. The method therefore protects against misspecifications of one of the two working models, although not against simultaneous misspecifications of both. Debates are ongoing regarding the relative merits of likelihood-based methods and weighting methods (e.g., see the discussions in Kang and Schafer 2007). We note three relevant points here. First, except for the special case of monotone missingness, finding the most efficient estimator in weighting methods is difficult, whereas a likelihood-based approach, if correctly specified, is most efficient. Second, for a general pattern of missingness, correctly specifying missing data models for all missing covariates is difficult, if not impossible. When the missing data models are misspecified, the validity of a weighting approach, similar to that of a likelihood-based approach, also depends on the robustness of the covariate model. Thus in the weighting approach, the robustness of covariate modeling is also important in achieving a good property. Third, because likelihood-based approaches are familiar to and used frequently by researchers for various reasons, it is highly relevant to develop robust methods for missing covariate problems within the likelihood-based framework.

As is reviewed above, to handle missing covariate problems in marketing models, there is a clear need for further research to find a method that is more robust and flexible yet also general enough and relatively simple to use. To that end, we contribute to the literature by proposing a new distribution-free

[3] For example, the multiple imputation procedure in SAS, PROC MI, uses the MVN model as the working model.

[4] For example, Kamakura and Wedel's data fusion approaches address problems for which all variables are treated equally, and no regression model in the form of $f(y \mid x)$ is considered. In contrast, in missing covariate problems, the regression model $f(y \mid x)$ is of primary interest, and it is important to use $f(y \mid x)$ to impute missing covariate values. Gilula et al. (2006) do employ a regression model $f(y \mid x)$. Their direct data fusion approach assumes that all covariates in x are fully observed and thus also does not address the missing covariate problems.

Bayesian approach that overcomes several important limitations of existing methods for missing covariates. Our approach builds on the novel semiparametric odds ratio model, first proposed by Chen (2004), and extends it to a Bayesian framework. In the extension, we study the Bayesian inference and carefully deal with the issue of efficient sampling algorithms under the unique semiparametric model. As such, our method shares many benefits of Chen's approach while overcoming its limitation in handling high-dimensional missing covariate problems and/or complex models. Because such problems are common in marketing applications, the proposed method is applicable to a much wider range of such applications. We offer a more detailed discussion on the features and benefits of the proposed method in §4.

3. Model and Notation

Following the notation in §1, let $f_\theta(Y \mid X)$ denote the density function of a parametric model, with its parameters θ being the main interest of the study. Below are some examples of the commonly used parametric regression models in marketing applications.

- *Generalized Linear Model (GLM).* The GLM assumes that the outcome Y_i is independently drawn from a distribution in the exponential family whose density function is

$$f_\theta(y_i \mid x_i) = \exp\left\{\frac{y_i \Psi_i(\beta, x_i) - b(\Psi_i(\beta, x_i))}{a(\tau)} + c(y_i, \tau)\right\}, \quad (2)$$

where Ψ_i is the canonical parameter as a function of regression parameter β; functions $b(\cdot)$ and $c(\cdot,\cdot)$ determine a particular distribution in the exponential family; and $a(\tau) = \tau/w$, where τ is the dispersion parameter and w is a known weight. The GLM includes normal, binomial, Poisson, Gamma, and inverse Gaussian models as special cases. It is frequently used in data analysis and forms the foundation for many more advanced marketing models.

- *Discrete-Choice Model* and *Conjoint Model.* Built from underlying marketing and economic theories (e.g., utility maximization), these models are well suited for estimating consumer preferences and sensitivity to marketing mix variables, market segmentation, and policy forecast.
- *Duration Model.* This model is useful for studying consumer purchase incidence behavior. Seetharaman and Chintagunta (2003) demonstrate examples of parametric duration models.
- *Models with heterogeneity.* All of the above models can be extended to incorporate consumer heterogeneity, a critical feature in marketing applications (Allenby and Rossi 1999).

As explained in §1, the covariate X can be subject to missingness in many marketing applications. To handle the problem, one needs to posit a covariate model, $f_\phi(X)$. We review a novel semiparametric odds ratio model below, first proposed by Chen (2004) and adopted here for covariate modeling. To illustrate the idea, we start with a simple case in which X contains only two variables, X_1 and X_2, which could be either continuous or discrete. Let $f(x_1, x_2)$ be the joint density function when $(X_1, X_2) = (x_1, x_2)$. Let (x_{10}, x_{20}) be a fixed and prespecified point in the sample space of X. The odds ratio is

$$\eta(x_2, x_{20}; x_1, x_{10}) = \frac{f(x_2 \mid x_1) f(x_{20} \mid x_{10})}{f(x_2 \mid x_{10}) f(x_{20} \mid x_1)}. \quad (3)$$

The odds ratio, as defined above, captures the dependence between X_1 and X_2. When X_2 is independent of X_1, the odds ratio $\eta(x_2, x_{20}; x_1, x_{10})$ is one for all possible values of x_1 and x_2. Chen (2004) shows that the conditional distribution can be reexpressed as

$$f(x_2 \mid x_1) = \frac{\eta(x_2, x_{20}; x_1, x_{10}) f(x_2 \mid x_{10})}{\int \eta(x_2, x_{20}; x_1, x_{10}) f(x_2 \mid x_{10})\, dx_2}.$$

As shown above, the main idea of the modeling approach is to decompose the conditional density $f(x_2 \mid x_1)$ into two parts: a conditional density function $f(x_2 \mid x_{10})$ and an odds ratio function $\eta(x_2, x_{20}; x_1, x_{10})$. These two parts can then be modeled separately. $f(x_2 \mid x_{10})$ is the density function of $X_2 = x_2$ at a fixed value, $X_1 = x_{10}$. Although it is not the same as the marginal density function $f(x_2)$, it behaves like a marginal density function instead of a conditional distribution, as will be shown later in this section. We call such a density a marginal-like density function. Using the odds ratio representation, the joint density for (x_1, x_2) is

$$\begin{aligned} f(x_1, x_2) &= f(x_1) f(x_2 \mid x_1) \\ &= \frac{\eta(x_2, x_{20}; x_1, x_{10}) f(x_2 \mid x_{10})}{\int \eta(x_2, x_{20}; x_1, x_{10}) f(x_2 \mid x_{10})\, dx_2} f(x_1). \end{aligned}$$

This idea can be extended to the case in which X contains more than two variables by using conditioning. Let $X = (X_1, \ldots, X_K)$ denote the K covariates. Its joint density function is

$$\begin{aligned} f_\phi(x_1, \ldots, x_K) &= f_{\phi_1}(x_1) \prod_{k=2}^{K} f_{\phi_k}(x_k \mid x_{k-1}, \ldots, x_1) \\ &= f_{\phi_1}(x_1) \prod_{k=2}^{K} \Big(\eta_{\gamma_k}(x_k, x_{k0}; x_{k-1}, \ldots, x_1, x_{(k-1)0}, \ldots, x_{10}) \\ &\qquad \cdot f_{\lambda_k}(x_k \mid x_{(k-1)0}, \ldots, x_{10})\Big) \\ &\quad \cdot \left(\int \eta_{\gamma_k}(x_k, x_{k0}; x_{k-1}, \ldots, x_1, x_{(k-1)0}, \ldots, x_{10}) \right. \\ &\qquad \left. \cdot f_{\lambda_k}(x_k \mid x_{(k-1)0}, \ldots, x_{10})\, dx_k\right)^{-1}, \end{aligned} \quad (4)$$

where each conditional distribution $f_{\phi_k}(x_k \mid x_{k-1}, \ldots, x_1)$ is reexpressed as a function of an odds ratio function and a marginal-like density function; ϕ_1 and ϕ_k denote the parameters in the marginal density function of X_1 and in the conditional density function of X_k, respectively; and γ_k and λ_k denote the parameters in the odds ratio function and the marginal-like density function for X_k, respectively. Let $x_{k1}, \ldots, x_{kN_k}$ be the unique observed values in the data set for X_k. A nonparametric model assigns probability mass $p_k = (p_{k1}, \ldots, p_{kN_k})$ to $f(x_k \mid x_{(k-1)0}, \ldots, x_{10})$, where a constraint is that $\sum_{l=1}^{N_k} p_{kl} = 1$ for every k. To relax the constraint, we reparameterize p_k as $\lambda_k = (\lambda_{k1}, \ldots, \lambda_{kN_k})$, such that $\lambda_{kl} = \ln(p_{kl}/p_{kN_k})$ for $l = 1, \ldots, N_k$. Thus, $p_{kl} = \exp(\lambda_{kl})/(\sum_{u=1}^{N_k} \exp(\lambda_{ku}))$. With this nonparametric model for $f(x_k \mid x_{(k-1)0}, \ldots, x_{10})$, the joint density function for X can be expressed as in Equation (4), where

$$
\begin{aligned}
f_{\phi_k}&(x_k \mid x_{k-1}, \ldots, x_1) \\
&= \frac{\eta_{\gamma_k}(x_k, x_{k0}; x_{k-1}, \ldots, x_1, x_{(k-1)0}, \ldots, x_{10}) f_{\lambda_k}(x_k \mid x_{(k-1)0}, \ldots, x_{10})}{\int \eta_{\gamma_k}(x_k, x_{k0}; x_{k-1}, \ldots, x_1, x_{(k-1)0}, \ldots, x_{10}) f_{\lambda_k}(x_k \mid x_{(k-1)0}, \ldots, x_{10})\, dx_k} \\
&= \frac{\sum_{l=1}^{N_k} 1_{\{x_k = x_{kl}\}} \eta_{\gamma_k}(x_{kl}, x_{k0}; x_{k-1}, \ldots, x_1, x_{(k-1)0}, \ldots, x_{10}) \exp(\lambda_{kl})}{\sum_{l=1}^{N_k} \eta_{\gamma_k}(x_{kl}, x_{k0}; x_{k-1}, \ldots, x_1, x_{(k-1)0}, \ldots, x_{10}) \exp(\lambda_{kl})}. \qquad (5)
\end{aligned}
$$

$f_{\phi_1}(x_1)$ can also be written in the above format by letting $\eta_{\gamma_1} = 1$. Equation (5) assigns probability mass $p'_k = (p'_{k1}, \ldots, p'_{kN_k})$ to $(x_{k1}, \ldots, x_{kN_k})$, where $p'_{kl} = f_{\phi_k}(x_k = x_{kl} \mid x_{k-1}, \ldots, x_1)$, $l = 1, \ldots, N_k$. As shown in Equation (5), the integral in the denominator of $f_{\phi_k}(x_k \mid x_{k-1}, \ldots, x_1)$ is replaced with a summation over a finite number of observed data values. This simplifies computation by avoiding the evaluation of integrals. In this modeling framework, the marginal-like distribution has been modeled nonparametrically, thus enhancing the robustness of the method. We follow Chen (2004) by using the following simple bilinear form for odds ratio functions:

$$
\begin{aligned}
\ln \eta_{\gamma_k}&(x_k, x_{k0}; x_{k-1}, \ldots, x_1, x_{(k-1)0}, \ldots, x_{10}) \\
&= \sum_{v=1}^{k-1} \gamma_{kv}(x_k - x_{k0})(x_v - x_{v0}). \qquad (6)
\end{aligned}
$$

As noted in Chen (2004), using the above simple bilinear form for odds ratio makes it easy to see that the model nests the popular generalized linear model as a special case. To see this, let x_k follow a GLM, as in Equation (2); its mean, μ_k, is

$$g(\mu_k) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{k-1} x_{k-1},$$

where $g(\cdot)$ is the canonical link. The corresponding odds ratio function can be shown to be

$$
\begin{aligned}
\ln \eta_{\gamma_k}&(x_k, x_{k0}; x_{k-1}, \ldots, x_1, x_{(k-1)0}, \ldots, x_{10}) \\
&= \sum_{v=1}^{k-1} \frac{\alpha_v}{a(\tau)} (x_k - x_{k0})(x_v - x_{v0}).
\end{aligned}
$$

The GLM model therefore has a bilinear form of odds ratio function, and the log-odds parameter γ_k is a reparametrization of the parameters in the familiar generalized linear model. On the other hand, the marginal-like density function, $f_{\lambda_k}(x_k \mid x_{(k-1)0}, \ldots, x_{10})$, is modeled parametrically in a GLM, whereas it is modeled nonparametrically in the distribution-free method. Thus, with the bilinear form of the odds ratio function, it is readily seen that the above distribution-free model nests the commonly used parametric GLM as a special case by eschewing the distributional assumptions. It is important to note that despite its full freedom from distributional assumptions, the distribution-free model is as flexible as a classical regression model in modeling relationships among variables. For example, similar to GLM, higher-order terms can be included in the odds ratio functions to model more complex relationships; nominal variables can be included in the odds ratio functions by using the dummy-variable technique.

We have developed an efficient Markov chain Monte Carlo (MCMC) algorithm to make inference of the joint model. One key step in the algorithm is to update the parameters in the semiparametric odds ratio model for covariates. The semiparametric odds ratio model is distinct from the conventional parametric models. A difficulty to overcome in Bayesian inference is finding an efficient method for posterior sampling. There is no conjugate prior to updating these parameters. The random-walk Metropolis-Hastings algorithm encounters the slow mixing problem because of the potentially high correlations between parameters. We employ a hybrid Monte Carlo (HMC) sampler (Duane et al. 1987) to update these parameters. The sampler exploits the local dynamics of the target distribution to propose a candidate draw, thus leading to a higher acceptance rate with fast mixing of the posterior draws. The details of the estimation algorithm are described in Appendix A of the electronic companion, available as part of the online version that can be found at http://mktsci.pubs.informs.org/.

4. Features and Benefits of the Proposed Method

To motivate our research, we conducted extensive simulation studies covering various covariate distributions and a wide range of types of regression models, including GLMs, mixed multinomial logit, and multivariate probit discrete-choice models, and hierarchical probit purchase incidence models with and without autocorrelation. Simulation results illustrate the limitations of the existing approaches to missing covariate problems and demonstrate the unique features and benefits of the proposed method. Because of the space limitation, we summarize the main conclusions below and move the details of simulation studies to Appendix B of the electronic companion.

4.1. Robustness

The simulation studies demonstrate the importance of robust covariate modeling. As shown in Appendix B.1 of the electronic companion, when the parametric distributional assumptions are correct, parametric covariate modeling approaches to handling missing covariates work well and can remove the large bias and substantial loss in estimation efficiency that occurred in complete-case analysis. When the covariate distribution is misspecified, however, a sizable bias and poor coverage rate can arise in the estimates of outcome regression parameters. In contrast, the proposed distribution-free method works well over different shapes of covariate distributions. The simulation studies detailed in Appendices B.5 and B.6 of the electronic companion confirm the robustness of the proposed method for other types of marketing models, and these studies show that the proposed method can correct for the bias in the estimates of consumer preferences and marketing mix sensitivities as a result of the misspecification of distributional assumptions in parametric covariate modeling approaches. These studies demonstrate the value of a nonparametric procedure as a robust approach to minimize the impact of distributional assumptions in handling missing covariate problems.

4.2. Simplicity

We investigate two aspects of simplicity: computational and modeling. We first discuss computational simplicity. Recall that the likelihood function of the joint model as specified in Equation (1) involves integration with respect to missing covariates. For a nonnormal regression model such as Poisson, the integral has no closed-form solution with those parametric covariate models. This leads to computational difficulty, as detailed in Appendix B.2 of the electronic companion. In contrast, updating X^{mis} in our proposed Bayesian approach is simple in that its conditional distribution is a closed-form multinomial distribution on a set of known values, and the probabilities in the multinomial distribution can be readily evaluated. As shown in simulation studies and in real data analysis, this computational simplicity helps to improve the convergence properties of the MCMC algorithm and to reduce the computational time compared with the parametric covariate modeling approach.

Next, we discuss modeling simplicity, which in our opinion is no less important than computational simplicity. As noted above, because selecting suitable distributions for missing covariates is important, careful modeling must be done. This presents a number of difficulties, however. A common and convenient parametric model might not be general enough to fit the unknown distributional shapes of missing covariates. The attempt to verify the parametric distributional assumptions for all covariates is also cumbersome—if not infeasible—when covariates are subject to missingness. The finite mixture modeling approach is more flexible in that it tries different parametric distributions for covariates and chooses the one that fits the data best. As shown in the simulation studies, its success depends critically on correctly specifying the number of mixture components. In general, selecting the correct number of mixture components is not an easy task and remains an active research area. This issue may be more challenging for missing covariate problems. Furthermore, avoiding local optima often requires that researchers repeatedly fit the joint model with a wide range of values for the number of mixture components. The workload can become computationally burdensome with a more complicated outcome regression model and in a high-dimensional missing covariate problem.

One important value of the proposed distribution-free approach for missing covariate problems, as a nonparametric procedure, is its *automatic* modeling feature; it does not require researchers to specify parametric covariate distributions and can automatically generate suitable distributions for missing covariates. The automatic feature of the proposed method frees up researchers' time to look for proper covariate distributions, and it allows them to invest their valuable time in other aspects of modeling, such as developing a more intelligent outcome regression model, thereby increasing the efficiency of their research activities.

Last but not least, our approach conditions on any fully observed covariates. This important feature has the advantage of further reducing the computational burden and increasing the modeling robustness compared with an approach based on a joint normal model or a mixture of joint normal models.

4.3. Efficiency

Our proposed method is also efficient in that the variability of regression parameter estimates using the proposed method is almost the same as that using the correct parametric covariate model.[5] This might at first seem counterintuitive to the concept that a highly parameterized model should have less variability in estimation than a nonparametric method. This should not be unexpected, however, because the regression parameters are high-level functionals of the probability density function. Such high-level functionals

[5] Note that the sample size used in the simulation study is only moderately large for marketing applications. For greater sample sizes, the efficiency loss would vanish.

undergo smoothing operations of integration of the probability density functions. As a result, the regression parameters theoretically can be estimated more efficiently with a nonparametric procedure than low-level functionals, such as the density function itself.[6]

In summary, by using the proposed distribution-free procedure, researchers would expect little or no loss of efficiency in typical marketing applications, and the method would perform as well as a correctly specified parametric covariate model. On the other hand, when the parametric covariate model is misspecified, a substantial bias can result. Because the nonparametric distribution-free method is still consistent in this case, the proposed distribution-free method will outperform the parametric method.

4.4. Convergence

The general convergence properties of the MCMC algorithm are established in Tierney (1994). Our approach uses proper priors, and the Markov chain in our algorithm is defined by a strictly positive transition kernel that is irreducible and aperiodic. Specifically, the samplers used in our algorithm to update model parameters are known to have these properties. The imputation step involves drawing from the set of observed values, each of which has a strictly positive probability to be visited. Therefore, theoretically, the MCMC chain will converge to its stationary distribution. Empirically, we use both time-series plot and numerical Geweke's diagnostic statistics (Geweke 1992) to check convergence. These statistics show that the Markov chains converge to the stationary distribution. In particular, the distributions of the Geweke's diagnostic statistics over the simulated data sets follow the null distribution of the test statistics. Chen (2004) establishes that under general regularity conditions, the maximum likelihood estimator (MLE) based on the semiparametric odds ratio covariate model is consistent and asymptotically normally distributed. Because of the asymptotic equivalence of the frequentist procedure and a Bayesian approach, we expect our Bayesian procedure to share this general property.

4.5. Flexibility

An important strength of the proposed method is its modeling flexibility, which refers to its ability to model the potentially complex dependence structure among covariates. The simulation studies in Appendix B.5 of the electronic companion show the importance of modeling the dependence among covariates; assuming independence in covariate modeling when covariates in fact depend on each other can lead to biased estimates in a brand-choice model. As reviewed in §2, despite its convenience in some special cases, the commonly used MVN model has limitations in accounting for a nonlinear relationship (e.g., a quadratic relationship or a relationship with interaction) among covariates. The simulation study described in Appendix B.3 of the electronic companion demonstrates that when the MVN covariate model is used in the presence of underlying nonlinear relationships among covariates, a significant amount of bias in outcome regression model estimates can occur. It is reasonable to believe that such bias would also exist when a latent MVN model is used to model discrete covariates, if nonlinear relationships exist among these covariates. In contrast, the odds ratio model is flexible enough to allow for such nonlinear relationships while not making any parametric distributional assumptions.

4.6. Generality

The proposed method is general in that it can handle a wide variety of types of continuous and discrete variables. It can also handle a general pattern of missingness.

4.7. Comparison with the Method of Chen (2004)

Our Bayesian method also compares favorably in terms of scalability to higher-dimensional missing data problems and to more complex models with the MLE method developed by Chen (2004). The MLE method requires evaluating the model likelihood. Although the integration in the likelihood is replaced by the summation over finite points, the number of terms to evaluate can become large with multiple missing covariates, which makes the MLE method computationally expensive. In contrast, the MCMC algorithm used in our Bayesian approach avoids evaluating likelihood, and thus it can handle much higher-dimensional missing data problems, commonly seen in marketing applications. The simulation study in Appendix B.4 of the electronic companion shows that the computational time increases exponentially for MLE as the number of missing covariates increases but only linearly with the Bayesian approach. The computational advantage would be even more dramatic for more complex models, such as when the covariate model involves interaction effects or when the outcome regression model becomes more complex. Because of computational difficulty, certain important data features that cannot be incorporated using the MLE approach can be handled with relative ease using the proposed Bayesian approach.

[6] Meier et al. (2004) had a similar finding in a different context. They found that when estimating mean survival time, a special type of regression parameter, the loss of efficiency of the nonparametric Kaplan–Meier procedure relative to parametric approaches is negligible.

Table 1 Summary Statistics of Marketing Variables

		Price				Coupon	
Brand	Choice share (%)	Mean	SD and correlation matrix			Proportion of no coupon (%)	Mean of no-zero coupon
Heinz	27.2	1.17	0.16	−0.43	−0.30	78.9	0.473
Hunt's	32.1	1.01		0.15	0.21	88.5	0.479
Store brand	20.8	0.71			0.10	100	NA

Other benefits of the proposed Bayesian approach compared with the MLE approach include exact inference in small samples, easy incorporation of useful information from other sources through prior specifications, and proper and convenient estimation of unit-level quantities. Using the proposed Bayesian approach for missing covariates, all these benefits become readily available to researchers.

5. Applications

5.1. Ketchup Data Set

Our first example is the ketchup data in the ERIM scanner panel data set provided by ACNielsen. We include three brands in the analysis: Heinz, Hunt's, and the store brand. These three major brands account for more than 80% share in the market. The sample period is from 1985 to 1987, covering two and a half years of transactions. Our analysis considers those purchases in the dominant package size of 32 ounces. The analysis sample contains 171 households that made a total of 1,093 purchases from one store in the Springfield market. Table 1 contains the summary statistics of the data set.

We employ a discrete-choice model to estimate the effect of pricing and coupon availability on the demand of ketchup. Let u_{itj} be the utility function of the jth brand for the ith consumer at purchase occasion t, and

$$u_{itj} = \psi_{0ij} + X_{itj}^T \psi_i + \epsilon_{itj}, \quad i = 1, \dots, N, \tag{7}$$

where the brand index $j = 1, 2, 3$ represents the Heinz, Hunt's, and store brands, respectively. In the utility function, ψ_{0ij} is the individual-specific preference for brand j, where ψ_{0i3} is normalized to be zero for identification purposes. The term X_{itj} is a vector of brand characteristics, and in our application, $X_{itj} = (P_{itj}, C_{itj})$, where P_{itj} and C_{itj} denote the price and coupon values, respectively, for the jth brand faced by the ith consumer at purchase occasion t. The parameter $\psi_i = (\psi_{1i}, \psi_{2i})$, where ψ_{1i} and ψ_{2i} are the individual-specific sensitivity coefficients for price and coupon, respectively. The term ϵ_{itj} is the idiosyncratic error term, unobservable to researchers. The researchers observe the consumers' choices among the brands. Let $Y_{it} = (Y_{it1}, \dots, Y_{itJ})$ be a vector of binary variables, where $Y_{itj} = 1$ if the consumer i chooses brand j at the purchase occasion t, and $Y_{itj} = 0$ otherwise. The random utility model assumes that Y_{it} is determined by the latent utility in the following way:

$$Y_{itj} = 1 \quad \text{iff } u_{itj} > u_{itj'} \quad \forall j' \neq j.$$

We assume that ϵ_{itj} follows an independent and identically distributed (iid) Type I extreme value distribution across purchase occasions, brands, and consumers. The probability for the choice of consumer i observed at time t is

$$f_{\beta_i}(Y_{it} \mid X_{it}) = \frac{\sum_{j=1}^{J} Y_{itj} \exp(V_{itj})}{\sum_{j=1}^{J} \exp(V_{itj})}, \quad \text{and}$$

$$V_{itj} = \psi_{0ij} + \psi_{1i} P_{itj} + \psi_{2i} C_{itj}.$$

We model consumer heterogeneity $\beta_i = (\psi_{0i1}, \dots, \psi_{0i,J-1}, \psi_{1i}, \psi_{2i})$, $J = 3$, as follows:

$$\beta_i \sim \mathrm{N}(\Pi Z_i, \Lambda^{-1}), \tag{8}$$

where Π is an $n_r \times n_z$ matrix, Z_i is a vector of length n_z containing consumer-level characteristics, and Λ is an $n_r \times n_r$ precision matrix. Π and Λ contain hyperparameters that describe the population distribution of the subject-specific parameters β_i.

The mixed multinomial logit (MNL) model specified above and its variant have been well studied and widely applied in economics and marketing (e.g., Guadagni and Little 1983, Kamakura and Russell 1989, Chintagunta et al. 1991, Gönül and Srinivasan 1993).[7] The Bayesian approach and the MCMC algorithm for the mixed MNL estimation are well established (Rossi and Allenby 1993, Allenby and Lenk 1994).

In practice, some of the important marketing mix variables are subject to missingness. The pioneering

[7] We use the standard form of the mixed MNL model for the following reasons: (1) This allows us to investigate and demonstrate the effects of missing covariates, the main theme of this paper, and to contrast our approach with the prior approach of EKS (1999) to the same missing covariate problem in a relatively straightforward setting. (2) It is reasonable to believe that more complicated brand choice models would not affect the relative performance of the methods in dealing with the missing covariates here.

and insightful work of EKS (1999) points out the problem of missing covariates in a discrete-choice model, where the prices and coupon availability are missing for the brands not purchased by any customer in the scanner panel data. They presented an econometric approach to correct for the self-selection bias that results from missing covariates. Specifically, they posited a model for the price and coupon process and based the inference on the likelihood

$$L(\Pi, \Lambda, \phi \mid Y, X^{\text{obs}}, Z) = \prod_{i=1}^{N} \int \Big[\int \prod_{t \in T_i} f_{\beta_i}(Y_{it} \mid X_{it}^{\text{obs}}, X_{it}^{\text{mis}}) \cdot f_{\phi}(X_{it}^{\text{obs}}, X_{it}^{\text{mis}})\, dX_{it}^{\text{mis}} \Big] f_{\theta}(\beta_i \mid Z_i)\, d\beta_i, \quad (9)$$

where T_i is the set of purchase occasions for consumer i, and $X_{it} = (\{P_{itj}\}, \{C_{itj}\}), j = 1, \ldots, J$. Their approach assumes that the covariates are independent: $f_\phi(X_{it}) = \prod_j f_{\phi_{pj}}(P_{itj}) \prod_j f_{\phi_{cj}}(C_{itj})$, where each density is separately modeled as a polynomial function.

When applying the proposed method to the brand choice model, our approach can be viewed as an extension of EKS in the following ways. First, we attempt to make the above approach more robust. To increase the modeling robustness, a nonparametric distribution function is applied to model each covariate. Because the parameters in the discrete-choice model are of primary interest, and the parameters in the covariate model are rarely of interest, a robust model with fewer assumptions about the covariate distribution is desirable. Second, we relax the assumption of independence among covariates. As will be shown later in this section, allowing for the correlations between covariates can further improve the estimates of brand preferences and sensitivity to marketing mix variables. Moreover, our generalization in this aspect makes the method applicable to other cases wherever correlations exist among covariates, as is shown in the second application of the paper. Third, our development uses a Bayesian framework, which has well-known advantages in the individual-level parameter estimation (Allenby and Rossi 1999).[8]

Strictly speaking, the discrete-choice model could not be directly estimated using available data because the price and coupon variables face serious missingness problems. In fact, no purchase transaction has all the values of P_{itj} and C_{itj} observed. Some sort of imputation method is required for filling in the missing values to estimate the discrete-choice model. We believe that a valid imputation method needs to take into account the dependence between the choice outcome and the missing covariate values, as well as the dependence among the covariates.

Here, we consider four imputation methods. The first method is a conventional simple imputation (SI) method as documented in EKS (1999). For any nonbought brand in a purchase, the conventional method searches in the database for any other consumer who bought this brand in the same store on the same day. If such a customer exists, the price at which that consumer bought the brand is used to fill in the missing price. If no such customer exists, we will fill in the missing price with the average weekly price. If there is no other weekly sale for this product, the average price on the nonpromotion days in the study period is used to fill in the missing price values. For a coupon, the SI method assumes that the coupon value is zero for any nonbought brand. It is important to note that this procedure only uses the observed price and coupon values to fill in missing values. Although this type of simple imputation procedure is commonly used in practice, e.g., by scanner panel data provider to fill in missing prices, it does not consider the potentially strong dependence between the choice outcome and these marketing mix variables, and thus it can lead to a strong self-selection bias.

The second and third imputation methods apply the proposed distribution-free method to model the price and coupon distributions. Let the covariate $X_{it} = (P_{it1}, P_{it2}, P_{it3}, C_{it1}, C_{it2}, C_{it3})$, where the third subscript takes a value of 1, 2, or 3 representing the Heinz, Hunt's, and the store brands, respectively. A semiparametric odds ratio model as specified in Equations (4) and (5) is applied to model X_{it}, with the following bilinear forms of the odds ratio functions:

$$\ln \eta(P_{itj}; P_{itj'}, P_{itj0}) = \sum_{j'=1}^{j-1} \gamma_{jj'}^{P} (P_{itj} - P_{j0})(P_{itj'} - P_{j'0}) + \gamma_{j0}^{P} (P_{itj} - P_{j0})(P_{itj0} - P_{j00}),$$

$$\ln \eta(C_{itj}; P_{itj}) = \gamma_j^C (C_{itj} - C_{j0})(P_{itj} - P_{j0}).$$

[8] Other related work includes Chiang (1995) and Musalem et al. (2008). Chiang (1995) also recognized the problem of missing marketing mix variables, but explicit modeling of the problem is not the emphasis of that work. A general issue with such ad hoc approaches to handling missing covariates is that they do not account for dependence between the regression outcome and missing covariates, which can lead to selection bias in outcome regression estimates. Furthermore, assumptions involved in such ad hoc approaches often are hidden, which makes it difficult to assess the validity of these methods. Musalem et al. (2008) developed a new Bayesian method to estimate demand models when only aggregate data are available. Their approach is to simulate latent (i.e., entirely missing) consumer-level data that are consistent with the aggregate data. The missing covariates are of binary types. Our approach considers more detailed data (e.g., coupon face values instead of coupon usage indicators) where covariates can contain a mixture of continuous and discrete variables; this requires more careful modeling.

Practical applications pose situations in which other observations can provide useful information about the missing values of a covariate. An important strength of the proposed method is its flexibility in allowing such information to be used through the odds ratio functions, despite its full freedom from distributional assumptions. For example, the prices of different brands might be correlated because of price competition or price conformity. This price correlation might be related to the market structure. The correlations among prices of the three brands are significant. Table 1 shows that the correlation coefficients among prices of the three brands, using the observed data, are as follows: -0.43 (between Heinz and Hunt's), -0.30 (between Heinz and the store brand), and 0.21 (between Hunt's and the store brand). In the above odds ratio model, the log odds parameter $\gamma_{jj'}^{P}$ captures the correlations among the prices of different brands and therefore allows the price of a brand not purchased by a customer to be informed by that of the customer's purchased brand. The price and coupon values may also be correlated, and the parameter γ_j^C captures such potential correlations. Because we never simultaneously observe the coupon availability and face values of all brands for any consumer, we opt for a simpler analysis that assumes independence among coupon values of different brands and that fixes the corresponding parameters in the odds ratio functions at zero. Observations from other consumers may provide useful information about the missing price values. For example, even though a price for a nonpurchased brand might not be observed for a consumer, another consumer might purchase that brand during the same time period. To incorporate such information, we create a new variable, P_{itj0}, which denotes the price paid for the brand j at the time t by a customer other than consumer i. If no such customer exists at time t, we search for the customer who purchased the brand in the nearest time and use that price as P_{itj0}. Our model then allows the covariate distribution to depend on such information through the odds ratio functions. In the above model, the fixed and prespecified points for each variable are chosen to be the smallest observed values for the price and coupon variables, respectively.[9] More details about the distribution-free procedure and its estimation algorithm can be found in Appendix A.2 of the electronic companion.

The above model is named "DF Model II." For a comparison, we fit a model named "DF Model I," which assumes all the covariates in X_{it} are independent of each other. This is equivalent to setting all the log odds parameters at zero in the above odds ratio functions. In this aspect, DF Model I is akin to the analysis of EKS (1999) in that it ignores the potential dependence among marketing mix variables. On the other hand, DF Model I assumes a nonparametric distribution for each covariate instead of the parametric polynomial distribution used in EKS (1999). We estimate the model using the priors and the MCMC algorithm described in Appendix A.2 of the electronic companion. For the purpose of comparison, we also fit a parametric MVN covariate model for (X_{it0}, X_{it}), where $X_{it0} = (P_{it10}, P_{it20}, P_{it30})$. It is important to note that the MVN method models X_{it0}, whereas DF Models I and II condition on it. The DF models therefore further reduce the computational workload while increasing the modeling robustness. All the models run the MCMC sampler, which discards the first 30,000 iterations as the burn-in period and keeps every 10th draw for the next 500,000 iterations. We use Geweke's diagnostic to check the convergence. The chains were found to converge well, except for the parameters related to the coupon variable in the SI model.[10] The computational times to obtain 1,000 effectively independent draws for the population regression parameters are 1 hour and 40 minutes, 35 minutes, and 38 minutes for the MVN model, DF Model I, and DF Model II, respectively. The ratios of the average f statistics of the population regression parameters in the Markov chains, relative to those from the MVN model, are 0.91 and 0.93 for DF Models I and II, respectively.[11] Because of its computational simplicity, as explained in §4.2, we can see that the proposed distribution-free method takes significantly less time than the parametric MVN model with a somewhat smaller autocorrelation.[12]

[9] Theoretically, the choice of the fixed points can be arbitrary. For example, when we use the largest values instead of the smallest values, the estimation results have negligible changes in that the changes of all parameter estimates are well within 3% of those estimates using the smallest values. Practically speaking, an absurd choice of these points, such as points extraordinarily remote from observed data points, could lead to computational instability. We recommend using a fixed-point value within the smallest and largest observed values.

[10] As explained later in this section, the coupon variable does not converge in the SI model because the ad hoc method to fill in the coupon variable in the SI method creates a strong self-selection bias. EKS (1999) also note this problem.

[11] Because we use the same sampler to update parameters in the outcome regression models for all methods, the difference (or ratio) of computational times (or f statistics) among methods can be attributed to differences in methods for dealing with missing covariates. For high-dimensional missing covariate problems, the importance-sampling-type algorithm is infeasible, so we use the data augmentation algorithm described in Appendix B.2 of the electronic companion for the MVN model. The f statistic, as defined in Rossi et al. (2005, Chapter 3.10.3), measures the strength of autocorrelation in a Markov chain, with a higher f value indicating stronger autocorrelation.

[12] Given that a finite mixture of the MVN model is more complicated than the MVN model, it is expected that the approach would

Table 2 Estimation Results in the Ketchup Purchase Data

Parameter	SI model	MVN model	DF Model I	DF Model II
		Choice outcome model		
Intercept (Heinz)	1.8 (0.28)	3.5 (0.44)	3.7 (0.45)	3.0 (0.36)
Intercept (Hunt's)	1.6 (0.20)	3.1 (0.36)	3.3 (0.36)	2.8 (0.31)
Price	−3.4 (0.50)	−6.1 (0.76)	−6.6 (0.86)	−5.4 (0.66)
Coupon	53.6 (3.32)	2.4 (0.56)	4.4 (1.28)	3.5 (1.24)
Σ_{11}	2.2 (1.1)	4.2 (1.5)	6.1 (2.6)	4.8 (1.8)
Σ_{22}	1.4 (0.6)	3.7 (1.1)	4.9 (1.7)	4.0 (1.2)
Σ_{33}	2.8 (2.1)	15.4 (5.8)	18.5 (8.2)	15.2 (5.9)
Σ_{44}	3.3 (8.8)	1.44 (1.21)	6.9 (6.4)	3.2 (2.1)
Σ_{12}	0.98 (0.77)	3.1 (1.2)	4.6 (2.0)	3.5 (1.4)
Σ_{13}	−1.56 (1.48)	2.8 (2.3)	1.4 (4.2)	2.9 (2.7)
Σ_{14}	0.24 (1.87)	0.3 (0.4)	1.6 (3.8)	0.4 (2.1)
Σ_{23}	−0.91 (1.04)	3.6 (2.1)	2.1 (3.5)	3.2 (2.4)
Σ_{24}	0.28 (1.28)	0.6 (0.8)	1.5 (3.0)	0.3 (1.8)
Σ_{34}	−0.18 (2.96)	−0.4 (0.2)	−2.2 (7.8)	−0.8 (3.9)
		Covariate model[a]		
(1) Price model				
Heinz				
p_{11}			0.21 (0.02)	0.20 (0.02)
p_{12}			0.48 (0.03)	0.49 (0.03)
p_{13}			0.19 (0.02)	0.22 (0.02)
p_{14}			0.008 (0.005)	0.004 (0.005)
p_{15}			0.054 (0.014)	0.051 (0.02)
p_{16}			0.053 (0.014)	0.035 (0.02)
Hunt's				
p_{21}			0.31 (0.02)	0.31 (0.02)
p_{22}			0.24 (0.02)	0.27 (0.02)
p_{23}			0.36 (0.03)	0.37 (0.04)
p_{24}			0.03 (0.01)	0.02 (0.01)
p_{25}			0.0095 (0.007)	0.005 (0.01)
p_{26}			0.025 (0.01)	0.016 (0.01)
p_{27}			0.022 (0.01)	0.018 (0.01)
Store brand				
p_{31}			0.11 (0.02)	0.10 (0.02)
p_{32}			0.66 (0.03)	0.66 (0.03)
p_{33}			0.134 (0.02)	0.164 (0.02)
p_{34}			0.024 (0.01)	0.020 (0.02)
p_{35}			0.08 (0.02)	0.06 (0.02)
(2) Coupon model				
Heinz				
c_{11}			0.87 (0.02)	0.84 (0.02)
c_{12}			0.005 (0.003)	0.003 (0.002)
c_{13}			0.040 (0.009)	0.047 (0.01)
c_{14}			0.035 (0.009)	0.039 (0.01)
c_{15}			0.002 (0.002)	0.0011 (0.002)
c_{16}			0.050 (0.011)	0.066 (0.02)
c_{17}			0.0027 (0.002)	0.0029 (0.002)
Hunt's				
c_{21}			0.93 (0.012)	0.90 (0.01)
c_{22}			0.002 (0.002)	0.002 (0.002)
c_{23}			0.002 (0.002)	0.001 (0.002)
c_{24}			0.005 (0.003)	0.006 (0.004)
c_{25}			0.06 (0.01)	0.088 (0.02)
c_{26}			0.002 (0.002)	0.004 (0.003)
Store brand				
c_{31}			1.00 (0.00)	1.00 (0.00)
(3) Dependence				
γ_{10}^{P}				39.1 (3.7)
γ_{20}^{P}				41.0 (8.0)
γ_{21}^{P}				−22.3 (5.1)
γ_{30}^{P}				28.4 (7.9)
γ_{31}^{P}				0.82 (4.0)
γ_{32}^{P}				12.97 (6.5)
γ_{1}^{C}				2.4 (1.8)
γ_{2}^{C}				9.2 (1.8)
Marginal LL	−1,164.40	−1,122.63	−3,428.16	−2,631.89

Notes. Presented are the posterior means (posterior SD) for each parameter. The parameter p_{bl} in the price model is the estimated marginal probability mass at the lth price value of brand b, where these price values, in the order presented in the table, are as follows: for Heinz, 0.99, 1.19, 1.39, 1.45, 1.49, 1.59; for Hunt's, 0.89, 0.99, 1.19, 1.39, 1.45, 1.49, 1.59; and for the store brand, 0.59, 0.69, 0.89, 0.95, 0.99. The parameter c_{bl} in the coupon model is the estimated marginal probability mass at the lth coupon value of brand b, where these coupon values, in the order presented in the table, are as follows: for Heinz, 0.00, 0.25, 0.30, 0.36, 0.40, 0.50, 0.90; for Hunt's, 0.00,0.30, 0.36, 0.40, 0.50, 1.00; and for the store brand, 0.00. LL, log likelihood.

[a] Because of space limitations, we report the estimates of the MVN covariate model in Appendix Table 6 of the electronic companion.

Table 2 summarizes the model estimation results. The result shows that the population price sensitivity parameter β_1 is estimated to be -3.4, -6.1, -6.6, and -5.4 by the SI method, MVN model, DF Model I, and DF Model II, respectively. The price sensitivity estimate from the SI method is substantially smaller. This is so because the SI method uses the accepted (i.e., observed) prices to fill in those missing price values. As the accepted price tends to be on the lower end of the underlying price distribution, this tends to underestimate the missing prices, therefore leading to an underestimation of the price sensitivity. The estimates of the price sensitivity parameter for the MVN model, DF Model I, and DF Model II are closer to each other, although the estimate from DF Model II is noticeably smaller in size.

A more serious selection bias for the coupon effect estimate can occur if it is not accounted for properly. The estimate for the population coupon effect β_2 is 53.6, 2.4, 4.4, and 3.5 for the SI method, MVN model, DF Model I and DF Model II, respectively. As noted in EKS (1999), the MLE estimate under the SI method theoretically will be infinity because the ad hoc method for filling in coupon values used in the SI method creates a strong self-selection bias. Because the Bayesian approach for the SI method puts a prior with mean at zero and a finite variance that helps stabilize the model estimation, however, the posterior mass of the parameter lies on a large number instead of on infinity.[13] The other three methods yield much more comparable coupon sensitivity estimates.

It is also important to note that the standard errors of the estimates in DF Model II are smaller than those in DF Model I, in some cases by about 50%. This is because DF Model II has allowed correlations in marketing mix variables, therefore reducing the variability when imputing the missing marketing mix values. As a result, the estimation efficiency increases.

Table 2 also reports the logarithm of the marginal density (LMD) of the data for four models, using the method of Raftery et al. (2007).[14] The LMD from the SI method cannot be compared to those from the two DF models because it does not model the covariate—neither can that of the MVN method because it views covariates as data types different from the DF models. The LMD from DF Models I and II can be compared with each other. We find DF Model II has a substantially larger marginal likelihood, with the difference of the LMD being 796.27, indicating that DF Model II fits data significantly better than DF Model I.

In Table 2 we also report the estimated nonparametric marginal distributions of price and coupon for each brand. In DF Model I, the distributions are calculated as the probability mass $p_{kl} = \exp(\lambda_{kl})/(\sum_{u=1}^{N_k}\exp(\lambda_{ku}))$ for the lth unique observed value of the kth covariate, where $\{\lambda_{kl}\}$ are the parameters defined in Equation (5). In DF Model II, this corresponds to the marginal-like distribution. We use the simulation method to calculate the marginal distribution in DF Model II. In the analysis, the observed price and coupon values are rounded to the second decimal place. The results change little when holding more decimal places. We find in Table 2 that the marginal distributions of price and coupon values are clearly nonnormal, which shows that the MVN model is not adequate for such data. Furthermore, we find significant dependence structures among marketing mix variables, as shown by the log odds parameters, most of which have 95% credible intervals excluding zero. Therefore, a model that does not account for such strong correlations, such as DF Model I, is inadequate in this aspect. In Appendix B.5 of the electronic companion, we further conduct simulation studies that demonstrate the advantages of the proposed method in repeated samples.

5.2. Managerial Implications

The above analyses demonstrate the potential bias of parameter estimates, caused by improper imputations of missing price and coupon values. Such bias can translate into substantial bias when the impact of a managerial policy of interest is assessed. As an example, Table 3 reports a simulation result that investigates the effects of a 20% cut in Hunt's price on market shares. Starting from almost identical market shares before the price cut, the percentage increases of the Hunt's market share for such a price cut are predicted to be 33%, 59%, 61%, and 53% for the SI method, MVN model, DF Model I, and DF Model II, respectively.

Table 3 The Impact of Price Cut on Market Shares

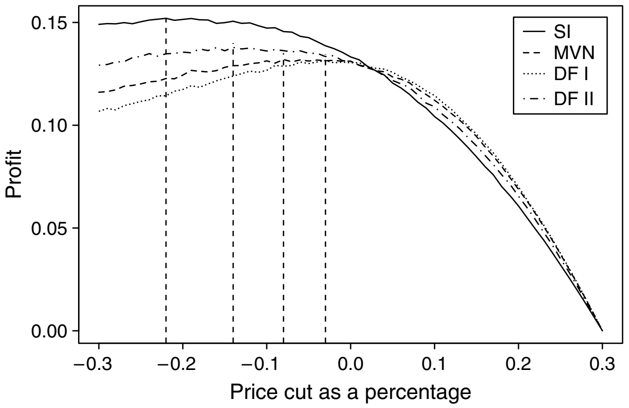
	SI (%)		MVN (%)		DF Model I (%)		DF Model II (%)	
Brand	Baseline	After cut	Baseline	After cut	Baseline	After cut	Baseline	After cut
Heinz	35.5	27.2	33.8	22.5	33.2	21.3	33.0	23.4
Hunt's	41.2	55.1	40.1	63.9	40.3	65.1	40.3	61.5
Store	23.2	17.7	26.1	13.6	26.5	13.6	26.7	15.1

take considerably more time, particularly if the correct number of mixture components needs to be assessed.

[13] Another ad hoc approach is to impute missing coupon values for a brand with the average of coupon face values used by customers who bought the brand, instead of zero. This creates selection bias in the opposite direction: because the coupon value used for a purchased brand tends to be on the high end of its distribution, one tends to overestimate the coupon values of nonpurchased brands, leading to a downward bias in coupon effect estimates. Indeed, because of the downward bias, the sign of the coupon effect estimate when using this ad hoc approach to impute missing coupon values in the data set becomes negative, which is highly implausible and inconsistent with the economic theory of the coupon effect.

[14] We thank the associate editor for pointing out the availability of the new method.

Figure 1 Comparison of Optimal Prices Determined from Different Approaches



Notes. There are substantial differences in the optimal price suggested by the different methods. The suggested optimal price cuts are −22%, −8%, −3%, and −14% for the SI method, MVN model, DF Model I, and DF Model II, respectively.

In practice, the above results can inform the optimal price for a manufacturer. To achieve this goal, we consider the optimal price that maximizes the profit. Specifically, we consider the profit function $p_j = M_j(Price_j - Cost_j)$, where M_j is the market share for alternative j at a specified price value $Price_j$, and $Cost_j$ is the cost of the alternative j. We calculate the market shares and profits for Hunt's for a grid of values of potential price cuts. We assume the Hunt's cost is 70% of its original price. Figure 1 presents the profit functions for a range of price-cut values based on the estimation results of the different methods. As we can see, different methods suggest substantially different optimal prices: the suggested optimal price cuts are −22%, −8%, −3%, and −14% for the SI method, MVN model, DF Model I, and DF Model II, respectively. Compared with DF Model II, other methods lead to substantially different pricing suggestions. As shown in the figure, substantial differences also exist in the profits predicted by different methods.

5.3. Retail Store Purchase Incidence Data Set

Our second application illustrates the method in a purchase incidence model using a data set in a frequent shopper database from a retail store in China. One managerial question is to study what affects purchase incidence of customers and to profile the customers based on some identifiable variables. Such analysis is often of interest for customer relationship management (CRM) and market segmentation. Our sample contains purchase records during four years for 455 frequent shoppers who made their initial purchases within the first year. The household characteristics considered important in profiling include *Firstbuy* (the purchase amount at the first visit), *Age* (age of the consumer), *Marriage* (marital status of the consumer), *Income* (household income), *Kidslt18* (the number of children at home younger than 18), and *DTS* (travel distance to the store). In the data set, because of item nonresponse, *Kidslt18* is missing 8.7%, *Age* is missing 19.5%, *Income* is missing 7.6%, and *DTS* is missing 9.7% of their values. There is no missingness for *Firstbuy* and *Marriage*.

We employ a discrete-time purchase incidence model to study the interpurchase time of these consumers (Gupta 1991, Wedel et al. 1995). A distinct feature of the discrete-time survival model is its ability to model explicitly the effects of marketing mix variables on consumer behavior at the times when they did not visit the store (Seetharaman and Chintagunta 2003). Let u_{it} denote consumer i's utility to purchase in the store at month t. We assume

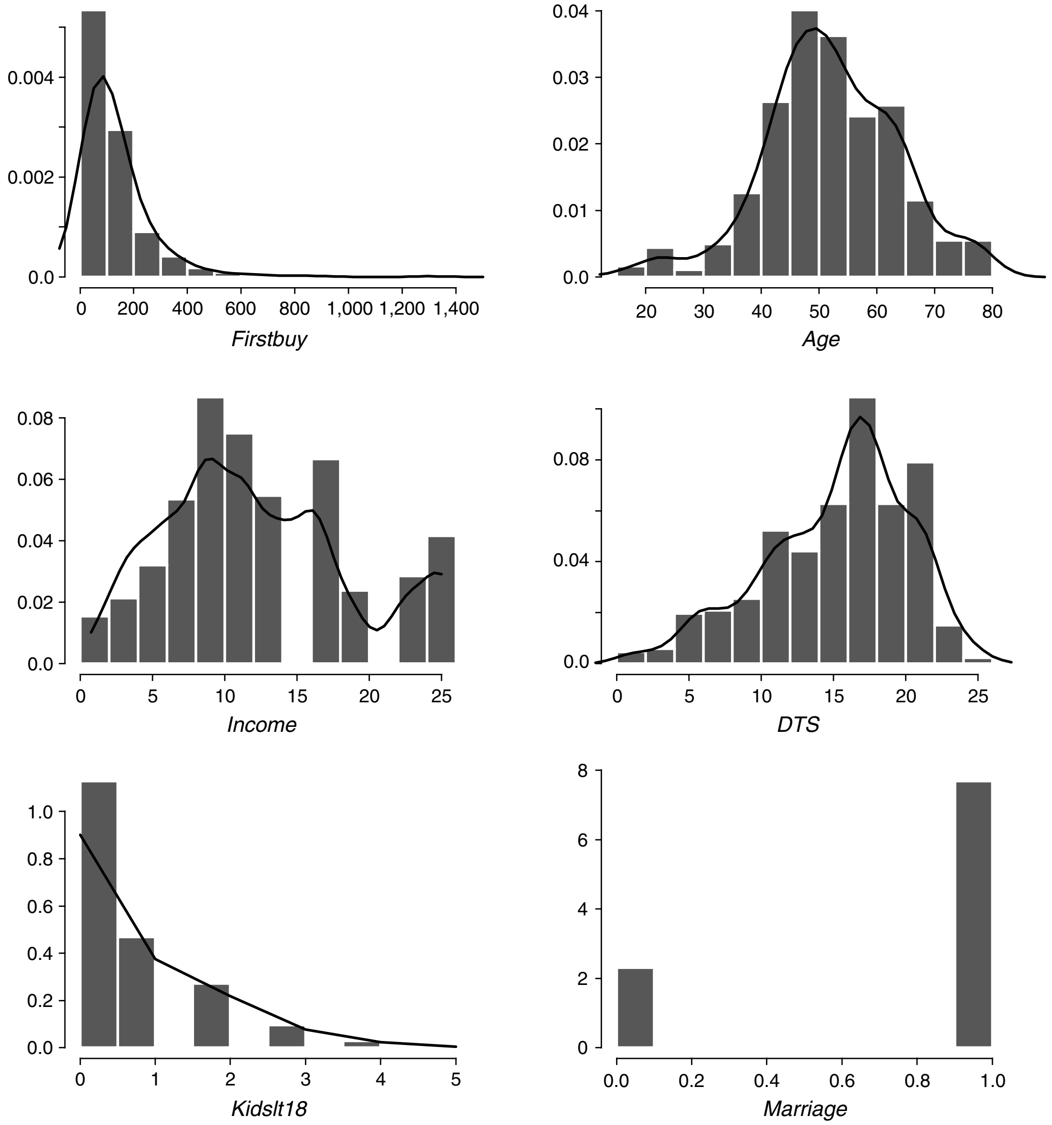
$$u_{it} = \beta_{0i} + \beta_{1i} e_{it} + \beta_{2i} Prom_t + \beta_{m_t} m_t + \epsilon_{it}, \qquad (10)$$

where e_{it} is the elapsed time since the last purchase by the consumer; $Prom_t$ is the variable capturing the monthly promotional activity of the store; m_t is the dummy for calendar month, where $m_t \in (1, \ldots, 11)$ denotes months from January to November; β_{m_t} is the corresponding fixed effect; and ϵ_{it} is the iid unobserved idiosyncratic factor affecting the utility to purchase, which follows a standard normal distribution. The consumer's decision to shop or not to shop at the store, Y_{it}, is a binary variable, and $Y_{it} = 1$ if $u_{it} > 0$ and $Y_{it} = 0$ otherwise. We model consumer heterogeneity $\beta_i = (\beta_{0i}, \beta_{1i}, \beta_{2i})$ as in Equation (8), where Z_i includes *Age*, *Income*, *Marriage*, *Kidslt18*, *DTS*, and *Firstbuy*.

As indicated above, some components in Z are missing. The complete-case analysis excludes about one-third of the customers who have a missing value in *any* of these variables. The method is thus inefficient, as the excluded consumers supply valuable information in other observed variables. More importantly, it may still be of interest for the manager to draw inferences about these consumers' preferences and sensitivities to manage relationships with them. In this case, excluding consumers from the analysis is managerially undesirable. In addition, the missing values of certain customers may result from systematic differences between these customers and the observed ones. Such selection effects must be accounted for in the analyses.

Here, we apply our Bayesian method to the data set, which includes all the consumers in the analysis. Note that Z contains a mixture of discrete and continuous variables, in which *Marriage* is dichotomous, *Kidslt18* is a count variable, and the other variables are continuous. Figure 2 shows these variables have features of nonnormality: skewness, multimodality, or discreteness. *Kidslt18* is a count variable naturally modeled as a Poisson outcome—or perhaps

Figure 2 Histograms of Variables in the Retail Store Data Set Using the Observed Data



more appropriately, a zero-inflated Poisson outcome. It is difficult to specify a joint parametric model to account simultaneously for all the features in these variables. The effect of covariate model misspecifications could be complicated because of the intercorrelations among covariates. Our method augments the above purchase incidence model with a semiparametric odds ratio covariate model for Z, as specified in Equations (4) and (5). In the approach, each covariate's marginal-like distribution is modeled nonparametrically, thereby automatically allowing for all the aforementioned data features. Thus, our approach reduces data analysts' modeling efforts while helping guard against model misspecifications. The dependence among the variables in Z is modeled by the parametric odds ratio functions.

Model estimation is performed through an MCMC algorithm that uses the approach of Albert and Chib (1993) to update parameters in the purchase incidence model and the HMC sampler to update parameters in the covariate model. The imputation of missing covariate values accounts for any potentially important dependence between the covariates and the purchase incidence outcomes. For the purpose of comparison, we conduct the complete-case analysis and the analysis based on an MVN covariate model. It is important to note that the MVN models all the covariates, including the two fully observed variables, *Firstbuy* and *Marriage*, whereas the DF model conditions on these two variables and models only the variables subject to missingness. All analyses run the MCMC sampler for 12,000 iterations, and the first 2,000 iterations are discarded as the burn-in period.[15]

[15] Because of the amenability of the outcome regression model to Bayesian analysis and the high efficiency of the HMC sampler,

Table 4 Estimation Result of a Purchase Incidence Model in the Retail Store Data Set

Parameter	Complete-case	MVN	DF
Intercept	−0.44 (0.11)$^+$	−0.52 (0.080)$^+$	−0.50 (0.079)$^+$
Firstbuy	0.02 (0.04)	0.016 (0.033)	0.012 (0.033)
Marriage	0.065 (0.11)	0.11 (0.083)	0.09 (0.080)
Age	0.035 (0.05)	0.053 (0.038)	0.086 (0.044)
Income	0.011 (0.043)	0.023 (0.036)	0.055 (0.039)
Kids	−0.073 (0.042)	−0.022 (0.035)	−0.017 (0.038)
DTS	−0.071 (0.05)	−0.095 (0.034)$^+$	−0.11 (0.035)$^+$
Prom	4.11 (1.12)$^+$	3.57 (0.77)$^+$	3.36 (0.78)$^+$
Prom ∗ *FirstBuy*	0.37 (0.30)	0.31 (0.26)	0.35 (0.26)
Prom ∗ *Marriage*	−0.02 (0.90)	0.30 (0.65)	0.49 (0.63)
Prom ∗ *Age*	−0.15 (0.39)	−0.21 (0.27)	−0.35 (0.36)
Prom ∗ *Income*	−0.74 (0.36)$^+$	−0.47 (0.27)	−0.69 (0.29)$^+$
Prom ∗ *Kids*	0.14 (0.36)	0.05 (0.27)	0.01 (0.31)
Prom ∗ *DTS*	−0.83 (0.38)$^+$	−0.49 (0.27)	−0.59 (0.27)$^+$
e_{it}	0.043 (0.027)	0.028 (0.017)	0.027 (0.016)
e_{it} ∗ *FirstBuy*	−0.002 (0.01)	0.001 (0.008)	0.001 (0.008)
e_{it} ∗ *Marriage*	−0.02 (0.03)	−0.005 (0.02)	−0.005 (0.02)
e_{it} ∗ *Age*	0.0069 (0.013)	0.004 (0.008)	0.002 (0.008)
e_{it} ∗ *Income*	0.004 (0.012)	0.003 (0.008)	0.002 (0.008)
e_{it} ∗ *Kids*	0.01 (0.012)	0.003 (0.008)	0.003 (0.008)
e_{it} ∗ *DTS*	−0.004 (0.014)	0.000 (0.008)	0.001 (0.008)
Σ_{11}	0.25 (0.03)$^+$	0.28 (0.03)$^+$	0.27 (0.03)$^+$
Σ_{22}	6.38 (2.85)$^+$	6.06 (2.26)$^+$	5.49 (2.08)$^+$
Σ_{33}	0.029 (0.0025)$^+$	0.020 (0.0014)$^+$	0.019 (0.0014)$^+$
Σ_{12}	−0.22 (0.21)	−0.20 (0.16)	−0.17 (0.15)
Σ_{13}	−0.012 (0.006)$^+$	−0.013 (0.004)$^+$	−0.013 (0.004)$^+$
Σ_{23}	−0.010 (0.029)	−0.0073 (0.020)	−0.0067 (0.019)
Subjects/Obs.	292/14,357	455/22,340	455/22,340

Note. Presented are the posterior means (posterior SD) of each parameter.
$^+$95% credible interval excludes zero.

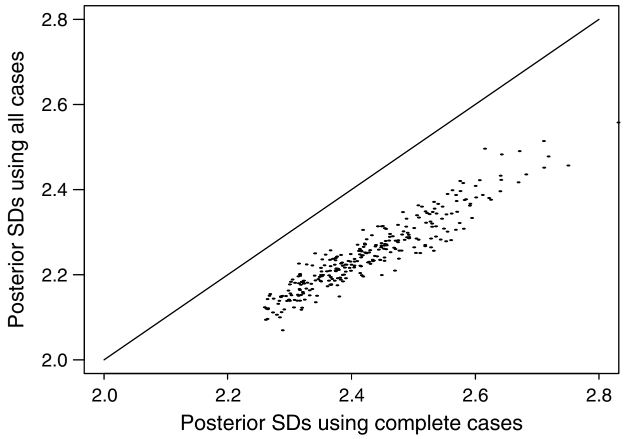
All the observed decimal places of the covariate values are kept in the analysis to form unique covariate values. Table 4 presents estimation results for the parameters in the purchase incidence model, which are of primary interest in the study. The covariates in Z are standardized before entering the model. The parameter estimates for *Intercept*, *Prom*, and e_{it} therefore represent their effects for an average consumer in the population. The comparison shows that the proposed method improves the estimation efficiency compared with the complete-case analysis by using all the available information. The posterior standard deviations (SDs) for all parameters are substantially smaller than those from the complete-case analysis—in some incidences, by a half. Also, some parameter estimates have substantial differences. For example, the estimated effect of *DTS* in our proposed model is larger than that in the complete-case analysis and is found to become statistically significant. The estimation results using the MVN and DF model are more similar to each other. Noticeable differences in some model estimates still remain, however. For example, the 95% credible interval for *Prom* ∗ *DTS* is found to exclude zero under the DF models but to include zero under the MVN model. We further conduct simulation studies in Appendix B.6 of the electronic companion that demonstrate the advantages of the DF method in repeated samples for purchase incidence models relative to the other two methods.

We now turn to the estimation of customer-level estimates when some variables are missing from a consumer's profile. In individually targeted marketing, it is often useful to make inferences on the consumer-specific parameters and then adopt differential marketing strategies based on the estimates of these parameters. A Bayesian approach is well suited for such individual-level analysis. Allenby and Rossi (1999) consider such estimation when covariates are fully observed. For our case in which covariates are subject to missingness, the posterior distribution of β_j in a fully Bayesian approach is given by

$$\begin{aligned}&\pi(\beta_j \mid Y, Z^{\text{obs}}, X)\\&\propto \int \pi(\{\beta_i\}, \Pi, \Lambda, \lambda, \gamma,\\&\qquad\qquad Z^{\text{mis}} \mid Y, Z^{\text{obs}}, X)\, d\beta_{-j}\, d\Pi\, d\Lambda\, d\lambda\, d\gamma\, dZ^{\text{mis}},\end{aligned}$$

a relatively small number of iterations is required for convergence. Geweke's diagnostic statistic confirms the quick convergence of the MCMC sampling.

Figure 3 **Comparison of Posterior SD for $\{\beta_{2i}\}$ on the Subset of Consumers with the Complete Data on Variables in Z_i**



Notes. The X axis gives the posterior SDs when the complete-case analysis is used. The Y axis gives the posterior SDs from our proposed Bayesian DF model analysis, where all the consumers are used to fit the model.

where β_{-j} denotes all the consumer-specific parameters except for the jth consumer. One benefit of the Bayesian approach is that the entire posterior distribution for β_j can be obtained as a by-product of the MCMC algorithm. It is important to note that our fully Bayesian approach also automatically accounts for the uncertainty in the imputation of missing covariates.

Figure 3 plots the posterior SDs of $\{\beta_{2i}\}$ obtained from the complete-case analysis versus those obtained from our proposed method. It shows that the estimates using the proposed method have smaller SDs because the proposed method uses all the consumers in the data, whereas the complete-case analysis discards information contained in those incomplete cases. Such reduction in the estimation variability of consumer-specific parameters can be valuable, as these individual-level parameters tend to be less accurately estimated. What are not shown in the figure are those incomplete cases whose individual-level estimates are not available in the complete-case analysis but are available as a by-product of the fitting of our proposed model.

5.4. Managerial Implications

In this subsection, we investigate the managerial implications of the above estimation results. Specifically, we investigate the differences of targeting and profiling consumers based on observed characteristics of consumers using the above estimation results. Such profiling on actionable consumer characteristics can be managerially very useful to take findings from one store to another similar store (e.g., Singh et al. 2006). Similar to the approach of Singh et al. (2006), we calculate the marginal effects of observed household characteristics. More specifically, we calculate the effect of a 30% price promotion on the purchase incidence probability for a population with covariate vector $Z = \bar{z}$, $z_{j,5\%}$, $z_{j,95\%}$. In the notation, $\bar{z}$ is the sample average of the covariate vector Z; $z_{j,5\%}$ $(z_{j,95\%})$ is the same as $\bar{z}$, except that the jth covariate is set to be the 5th (95th) percentile of its distribution. Based on the model estimates from different methods, we simulate populations of consumers and calculate the average purchase incidence probabilities before and after the promotion when the variables of interest are *DTS* and *Age*.

Table 5 presents the simulation results. It shows that the estimates of the changes in purchase probabilities before and after the promotion vary under different methods. The complete-case (CC) analysis tends to overestimate considerably the promotional effects in this data set. For example, at $Age_{95\%}$, the overestimation, compared with DF, is about 35%. Thus, the complete-case analysis could lead a manager to misjudge the effect of promotion on purchase incidence probability and, in turn, the profitability of the potential promotion activity. For example, using the complete-case analysis, the manager might conclude that the increase in the purchase incidence probability outweighs the loss of money values because of promotion and thus may conclude that the promotion is profitable. An analysis using DF may show, however, that the increase in the purchase incidence is not large

Table 5 **Moderating Effect of Covariates on the Promotional Effects**

	CC			MVN				DF			
Z	Before	After	Δ_{CC}	Before	After	Δ_{MVN}	$\frac{\Delta_{MVN}-\Delta_{CC}}{\Delta_{CC}}$	Before	After	Δ_{DF}	$\frac{\Delta_{DF}-\Delta_{CC}}{\Delta_{CC}}$
$DTS_{5\%}$	0.39	0.85	0.46	0.38	0.76	0.38	−17	0.39	0.77	0.38	−17
DTS_{Z}	0.35	0.72	0.37	0.32	0.66	0.34	−8	0.33	0.65	0.32	−14
$DTS_{95\%}$	0.31	0.57	0.26	0.29	0.54	0.25	−4	0.28	0.50	0.22	−14
$Age_{5\%}$	0.32	0.73	0.40	0.29	0.66	0.37	−8	0.28	0.66	0.38	−5
Age_{Z}	0.35	0.72	0.37	0.32	0.66	0.34	−10	0.33	0.65	0.32	−14
$Age_{95\%}$	0.37	0.72	0.35	0.35	0.65	0.30	−16	0.38	0.64	0.26	−26

enough to offset the loss because of promotion, leading the manager to conclude that the promotion is not profitable. The MVN and DF models for covariates lead to more comparable results. In Table 5, we report the percentage change of the promotional effect on purchase incidence probabilities from that predicted using the complete-case analysis for the MVN and DF models, respectively. These numbers indicate that the MVN model also overestimates the promotional effects at $DTS_{95\%}$ and $Age_{95\%}$.

6. Conclusion

With the increasing popularity of database marketing, CRM, and individualized marketing, companies face a greater need to provide customized marketing solutions based on consistent and precise elasticity estimates on marketing mix variables. On the other hand, the real applications above indicate that the covariates in a marketing regression model often are subject to missingness. The convenient complete-case analysis can lead to strong self-selection bias and to substantial loss of estimation efficiency. More advanced methods for overcoming these drawbacks commonly assume parametric models for covariate distributions. One limitation of the parametric modeling approach is its nonrobustness; when the parametric covariate model is misspecified, a substantial bias can arise in the estimation of marketing outcome models. The issue is further exacerbated by the difficulty in assessing the validity of distributional assumptions in modeling covariates with missing values. Furthermore, the extra computational and modeling workload can be high, which have hindered the routine use of these methods for dealing with high-dimensional missing covariate problems. Therefore, how to extract useful information from the available data in a robust, efficient, and simple manner is an issue pertinent to current marketing research activities.

To this end, we have developed a distribution-free Bayesian method to handle missing covariate problems. Our development of an efficient MCMC algorithm overcomes an important limitation of Chen (2004) and enables one to handle high-dimensional missing covariate problems and/or complex models commonly seen in marketing applications. Some other key benefits of the method to marketing researchers are (1) its distribution-free feature, which enables robust modeling of covariate distributions and minimizes the impact of distributional assumptions in covariate modeling; (2) its flexibility, which allows for complex dependence among covariates and to incorporate any useful information for covariate distribution; and (3) its simplicity in modeling and computation, which substantially reduces the workload associated with careful modeling of covariates compared with alternative parametric approaches.

The applications of the proposed method yield some interesting empirical findings. In the ketchup example, we confirm that because they do not account for the dependence between choice outcomes and missing marketing mix variables, conventional ad hoc approaches to imputing missing marketing mix values can create strong selection bias in the estimation of brand-choice models. The joint modeling of the choice outcomes and missing covariates provides a principled approach to correcting the bias. Our analysis shows that to remove the bias, it is important to properly model distributional features of marketing mix variables and to account for the dependence among them. Ignoring either feature can lead to sizable bias in estimation and, consequently, to suboptimal managerial decisions. As demonstrated in the empirical application and simulation studies, the proposed method improves model estimates of consumer preferences and sensitivities to marketing mix variables when compared with prior approaches to the problem. The retail store example shows that the interrelated covariates often exhibit various features such as discreteness, skewness, multimodality, semicontinuity, and zero-inflation. The proposed method accounts automatically for these important data features and helps guard against model misspecifications in a parametric covariate modeling approach. Our Bayesian approach also enables straightforward estimation and inference of consumer-level parameters if some components of the consumer's profile are missing. This approach ensures that no customer is left behind in the analyses and in the subsequent managerial inferences.

Several issues and opportunities for future research remain. First, in our covariate model, although the univariate marginal-like distributions are modeled nonparametrically, the odds ratio functions are modeled parametrically. As noted in Chen (2004), the assumption can be wrong only for the higher-order terms of dependence structure. It is therefore reasonable to believe that the potential effect is relatively minor when the lower-order terms capture the majority of the associations among covariates. To control for such potential misspecification more thoroughly, however, it would be valuable in future research to develop a formal procedure for choosing proper terms in the odds ratio functions.

Second, our analysis assumes that the covariates are missing at random. Although this is a standard assumption in missing data analysis and is either known to hold or considered reasonable in many marketing applications, it could be tenuous in some marketing applications. When covariates are missing not at random, all methods based on the MAR assumption, including the proposed method in this paper, are potentially invalid. Extending the method to account for such potentially nonignorable missing data would be very valuable.

In this paper we have investigated the use of the method for a wide range of types of commonly used marketing models. The proposed method is not limited to these marketing models, however. An essentially infinite number of parametric marketing models exists, and new ones are proposed constantly. In many of these models, a Bayesian approach is a preferred method for estimation and inference. It would be interesting to combine the proposed method with these other types of marketing regression models to make better use of available data sets.

7. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://mktsci.pubs.informs.org/.

Acknowledgments

The authors thank the editors, area editor, and two anonymous referees whose constructive comments helped to substantially improve the earlier versions of this paper. They thank Ulf Böckenholt, Eric Bradlow, Yuxin Chen, Hua Yun Chen, Preyas Desai, Manuel Hermosilla, and Peter Rossi for their very helpful comments on the earlier versions of the paper. The second author received financial support from National Institutes of Health Grant UL1RR029879.

References

Albert, J. H., S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88**(422) 669–679.

Allenby, G. M., P. J. Lenk. 1994. Modeling household purchase behavior with logistic normal regression. *J. Amer. Statist. Assoc.* **89**(428) 1218–1231.

Allenby, G. M., P. E. Rossi. 1999. Marketing models of consumer heterogeneity. *J. Econometrics* **89**(1–2) 57–78.

Blattberg, R. C., B.-D. Kim, S. A. Neslin. 2008. *Database Marketing: Analyzing and Managing Customers*. Springer, New York.

Bradlow, E. T., A. M. Zaslavsky. 1999. A hierarchical latent variable model for ordinal data from a customer satisfaction survey with "no answer" responses. *J. Amer. Statist. Assoc.* **94**(445) 43–52.

Bradlow, E. T., Y. Hu, T.-H. Ho. 2004. A learning-based model for imputing missing levels in partial conjoint profiles. *J. Marketing Res.* **41**(4) 369–381.

Chen, H. Y. 2004. Nonparametric and semiparametric models for missing covariates in parametric regression. *J. Amer. Statist. Assoc.* **99**(468) 1176–1189.

Chiang, J. 1995. Competing coupon promotions and category sales. *Marketing Sci.* **14**(1) 105–122.

Chintagunta, P. K., D. C. Jain, N. J. Vilcassim. 1991. Investigating heterogeneity in brand preferences in logit models for panel data. *J. Marketing Res.* **28**(4) 417–428.

Daniels, M. J., J. W. Hogan. 2008. *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman & Hall, London.

Duane, S., A. D. Kennedy, B. J. Pendleton, D. Roweth. 1987. Hybrid Monte Carlo. *Phys. Lett. B* **195**(2) 216–222.

Erdem, T., M. P. Keane, B. Sun. 1999. Missing price and coupon availability data in scanner panels: Correcting for the self-selection bias in choice model parameters. *J. Econometrics* **89**(1–2) 177–196.

Feit, E. M., M. A. Beltramo, F. M. Feinberg. 2010. Reality check: Combining choice experiments with market data to estimate the importance of product attributes. *Management Sci.* **56**(5) 785–800.

Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith, eds. *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*. Clarendon Press, Oxford, UK, 169–194.

Gilula, Z., R. E. McCulloch, P. E. Rossi. 2006. A direct approach to data fusion. *J. Marketing Res.* **43**(1) 73–83.

Gönül, F., K. Srinivasan. 1993. Modeling multiple sources of heterogeneity in multinomial logit models: Methodological and managerial issues. *Marketing Sci.* **12**(3) 213–229.

Guadagni, P. M., J. D. C. Little. 1983. A logit model of brand choice calibrated on scanner data. *Marketing Sci.* **2**(3) 203–238.

Gupta, S. 1991. Stochastic models of interpurchase time with time-dependent covariates. *J. Marketing Res.* **28**(1) 1–15.

Ibrahim, J. G., M.-H. Chen, S. R. Lipsitz, A. H. Herring. 2005. Missing-data methods for generalized linear models: A comparative review. *J. Amer. Statist. Assoc.* **100**(469) 332–346.

Kamakura, W. A., G. J. Russell. 1989. A probabilistic choice model for market segmentation and elasticity structure. *J. Marketing Res.* **26**(4) 379–390.

Kamakura, W. A., M. Wedel. 1997. Statistical data-fusion for cross-tabulation. *J. Marketing Res.* **34**(4) 485–498.

Kamakura, W. A., M. Wedel. 2000. Factor analysis and missing data. *J. Marketing Res.* **37**(4) 490–498.

Kang, J. D. Y., J. L. Schafer. 2007. Demystifying double robustness: A comparison of alternative strategies for estimating population mean from incomplete data (with discussion). *Statist. Sci.* **22**(4) 523–539.

Khan, R., M. Lewis, V. Singh. 2009. Dynamic customer management and the value of one-to-one marketing. *Marketing Sci.* **28**(6) 1063–1079.

Little, R. J. A. 1992. Regression with missing X's: A review. *J. Amer. Statist. Assoc.* **87**(420) 1127–1137.

Little, R. J. A., D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Meier, P., T. Karrison, R. Chappell, H. Xie. 2004. The price of Kaplan–Meier. *J. Amer. Statist. Assoc.* **99**(467) 890–896.

Musalem, A., E. T. Bradlow, J. S. Raju. 2008. Who's got the coupon? Estimating consumer preferences and coupon usage from aggregate information. *J. Marketing Res.* **45**(6) 715–730.

Qian, Y. 2007. Do national patent laws stimulate domestic innovation in a global patenting environment? A cross-country analysis of pharmaceutical patent protection, 1978–2002. *Rev. Econom. Statist.* **89**(3) 436–453.

Qian, Y., H. Xie. 2010. Measuring the impact of nonignorability in panel data with non-monotone nonresponse. *J. Appl. Econometrics*, ePub ahead of print February 1, http://onlinelibrary.wiley.com/doi/10.1002/jae.1157/abstract.

Raftery, A. E., M. A. Newton, J. M. Satagopan, P. N. Krivitsky. 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statist.* **8** 1–45.

Rossi, P. E., G. M. Allenby. 1993. A Bayesian approach to estimating household parameters. *J. Marketing Res.* **30**(2) 171–182.

Rossi, P. E., G. M. Allenby, R. McCulloch. 2005. *Bayesian Statistics and Marketing*. John Wiley & Sons, London.

Rubin, D. B. 1976. Inference and missing data (with discussion). *Biometrika* **63**(3) 581–592.

Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Schafer, J. L., J. W. Graham. 2002. Missing data: Our view of the state of the art. *Psych. Methods* **7**(2) 147–177.

Seetharaman, P. B., P. K. Chintagunta. 2003. The proportional hazards model for purchase timing: A comparison of alternative specifications. *J. Bus. Econom. Statist.* **21**(3) 368–382.

Singh, V. P., K. T. Hansen, R. C. Blattberg. 2006. Market entry and consumer behavior: An investigation of a Wal-Mart supercenter. *Marketing Sci.* **25**(5) 457–476.

Tierney, L. 1994. Markov chains for exploring posterior distributions. *Ann. Statist.* **22**(4) 1701–1728.

Tsiatis, A. A. 2006. *Semiparametric Theory and Missing Data*. Springer, New York.

Wedel, M., W. A. Kamakura, W. S. Desarbo, F. T. Hofstede. 1995. Implications for asymmetry, nonproportionality, and heterogeneity in brand switching from piece-wise exponential mixture hazard models. *J. Marketing Res.* **32**(4) 457–462.

Yang, S., Y. Zhao, R. Dhar. 2010. Modeling the underreporting bias in panel survey data. *Marketing Sci.* **29**(3) 525–539.

Ying, Y., F. Feinberg, M. Wedel. 2006. Leveraging missing ratings to improve online recommendation systems. *J. Marketing Res.* **43**(3) 355–365.

Web Appendix to
"No Customer Left Behind: A Distribution-Free Bayesian Approach to Accounting for Missing Xs in Marketing Models"

This Appendix contains two main sections: A and B. In Web Appendix A, we provide the overview of estimation algorithm and the details of the algorithm when the outcome follows a discrete choice model. In Web Appendix B, we provide the details of extensive simulation studies from which the main conclusions are summarized in the main paper. The simulation studies investigate the performance of our proposed method for various models and covariate distributions, and compare our method with various other methods to treat missing covariate problems.

A. Estimation using an MCMC algorithm

A.1 *An overview*

We develop an MCMC procedure to sample from the following posterior distribution of model unknowns

$$\pi(\theta, \phi, X^{mis}|Y, X^{obs}) \propto f_\theta(Y|X^{obs}, X^{mis}) f_\phi(X^{obs}, X^{mis})\pi(\theta)\pi(\phi),$$

where $\pi(\theta)$ and $\pi(\phi)$ are the priors of θ and ϕ, respectively. Below are the major steps of this MCMC procedure.

- Impute X^{mis}. The imputation of X^{mis} can be achieved componentwisely.[1] That is, we draw a value for a missing component of X^{mis}, given all the observed components and the other missing components at their current values. The updating of any missing component involves a draw from a multinomial distribution on a finite number of values. Details on calculating the probabilities in the multinomial distribution for discrete choice outcome models are given in Section A.2.

- Draw (λ, γ) from $\lambda, \gamma|X^{obs}, X^{mis}$. Once X^{mis} is imputed, one can update the values of (λ, γ) from their distribution conditioning on the complete data for the covariates. Because conjugate priors for this updating step do not exist, a Metropolis-type algorithm is needed. The dimension of (λ, γ) is relatively large and these parameters are significantly correlated. This renders a Random-Walk Metropolis algorithm not working well since the proposal draws have low acceptance rates. To overcome this challenge, we adopt a Hybrid Monte Carlo (HMC) algorithm, where the derivatives of the joint density function of X with respect to (λ, γ) are calculated and used to generate proposal draws. We find that the HMC algorithm handles the updating of multiple correlated parameters efficiently.

[1] An alternative approach is to update missing covariates in an observation all together. This reduces autocorrelation of Markov Chain but it can involve evaluating summations with a large number of combinatorial terms which is time consuming. Our experiences in the applications and simulation studies are that imputing componentwisely takes less overall computational time and the inflation of autocorrelation is manageable. Another alternative approach is a hybrid updating algorithm which periodically switches between imputation componentwisely and imputation all together. Although we do not find the need to use this approach in our applications, it could be a viable updating approach in future applications.

Because the HMC algorithm uses the derivative information to make a trial move, it produces draws that are more likely to be accepted while still being able to quickly explore the target distribution (Liu 2001). Therefore the target acceptance rate in HMC sampler is significantly higher than that in a RW-MH sampler, especially in high-dimensional parameter space. Recent theoretical work (Beskos et al. 2010) shows that the acceptance rate under optimal tuning of an HMC sampler is 0.651. We therefore set the acceptance rate of our HMC sampler to be at the level between 0.6 and 0.7.

- Once the missing covariates X^{mis} are imputed, one can use a conventional algorithm to update the parameters in the outcome model, which depends on the specific outcome model used.

A.2 *MCMC Algorithm for a Discrete Choice Model with Missing Covariates.*

The model parameters are $\theta = (\Pi, \Lambda)$ and $\phi = (\lambda, \gamma)$, where $\lambda = (\lambda_1, \cdots, \lambda_K)$, $\gamma = (\gamma_1, \cdots, \gamma_K)$ and $\gamma_1 = 0$. As we adopt the Bayesian approach, to complete the model specification, we need to assign priors for model parameters $(\Pi, \Lambda, \lambda, \gamma)$. We assign the priors in the following forms:

$$\begin{aligned} vec(\Pi) \sim MVN(\mu_\Pi, \Lambda_\Pi^{-1}) \quad &, \quad \Lambda \sim W(\nu, S) \\ \lambda \sim MVN(\mu_\lambda, \nu_\lambda I_{n_\lambda}) \quad &, \quad \gamma \sim MVN(\mu_\gamma, \nu_\gamma I_{n_\gamma}), \end{aligned} \tag{1}$$

where $W(\nu, S)$ is a Wishart distribution with ν being the degrees of freedom and S being the scale matrix. As will be shown later, we assign the values to the constants in the above priors in such a way that the priors are non-informative relative to the data.

The posterior distribution of the model parameters involves integration over the random effects $\{\beta_i\}, i = 1, \cdots, N$ and the missing data X^{mis}. To facilitate the posterior sampling, we first apply the data augmentation method (Tanner and Wong 1987), where the model parameters $(\Pi, \Lambda, \lambda, \gamma)$ are augmented with the other unknowns $(\{\beta_i\}, X^{mis})$ in the model. Their joint posterior distribution is given below.

$$\pi(\Pi, \Lambda, \lambda, \gamma, \{\beta_i\}, X^{mis}|Y, X^{obs}, Z) \propto$$
$$\left\{ \prod_{i=1}^{N} \left[\prod_{t\in T_i} f_{\beta_i}(Y_i|X_i^{obs}, X_i^{mis}) f_{\lambda,\gamma}(X_{it}^{obs}, X_{it}^{mis}) \right] f_{\Pi,\Lambda}(\beta_i|Z_i) \right\} \pi(\Pi)\pi(\Lambda)\pi(\lambda)\pi(\gamma), \tag{2}$$

where $\pi(\Pi)$, $\pi(\Lambda)$, $\pi(\lambda)$ and $\pi(\gamma)$ are the priors for these model parameters. Below are the details of each updating step in the MCMC algorithm.

(1) Updating $X^{mis}|X^{obs}, \{\beta_i\}, Y, \lambda, \gamma$.

Once we have draws of the model parameters, we can impute the missing price and coupon values. One approach imputes missing values in $X_{it} = (\{P_{itj}\}, \{C_{itj}\})$ one component at a time. Suppose that the kth component of X_{it} is missing. Then by the Bayes Rule, the formula for imputing X_{itk}^{mis} is:

$$p(X_{itk}^{mis} \quad = \quad x_{kl}|y_{it}, x_{it1}, ..., x_{it(k-1)}, x_{it(k+1)}, ..., x_{itK}) \quad = \quad \frac{f_{\beta_i}(y_{it}|x_{it}^{kl}) f_{\lambda,\gamma}(x_{it}^{kl})}{\sum_{l'=1}^{N_k} f_{\beta_i}(y_{it}|x_{it}^{kl'}) f_{\lambda,\gamma}(x_{it}^{kl'})},$$

where $x^{kl}_{it} = (x_{it1}, ..., x_{it(k-1)}, X^{mis}_{itk} = x_{kl}, x_{it(k+1)}, ..., x_{itK})$ is the vector of covariate values for x_{it}. X^{mis}_{itk} is imputed with x_{kl} in x^{kl}_{it}, and x_{kl} denotes the lth unique observed value in the dataset for the kth component of X_{it} $(l = 1, ..., N_k)$. In x^{kl}_{it}, all the missing values except the kth component take the imputed values in the previous iteration. The starting values can be obtained by drawing from their empirical distributions on the observed data.

The above imputation step assigns discrete probabilities to those unique observed data values, and make draws from this discrete probability distribution. It is clear from the above imputation equation that when imputing X^{mis}_{itk}, one should account for both the dependence between the outcome Y and covariates as quantified by the conditional density function $f_{\beta_i}(\cdot|\cdot)$, as well as the dependence among the covariates as quantified by the density function $f_{\lambda,\gamma}(\cdot)$. When the covariate X_{itk} is independent of the other covariates, the above imputation formula reduces to a simpler form:

$$p(X^{mis}_{itk} = x_{kl}|y_{it}, x_{it1}, ..., x_{it,k-1}, x_{it,k+1}, ..., x_{itK}) = \frac{f_{\beta_i}(y_{it}|x^{kl}_{it}) f_{\lambda_k,\gamma_k=0}(X^{mis}_{itk} = x_{kl})}{\sum_{l'=1}^{N_k} f_{\beta_i}(x^{kl'}_{it}) f_{\lambda_k,\gamma_k=0}(X^{mis}_{itk} = x_{kl'})}.$$

The simpler imputation formula is intuitive: when the X_{itk} is independent of all the other covariates, these other variables do not contribute to its imputation through the covariate density function $f_{\lambda,\gamma}(\cdot)$.

As noted in EKS (1999), it is possible that some non-purchased brands by a consumer might be purchased by other consumers on the same day at the same store. Therefore, the prices for these brands at that day are known. When we use this information, we can form the posterior distribution using the full-information likelihood (EKS 1999), while the likelihood given in Equation (9) of the main paper can be considered as the limited information likelihood. Note that the price values for brands not purchased by any consumer in the store at a given day are still missing and need to be imputed. Let $(x_{it1}, ..., x_{itJ})$, the first J components of x_{it}, denote the pricing variables. Then to impute the missing pricing variables with the full-information likelihood, we use:

$$p(X^{mis}_{i^t_1 tk} = \cdot\cdot = X^{mis}_{i^t_{n_t} tk} = x_{kl}|\left\{y_{it}\right\}_{i\in s(t)}, \left\{x^{kl}_{it}\right\}_{i\in s(t)}) =$$
$$\frac{\prod_{i\in s(t)} f_{\beta_i}(y_{it}|x^{kl}_{it}) f_{\phi_{J+1},...,\phi_K}(x_{it,J+1}, ..., x_{itK}|x^{kl}_{itP}) f^{1/n_t}_{\phi_1,...,\phi_J}(x^{kl}_{itP})}{\sum_{l'=1}^{N_k}\left[\prod_{i\in s(t)} f_{\beta_i}(y_{it}|x^{kl}_{it}) f_{\phi_{J+1},...,\phi_K}(x_{it,J+1}, ..., x_{itK}|x^{kl'}_{itP}) f^{1/n_t}_{\phi_1,...,\phi_J}(x^{kl'}_{itP})\right]},$$

where $x^{kl}_{itP} = (x_{it1}, ..., x^{mis}_{itk} = x_{kl}, ..., x_{itJ})$ denotes the vector of prices at purchase occasion t within which x^{mis}_{itk} is imputed by x_{kl} and $k \leq J$; $s(t) = (i^t_1, ..., i^t_{n_t})$ is the set of indices for the n_t consumers who made purchase on the same occasion t. For these consumers, the pricing values faced by them are assumed to be the same for the full-information likelihood. Therefore, essentially only one of the consumers contributes likelihood for price variables. This is equivalent to having each consumer contribute $f^{1/n_t}_{\phi_1,...,\phi_J}(x^{kl}_{itJ})$ to the likelihood, as used in the above imputation formula. In our application we use the imputation formula based on the full-information likelihood.

(2) Updating $\lambda, \gamma|X^{obs}, X^{mis}$.
Once the missing values in X are all imputed, we can update the parameters $\phi = (\lambda, \gamma)$ given these imputed covariate data. Note that $f_\phi(X_{it1}, ..., X_{itK}) = \prod_{k=1}^K f_{\phi_k}(X_{itk}|X_{it(k-1)}, ..., X_{it1})$. Thus, given that the prior for ϕ factorizes, i.e., $\pi(\phi) = \prod_{k=1}^K \pi_k(\phi_k)$, ϕ_k can be updated independently of $(\phi_1, \cdots, \phi_{k-1}, \phi_{k+1}, \cdots, \phi_K)$. For ϕ_k, the density function for its conditional

distribution is given as:

$$\pi_c(\phi_k) = \left[\prod_i f_i(\phi_k)\right] \pi_k(\phi_k),$$

where $\pi_k(\phi_k)$ is the prior and

$$f_i(\phi_k) \quad = \quad \frac{\prod_t \sum_{l=1}^{N_k} 1_{(x_{itk}=x_{kl})}\eta_{\gamma_k}(x_{itk}, x_{k0}; x_{it(k-1)}, ..., x_{it1}, x_{(k-1)0}, ..., x_{10}) \exp(\lambda_{kl})}{\prod_t \sum_{l'=1}^{N_k} \eta_{\gamma_k}(X_{itk} = x_{kl'}, x_{k0}; x_{it(k-1)}, ..., x_{it1}, x_{(k-1)0}, ..., x_{10}) \exp(\lambda_{kl'})}. \quad (3)$$

With the above odds-ratio covariates model, there does not exist conjugate prior for $\pi_k(\phi_k)$ and thus the conditional distribution of ϕ_k does not have a closed form. One general approach is to apply the Metropolis-Hasting type algorithm. We find the Random-Walk Metropolis-Hasting (RW-MH) is inefficient since there are a relatively large number of parameters, which tend to be highly correlated. We adopt the Hybrid Monte Carlo method to make the conditional draws.

Duane et al. (1987) proposed the method of Hybrid Monte Carlo (HMC) that combines the idea of Molecular Dynamics (MD) proposal and the Metropolis acceptance-rejection method to sample from a target distribution. In the HMC algorithm, the trial move is generated by the MD simulation method. The MD method obeys the Newton's law of motion, and is known to work well in simulating a complex physical system. When making the proposal draws, the local dynamics of the target distribution, quantified by the gradient of the log-density function of the target distribution, is utilized. Therefore, unlike the RW-MH algorithm, the randomness in proposing new draws is suppressed in the HMC algorithm. The result is that the HMC sampler produces draws that are more likely to be accepted and more quickly reach the high mass area of the target distribution by adapting to its local dynamics. The method handles the correlation between parameters more efficiently than the standard Metropolis algorithm.

Operationally, the HMC algorithm works in the following way. To sample from a target distribution $\pi_c(\phi_k)$, we first express $\pi_c(\phi_k)$ as $\exp[-U_k(\phi_k)]$. We then augment the parameter ϕ_k with a vector of auxiliary momentum variables p_k which has the same dimension as ϕ_k. The guide Hamiltonian is given as

$$H_k(\phi_k, p_k) = U_k(\phi_k) + \varphi_k(p_k),$$

where $\varphi_k(p_k) = -\frac{1}{2}p_k^T p_k$. To run the algorithm, we first need to initialize the system. Let ϕ_k^{old} be the current value of ϕ_k, and let the initial value $\phi_k^0 = \phi_k^{old}$. Generate p_k' from a standard Gaussian distribution and then assign to the system an initial momentum: $p_k^0 = p_k' - \frac{\delta_k}{2}U_k'(\phi_k^0)$, where $U_k'(\phi_k^0)$ is the derivative of $U_k(\cdot)$ with respect to its argument, and δ_k is a user-specified stepsize. Starting from the initial phase space (ϕ_k^0, p_k^0), an approximate molecular dynamic algorithm, called leap-frog algorithm, is run L steps to generate a new state (ϕ_k^L, p_k^L) in the phase space (Duane et al. 1987, Liu 2001), where

$$\begin{aligned} \phi_k^l &= \phi_k^{l-1} + \delta_k p_k^{l-1}, \\ p_k^l &= p_k^{l-1} - \delta_k^l U_k'(\phi_k^l), \end{aligned}$$

$l = 1, ..., L$, $\delta_k^l = \delta_k$ for $l < L$ and $\delta_k^L = \frac{\delta_k}{2}$. At the end of the leap-frog steps, let the candidate draw $(\phi_k^{prop}, p_k^{prop}) = (\phi_k^L, p_k^L)$. The algorithm accepts the candidate draw according to the following probability

$$min(1, \exp\left\{-H_k(\phi_k^{prop}, p_k^{prop}) + H_k(\phi_k^{old}, p_k^{old})\right\}).$$

If the candidate draw is accepted, ϕ_k^{prop} becomes the new draw; otherwise, ϕ_k^{old} becomes the new draw. At the end of the MCMC run, the draws for ϕ_k are collected and the draws for the momentum variable p_k can be discarded. The algorithm requires evaluating the derivative of $U_k(\phi_k)$ with respect to ϕ_k. For the semiparametric odd ratio covariate model, we have

$$\frac{\partial U_k(\phi_k)}{\partial \phi_k} = -\sum_i \frac{\partial \ln f_i(\phi_k)}{\partial \phi_k} - \frac{\partial \ln \pi(\phi_k)}{\partial \phi_k}, \tag{4}$$

where each component of the derivative is given below.

$$\frac{\partial \ln f_i(\phi_k)}{\partial \lambda_{kl}} = \sum_t \left[1_{(x_{itk}=x_{kl})} - \frac{\eta_{\gamma_k}(x_{kl}, x_{k0}; x_{it(k-1)}, ..., x_{it1}, x_{(k-1)0}, ..., x_{10}) \exp(\lambda_{kl})}{\sum_{l'=1}^{N_k} \eta_{\gamma_k}(x_{kl'}, x_{k0}; x_{it(k-1)}, ..., x_{it1}, x_{(k-1)0}, ..., x_{10}) \exp(\lambda_{kl'})} \right],$$

$$\frac{\partial \ln f_i(\phi_k)}{\partial \gamma_{kv}} = \sum_t \left[(x_{itk} - x_{k0})(x_{itv} - x_{v0}) - \frac{\sum_{l'=1}^{N_k} \eta_{\gamma_k}(x_{kl'}, x_{k0}; x_{it(k-1)}, ..., x_{it1}, x_{(k-1)0}, ..., x_{10}) \exp(\lambda_{kl'})(x_{kl'} - x_{k0})(x_{itv} - x_{v0})}{\sum_{l'=1}^{N_k} \eta_{\gamma_k}(x_{kl'}, x_{k0}; x_{it(k-1)}, ..., x_{it1}, x_{(k-1)0}, ..., x_{10}) \exp(\lambda_{kl'})} \right],$$

$$\frac{\partial \ln \pi(\phi_k)}{\partial \lambda_{kl}} = -\frac{\lambda_{kl} - \mu_{\lambda,kl}}{\nu_\lambda}, \quad \frac{\partial \ln \pi(\phi_k)}{\partial \gamma_{kv}} = -\frac{\gamma_{kv} - \mu_{\gamma,kv}}{\nu_\gamma},$$

where $\mu_{\lambda,kl}$ is the corresponding element for λ_{kl} in μ_λ (the mean vector in the prior for λ), $\mu_{\gamma,kv}$ is the corresponding element for γ_{kv} in μ_γ (the mean vector in the prior for γ) and ν_λ, ν_γ are the variances in the prior for λ, γ, as shown in Equation (1). In our applications, we set the constants in the priors for λ and γ as follows: μ_λ and μ_γ are vectors of zeros with length n_λ and n_γ, respectively; $\nu_\lambda = \nu_\gamma = 10^4$. Note that in the derivatives shown in the above, the larger ν_λ and ν_γ, the smaller the contribution of the priors in the updating. As long as v_λ and v_γ are sufficiently large, the contribution of the prior to updating is negligible and the prior becomes noninformative relative to data. Further increased values of ν_λ and ν_γ lead to negligible change in the estimation results.

For the pricing variables, let $(X_{it1}, ..., X_{itJ})$ denote the covariates for prices of the J brands. Since only one consumer in $s(t)$ contributes likelihood of x_{it}, the derivatives for the kth price, where $1 \leq k \leq J$, is given as:

$$\frac{\partial \ln f_i(\phi_k)}{\partial \lambda_{kl}} = \sum_t \frac{1}{n_t} \left[1_{(x_{itk}=x_{kl})} - \frac{\eta_{\gamma_k}(x_{kl}, x_{k0}; x_{it(k-1)}, ..., x_{it1}, x_{(k-1)0}, ..., x_{10}) \exp(\lambda_{kl})}{\sum_{l'=1}^{N_k} \eta_{\gamma_k}(x_{kl'}, x_{k0}; x_{it(k-1)}, ..., x_{it1}, x_{(k-1)0}, ..., x_{10}) \exp(\lambda_{kl'})} \right],$$

$$\frac{\partial \ln f_i(\phi_k)}{\partial \gamma_{kv}} = \sum_t \frac{1}{n_t}\Bigg[(x_{itk} - x_{k0})(x_{itv} - x_{v0})$$
$$- \frac{\sum_{l'=1}^{N_k} \eta_{\gamma_k}(x_{kl'}, x_{k0}; x_{it(k-1)}, ..., x_{it1}, x_{(k-1)0}, ..., x_{10}) \exp(\lambda_{kl'})(x_{kl'} - x_{k0})(x_{itv} - x_{v0})}{\sum_{l'=1}^{N_k} \eta_{\gamma_k}(x_{kl'}, x_{k0}; x_{it(k-1)}, ..., x_{it1}, x_{(k-1)0}, ..., x_{10}) \exp(\lambda_{kl'})}\Bigg],$$

where n_t is the number of consumers who made purchase on the occasion t in the same store.

(3) Updating $\beta_i | Y_i, X_i^{obs}, X_i^{mis}, \Pi, \Lambda$.
Random-Walk Metropolis-Hasting algorithm is used to update the individual-specific parameter β_i. Draw a proposal β_i^{prop} from $MVN(\beta_i^{old}, \kappa\Lambda^{-1})$, where β_i^{old} and Λ are parameter draws at the previous iteration, and the scaling parameter κ is adjusted in the RWMH algorithm to achieve approximately 30% acceptance rate. Calculate $p(\beta_i^{prop}) = \prod_{t\in T_i} f_{\beta_i=\beta_i^{prop}}(Y_{it}|X_{it}) f_\theta(\beta_i^{prop})$ and $p(\beta_i^{old}) = \prod_{t\in T_i} f_{\beta_i=\beta_i^{old}}(Y_{it}|X_{it}) f_\theta(\beta_i^{old})$, where X_{it} is the complete covariates with the missing components X_{it}^{mis} imputed in the previous iteration. Then with probability $q_i = min(1, p(\beta_i^{prop})/p(\beta_i^{old}))$, $\beta_i^{new} = \beta_i^{prop}$, and with probability $1 - q_i$, $\beta_i^{new} = \beta_i^{old}$.

(4) Updating $\Pi, \Lambda | \beta_i, Z_i$.
Let $\Theta = vec(\Pi')$. Our priors for the hyperparameter Θ and Λ are of the following form:

$$\Theta \sim \text{MVN}(\mu_\Pi, \Lambda_\Pi^{-1}), \quad \Lambda \sim \text{W}(\nu, S).$$

The prior distributions are standard for multivariate normal. In our applications, we set the constants in the priors as follows: μ_Π is a $n_r \times n_z$ vector of zeros; $\Lambda_\Pi = 0.01 * I_{n_r \times n_z}$; $\nu = n_r + 4$; $S = \nu I_{n_r}$. These priors are chosen to be noninformative relative to the data so that the resulting inference is dominated by data rather than by priors. With the priors, we follow the standard approach (Gelman et al. 2004, Rossi et al. 2005) to obtain the conditional draws as follows:

$$p(\Lambda|\Theta) = \text{W}\Big(\nu + N, \big(S^{-1} + \sum_{i=1}^{N} e_i e_i'\big)^{-1}\Big),$$

where $e_i = \beta_i - \Pi Z_i$, and

$$p(\Theta|\Lambda) = \text{N}\Big(\Sigma_\Pi\big[(\Lambda \otimes I_{n_z})H_{z\beta} + \Lambda_\Pi \mu_\Pi\big], \Sigma_\Pi\Big),$$

N is the number of subjects, $\Sigma_\Pi = \big[(\Lambda \otimes H_{zz}) + \Lambda_\Pi\big]^{-1}$, and

$$H_{zz} = \sum_{i=1}^{N} Z_i Z_i', \quad H_{z\beta} = \begin{bmatrix} H_{z\beta_1} \\ H_{z\beta_2} \\ \vdots \\ H_{z\beta_{n_r}} \end{bmatrix},$$

and $H_{z\beta_j} = \sum_{i=1}^{N} Z_i \beta_{ij}$, for $j = 1, ..., n_r$.

B. Simulation Studies

To motivate the research, we conduct a series of simulation studies that illustrate the limitations of the existing approaches to handling missing covariate problems and demonstrate the unique features and benefits of the proposed method.

B.1 *Robustness*

To demonstrate the importance of robust covariate modeling, we conduct the following simulation study to compare the performance of the proposed approach with that of fully Bayesian parametric approaches to handling missing covariates. We simulated the outcome y from a Poisson distribution with its mean λ_y following a regression model

$$\ln \lambda_y \;=\; \beta_0 + \beta_1 x.$$

We generated the covariate x from five different distributions. Figure 1 plots the shapes of these distributions.[2] The simulated complete datasets have roughly the same range of the values for the outcome and the covariate among the five different distributions. To create missingness, we divide each simulated complete dataset into two halves where in the first half all observations have the outcome larger than its median value. In the first half, the covariate x is set to be missing with a probability of 0.6, and in the second half with a probability of 0.2. For each generated dataset with missing covariate, we apply four methods to deal with the missing covariate problem. The first method is the complete-case (CC) analysis. The second method is to assume a normal model (NM) for the covariate X. That is, we assume that $X \sim N(\mu, \sigma^2)$. The third method applies a finite mixture of normal model (MNM) for X, whose density function is

$$f(x) = \sum_{j=1}^{J} \pi_j f_j(x), \quad \text{where } f_j(x) = N(\mu_j, \sigma_j^2) \text{ and } \sum_{j=1}^{J} \pi_j = 1, 0 \leq \pi_j \leq 1 \; \forall j,$$

where J is the number of components in the mixture, $f_j(x)$ is the jth component density of the mixture and π_j is the mixing weight for the jth component. The MNM is more flexible than the NM to accommodate different shapes of distributions. We assign the following proper priors for model parameters:

$$\mu_j \sim N(\mu_0, \sigma_0^2), \;\; \sigma_j^2 \sim IG(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_{j0}^2), \;\; (\pi_1, \cdots, \pi_J) \sim D(\alpha_1, \cdots, \alpha_J),$$

where IG denotes inverse-Gamma distribution, $D(\alpha_1, \cdots, \alpha_J)$ denotes Dirichlet distribution. The constants in these priors are given values such that the resulting priors are diffuse. In our analysis, we use a homogeneous mixture model which sets σ_j^2 to be equal in all mixture components since we find nonconvergence issues with heterogeneous variances as documented in the literature. Because there are at most two modes for the covariate distributions considered, we assume a mixture of two normal distributions (i.e. J=2). We develop two algorithms to sample the posterior distributions of joint model parameters under both parametric models. One is an importance sampling type algorithm and the other one uses

[2] To simulate from the Basin Shape distributions, we divide the interval (0,1) into 10 equal-length segments and assign the probability vector p_c on the 10 segments. Then within each segment we use a uniform distribution to generate x. The simulated values approximate the basin-shape distributions. For Basin Shape distribution I, $p_c = (0.3, 0.1, 0.05, 0.03, 0.02, 0.03, 0.05, 0.1, 0.3)$ and for basin shape distribution II, $p_c = (0.1, 0.1, 0.05, 0.03, 0.02, 0.03, 0.05, 0.1, 0.5)$.

a Metropolis-Hasting algorithm.[3] The fourth method applies our proposed distribution-free method (DFM). In addition to the above four methods to handle missing covariates, we also consider the analysis using the full data as a benchmark. For each of the five covariate distributions, we simulated 100 datasets. With the resulting sample of estimates, we calculate the bias, standard deviation (SD), the mean squared error (MSE) as well as 95% credible intervals for each Bayesian estimator.

The simulation results are summarized in Table 1. The two sampling algorithms for the parametric approaches give similar estimation results and the importance sampling type algorithm gives slightly more accurate results which are reported in Table 1. The results show that the complete-case analysis has a substantial amount of bias, and their coverage rates are poor. In addition, the method also has the largest variabilities (i.e. largest SDs) among all estimators, indicating the inefficiency of the method. A second message is that when the parametric assumption is correct, the parametric approach to handling the missing covariates can work well. For example, the NM performs well when the covariate is indeed normally distributed in that there is no bias and the coverage rates of credible intervals are close to the nominal rate. So does the MNM when the covariate is in fact a mixture of two normals. However, when the covariate distribution is misspecified, sizable bias and poor coverage rate can occur in the estimates of outcome regression parameters. For example, the NM can perform poorly when the covariate is not normally distributed in that there is sizable bias and coverage rate is far from the 95% nominal rate. This is so also for MNM when the covariate is not distributed as a mixture of two normals. Interestingly, when the covariate is normally distributed, i.e. a mixture of one normal, the coverage rate is noticeably different from the 95% nominal rate. As noted in Kamakura and Wedel (1997), selecting a proper number of components in mixture modeling is important. When using more components than needed, the number of observations in each component will decrease, which in turn may lead to computational instability and imprecise parameter estimates. On the other hand, when using less components than needed, the mixture model is not able to model the shape of the distribution adequately which can lead to bias in the estimation of outcome regression model parameters. This points to the importance of selecting the correct number of components in mixture modeling, which is not an easy task in the context of missing covariate problems. In contrast, the proposed distribution-free method does not require specifying a parametric covariate distribution or selecting the correct number of mixture components, and works well over different shapes of distributions. This demonstrates the value of a nonparametric procedure as an approach to minimize the impact of distributional assumptions in handling missing covariate problems.

B.2 *Computational Simplicity*

Recall that the posterior distribution contains the likelihood function of the joint model which involves integration with respect to the missing covariates:

$$L(\theta, \phi; Y, X^{obs}) \quad \propto \quad \int f_\theta(Y|X^{obs}, X^{mis}) f_\phi(X^{obs}, X^{mis}) dX^{mis}. \tag{5}$$

For a nonnormal regression model, such as the Poisson regression model, the above integral has no closed-form solution with a normal or a mixture of normal covariate model. For these two parametric covariate models, we have considered two algorithms to sample from the

[3] More details about the two algorithms are given in the following subsection.

posterior distribution of (θ, ϕ). The first one is an independence Metropolis sampler (referred to as the importance sampling type algorithm in the subsection above due to their similarity), where the proposal distribution used is a multivariate Student t distribution with 6 degrees of freedom, the location parameter being the MLEs of (θ, ϕ), and the scale parameter being the inverse of the negative Hessian matrix evaluated at the MLEs. The sampler requires MLEs. In this case, a very accurate numerical method for evaluating the integral is required. We have used an adaptive quadrature method for its evaluation. Note that this method can involve high-dimensional integrations and become computationally infeasible with multiple missing covariates. The second method is to view X^{mis} as latent variables and update X^{mis} with other model parameters. For the missing covariate of unit i, its conditional distribution is proportional to the product of the unit-level models: $f_\theta(y_i|x_i^{obs}, x_i^{mis})$ and $f_\phi(x_i^{obs}, x_i^{mis})$. For a nonlinear regression model with a normal or a mixture of normal covariate model, the conditional distribution is an unknown form of multivariate distribution. Therefore its updating requires a Metropolis-type algorithm, whose performance depends critically on the selection of proposal distributions. A reasonable proposal distribution should be customized to the shapes of unit-specific conditional distribution which can be very different from unit to unit. We have implemented a Hybrid Monte Carlo method to make these conditional draws because the method exploits the derivative information for each unit-level conditional distribution. Therefore the draws have a higher rate of acceptance with relative low autocorrelations.

In contrast, Section 4 shows that updating X^{mis} in our proposed Bayesian approach is simple in that its conditional distribution is a closed-form multinomial distribution on a set of known values, in which the probabilities in the multinomial distribution can be readily evaluated. Unlike the Independence Metropolis algorithm for parametric covariate modeling approaches, no numerical evaluation of the integrals is required. Unlike the Metropolis step of the second algorithm for parametric approaches, our updating step has an acceptance rate of 100%. This helps improve the convergence properties of the MCMC algorithm. In our simulation study, the computational times for obtaining 1000 effectively independent draws of the Poisson regression parameteres for NM, MNM and DFM are on average 0.25 min, 0.70 min and 0.15 min, respectively, demonstrating the computational simiplicity of the proposed distribution-free approach.[4]

B.3 *Flexibility*

An important strength of the proposed method is its modeling flexibility, which refers to the ability to incorporate the potentially complex dependence structure among covariates. As reviewed in Section 2 of the main paper, albeit its convenience in some special cases, the commonly-used multivariate normal model has limitations in accounting for potentially nonlinear relationships (e.g. quadratic or interaction effects) among covariates. To illustrate the point, we conduct the following simulation study. We simulate the outcome Y_i from $N(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}, 1), i = 1, \cdots, 500$. The covariates are simulated from the following distributions. We simulate x_{1i} from $N(0,1)$, x_{2i} from $N(\alpha_{20} + \alpha_{21} x_{1i}, 1)$ where $\alpha_{20} = 0$ and $\alpha_{21} = 0.5$, and x_{3i} from $N(\alpha_{30} + \alpha_{31} x_{1i} + \alpha_{32} x_{2i} + \alpha_{33} x_{1i} x_{2i}, 1)$ where $\alpha_{30} = 0$, $\alpha_{31} = \alpha_{32} = 0.5$. Note that α_{33} captures the nonlinear interaction effect of x_1 and x_2 on x_3. We consider two scenarios. In the first scenario we set $\alpha_{33} = 0$ which implies that there is no

[4] Our computations are coded in Fortran language and executed using a 2.7GHz Intel Xeon Processor with 4GB memory. The computational times for NM and MNM are based on the faster one of the two posterior sampling algorithms described in the above.

interaction effect and thus the covariates are jointly normal. In the second scenario we set $\alpha_{33} = 0.5$ in which case there is no equivalent joint normal distribution for these covariates. On the other hand, the proposed method can allow for such nonlinear relationships through flexible specification of odds ratio functions.

We then create the missingness in the covariates as follows. The covariate x_1 is always observed. We then divide the data into two halves according to the median value of the outcome y. In the first strata, the covariate x_2 is subject to missingness according to the following rule: $logit(Prob(G_2 = 1)) = \gamma_{20} + \gamma_{21}x_1 + \gamma_{22}x_3$, where G_2 is indicator for X_2 being observed. In the second half, the covariate x_3 is subject to missingness according to the following rule: $logit(Prob(G_3 = 1)) = \gamma_{30} + \gamma_{31}x_1 + \gamma_{32}x_2$. In simulations, we set $\gamma_{20} = 4, \gamma_{21} = 4, \gamma_{22} = 3, \gamma_{30} = -2, \gamma_{31} = -4, \gamma_{32} = 2$. This creates about 25% missingness for x_2 and 15% missingness for x_3. Then for each dataset with missingness, we compare the following three methods: the complete case analysis (CC), the method based on a joint normal model for covariates, and our proposed distribution-free method.

Table 2 summarizes the simulation results over 100 simulated datasets. As we can see from the table, the complete case analysis has large biases and the coverage rates are poor. When $\beta_{33} = 0$, both the method based on the MVN covariate model and our proposed method work equally well in that biases are removed and coverage rates are very close to nominal 95% confidenc rate. However, when $\beta_{33} = 0.5$, there is a substatial bias for the method based on the MVN covariate model and covarage rates for some parameters deteoriate. In contrast, the proposed method allows for such nonlinear relationship among covariates and continues to perfrom well. The above simulation study demonstrates that despite the fact that the multivariate normal model is a common model choice for missing covariates, it suffers from modeling inflexibility in that it cannot model nonlinear relationships. When using the MVN model in the presence of underlying nonlinear relationships among covariates, a significant amount of bias in the outcome regression model estimates can occur. It is reasonable to believe that such bias would also exist when a multivariate normal model is used for latent data to model discrete covariates if nonlinear relationships exist among these covariates. In contrast, the odds-ratio model is flexible enough to allow for such nonlinear effects, while not making any parametric distributional assumptions.

B.4 *Comparison with the Method of Chen (2004)*

Our Bayesian method also compares favorably in terms of its scalability to larger dimensional missing data problems and its ability to handle more complex models with the MLE method previously developed by Chen (2004). The MLE method requires evaluating the model likelihood. Although the integration in the likelihood is replaced by the summation over the finite points, the number of terms to evaluate can become large with multiple missing covariates. This can make the MLE method computationally expensive. To demonstrate the point, we conduct the following simulation study. We simulated data from a normal regression model in which $y_i \sim N(\beta_0 + \sum_{j=1}^{J} \beta_j x_{ji}, 1)$, where $i = 1, \cdots, 500$. In the simulation, $\beta_0 = -3$ and $\beta_j = 2, \forall\ j$. The covariate x_{ji} is simulated from a distribution which assigns equal probability mass on 10 points, $\{k/10\}, k = 1, \cdots, 10$. We simulated three datasets where the total number of covariates, J, is 1,2,3, respectively. For each generated dataset, we random select 40% of observations where we set the covariates to be missing. We then apply both the MLE approach and our proposed Bayesian approach to the simulated datasets. This would allow us to investigate the performance of two methods with increasingly larger dimensional missing data problems. Both methods give very similar estimation results, and our comparison focuses on computational time.

Figure 2 plots the comparison of computation times between two methods. It shows that the Bayesian approach takes less computational time. More importantly, the computational time appears to increase exponentially for MLE as the number of covariates increases but only linearly with the Bayesian approach. Note that with J missing covariates, the number of terms to evaluate for the summations in the likelihood is in the order of 10^J, which increases exponentially with J. In this case, the MLE can become computationally very expensive. In contrast, the MCMC algorithm employed in our Bayesian approach avoids evaluating likelihood and as a result it is more scalable and can handle much higher-dimensional missing data problems, commonly seen in marketing applications. The computational advantage would be even more dramatic for more complex models, such as when the covariate model involves interaction effects or when the outcome regression model becomes more complex in which cases the MLE approach would encounter even more difficulty. In summary, the study demonstrates significant computation advantage of the proposed Bayesian approach over Chen's MLE approach. Because of computational difficulty, there can be important data features that cannot be incorporated using the MLE approach, but can be handled with relative ease using the proposed Bayesian approach.

B.5 *Simulation Studies for Discrete Choice Models*

In this subsection, we use simulated datasets to illustrate the potential bias when not properly accounting for missing covariates in a brand choice model and evaluate the effectiveness of the proposed method to remove the bias. In this simulation study, we generate choice outcomes for 150 consumers over 300 days on three brands from a mixed multinomial logit model. Each consumer has three occasions to make choices. The choice outcome is simulated from the following discrete choice model, where the utility function of brand j for consumer i at the purchase occasion t is

$$u_{itj} = \psi_{01i} + \psi_{02i} + \psi_{1i} Price_{itj} + \epsilon_{itj},$$

where ψ_{01i}, ψ_{02i} are the individual-specific intercepts for brand 1 and 2 respectively, ψ_{1i} is the individual-specific price sensitivity, and ϵ_{ijt} is generated from *iid* Type-I Extreme value distribution. We simulate the consumer-specific parameters $(\psi_{01i}, \psi_{02i}, \psi_{1i})$ from a multivariate normal distribution with mean $(\pi_{01} = 3, \pi_{02} = 2, \pi_1 = -5)$ and a diagonal variance-covariance matrix whose diagonal elements are $(\sigma_{01}^2 = 0.6^2, \sigma_{02}^2 = 0.6^2, \sigma_1^2 = 1.6^2)$. We generate price values for three brands from three distributions: (1) a multivariate normal with the mean vector $(\mu_1, \mu_2, \mu_3) = (1.4, 1.2, 0.8)$ and a diagonal variance-covariance matrix Σ whose diagonal elements (i.e. variances) are 0.2^2; (2) the same distribution as in (1) except that we set the off-diagonal correlations to be $(\rho_{12}, \rho_{13}, \rho_{23}) = (-0.5, -0.5, 0.5)$, instead of zeros as in (1); (3) nonnormal distributions, in which the price values for each brand are simulated from a Basin Shape distribution. [5] These values are motivated by the ketchup dataset. We then set the price values to be missing for non-purchased brands at any day. For each simulated dataset, we fit the discrete choice model using the four methods (SI, MVN model, DF model I and DF model II) as explained in the analysis of the ketchup

[5] These distributions are simulated in a similar way as in Section 5.1. We divide the interval of price values (a, b) into 10 equal-length segments and assign the probability vector p_c on the 10 segments. Then within each segment we use a uniform distribution to generate price values. In simulations, $p_c = (0.6, 0.05, 0.05, 0.03, 0.02, 0.02, 0.03, 0.05, 0.05, 0.1)$. The price intervals (a, b) are (0.9,1.6), (1.0,1.4), (0.7,1.0) for three brands, respectively.

dataset. We repeat the analyses for 50 simulated datasets and the results are summarized in Table 3.

As shown in Table 3, the population parameter estimates (posterior means) in the brand utility function when using the simple imputation method have severe biases with the size of the average percentage bias above 40%. The precentage bias for an estimtor, $\hat{\pi}$, of a parameter, π, is defined as $1 - E(\hat{\pi})/\pi$. In particular, the estimates of the price sensitivity have a percentage bias of about 40%. This is because the SI method does not consider the dependence between the choice outcome and covariates when imputing prices for non-purchased brands and thus can lead to serious selection bias. The MVN model performs well when the covariates are generated from a multivariate normal distribution but can have sizable bias for nonnormal distirbutions. The DF Model I accounts for the dependence between the choice outcome and covariates but ignores the dependence between prices of different brands, when imputing prices. It also performs much better than the simple imputation method but still has sizable biases when prices among brands are correlated (e.g. 12% in the price sensitivity estimates). In comparison, the DF Model II accounts for both the correlations between covariates and the nonnormal feature of covariates. It recovers the model parameters well and removes the remaining bias in the MVN model and DF model I. Furthermore, the DF model II is efficient in that its estimates have essentially the same variability as those using correct parametric model assumptions. Overall, the simulation study using the repeated samples demonstrates that a simple imputation that does not account for the dependence between the brand choice outcome and the prices can lead to serious bias in the parameter estimations. It is also important to account for both the non-standard distributional shapes and correlations between covariates. Our proposed method can meet these challenges and recover the true parameter values well.

In practice, it is possible for a consumer to purchase more than one brand at a single purchase occasion. Although this does not occur in our ketchup dataset, we conduct a simulation study to evaluate the performance of our method in this case. Specifically we apply a multivariate probit model to account for multiple brand purchase and use our method to handle the missing covariate problem. The utility function is specified as $u_{itj} = \psi_{01i} + \psi_{02i} + \psi_{03i} + \psi_{1i} Price_{ijt} + \epsilon_{itj}$, where the error term $\epsilon_{it} = (\epsilon_{it1}, \epsilon_{it2}, \epsilon_{it3}) \sim N(0, \Sigma)$, and Σ is a correlation matrix. Unlike the mixed multinomial logit model, the observed choice outcome Y_{itj} is formed as follows:

$$y_{itj} = \begin{cases} 1, & \text{if } u_{itj} > 0 \\ 0, & \text{if } u_{itj} \leq 0. \end{cases}$$

The above model specification results in a multivariate probit model. In our simulation, we set the variance-covariance matrix Σ as an equi-correlation matrix with correlation coefficient being 0.5. We simulate the unit specific parameters $(\psi_{01i}, \psi_{02i}, \psi_{03i}, \psi_{1i})$ from a multivariate normal distribution with mean $(\pi_{01} = 6, \pi_{02} = 5, \pi_{03} = 4, \pi_1 = -5)$ and a diagonal variance-covariance matrix with diagonal elements $(\sigma_{01}^2 = \sigma_{02}^2 = \sigma_{03}^2 = 0.6^2, \sigma_1^2 = 1.6^2)$. We then follow the procedure described above to generate the covariates and to set missing values. We use the method proposed by Chib and Greenberg (1998) to update parameters in the multivariate probit model.

Table 4 summarizes the simulation results using 50 simulated datasets. The conclusions are broadly similar to the simulation result for the mixed multinomial logit model. The simple imputation results in biased parameter estimates in the utility function, including the correlation coefficient in the variance-covariate matrix Σ. The MVN covariate model works well when the covariates are indeed normally distributed but can have bias for nonnormal distributions. In this case, the distribution free method can help protect the analysis from being biased.

B.6 *Simulation Studies for Purchase Incidence Models*

In this subsection, we conduct a simulation study to demonstrate the potential bias when not properly accounting for missing covariates in a purchase incidence model and evaluate the performance of the proposed method to remove the bias. We simulate the purchase incidence outcome from the following hierarchical probit model. The utility to purchase in the store is $u_{it} = \beta_{0i} + \epsilon_{it}$, $t = 1, \cdots, n_i$. We simulate the within-unit error terms from the following multivariate distribution: $\epsilon_i = (\epsilon_{i1}, \cdots, \epsilon_{in_i}) \overset{ind}{\sim} N(0, \Sigma_i)$. We consider two forms of Σ_i. The first one is $\Sigma_i = I_{n_i \times n_i}$. That is, the error terms are independent standard normal across purchase occasions. In the second scenario, Σ_i follows an AR-1 form, where the diagonal entries of Σ_i are all ones, and the off-diagonal element of Σ_i is $\sigma_{ij} = \rho^{|k-j|}$ for the entry at kth row and jth column of Σ_i.[6] In our simulation, we set the value of the auto-correlation coefficient, ρ, as 0.5. The observed purchase outcome $Y_{it} = 1$ if $u_{it} > 0$ and $Y_{it} = 0$ otherwise. In the second level of the model, $\beta_{0i} = N(\pi_0 + \pi_1 Z_{1i} + \pi_2 Z_{2i}, \sigma^2_\beta), i = 1, \cdots, N$. The covariates are simulated from two scenarios. In the first scenario, Z_1 and Z_2 are simulated from standard normal distributions. In the second scenario, the covariate Z_1 is simulated from an exponential distribution with rate 1 and Z_2 from another exponential distribution with rate 1. In the simulation, we set parameters as following: $\pi_0 = -0.3, \pi_1 = 0.1$, $\pi_2 = 0.1$ and $\sigma_\beta = 0.3$. We set the number of consumers $N = 400$ and the number of observations per consumer $n_i = 50$. These settings are similar to those in the above retail store application.

We follow the procedure below to create missingness in the covariates. First, we calculate the purchasing rate for each consumer, that is, $\sum_j Y_{ij}/n_i$. We then divide the simulated complete dataset into two halves where in the first half all consumers have their purchase rates larger than the median of all consumers' purchase rates. In the first half, Z_1 is missing subject to the following missingness probability: $\text{logit}(P(Z_1 \text{ is observed})) = \gamma_{10} + \gamma_{11} Z_{2i}$, and in the second half, Z_2 is missing according to the missingness probability: $\text{logit}(P(Z_2 \text{ is observed})) = \gamma_{20} + \gamma_{21} Z_{1i}$. In the simulation, we set $\gamma_{10} = 0.5, \gamma_{11} = 1$ and $\gamma_{20} = -0.5, \gamma_{21} = 1$. For each resulting simulated data with missingness, we conduct the following analyses: (1) Fit the above hierarchical probit purchase incidence model using only complete cases; (2) Fit a joint model of purchase incidence outcome and covariates with all consumers using a MVN for the covariates; and (3) Fit a joint model of purchase incidence outcome and covariates with all consumers using our proposed DF approach. We also fit a hierarchical probit model to the complete dataset before we create missingness.

Table 5 summarizes the results on 100 simulated datasets. The complete-case analysis results in substantially biased estimates. Our proposed Bayesian approach removes the bias and recovers the true parameter well for both independent and correlated error terms and both normal and nonnormal covariate distributions. When the parametric MVN covariate model is used, it recovers the parameters equally well as our approach when the covariate is truly jointly normal. On the other hand, for nonnormal covariate distributions, substantial bias arises for the parametric approach, whereas the DF approach continues to perform well.

[6] When simulating the posterior distribution of model unknowns, a Metropolis-Hasting step was used to update the auto-correlation coefficient ρ.

References

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J. and Stuart, A. (2010) "Optimal Tuning of the Hybrid Monte-Carlo Algorithm,", arXiv:1001.4460v1 [math.PR].

Chen, H. Y. (2004) "Nonparametric and Semiparametric Models for Missing Covariates in Parametric Regression," *Journal of the American Statistical Association,* **99**, 1176–1189.

Chib, S. and Greenberg, E. (1998) "Analysis of Multivariate Probit Models," *Biometrika,* **85**, 347–361.

Duane, S., Kennedy, A.D., Pendleton, B.J. and Roweth, D. (1987) "Hybrid Monte Carlo," *Physics Letters B,* **195**, 216-222.

Erdem, T., Keane, M.P. and Sun, B. (1999) "Missing Price and Coupon Availability Data in Scanner Panels: Correcting for the Self-selection Bias in Choice Model Parameters," *Journal of Econometrics,* **89**, 177–196.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004), *Bayesian Data Analysis, 2nd Ed.* Chapman & Hall, London.

Kamakura, W.A. and Wedel, M (1997), "Statistical Data Fusion for Cross-Tabulation," *Journal of Marketing Research,* **34**: 485–98.

Liu, J. (2001) *Monte Carlo Strategies in Scientific Computing,* Springer-Verlag, New York.

Rossi, P.E., Allenby, G., McCulloch, R. (2005) *Bayesian Statistics and Marketing,* Wiley, London.

Tanner, M.A. and Wong, W.H. (1987) "The Calculation of Posterior Distributions by Data Augmentation (with discussion)," *Journal of the American Statistical Association,* **82**, 528–550.

Table 1: Simulation Results on the Impact of Covariate Distributional Assumption. FD: full data analysis; CC: complete-case analysis; NM: normal covariate model; MNM: mixture of normal model; DFM: distribution-free model.

Covariate	Method	$\beta_0 = -3$				$\beta_1 = 4.5$			
Distribution		Bias	SD	MSE	CR	Bias	SD	MSE	CR
Normal	FD	-0.02	0.21	0.21	95%	0.02	0.31	0.31	93%
	CC	-0.95	0.34	1.01	12%	0.95	0.47	1.06	38%
	NM	-0.04	0.24	0.25	92%	0.05	0.35	0.35	93%
	MNM	-0.07	0.25	0.26	86%	0.08	0.37	0.38	84%
	DFM	-0.05	0.26	0.26	94%	0.05	0.37	0.37	95%
Normal Mixture	FD	-0.03	0.22	0.22	93%	0.04	0.28	0.28	95%
	CC	-0.91	0.45	1.02	20%	0.93	0.58	1.10	49%
	NM	0.39	0.20	0.44	52%	-0.52	0.27	0.58	53%
	MNM	-0.06	0.28	0.29	94%	0.08	0.36	0.37	93%
	DFM	-0.06	0.29	0.29	96%	0.07	0.37	0.37	94%
Uniform	FD	0.01	0.20	0.20	91%	-0.01	0.24	0.24	93%
	CC	-0.91	0.37	0.99	11%	0.93	0.45	1.03	30%
	NM	0.28	0.23	0.36	68%	-0.34	0.28	0.44	69%
	MNM	0.05	0.26	0.27	90%	-0.06	0.31	0.32	89%
	DFM	-0.01	0.26	0.26	94%	0.02	0.31	0.31	94%
Basin Shape I	FD	-0.02	0.20	0.20	97%	0.02	0.22	0.22	97%
	CC	-1.93	0.66	2.04	0%	1.90	0.73	2.03	5%
	NM	1.19	0.15	1.20	0%	-1.22	0.16	1.24	0%
	MNM	0.25	0.25	0.35	85%	-0.26	0.27	0.38	85%
	DFM	0.07	0.33	0.34	95%	-0.06	0.35	0.36	93%
Basin Shape II	FD	0.01	0.27	0.27	91%	-0.02	0.29	0.29	91%
	CC	-0.42	0.58	0.71	84%	0.13	0.68	0.68	83%
	NM	1.80	0.16	1.80	0%	-2.05	0.17	2.06	0%
	MNM	0.61	0.21	0.65	23%	-0.71	0.23	0.75	15%
	DFM	0.07	0.29	0.30	92%	-0.08	0.31	0.33	91%

Table 2: Simulation Results on the Flexibility of Modeling Dependence Structure.

Parameter	FD				CC				MVN				DFM			
	Bias	SD	MSE	CR	Bias	SD	MSE	CR	Bias	SD	MSE	CR	Bias	SD	MSE	CR
$\alpha_{33} = 0$																
$\beta_0 = 0$	-0.01	0.10	0.10	92%	0.36	0.12	0.38	12%	-0.00	0.10	0.10	93%	0.01	0.10	0.10	96%
$\beta_1 = 0.5$	0.01	0.10	0.10	98%	0.50	0.13	0.51	3%	0.01	0.11	0.11	98%	-0.00	0.11	0.11	96%
$\beta_2 = 0.5$	-0.01	0.10	0.10	96%	-0.30	0.12	0.31	29%	-0.01	0.12	0.12	95%	-0.01	0.11	0.11	97%
$\beta_3 = 0.5$	-0.00	0.10	0.10	93%	0.10	0.10	0.15	81%	-0.00	0.10	0.10	94%	0.00	0.09	0.09	94%
$\alpha_{33} = 0.5$																
$\beta_0 = 0$	-0.01	0.10	0.10	91%	0.43	0.12	0.45	4%	-0.05	0.10	0.11	82%	-0.01	0.10	0.10	92%
$\beta_1 = 0.5$	0.01	0.09	0.09	98%	0.55	0.13	0.57	1%	0.04	0.10	0.11	93%	0.00	0.11	0.11	93%
$\beta_2 = 0.5$	-0.01	0.09	0.09	97%	-0.30	0.12	0.32	28%	-0.10	0.12	0.15	81%	-0.01	0.12	0.12	95%
$\beta_3 = 0.5$	-0.01	0.08	0.08	92%	-0.02	0.10	0.10	89%	0.04	0.09	0.10	88%	0.00	0.10	0.10	92%

Table 3: Simulation Result B.5.A. The first row for each parameter in a mixed multinomial logit discrete choice model lists the average of the estimates (posterior means) over all the simulations and their percentage bias in the parenthesis. The second row for each parameter lists the standard deviation of the estimates over all the simulations.

Case I: Independent MVN

Parameter	True Value	SI	MVN	DF Model I	DF Model II
Intercept$_{\text{brand1}}$ (π_{01})	3	1.71(-43%)	3.06 (2%)	3.05 (2%)	2.98 (-1%)
		0.41	0.47	0.46	0.47
Intercept$_{\text{brand2}}$ (π_{02})	2	1.09 (-46%)	2.05 (3%)	2.07 (3%)	2.03 (2%)
		0.30	0.36	0.36	0.37
Price (π_1)	-5	-2.67 (47%)	-5.07 (-2%)	-5.06 (-1%)	-4.94 (2%)
		0.64	0.74	0.72	0.73

Case II: Correlated MVN

Parameter	True Value	SI	MVN	DF Model I	DF Model II
Intercept$_{\text{brand1}}$ (π_{01})	3	1.84(-39%)	3.12 (4%)	3.42 (14%)	3.11 (3%)
		0.32	0.42	0.44	0.40
Intercept$_{\text{brand2}}$ (π_{02})	2	1.11 (-45%)	2.03 (2%)	2.24 (12%)	2.07 (3%)
		0.21	0.33	0.35	0.31
Price (π_1)	-5	-2.84 (43%)	-5.13 (-3%)	-5.61 (-12%)	-5.22 (4%)
		0.49	0.66	0.75	0.67

Case III: Non-normal Distribution

Parameter	True Value	SI	MVN	DF Model I	DF Model II
Intercept$_{\text{brand1}}$ (π_{01})	3	2.01(-33%)	3.38 (13%)	3.09 (3%)	3.10 (3%)
		0.49	0.42	0.34	0.33
Intercept$_{\text{brand2}}$ (π_{02})	2	1.34 (-34%)	2.37 (18%)	2.06 (3%)	1.99 (-1%)
		0.51	0.47	0.35	0.34
Price (π_1)	-5	-3.29 (34%)	-6.26 (-25%)	-5.12 (-2%)	-4.95 (1%)
		0.86	1.23	0.78	0.75

Table 4: Simulation Result B.5.B. The first row for each parameter in a multivariate probit outcome model lists the average of the estimates (posterior means) over all the simulations and their percentage bias in the parenthesis. The second row for each parameter lists the standard deviation of the estimates over all the simulations.

Case I: Independent MVN

Parameter	True Value	SI	MVN	DF Model I	DF Model II
Intercept$_{\text{brand1}}$ (π_{01})	6	5.18(-14%)	6.21 (4%)	6.15 (3%)	6.15 (3%)
		0.41	0.41	0.40	0.39
Intercept$_{\text{brand2}}$ (π_{02})	5	4.22 (-16%)	5.21 (4%)	5.11 (3%)	5.11 (3%)
		0.36	0.40	0.39	0.38
Intercept$_{\text{brand3}}$ (π_{03})	4	3.41 (-15%)	4.18 (5%)	4.13 (3%)	4.12 (3%)
		0.28	0.29	0.28	0.27
Price (π_1)	-5	-4.47 (11%)	-5.25 (-5%)	-5.25 (-5%)	-5.13 (-3%)
		0.38	0.35	0.32	0.32
ρ	0.5	0.40 (-20%)	0.51 (-2%)	0.51 (-2%)	0.50 (1%)
		0.07	0.08	0.07	0.07

Case II: Correlated MVN

Parameter	True Value	SI	MVN	DF Model I	DF Model II
Intercept$_{\text{brand1}}$ (π_{01})	6	5.30(-12%)	6.27 (4%)	6.26 (4%)	6.18 (3%)
		0.61	0.45	0.57	0.52
Intercept$_{\text{brand2}}$ (π_{02})	5	4.39 (-12%)	5.20 (4%)	5.19 (4%)	5.14 (3%)
		0.52	0.40	0.48	0.46
Intercept$_{\text{brand3}}$ (π_{03})	4	3.54 (-12%)	4.19 (5%)	4.18 (4%)	4.17 (4%)
		0.35	0.28	0.32	0.30
Price (π_1)	-5	-4.61 (8%)	-5.21 (-4%)	-5.20 (-4%)	-5.20 (-4%)
		0.52	0.37	0.45	0.45
ρ	0.5	0.40 (-20%)	0.51 (2%)	0.51 (2%)	0.51 (2%)
		0.09	0.08	0.09	0.09

Case III: Non-normal Distribution

Parameter	True Value	SI	MVN	DF Model I	DF Model II
Intercept$_{\text{brand1}}$ (π_{01})	6	5.42(-10%)	6.54 (9%)	6.21 (4%)	6.18 (3%)
		0.58	0.52	0.51	0.52
Intercept$_{\text{brand2}}$ (π_{02})	5	4.52 (-10%)	5.47 (9%)	5.16 (3%)	5.13 (3%)
		0.61	0.50	0.47	0.49
Intercept$_{\text{brand3}}$ (π_{03})	4	3.65 (-9%)	4.31 (8%)	4.12 (3%)	4.11 (3%)
		0.45	0.38	0.36	0.38
Price (π_1)	-5	-4.52 (10%)	-5.48 (-10%)	-5.15 (-3%)	-5.11 (2%)
		0.59	0.50	0.47	0.48
ρ	0.5	0.42 (-16%)	0.50 (-0%)	0.51 (2%)	0.51 (2%)
		0.05	0.06	0.05	0.06

Table 5: Simulation Result B.6. The first row for each parameter in a hierarchical Probit model lists the average of the estimates (posterior means) over all the simulations and their percentage bias in the parenthesis. The second row for each parameter lists the standard deviation of the estimates over all the simulations.

Parameter	True Value	FD	CC	MVN model	DF model
		Independent Error Term and Normal Covariate Distributions			
Intercept (π_0)	-0.3	-0.302 (-1%)	-0.250 (17%)	-0.302 (-1%)	-0.301 (-0%)
		0.029	0.037	0.029	0.030
Z1 (π_1)	0.1	0.099 (-1%)	0.068 (-32%)	0.099 (-1%)	0.099 (-1%)
		0.015	0.018	0.017	0.015
Z2 (π_2)	0.1	0.102 (2%)	0.120 (20%)	0.099 (-1%)	0.098 (-2%)
		0.018	0.022	0.018	0.019
σ_β	0.3	0.302 (1%)	0.298 (-1%)	0.298 (-1%)	0.299 (-0%)
		0.015	0.020	0.015	0.015
		Independent Error Term and Non-normal Covariate Distributions			
Intercept (π_0)	-0.3	-0.305 (-2%)	-0.242 (19%)	-0.278 (8%)	-0.303 (-1%)
		0.029	0.038	0.028	0.030
Z1 (π_1)	0.1	0.103 (3%)	0.072 (-28%)	0.089 (-11%)	0.104 (4%)
		0.017	0.020	0.019	0.018
Z2 (π_2)	0.1	0.102 (2%)	0.112 (12%)	0.121 (21%)	0.100 (0%)
		0.018	0.021	0.015	0.019
σ_β	0.3	0.299 (-0%)	0.297 (-1%)	0.278 (-8%)	0.299 (-0%)
		0.015	0.020	0.015	0.015
		AR(1) Error Term and Normal Covariate Distributions			
Intercept (π_0)	-0.3	-0.303 (-1%)	-0.233 (22%)	-0.302 (-1%)	-0.301 (-0%)
		0.020	0.031	0.021	0.021
Z1 (π_1)	0.1	0.099 (-1%)	0.019 (-81%)	0.100 (0%)	0.101 (1%)
		0.022	0.026	0.024	0.024
Z2 (π_2)	0.1	0.097 (-3%)	0.143 (43%)	0.099 (-1%)	0.098 (-2%)
		0.018	0.024	0.022	0.022
σ_β	0.3	0.296 (-1%)	0.263 (-12%)	0.294 (-2%)	0.293 (-2%)
		0.022	0.042	0.023	0.023
ρ	0.5	0.500 (0%)	0.503 (1%)	0.500 (0%)	0.500 (0%)
		0.010	0.016	0.010	0.010
		AR(1) Error Term and Non-normal Covariate Distributions			
Intercept (π_0)	-0.3	-0.301 (-0%)	-0.228 (24%)	-0.284 (6%)	-0.300 (0%)
		0.035	0.046	0.031	0.037
Z1 (π_1)	0.1	0.102 (2%)	0.066 (-34%)	0.086 (-14%)	0.102 (2%)
		0.020	0.023	0.021	0.021
Z2 (π_2)	0.1	0.101 (1%)	0.113 (13%)	0.131 (31%)	0.100 (0%)
		0.021	0.023	0.016	0.023
σ_β	0.3	0.292 (-3%)	0.288 (-4%)	0.266 (-12%)	0.293 (-2%)
		0.018	0.024	0.021	0.019
ρ	0.5	0.502 (0%)	0.502 (-0%)	0.502 (0%)	0.502 (0%)
		0.010	0.013	0.010	0.010

Table 6: Estimation Result in the Ketchup Purchase Data.

Table presents posterior mean (posterior SD) of each parameter. SI stands for the simple imputation model. MVN stands for multivariate normal covariate model. DF stands for the distribution-free covariate model. The parameter p_{bl} in the price model is the estimated marginal probability mass at the lth price value of brand b, where these price values, in the order presented in the table, are as follows: for Heinz: 0.99, 1.19, 1.39,1.45, 1.49,1.59; for Hunt's: 0.89, 0.99,1.19, 1.39, 1.45, 1.49, 1.59; for Store brand: 0.59, 0.69, 0.89,0.95, 0.99. The parameter c_{bl} in the coupon model is the estimated marginal probability mass at the lth coupon value of brand b, where these coupon values, in the order presented in the table, are as follows: for Heinz: 0.00, 0.25, 0.30, 0.36, 0.40, 0.50, 0.90; for Hunt's: 0.00,0.30, 0.36, 0.40, 0.50, 1.00; for Store brand: 0.00.

Parameter	SI Model	MVN Model	DF Model I	DF Model II
Choice Outcome Model				
Intercept (Heinz)	1.8(0.28)	3.5 (0.44)	3.7 (0.45)	3.0 (0.36)
Intercept (Hunts)	1.6(0.20)	3.1 (0.36)	3.3 (0.36)	2.8 (0.31)
Price	-3.4(0.50)	-6.1(0.76)	-6.6(0.86)	-5.4(0.66)
Coupon	53.6 (3.32)	2.4 (0.56)	4.4 (1.28)	3.5 (1.24)
Σ_{11}	2.2 (1.1)	4.2 (1.5)	6.1(2.6)	4.8 (1.8)
Σ_{22}	1.4(0.6)	3.7 (1.1)	4.9(1.7)	4.0 (1.2)
Σ_{33}	2.8 (2.1)	15.4 (5.8)	18.5(8.2)	15.2 (5.9)
Σ_{44}	3.3 (8.8)	1.44 (1.21)	6.9(6.4)	3.2 (2.1)
Σ_{12}	0.98(0.77)	3.1 (1.2)	4.6 (2.0)	3.5 (1.4)
Σ_{13}	-1.56(1.48)	2.8 (2.3)	1.4 (4.2)	2.9 (2.7)
Σ_{14}	0.24(1.87)	0.3 (0.4)	1.6(3.8)	0.4 (2.1)
Σ_{23}	-0.91(1.04)	3.6 (2.1)	2.1(3.5)	3.2 (2.4)
Σ_{24}	0.28(1.28)	0.6 (0.8)	1.5 (3.0)	0.3 (1.8)
Σ_{34}	-0.18 (2.96)	-0.4 (0.2)	-2.2(7.8)	-0.8 (3.9)
Covariate Model (MVN)				
μ_1		1.23 (0.02)		
μ_2		1.05 (0.02)		
μ_3		0.74 (0.02)		
μ_4		1.26 (0.03)		
μ_5		1.16 (0.03)		
μ_6		0.82 (0.03)		
μ_7		0.11 (0.02)		
μ_8		0.10 (0.02)		
μ_9		0.01 (0.00)		
Σ_{11}		0.027 (0.001)		
Σ_{22}		0.026 (0.001)		
Σ_{33}		0.009 (0.001)		
Σ_{44}		0.028 (0.002)		
Σ_{55}		0.027 (0.002)		
Σ_{66}		0.014 (0.002)		
Σ_{77}		0.023 (0.001)		
Σ_{88}		0.024 (0.001)		
Σ_{99}		0.002 (0.000)		
Σ_{12}		-0.009 (0.001)		
Σ_{13}		-0.006 (0.001)		
Σ_{14}		0.021 (0.001)		

continued on next page

Table 6: *continued*

Parameter	SI Model	MVN Model	DF Model I	DF Model II
Σ_{15}		-0.007 (0.001)		
Σ_{16}		-0.005 (0.001)		
Σ_{17}		0.002 (0.001)		
Σ_{18}		-0.004 (0.001)		
Σ_{19}		-0.001 (0.001)		
Σ_{23}		0.006 (0.001)		
Σ_{24}		-0.007 (0.001)		
Σ_{25}		0.020 (0.001)		
Σ_{26}		0.007 (0.001)		
Σ_{27}		0.003 (0.001)		
Σ_{28}		0.007 (0.001)		
Σ_{29}		0.000 (0.001)		
Σ_{34}		-0.006 (0.001)		
Σ_{35}		0.006 (0.001)		
Σ_{36}		0.008 (0.001)		
Σ_{37}		0.002 (0.001)		
Σ_{38}		0.004 (0.001)		
Σ_{39}		0.000 (0.001)		
Σ_{45}		-0.009 (0.001)		
Σ_{46}		-0.007 (0.001)		
Σ_{47}		0.003 (0.001)		
Σ_{48}		-0.005 (0.001)		
Σ_{49}		-0.000 (0.001)		
Σ_{56}		0.004 (0.001)		
Σ_{57}		0.003 (0.001)		
Σ_{58}		0.009 (0.001)		
Σ_{59}		-0.000 (0.001)		
Σ_{67}		0.002 (0.001)		
Σ_{68}		0.003 (0.001)		
Σ_{69}		0.001 (0.001)		
Σ_{78}		0.002 (0.001)		
Σ_{79}		-0.001 (0.001)		
Σ_{89}		-0.001 (0.001)		
Covariate Model (DF)				
(1) Price Model				
Heinz				
p_{11}			0.21(0.02)	0.20(0.02)
p_{12}			0.48(0.03)	0.49(0.03)
p_{13}			0.19(0.02)	0.22(0.02)
p_{14}			0.008(0.005)	0.004(.005)
p_{15}			0.054(0.014)	0.051 (0.02)
p_{16}			0.053(0.014)	0.035(0.02)
Hunts				

continued on next page

Table 6: *continued*

Parameter	SI Model	MVN Model	DF Model I	DF Model II
p_{21}			0.31(0.02)	0.31(0.02)
p_{22}			0.24(0.02)	0.27 (0.02)
p_{23}			0.36(0.03)	0.37(0.04)
p_{24}			0.03(0.01)	0.02(0.01)
p_{25}			0.0095(0.007)	0.005(.01)
p_{26}			0.025(0.01)	0.016 (0.01)
p_{27}			0.022(0.01)	0.018 (0.01)
Store Brand				
p_{31}			0.11(0.02)	0.10 (0.02)
p_{32}			0.66(0.03)	0.66 (0.03)
p_{33}			0.134(0.02)	0.164 (0.02)
p_{34}			0.024(0.01)	0.020 (0.02)
p_{35}			0.08(0.02)	0.06 (0.02)
(2) Coupon Model				
Heinz				
c_{11}			0.87(0.02)	0.84 (0.02)
c_{12}			0.005(0.003)	0.003(0.002)
c_{13}			0.040(0.009)	0.047 (0.01)
c_{14}			0.035(0.009)	0.039 (0.01)
c_{15}			0.002(0.002)	0.0011 (0.002)
c_{16}			0.050(0.011)	0.066(0.02)
c_{17}			0.0027(0.002)	0.0029 (0.002)
Hunts				
c_{21}			0.93(0.012)	0.90 (0.01)
c_{22}			0.002(0.002)	0.002(0.002)
c_{23}			0.002(0.002)	0.001 (0.002)
c_{24}			0.005(0.003)	0.006 (0.004)
c_{25}			0.06(0.01)	0.088 (0.02)
c_{26}			0.002(0.002)	0.004 (0.003)
Store Brand				
c_{31}			1.00 (0.00)	1.00 (0.00)
(3) Dependence				
γ^P_{10}				39.1 (3.7)
γ^P_{20}				41.0 (8.0)
γ^P_{21}				-22.3 (5.1)
γ^P_{30}				28.4(7.9)
γ^P_{31}				0.82(4.0)
γ^P_{32}				12.97 (6.5)
γ^C_{1}				2.4 (1.8)
γ^C_{2}				9.2 (1.8)
Marginal LL	-1164.40	-1122.63	-3428.16	-2631.89

[hp]

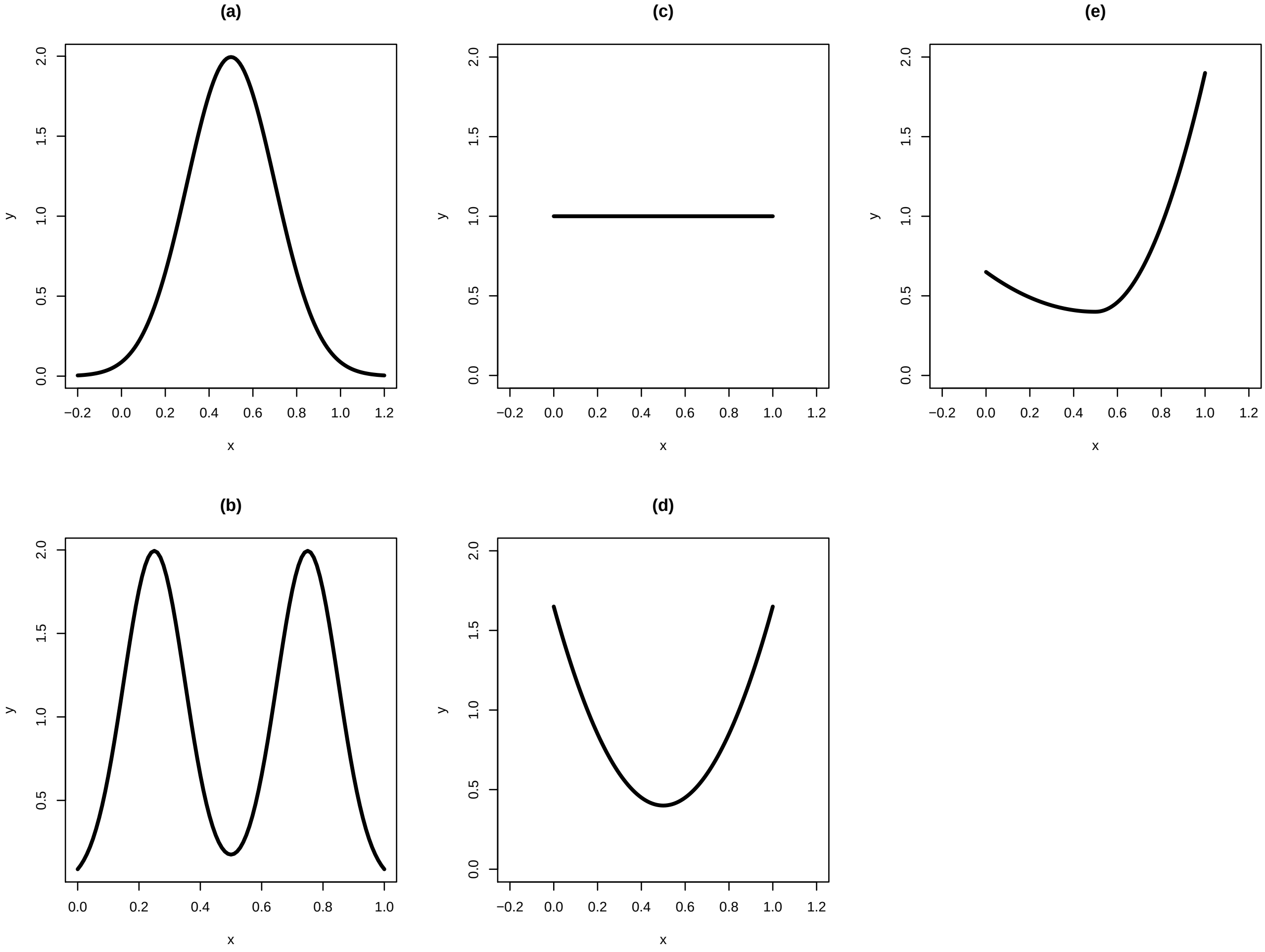


Figure 1. The Shapes of the Different Covariate Distributions. (a): N(0.5, 0.2^2). (b): 0.5*N(0.25,0.1^2)+0.5*N(0.75, 0.1^2). (c) Uniform(0,1). (d) Basin Shape I. (e) Basin Shape II.

[hP]

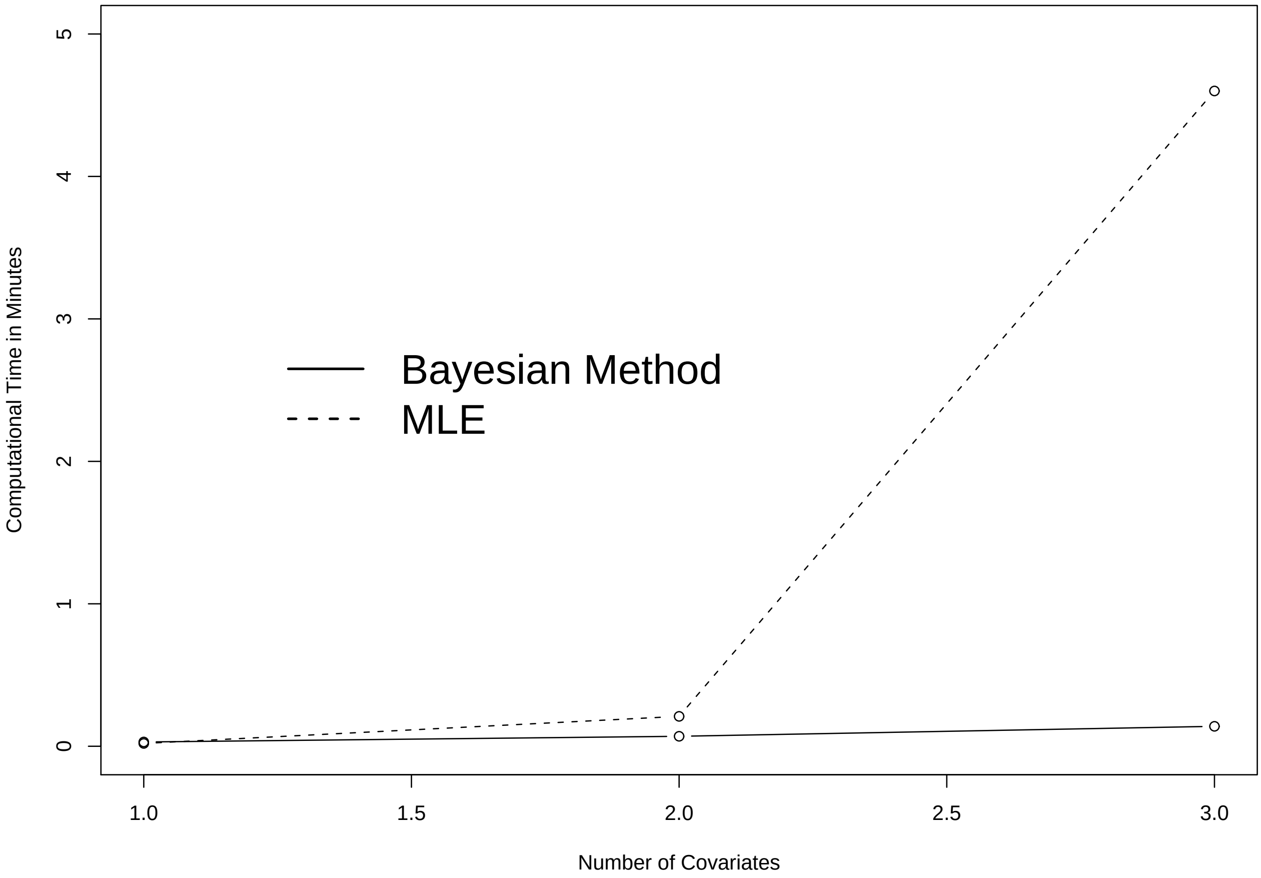


Figure 2. Comparison of Computational Times between Bayesian Method and the MLE method. The computation time for the Bayesian method is the time to obtain 1000 effectively independent draws plus the time for burn-in.