



University of Pennsylvania
ScholarlyCommons

Publicly Accessible Penn Dissertations

1-1-2015

Bayesian Modeling of Consumer Behavior in the Presence of Anonymous Visits

Julie Novak

University of Pennsylvania, julnovak@wharton.upenn.edu

Follow this and additional works at: <http://repository.upenn.edu/edissertations>

 Part of the [Statistics and Probability Commons](#)

Recommended Citation

Novak, Julie, "Bayesian Modeling of Consumer Behavior in the Presence of Anonymous Visits" (2015). *Publicly Accessible Penn Dissertations*. 1107.

<http://repository.upenn.edu/edissertations/1107>

This paper is posted at ScholarlyCommons. <http://repository.upenn.edu/edissertations/1107>

For more information, please contact libraryrepository@pobox.upenn.edu.

Bayesian Modeling of Consumer Behavior in the Presence of Anonymous Visits

Abstract

Tailoring content to consumers has become a hallmark of marketing and digital media, particularly as it has become easier to identify customers across usage or purchase occasions. However, across a wide variety of contexts, companies find that customers do not consistently identify themselves, leaving a substantial fraction of anonymous visits. We develop a Bayesian hierarchical model that allows us to probabilistically assign anonymous sessions to users. These probabilistic assignments take into account a customer's demographic information, frequency of visitation, activities taken when visiting, and times of arrival. We present two studies, one with synthetic and one with real data, where we demonstrate improved performance over two popular practices (nearest-neighbor matching and deleting the anonymous visits) due to increased efficiency and reduced bias driven by the non-ignorability of which types of events are more likely to be anonymous. Using our proposed model, we avoid potential bias in understanding the effect of a firm's marketing on its customers, improve inference about the total number of customers in the dataset,

and provide more precise targeted marketing to both previously observed and unobserved customers.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Statistics

First Advisor

Eric T. Bradlow

Second Advisor

Shane T. Jensen

Subject Categories

Statistics and Probability

BAYESIAN MODELING OF CONSUMER BEHAVIOR IN THE PRESENCE OF
ANONYMOUS VISITS

Julie Esther Novak

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2015

Supervisor of Dissertation

Co-Supervisor of Dissertation

Shane T. Jensen

Associate Professor, Statistics

Eric T. Bradlow

Professor of Marketing, Statistics and
Education

Graduate Group Chairperson

Eric T. Bradlow, Professor of Marketing, Statistics, and Education

Dissertation Committee

Elea McDonnell Feit, Assistant Professor, Marketing, Drexel University

Edward George, Professor of Statistics

Acknowledgements

I was very fortunate to have Shane Jensen as my co-advisor. I appreciate all of your support and encouragement over the last four years at Wharton. It was your class that inspired my thesis work to be in Bayesian Statistics. I hope we continue working together for years to come.

Eric Bradlow has been a demanding co-advisor and inspiring mentor. Thank you for having such high expectations- they led me to learn and develop greatly over the last few years, both as a researcher and as a person.

Elea Feit has made a sizable impact on this dissertation and in my academic development. Thank you for your encouragement and the one-on-one time you spent with me.

Ed George, thank you for being so supportive, optimistic, and always believing in me.

I owe a very special thanks to Warren Ewens. You have been my role model as a teacher, an academic, and most importantly, as a person.

I want to thank my wonderful group of colleagues in the Statistics Department at Wharton. Tung, Justin, Ville, Colin, and Kory, I was very lucky to have such a collaborative and supportive class. Our camaraderie helped me make a smooth transition from undergraduate to graduate life. Sameer, Veronika, and Sivan, you were wonderful colleagues that have now become my lifelong friends. Carol, I will

miss your wit, sarcasm, and realistic outlook on life.

My friends have played a tremendous role in my life as a PhD student. First and foremost, thank you Clara. I would have never made it through without our life saving study Skype sessions over the last four years. Neda, our long evening walks in Rittenhouse Square and cooking lessons at your house made Philly feel like a second home to me. Alison, our Euro-trip was the perfect break to refuel me for the second half of my PhD. Lieke, our bi-weekly coffee dates made my first two years a lot more fun. Scott, thank you for always being there for me.

Finally, my family deserves a trophy for how much they tolerate and deal with me. I genuinely believe that I have the most caring parents and brother in the world. It's encouraging to know that I have your genes. Mom, Dad, Andy (and my dog, Duke). Thank you.

ABSTRACT

BAYESIAN MODELING OF CONSUMER BEHAVIOR IN THE PRESENCE OF ANONYMOUS VISITS

Julie Esther Novak

Shane T. Jensen

Eric T. Bradlow

Tailoring content to consumers has become a hallmark of marketing and digital media, particularly as it has become easier to identify customers across usage or purchase occasions. However, across a wide variety of contexts, companies find that customers do not consistently identify themselves, leaving a substantial fraction of anonymous visits. We develop a Bayesian hierarchical model that allows us to probabilistically assign anonymous sessions to users. These probabilistic assignments take into account a customer’s demographic information, frequency of visitation, activities taken when visiting, and times of arrival. We present two studies, one with synthetic and one with real data, where we demonstrate improved performance over two popular practices (nearest-neighbor matching and deleting the anonymous visits) due to increased efficiency and reduced bias driven by the non-ignorability of which types of events are more likely to be anonymous. Using our proposed model, we avoid potential bias in understanding the effect of a firm’s marketing on its customers, improve inference about the total number of customers in the dataset, and provide more precise targeted marketing to both previously observed and unobserved customers.

Contents

Title Page	i
Acknowledgements	ii
Abstract	iv
Table of Contents	v
List of Tables	ix
List of Illustrations	x
1 Introduction and Motivation	1
2 Exploratory Data Analysis	6
2.1 Description of the Data	6
2.2 Marketing Actions	9
3 Bayesian Hierarchical Model	13
3.1 General Model for Customer Visits	13
3.2 Estimation via Markov Chain Monte Carlo Sampling	21
3.3 Allowing for Completely Unknown Users	24
4 Approaches to Missing Data	27

4.1	Case-Deletion Method	28
4.2	Nearest-Neighbor Matching	28
5	Summary for Behavior of Modeling Approaches under Different Data Settings	31
6	Synthetic Data Evaluation	34
6.1	Correlation between Missingness and Propensity to Visit in Response to Marketing	35
6.2	Correlation between Missingness and Effect of Marketing on the Propensity to Engage in an Activity	41
6.3	Correlation between Missingness and Overall Propensity to Engage in Activities	46
6.4	Estimating the Number of Unique Customers	51
6.5	Summary of Synthetic Data Evaluations	53
7	Application to a Retail Website	55
7.1	Relationship between Missingness and Propensity to Visit in Response to a Marketing Action	56
7.2	Application to a Retail Dataset with its True Missingness Pattern . .	63
8	Conclusion and Future Work	65
8.1	Computational Efficiency	68
	Appendix A Gibbs Sampler	72
A.1	Prior Distributions on Global Parameters	72
A.2	Gibbs Sampler Steps 1 through 8	73
	Appendix B Computational Details	78

B.1	Parameter Recovery	78
B.2	Demonstration that Subsampling Works	79
	References	84

List of Tables

2.1	Frequency of Number of Visits Taken by Identified Customers	7
2.2	Examples of Customer Emails	9
3.1	A Typical Data Table with Anonymous Visits	13
3.2	Parts of the Likelihood for the Rates of Arrival Across Customers . .	18
6.1	Model Comparison in the Setting Where Missingness is Correlated with the Propensity to Visit in Response to a Marketing Action	36
6.2	Rank Ordering Customers in the Setting Where Missingness is Corre- lated with the Propensity to Visit in Response to a Marketing Action	40
6.3	Model Comparison in the Setting Where Missingness is Correlated With Effect of Marketing Action	42
6.4	Rank Ordering Customers in the Setting Where Missingness is Corre- lated with Effect of Marketing Action	45
6.5	Model Comparison in the Setting Where Missingness is Correlated With the Propensity to Undertake an Activity	47
6.6	Rank Ordering Customers in the Setting Where Missingness is Corre- lated with the Propensity to Undertake an Activity	51
7.1	Model Comparison in the Setting Where Missingness is Correlated with the Propensity to Visit in Response to a Marketing Action in the Real Data	58

7.2	Model Comparison of the Baseline Rate in the Setting Where Missingness is Correlated with the Propensity to Visit in Response to a Marketing Action in the Real Data	60
7.3	Rank Ordering Customers in the Setting Where Missingness is Correlated with the Propensity to Visit in Response to an Email	62
7.4	Model Comparison under the True Missingness Pattern in the Real Data	64
B.1	Parameter Recovery for a Representative Sample of Parameters in the Model	79
B.2	Parameter Estimates in the Full Model versus Model with Subsampling	80
B.3	Rank Ordering Customers in the Full Model versus Model with Subsampling	81

List of Figures

2.1	Percentages of Purchases from Each Category	8
2.2	Visitation Rates with Emails Versus without Emails	11
3.1	An Example of A Customer's Rate of Arrival	17
6.1	An Estimate of the Total Number of Unique Customers under Two Different Missingness Patterns	53

Chapter 1

Introduction and Motivation

An important aspect of marketing practice is the targeting of consumers for differential promotional activity [19] [5]. Recent advancements in digital marketing and loyalty card programs have expanded companies' ability to track customers, thus increasing the popularity of targeted marketing [14] [25]. However, despite the advancements in tracking technologies, companies still find that a large number of their interactions with their customers can not be matched to a particular customer and remain "anonymous." [9] [6] Marketers have long recognized this problem and have established generous incentive programs and other strategies to reduce anonymous visits [16]. For example, online retailers encourage customers to sign up for loyalty programs in order to receive special promotional emails [24] [8]. Yet, with few exceptions, companies consistently report that large proportions of visits cannot be tracked back to an existing customer.

There are numerous examples in everyday life where anonymous visits arise. A daily frequenter of a coffee shop might often pay with her credit card, allowing the shop to keep track of her purchases over time. This allows the coffee shop to send her tailored discounts and product offerings to the address associated with the card. However, some days she may prefer to pay with cash, resulting in a record of her

purchase that is not tied to her customer ID. Another case where anonymous visits may occur is when an online customer signs in with a unique user ID while frequenting a clothing retailer’s website. Although the customer has a user ID, some days she may browse the website while not logged on, which in turn will cause the clothing retailer to lose valuable information about this customer’s interests. Many companies ignore anonymous visits when analyzing customer visits and so information on customer preferences is seemingly lost.

We should point out that the systems for tracking users, and hence the potential for the prevalence of anonymous visits, varies across different situations. For example, in the web example, an identified customer can be tracked through cookies, through his IP address, or by clicking an ad in an email sent to him. In the coffee shop example, the identified customer can be tracked through the credit card number. The method we develop is agnostic about the tracking technology; so long as users make “visits” and during those visits engage in a number of activities, e.g. purchasing in certain categories, visiting certain pages, etc, this research can be applied.

When companies compile customers’ behavioral patterns over time to provide direct marketing, they do not typically attempt to link the anonymous visits to the other visits. But, there is a lot of potential information in anonymous visits; the data on anonymous visits still includes the time of visitation, as well as the activities that the unknown customer engaged in. We propose a Bayesian hierarchical model that aims to probabilistically assign anonymous visits to customers based on previous records of users’ behavioral patterns within company databases. This assignment is based on the time of the visit (relative to the timing of all customer’s observed visits) as well as the set of activities that the customer engages in during the visit (relative to the activities that all customers have engaged in).

Using our model, companies can better track the behavior of their customers,

allowing them to better target those customers during future identified or probabilistically identified visits. Our approach could, under some circumstances, even be used to target a customer during an anonymous visit, based on the probabilistic inference about “who they are.” Our methodology will allow us to deepen our knowledge of each customer by probabilistically assigning anonymous visits to customers, which increases the precision of targeted advertising not only to the unidentified customers, but to the identified ones as well. In addition, our model allows us to account for the anonymous visits when estimating overall features of a company’s customer base, and, as we will show, failing to account for anonymous visits can lead to erroneous inferences about critical business questions like, “how many unique users do I have?” by erroneously assigning the anonymous visits to either previously seen users or new ones.

To evaluate the ability of our model to recover the identity of unidentified users, we present a study with simulated data. We then demonstrate the performance of our model with real data from a large specialty retailer where the true visits are known, but we non-ignorably delete visits as a demonstration of how a firm can use our approach to provide improved direct marketing to customers.

As we will discuss, there is great potential that ignoring anonymous visits, as is common practice, not only reduces efficiency, but also may cause bias. For example, customers who tend to make anonymous visits may engage in different activities or be differentially affected by marketing. When the propensity to remain anonymous is correlated with the activities that may occur during the visit, the missing information is non-ignorable [12] for the inferential goals that companies might be interested in. For example, the company may be interested in knowing how effective their discount emails are in increasing their customer’s visitation rate to their website. Not using the anonymous visits can cause bias in their estimate of the effect of marketing,

potentially leading to the firm sending too many or too few emails. By accounting for the anonymous visits, we will show that our model will obtain more precise estimates of the effects of marketing actions on customers than commonly-used alternative approaches.

Our research can be considered an application of Bayesian missing data methods in marketing. In particular, our work is closely tied to extant studies which compare complete-case analysis (i.e., only keep the records with fully observed data) to incomplete case methods that impute values using Bayesian data augmentation (as done here) [23].

Previous works in marketing have used Bayesian data augmentation to handle partial information. Data augmentation has been used to handle the situation in which each of the datasets that the authors fuse together contains different demographic information [2]. It has also been used when the covariate information is only available in the aggregate [15], and to address the issue of having some outcomes observed at the individual level and others in the aggregate [3]. Unlike previous work, we will be using data augmentation to impute identification of unobserved customers based on their observed behavior and demographics.

The remainder of the dissertation is as follows. In Chapter 2, we present an exploratory data analysis of the clothing retailer’s data that motivated this thesis. We describe the data and show the effects of marketing on the customer’s rate of visitation.

In Chapter 3, Section 1, we first develop a general model for customer visits that can be applied in many contexts. This model provides the basis to lay out the likelihood for observing a particular anonymous visit. In Chapter 3, Section 2, we lay out extensions to the model to accommodate anonymous visits, which we treat as missing data in our Bayesian approach. In Chapter 3, Section 3, we show how one

can infer, from just data on visits, how many of the visits come from new, previously unobserved customers and we explain how to sample covariates for imputed customers that were not originally in the dataset. In Chapter 4, we describe two alternative approaches for handling missing data, and give examples of how they are used.

In Chapter 5, we summarize the behavior of modeling approaches under different hypothetical data scenarios. In Chapter 6, we demonstrate our ability to recover parameters and infer which user made anonymous visits using synthetic data. We evaluate our model’s performance as compared to two ‘competitor’ models under three missingness patterns; when there is a relationship between the propensity to be missing and the propensity to visit in response to marketing, when there is a relationship between the propensity to be missing and the effect of marketing on engaging in an activity, and when there is a relationship between the propensity to be missing and overall propensity to engage in an activity. As we will see, subtle factors such as the heterogeneity in the distribution of the propensity to be missing across individuals will affect the ability of our method to recover parameters.

In Chapter 7, we apply our methodology to a specialty retailer’s dataset. As the evidence suggests that the marketing’s effect is limited to the rate of arrival, we compare the methods in the setting where there is a relationship between missingness and the propensity to visit in response to an email (as in simulated section 6.3). We then apply our method and the alternative methods on a subsample of the complete dataset.

In Chapter 8, we conclude with a summary of findings and discuss future research directions. We discuss the issue of computational efficiency, and propose two ways of improving computational speed: subsampling and use of the expectation maximization (EM) algorithm.

Chapter 2

Exploratory Data Analysis

2.1 Description of the Data

Our modeling approach was motivated by a transactions dataset provided by a large specialty clothing retailer. In this dataset, each visit represents a purchase occasion and activities represent categories from which the customer purchased. There are 24,000 customers with known identification, and they are selected to be in the dataset because they made a purchase from the retailer within the two year period of recorded transactions. There is a median of 6 visits (purchases) per customer, with a minimum of 1 visit and a maximum of 270 visits. A visit may consist of purchases in any/all of the 21 categories. On average, customers purchase in 2.15 categories in each visit. Below we provide a table of the frequency of the number of visits taken by identified customers in order to give the reader a sense of the total number of transactions (times and purchase history) per customer.

Table 2.1: Frequency of Number of Visits Taken by Identified Customers

Number of Visits	Number of Total Customers
1	7190
2	3958
3	2953
4-5	3790
6-10	4820
11-20	2951
21-30	742
31-100	421
over 100	15

The purchase channel can be either direct or retail, and the categories from which customers can purchase are accessories, entertainment, holiday, home furnishings, home textiles, intimate apparel, jewelry, kitchen bar, leather goods, mens accessories, mens bottoms, mens knit tops, mens shoes, mens wovens, misc, home accessories, womens bottoms, womens knit tops, womens other, womens shoes, and womens woven tops.

Below we provide a figure of the percentage of purchases from each category out of the total transactions in the dataset.

Percentages of Trips with a Purchase in a Given Category

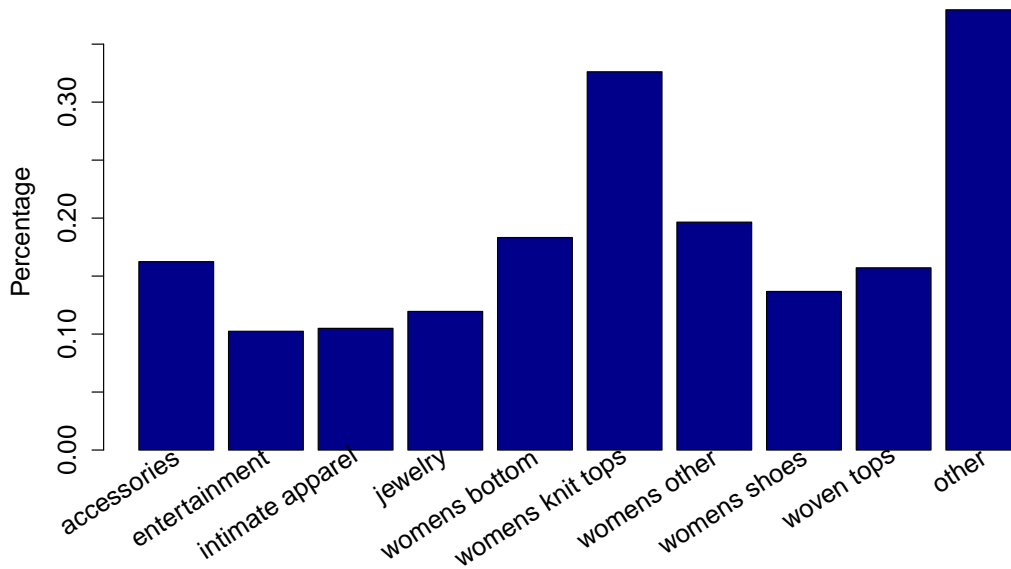


Figure 2.1: Percentages of Purchases from Each Category

A bar plot of the percentage of purchases coming from each category out of the 158,911 transactions in the dataset. We have combined the remaining 12 categories into one category called “other”.

We see that women’s knit tops is the most popular category. Out of the 158,911 total visits in the clothing retailer’s dataset, there were 51,826 visits that included a purchase from this category. In addition, there were 25,809 visits that included a purchase from the accessories category, and 16,271 visits that included a purchase from the entertainment category.

Customer characteristics include age, gender, whether the customer has a wishlist, and distance from nearest store to place of residence. Of the known customers, the mean age is 38 and the median age is 34 years old. 85% of the customers are women, and 15% are men. 20% of customers have a wishlist, and 80% do not.

Anonymous transactions exist in this clothing retailer’s dataset. There are a total of 24,000 customers with known identification numbers making anywhere from 1 to 50 visits, and 2,100 anonymous visits. In other words, if we assume that each anonymous

visit comes from a unique (or different) customer, there could be up to 9 percent of their customers that always remain anonymous when visiting.

2.2 Marketing Actions

The marketing actions in this application are emails sent to customers. This application contained a variety of different kinds of emails: “new arrivals” emails, promotional emails for specific categories, discount emails (either for particular categories and for the entire store), and new clothing promotional emails for each season. In addition, there were purchase confirmation receipt emails and return of item emails. Table 2.2 gives a sample of emails that customers received.

Table 2.2: Examples of Customer Emails

Email ID	Email Offer Name
208751	February 2012-02 Catalog
235421	Wednesday Free Shipping Ends
245651	New to Sale
285411	Must Have Shoes
270001	Women’s Holiday Preview

Since some of these emails should not be considered marketing actions, we only focused on discount emails in our application.

To estimate an effect of marketing action, we define an ‘email visit’ to be one that occurred within one week of receiving an email, and a ‘non-email visit’ to be one that occurred without receiving an email in the week prior to visitation. Previous marketing literature has shown that the effect of an email typically lasts approximately one week [26]. We focus our analysis on customers in the dataset who have at least

two ‘email’ and two ‘non-email’ visits.

Since there were many categories of purchase for each visit, we expect the marketing action to affect arrival rates, but not necessarily the propensity to purchase in one of the specific categories. As expected, the ‘sale’ emails did not have an effect on purchasing in one type of category, however they had a strong effect on visitation rate.

In Figure 2.2, we examine the empirical effect of email on arrival rate. In this figure, we plot the ratio of the rate of arrival in periods when emails are in effect versus the rate of arrival in periods when emails are not in effect for each customer in our subsampled dataset. We compute the rate of arrival in periods when emails are in effect to be the total number of ‘email visits’ out of the total number of periods in the dataset when emails are in effect. Similarly, the rate of arrivals when emails are not in effect is the total number of ‘non-email visits’ out of the total number of periods in the dataset when emails are not in effect. The horizontal line at Ratio=1 is the point at which the two rates are equal. The customers with points below the line arrive more frequently without emails, and the customers with points above the line arrive more frequently when they receive emails.

Most customers in the dataset arrive faster when an email is in effect than when it is not. Performing a binomial test with the null hypothesis being that the rates with and without emails are equal, we obtain a p-value < 0.001 that the customers arrive as frequently with emails as without them.

Ratio of the Rate of Visitation with Email versus without Email

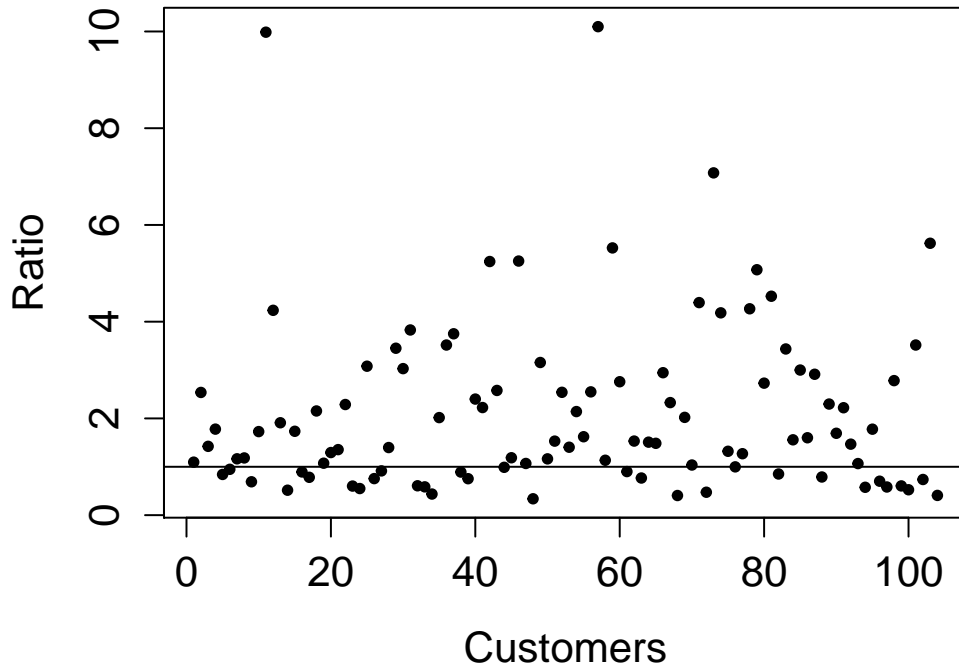


Figure 2.2: Visitation Rates with Emails Versus without Emails

For each customer, we plot the ratio of the number of visits when email was in effect divided by total time email was in effect to the number of visits when email was not in effect divided by the total time emails was not in effect. We add a horizontal line at Ratio=1 to indicate the point at which the rate of arrival when email is in effect is equal to the rate of arrival when email is not in effect. The customers with points below the line arrive more frequently without emails, and the customers with points above the line arrive more frequently when they receive emails.

As we see in this company's dataset, the marketing action they send to their customers affects the visitation rate of those customers. In addition, as we have shown in the previous section, this retailer has a lot of transactions made by unidentified clients. It would benefit the company to take advantage of this 'missing' data to obtain a more accurate understanding of the effects of their emails.

Chapter 3

Bayesian Hierarchical Model

3.1 General Model for Customer Visits

We begin by characterizing customer “visits” with a very general data structure like the one shown in Table 3.1.

Table 3.1: A Typical Data Table with Anonymous Visits

We provide an example of a typical incomplete data table below. When a customer is identified with a User ID, we have their time stamp, their ID number, whether or not they participated in the activities, and their covariate information. When a customer is not identified, we still have their time stamp and which activities they participated in; however, we no longer have their covariate information or their User ID.

Time j	User ID U_j	Shoes y_{j1}	Pants y_{j2}	Age Z_{j1}	Gender Z_{j2}
2010-01-01 12:46:49	16	1	0	34	0
2010-01-01 12:50:47	19	1	1	17	1
2010-01-01 13:20:54	3	0	0	19	0
2010-01-01 13:24:24	?	1	0	?	?
2010-01-01 13:25:00	27	0	1	45	1
2010-01-01 13:26:07	5	1	1	20	1
2010-01-01 14:10:09	16	1	0	34	0
2010-01-01 15:12:00	12	0	0	12	0

Let j index a set of observed customer visits, where there are n visits in total, so $j = 1, \dots, n$. At each visit we observe a set of discrete variables y_{j1}, \dots, y_{jM} indicating which activities the user engaged in during that visit, and $U_j \in \{1, \dots, I\}$, which indicates the user that made visit j where there are up to I potential unique visitors who could have visited the website.

Note that this is a very general data structure that could apply to users visiting websites, and engaging with certain features of the site, or video service subscribers watching certain movies during a session. In our retailer example, y_{j1}, \dots, y_{jM} denotes the categories from which the customer purchased, such as women's shoes, housewares, etc. where y_{jm} takes the values 0 or 1, indicating whether or not the customer purchased from category m . In other applications, y_{jm} could be ordinal counts or continuous, and in that case, we would substitute an appropriate link function.

We model the observed vector \mathbf{y}_j of indicators for the activities that the customer engaged in during the visit using a multivariate probit regression model [18] [1],

$$y_{jm} = \begin{cases} 1 & \text{if } y_{jm}^* > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (3.1)$$

where y_{jm}^* is customer U_j 's latent underlying utility to engage in activity m on visit j .

Using a multivariate hierarchical framework, we model $\mathbf{y}^*_j = (y_{j1}^*, \dots, y_{jM}^*)$ as

$$y_{jm}^* = \nu_{U_j, m} + \boldsymbol{\beta}_{U_j, m}^T \mathbf{X}_{jm} + e_{jm} \quad (3.2)$$

and

$$\mathbf{e}_j \sim N(0, \boldsymbol{\Sigma}) \quad (3.3)$$

where \mathbf{X}_{jm} are the visit-specific marketing actions for that visit across each of the M activities. More specifically, \mathbf{X}_{jm} is a length P_x vector (where P_x is the total number of

different marketing actions the firm can potentially take) for each of the M activities. For example, if there was a sale on shoes during visit j , \mathbf{X}_{jm} for the activity, which in our case is the purchase of shoes, would take the value 1 where there was a sale and 0 otherwise. The $\beta_{U_j,m}$ are the user U_j specific coefficients corresponding to activity m . Modeling y_{jm}^* in such a manner allows for a full correlation structure, Σ , among all the activities (as was done in [13]) to accommodate the possibility that some activities tend to occur together, e.g., purchasing women’s tops and women’s skirts.

The user specific coefficients consist of M individual level intercepts, $\nu_{U_j,m}$, which characterize individual U_j ’s overall propensity to engage in activity m , and $M \times P_x$ individual-level coefficients, β_{U_j} , which characterize each customer’s response to visit-specific marketing actions, $p = 1, \dots, P_x$. For example, the underlying propensity for user U_j to purchase shoes, without any form of enticement taken by the store, is $\nu_{U_j,\text{shoes}}$. If the store sends this user an advertisement, her underlying utility for purchasing these shoes would increase by $\beta_{U_j,\text{shoes,ad}}$.

Note that in Table 3.1 we also observe a time stamp for each visit. To model rate of visitation, we let a_{U_j,t_j} denote the waiting time between the $t_j - 1^{\text{th}}$ visit and the t_j^{th} visit by user U_j . While j indexes the visits among all the users in the dataset, t_j are the visits that correspond to a specific user U_j . We assume that the inter-arrival times follow a heterogeneous covariate-driven exponential distribution given by

$$a_{U_j,t_j} \sim \text{Exponential}(\lambda_{U_j,t_j}) \tag{3.4}$$

The arrival rate λ_{U_j,t_j} is comprised of two components: (i) $\omega_{U_j,0}$, a baseline arrival rate for each customer (independent of time), reflecting heterogeneity in visiting propensities and (ii) a person-specific covariate vector H_{U_j,t_j} reflecting marketing actions taken by the firm that may affect visitation rates.

$$\log \lambda_{U_j, t_j} = \omega_{U_j, 0} + \boldsymbol{\omega}_{U_j, 1} H_{U_j, t_j} \quad (3.5)$$

Note that H_{U_j, t_j} may be the same as \mathbf{X}_{jm} , if there is a marketing action that affects the rate and activities simultaneously. Let $\boldsymbol{\omega}_{U_j, 1}$ be an P_H -dimensional vector of regression coefficients corresponding to the P_H marketing actions, H_{U_j, t_j} . We now develop the rate of arrival part of the likelihood. For simplicity of exposition, we will focus on one marketing action ($P_H = 1$). User U_j visits the website at a constant underlying baseline rate $\omega_{U_j, 0}$. Upon receiving a marketing action, the customer's underlying visitation rate immediately changes to $\omega_{U_j, 0} + \omega_{U_j, 1} H_{U_j, t_j}$, and continues at this rate for a certain interval of time (during which this effect lasts). Once this time interval is over, the user's rate of visitation drops back to their baseline, $\omega_{U_j, 0}$. While one could choose to model the effect length differently (e.g., exponential decay), our straightforward approach is suitable for the purpose of accounting for the marketing action when we make our anonymous visit imputation.

In Figure 3.1, we give an example of a customer U_j 's arrivals and marketing action effects over a fixed period of time, T . For this example, we can construct the likelihood $L_{\lambda_{U_j}}$ for the sequence of arrivals, by taking a product over all the consecutive periods between the start of the dataset and time T .

$$\begin{aligned} L_{\lambda_{U_j}} &= (\omega_{U_j, 0} \exp[-\omega_{U_j, 0} t_0]) \times (1 - \exp[-\omega_{U_j, 0} t_1]) \times (1 - \exp[-(\omega_{U_j, 0} + \omega_{U_j, 1}) t_2]) \\ &\quad \times (\omega_{U_j, 0} \exp[-\omega_{U_j, 0} t_3]) \times (1 - \exp[-\omega_{U_j, 0} t_4]) \\ &\quad \times ([\omega_{U_j, 0} + \omega_{U_j, 1}] \exp[-[\omega_{U_j, 0} + \omega_{U_j, 1}] t_5]) \\ &\quad \times (1 - \exp[-(\omega_{U_j, 0} + \omega_{U_j, 1}) t_6]) \times (1 - \exp[-\omega_{U_j, 0} t_7]) \end{aligned} \quad (3.6)$$

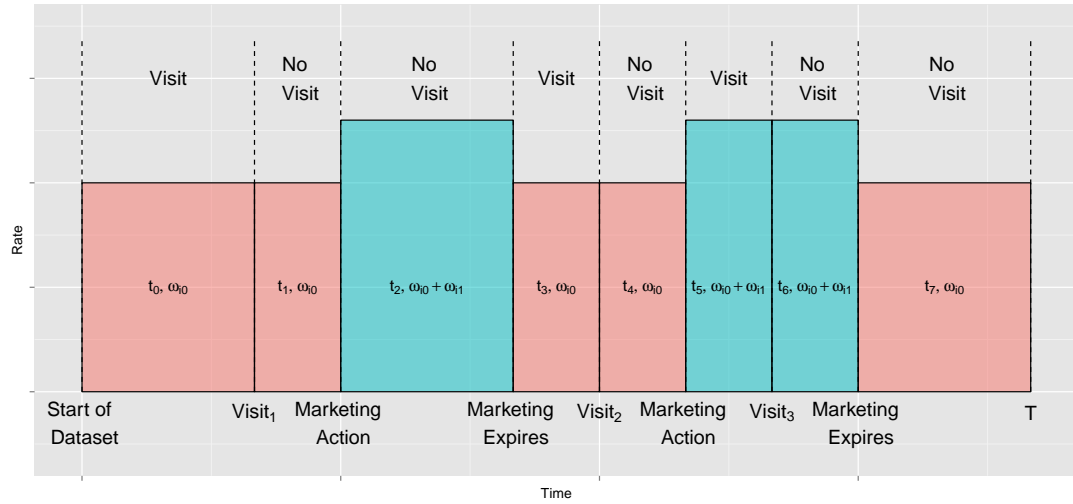


Figure 3.1: An Example of A Customer’s Rates of Arrival

We split each customer’s lifespan in the dataset into a series of periods. These periods can start and end with any of the following: a start of a marketing action, an end of a marketing action, and a visit. We take the product of the likelihood for all such events for each customer, and obtain the arrival likelihood.

To formulate the construction of the likelihood in the general case, we must segment time into intervals, considering all possible ‘start’ and ‘end’ events: a visit, a start to an effect of marketing action, and an end to an effect of marketing action, as shown in Table 3.2.

	End: MA	End: End of MA	End: Visit
Start: MA	DNE	$1 - \exp[-(\omega_{U_j,0} + \omega_{U_j,1})t_j]$	$(\omega_{U_j,0} + \omega_{U_j,1}) \times \exp[-(\omega_{U_j,0} + \omega_{U_j,1})t_j]$
Start: End of MA	$1 - \exp[-\omega_{U_j,0}t_j]$	DNE	$\omega_{U_j,0} \exp[-\omega_{U_j,0}t_j]$
Start: Visit	$1 - \exp[-\omega_{U_j,0}t_j]$	$1 - \exp[-(\omega_{U_j,0} + \omega_{U_j,1})t_j]$	$\omega_{U_j,0} \exp[-\omega_{U_j,0}t_j]$, if not within MA effect $(\omega_{U_j,0} + \omega_{U_j,1}) \times \exp[-(\omega_{U_j,0} + \omega_{U_j,1})t_j]$, if within MA effect

Table 3.2: Parts of the Likelihood for the Rates of Arrival Across Customers

In this table, MA is an abbreviation for a marketing action. Without receiving a marketing action, customer i has an underlying rate of arrival of $\omega_{i,0}$. However, upon receiving a marketing action, customer i 's rate increases to $\omega_{i,0} + \omega_{i,1}$ for a fixed length of time.

We now explain the two ‘‘DNE’’s in Table 3.2. Suppose an email has an effect that lasts one week, and suppose a customer received an email on Thursday and another one the following Monday. This customer will have an accelerated rate of arrival from Thursday until the Monday ten days later. For these overlapping marketing actions (that come prior to the end of the effect of the previous marketing actions), we remove all of the intermediate events except for the first marketing action. There may be a cumulative effect of receiving multiple emails, but for the purpose of accounting for the marketing action when we make our anonymous visit imputation, we choose to ignore these and assume a fixed effect. Likewise for overlapping marketing expiration events, we remove all intermediate events except for the last marketing expiration. That way, we eliminate the possibility of starting and ending with a marketing action, or starting and ending with an ‘‘end of marketing action’’.

With the likelihood for activities defined in equations 3.1-3.3, and the likelihood for arrivals defined in equations 3.4-3.6, we can now write the complete likelihood function for customer visits is given by equation 3.7. Note that arrival times are censored given that no arrivals are observed after a terminal time point T .

$$P(\mathbf{y}, \mathbf{A}, \mathbf{U} | \boldsymbol{\beta}, \boldsymbol{\nu}, \boldsymbol{\Sigma}, \mathbf{y}^*) = \prod_{j=1}^n \prod_{U_j=1}^I \left[\int_{G_{U_j, M}} \dots \int_{G_{U_j, 1}} \Phi_M \{ \mathbf{y}^*_j | \boldsymbol{\nu}_{U_j} + \boldsymbol{\beta}_{U_j}^T \mathbf{X}_j, \boldsymbol{\Sigma} \} d\mathbf{y}^*_j \right] L_{\lambda_{U_j}}]^{I(U_j)} \quad (3.7)$$

We now address the main question of interest: how to account for the anonymous visits in the model. $U_j = i$ represents the user ID for visit j which is known. When U_j is unknown, it can be any of the $i = 1, \dots, I$ potential unique users in the dataset. We define a missing data indicator $V_j = 1$ if the user for visit j is unknown, and 0 otherwise and let δ_i be the probability that user i will be anonymous, i.e. the probability that $V_j=1$ conditional on $U_j = i$. Our goal is to simultaneously estimate both the missing U_j and the parameters of the model using a Bayesian data augmentation approach [23].

In a Bayesian approach, we must first specify priors on the individual-level parameters $\boldsymbol{\theta}_i^T = (\boldsymbol{\nu}_i, \boldsymbol{\beta}_i, \omega_{i0}, \boldsymbol{\omega}_{i,1})$ as a function of both user-specific demographic covariates $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, \dots, Z_{iS})$ and population-level regression coefficients, $\boldsymbol{\Gamma}$, where S indicates the total number of user-specific covariates for each user i . For example, in a retail setting as described here, the user-specific demographic vector \mathbf{Z}_i could be that a customer is a female, her age is 27, and she does not have a loyalty card for the website, and $\boldsymbol{\Gamma}_{\text{age}, \boldsymbol{\nu}_{\text{shoes}}}$ would indicate the population-level baseline propensity to purchase shoes for a given age. Given this structure we model each customer's parameter vector, $\boldsymbol{\theta}_i$, with a hierarchical multivariate regression.

To allow for the possibility that these user specific parameters $\boldsymbol{\theta}_i$ (including missingness rate δ_i) could be correlated with each other, we specify a hierarchical

multivariate regression model as follows

$$\boldsymbol{\theta}_i = \begin{pmatrix} \boldsymbol{\nu}_i \\ \boldsymbol{\beta}_i \\ \omega_{i0} \\ \boldsymbol{w}_{i1} \\ \text{logit } \delta_i \end{pmatrix} \sim MVN(\boldsymbol{\Gamma}\boldsymbol{Z}_i, \boldsymbol{\Omega}) \quad (3.8)$$

Note that the dimension of the $\boldsymbol{\theta}_i$ vector is $M + P_x \times M + (1 + P_H) + 1$. This is because $\boldsymbol{\nu}_i$ is $M \times 1$, $\boldsymbol{\beta}_i$ is $P_x \times M$, $\boldsymbol{w}_{i,1}$ is $P_H \times 1$, and ω_{i0} and $\text{logit } \delta_i$ are scalars. \boldsymbol{Z}_i is an $S \times 1$ vector of user specific demographics (such as age and gender), $\boldsymbol{\Gamma}$ is the regression coefficient matrix that describes how these demographics relate to activity preferences, arrival rates, and marketing responses, and $\boldsymbol{\Omega}$ is the covariance matrix that characterizes heterogeneity across customers. $\boldsymbol{\Gamma}$ consists of $[M + P_x \times M + (1 + P_H) + 1] \times S$ regression coefficients, thereby allowing all S individual specific demographics to affect the value of the $\boldsymbol{\theta}_i$'s of that individual.

Returning to the issue of missing user IDs, let \boldsymbol{U}^{obs} be the subset of \boldsymbol{U} when $V_j = 0$, and let \boldsymbol{U}^{mis} be the subset of \boldsymbol{U} when $V_j = 1$. We will infer the \boldsymbol{U}^{mis} with a Bayesian approach where we estimate the joint posterior distribution of \boldsymbol{U}^{mis} simultaneously with the model parameters as given in equation 3.9.

$$\begin{aligned} P(\boldsymbol{\theta}, \boldsymbol{Z}, \boldsymbol{\Sigma}, \boldsymbol{U}^{mis} | \boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{U}^{obs}) &\propto \prod_{j=1}^n \prod_{i=1}^I \left[\int_{G_{U_j, M}} \dots \int_{G_{U_j, 1}} \Phi_M\{\boldsymbol{y}^*_j | \boldsymbol{\nu}_{U_j} + \boldsymbol{\beta}_{U_j}^T \boldsymbol{X}_j, \boldsymbol{\Sigma}\} d\boldsymbol{y}^*_j \right] \\ &\times L_{\lambda_{U_j}} \times \delta_{U_j}^{(V_j=0)} (1 - \delta_{U_j})^{(V_j=1)}]^{I_{(U_j=i)}} P(\boldsymbol{\theta}, \boldsymbol{Z}, \boldsymbol{\Sigma}) \end{aligned} \quad (3.9)$$

By accounting for anonymous visits, we will avoid potential bias in the estimate of

the effect of a marketing action on visitation, the estimate of the underlying propensity for customers to partake in specific activities, and in the estimate of the effect of a marketing action on the propensity to undertake an activity. In addition, it will allow us to make inference about the total number of customers in the dataset. Finally, the company can now provide more precise targeted marketing to both previously observed and unobserved customers, using a much richer knowledge about their preferences and potential interests.

3.2 Estimation via Markov Chain Monte Carlo Sampling

Our approach relies on Bayesian missing data methods [12]. The key idea is that any missing data (such as the anonymous visits in our application) can be treated in precisely the same fashion as model parameters. In particular, if \mathbf{U} is the “complete” set of \mathbf{U} ’s composed of observed and missing components, $\mathbf{U} = (\mathbf{U}^{mis}, \mathbf{U}^{obs})$, and $\boldsymbol{\gamma}$ are the remaining model parameters, then \mathbf{U}^{mis} can be integrated out of the posterior as follows [21]:

$$\begin{aligned} p(\boldsymbol{\gamma}|\mathbf{U}^{obs}) &\propto p(\mathbf{U}^{obs}|\boldsymbol{\gamma})p(\boldsymbol{\gamma}) \\ &= \int p(\mathbf{U}^{obs}, \mathbf{U}^{mis}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})d\mathbf{U}^{mis} \end{aligned} \quad (3.10)$$

In the Bayesian MCMC framework, this integration is accomplished by drawing the missing \mathbf{U} ’s at each iteration of the sampler conditional on the parameters [22]. We impose conjugate multivariate normal and inverse Wishart prior distributions on the global model parameters, $\boldsymbol{\Gamma}$, $\boldsymbol{\Omega}$, and $\boldsymbol{\Sigma}$ [4]. We use multiple Metropolis-Hastings

steps [11] when sampling $\boldsymbol{\theta}$, since this is sampled from a non-standard distribution. We first sample $(\boldsymbol{\nu}_i, \boldsymbol{\beta}_i)$ from a conjugate multivariate normal distribution, holding the remainder of the parameters fixed, and then we sample each of the ω and logit δ parameters separately using a Metropolis-Hastings algorithm, holding the remainder of the parameters fixed. To illustrate how inference is made about \mathbf{U}_j^{mis} , we go through the details of sampling a customer for an unassigned visit. Please see Appendix A: Gibbs Sampler for details for the remainder of our Gibbs Sampling algorithm. We will discuss sampling \mathbf{Z} , covariates, when missing, at the end of this section.

We sample a specific user for each missing U_j from a multinomial distribution where the probability of visit j being made by user i is:

$$P(U_j^{mis} = k | \mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\Sigma}, \mathbf{y}^*) = \quad (3.11)$$

$$\frac{\left(\int_{G_{kM}} \cdots \int_{G_{k1}} \Phi_M \{ \mathbf{y}^*_j | \boldsymbol{\nu}_{U_j, k} + \boldsymbol{\beta}_{U_j, k}^T \mathbf{X}_{jk}, \boldsymbol{\Sigma} \} d\mathbf{y}^*_j \right) L_{\lambda_{U_k}} \delta_k^{(V_j=1)} (1 - \delta_k)^{(V_j=0)}}{\sum_{i=1}^I \left(\int_{G_{iM}} \cdots \int_{G_{i1}} \Phi_M \{ \mathbf{y}^*_j | \boldsymbol{\nu}_{U_j, i} + \boldsymbol{\beta}_{U_j, i}^T \mathbf{X}_{ji}, \boldsymbol{\Sigma} \} d\mathbf{y}^*_j \right) L_{\lambda_{U_i}} \delta_i^{(V_j=1)} (1 - \delta_i)^{(V_j=0)}}$$

where $i = 1, \dots, I$ are the total potential users that could be assigned to an anonymous visit. In this way, we can draw the customer who has a high probability of making the anonymous visit based on the time of arrival, their demographic information, and the targeted advertisements they received.

Then, once \mathbf{U}_j^{mis} is sampled, the Gibbs sampler proceeds, sampling the other parameters from their full conditionals. The procedure continues iteratively sampling \mathbf{U}_j^{mis} alternately with the parameters. In this way, we simultaneously obtain draws from the posterior of \mathbf{U}_j^{mis} and the model parameters. Thus we can incorporate the anonymous visits in our model estimation in a way that utilizes all the information from both observed and anonymous visits.

We demonstrate how the imputation method works (compared to the nearest-neighbor and case-deletion methods) by going through a detailed example. Referring

back to the anonymous visit in the fourth row of Table 3.1, we are looking to impute which customer made that visit. We do not know the customer identification or the demographic information for the anonymous visit. However, we do know that the customer arrived at 2010-01-01 at 13:24:24, and that the customer purchased shoes but did not purchase pants. We see that customer 16 visited twice, and in both of their visits, they purchased shoes but did not purchase pants, as did the anonymous visitor. In addition, this customer visits more frequently than everyone else, making it even more likely that he/she was the anonymous visitor. The remainder of the customers had different behavior when on the website. Both customers 3 and 12 did not purchase in either category, customer 27 purchased pants but not shoes, whereas customers 19 and 5 purchased from both categories.

Our method estimates the probability that each customer made a particular anonymous visit. In this example, customer 16 would have the highest probability of making the anonymous visit according to equation 3.11. The sampler draws a new assignment at each iteration of the sampler resulting in a posterior distribution for the missing U_j .

In contrast to our method, the two alternative methods make a single assignment for the anonymous visit prior to any parameter estimation. The case-deletion method simply eliminates the entire anonymous visit from the dataset. The nearest-neighbor method assigns a previously observed customer to the anonymous visit by deterministically selecting the customer with the most similar observed activities. In this example, nearest-neighbor would assign customer 16 and not consider the possibility that a different user may have made the visit. This assignment will remain fixed, and the parameters would be estimated using standard approaches. To facilitate comparisons between these alternative approaches, in our example we estimate the proposed model using our same MCMC implementation but without the anonymous imputation step.

3.3 Allowing for Completely Unknown Users

To better gauge their presence in the global marketplace, firms need to be able to determine how many unique customers visit a firm’s store, be it digital or otherwise. Given the impact that targeted advertising can have on a firm’s bottom line, it is important for companies to distinguish between new and repeat customers in measuring customer lifetime value, churn rate, and company value.

We create R potential “new” customers, who can potentially be assigned to the R anonymous visits. We can then use the sampler to estimate the number of these “new” customers, who are assigned to a visit, as our estimate of the total number of unique customers. Let L be the true number of unique customers. This can be no more than the number of observed customers, Q , plus the total number of anonymous visits in the dataset, R .

In each iteration of the sampler, for every customer for whom there are no assigned visits, we will sample his parameter vector, θ_i , from the population prior distribution, which is updated given the current observed data (i.e. the θ_i ’s of all the customers who were assigned to a visit in the current iteration). If a customer is currently unassigned to any visit in the dataset, we do not have any observed data about him, and therefore we must sample θ_i from the prior $p(\theta_i)$.

For example, suppose that at iteration r , $d = 1, \dots, D$ out of the R total potential “new” users have not been assigned any visits. For each of those D users at this iteration, sample $\theta_d^T = (\nu_d, \beta_d, \omega_{d,0}, \omega_{d,1}, \text{logit}(\delta_d))$ from the prior.

It is important that the θ_d vector not be held fixed at the same values as in the previous iteration for two reasons. First, the parameters from which the individual level propensity vector is sampled change from iteration to iteration. This causes the global mean and variance structure to change. Second, by resampling θ_d , we

allow these users to be recycled. A customer for whom there are no assigned visits lacks the ‘right’ individual-level parameters to be assigned to any of the visits. By redrawing new users parameters from the prior, we may eventually create a customer that would be best suited for a visit, which in turn improves parameter estimation for the remainder of the model parameters. If an anonymous visit is assigned to a previously unassigned customer, then we update and include his parameter vector among the set of current users.

When assigning anonymous visits to previously unobserved users, we note that these unobserved users have missing demographic information that can actually be inferred from the customer’s behavior on the visit. That is, in the same way we probabilistically assign the user ID, we can probabilistically assign a demographic profile to new users. Estimating their user-specific characteristics may in some cases provide the company with a more accurate assessment of the demographics of their customer base, helping them optimize their assortment of products, target advertising, etc. At an iteration of the sampler, if one of the R total potential customers in the sampler is not assigned to any visit, we can estimate his or her covariate vector \mathbf{Z}_i . In fact, we must do that, as we will need to condition on \mathbf{Z}_i in the remaining steps of the sampler. We take advantage of the relationship $\boldsymbol{\theta}_i \sim N(\boldsymbol{\Gamma}\mathbf{Z}_i, \boldsymbol{\Omega})$ from our model. In the sampler, we drew $\boldsymbol{\Gamma}$ using a Bayesian regression, as a function of the parameters $\mathbf{Z}_i, \boldsymbol{\theta}_i$, and $\boldsymbol{\Omega}$.

Following the usual approach to missing regressors, when sampling \mathbf{Z}_i , we can use the usual draws for a multivariate regression by treating the matrix $\boldsymbol{\Gamma}$ as the regressors, and the \mathbf{Z}_i as the parameter vector, switching what we considered the covariates and regression coefficients. We sample \mathbf{Z}_i for each currently unassigned customer at iteration k as follows:

$$\mathbf{Z}_i | \Omega, \mathbf{U}, \Gamma, \boldsymbol{\theta}_i \sim MVN(\hat{\mathbf{Z}}_*, \mathbf{V}_{\mathbf{Z}_*}) \quad (3.12)$$

where $\hat{\mathbf{Z}}_* = (\Gamma^T \Omega^{-1} \Gamma + \mathbf{P}_0)^{-1} (\Gamma^T \Omega^{-1} \boldsymbol{\theta}_i + \mathbf{P}_0 \boldsymbol{\xi}_0)$ and $\mathbf{V}_{\mathbf{Z}_*} = (\Gamma^T \Omega^{-1} \Gamma + \mathbf{P}_0)^{-1}$ and where \mathbf{P}_0 and $\boldsymbol{\xi}_0$ are the prior parameters.

Chapter 4

Approaches to Missing Data

We will go through two alternative approaches that companies may use to handle missing customer IDs.

1. Case-Deletion
2. Nearest-Neighbor Matching

These alternative approaches can be classified into two types: complete-case analysis (approach 1) and imputation methods (approach 2). Imputation methods are ways of filling in missing variables. Approach 2 uses single imputation, which imputes one value for each missing variable. We lay out and explain each alternative modeling strategy in detail here. We also provide examples of uses for these strategies, and discuss their advantages and disadvantages. We will then use these methods as “benchmark” models in Chapters 6 and 7. They will be used for model comparison in both the simulation evaluation and real data analysis. After the model comparison, we discuss the settings under which it may be preferable to use one of these methods instead of our proposed approach.

4.1 Case-Deletion Method

The most straightforward method one can use to analyze missing data is complete-case analysis, which we refer to as ‘case-deletion.’ This method restricts the analysis to only the cases (or rows of data) in which there are no missing data, and deletes the rest. For the particular problem of anonymous visits, this means deleting the observations where the user ID is unknown.

The advantage of using complete-case analysis is that it is straightforward. We can use standard statistical methodology without any alterations to take the missing data into account. However, as we demonstrate in later chapters, a major disadvantage of deleting the missing visits can arise when missing data are non-ignorable. In other words, the missingness pattern may depend on the data in a way that affects inference [20]. This can lead to loss of precision and bias in parameter estimates, as the complete-cases may not be a representative sample of all possible data [12] [7]. Case-deletion would work well in the setting where the missing data add no additional information to the complete-cases. This is more likely to be the case when the proportion of cases that are missing is very small.

4.2 Nearest-Neighbor Matching

One imputation method that can be used to analyze data with anonymous visits is nearest-neighbor matching, which matches anonymous units to the closest non-anonymous unit based on observed variables [10]. In the case of anonymous visits, this means matching anonymous visits to known customers based on the activities undertaken in the anonymous visits and the activities that the known customer typically undertakes.

In order to match anonymous visits to the “closest” observed customers, one must

first define a distance metric, d . Let the M measures corresponding to unit j to be $y_{j,1}, \dots, y_{j,M}$. In our current application the measures will correspond to the categories that a customer purchased from. One example of a distance metric is the minimum deviation,

$$d(j, k) = \min_M |y_{j,M} - y_{k,M}| \quad (4.1)$$

Another example of a possible distance metric is the Mahalanobis distance,

$$d(j, k) = (y_j - y_k)^T S_{yy}^{-1} (y_j - y_k) \quad (4.2)$$

where S_{yy} is an estimate of the covariance matrix of y_j . Incorporating the covariance matrix means that categories which have high variability will carry less weight when finding potential candidates. For example, if an anonymous user has similar behavior to a known user in several categories but differs in one with high variance, that known user will still be a good candidate for a possible match.

In the analysis below, we will be using the Mahalanobis distance metric in our transactional data setting. In equation 4.2 above, y_j would be the set of M activities corresponding to visit j (an anonymous customer visit) in our transactional dataset. y_k would be the set of M activities corresponding to visit k (a known customer visit) in the dataset. We match all the anonymous visits to known customer visits with the most similar set of activities undertaken, where we define the most similar set of activities to be the ones with the smallest Mahalanobis distance between them.

More specifically, since known customers may have multiple observed visits, we select a customer for a match when their average Mahalanobis distance across all their visits is smallest. This would be the customer whose behavior is most consistent with the anonymous visit. For example, if a customer who is being considered for a match

has 10 observed visits, we would first compute the Mahalanobis distance between each of these observed visits and the anonymous visit. We would then take the average of these 10 distances, and use that average distance when considering this customer for the match.

We allow a known customer to be matched with as many anonymous visits as the algorithm chooses, and an anonymous visit can only be matched to a known customer. The method assigns matches deterministically.

Chapter 5

Summary for Behavior of Modeling Approaches under Different Data Settings

The goal of the simulations we present in Chapter 6 is to illustrate where the proposed method performs well relative to our benchmarks. We consider different settings, where we vary the total amount of missingness, the heterogeneity in distribution of missingness across customers, and the correlation between missingness and a parameter of the model that a company would be interested in. In this section, we first describe some general hypotheses about when the proposed method will work well.

When there is no relationship between the missingness process and the other model parameters (and the missing visits are ignorable), our method will correctly recover all the model parameters. Whether missingness is correlated with the effect of marketing on engaging in an activity, the propensity to engage in an activity, or with the effect of marketing on visitation, our method recovers this effect when the heterogeneity in the propensity to be missing is low. However, if in one of the three missingness settings

above the heterogeneity in the propensity to be missing is high, it will be more difficult for our method to recover the effect, and without enough individual-level data for certain people it will not be able to do so. Suppose that we are in the setting where the propensity to be missing is correlated with the effect of marketing on visitation and the heterogeneity in missingness is high. In this setting, suppose that customers whose rate increases upon receiving the marketing do not sign-in when they visit (due to the correlation). We would only obtain information about the customers who do not have an effect (since they are the only ones that sign-in). However, there would be a subset of the customer base for whom we would have barely any information, and these would be the customers with the high effect of the marketing action. Therefore, there would not be sufficient information in the visits with user IDs about the variation across the different types of customers. And so in this setting our method would underestimate the effect of marketing on visitation.

The case-deletion method will over or underestimate the effect of marketing action depending on the correlation structure between missingness and the parameter of interest. More specifically, if the correlation is positive, then case-deletion will underestimate the marketing effect (and vice versa). In the setting with a positive correlation, a larger proportion of anonymous visits come from customers who have a high effect of the marketing action, while more of the known visits come from customers with a smaller effect. Therefore, a larger proportion of data that is left comes from customers with a smaller effect, resulting in an underestimate of the effect of the marketing action.

The nearest-neighbor method will also over or underestimate the effect of marketing action depending on the distribution of marketing actions across customers. This method matches anonymous visits to previously identified customers based on their observed activities, not taking into account whether or not they received marketing

actions or the customer's typical arrival rate. So, this method will over or underestimate the effect based on the proportion of marketing action to non-marketing action visits that will result from the matching. For example, suppose five anonymous visits were matched to a known customer. During those five arrival weeks, the known customer had never obtained a marketing action. This would make it appear as though this customer visited more frequently during non-marketing action weeks, making the marketing action appear less effective.

In summary, there are several data scenarios under which our method would perform better than the other approaches in parameter recovery. Firstly, a correlation between missingness and the parameter we are interested in estimating would actually help our method recover the parameter, since this correlation would provide more distinguishing information about customers than in the no correlation setting. However, this correlation implies that the anonymous visits are not missing at random, which will negatively impact parameter recovery of the alternative methods. The second setting under which our method performs better is when there is low heterogeneity in the underlying distribution of missingness across individuals (and again a correlation). In the low heterogeneity setting, we would have partial data across a large proportion of the customer-base, providing us with information across all 'types' of customers (with high and low effects). Since there is still a correlation here, the missing visits would have a pattern, and so the alternative methods would again perform worse.

Chapter 6

Synthetic Data Evaluation

In the first set of synthetic data studies, we generate datasets in which missingness (δ_i) is correlated with the propensity to visit in response to a marketing action (ω_{i1}). For example, suppose a firm sends promotional emails to their customer base. The customers may click on the email, see that there is a huge sale going on now, and visit the website or store to make a purchase. In our first set of simulations, we assume that those customers who visit more frequently as a result of promotions are also the customers who tend not to sign in when visiting. We show that when there is a low heterogeneity in the distribution of missingness, our method does best in recovering the correlated parameter which governs the population-level propensity to visit in response to marketing.

In the second set of synthetic data studies, we focus on simulated datasets in which missingness (δ_i) is correlated with the effect of a marketing action on the propensity to engage in an activity (β_{im}). That is, the propensity for the customers to react positively to marketing by engaging in activity m is particularly high among those who tend not to sign in when they visit. We show that our method does best in recovering the correlated parameter which governs the population-level propensity to

engage in activity m in response to marketing.

In the third set of synthetic data studies, we generate datasets in which customers propensity to be missing (δ_i) is correlated with the underlying propensity to undertake an activity (ν_{im}). So, for example, customers who purchase pants more often might be more likely to remain anonymous. In this scenario, we will show that ignoring the anonymous visits leads to incorrect inference about how many customers are interested in the activity. If a firm is looking to stock a quantity of pants for the next month, for example, they may misgauge the population-level propensity to buy pants using all of the data they have appropriately.

6.1 Correlation between Missingness and Propensity to Visit in Response to Marketing

We begin by comparing models in the setting where missingness (δ_i) is correlated with the propensity to visit in response to receiving a firm's marketing (ω_{i1}).

We chose a large number of activities and one marketing action to stay consistent with the clothing retailer's dataset. The large number of activities also improves inference about which customers made the anonymous visit since customers can have a wide variety of preferred activities. We generated data with an average of 15 visits per customer, spanning a range of anywhere from 8 to 22 visits per customer. There are a total of 25 activities that a customer can undertake during a visit, and one marketing action that the firm can send to its customers. The times when the firms sends the marketing differs across customers. The frequency with which the firm sends it differs across customers as well. Customers visit on average every second week without marketing and four times per week with marketing.




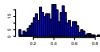
	Case 1 (baseline)	Case 2	Case 3	Case 4
Total Amount of Missingness	45%	45%	45%	30%
Heter. Missingness across Individuals	moderate	mod.	high	mod.
				
Correlation (missingness, visit) (δ_i, ω_{i1})	0.0	0.9	0.9	0.9
True value ($\mathbf{\Gamma}_{MP+M+2,1}$)	3.83	3.83	3.83	3.83
Visit Prop., our method ($\hat{\mathbf{\Gamma}}_{MP+M+2,1}$)	3.86 (3.55,4.12) 0%	3.67 (3.44,3.93) 4%	3.41 (3.19,3.65) 11%	3.73 (3.52,4.04) 2%
Visit Prop., case deletion ($\hat{\mathbf{\Gamma}}_{MP+M+2,1}$)	3.96 (3.69,4.23) 3%	3.51 (3.19,3.81) 8%	3.52 (3.18,3.82) 8%	3.84 (3.52,4.12) 0%
Visit Prop., nearest neighbor ($\hat{\mathbf{\Gamma}}_{MP+M+2,1}$)	3.75 (3.53,3.98) 2%	3.56 (3.30,3.82) 7%	3.37 (3.11,3.65) 12%	3.63 (3.42,3.84) 5%

Table 6.1: Model Comparison in the Setting Where Missingness is Correlated with the Propensity to Visit in Response to a Marketing Action

Recovery of the parameter which governs the population-level propensity to visit in response to a marketing action ($\mathbf{\Gamma}_{MP+M+2,1}$) in the setting where missingness (δ_i) is potentially correlated with the individual-level propensity to visit in response to a marketing action (ω_{i1}). Gray indicates that the true parameter was covered by the posterior interval. The last row in each cell indicates the percent bias.

6.1.1 Parameter Recovery

We present parameter recovery results across four data settings in Table 6.1. We consider Case 1 the “baseline case”, in which there is no correlation between the propensity to be missing and an effect of marketing on visitation. In addition, 45 percent of the visits are anonymous and there is moderate heterogeneity in the distribution of the propensity to be missing across individuals.

In Case 2, we increase the correlation between the propensity to be anonymous and the propensity to visit in response to marketing to 0.9, leaving the remainder of the settings the same as in Case 1. In a firm’s real dataset, the correlation may be less extreme, but the pattern of results are likely to be robust to large values of the correlation. In Case 3, we keep the correlation and total amount of missingness the same as in Case 2, but we change the heterogeneity of the distribution of the propensity to be anonymous from moderate to high. In Case 4, we keep the correlation and the distribution of the propensity to be anonymous the same as in Case 2; however, we reduce the total amount of missingness in the dataset to 30 percent.

We begin by presenting the results for our method in the row called “Visit Prop., our method”. The table shows estimates under each method of the population-level effect of marketing on visit propensity, which can be compared to the true value in the first row. Our method can recover the parameter under most settings. In Case 1, when there is no correlation between the propensity to be missing and the effect of marketing on visitation, our method obtains an unbiased estimate of the effect, 3.86 (versus the truth of 3.83). Our method obtains coverage of the remaining model parameters as well. In Cases 2 and 4, when there is a correlation and a moderate heterogeneity of missingness across individuals, our method still obtains coverage of the effect. It slightly underestimates it, with estimates of 3.67 and 3.73. However, in Case 3, when the heterogeneity in the distribution of missingness across individuals

is extreme, our method is no longer able to obtain coverage of the true value. In this case, the anonymous visits come from customers for whom marketing is highly effective and the known visits correspond to customers less affected by marketing. Our method is not able to recover the effect in this setting, as it does not have enough information on customers who visit more frequently in response to the marketing. It obtains an estimate of 3.41 in Case 3. This suggests that for firms who would want to use our method, it would make sense to do so if they believe that the heterogeneity in the distribution of missingness across their customers is moderate.

In the next row of Table 6.1, we present the recovery of population-level marketing effect ($\Gamma_{MP+M+2,1}$) for the case-deletion method. In Case 1, when there is no correlation between the effect of marketing and the propensity to be missing, the structure of the anonymous visits is the same as the structure of the identified visits. Ignoring the anonymous visits in this setting, as case-deletion does, does not have an effect on the parameter estimate. The case-deletion method is able to obtain coverage of the parameter, with an estimate of 3.96. However, when the synthetic data correlates missingness with the effect of marketing, case-deletion is no longer able to recover the true parameter value. In both Cases 2 and 3, the case-deletion method underestimates the effect, obtaining estimates of 3.51 and 3.52, respectively. In addition, the intervals are 27 and 39 percent wider here, respectively, than in our method due to the smaller amount of data used. In both of these cases, the visits with a high effect are anonymous, and case-deletion deletes them, resulting in underestimation of the effect. In Case 4, there is less total missingness, and in this setting case-deletion has enough information in the known visits to recover the parameter, with an unbiased estimate of 3.84. Even though there is a high correlation in Case 4, there are enough observed visits per customer for the case-deletion method to recover the effect.

In the final row of the table, we present the results for the nearest-neighbor method.

This method matches people based on observed activities, not taking into account marketing actions or rate of visitation. The method is unable to recover the effect when there is a large total percent missingness and a correlation between missingness and the propensity to visit in response to marketing. It underestimates the effect in both of these cases, obtaining estimates of 3.56 and 3.37 in Cases 2 and 3, respectively. Given the way that this particular dataset was generated, the anonymous visits that were matched to previously observed customers were assigned during times at which the marketing appeared to have less of an effect. In Case 1, the nearest-neighbor method is able to recover the effect, obtaining an estimate of 3.75. Since there is no correlation between the effect of marketing and the propensity to be missing, the structure of the anonymous visits is the same as the structure of the identified visits. In Case 4, there is less total missingness in the dataset, so the method's performance improves. However, it still underestimates the effect, obtaining an estimate of 3.63.

In summary, in the setting where the propensity to be missing is correlated with the effect of marketing on the propensity to visit, our method does especially well in the low heterogeneity setting, and obtains the best estimates of the marketing effect across all four data settings.

6.1.2 Targeted Marketing Results

Next we evaluate the models in terms of identifying the customers with the highest propensity to visit in response to marketing, using Cases 1 and 2 from Table 6.1.

Imagine that a company is looking to send targeted mail advertisements or promotions to the customers that will visit the fastest in response to advertisements (in other words, have the largest $\omega_{i0} + \omega_{i1}$). Both cases have 45 percent of the visits missing, and low heterogeneity in the distribution of missingness across customers. Case 1 has no correlation between the effect of marketing on the propensity to visit and the propensity to be missing, and Case 2 has a high correlation between the effect of marketing on the propensity to visit and the propensity to be missing.

Number of Customers Selected out of the Top 100	Case 1	Case 2
model	72	75
case deletion	40	46
nearest neighbor	60	36

Table 6.2: Rank Ordering Customers in the Setting Where Missingness is Correlated with the Propensity to Visit in Response to a Marketing Action

Each model produces its own rank ordering of customers in terms of highest propensity to visit in response to receiving a marketing action ($\omega_{i0} + \omega_{i1}$) in the setting where the change in propensity to visit in response to receiving a marketing action, ω_{i1} , is correlated with missingness (δ_i). For each of the methods, we first select the top 100 customers that would have the highest propensity to visit in response to marketing. We compared each model's selection to the true generated top 100 customers with the highest response to see how many are correctly chosen.

Each model produces its own rank ordering of customers in terms of strongest propensity to visit in response to a marketing action, when that propensity is correlated with missingness. For each of the methods, we select the top 100 customers (out of the 400) that would have the strongest propensity to visit. We compared each model's selection to the originally generated top 100 customers with the highest propensity to

see how many were correctly chosen. In Table 6.2, we report these results for the true top 100 of customers with the highest propensity to visit in response to a marketing action ($\omega_{i0} + \omega_{i1}$).

Our method performs best in both Cases 1 and 2, selecting 75 of the true top 100 customers in the correlated setting and 72 of the top 100 in the uncorrelated setting. The nearest-neighbor method performs less well, selecting 60 out of the true top 100 customers in the uncorrelated setting and 36 in the correlated setting. The case-deletion method also performs worse than our method, only choosing 40 of the true top 100 customers in the uncorrelated setting, and 46 in the correlated setting.

6.2 Correlation between Missingness and Effect of Marketing on the Propensity to Engage in an Activity

In the next simulation setting, we generate data where there is a correlation between missingness (δ_i) and effect of marketing on a customer's propensity to engage in activity m (β_{im}). In other words, suppose that this marketing action works best on people who prefer not to sign-in when they visit. The firm would like to gauge how effective this marketing action is; i.e. whether it is worth continuing with this type of advertisement.

We keep the same simulation settings as in the previous synthetic data section. As in Section 6.1, the times at which the firms sends the marketing differs across customers and the frequency with which the firm sends it differs across customers as well. Customers engage in activity m 20 percent of the time without marketing and




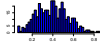
	Case 1 (baseline)	Case 2	Case 3	Case 4
Total Amount of Missingness	45%	45%	45%	30%
Heter. Missingness across Individuals	moderate	mod.	high	mod.
				
Correlation (missingness, marketing action) (δ_i, β_{im})	0	0.9	0.9	0.9
True value $(\mathbf{\Gamma}_{M+m,1})$	0.70	0.70	0.70	0.70
Marketing effect, our method $(\hat{\mathbf{\Gamma}}_{M+m,1})$	0.78 (0.64,0.97) 12%	0.75 (0.56,0.90) 7%	0.67 (0.51,0.84) 4%	0.75 (0.60,0.91) 7%
Marketing effect, case deletion $(\hat{\mathbf{\Gamma}}_{M+m,1})$	0.72 (0.51,0.92) 19%	0.72 (0.53,0.92) 3%	0.39 (0.20,0.59) 44%	0.72 (0.55,0.89) 3%
Marketing effect, nearest neighbor $(\hat{\mathbf{\Gamma}}_{M+m,1})$	-0.14 (-0.27,-0.02) 120%	0.31 (0.15,0.46) 56%	0.02 (-0.13,0.14) 97%	-0.34 (-0.50,-0.19) 149%

Table 6.3: Model Comparison in the Setting Where Missingness is Correlated With Effect of Marketing Action on an Activity

Recovery of the parameter which governs the population-level propensity to engage in activity m in response to a firm’s marketing $(\mathbf{\Gamma}_{M+m,1})$ in the setting where missingness (δ_i) is potentially correlated with the individual-level propensity to engage in activity m in response to marketing (β_{im}) . Gray indicates that the true parameter was covered by the posterior interval. The last row in each cell indicates the percent bias.

70 percent of the time with marketing on average.

6.2.1 Parameter Recovery

We consider four data settings in Table 6.3, which are structured in the same manner as in the previous section.

Table 6.3 shows how well the various methods recover the effect of marketing on a particular activity when missingness propensity (δ_i) is correlated with that same parameter. We begin by presenting the results for our method in the row called “marketing effect, our method.” Across all four data settings, our posterior intervals cover the parameter estimate. Our method is the only one that obtains coverage across all the four cases. Our method obtains estimates of 0.78, 0.75, 0.67, and 0.75 in cases 1 through 4, respectively (versus a true value of 0.70). Overall, our method performs well and is able to adjust for the missingness pattern regardless of the total amount of missingness and the distribution of the propensity to be anonymous across customers.

In the next row, we present recovery of the marketing parameter for the case-deletion method. Case-deletion can only recover the effect under certain settings. In Case 1, there is no correlation between the effect of marketing and the propensity to be missing, and so the structure of the missing data is no different from the structure of the known data. Ignoring the anonymous visits has no effect on the parameter estimates, resulting in a nearly unbiased estimate of the marketing effect of 0.72 for the case-deletion method. As in the uncorrelated case, in the correlated dataset case-deletion is able to perform well, giving an estimate of 0.72. Since the heterogeneity in the missingness parameter across individuals is low, we have enough known visits per customer to be able to estimate the effect accurately, even when ignoring the missing visits. In Case 3, case-deletion strongly underestimates the effect (0.39 versus a true value of 0.70). Here, the anonymous visits come from the customers who have a larger effect of marketing. The known visits correspond to customers without an

effect. Therefore, the only data that case-deletion uses to estimate the effect is that of the customers who don't actually have an effect, resulting in the underestimate. With a reduced amount of missingness in the final dataset (Case 4), the case-deletion method performs well (with an estimate of 0.72). There is enough information in the known visits for this method to obtain accurate estimates.

In the last row, we present the results for the nearest-neighbor method. Across all four data settings, this method heavily underestimates the effect of marketing. This occurs because the nearest-neighbor method matches visits based on the activities that the customers engaged in during the visit without taking the firm's marketing into account. For example, it will match an anonymous visit where the customer participated in activities y_m to a known visit where the customer participated in the same or similar activities, y_m . The method does not take into account that the customer in the anonymous visit may have engaged in those activities because of the firm's marketing, and would not have done so otherwise. Therefore, the nearest-neighbor method is not able to properly estimate the baseline propensity to engage in activities versus the propensity to engage in activities in response to marketing.

In summary, in the setting where the propensity to be missing is correlated with the effect of marketing on the propensity to engage in an activity, our method performs better than any of the competitor methods, as it obtains the best estimate of the effect across all four data settings.

6.2.2 Targeted Marketing Results

We again evaluate how well each method performs at selecting the individuals with the highest propensity to buy once they are sent marketing.

A firm typically wants to send mail advertisements or promotions to the customers with the highest propensity to buy once they are sent marketing ($\nu_{im} + \beta_{im}$). We use

the datasets generated in Cases 1 and 2 in Table 6.3 for this analysis. Both cases have 45 percent of the data missing, and moderate heterogeneity in the distribution of missingness across customers. The key difference is that Case 1 has 0 correlation between the effect of marketing and the propensity to be missing, while Case 2 has a correlation of 90 percent between the two.

Number of Customers Selected out of the Top 100	Case 1	Case 2
model	64	55
case deletion	27	27
nearest neighbor	26	26

Table 6.4: Rank Ordering Customers in the Setting Where Missingness is Correlated with Effect of Marketing Action

Each model produces its own rank ordering of customers in terms of strongest total reaction to the advertisement ($\nu_{im} + \beta_{im}$) in the setting where the change in reaction, β_{im} , is correlated with missingness (δ_i). For each of the four models, we first select the top 100 customers that would have the strongest total reaction (or the highest propensity to buy) after receiving the advertisement. We compared each model’s selection to the true generated top 100 customers with the highest response to see how many are correctly chosen.

As in the previous section, each model produces its own rank ordering of customers in terms of highest propensity to buy in response to marketing ($\nu_{im} + \beta_{im}$) when the change in propensity to buy in response to marketing (β_{im}) is correlated with missingness (δ_i). For each of the methods, we select the top 100 customers that would have the highest propensity to engage in activity m after receiving the advertisement. We compared each model’s selection to the originally generated “true” top 100 customers with the highest response to see how many were correctly chosen. In Table 6.4, we report these results for the top 100 customers with the highest propensity to buy in response to a marketing action ($\nu_{im} + \beta_{im}$).

In Table 6.4, in both Cases 1 and 2, our model consistently selects the highest number of the true top 100 customers to send the marketing actions. This implies that the firm would send its targeted advertising to the ‘best’ possible customers when

the firm would use our model (as opposed to the case-deletion and nearest-neighbor methods) in terms of finding the customers with the highest propensity to buy.

6.3 Correlation between Missingness and Overall Propensity to Engage in Activities

After having analyzed the implications of a correlation between the propensity to be missing and the propensity to engage in an activity in response to a marketing action, we proceed to compare the different methods in the setting where the propensity to be missing (δ_i) is correlated with the underlying propensity to engage in activity m (ν_{im}). In this setting, we focus on the propensity to engage in an activity when there is no marketing action sent to customer i . We keep the same simulation settings as in Section 6.2. Again, customers engage in activity m 20 percent of the time without marketing and 70 percent of the time with marketing on average.

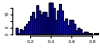
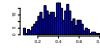
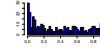
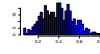
	Case 1 (baseline)	Case 2	Case 3	Case 4
Total Amount of Missingness	45%	45%	45%	30%
Heter. Missingness across Individuals	moderate	mod.	high	mod.
				
Correlation (missingness, undertake activity) (δ_i, ν_{im})	0.0	0.9	0.9	0.9
True value ($\mathbf{\Gamma}_{m,1}$)	-0.50	-0.50	-0.50	-0.50
Undertake Activity our method ($\hat{\mathbf{\Gamma}}_{m,1}$)	-0.39 (-0.51,-0.26) 22%	-0.40 (-0.51,-0.27) 20%	-0.43 (-0.56,-0.30) 14%	-0.44 (-0.54,-0.32) 12%
Undertake Activity case deletion ($\hat{\mathbf{\Gamma}}_{m,1}$)	-0.44 (-0.60,-0.30) 12%	-0.60 (-0.76,-0.44) 20%	-0.88 (-1.02,-0.70) 76%	-0.58 (-0.71,-0.40) 16%
Undertake Activity nearest neighbor ($\hat{\mathbf{\Gamma}}_{m,1}$)	-0.14 (-0.25,-0.02) 72%	-0.21 (-0.33,-0.10) 58%	-0.36 (-0.52,-0.21) 28%	-0.20 (-0.30,-0.10) 60%

Table 6.5: Model Comparison in the Setting Where Missingness is Correlated With the Propensity to Undertake an Activity

Recovery of the parameter which governs the population-level propensity to engage in activity m ($\mathbf{\Gamma}_{m,1}$) in the setting where missingness (δ_i) is potentially correlated with the individual-level underlying propensities to engage in activity m (ν_{im}). Gray indicates that the true parameter was covered by the posterior interval. The last row in each cell indicates the percent bias.

6.3.1 Parameter Recovery

Table 6.5 compares the ability of the three methods to recover the population-level propensity to engage in an activity under a scenario where missingness is correlated with propensity to undertake an activity. So, for example, these cases might represent the situation where people who like to buy shirts are also likely to be anonymous. We consider four data settings in Table 6.5 which are structured in the same manner as in the previous synthetic data sections.

We begin by presenting the results for our method in the row called “Undertake Activity, our method”. Across all four data settings, we obtain coverage of the parameter estimate. Our method is the only one that obtains coverage of the true parameter across all four cases. It obtains estimates of -0.39, -0.40, -0.43, and -0.44 in Cases 1 through 4, respectively (versus a true value of -0.50). Despite this slight positive bias in all the cases, our method performs well and is able to adjust for the missingness pattern regardless of the total amount of overall missingness and the distribution of the propensity to be anonymous across customers.

We present the case-deletion results in the next row of the table. Similar to the previous section where the propensity to be missing is correlated with the effect of marketing on engaging in an activity, in this setting, the case-deletion method can only recover the effect under certain settings. In Case 1, since there is no correlation between the propensity to engage in activity m and the propensity to be missing, the missing data has the same structure as the known data. Ignoring the anonymous visits has no effect on the parameter estimates, resulting in a low bias estimate of the propensity to engage in activity m , -0.44 (versus a true value of -0.50). In Case 2, we impose a high correlation between the propensity to engage in activity m and the propensity to be missing and the case-deletion method is able to recover the parameter in this setting. The heterogeneity in the distribution of missingness across customers

is moderate, so we have enough known visits for each customer to be able to estimate the population level effect accurately even without taking the anonymous visits into account. In Case 3, where we impose a high heterogeneity in the distribution of missingness, the case-deletion method is no longer able to obtain coverage of the parameter which governs the propensity to engage in activity m . In this case, the known visits correspond to customers who rarely engage in activity m , while the anonymous visits correspond to the customers who often engage in it. By ignoring the anonymous visits, the case-deletion method will heavily underestimate the propensity to engage in the activity, with an estimate of -0.88 (versus the truth of -0.50). In Case 4, we return to moderate heterogeneity in the distribution of missingness and decrease the total amount of missingness. In this setting, case-deletion is able to cover the truth, with an estimate of -0.58.

The nearest-neighbor method overestimates the baseline effect across all four data settings. Regardless of setting, this method matches customers on the observed behavior, without taking the marketing actions into account. By doing so, it underestimates the effect of the marketing actions, which in turn results in an overestimate of the baseline propensity to engage in the activity. It obtains overestimates of -0.14, -0.21, -0.36, and -0.20 in Cases 1 through 4, respectively versus a true value -0.50.

In summary, in the setting where the propensity to be missing is correlated with the propensity to engage in an activity, our method performs better than any of the four competitor methods, as it obtains the best estimate of the effect across all four data settings.

6.3.2 Targeted Marketing Results

We now evaluate how well our method would perform at selecting the individuals with the highest propensity to engage in activity m .

A firm is interested in understanding which customers prefer a specific category on their website without sending any advertisements. They may want to evaluate whether that category is popular amongst its customers, and if so, amongst which customers exactly so that they can target special offers to those customers. This may lead them to decide whether it is worth keeping that category. In other words, they want to select the customers with the highest propensity to engage in that activity.

We use the datasets generated in Cases 1 and 2 from Table 6.5. Both cases have 45 percent of the data missing, and low heterogeneity in the distribution of missingness across customers. The key difference is that Case 1 has 0 correlation between the propensity to engage in activity m and the propensity to be missing, while Case 2 has a correlation of 90 percent between the two.

Each model produces its own rank ordering of customers in terms of highest baseline propensity to engage in activity m , that is correlated with missingness. In a similar manner to what we did in Tables 6.2 and 6.4, in Table 6.6 we include the top 100 customers that would have the highest propensity to undertake a certain activity. We compare each model's selection to the originally generated "true" top 100 customers (or top 25 percent) with the highest propensity to engage in the activity to see how many were correctly chosen.

Number of Customers Selected out of the Top 100	Case 1	Case 2
model	59	57
case deletion	20	21
nearest neighbor	22	23

Table 6.6: Rank Ordering Customers in the Setting Where Missingness is Correlated with the Propensity to Undertake an Activity

Each method produces its own rank ordering of customers in terms of highest propensity to undertake a certain activity (ν_{im}) in the setting where the change in propensity to undertake activity m , ν_{im} , is correlated with missingness (δ_i). For each of the three methods, we select the top 100 customers that would have the highest propensity to undertake activity m . We compared each method’s selection to the true generated top 100 customers with the highest response to see how many are correctly chosen.

Once again, in both Cases 1 and 2, we consistently select the largest number of customers $i = 1, \dots, 100$ (or top 25 percent) with the highest propensities (ν_{im}) to undertake activity m , which is correlated with the propensity to be missing (δ_i). Our method selects 59 of the true top 100 customers in Case 1, and 57 in Case 2. The other two competitor methods perform significantly worse. The case-deletion method only selects the true top 20 in Case 1, and true top 21 in Case 2. The nearest-neighbor method only selects the true top 22 in Case 1, and true top 23 in Case 2.

6.4 Estimating the Number of Unique Customers

In order for a firm to gauge the size of their customer base, we provide estimates of the total number of unique customers that have visited using the various methods. In the simulated setting, we know both the number of observed customers and the true number of customers. We generated two datasets with 15 percent of the visit IDs missing. We induce a correlation between missingness and underlying propensity to undertake an activity of 0.9. In this dataset, 377 customers are observed, whereas there are 398 total true visitors.

The difference between the two synthetic datasets is the underlying distribution of missingness. In the first dataset, the heterogeneity in the distribution of the missingness parameter is high: we generate many people that never sign-in, and many people that always sign-in. In the second dataset, the heterogeneity in the distribution of the missingness parameter is low: most of the customers sign-in approximately half of the time that they visit, with very few that always or never sign-in.

To estimate the number of unique customers, at every iteration of the Gibbs sampler, we compute the total number of unique customers. We provide a histogram of the number of times we obtained each total number of unique customers, along with lines for the number of observed and the number of true customers.

The histogram for data setting 1, where the heterogeneity in missingness is high, is shown on the left side of Figure 6.1. Under this setting of high heterogeneity, we obtain the most accurate estimate out of the competing models we considered, with the nearest-neighbor and case-deletion methods by construction estimating the number of unique users to be the number of observed customers (377).

This histogram for data setting 2, where the heterogeneity in missingness is low, is shown on the right side of Figure 6.1. With lower heterogeneity in missingness, we obtain coverage of the true total number of customers. Just like in the previous case, the nearest-neighbor and case-deletion methods underestimate the true number. Depending on the distribution of missingness, we are able to estimate the total number of unique customers in the dataset with more or less accuracy, and we always obtain the best estimate across the three methods.

We obtain a better estimate of the number of unique customers in the low heterogeneity setting. When there is a low heterogeneity in the distribution of missingness across customers, all the customers sign-in at a similar rate when they visit. Therefore with a high probability, they would all sign-in at least once when they visit. In the

high heterogeneity setting, there is a group of customers who are likely to always remain anonymous (since they have a very low probability of signing-in when they visit). This implies that in the low heterogeneity setting, we observe a larger fraction of the entire customer base, and so the number of observed customers is closer to the number of unique customers (than in the high heterogeneity setting). By construction, this makes it easier for the methods to estimate the number of unique customers in the low heterogeneity setting.

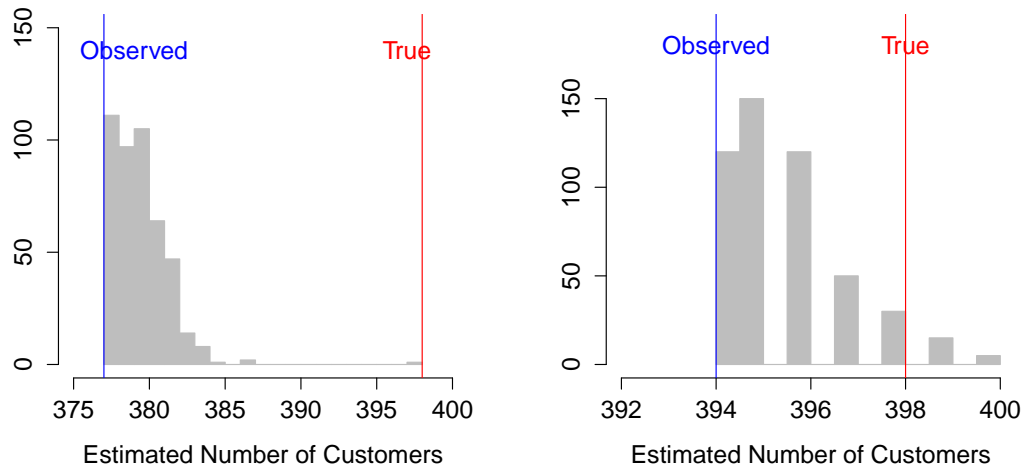


Figure 6.1: An Estimate of the Total Number of Unique Customers under Two Different Missingness Patterns

Total number of unique customers when the underlying heterogeneity in missingness is high (left) and low (right). Posterior samples of the number of customers from our model for a single dataset along with the true number of customers. The number of observed customers, which is the estimate for case-deletion and nearest-neighbor, is indicated by the red line.

6.5 Summary of Synthetic Data Evaluations

For a company, the above simulation studies demonstrate the risk of not linking the anonymous visits to previously observed or unobserved customers. In almost all

synthetic data settings, a company would obtain the most accurate estimates of the effects of marketing actions on their customers by using our model.

More specifically, if the correlation structure is between missingness and the propensity to visit in response to a marketing action, our model does best in the low heterogeneity setting.

If there is a correlation between missingness and the effect of a marketing action on the propensity to engage in an activity, our method obtains the most accurate estimates across all the synthetic data scenarios we tested. For example, suppose the company wanted to understand the effects of email advertising on their customer base. Without linking the anonymous visits to customers, they would obtain less accurate estimates of their effects, which may cause them to stop using that method of marketing, whereas in fact it could be effective. If they are interested in targeting the right customers, our model also provides the most accurate targeting choices.

Finally, if the correlation structure is between the propensity to undertake an activity and missingness, our model is the only one to obtain coverage of the correlated propensity parameter under all data settings. By not using our model, the company may risk mis-estimating whether it is worth keeping that activity, or misallocate how much of it to stock up on for the next season.

Understanding the size of the customer base is important as well. We demonstrated that our model provides the firm with an accurate estimate of the magnitude of their customer base. This estimate was better than that of the case-deletion, nearest-neighbor, or either of the unique-customer methods.

We cannot observe the true underlying correlation structure with missingness or the heterogeneity in missingness. However, through these simulation studies, we demonstrate that in most of the settings, our model obtained the most accurate estimates.

Chapter 7

Application to a Retail Website

In this chapter, we use a dataset from a clothing retailer’s website. A visit will be a transaction from the retailer’s website, and an activity will be a particular clothing category from which the transaction was made. The single marketing action in this data is a discount email. Following previous research, we assume that the effect of the discount email lasts for a week [26].

In the simulation studies in the previous chapter, since we generate the underlying data, we know the “true” parameter values (which we set). We use these “true” values to see how well the methods perform under different missing data settings at recovering this “truth”. However, in the real clothing retailer’s dataset, we do not know the underlying truth.

To remedy this problem, we consider as the complete data the subset of the data where the customer identification is known. We run the hierarchical model with no missing data, to obtain what we consider the “true” parameter estimates. After running the hierarchical model with no missing data, we obtain an effect of discount emails on the propensity to visit (or in our setting, to make a transaction). As we did in Section 6.1, we induce a correlation between the propensity to be missing and

the propensity to visit in response to the discount email by anonymizing user IDs accordingly.

In Section 7.2, we select a random subsample from the entire dataset (with the true missingness pattern), and compare the results across the methods.

7.1 Relationship between Missingness and Propensity to Visit in Response to a Marketing Action

After running the hierarchical model with no missing data, we find one strictly positive effect in the full data setting. Consistent with the EDA in Section 3.2, the discount email has a population-level effect on the propensity to make a transaction.

To follow the structure of our simulated results, we create four types of missing data settings. In the first data setting (Case 1), we impose no correlation between missingness and the propensity to make a transaction in response to the discount email. The total amount of missingness in the dataset is 45 percent, and the heterogeneity of missingness across individuals is low. In the second data setting (Case 2), we impose a high correlation of 0.9 between missingness and the effect of the discount email on making a transaction. The total amount of missingness across the dataset is 45 percent, and the heterogeneity of missingness across individuals is low. In the third data setting (Case 3), we change the heterogeneity of missingness across individuals to be high. We leave the other two settings the same as in Case 2. The correlation is 0.9 between the propensity to be missing and the effect of the discount email on making a transaction, and a total amount of missingness is 45 percent. In the fourth data setting (Case 4), we reduce the total missingness to 30 percent. We leave the

other two settings the same as in Case 2. We continue to have a correlation of 0.9 between the propensity to be missing and the effect of the discount email on making a transaction, and a low heterogeneity of missingness across individuals.

We then run the different methods using our induced missingness pattern across all four data structures, and compare the parameter estimates to those of the model estimated on the data with no missingness. In Table 7.1, we provide estimates of the population-level effect of discount emails and posterior intervals across all the methods and settings. We also provide the percent bias in each case. The ‘true’ population-level effect of the discount email on the propensity to visit is 2.16, with a posterior interval of [2.06,2.33]. The ‘true’ population-level baseline rate of arrival (without having received a discount email) is 0.62, with a posterior interval of [0.56,0.68]. This means that on average, a customer visits approximately once every second week, whereas upon receiving an email, the customer will visit approximately twice in one week.




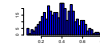
	Case 1 (baseline)	Case 2	Case 3	Case 4
Total Amount of Missingness	45%	45%	45%	30%
Heter. Missingness across Individuals	moderate	mod.	high	mod.
				
Correlation (missingness, visit) $(\delta_i, \omega_{i,1})$	0	0.9	0.9	0.9
True value $(\mathbf{\Gamma}_{MP+M+2,1})$	2.16 (2.06,2.33)	2.16 (2.06,2.33)	2.16 (2.06,2.33)	2.16 (2.06,2.33)
Visit Prop., our method $(\hat{\mathbf{\Gamma}}_{MP+M+2,1})$	2.05 (1.79,2.51)	2.18 (1.99,2.37)	1.65 (1.52,1.76)	2.19 (1.95,2.45)
	5%	0%	23 %	1%
Visit Prop., case deletion $(\hat{\mathbf{\Gamma}}_{MP+M+2,1})$	1.97 (1.87,2.06)	1.83 (1.69,1.98)	1.84 (1.65,2.05)	2.17 (1.83,2.47)
	9%	15%	15%	0%
Visit Prop., nearest neighbor $(\hat{\mathbf{\Gamma}}_{MP+M+2,1})$	2.15 (1.95,2.35)	1.97 (1.86,2.10)	1.86 (1.73,1.99)	2.05 (1.90,2.19)
	0%	9%	14%	5%

Table 7.1: Model Comparison in the Setting Where Missingness is Correlated with the Propensity to Visit in Response to a Marketing Action in the Real Data

Recovery of the parameter which governs the population-level propensity to visit in response to a marketing action $(\mathbf{\Gamma}_{MP+M+2,1})$ in the setting where missingness (δ_i) is potentially correlated with the individual-level propensity to visit in response to a marketing action $(\omega_{i,1})$ in the real dataset for all three methods. Gray indicates that the true parameter was covered by the posterior interval. The last row in each cell indicates the percent bias.

7.1.1 Parameter Recovery

We begin by presenting the results for our method in the row called “Visit Prop., our method”. Our method behaves the same as it did in the simulated setting where the propensity to be missing is correlated with the propensity to visit in response to marketing. It is able to recover the effect parameter under the low heterogeneity settings. In Case 1, when there is no correlation between the propensity to be missing and the effect of a discount email on visitation, our method obtains an unbiased estimate of the effect, 2.05 (versus the truth of 2.16). In Cases 2 and 4, when there is a correlation and a moderate heterogeneity of missingness across individuals, our method still covers the true parameter, though with slight overestimation. In Case 3, when the heterogeneity in the distribution of missingness is high, our method is no longer able to obtain recover the true effect. The anonymous visits come from customers who truly have an effect of the discount email on making a transaction. The known visits correspond to customers who have a smaller effect. Our method is not able to recover the effect in this setting, as it does not have enough information on customers who visit faster in response to the marketing.

In the next row of Table 7.1, we present the recovery of the discount email effect for the case-deletion method. Overall, the method is not able to recover the effect in any setting where there is 45 percent of the data missing. It is only able to recover it when there is 30 percent of the data missing in Case 4. It obtains estimates of 1.97, 1.83, and 1.84 in Cases 1 through 3, respectively. In Case 4, the case-deletion method obtains coverage of the truth, with an unbiased estimate of 2.17.

In the final row of the dataset, we present results for the nearest-neighbor method. This method matches people based on observed behavior, not taking into account the discount emails. Similarly to the synthetic data results, the nearest-neighbor method is unable to recover the effect when there is high total missingness and a correlation

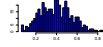
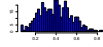
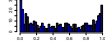
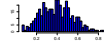
	Case 1 (baseline)	Case 2	Case 3	Case 4
Total Amount of Missingness	45%	45%	45%	30%
Heter. Missingness across Individuals	moderate	mod.	high	mod.
				
Correlation (missingness, visit) $(\delta_i, \omega_{i,1})$	0	0.9	0.9	0.9
True value $(\mathbf{\Gamma}_{MP+M+1,1})$	0.62 (0.56,0.68)	0.62 (0.56,0.68)	0.62 (0.56,0.68)	0.62 (0.56,0.68)
Base Visit Prop, our model $(\hat{\mathbf{\Gamma}}_{MP+M+1,1})$	0.61 (0.53,0.69)	0.69 (0.62,0.75)	0.69 (0.61,0.79)	0.66 (0.58,0.74)
Base Visit Prop, case deletion $(\hat{\mathbf{\Gamma}}_{MP+M+1,1})$	0.53 (0.46,0.60)	0.51 (0.44,0.57)	0.49 (0.43,0.55)	0.57 (0.51,0.64)
Base Visit Prop, nearest neighbor $(\hat{\mathbf{\Gamma}}_{MP+M+1,1})$	0.56 (0.48,0.63)	0.58 (0.50,0.66)	0.59 (0.52,0.66)	0.58 (0.52,0.65)
	10%	6%	5%	6%

Table 7.2: Model Comparison of the Baseline Rate in the Setting Where Missingness is Correlated with the Propensity to Visit in Response to a Marketing Action in the Real Data Recovery of the parameter which governs the population-level baseline rate of arrival $(\mathbf{\Gamma}_{MP+M+1,1})$ in the setting where missingness (δ_i) is potentially correlated with the individual-level propensity to visit in response to a marketing action $(\omega_{i,1})$ in the real dataset for all three methods. Gray indicates that the true parameter was covered by the posterior interval and no highlight that it was not. The last row in each cell indicates the percent bias.

between missingness and the propensity to visit in response to marketing. It obtains estimates of 1.97 and 1.86 in Cases 2 and 3, respectively. When there is less total missingness in Case 4, it is able to recover the effect. It still underestimates it, with an estimate of 2.05. When there is no correlation in Case 1, the nearest-neighbor method obtains an unbiased estimate of the effect, 2.15 (versus a truth of 2.16).

We now present results from Table 7.2. These results come from the same datasets as in Table 7.1, however here we present recovery of the population-level baseline rate parameter.

We begin by presenting the results for our method in the row called “Base Visit Prop., our method”. We see that in all four data settings, our method is able to recover the parameter that governs the propensity to visit without receiving a discount email. It obtains estimates of 0.61, 0.69, 0.69, and 0.66 (versus a truth of 0.62).

Next, we see that case-deletion underestimates the visitation rate without emails, as expected, since it deletes a large number of anonymous visits. When the total amount of missingness in the dataset is 45 percent, it heavily underestimates the visitation rate without discount emails. It obtains estimates of 0.53, 0.51, and 0.49 in Cases 1 through 3, respectively. However when the total missingness decreases to 30 percent in Case 4, the estimate increases to 0.57 and the method obtains coverage of the truth, since there appear to be more frequent visits per customer now.

In the last row of the table, similar to our method, we see that the nearest-neighbor method is also able to obtain coverage of the parameter that governs the propensity to visit without receiving discount emails. It obtains estimates of 0.56, 0.58, 0.59, and 0.58 in Cases 1 through 4, respectively.

In summary, in the setting where the propensity to be missing is correlated with the effect of marketing on the propensity to visit, we recommend using our method, as it obtained the least biased results in the setting where there is low heterogeneity in missingness and a high correlation between the propensity to be missing and the propensity to visit in response to receiving a discount email.

7.1.2 Targeted Marketing Results

Next we want to consider how well our method would perform at selecting the individuals with the highest propensity to visit in response to marketing.

We used the datasets generated in Cases 1 and 2 from Table 7.1. Both cases have 45 percent of the data missing, and low heterogeneity in the distribution of missingness

across customers. Case 1 has no correlation between the effect of marketing on the propensity to visit and the propensity to be missing, and Case 2 has a high correlation between the effect of marketing on the propensity to visit and the propensity to be missing.

Each model produces its own rank ordering of customers in terms of strongest propensity to visit in response to a marketing action, when that propensity is correlated with missingness. For each of the methods, we select the top 25 customers (out of the 100 true customers) that would have the strongest propensity to visit. We compared each model’s selection to the originally generated top 25 customers with the highest propensity to see how many were correctly chosen.

Number of Customers Selected out of the Top 25	Case 1	Case 2
model	15	18
case deletion	12	2
nearest neighbor	4	5

Table 7.3: Rank Ordering Customers in the Setting Where Missingness is Correlated with the Propensity to Visit in Response to an Email

Each model produces its own rank ordering of customers in terms of highest propensity to visit in response to an email ($\omega_{i,0} + \omega_{i,1}$) in the four settings in Table 7.1. We report the results for Case 1 (when the change in reaction, $\omega_{i,1}$ is independent of missingness, δ_i) and Case 2 (where the change in reaction, $\omega_{i,1}$, is correlated with missingness, δ_i). For each of the four models, we first select the top 25 customers that would have the strongest total reaction (or the highest propensity to buy) after receiving the advertisement. We compared each model’s selection to the true top 25 customers with the highest response to see how many were correctly chosen.

In Table 7.3, we report these results for the top 25 customers (or top 25 percent of the dataset). In other words, these are the 25 customers with the highest propensity to visit in response to a marketing action ($\omega_{i0} + \omega_{i1}$).

Our method and the nearest-neighbor method perform equally well in the uncorrelated setting, while our method performs best at selecting the top customers in the

the correlated setting. In Case 1, our method selects 15 out of the top 25 customers, and in Case 2, it selects 18 of the top customers. The case-deletion method performs the worst in both cases, only choosing 12 of the top customers in Case 1 and 2 of the top customers in Case 2. The nearest-neighbor method only selects 4 out of the top 25 customers in the uncorrelated setting, and 5 customers in the correlated setting.

In summary, if a firm is looking to selected the customers who would visit the fastest in response to receiving a discount email, when there is no correlation between propensity to be missing and the propensity to visit in response to the discount email, there is no one clear method that would perform best. However, if there is a correlation, our method consistently selects the highest number of the most responsive customers.

7.2 Application to a Retail Dataset with its True Missingness Pattern

After demonstrating what happens with our imposed missingness patterns in Section 7.1, we now run the different methods on a random subset of the entire dataset. In the previous section, we only subsampled customers with known user identification, and then ‘pretended’ that some of the user ID’s were missing. In this section, by randomly selecting customers with or without known user identification, we obtain a subsampled dataset with a more reflective pattern of missingness.

As is the case in the full dataset, in our subsampled dataset, 10 percent of the visits are anonymous. We use the same marketing action as in the previous section, whether or not the customer obtained a discount email in the week prior to purchase. To stay consistent with the previous section, we summarize the results for the same parameter as before, the effect of a discount email on the propensity to make a transaction.

Our method obtains an estimate of 1.36, with a posterior interval of [1.11,1.66].

The case-deletion method obtains an estimate of 2.41, with a posterior interval of [2.13,2.70]. The nearest-neighbor method obtains an estimate of 2.37, with a posterior interval of [2.12,2.63].

In this dataset, we see that our method obtains a lower parameter estimate than the case-deletion and the nearest-neighbor method. In Section 7.1, we saw a similar pattern when the heterogeneity in the distribution of missingness across individuals was high. In that case, the case-deletion and nearest-neighbor method both obtained similar and higher estimates than our method. When the heterogeneity in the distribution of missingness was low, our method typically obtained a higher estimate. Also, in the previous section, we induced a positive correlation between the effect of email on the propensity to visit and the propensity to be missing.

In this dataset, there is a positive correlation between δ_i and $\omega_{i,1}$, but the heterogeneity in the distribution of missingness is high.

Effect of Email On Rate of Arrival ($\hat{\Gamma}_{M+MP+2,1}$)	
Marketing effect, our method	1.36 (1.11,1.66)
Marketing effect, case deletion	2.41 (2.13,2.70)
Marketing effect, nearest neighbor	2.37 (2.12,2.63)

Table 7.4: Model Comparison under the True Missingness Pattern in the Real Data

We run the model with the true anonymous and known visitors, and compare results for the effect of discount emails on the propensity to make a transaction.

Chapter 8

Conclusion and Future Work

We proposed a Bayesian hierarchical model to link anonymous visits in a company's database to either previously observed or new customers. The model incorporates more complete information about both the anonymous and known customers. This information includes customers' times of arrival, whether the customers signed in or not, the activities the customers had engaged in while visiting, their demographic information, and information on marketing actions taken by the firm. Our model probabilistically imputes particular customers into the anonymous visits based on similarity of observed behavior. We then compare our proposed model to several benchmark methods to determine under which circumstances our model performs best. The implications of the model can then be used to guide the companies' marketing actions.

By conducting several synthetic data studies, we are able to evaluate model performance under three different missingness structures: missingness correlated with the propensity to visit in response to a marketing action, missingness correlated with the effect of a marketing action on the propensity to engage in a particular activity, and missingness correlated with the propensity to undertake an activity. Under each

of these settings, we vary the percentage of missingness, the underlying distribution of missingness, and the underlying correlation structure between missingness and the parameter of interest. We demonstrate the myriad of consequences that can occur for a company opting to not link anonymous visits to customers, i.e., by not using a model such as ours. Should a firm choose our model, it can be used to provide the following information:

- (1) Improved performance in estimating the effects of any marketing action on the propensity of the firm's customers to partake in a particular activity relative to the benchmark methods;
- (2) Improved performance in estimating the firm's customers' underlying propensity to partake in an activity relative to the benchmark methods;
- (3) Improved performance in estimating the effects of any marketing action on the propensity of the firm's customer's to visit relative to the benchmark methods;
- (4) Improved targeted advertising to individuals in any correlation setting;
- (5) Improved estimates of the number of unique customers;

When estimating the effect of a marketing action on the propensity to visit, our method performs best when there is moderate heterogeneity of missingness across individuals. More specifically, under the moderate heterogeneity setting, our method obtains the most accurate estimates.

When estimating the effect of a marketing action on the propensity to engage in an activity, our model consistently provides the most accurate estimates of the effect in our simulated dataset across all four cases. The case-deletion method performs well in the low heterogeneity settings, but is not able to recover the effect in the high

heterogeneity setting. The nearest-neighbor method is unable to recover the parameter in any of the data settings.

When estimating the customers' propensity to engage in an activity, our method is the only one to obtain coverage of the parameter of interest in all the data settings again. The case-deletion method performs well in the low heterogeneity settings, but is not able to recover the effect in the high heterogeneity setting. The nearest-neighbor method is unable to recover the parameter in almost any of the data settings.

If a company wants to target the customers that respond best to advertisements, we demonstrate that overall our method is the most effective at selecting those customers in all data settings. It selects the largest proportion of the 'true' most responsive customers, in the setting where the missing data is ignorable or non-ignorable.

For a firm hoping to gauge the size of their customer base, our model provided the most accurate estimate across all the methods. Assuming each anonymous visit is a unique-customer always overestimates the number of customers, while the nearest-neighbor and case-deletion methods always underestimates the total. By construction, the latter two methods assume that the number of customers is the number of observed customers.

We then tested our model on a dataset from a specialty retailer of consumer goods. First, we induced a correlation between missingness and the effect of a discount email on the propensity to make a transaction. The results we obtain are consistent with the synthetic data under this correlation structure. As before, our method performs best in the low heterogeneity setting. However, under the high heterogeneity setting, the case-deletion and the nearest-neighbor methods obtain less biased estimates.

We would now like to address the low performances of the methods in the high heterogeneity setting. In Chapters 6 and 7, we attribute the low performance in this setting to insufficient information from customers who rarely sign-in. We would like to

acknowledge that this low performance may also be due to a sparse number of visits per customer. If there are very few visits for each of the customers in the dataset, none of the methods will have sufficient observations at the individual-level to recover the parameters that the firm is interested in. In addition, the low performance may be caused by a large number of customers completely deleted. If a large proportion of the total customer base is entirely unobserved, none of the methods will have enough signal to infer either the number of unique customers or the individual-level behavior of the unobserved customers (who are a large proportion of the customer base), thereby poorly estimating the parameters of interest to the firm.

Regardless of whether or not there is any correlation between missingness and the effect of a marketing action on a company’s customer base, it will often make sense for a firm to use our model instead of either of the benchmark models (case-deletion or nearest-neighbor matching). By using all the available data, our model provides the most accurate estimates of the parameters across many different scenarios. It provides less biased estimates and makes a better selection of those customers at whom a firm should target its marketing efforts. It also obtains the most accurate estimate of the total number of unique customers in a firm’s database.

8.1 Computational Efficiency

A future focus of this research will be more computationally efficient strategies for implementing this type of Bayesian hierarchical model. The slow speed of the MCMC computation is a recurring issue, making the model hard to scale to a company’s full dataset. The computation time was very large because of the missing visit imputation step. Here, for every iteration of the sampler and for each anonymous visit, we had to compute the likelihood of every customer being assigned to that missing visit, prior to

sampling a single customer from a multinomial distribution.

In the future, we will examine alternative implementation strategies, with a focus on the expectation maximization (“EM”) algorithm. Within the EM estimation algorithm, there are further potential computational gains, which include only considering a representative subsample of the customers in the probabilistic imputation (which is obtained through randomly sampling visits). Another strategy for improving the computational time would be replacing the likelihood with an approximation.

8.1.1 Expectation-Maximization Algorithm

In the EM algorithm’s expectation step, for each anonymous visit, a fraction of that visit is assigned to each potential customer corresponding proportionally to the probability that such a visit pertains to that customer.

$$\begin{aligned}
 E(I_{U_j=i}) &= P(U_j^{mis} = i | \mathbf{Y}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}, \hat{\mathbf{y}}^*) \\
 &= \frac{\Phi_M\{\hat{\mathbf{y}}_j^* | \hat{\boldsymbol{\mu}}_{U_j,i}, \hat{\boldsymbol{\Sigma}}\} L_{\lambda_{U_j}} \hat{\delta}_i^{(V_j=1)} (1 - \hat{\delta}_i)^{(V_j=0)}}{\sum_{d=1}^D \Phi_M\{\hat{\mathbf{y}}_j^* | \hat{\boldsymbol{\mu}}_{U_j,d}, \hat{\boldsymbol{\Sigma}}\} L_{\lambda_{U_j}} \hat{\delta}_i^{(V_j=1)} (1 - \hat{\delta}_i)^{(V_j=0)}} \quad (8.1)
 \end{aligned}$$

We would replace the missing visits with their conditional expectation, which in this case is the fractional customer visits. Then, using these fractional assignments of customers, the remaining parameters are maximized using numerical methods, such as Newton-Raphson, and conjugate MAP estimates.

A question that would need to be immediately addressed is that of a truncation point for fractional visits. Since the number of customers whose fractional visits would be assigned in the Expectation step could be thousands or millions, we would have that many fractional visits to be assigned. Since we would need to use these fractional assignments to maximize the remaining parameters, this would make the computation

here even more burdensome than in the Bayesian hierarchical approach. Therefore, we would need to find an optimal truncation point for the fractional visits through synthetic data analysis. We would do so by trying a variety of different truncation points. We would select the point with the smallest number of fractional assignments that would still result in accurate estimation.

After determining a truncation point, there would be two more questions that would need to be addressed via simulation. For population-level parameters, how much total weight of visits needs to be assigned to a customer in order to include him as an individual for estimating population-level parameters? And for individual-level parameters, how much total weight of a visit needs to be assigned to a customer to include that fraction of a visit in estimating the individual-level parameters?

8.1.2 Representative Subsampling

As a further potential solution, we will not consider every observed and potential ‘new’ customer for an anonymous visit assignment. Instead, we only consider a random, representative subsample of customers for assignment, leading to a much smaller number of likelihood evaluations for each anonymous visit.

For example, if there are 1,000 potential customers in the firm’s database, that would require 1,000 likelihood computations for every anonymous visit (in both approaches). Instead of computing 1,000 likelihoods, we select a random representative subsample of customers, and only consider them for a visit assignment.

Since this representative subsample is random, it would be different at every iteration. This strategy is applicable for both the EM algorithm and the Bayesian hierarchical approach.

8.1.3 Likelihood Approximation

Another model improvement that might reduce the computational complexity is a more computationally efficient approximation of the likelihood. In the current Bayesian hierarchical model, we must compute the following likelihood for every customer for every anonymous visit:

$$\frac{\Phi_M\{\hat{\mathbf{y}}_j^*|\hat{\boldsymbol{\mu}}_{U_j,i},\hat{\boldsymbol{\Sigma}}\}L_{\lambda_{U_j}}\hat{\delta}_i^{(V_j=0)}(1-\hat{\delta}_i)^{(V_j=1)}}{\sum_{d=1}^D\Phi_M\{\hat{\mathbf{y}}_j^*|\hat{\boldsymbol{\mu}}_{U_j,d},\hat{\boldsymbol{\Sigma}}\}L_{\lambda_{U_j}}\hat{\delta}_i^{(V_j=0)}(1-\hat{\delta}_i)^{(V_j=1)}} \quad (8.2)$$

where $L_{\lambda_{U_j}}$ is part of the likelihood corresponding to the rate of arrival for user U_j . $L_{\lambda_{U_j}}$ is a complicated product that could be approximated by an average rate, where we take the number of visits during marketing action weeks over the total number of marketing action times as an estimate of the rate of arrival during marketing action times. Likewise, we would estimate the rate of arrival during non-marketing action times as the number of visits during non-marketing action times over the total non-marketing action times.

In summary, a major disadvantage to using our method over one of the alternative methods is the computational speed, since the imputation step in our Gibbs Sampler slows it down. That said, we discuss here three potential solutions to improving speed: representative subsampling, likelihood approximation, and the Expectation-Maximization algorithm. If these solutions increase the speed, we would eliminate the largest drawback from recommending a model like ours, making it even more appealing for firms with missing data to use.

Appendix A

Gibbs Sampler

A.1 Prior Distributions on Global Parameters

- (1) The first prior to the population-level regression coefficients, Γ , is

$$\Gamma_h | \mathbf{\Gamma}_0, \boldsymbol{\gamma}_0 \sim MVN(\boldsymbol{\gamma}_0, \mathbf{\Gamma}_0) \quad (\text{A.1})$$

where $h = 1, \dots, S$ indexes a row of $\mathbf{\Gamma}$ and where $\boldsymbol{\gamma}_0$ and $\mathbf{\Gamma}_0$ are fixed hyperparameters.

- (2) The prior to the population-level variance-covariance matrix that characterizes heterogeneity across the customers, $\mathbf{\Omega}$, is

$$\mathbf{\Omega} \sim InvWish_{\eta_0}(\mathbf{\Lambda}_0) \quad (\text{A.2})$$

where η_0 , and $\mathbf{\Lambda}_0$ are fixed hyperparameters.

- (3) The prior on the global correlations amongst the activities within visits, $\mathbf{\Sigma}$, is

$$\Sigma \sim \text{InvWish}_{\eta_0}(\mathbf{T}_0) \quad (\text{A.3})$$

for some fixed hyperparameters η_0 and \mathbf{T}_0 .

A.2 Gibbs Sampler Steps 1 through 8

- (1) We sample a specific user for each missing U_j from a multinomial distribution where the probability of visit j being made by user i is:

$$P(U_j^{\text{mis}} = k | \mathbf{Y}, \boldsymbol{\theta}, \Sigma, \mathbf{y}^*) = \frac{(\int_{G_{kM}} \cdots \int_{G_{k1}} \Phi_M\{\mathbf{y}^*_j | \boldsymbol{\nu}_{U_j, k} + \beta_{U_j, k}^T \mathbf{X}_j, \Sigma\} d\mathbf{y}^*_j) L_{\lambda_{U_j, k}} \delta_k^{(V_j=1)} (1 - \delta_k)^{(V_j=0)}}{\sum_{i=1}^I (\int_{G_{iM}} \cdots \int_{G_{i1}} \Phi_M\{\mathbf{y}^*_j | \boldsymbol{\nu}_{U_j, i} + \beta_{U_j, i}^T \mathbf{X}_j, \Sigma\} d\mathbf{y}^*_j) L_{\lambda_{U_j, i}} \delta_i^{(V_j=1)} (1 - \delta_i)^{(V_j=0)}}$$

where $i = 1, \dots, I$ are the total potential users that could be assigned to an anonymous visit and $L_{\lambda_{U_j, i}}$ is the part of the likelihood that corresponds to the rate of arrival for user i .

- (2) Sample our parameters for the underlying propensity to visit a page on a particular visit, \mathbf{y}^* , for all pages M and all rows n from a truncated multivariate normal distribution,

$$y_{jm}^* | \boldsymbol{\theta}_i, \Sigma, \mathbf{U}, \mathbf{y}_{j(-m)}^* \sim e^{(-\frac{1}{2}(\mu_i^*)'(\Sigma^*)^{-1}\mu_i^*)} \times \{I(y_{jm}^* > 0)I(y_{jm} = 1) + I(y_{jm}^* < 0)I(y_{jm} = 0)\} \quad (\text{A.4})$$

and where

$$\mu_i^* = (\nu_{U_j} + \beta_{U_j}^T X_j)^* = (\mu_i)_m + \Sigma_{12} \Sigma_{22}^{-1} (\mathbf{y}^*_{j(-m)} - (\boldsymbol{\mu}_{i(-m)})) \quad (\text{A.5})$$

and

$$\Sigma^* = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \quad (\text{A.6})$$

We use the ‘‘star’’ notation to mean the Schur compliment. For example, Γ^* and Ω^* for the m th page would mean

$$\Gamma^* = (\Gamma)_m + \Omega_{12} \Omega_{22}^{-1} (\boldsymbol{\theta}_{i(-m)} - (\boldsymbol{\Gamma})_{(-m)})$$

$$\Omega^* = \Omega_{11} - \Omega_{12} \Omega_{22}^{-1} \Omega_{21},$$

and

$$\boldsymbol{\Gamma} = \begin{pmatrix} (\Gamma)_m \\ (\boldsymbol{\Gamma})_{(-m)} \end{pmatrix}, (\Gamma)_m \text{ is } 1 \times 1, (\boldsymbol{\Gamma})_{(-m)} \text{ is } (M+1) \times 1$$

$$\text{and } \boldsymbol{\Omega} = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix} \text{ with size } \begin{pmatrix} 1 \times 1 & 1 \times (M+1) \\ (M+1) \times 1 & (M+1) \times (M+1) \end{pmatrix}$$

where we denote Ω_{11} as the variance for the m^{th} entry.

- (3) Sample our user specific parameters, $\boldsymbol{\theta}_i$. This consists of three parts. First we sample the β_i 's and ν_i 's,

$$\beta_i, \nu_i | Z_{U_j=i}, \boldsymbol{\Gamma}^*, \boldsymbol{\Omega}^*, \Sigma \sim MVN(\hat{\beta}_*, \mathbf{V}_{\beta_*}) \quad (\text{A.7})$$

$$\text{where } (\mathbf{y}^*)_* = \begin{pmatrix} \mathbf{y}^*_{U_j} \\ [\boldsymbol{\Gamma} Z_i]^* \end{pmatrix}, \mathbf{X}_* = \begin{pmatrix} \mathbf{X} \\ \mathbf{I}_p \end{pmatrix}, \text{ and } \Sigma_* = \begin{pmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \Omega^* \end{pmatrix}$$

$$\text{and } \hat{\beta}_* = (X_*^T \Sigma_*^{-1} X_*)^{-1} X_*^T \Sigma_*^{-1} (\mathbf{y}^*)_*,$$

$$\mathbf{V}_{\beta_*} = (X_*^T \Sigma_*^{-1} X_*)^{-1}.$$

We use the notation $[\Gamma Z_i]^*$ and Ω^* as we did above.

Next, for $(\theta_{i,(M+M \times P+1)}, (\theta_{i,(M+M \times P+2)})) = (\omega_{i,0}, \omega_{i,1})$, we must do two Metropolis steps since we have non-standard distributions.

First for $\omega_{i,0}$, we have a proposal,

$$\omega'_{i0} \sim N(\omega_{i0}, \zeta^2) \tag{A.8}$$

where ζ^2 is a tuning parameter and do a Metropolis steps with

$$P(\lambda_i | \boldsymbol{\lambda}_{-i}, \mathbf{y}, \mathbf{U}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}) \propto \prod_{j=1}^n L_{\lambda_{i,t_j}} \times e^{-\frac{1}{2}(\theta_{(M+M \times P+1)i} - ((\Gamma Z_i)^*)'(\Omega^*)^{-1}(\theta_{(M+M \times P+1)i} - (\Gamma Z_i)^*))^2} \tag{A.9}$$

where $\log(\lambda_{i,t_j}) = \omega_{i,0} + \omega_{i,1} H_{i,t_j}$, and $L_{\lambda_{i,t_j}}$ is the product of the parts of the likelihood that correspond to user i .

Next, for $\theta_{i,(M+M \times P+2)} = \omega_{i,1}$, we use the same density function for the Metropolis step (as for $\omega_{i,0}$), except we now hold $\theta_{i,(M+M \times P+1)}$ fixed. We use the same tuning parameter, ζ^2 , and draw a proposal

$$w'_{i,1} \sim N(w_{i,1}, \zeta^2) \tag{A.10}$$

For $\theta_{i,(M+M \times P+3)} = \text{logit } \delta_i$, we must also do a Metropolis step. We use η^2 for the tuning parameter, and draw

$$\delta'_i \sim N(\delta_i, \eta^2) \tag{A.11}$$

and do a Metropolis step with

$$P(\delta_i | \boldsymbol{\delta}_{-i}, \mathbf{Y}, \mathbf{U}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}) \propto \prod_{j=1}^n [\delta_i^{(V_j=1)} (1 - \delta_i)^{(V_j=0)}] \times e^{-\frac{1}{2}(\boldsymbol{\theta}_{(M+M \times P+3)i} - (\boldsymbol{\Gamma} \mathbf{Z}_i)^*)' (\boldsymbol{\Omega}^*)^{-1} (\boldsymbol{\theta}_{(M+M \times P+3)i} - (\boldsymbol{\Gamma} \mathbf{Z}_i)^*)} \quad (\text{A.12})$$

where $V_j = 0$ if user i is known at visit j , and $V_j = 1$ if user i is anonymous.

(4) Sample $\boldsymbol{\Gamma}$,

$$\boldsymbol{\Gamma} | \boldsymbol{\Omega}, \mathbf{U}, \boldsymbol{\theta} \sim \text{MVN}(\hat{\boldsymbol{\Gamma}}_*, \mathbf{V}_{\boldsymbol{\Gamma}*}) \quad (\text{A.13})$$

$$\text{where } \theta_* = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_I \\ \Gamma_0^1 \\ \vdots \\ \Gamma_0^S \end{pmatrix} \quad X_* = \begin{pmatrix} Z_1 & 0 & 0 \\ 0 & Z_1 & 0 \\ 0 & 0 & Z_1 \\ \vdots & & \\ Z_I & 0 & 0 \\ 0 & Z_I & 0 \\ 0 & 0 & Z_I \\ I_{(M+M \times P+3)*S} \end{pmatrix}, \text{ and } \Omega_* = \begin{pmatrix} \Omega & 0 \\ 0 & \Gamma_0 \end{pmatrix}$$

$$\text{and } \hat{\boldsymbol{\Gamma}}_* = (X_*^T \Omega_*^{-1} X_*)^{-1} X_*^T \Omega_*^{-1} \theta_*,$$

$$\mathbf{V}_{\boldsymbol{\Gamma}*} = (X_*^T \Omega_*^{-1} X_*)^{-1}.$$

(5) Sample $\boldsymbol{\Sigma}$,

$$\boldsymbol{\Sigma} | \mathbf{U}, \boldsymbol{\Gamma}, \boldsymbol{\Omega} \sim \text{InvWish}(\eta_0 + n, \mathbf{S}) \quad (\text{A.14})$$

$$\text{where } \mathbf{S} = \mathbf{T}_0 + \sum_{j=1}^n (\mathbf{y}^*_j - \boldsymbol{\mu}_{\mathbf{U}_j})(\mathbf{y}^*_j - \boldsymbol{\mu}_{\mathbf{U}_j})^T.$$

(6) Sample $\boldsymbol{\Omega}$,

$$\boldsymbol{\Omega}|\nu_0, \boldsymbol{\Lambda}_0, \kappa_0, \boldsymbol{\Gamma}, \boldsymbol{\theta} \sim \text{InvWish}(\nu_0 + I, \boldsymbol{\Lambda}_n) \quad (\text{A.15})$$

where $\boldsymbol{\Lambda}_n = \boldsymbol{\Lambda}_0 + \sum_{i=1}^I (\boldsymbol{\theta}_i - \boldsymbol{\Gamma}\mathbf{Z}_i)(\boldsymbol{\theta}_i - \boldsymbol{\Gamma}\mathbf{Z}_i)^T$

(7) Sample \mathbf{Z}_i ,

$$\mathbf{Z}_i|\boldsymbol{\Omega}, \mathbf{U}, \boldsymbol{\theta}_i \sim \text{MVN}(\hat{\mathbf{Z}}_*, \mathbf{V}_{\mathbf{Z}_*}) \quad (\text{A.16})$$

where $\hat{\mathbf{Z}}_* = (\boldsymbol{\Gamma}^T \boldsymbol{\Omega}^{-1} \boldsymbol{\Gamma} + \mathbf{P}_0)^{-1} (\boldsymbol{\Gamma}^T \boldsymbol{\Omega}^{-1} \boldsymbol{\theta}_i + \mathbf{P}_0 \boldsymbol{\xi}_0)$

and $\mathbf{V}_{\mathbf{Z}_*} = (\boldsymbol{\Gamma}^T \boldsymbol{\Omega}^{-1} \boldsymbol{\Gamma} + \mathbf{P}_0)^{-1}$

and where \mathbf{P}_0 and $\boldsymbol{\xi}_0$ are the prior parameters.

Appendix B

Computational Details

B.1 Parameter Recovery

Before comparing our model to the competitor models, we demonstrate that our model is recovering the parameters. In Table B.1, we present full parameter recovery across a representative sample of parameters in the model. We ran a simulated dataset for 5,000 iterations, used the first 1,000 as burn-in, and thinned every 10 iterations. We generated 400 true underlying customers, and had approximately 30 percent of the data missing, and had customers arriving on average 5 times in the dataset, ranging from 1 arrival to 15 arrivals per customer. We see that all the parameters are well estimated and covered by their 95 percent posterior intervals.

Parameter	True Value	Estimate	Interval Width
$\Gamma_{1,1}$	-0.84	-0.83	(-0.95,-0.71)
$\Gamma_{2,1}$	0	0.02	(-0.08,0.12)
$\Gamma_{3,1}$	0	-0.04	(-0.15,0.07)
$\Gamma_{4,1}$	0.70	0.69	(0.59,0.79)
$\Gamma_{11,1}$	0.025	-0.14	(-0.26,-0.02)
$\Gamma_{12,1}$	1.02	1.23	(1.01,1.45)
$\Gamma_{13,1}$	1.01	1.15	(0.94,1.36)
$\Gamma_{14,1}$	-1.50	-1.46	(-1.61,-1.31)
$\Omega_{1,1}$	0.20	0.23	(0.19,0.27)
$\Omega_{2,2}$	0.50	0.62	(.46,.76)
$\Omega_{3,3}$	0.50	0.59	(.46,.72)
$\Sigma_{1,1}$	1.0	1.06	(0.98,1.14)
$\Sigma_{2,2}$	1.0	1.07	(.99,1.15)
$\Sigma_{3,3}$	1.0	1.06	(0.98,1.14)
$\theta_{1,1}$	-1.04	-1.05	(-1.27,-0.83)
$\theta_{2,3}$	-0.44	-0.43	(-1.17,0.29)
$\theta_{5,6}$	0.80	0.55	(-0.18,1.28)
$\theta_{15,2}$	0.07	-0.27	(-1.16,.62)

Table B.1: Parameter Recovery for a Representative Sample of Parameters in the Model

B.2 Demonstration that Subsampling Works

Since the computational time is slow, we implement a subsampling strategy on the real data. However, before we do so, we first demonstrate that it works via simulation.

In the subsampling strategy, we randomly select 10 percent of the total, potential population of customers for each anonymous visit at each iteration. In the imputation step, we only consider imputing one of the randomly selected subset of customers into each anonymous visit.

We consider the setting where missingness (δ_i) is correlated with the propensity to visit in response to a marketing action (ω_{i1}) from the simulation studies. We consider Case 2, where there is 45 percent missingness, low heterogeneity, and a high correlation of 0.9 between missingness and the propensity to visit in response to a marketing

action.

We compare the population-level parameter estimates (Γ) between our model and our model with subsampling. In Table B.2, we include the baseline parameter estimates as well as their posterior intervals. We highlight the ones that overlap between the two models in gray.

Model Estimate	Model Interval Width	Sub Model Estimate	Sub Model Interval Width
0.10	(-0.08,0.30)	0.11	(-0.05,0.27)
-0.05	(-0.26,0.13)	0.05	(-0.11,0.23)
-0.17	(-0.34,-0.01)	-0.09	(-0.26,0.05)
0.63	(0.42,0.81)	0.65	(0.51,0.80)
0.67	(0.49,0.83)	0.62	(0.48,0.78)
0.70	(0.53,0.88)	0.67	(0.52,0.86)
0.02	(-0.17,0.20)	0.07	(-0.09,0.23)
-0.09	(-0.31,0.09)	-0.07	(-0.29,0.07)
-0.01	(-0.20,0.18)	0.00	(-0.20,0.13)
0.00	(-0.18,0.20)	0.01	(-0.15,0.19)
-0.96	(-1.15,-0.79)	-0.88	(-1.04,-0.74)
-0.84	(-1.01,-0.67)	-0.77	(-0.95,-0.63)
-0.78	(-0.95,-0.63)	-0.76	(-0.86,-0.58)
-0.76	(-0.84,-0.69)	-0.71	(-0.84,-0.68)
-0.77	(-0.93,-0.61)	-0.76	(-0.84,-0.56)
-0.74	(-0.92,-0.59)	-0.71	(-0.85,-0.57)
-0.97	(-1.16,-0.81)	-0.84	(-1.01,-0.70)
0.09	(-0.06,0.24)	0.09	(-0.04,0.21)
-0.10	(-0.26,0.03)	-0.08	(-0.22,0.03)
0.02	(-0.13,0.17)	0.03	(-0.10,0.17)
-0.01	(-0.06,0.02)	-0.01	(-0.06,0.02)
0.11	(0.04,0.17)	0.12	(0.06,0.19)
-0.53	(-0.64,-0.44)	-0.47	(-0.59,-0.33)

Table B.2: Parameter Estimates in the Full Model versus Model with Subsampling

A comparison of parameter estimates between the full model and the model with subsampling. We include parameter estimates and interval widths. We highlight the intervals that overlap in gray.

Next, we consider customer rankings between the two models. We look at the

selections of the six models of the top customers that would have the strongest reaction (or the highest propensity to buy) after receiving a marketing action. We see that even though our model does not select the largest number of correct customers in every case, our model and the subsampled model consistently produce similar results across all the cases.

Number of Customers	Top 50	Top 100	Top 150
full model	10	44	108
subsampled model	8	40	106
case-deletion	13	55	96
nearest-neighbor	12	51	94

Table B.3: Rank Ordering Customers in the Full Model versus Model with Subsampling

Each model produces its own rank ordering of customers in terms of strongest propensity to visit in response to a marketing action (ω_{i1}) in the setting where the propensity to visit in response to a marketing action, ω_{i1} , is correlated with missingness (δ_i). For each of the methods, we first select the top 100 customers that would have the strongest propensity to visit in response to a marketing action. We compared each model’s selection to the originally generated top 100 customers with the highest response to see how many were correctly chosen.

Bibliography

- [1] S. Chib and E. Greenberg. Analysis of multivariate probit models. *Biometrika*, 85(85):347–361, 2 1998.
- [2] E. M. Feit, M. A. Beltramo, and F. M. Feinberg. Reality check: combining choice experiments with market data to estimate the importance of product attributes. *Management Science*, 56(5):785–800, 5 2010.
- [3] E. M. Feit, P. Wang, E. T. Bradlow, and P. S. Fader. Fusing aggregate and disaggregate data with an application to multiplatform media consumption. *Journal of Marketing Research*, 50(3):348–364, 6 2013.
- [4] A. Gelman, J. B. Carlin, Hal S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. CRC Press, Boca Raton, FL, 2003.
- [5] K. T. Gordon. The newest insights into loyalty programs reveal the best ways to engage customers, March 2010. [Online; posted 12-March-2010].
- [6] Mariusz Grabowski. Handling missing values in marketing research using som. In Daniel Baier and Klaus-Dieter Wernecke, editors, *Innovations in Classification, Data Science, and Information Systems*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 322–329. Springer Berlin Heidelberg, 2005.
- [7] J. W. Graham, S. M. Hofer, and A. M. Piccinin. Analysis with missing data in

- drug prevention research. In L. Collins and L. Seitz, editors, *National Institute on Drug Abuse Research Monograph Series*, volume 142, pages 201–213. National Institute on Drug Abuse, Washington, D.C., 1994.
- [8] K. B. Grant. Retailers expand customer-loyalty programs, March 2008. [Online; posted 31-March-2008].
- [9] R. Grover and M. Vriens. *The Handbook of Marketing Research: Uses, Misuses, and Future Advancements*. Sage Publications, Thousand Oaks, CA, 2006.
- [10] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics, New York, NY, 2009.
- [11] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.
- [12] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics, Hoboken, NJ, 2002.
- [13] P. Manchanda, A. Ansari, and S. Gupta. ‘the shopping basket’: A model for multicategory purchase incidence decisions. *Marketing Science*, 18(2):95–114, 2 1999.
- [14] F. Mulhern. Integrated marketing communications: From media channels to digital connectivity. *Journal of Marketing Communications*, 15(2):85–101, 2 2009.
- [15] A. Musalem, E. T. Bradlow, and J. S. Raju. Who’s got the coupon: Estimating consumer preferences and coupon usage from aggregate information. *Journal of Marketing Research*, 45(2):715–730, 12 2008.
- [16] J. C. Nunes and X. Dreze. Your loyalty program is betraying you. *Harvard Business Review*, 84(4):124–131, 5 2006.

- [17] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 4 1983.
- [18] P. E. Rossi, G. M. Allenby, and R. McCulloch. *Bayesian Statistics and Marketing*. Wiley Series in Probability and Statistic, Hoboken, NJ, 2005.
- [19] P. E. Rossi, R. E. McCulloch, and G. M. Allenby. The value of purchase history data in targeted marketing. *Marketing Science*, 15(4):321–340, 1 1996.
- [20] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1 1976.
- [21] D. B. Rubin. *Multiple Imputations in Sample Surveys- A Phenomenological Bayesian Approach to NonResponse*. Proceedings of the International Statistical Institute, Manila, 1978.
- [22] D. B. Rubin and N. Schenker. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81(1):366–374, 1 1986.
- [23] M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 6 1987.
- [24] C. Tode. Nordstrom upgrades loyalty program experience, May 2007. [Online; posted 04-May-2007].
- [25] R. S. Winer. New communications approaches in marketing: Issues and research directions. *Journal of Interactive Marketing*, 23(2):108–117, 2 2009.
- [26] D. Zantedeschi, E. M. Feit, and E. T. Bradlow. Measuring multi-channel advertising effectiveness using consumer-level advertising response data. 11 2013.