BAYESIAN NONPARAMETRIC METHODS IN MARKETING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Saisandeep Reddy Satyavolu

May 2016

BAYESIAN NONPARAMETRIC METHODS IN MARKETING

Saisandeep Reddy Satyavolu, Ph.D.

Cornell University 2016

The proliferation of available data in marketing has placed an emphasis on the applicability of extant marketing models to big data. To tackle this problem, methods from machine learning have been increasingly applied by the marketing community. This line of research is a subset of research in marketing that is becoming interdisciplinary. A number of marketing researchers have successfuly adopted methods from other seemingly unrelated fields in their research. In that vein, this thesis examines the applicability of Bayesian Nonparametric methods (from the field of machine learning) to marketing.

The first chapter of this thesis provides a very brief survey of marketing research papers that have enhanced pure marketing models using methods from machine learning. The second chapter describes the Dirichlet Process, a key component of Bayesian Nonparametric analysis and provides two synthetic data applications. Going forward, we study the applicability of Bayesian Nonparametric methods to model Heterogeneity across multiple markets. Bayesian Nonparametric methods have been used in marketing and economics literature to model heterogeneity in discrete choice models, but past applications have only been limited to data from a single market. So as to compare heterogeneity in consumer preferences across multiple markets, we use the Hierarchical Dirichlet Process (HDP) which lets multiple "groups" of data "share statistical strength".

Heterogeneity across multiple markets is modeled using the HDP in two different contexts (B2C and B2B) in this thesis. Our work shows that the HDP provides a convenient "middle ground" to other extreme modeling options, which are (1) ignore heterogeneity of preferences across markets and (2) model each market separately. Another aspect of the HDP is the ease with which it can be incorporated into models of discrete choice. The models developed and estimated in this thesis are also helpful for the marketing manager. In the B2C application, the results of the model provide the manager with a practical way of tailoring targeting activities towards consumers with varying preferences. Finally, in the B2B application, we find that based on the Stage of the selling process, some marketing activities play a larger role than others in converting sales leads into clients. These results provide a data driven basis for the manager to appropriately allocate marketing dollars to activities based on the selling process.

BIOGRAPHICAL SKETCH

Saisandeep (Sandeep) R. Satyavolu was born on February 11, 1986 in Nellore, India to Sudheera Reddy Satyavolu and Hariprasad Reddy Satyavolu. From a young age, he was enamored with numbers, and a common past time of his was to add up numbers he saw on anything (car number plates, for example) and check if they were divisble by the digits from 2 to 9. He was able to come up with an empirical measure for the chance that a given car number was divisible by these digits (basically an empirical distribution function, though he didn't know what those were called at that time).

His interest in numbers was misconstrued as an interest in engineering and he pursued a degree in Civil Engineering at the Indian Institute of Technology Bombay from 2003 to 2007. He showed more interest in statistical methods used to analyze remote sensing and GPS data while there. Pursuing that interest in statistical methods took him to Stanford University, where he obtained a Master's degree in Management Science and Engineering.

Sandeep started his Ph.D. program in Cornell University in the fall of 2010. He has been fortunate enough to share ideas with Prof. Vithala Rao, Prof. Vrinda Kadiyali, Prof. Francesca Molinari and Prof. Sachin Gupta, all of whom have helped and supported him during his Ph.D. journey. He hopes to be a true Bayesian and upate his research perspectives based on theirs in the future. In June 2016, he will join the IT department at Procter and Gamble Corporation as a Data Scientist.

Sandeep married Swarnalatha on April 18, 2014.

I dedicate this thesis to my family.

ACKNOWLEDGEMENTS

I would like to thank my advisor Professor Vithala R. Rao for his excellent guidance and support during my Ph.D. His productivity and dedication to research is something I wish to replicate in my own professional journey. His extraordinary breadth of knowledge and insight have helped me time and again when I hit roadblocks in my research. I also sincerely thank Prof. Vrinda Kadiyali and Prof. Sachin Gupta for encouraging me to pursue difficult topics on my own emphasizing the importance of learning new things, and always keeping their door open for me no matter how busy their schedules were.

Pursuing nonparametric techniques would not have been possible if it weren't for Prof. Francesca Molinari's encouragement. Her approach of gently presenting very difficult ideas made it easier to retain key ideas from the field, something which would've easily taken me longer to do on my own. I also thank Prof. Martin Burda and Prof. Andrew Ching for introducing me to Bayesian methods, while Prof. Ching was hosting me at the University of Toronto.

I thank all the wonderfull teachers Cornell has given me during my time here. Thanks also to Prof. Sanjog Misra, Prof. Paul Ellickson and Prof. Minjae Song for letting me attend their course at the University of Rochester. Thank you to the support staff in Sage 304 for all the help during my Ph.D., and the folks at the Johnson IT office for supporting me with all my crazy computing requests. Thanks also to all of the Ph.D.'s I shared office space with during my time here. I'm sure I have not been able to thank everyone who has made my experience at Cornell a memorable one, and for that, I apologize.

This journey wouldn't have been possible without the encouragement of my family and friends back home in India. And last, but not least, thanks to my wife

Swarna, who has made incredible sacrifices to support me wholeheartedly in all of my endeavors. She believed in me at times I didn't even believe in myself. Thank you for being the person you are.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

MARKETING AND MACHINE LEARNING: A BRIEF OVERVIEW

"We are drowning in information but starved for knowledge" – John Naisbitt

"Data are becoming the new raw material of business" – Craig Mundie

"Not everything that can be counted counts, and not everything that counts can be counted" – William Bruce Cameron

1.1 Introduction

As Kevin Murphy [79] puts it, machine learning can be thought of as a set of tools for automated data analysis (i.e. a collection of algorithms that can automatically detect deep patterns in data). The need (and also the cause for the growth) of machine learning comes from the abundance of user data available from the Internet and company databases [79], making traditional methods of data analysis cumbersome. Hal Varian [113] acknowledges this phenomenon and advocates the use of machine learning in econometrics, citing the need to scale extant estimation methods to big data.

Many of the methods put forward in machine learning have actually been used in economics and marketing, but they were just given a different name. For example, the problem of classification (a fundamental concept in machine learning), where different data points are allocated to groups or "clusters" based on certain attributes, lies at the core of many marketing models (as a strategy to model heterogeneity). Some of these methods have also been discussed in [113].

In the past 40 years, there has been a movement towards the use of sophis-

ticated models that appropriately explain the data, rather than apply reduced form models without testing whether they are a good approximation [89]. With sophistication, there is always the issue of computational burden, and with the advent of big data, this problem is compounded. However, this problem can be alleviated with the use of machine learning techniques in conjunction with marketing (or econometric) mdoels. This helps managers make informed decisions based on data analysis rather than gut feel, as described in [102]. In [102], the authors use Support Vector Machine and Neural Network algorithms to classify customers who were most likely to convert into mobile internet users. They found that their algorithm came up with conversion rates far better than the current best marketing practices adopted by mobile network operators. Dzyabura and Hauser [31] develop an "active machine learning" method to select questions adaptively, when consumers use heuristic decision rules and from their analysis, find that this startegy of using adaptive questions provides better information about these decision rules than existing methods.

This chapter is a very brief note on the use of machine learning concepts in marketing (and economics), citing work done in the past, and the possibility of future work along those lines. The next sections introduce machine learning concepts, and their use in marketing literature. This is by no means a comprehensive list (or a description of concepts), as that would warrant a book by itself.

1.2 Convex Optimization

Even though convex optimization is more of an optimization technique than a machine learning concept, its uses in machine learning are widespread [20]. Indeed, the topics in optimization and machine learning are acknowledged to be "intertwined" [12]. Although by definition, convex optimization appears to apply to a very narrow range of problems (i.e. problems where the objective and constraint functions are convex), it is surpisingly applicable to a larger set of problems [53, 54, 19, 64]. In the field of partial identification, Beresteanu, Molchanov and Molinari [13] study econometric models with convex moment predictions, and exploit the convexity property to use algorithms in convex optimization to recover the partially identified set of parameters. Empirical Static games [9, 32] also belong to the category of models with convex moment predictions, and Beresteanu, Molchanov and Molinari [13] apply their methodology there too. Convex optimization has also been creatively used to model heterogeneity in conjoint analysis models [37, 109]).

1.3 Dynamic Programming

To repeat the idea from the previous section, dynamic programming could be interpreted as more of a (recursive) optimization technique than a concept in machine learning [14, 11]. There are various flavors of this idea, the more relevant one (in terms of being considered a machine learning concept) being that of reinforcement learning [111, 10, 88]. The use of dynamic programming in structural economics began with the seminal papers of Wolpin, [117], Pakes [85] and Rust [92]. The idea was extended (empirically) to a mulple player setting

(dynamic games) in Pakes and McGuire [86] and Ericson and Pakes [34]. A survey of these methods is given in Aguirregabiria and Mira [1]. Marketing models with forward looking consumers also use this solution concept, with the dynamic link between periods being the idea that consumers buy products with an intent to reap rewards later (i.e. frequent flyer programs on airlines). In marketing, the applications of these models have been diverse, to the sales of digital cameras [95, 99], video games [80], supermarket pricing strategies [33] and salesforce compensation [73] to name a few.

Warren Powell [88] discusses various strategies of reducing the computational burden in dynamic programming problems, approximating value functions being one of them (this borrows a lot from machine learning and statistical learning theory). These methods were applied in replicating the results of Ericson and Pakes [34], by Farias, Saure and Weintraub [39]. Bhat, Farias and Moallemi [15] extend the ideas in approximate dynamic programming to a "practical non-parametric" version, which is a dimension-independent approximation.

In the recent past, there has been a push towards a bayesian approach to dynamic structural models. A key advantage of this method is obviating the need to use dynamic programming methods [44]. These ideas were further developed in Imai, Jain and Ching [56] and Norets [82]. Norets [83] uses MCMC and Artifical Neural Networks to approximate the Dynamic Programming solution as a function of the parameters and state variables so as to avoid solving the dynamic program at each iteration (he also shows that this methodology is applicable in a variety of situations).

One of the many issues encountered in the area of reinforcement learning

is the trade-off between exploration and exploitation (i.e. "when does one stop learning and start exploiting the information they've learnt in the past?") [27]. This is also faced in a class of problems called bandit problems. A number of solutions have been proposed in the past, one of the early ones being provided by Gittins [46] and Gittins and Jones [47]. Recently, these methods have been used in problems of experience goods provision as applied to antidepressants, [30]) and learning from experience, as applied to diaper sales [71].

1.4 Support Vector Machines

Support Vector Machines (SVM) are a special case of sparse kernel machines[1] [79]. The idea is to define a loss function for data analysis (using kernel methods), which would ensure a sparse solution, and that would mean that predictions would only depend on a select set of training data, called support vectors (and hence the name support vector machines for kernel models with modified loss functions ensuring sparse solutions). In marketing literature, Cui and Curry [28] test the predictive power of SVMs, and assess their strengths and weaknesses against traditional marketing models. Evgeniou, Pontil and Toubia [37] briefly mention how modifying some of their (conjoint) model parameters gives rise to a SVM specification [2]. In economics, adding to the work of partial identification in models with convex moment predictions in [13], Molinari and Bar [77] provide a solution for the issue of finding sharp identification regions for high dimensional parameters, using SVMs.

[1] Loosely put, a kernel machine is a Generalized Linear Model (GLM) with the input feature vector being a function of the data point and the chosen "centroid" of the data

[2] [109] gives a thorough description of machine learning and optimization concepts as applied to conjoint analysis estimation and question design.

1.5 Latent Dirichlet Allocation

First introduced in Blei, Ng and Jordan [18], Latent Dirichlet Allocation (LDA) provides an unsupervised method to extract latent dimensions (among other things) for "collections of discrete data such as text corpora". The LDA is structured as a three-level hierarchical bayesian model, where depending on the hierarchy, variables are modeled as infinite or finite mixtures over the underlying set of probabilities. Tirunillai and Tellis [106] use LDA to extract the valence expressed in user generated content, and extend the model to study context specific valence in product reviews across 15 firms in 5 markets over 4 years (this exercise also shows the scalability of LDA to big data).

1.6 Bayesian Nonparametrics

In Bayesian Nonparametrics, the term "nonparametrics" is a misnomer, since it would imply that this class of models is parameter free. Instead, these models allow for infinitely many parameters, choosing relevant ones as data become available [84]. Given the structure of bayesian models, defining a prior on a set of possible distributions (a feature of nonparametric problems) which was large enough and resulted in a posterior distribution which was tractable (analytically or computationally) was a challenge. Ferguson [40] was one of the first to come up with an attractive solution (that of a Dirichlet Process prior), the properties of which are described in Teh [104]. One of the main advantages of a bayesian nonparametric paradigm (especially dirichlet process mixture models) is the ability of these models to infer the number of clusters (or latent groups) in a given dataset, effectively modeling for heterogeneity [43].

Bayesian nonparametric methods have been used in different contexts in marketing and economics in the past. Burda, Harding and Hausman [23] model supermarket choices, allowing a few key individual and alternative specific parameters of interest to be nonparametric, whereas the others are drawn from a multivariate normal, and this eliminates the independence of irrelevant alternatives assumption. They find a complex multi-modal preference distribution, differentiating between customers who strongly value lower prices (or shopping convenience) and the others. Li and Ansari [69] apply centered Dirichlet Process Mixtures to model endogeneity and heterogeneity in discrete choice models and they find the semiparametric model to outperform traditional modeling specifications. Dzyabura and Hauser [31] apply variational techniques to approximate the posterior in a setting where consumers are using heuristic decision rules when answering questions.

1.7 Structure of The Thesis

As mentioned earlier, this particular chapter was not meant to be a comprehensive description of machine learning methods in marketing. However, this serves as a starting point, in that it shows the concepts of machine learning that have been employed so far in marketing research. This thesis shares the same theme – of using an idea from machine learning, namely Bayesian Nonparametrics, and applying it in two different contexts in marketing literature. The next chapter introduces the Dirichlet Process, and shows two synthetic data applications which demonstrate the flexibility of the Dirichlet Process prior and its ease in being "embedded" in more involved marketing models. Chapter 3 introduces the Hierarchical dirichlet Process [105], and decribes an application

where across market heterogeneity can be modeled flexibly while maintaining model parsimony. This thesis concludes with Chapter 4, which also uses the same technique to model heterogeneity across markets, but in a B2B context, where the selling process of the firm being studied plays a key role in model development.

CHAPTER 2

THE DIRICHLET PROCESS

Very loosely put, the Dirichlet Process is the infinite dimensional version of the dirichlet distribution. In one of its earliest applications, Ferguson [40] used it as a prior distribution that was key to extending the framework of Bayesian analysis to common nonparametric problems. To elaborate, consider an example where we have N data points, and we know before hand that these data were generated from a normal distribution. In this example, the data are completely determined by the posterior mean (μ) and variance (σ^2) parameters, as shown in the Directed Acyclic Graph (DAG henceforth), 2.1 below (in Figure 2.1, the rectangle enclosing the shaded data bubble (x_i) is called a plate [16], which indicates the number of data points observed). However, very rarely do we know beforehand what the the underlying distribution for a given data sample is, and the DAG in Figure 2.1 could very well be an oversimplification of the problem. In the past, nonparametric methods (in a frequentist context) have been proposed to deal with this issue [68].

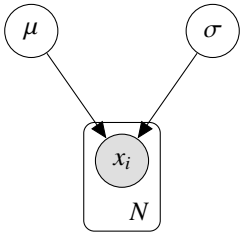

Figure 2.1: A Simple Bayesian Model

Modeling data nonparametrically in a Bayesian context is a little more involved, since ideally, according to Feguson [40] the prior being used should have a "large" support, and the posterior distributions from a data fitting exer-

cise should be "manageable analytically". The Dirichlet process, is considered to be a "distribution over distributions" (Antoniak [8]) which satisfies both the aforementioned conditions [40]. Hence, a draw from a dirichlet process is a probability distribution. Going back to our simple example in Figure 2.1, we now consider the case where we leave the distribution of the data unspecified, and place a Dirichlet Process prior on it. A Dirichlet Process is specified by a concentration parameter $\alpha > 0$ and a base distribution H [101, 42]. A DAG representing this model is given in Figure 2.2

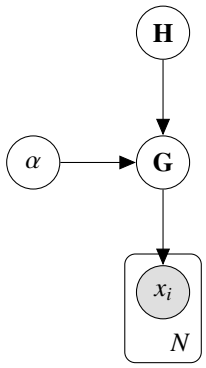

Figure 2.2: A Simple Dirichlet Process Model

G is a random distribution, specifically, a draw from a Dirichlet Process with concentration parameter α and base distribution H, denoted $G \sim DP(\alpha, H)$ (hence, G has a Dirichlet Process Prior [108]). The data x_i are draws from G. By construction, G is a discrete distribution [42], i.e. it places point masses at different distinct values of x_i, and there is a chance of repetition of x_i values with probability 1 (also a reason why a model specified according to the DAG in Figure 2.2 might not be appropriate for our simple example). To understand what G looks like, we need to understand what structure the Dirichlet Process prior places on the distribution. Following the notation in Tomlinson [108], let χ be the support of G. If $B_1, B_2, \ldots, B_K$ is an arbitrary partition of χ, then the (discrete)

probability vector $G = (G(B_1), G(B_2), \ldots, G(B_K))$, by definition is distributed according to [107, 108]:

$$(G(B_1), G(B_2), \ldots, G(B_K)) \sim Dir(\alpha H(B_1), \alpha H(B_2), \ldots, \alpha H(B_K)) \tag{2.1}$$

Where $H(B_i) = \int_{x \in B_i} dH(x)$. We began this section by saying that the Dirichlet Process is the infinite dimensional version of the dirichlet distribution. However, once we take a finite partition of the support of the random distribution G, the probabilities assigned to these partitions are distributed according to a dirichlet distribution (with the dimensionality equaling the size of the partition). Note that by definition, since $B_1, B_2, \ldots, B_K$ is a partition of χ, $\sum_{i=1}^{K} G(B_i) = 1$. Since the vector of probabilities follows a dirichlet distribution, all the properties derived in the appendix apply. Specifically, for a given partition B_i, we have the following:

$$G(B_i) \sim Beta\left(\alpha H(B_i), \sum_{j \neq i} \alpha H(B_j)\right) = Beta\left(\alpha H(B_i), \alpha(1 - H(B_i))\right) \tag{2.2}$$

$$E(G(B_i)) = \frac{\alpha H(B_i)}{\alpha H(B_i) + \sum_{j \neq i} \alpha H(B_j)} = H(B_i) \tag{2.3}$$

$$Var(G(B_i)) = \frac{\alpha H(B_i)\left(\sum_{j \neq i} \alpha H(B_j)\right)}{\alpha^2(\alpha + 1)} = \frac{H(B_i)(1 - H(B_i))}{\alpha + 1} \tag{2.4}$$

$$Cov(G(B_i), G(B_j)) = -\frac{\alpha H(B_i)\alpha H(B_j)}{\alpha^2(\alpha + 1)} = -\frac{H(B_i)H(B_j)}{\alpha + 1} \tag{2.5}$$

While 2.2 follows from theorem 3 in the appendix, the rest follow from theorem 5 and by using the fact that $\sum_{i=1}^{K} H(B_i) = 1$. Note that in 2.5, the probabilities assigned to different partitions are correlated negatively, irrespective

of the distance between these partitions. This property is referred to as the "lack of smoothness" [42]. If G were smooth, then two adjacent partitions would be more strongly dependent that two partitions which are far away and this is violated in 2.5. Given this prior distribution (under the current partition), it is fairly straightforward to get the posterior distribution of G, given observed data x_i. Given a data point $x_1 \in B_i$, we have under the prior P, $Pr(x_1|G) = G(B_i)$. The posterior is given by $Pr(G|x_1) \propto Pr(x_1|P)Pr(P)$. Since from 2.1, $G \sim Dir(\alpha H(B_1), \ldots, \alpha H(B_K))$, $Pr(G|x_1)$ is given by:

$$
\begin{aligned}
Pr(G|x_1) &\propto G(B_i) * \frac{\Gamma(\alpha)}{\prod_{j=1}^{K} \Gamma(\alpha H(B_j))} G(B_1)^{(\alpha H(B_1)-1)} \ldots\ldots G(B_K)^{(\alpha H(B_K)-1)} \\
\implies Pr(G|x_1) &\propto G(B_i) * G(B_1)^{(\alpha H(B_1)-1)} \ldots G(B_i)^{(\alpha H(B_i)-1)} \ldots G(B_K)^{(\alpha H(B_K)-1)} \\
\implies Pr(G|x_1) &\propto G(B_1)^{(\alpha H(B_1)-1)} \ldots G(B_i)^{(\alpha H(B_i))} \ldots G(B_K)^{(\alpha H(B_K)-1)} \\
\implies G|x_1 &\sim Dir(\alpha H(B_1), \ldots, \alpha H(B_i) + 1, \ldots, \alpha H(B_K)) \qquad (2.6)
\end{aligned}
$$

Extending the result in 2.6 for N such data points x_i, we get:

$$
G|x_1, \ldots, x_N \sim Dir\left(\alpha H(B_1) + \sum_{i=1}^{N} \delta_{x_i \in B_1}, \ldots, \alpha H(B_K) + \sum_{i=1}^{N} \delta_{x_i \in B_K}\right) \qquad (2.7)
$$

Where $\delta_{x_i \in B_i} = 1$ if $x_i \in B_i$ or 0 otherwise. The result in 2.7 can be extended further [42] to obtain:

$$
G|x_1, \ldots, x_N \sim DP\left(\alpha H + \sum_{i=1}^{N} \delta_{x_i}\right) \qquad (2.8)
$$

Note that in 2.8, the posterior for G is a Dirichlet Process with equal weights placed on the data points x_i. As α approaches 0, the posterior for G approaches

the empirical distribution function of the data [42]. Using the posterior distribution of G in 2.7, we now compute the posterior predictive distribution of a new data point x_{N+1}. Specifically, we compute $Pr(x_{N+1} \in B_k | x_1, \ldots, x_N)$ where B_k is a parition of χ defined earlier. This is given by:

$$Pr(x_{N+1} \in B_k | x_1, \ldots, x_N) = \int Pr(x_{N+1} \in B_k | G, x_1, \ldots, x_N) Pr(G | x_1, \ldots, x_N) dG \quad (2.9)$$

The integrand in Equation 2.9 is straightforward to calculate since conditioning on G, the probability that a new data point will belong to the partition B_k is given by $G(B_k)$. Using this result and the fact that the posterior distribution of G is given by 2.7, we get:

$$Pr(x_{N+1} \in B_k | x_1, \ldots, x_N) = \int G(B_k) * f * \prod_{j=1}^{K} G(B_j)^{(\alpha H(B_j) + \sum_{x_i \in B_j} \delta_{x_i} - 1)} dG$$

$$Pr(x_{N+1} \in B_k | x_1, \ldots, x_N) = \int f * G_{-k}(\cdot) G(B_k)^{(\alpha H(B_k) + \sum_{x_i \in B_k} \delta_{x_i})} dG \quad (2.10)$$

Where $f = \frac{\Gamma(\alpha+N)}{\prod_{j=1}^{K} \Gamma(\alpha H(B_j) + \sum_{x_i \in B_j} \delta_{x_i})}$ and $G_{-k}(\cdot) = \prod_{j \neq k} G(B_j)^{(\alpha H(B_j) + \sum_{x_i \in B_j} \delta_{x_i} - 1)}$. The integral in Equation 2.10, is given by a standard result from Euler [29].

$$Pr(x_{N+1} \in B_k | x_1, \ldots, x_N) = f * \frac{\prod_{j \neq k} \Gamma(\alpha H(B_j) + \sum_{x_i \in B_j} \delta_{x_i}) \, \Gamma(\alpha H(B_k) + \sum_{x_i \in B_k} \delta_{x_i} + 1)}{\Gamma(\alpha + N + 1)}$$

Substiuting for f above, we get:

$$Pr(x_{N+1} \in B_k | x_1, \ldots, x_N) = \frac{\Gamma(\alpha + N)}{\Gamma(\alpha + N + 1)} \frac{\Gamma(\alpha H(B_k) + \sum_{x_i \in B_k} \delta_{x_i} + 1)}{\Gamma(\alpha H(B_k) + \sum_{x_i \in B_k} \delta_{x_i})}$$

$$Pr(x_{N+1} \in B_k | x_1, \ldots, x_N) = \frac{\alpha H(B_k) + \sum_{x_i \in B_k} \delta_{x_i}}{\alpha + N}$$

$$Pr(x_{N+1} \in B_k | x_1, \ldots, x_N) = \frac{\alpha}{\alpha + N} H(B_k) + \frac{N}{\alpha + N} \sum_{x_i \in B_k} \frac{1}{N} \delta_{x_i} \quad (2.11)$$

The result in 2.11 can be further extended and re-written [42, 104] as:

$$x_{N+1} | x_1, \ldots, x_N \sim \frac{\alpha}{\alpha + N} H(B_k) + \frac{N}{\alpha + N} \sum_{x_i} \frac{1}{N} \delta_{x_i} \quad (2.12)$$

In Equation 2.12, $\delta_{x_i} = 1$ if $x_{N+1} = x_i$, or 0 otherwise. The form of the predictive distribution in 2.12 gives us a strategy to draw data points from the base distribution H of the Dirichlet Process directly, a procedure called the "Polya urn scheme" [17, 108, 104]. Following the convention in Tomlinson [108], we can draw data points directly from a Dirichlet Process with concentration α and H, without needing the random distribution G, as shown below:

$$
\begin{aligned}
x_1 &\sim H \\
x_2 &\sim \frac{\alpha}{\alpha+1}H + \frac{1}{\alpha+1}\delta_{x_1} \\
x_3 &\sim \frac{\alpha}{\alpha+2}H + \frac{1}{\alpha+2}(\delta_{x_1} + \delta_{x_2}) \\
&\vdots \\
x_{N+1} &\sim \frac{\alpha}{\alpha+N}H + \frac{1}{\alpha+N}\sum_{i=1}^{N}\delta_{x_i}
\end{aligned}
$$

The first data point is drawn from the base distribution directly. The next data point is now drawn from H with probability $\frac{\alpha}{\alpha+1}$, and it equals the previously drawn data point (x_1) with probability $\frac{1}{\alpha+1}$.This sampling scheme shows that there is a non-zero probability of data points being repeated, and this probability depends on the number of data points being drawn (N) and the concentration parameter (α). If α is large enough, initially, we might get a larger number of unique draws [17, 108, 104, 42], but as N increases, we might get repeated draws. This data drawing procedure also highlights another feature of the resultant draws: if a particular data point, say x_i, is drawn more than the others, then the next data point drawn, say x_{N+1} will more likely equal x_i. This phenomenon is referred to as the "rich gets richer" property [42] and also has a culinary analogy built around it, called the "chinese restaurant process" [104]. More importantly, the draws from a random probability measure G with a Dirichlet Process prior exhibit a clustering property (i.e. out of the N x_i's drawn

from G, only $k < N$ of those are unique values. Tomlinson [108] and Antoniak [8] derive the probability of k unique values in a sample of draws from G)

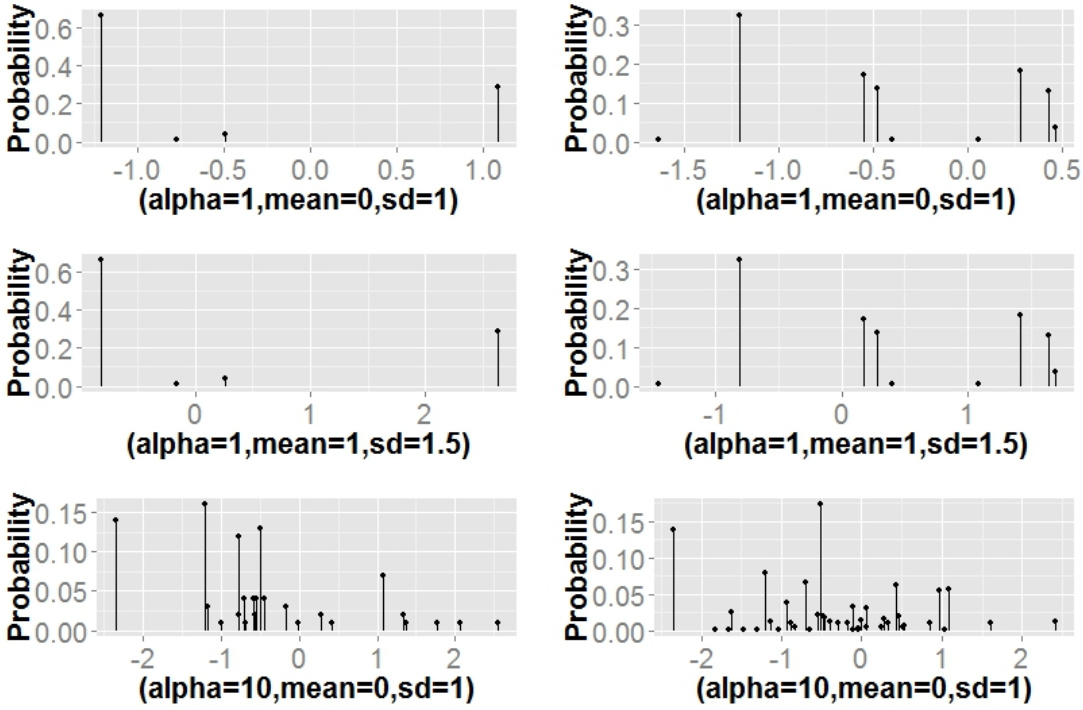

Figure 2.3: Draws From a Dirichlet Process

Figure 2.3 shows draws of data from a dirichlet process for different values of N, α and H, with the associated probability masses placed at the data points. In the plots shown, $N = 100$ data points were generated for each plot in the first column, and $N = 1000$ were generated for each plot in the second column. For the plots in the first row in Figure 2.3, $\alpha = 1$ and $H \sim N(0, 1)$ for both. It is clear that since $\alpha < N$ in this case, as we keep drawing more data points, we tend to pick out already existing data. For the plots in the second row, $\alpha = 1, H \sim N(1, 1.5^2)$ for both. To ensure comparability of plots from the first row to the second, data were generating by using the same random seed, due to which the plots in the first and second row look completely similar, except for a subtle

difference: the values of the x-axis "shift" rightwards, which makes sense since the mean of the base distribution H used to generate these data is higher than in the previous case. In the third row, $H \sim N(0, 1)$, but $\alpha = 10$ now. Since a higher value of α is used, more unique values of the data are generated, in contrast to the previous cases.

Since draws from a Dirichlet Process tend to be repeated, the process might not be ideally suited to model data directly, but can be used to model cluster memberships in mixture models [108, 104]. Instead of placing a Dirichlet Process Prior directly on the data, this prior is placed on the parameters of the distribution which is assumed to generate the x_i's. Figure 2.4 shows a simple Dirichlet Process Mixture model (alternatively referred to as a Bayesian Nonparametric mixture model). The only difference between the DAG in Figure 2.2, and this one is the θ_i, on which the Dirichlet Process prior is placed. In Figure 2.4, G is a random probability measure which is distributed according to a Dirichlet Process with concentration parameter α and base distribution H. The draws from G are represented by θ_i. Each θ_i has an associated data point x_i, that is generated from θ_i based on some distributional assumption. As an example, x_i can be modeled as being normally distributed with parameters $\theta_i = (\mu_i, \Sigma_i)$. Given the discreteness property of G discussed earlier, not all values of θ_i are unique. Let the unique values of $\{\theta_i\}_{i=1}^{N}$ be denoted by $\{\phi_l\}_{l=1}^{K}$, where K is the total number of unique values (or alternately, the total number of clusters). Hence, different values of x_i could "share" the value of the underlying parameter ϕ_l that generates them, and these particular values belong to the same cluster (ϕ_l is the cluster specific parameter). In addition, the values of x_i that do not share the same value of underlying ϕ_l belong to different clusters.

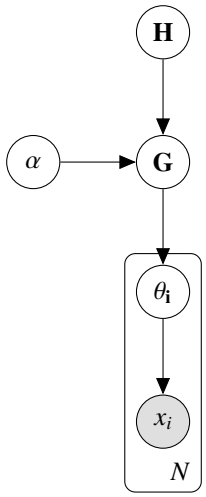

Figure 2.4: A Simple Dirichlet Process Mixture Model

In model form, this amounts to the following structure being placed on the data x_i:

$$G \sim DP(\alpha, H)$$

$$\theta_i | G \sim G$$

$$x_i | \theta_i \sim F(\theta_i)$$

Where $F(\cdot)$ could be any distribution function (usually a Normal distribution). As discussed earlier, the number of unique cluster specific parameters generated from G can be controlled by changing the value of α. When α is larger, more clusters (i.e. more ϕ_l values or unique values of θ_i) are created and vice versa. When applying these models to data, the number of clusters are chosen during the estimation process. This is convenient, since in typical frequentist models applied to mixture data, the researcher starts with a guess for the number of underlying clusters and then estimates the model [1]. This process

[1]However, as Orbanz [84] mentions, Bayesian nonparametric mixture models update themselves (by appropriately changing the number of clusters) as more data become available, and

is repeated till the "best" number of clusters is obtained [16]. The number of clusters determined by the Bayesian nonparametric mixture model depends on α, on which a Gamma prior is placed [35] or usually chosen to be equal to 1 [81, 23, 69].

The applications of the Dirichlet Process to mixture models was first discussed by Antoniak [8]. Since then, there have been many applications and extensions of the Dirichlet Process to mixture models in Statistics [35, 50, 42], Machine Learning [104], Neuroscience [43] and Marketing [63, 23, 69].

Two broad approaches exist in the literature of applying Dirichlet Process based models to data. Escobar and West [36] formally outline a procedure to incorporate Bayesian Nonparametric techniques in general hierarchical Bayesian models, while Neal [81] presents a comprehensive collection of MCMC methods for the implementation of Dirichlet Process models. These algorithms were also used in marketing applications by Kim, Menzefricke and Feinberg [63] and Burda, Harding and Hausman [23]. A key step in these algorithms, which Neal [81] discusses at length in his paper, is to treat the random distribution G as something that needs to be "integrated out" [2]. Due to this, the algorithms described in Neal [81] need some form of book keeping at each iteration of the sampling process, to check the number of "active clusters" (clusters which are non empty). In addition, since G is effectively removed from the sampling process, there is no inference being made on the structure of G (for example, we don't know the probability mass Gplaces on each of the θ's drawn from it).

Another approach to inference in Dirichlet Process mixture models is due to

would be misspecified when applied to data where the number of underlying clusters is assumed to be finite

[2]Quotations, since by definition, G is discrete

Sethuraman [93]. He defines a stick breaking construction of the draws from a Dirichlet Process. Since $G \sim DP(\alpha, H)$ is a discrete distribution, it places a some probability mass on points drawn from H, usually referred to as atoms [84]. These probability masses should sum to 1, so each one of these treated as part of a stick of unit length, so to speak. Each of these probabilities, once generated, is "broken off" from this stick of unit length. Each probability mass is constructed as follows:

$$\begin{aligned} V_1 &\sim Beta(1, \alpha) & P_1 &= V_1 \\ V_2 &\sim Beta(1, \alpha) & P_2 &= V_2(1 - V_1) \\ &\vdots \\ V_k &\sim Beta(1, \alpha) & P_k &= V_k \prod_{i=1}^{k-1} (1 - V_i) \\ &\vdots \end{aligned} \tag{2.13}$$

The V_i's are called the stick breaking weights [42]. Given these probability masses, $G \sim DP(\alpha, H)$ is then given by:

$$G(\cdot) = \sum_{j=1}^{\infty} P_j \delta_{\phi_j}(\cdot) \qquad \phi_j | H \sim H \tag{2.14}$$

Where ϕ_j are the unique values of θ_i, and $x_i \sim F(\theta_i)$ This representation enables us to make inferences on G directly, rather than getting rid of it [42]. By definition, the stick breaking probabilities P_k in Equation 2.13 should be computed for large k (as $k \to \infty$ in 2.13). Ishwaran and Zarepour [58] show that these stick breaking probabilities need not be computed for large k (in 2.13 note that in principle, the probabilites P_k are computed for $k \to \infty$). They show that when $\alpha = 1$, restricting k to 25 approximates G well (they show via plots that the probability left over, what they refer to as the "tail probability", is a very small number).

Hence, the stick breaking probabilities for $k = 25$ can be constructed by setting $P_k = 1 - \sum_{i=1}^{k-1} P_i$, which simplifies to:

$$
\begin{aligned}
& P_k = 1 - (P_1 + P_2 + \cdots + P_{k-1}) \\
\implies & P_k = 1 - \left(V_1 + V_2(1 - V_1) + \cdots + V_{k-1} \prod_{i=1}^{k-2} (1 - V_i) \right) \\
\implies & P_k = (1 - V_1) - \left(V_2(1 - V_1) + \cdots + V_{24} \prod_{i=1}^{k-2} (1 - V_i) \right) \\
\implies & P_k = (1 - V_1) - V_2(1 - V_1) - \left(V_3(1 - V_1)(1 - V_2) \cdots + V_{k-1} \prod_{i=1}^{k-2} (1 - V_i) \right) \\
\implies & P_k = (1 - V_1)(1 - V_2) - \left(V_3(1 - V_1)(1 - V_2) \cdots + V_{k-1} \prod_{i=1}^{k-2} (1 - V_i) \right) \\
& \vdots \\
\implies & P_k = \prod_{i=1}^{k-1} (1 - V_i) \qquad (2.15)
\end{aligned}
$$

In addition to the advantage of making inferences on G, Ishwaran and James [57] also argue that Gibbs sampling methods for stick breaking priors (a more general class of priors of which the one just described is a member) are simpler to understand and implement. Recently, in marketing literature, Li and Ansari [69] use a stick breaking approach to sample from a centered dirichlet process [118] in their application. In the rest of this thesis, Gibbs sampling algorithms for stick breaking priors are used.

2.1 Synthetic Data Applications

This section shows how Bayesian nonparametric models can be applied to model (1) mixture data and (2) heterogeneity in a mixed logit model, using syn-

thetic data. As we will see, the second example is an extension of the first.

2.1.1 Mixture Model

We first generate 1000 data points from a mixture of 3 gaussian distributions (of varying means and variances), as shown in Equation 2.16 [3]:

$$x \sim 0.1 * N(-1, 0.2^2) + 0.5 * N(0, 1) + 0.4 * N(1, 0.4^2) \tag{2.16}$$

We augment a Bayesian nonparametric mixture model to these data. A DAG representing the model is shown in Figure 2.5.

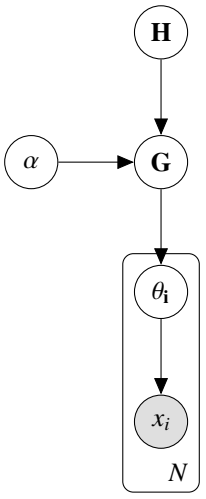

Figure 2.5: Dirichlet Process Mixture Model Applied to Synthetic Data

The priors and hyperparameters assumed are shown below:

$$\alpha = 1$$

$$H = N(\mu_0, \sigma^2/\kappa_0)\, InvGamma(\sigma^2|a/2, b/2)$$

[3] This is a slightly modified version of an exercise in chapter 23 of Gelman, Carlin, Stern, Dunson, Vehtari and Rubin [42]

Where $\mu_0 = 0$, $\kappa_0 = 0.01$ and $a = b = 1$. Since H is a Normal-Gamma prior, the θ's drawn from it contain a mean and variance parameter, i.e. $\theta = (\mu, \sigma^2)$. For the stick breaking construction of G, we choose $k = 25$ clusters, and construct the stick breaking probabilities as described in 2.15. To signify that these probabilities $P = \{P_1, \ldots, P_{25}\}$ are generated from a stick breaking process, the expression $P \sim stick(\alpha)$ is used henceforth. Define $z = \{1, 2, \ldots, k\}$. These are the cluster indicators. Note that by definition, $Pr(z = i) = P_i$, where P_i is the stick breaking probability as defined in 2.15. For convenience, let z_i denote the cluster a given data point x_i belongs to, and let ϕ_k be the unique cluster specific parameter for cluster k, drawn directly from the base distribution H. Then the underlying cluster specific parameter that generates x_i is given by ϕ_{z_i}. In Equation form:

$$\begin{aligned} \phi_k | H &\sim H \\ P &\sim stick(\alpha) \\ z_i &\sim discrete(1, 2, \cdots, 25) \\ x_i | \{\phi_k\}_{k=1}^{25}, z_i &\sim N(\phi_{z_i}) \qquad\qquad \phi_{z_i} = (\mu_{z_i}, \sigma^2_{z_i}) \end{aligned} \tag{2.17}$$

In Equation 2.17, z_i equals one of $\{1, 2, \ldots, k\}$ with probabilities given by the vector P. x_i is chosen to be distributed normally so to maintain the conjugacy of the prior and posterior. Before taking the model to data, we randomly assign data points to one of the 25 clusters. The Gibbs algorithm steps used for inference are shown below:

Algorithm 1 *At each iteration m,*

- *For each data point x_i, compute:*

$$Pr(z_i = k| \dots) = \frac{P_k \, p(x_i|\mu_j, \sigma_j^2)}{\sum_{l=1}^{25} P_l \, p(x_i|\mu_l, \sigma_l^2)}$$

  for $k = \{1, 2, \dots, 25\}$, where $p(\cdot|\mu, \sigma^2)$ is the normal density with mean μ and variance σ^2

- *Compute n_k which is the number of data points x_i assigned to each of the clusters $k = \{1, 2, \dots, 25\}$. This is $n_k = \sum_{i=1}^{N} \delta_{z_i}(k)$, where $\delta_{z_i}(k) = 1$ if $z_i = k$ or 0 otherwise.*

- *Get cluster specific posterior parameters. For each $k = \{1, 2, \dots, 25\}$*

$$a_{n_k} = a + n_k, \kappa_{n_k} = \kappa_0 + n_k$$
$$\hat{\mu}_{n_k} = \frac{\kappa_0 \mu_0 + n_k \bar{x}_{n_k}}{\kappa_{n_k}}, b_{n_k} = b + \frac{\kappa_0 n_k}{\kappa_{n_k}} (\bar{x}_{n_k} - \mu_0)^2 + \sum_{z_i = k} (x_i - \bar{x}_{n_k})^2$$

  where $\bar{x}_{n_k} = \frac{1}{n_k} \sum_{z_i = k} x_i$ is the cluster specific mean.

- *For each cluster k, draw the posterior means and variances:*

$$\sigma^2_{n_k} \sim InvGamma(a_{n_k}/2, b_{n_k}/2), \, \mu_{n_k} \sim N(\hat{\mu}_{n_k}, \sigma^2_{n_k}/\kappa_{n_k})$$

- *Update the stick breaking weights (V_k), using the following result:*

$$V_k \sim Beta(1 + n_k, \alpha + \sum_{l=k+1}^{25} n_l)$$

  Generate the stick breaking probabilities

$$P_1 = V_1, P_2 = V_2(1 - V_1), \dots, P_{25} = \prod_{l=1}^{24} (1 - V_l)$$

The results of the estimation are shown in Table 2.1. The actual number of clusters and the data points assigned in each are given in Table 2.2.

Table 2.1: Estimated Cluster Properties (SD = Standard Deviation)

Cluster	N	Probability	Posterior Mean	Posterior SD
1.00	489.00	0.51	0.27	0.92
2.00	349.00	0.31	0.93	0.39
3.00	143.00	0.16	-1.08	0.20
4.00	19.00	0.02	-1.79	0.62

Table 2.2: Actual Cluster Properties (SD = Standard Deviation)

Cluster	N	Probability	Mean	SD
1	110	0.10	-1.00	0.20
2	493	0.50	0.00	1.00
3	397	0.40	1.00	0.40

Though the labels of the clusters are switched (a common issue in mixture modeling [42], and an extra cluster is created (cluster 4), it is evident that the model does a good job of recovering the underlying clusters in the data. The estimated cluster means, standard deviations and cluster probabilities are all close to the actual values. A plot of the actual density of the data (red) and the estimated density(dotted blue) is shown in Figure 2.6, which again suggests a good fit.

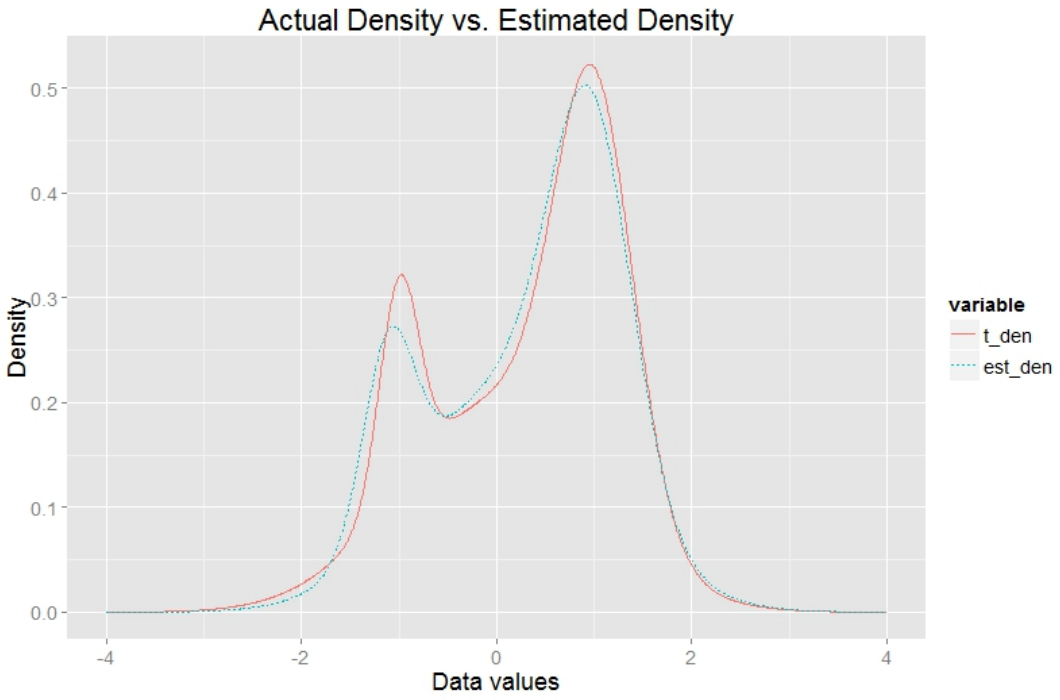

Figure 2.6: Plot of Actual vs. Estimated Density

To check for convergence, this model was estimated on the data from 20 different starting points (i.e. with different initial random cluster assignments in the data). The log-likelihood plot from the exercise (in Figure 2.7) shows that irrespective of the starting point, the Gibbs sampling algorithm converges to the same solution (the plot is shown for only the first 50 iterations since the model converges very quickly). The posterior parameters was also exhibit good mixing properties. A few sample trace plots are shown in Figure 2.8.

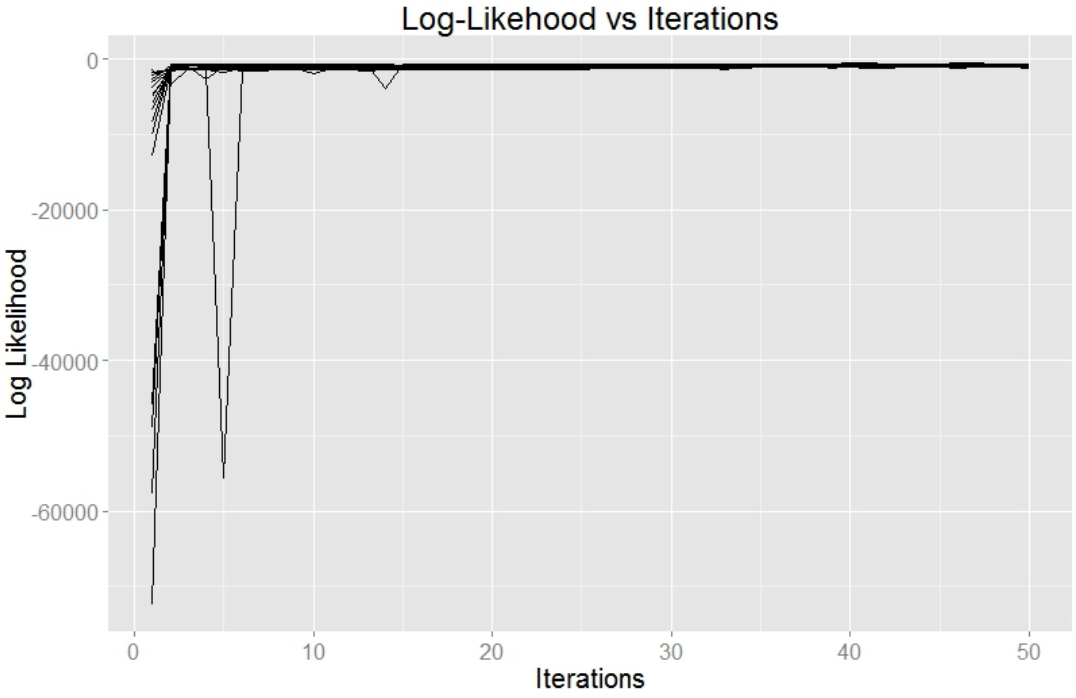

Figure 2.7: Plot of Log-Likelihood for Different Starting Points

2.1.2 Mixed Logit Model

In Marketing literature, the Dirichlet Process mixture model has been used to model heterogeneity in models of discrete choice [63, 23, 69]. In this section, we show how this is done in practice, using a very simple data generating process in 1D space. An extension to multiple dimensions is straightforward. To generate the choice data, 1500 individual level coefficients were generated as a mixture of 3 Gaussians as shown below:

$$\beta \sim 0.3 * N(1, 0.2^2) + 0.2 * N(2, 1) + 0.5 * N(3, 0.4^2) \tag{2.18}$$

Next, covariates were generated for 4 alternatives and for 100 choice situations per individual, from a uniform distribution on the range $[1, 3]$. Next, for all these choice situations, the utilties were computed (after generating an appropriate

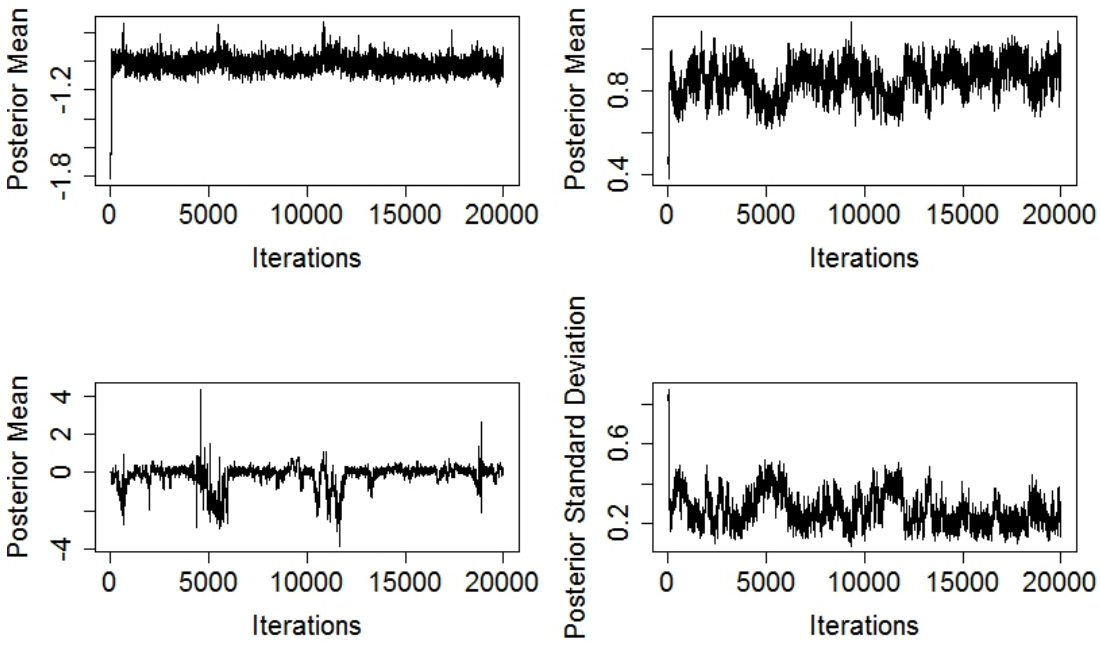

Figure 2.8: Sample Trace Plots

number of extreme value error terms and then adding them to the deterministic part of the utility) and the choices were determined (this becomes the dependent variable in the estimation exercise). Figure 2.9 shows the DAG representing the model for data augmentation (where $T = 100$ and $N = 1500$). Each β_i, or individual level coefficient, is an outcome of a Dirichlet Process mixture model, but unlike in the previous data example, we don't observe them. We only observe the outcomes y_{it}. An advantage of using a Hierarchical Bayesian framework to estimate a mixed logit model lies in the fact that the MCMC algorithm picks out the individual level parameters during estimation [110].

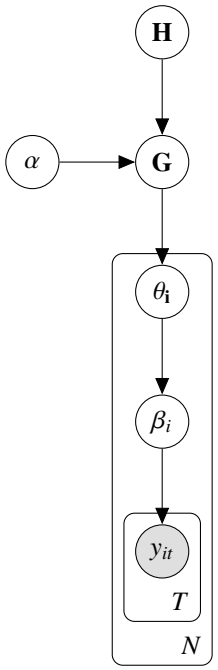

Figure 2.9: A Mixed Logit Model with a Dirichlet Process Mixture Prior on the Coefficients

In this exercise, we assume $\beta_i \sim F(\theta_{z_i})$, where $F(\cdot)$ is the normal distribution, as before. Since this is not conjugate with the form of the likelihood (the logistic function), we use a Random Walk Metropolis Hastings step to get a new draw of β_i [91]. Once we get draws of the β_i, these become the "data" for the mixture model, and all the steps in Algorithm 1 apply. The algorithm used for this case is Algorithm 2.

Algorithm 2 *At each iteration m,*

- *For each individual i, compute the likelihood:* $L_i(y|\beta_i^{m-1}) = \prod_{t=1}^{T} L_{it}$
- *Get a new draw of* β_i*:*

$$\beta_i^m = \beta_i^{m-1} + \rho\sigma_{z_i}\eta$$

  where $\eta \sim N(0,1)$, $\rho = 2.93$*, from [91] and* $\beta_i \sim N(\mu_{z_i}, \sigma_{z_i}^2)$

- *Compute the ratio:*

$$MHR = \frac{L_i(y|\beta_i^m)p(\beta_i^m|\mu_{z_i}, \sigma_{z_i}^2)}{L_i(y|\beta_i^{m-1})p(\beta_i^{m-1}|\mu_{z_i}, \sigma_{z_i}^2)}$$

  where $p(\cdot|\theta_{z_i})$ *is the normal density evaluated at* θ_{z_i}

- *Get* $u \sim U[0,1]$*. Accept* β_i^m *as a new draw only if* $u < MHR$*, otherwise set* $\beta_i^m = \beta_i^{m-1}$
- *Once* $\{\beta_i^m\}_{i=1}^{1500}$ *are obtained, these are modeled using a Dirichlet Process mixture, so follow all the steps in Algorithm 1 once (replace* β_i *instead of* x_i *in Algorithm 1).*

Repeat the above process till convergence.

The results of the estimation are shown in Table 2.3. Comparing the results in this with the actual values in Table 2.4, we find that though the model recovers the correct number of underlying clusters, the posterior mean for cluster 3 is way off (1.13) from the actual value (2). This could be attributed to the fact that the data generated from this cluster was incorrectly assigned to the other two clusters (since for this cluster, the variance is high compared to the other two). Kim et al. [63] face similar problems in their application, and suggest some post

processing to fix this cluster membership problem. However, a density plot of the estimated and actual data (Figure 2.10) shows that the estimated density approximates the actual density quite well. In addition, the trace plots of the individual level coefficients show good mixing (Figure 2.11), and the average acceptance ratio is 0.2639, which is also good [91].

Table 2.3: Estimated Cluster Properties (SD = Standard Deviation)

Cluster	N	Probability	Posterior Mean	Posterior SD
1.00	531.00	0.60	2.97	0.47
2.00	903.00	0.36	1.02	0.29
3.00	66.00	0.04	1.13	1.38

Table 2.4: Actual Cluster Properties (SD = Standard Deviation)

Cluster	N	Probability	Mean	SD
1	458	0.30	1.00	0.20
2	291	0.20	2.00	1.00
3	751	0.50	3.00	0.40

To conclude, the Dirichlet Process mixture model is a very flexible technique which can be used to model heterogeneity and can be easily "embedded" into commonly used marketing models [36]. However, as seen in the examples above, there are are some issues with the posterior estimates recovered from the data [63, 84]. To alleviate some of these issues and broaden the applicabilities of this framework, the Hierarchical Dirichlet Process [105] is introduced in the next chapter, and it is applied in the context of modeling Heterogeneity across multiple markets.

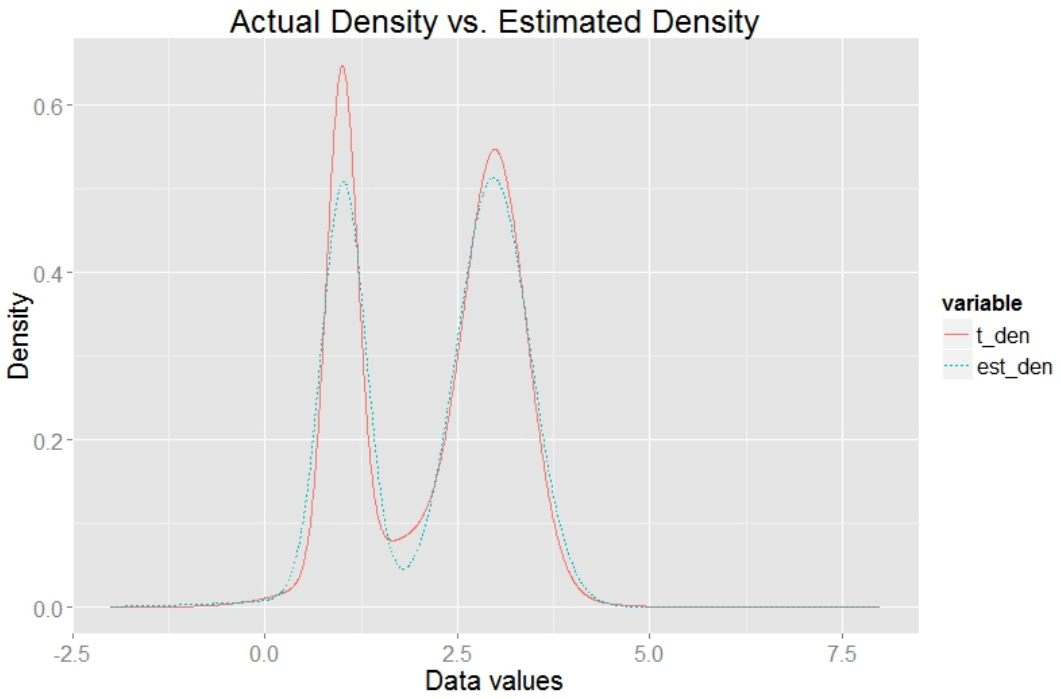

Figure 2.10: Actual vs Estimated Density for Individual Level Coefficients

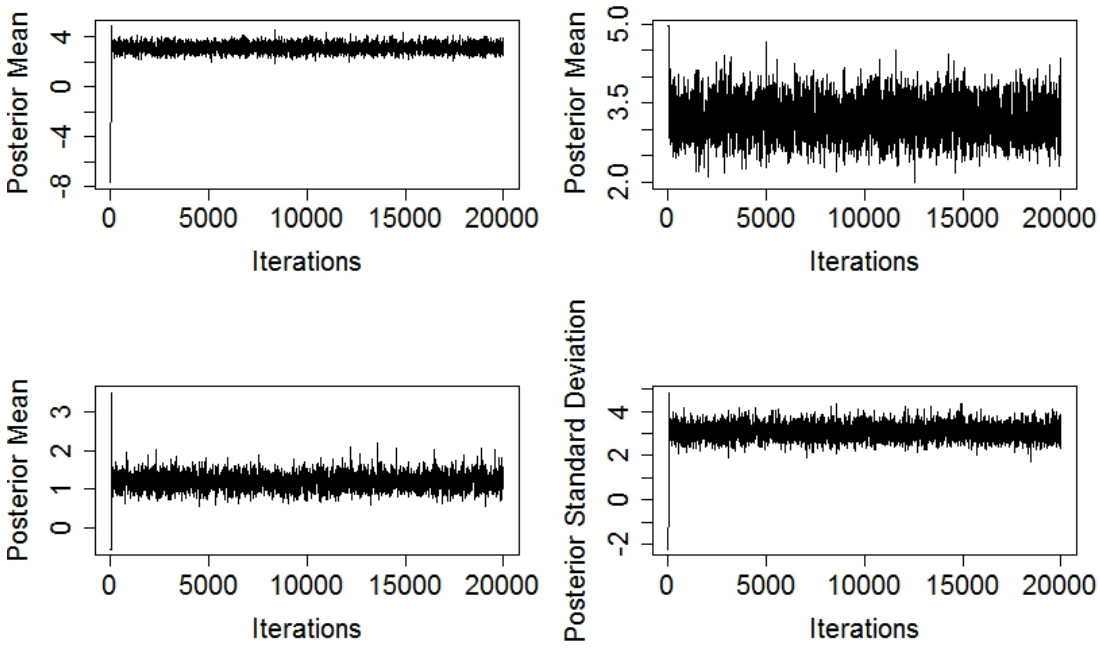

Figure 2.11: Trace Plots for Select Coefficients

CHAPTER 3

MODELING HETEROGENEITY ACROSS MULTIPLE MARKETS

3.1 Introduction and Literature Review

A central idea in the formulation of marketing strategy for brands is segmentation. Market or consumer segmentation involves grouping consumers such that consumers within a group are relatively homogeneous with respect to their brand preferences and their responsiveness to the marketing efforts of the firm. The segmentation exercise is usually followed by targeting and positioning activities, wherein the firm chooses one or more target segments to serve, and formulates marketing plans to address those segments. Most of the conceptual or theoretical literature in marketing implicitly assumes that the geographic segmentation, targeting and positioning exercise is being conducted for a single market. In practice, however, many brands are marketed in multiple regional or national markets and there may be substantive differences in the segmentation structure across these markets. Then a key question facing marketing managers is that of standardization versus adaptation of the positioning strategies across these markets.

In the context of international marketing, Szymanski et al. [103] note that the marketing strategy formulation process comprises a series of decisions pertaining to the business's (1) strategic orientation (standardization versus adaptation), (2) desired degree of standardization of the strategic resource mix (i.e. pattern of resource allocation among advertising, promotions, personal selling, etc.), and (3) the desired degree of standardization of the strategy content (i.e. decisions on product positioning, brand name, content of advertisements, etc.).

These same decisions also arise in the management of a national brand with regard to differences between regional markets.

In this chapter we propose a Hierarchical Bayesian model that can help the brand manager in developing segmentation, targeting, and positioning strategies for multiple markets. The model presents itself as a natural extension of extant work in modeling heterogeneity as an outcome of a finite mixture model. At one end of the spectrum, we impose homogeneity in consumer responses across markets. Such models are consistent with an approach of complete standardization of both the strategy content and the resource allocation across markets. At the other end of the spectrum are models that assume that the segmentation structure is idiosyncratic to each market. Intermediate options impose homogeneity in segment characteristics (such as brand preferences and marketing mix responsiveness), but allow for heterogeneity in the segment sizes across markets. These models are consistent with an approach of standardizing the strategic content across markets, but adapting the resource mix to the relative segment sizes in each market (or groups of markets).

Many different bases of segmentation have been proposed in the marketing literature [116]. In this chapter we focus on preference and response based segmentation [114], an approach that is especially relevant in frequently purchased packaged goods categories where market-place purchasing data can be used to infer differences in brand preferences and marketing-mix responsiveness of consumers. The data are collected via optical scanning of purchases of households that participate in large-scale longitudinal panels. In addition to purchases, the market research data providers also collect data on the prices and promotional conditions in the stores in which the panelist households made their pur-

chases. Consequently, the impact of marketing efforts of firms on households' purchases can be modeled and estimated. In the past, latent class models were used to implement preference and response based segmentation.

Latent class models [60, 26, 61] assume that differences between consumers in a market with respect to preferences for brands and their responsiveness to marketing activities can be adequately described by a small number of discrete groups or segments of consumers. Each consumer is assumed to be a member of one of these segments. The objective of the estimation methodology is to uncover the preferences and responsiveness of each of these groups, and the relative sizes of the groups. The latent class multinomial logit model and its extensions have been widely applied to the segmentation problem in the marketing literature. We discuss here a few of these studies to serve as a backdrop for the contribution made by our paper to the literature.

Kamakura and Russell [60] developed preference and response segments using brand choices of a panel of households. Chintagunta [25] extended the model to obtain brand positions on a product-market map in addition to the distribution of preferences across households, while accounting for the effects of marketing variables on choice behavior. Using a nested logit formulation, Bucklin and Gupta [21] identified preference and response segments on the basis of consumers brand choice and category purchase incidence behavior. Kamakura, Kim and Lee [59] used a finite mixture of nested logit models to identify segments that differ not only in their brand preferences and marketing mix responsiveness, but also in terms of their choice process. Gupta and Chintagunta [51] and Kamakura et al.[62] incorporated observable descriptors of consumers into the model, to simultaneously profile the preference and response segments.

All these applications have focussed on segmentation of consumers within a single market. In a survey paper, Steenkamp and Ter Hofstede, [100] examine the different methods used by researchers in the past to tackle the issue of segmentation across multiple markets. They acknowledge that finite mixture models present a powerful approach to model different segments, specifically mentioning the desirable properties of Bayesian sampling-based approaches. The key advantage of using a Bayesian approach is the capability to model within segment heterogeneity, something that is lacking in latent class models. In addition, Allenby and Rossi, [5], claim that a continuous representation of heterogeneity would be more useful, and Andrews, Ainslie, and Currim [7] find that a logit model with a continuous representation of heterogeneity tends to fit the data better. A drawback of a continuous representation of heterogeneity is the modeler's reliance on correctly specifying the underlying probability distribution [100]. However, that concern has been alleviated with the advent of Bayesian Nonparametric methods [35, 36, 104], and recent work in marketing has made use of these techniques [63, 69]. However, to the best of our knowledge, these concepts have not been used to model heterogeneity across multiple markets.

A key advantage of Nonparametric Bayesian methods is the fact that the researcher can leave the number of latent clusters unspecified and the model will decide the number of clusters based on the data. It is worth noting that this feature of Nonparametric Bayesian models is not a panacea, since researchers in machine learning have noted that these models tend to be misspecified [84] if there is reason to believe that the number of underlying clusters in the data are finite. A common issue is the creation of "extra" clusters by the model which might not exist in the data. Kim, Menzefricke and Feinberg [63] study this issue

using a Dirichlet Process prior to model heterogeneity in their data, and they find that some post processing of clusters after estimation is needed. To handle this issue while also accounting for the fact that there are multiple markets in our data, we use the Hierarchical Dirichlet Process [105] (HDP henceforth) as a prior on the coefficients in our model.

The HDP was developed in the machine learning literature to model topics in different document types (blogs, news, etc.), while allowing the documents to share common topics (a news article about college sports and college education might differ theme-wise, but might contain common topics, such as education, for example) but also differ in the content. This is what Teh et al. [105] call allowing the different groups (document types) to "share statistical strength". In the context of our application, the different markets play the role of the different document types, where customer types (topics) could be shared between markets. This feature of the HDP helps researchers find a suitable "middle-ground" between having to ignore heterogeneity across markets and modeling each market separately.

In section 3.2 we describe the HDP, which will be used to estimate the coefficients in the proposed model and the estimation algorithm. In section 3.3 we describe an application of the proposed model to household scanner panel data from 73 regional markets in the US. In section 3.4 we discuss estimation results and managerial implications of the empirical application. We summarize and conclude in section 3.5.

3.2 Model Details

3.2.1 The Hierarchical Dirichlet Process

The HDP was first discussed by Teh et al. [105], as a method to model multiple groups of documents, allowing the topics within them to be modeled separately while allowing these groups to "share statistical strength". This is accomplished by letting these groups share the "atoms" drawn from the base distribution. A DAG of the proposed HDP model is shown in Figure 3.1. The HDP simply

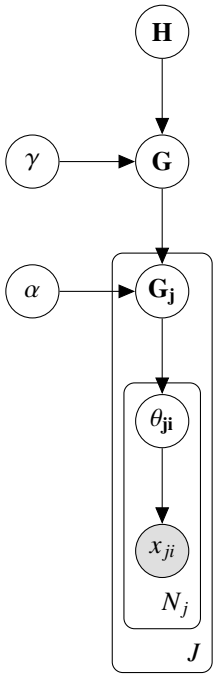

Figure 3.1: A Hierarchical Dirichlet Process Model

extends the Dirichlet Process discussed in the previous chapter by placing a Dirichlet Process Prior on the base distribution (G) that generates the unknown prior distribution (G_j), which is placed on the underlying parameters (θ_{ji}) that generates the observed data (x_{ji}). Similar to the stick breaking construction [93]

for the Dirichlet Process mentioned in chapter 2, Teh et al. [105] discuss a stick breaking construction of the HDP, which is explained here. We start with the random probability measure $G \sim DP(\gamma, H)$. A stick breaking construction of G would entail the following steps:

$$P \sim stick(\gamma) \tag{3.1}$$

$$\phi_k | H \sim H$$

$$G = \sum_{k=1}^{\infty} P_k \delta_{\phi_k}$$

Where the notation in Equation 3.1 indicates that P is generated by using a stick breaking construction, as described in chapter 2. The ϕ_k's are drawn directly from H, and these become the "atoms" of G [105]. Moving below G in the DAG in Figure 3.1, we now look at each of the G_j's ($j = 1$ to J). For each j, $G_j \sim DP(\alpha, G)$. Note that the concentration parameter α for each j is the same, and while this can be varied by j, we keep it fixed here (changing α by j changes the way G_j varies around G, as discussed in Teh et al. [105]). By construction, G_j and G share the atoms ϕ_k generated from H [105]. Let $\pi_j = (\pi_{jk})_{k=1}^{\infty}$ be the stick breaking weights of G_j, then by the stick breaking construction [93], we have:

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k} \tag{3.2}$$

We set about finding the distribution of the weights π_{jk} in Equation 3.2. Following the strategy in Teh et al. [105], we take a partition of the probability space on which G, and by construction G_j, are defined. Denote this partition as $B = \{B_1, \ldots, B_r\}$. From the Equation 2.1 defined in chapter 2, we have:

$$(G_j(B_1), G_j(B_2), \ldots, G_j(B_r)) \sim Dir(\alpha G(B_1), \alpha G(B_2), \ldots, \alpha G(B_K)) \tag{3.3}$$

Note that $G_j(B_l)$ adds the probabilities (π_{jk}) of its atoms (ϕ_k) lying in the partition B_l (the same concept applies to G, just that the probabilities in this case are given

by P_k). Let $K_l = \{k : \phi_k \in B_l\}$ for $l = 1, \ldots, r$. Then, $G_j(B_l) = \sum_{k \in K_l} \pi_{jk}$ and $G(B_l) = \sum_{k \in K_l} P_k$. Putting these results in 3.3 gives us:

$$(\sum_{k \in K_1} \pi_{jk}, \ldots, \sum_{k \in K_r} \pi_{jk}) \sim Dir(\alpha \sum_{k \in K_1} P_k, \ldots, \alpha \sum_{k \in K_r} P_k) \tag{3.4}$$

Extending the result in Equation 3.4 further, we get $\pi_j \sim DP(\alpha, P)$. Now that the distribution of the π_{ji}'s is obtained, we comment on the nature of the draws θ_{ji} (moving further down the DAG in Figure 3.1), from G_j. As θ_{ji}'s are draws from a random probability measure with a Dirichlet Process prior, they exhibit a clustering property, as discussed in chapter 2, and are sampled from the set of unique atoms $(\phi_k)_{k=1}^{\infty}$ with probabilty $\pi_j = (\pi_{jk})_{k=1}^{\infty}$. Each θ_{ji} is now the underlying parameter for the observed data x_{ji}. Denote z_{ji} as an indicator for the cluster to which x_{ji} belongs. By definition, if $z_{ji} = 3$, then $\theta_{ji} = \phi_{z_{ji}} = \phi_3$ and $x_{ji}|z_{ji}, (\phi_k)_{k=1}^{\infty} \sim F(\phi_{z_{ji}}) = F(z_3)$. Now we put all these results together to get:

$$\begin{aligned} P|\gamma &\sim stick(\gamma) & \phi_k|H &\sim H \\ \pi_j|\alpha, P &\sim DP(\alpha, P) & z_{ji}|\pi_j &\sim \pi_j \\ & x_{ji}|z_{ji}, (\phi_k)_{k=1}^{\infty} \sim F(\phi_{z_{ji}}) & & \end{aligned} \tag{3.5}$$

To get a stick breaking construction formulation for the π_{ji}'s from equation 3.4, we use standard properties of the Dirichlet Distribution, derived in the appendix. Consider a partition $\{B_1, B_2, B_3\}$, such that $(1, \ldots, k-1) \in B_1$, $k \in B_2$ and $(k+1, k+2, \ldots) \in B_3$ [105]. The distribution of these partitions is derived (by using the same idea as in equation 3.4) below:

$$(\sum_{l=1}^{k-1} \pi_{jl}, \pi_{jk}, \sum_{l=k+1}^{\infty} \pi_{jl}) \sim Dir(\alpha \sum_{l=1}^{k-1} P_l, \alpha P_k, \alpha \sum_{l=k+1}^{\infty} P_l) \tag{3.6}$$

Removing the first element, and by using Theorem 4 in the appendix, we get:

$$\frac{1}{1 - \sum_{l=1}^{k-1} \pi_{jl}} (\pi_{jk}, \sum_{l=k+1}^{\infty} \pi_{jl}) \sim Dir(\alpha P_k, \alpha \sum_{l=k+1}^{\infty} P_l) \tag{3.7}$$

Defining $\pi'_{jk} = \frac{\pi_{jk}}{1-\sum_{l=1}^{k-1}}$, we have:

$$
\begin{aligned}
(\pi'_{jk}, \sum_{l=k+1}^{\infty} \pi'_{jl}) &\sim Dir(\alpha P_k, \alpha \sum_{l=k+1}^{\infty} P_l) \\
\implies (\pi'_{jk}, 1-\pi'_{jk}) &\sim Dir(\alpha P_k, \alpha \sum_{l=k+1}^{\infty} P_l) \qquad (3.8) \\
\implies (\pi'_{jk}, 1-\pi'_{jk}) &\sim Dir\left(\alpha P_k, \alpha\left(1 - \sum_{l=1}^{k-1} P_l\right)\right) \\
\implies \pi'_{jk} &\sim Beta\left(\alpha P_k, \alpha\left(1 - \sum_{l=1}^{k-1} P_l\right)\right) \qquad (3.9)
\end{aligned}
$$

Where in Equation 3.8, we use the fact that $\sum_{l=k+1}^{\infty} \pi'_{jl} = 1 - \pi'_{jk}$ and in equation 3.9, that $1 - \sum_{l=1}^{k-1} P_l = \sum_{l=k+1}^{\infty} P_l$. To get π_{jk}, we generate π'_{jk} using Equation 3.9 and then set $\pi_{jk} = \pi'_{jk} \prod_{l=1}^{k-1}(1-\pi'_{jl})$ [105].

3.2.2 Estimation Algorithm

We place a HDP mixture prior on the coefficients of a standard logit model. This is done to capture the across-market variation in heterogeneity. This is an extension of the mixed logit model example studied in Chapter 2. As mentioned in the previous chapter, directly applying a Dirichlet Process mixture prior on the coefficients might not be appropriate in this situation. Orbanz [84] discusses how a Dirichlet Process mixture model could be misspecified when the number of underlying clusters is finite and unknown since more clusters are created as we keep collecting data. By placing a HDP mixture prior on the coefficients, we methodologically restrict the number of clusters created [105], while also acknowledging across-market (or group) variation. The HDP creates "market types" (indexed by j) based on the household level preference and response

coefficients β_{ji} that generate the observed choices in the data y_{jit}. A directed acyclic graph representing the model is shown in Figure 3.2. Moving from top to bottom, we have G, which is a random probability measure generated from a Dirichlet Process with base distribution H and concentration parameter γ. For each "market type" j, $G_j \sim DP(\alpha, G)$ is generated, and this becomes the base distribution from which θ_{ji} is generated. β_{ji} is modeled as distributed normally with parameters θ_{ji}. These β_{ji} are then used in the logit model to compute the choice probability. The algorithm used to estimate the posterior cluster and market-type specific parameters is a slight modification of algorithm 2 in Chapter 2. We have an extra step here, where we generate the G_j's for each market type j. The estimation algorithm is described below. As discussed in Train [110], household level parameters are modeled, since Hierarchical Bayesian methods are easier to augment with data in that setup.

Algorithm 3 *At each iteration m,*

- *For each individual i, compute the likelihood:* $L_{ji}(y|\beta_{ji}^{m-1}) = \prod_{t=1}^{T} L_{jit}$
- *Get a new draw of* β_{ji}*:*

$$\beta_{ji}^{m} = \beta_{ji}^{m-1} + \rho\sigma_{z_{ji}}\eta$$

  where $\eta \sim N(0, I_s)$*,* $\rho = 2.93/\sqrt{s}$*, where* $s = dim(\beta_{ji})$ *from [91] and* $\beta_{ji} \sim N(\mu_{z_{ji}}, \sigma^2_{z_{ji}})$

- *Compute the ratio:*

$$MHR = \frac{L_{ji}(y|\beta_i^m)p(\beta_i^m|\mu_{z_{ji}}, \sigma^2_{z_{ji}})}{L_{ji}(y|\beta_{ji}^{m-1})p(\beta_{ji}^{m-1}|\mu_{z_{ji}}, \sigma^2_{z_{ji}})}$$

  where $p(\cdot|\theta_{ji})$ *is the normal density evaluated at* θ_{ji}

- *Get* $u \sim U[0, 1]$*. Accept* β_{ji}^m *as a new draw only if* $u < MHR$*, otherwise set* $\beta_{ji}^m = \beta_{ji}^{m-1}$
- *Once* $\{\beta_{ji}^m\}_{i=1}^{10462}$ *are obtained, these are modeled using a HDP mixture, so follow all the steps in Algorithm 1 once. When using the stick breaking construction for* π_j*, use the beta distribution derived in Equation 3.9*
- *In addition, update* $P \sim stick(\gamma)$ *using the procedure outlined in Teh et al. [105]*

Repeat the above process till convergence.

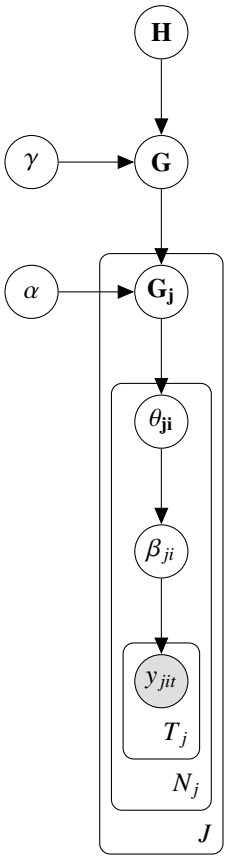

Figure 3.2: A Mixed Logit Model with a Hierarchical Dirichlet Process Mixture Prior on the Coefficients

3.3 Empirical Application

The proposed model was applied to the ACNielsen Company's national scanner panel data on purchases of wrapped cheese slices (known as"singles" in industry parlance). The national scanner panel is located in 73 regional markets in the US . We include for analysis purchases of a 10,462 households over a period of twelve quarters. The households in the data made 136,772 purchases in the category over this period. We restricted our analysis to the eight largest brand-sizes in the category; these items account for over 80% of category sales volume. Products are in two sizes - 12 ounces and 16 ounces. The eight items include six

national brand items and two private label items. The national brand items also differ on the fat-content attribute. The six items include three fat-free and light items (known in the industry as better for you (BFY) products), and four regular items. Detailed information on five marketing mix variables is available in the data to explain brand choices of consumers. These variables are:

1. Regular price (equalized to 12 oz. pack size)
2. Price cut (equalized to 12 oz. pack size). This variable is defined as (Regular price - Shelf price).
3. In-store Display only (indicator variable).
4. Feature ad only (indicator variable).
5. Feature ad with in-store display (indicator variable).

Descriptive statistics of the data are in Table 3.1[1] The four items of National Brand 1 jointly hold 52.17% market share, the private label items account for 32.05%, and the balance is accounted for by National Brand 2. We are unable to disclose brand names because of confidentiality requirements of the data provider. The three BFY items hold 20.15% market share. The national brands are generally priced higher than the private label brands. Also, the BFY items have a higher regular price on average. We find that price and non-price promotion activity occurs on both the national brands and the private labels.

[1]Share was computed based on 136,772 purchases in the data made by 10,462 households in 73 markets. Regular Price and Price Cut entries are normalized per 12 oz. pack. Average Price cut is based only on occasions when there was a price cut. Display Only, Feature Only and Display + Feature: percentage of category purchase occasions.

Table 3.1: Descriptive Statistics of Cheese Slices Data

Brand Type	Share (%)	Avg. Reg. Price (per 12 oz.) ($)	Avg. Price Cut (per 12 oz.) ($)	Dis Only (%)	Ftr Only (%)	Dis /Ftr (%)
Natl.Br.1 Reg. 12 oz.	28.09	2.13	0.114	1.377	3.351	1.489
Natl.Br.1 Reg. 16 oz.	9.44	2.37	0.098	0.209	0.707	0.151
Natl.Br.1 Fat Free 12 oz.	11.52	2.77	0.067	0.029	0.537	0.036
Natl.Br.1 Light 12 oz.	3.11	2.68	0.069	0.011	0.167	0.004
Natl.Br.2 Fat Free 12 oz.	5.52	2.72	0.040	0.004	0.243	0.003
Natl.Br.2 Reg. 12 oz.	10.26	2.17	0.134	0.763	0.931	0.121
Pvt. Lbl Reg. 12 oz.	21.84	1.74	0.095	0.780	1.527	0.429
Pvt. Lbl Reg. 16 oz.	10.21	1.82	0.064	0.075	0.650	0.088

N = 136,772 purchase occasions

3.4 Results

The model was estimated using Algorithm 3. For each household i, we estimate the posterior distribution of the brand preference and marketing variables response coefficients. Results are shown in Tables 3.2 and 3.3 . As can be seen, the model "chooses" $J = 3$ market types, with varying segment specific means and segment probabilities (π_j). For each segment, the segment specific posterior means are provided with the 95% credible interval. There are no segment specific posterior estimates for the Private Label Reg. 16 oz. pack since that is taken as a base item.

Table 3.2: Segment Probabilities for Different Market Types

Market Type (Type Size)	Segment I	Segment II	Segment III	Segment IV
Market Type I (0.604)	0.19	0.42	0.28	0.11
Market Type II (0.161)	0.33	0.51	0.16	-
Market Type III (0.235)	0.31	0.28	0.22	0.19

The model recovers "shared segments" from all market types and the segment probability placed on each of these segments. For ease of exposition, each segment is assigned a label, based on the characterisitics of the segment specific means recovered by the model. Segments I and IV prefer the National Brand Regular items (Segment I has stronger preferences than Segment IV) and so are called NBR1 and NBR2. Segment II prefers the private label items (since the

Table 3.3: Posterior Segment Means for Different Market Types

Covariates	Segment I	Segment II	Segment III	Segment IV
Natl.Br.1 Reg. 12	3.77 [3.46, 4.08]	-0.94 [-1.20, -0.68]	1.55 [1.31, 1.79]	2.09 [2.03, 2.17]
Natl.Br.1 Reg. 16	3.45 [3.18, 3.72]	-2.32 [-2.66, -1.98]	0.35 [0.22, 0.48]	1.11 [0.93, 1.29]
Natl.Br.1 F.F. 12	0.57 [0.29, 0.85]	-2.29 [-2.42, -2.18]	4.65 [4.51, 4.79]	0.18 [-0.11, 0.47]
Natl.Br.1 Light 12	-0.82 [-0.96, -0.68]	-3.53 [-3.76, -3.29]	2.97 [2.79, 3.15]	-0.31 [-0.55, -0.07]
Natl.Br.2 F.F. 12	-0.74 [-0.95, -0.53]	-3.33 [-3.64, -3.02]	3.39 [3.34, 3.46]	-0.36 [-0.50, -0.22]
Natl.Br.2 Reg. 12	2.71 [2.49, 2.93]	-1.45 [-1.63, -1.27]	0.61 [0.27, 0.95]	0.90 [0.81, 0.99]
Pvt. Lbl Reg. 12	0.92 [0.57, 1.27]	0.89 [0.77, 1.03]	1.02 [0.87, 1.17]	0.48 [0.25, 0.71]
Pvt. Lbl Reg. 16	-	-	-	-
Regular Price	-0.35 [-0.57, -0.13]	-0.98 [-1.27, -0.69]	-0.03 [-0.041, -0.019]	-0.61 [-0.72, -0.50]
Price Cut	1.21 [0.94, 1.48]	1.79 [1.44, 2.14]	0.68 [0.49, 0.87]	1.51 [1.38, 1.64]
Display Only	1.14 [1.01, 1.27]	1.03 [0.83, 1.23]	0.79 [0.71, 0.87]	0.56 [0.31, 0.81]
Feature Only	1.36 [1.22, 1.50]	1.04 [0.93, 1.15]	0.79 [0.68, 0.90]	0.69 [0.59, 0.80]
Display + Feature	2.08 [2.00, 2.16]	2.52 [2.23, 2.81]	2.01 [1.81, 2.23]	1.12 [0.81, 1.43]

fixed effect coefficient posterior means are all negative) and is called PL. Segment III shows the highest preference for the "Better For You" items, and is called BFY.

Comparing the posterior means of the segment specific coefficients, we find substantial differences in consumer preferences. As discussed above, each of the segments (with the exception of NBR1 and NBR2) prefers a different type of product. In addition, we also find that the PL segment is most price sensitive among all segments, which is understandable since this segment of consumers prefers the cheaper PL product (similarly, this segment is also sensitive to price cuts). The PL segment is followed by the NBR2 and NBR1 segments in terms of price sensitivity, and the BFY segment is least price sensitive.

The NBR1 and NBR2 segments represent consumers with varying preferences over the choice of NBR products. The HDP model "picked out" enough variation in the coefficients of the mixed logit model to create two segments that preferred NBR products, but to a varying extent. This can also be attributed to the fact that the National Brands account for 67.94% of all sales in the product category, allowing the co-existence of segments NBR1 and NBR2, which prefer the same product type, one more (NBR1) than the other (NBR2). Note that NBR2 does not prefer National Brand Regular products as strongly as NBR1, and are more sensitive to price and price cuts (when compared to NBR1). By comparing the coefficient of "Display + Feature" with the separate coefficients of "Display only" and "Feature only" we note that there is a positive interaction effect in segments PL and BFY, but a negative interaction effect in segments NBR1 and NBR2. The results indicate that within each market-type the nature of the interaction effect may vary by segment.

When examining the different market types suggested by the model, we interpret the probabilities placed by the model on the different segments (π_j) to be segment sizes, and we find substantial heterogeneity in these sizes between market types. In fact, we find that in market type II, the NBR2 segment does not exist (is of size 0). While the size of the PL segment is almost the same in market types I and II, it is nearly halved in market type III. Moreover, in market type III, there are more households that prefer the national brand regular type products, when compared to the Private Label or Better For You category. Market type II clearly has a bigger chunk of consumers who prefer the Private Label product, and can be labeled the price sensitive market (owing to the price sensitive characteristic of the PL segment). Market Type I is more of a mixed bag: though the PL segment is the largest, this market type has the largest BFY segment of all market types. We also estimate the "size" of the market types, as shown in Table 3.2. Market type I is the largest of all occuring market types, followed by market type III and market type II.

Since our model allows segment sizes to vary by market type, we're able to identify different market types, and also the variation in preferences within each of these market types. From a managerial perspective, this is important, since it gives the manager the flexibility to change the marketing mix based on market type and segment specific characteristics. Based on the the market type and segment properties, it is quite possible that the optimal marketing plan could lead to very different decisions for different segments. The results of the model help the manager compare segments by market type, and customize targeting activities accordingly.

3.5 Conclusion

In this chapter we propose a method to model heterogeneity to assist brand managers who need to adapt or standardize marketing strategy to multiple markets. The proposed models are generalizations of the latent class discrete choice model that is used commonly in the marketing econometrics literature. We place a HDP mixture prior on the coefficients of the logit model. This model presents a compromise between fitting a model where we ignore heterogeneity across markets and fitting a separate model by market while placing a Dirichlet Process prior on the model coefficients for heterogeneity. The proposed model places a restriction on the extent of heterogeneity between markets permitted in the segment characteristics and the segment probabilities. The restrictions enhance comparability of the segmentation structure between markets. The models are applied to brand choices of cheese slices made by households in 73 regional markets in the U.S. The results support the proposed model in terms of fit to the data, and help demonstrate the managerial value of the model in understanding similarities and differences across markets in the segmentation structure. There are plenty of ways in which the work presented here can be extended. While we model heterogeneity at the household level, we do not capture the time changing preferences of these households (β_{ji} is assumed to be constant with time). In addition, we also enforce the concentration parameter (α) for all market types j to be the same. Further extensions of this model could model the time variation of household level parameters, while also allowing for data driven α, by placing a diffuse Gamma prior on it and updating it as an extra step in the algorithm presented here [105].

CHAPTER 4

FROM SALES LEADS TO CUSTOMERS: TRACKING CONVERSION USING ONLINE AND OFFLINE ACTIVITY DATA IN A B2B SETTING

4.1 Introduction

A crucial aspect of a firms customer acquisition is to select the "right" customer to acquire. While the RFM criterion is frequently used for deciding whether or not a prospective customer is worth pursuing [67], it assumes the availability of prospective customer transaction data; this is usually not the case. With the proliferation of the Internet and the surge of Internet-based businesses in the last 20 years or so, empirical research has been growing on analyzing web browsing behavior to predict which customer is more likely to buy a product. Bucklin and Sismeiro [22] survey several papers that study clickstream data to predict customer conversion. Whereas these papers have contributed immensely to understanding consumer web browsing behavior, the extant literature has two gaps[1] : (1) The combination of offline and online activities is unavailable ; and (2) No focus on a B2B context [70]. This current research intends to bridge that gap, by using clickstream data on leads, in addition to their offline activities, to predict the probability of their conversion into a prospective customer, in a B2B context. Srinivasan [98] identifies several aspects in which B2B marketing differs from B2C marketing, and the lack of well-defined marketing metrics to measure the effectiveness of such marketing activities. In this paper, we account for these differences and study the effectiveness of several online and offline

[1]The sole exception to this claim is the paper by (Wiesel, Pauwels, and Arts [115]), who have both online and offline lead request and purchase data in a B2B context and study the effect of these marketing activities on profit

marketing activities in converting the sales leads into prospective customers for a mid-sized B2B software company. Based on lead response behavior to marketing activity, we predict the probability that a lead will convert (or become a customer of the firm), and also identify the marketing activities that increase the aforementioned probability.

This work helps B2B marketing managers in the two major ways:

1. Shift focus on the marketing activities that increase the chances of a lead becoming a customer.
2. Reduce the uncertainty in attributing a specific marketing action to a change in lead behavior, facilitating easier ROI calculation (in the case where a lead eventually becomes a customer).

Even though this study uses data from a mid-size B2B software firm, the issues faced by the firm in determining which marketing activities are effective in lead conversion are shared amongst other B2B firms in various industries [98]. The framework employed in this research project is not limited to the context of the firm being studied, and can be applied to any other B2B firm that keeps a database of online and offline activities of their leads.

4.2 Theoretical Background and Literature Review

As mentioned earlier, previous research that has studied customer acquisition using clickstream data has been focused primarily in a B2C setting [22]. Since the process of customer acquisition (via marketing communication) in a B2C

context varies markedly from that in a B2B context [98], the modeling approaches used in the B2C context must be modified appropriately. Table 4.1 (adapted from Anderson and Narus [6]) presents a generic example of the type of marketing communications that are used to acquire customers in a B2B context, which is similar to the strategy followed by the firm being studied in this project. In the B2C context, a lot of the work studied the antecedents of pur-

Table 4.1: Marketing Communications Mix of B2B firms

Communication Objectives	Potential Customers	Communication Tools
Awareness	Leads	Advertising, direct mail, publicity, industry conferences
Interest	Inquiries	Brochures, videos, recorded demonstrations, websites, trade shows
Evaluation	Prospects	Telemarketing, Field Sales Visits
Trial	New Customers	Field Sales Visits, inside sales calls
Purchase	Established Customers	Transactional and relationship teams, key account management, thought leadership

chase behavior on ecommerce websites. To ascertain the contribution of different types of predictors to the purchasing behavior at an online store, Van Den Poel and Buckinx, [112] include variables from four different categories in predicting online-purchasing behavior (general clickstream behavior, more detailed clickstream information, customer demographics and historical purchase behavior). They use backward and forward variable selection techniques as described in Furnival and Wilson [41] and they find predictors from all four cat-

egories prominently featured in the best subset of predictor variables (hence indicating the importance of clickstream data). Sismeiro and Bucklin [94] model online buying with clickstream data by framing the buying process as a completion of a sequence of tasks, instead of a single stage decision process, to find that visitor browsing experiences and navigational behavior predict "task completion" well. However, past research found that repeat visits to a website was not diagnostic of buying propensity (as found in Moe and Fader [75]s setting) and offering sophisticated decision aids did not guarantee increased conversion rates. Moe [74] modeled observed choices in two choice stages, products viewed and products purchased in clickstream data. She found that product attributes considered in stage 1 were different from that in stage 2, and that consumers tend to use simpler decision rules based on a subset of attributes in earlier stages. These "criterion" attributes along with the preference parameters estimated from her model, would prove important in designing targeting strategies. Moe and Fader [76] find that modeling customer conversion behavior as a combination of visit effects and purchasing threshold has better statistical properties when compared to logistic regression models used in past research[2] . Park and Fader [87] model browsing behavior at multiple websites to test whether using clickstream data from only one website could lead to biased results. They find that analysis of clickstream data collected from a single website lead to biased estimates, as they are inherently incomplete and do not capture shopping behavior across multiple websites. Montgomery, Li, Srinivasan, and Liechty [78] study detailed webpage transition choices (path data) and found that navigational behavior of customers was best explained by two modes of behavior (deliberation and browsing). Allowing for customers to switch between these

[2]Visit effects = store visits (that play different roles in the purchasing process). Purchasing threshold = psychological resistance to online purchasing, which depends on customer experience with the purchase process.

two modes in a single browsing session was found to be a better representation of customer browsing behavior.

Another gap in this literature is the absence of offline data. In the context of this research, offline data on a lead refers to any activities of the lead that did not involve the internet. Wiesel, Pauwels, and Arts [115], do have online and offline information during different stages of the purchase funnel, but the objectives of their research are different from ours, in that they study the impact of marketing communications on pre-defined purchase funnel metrics and firm profit. In our work, we have access to the offline activities of leads, and we incorporate that information into our modeling framework, with the goal of predicting the probability of conversion of a lead as a customer of the company.

As Srinivasan [98] points out, the "substantial differences" between the B2B and B2C marketplaces merits a sophisticated examination of the B2B marketplace, so that marketers are more careful when measuring the "performance of their marketing actions". Srinivasan [98] identifies key differences between the B2B and B2C marketplaces, of which the most relevant ones to this project are the following:

1. There are many stages in the buying process;
2. There are many participants in the buying process;
3. Usually, there are complex integrated marketing communications involved; and
4. Buying situations vary considerably.

Extending the research work done in customer acquisition from a B2C to a

B2B context will involve incorporating these key differences into our modeling framework. Specifically, we address the following:

1. Which marketing activities (online or offline) increase the chances of a lead to transition between the many Stages of a buying process? ;
2. Which marketing activities increase the speed of transition in different Stages of the buying process? ;
3. Given the outcome of a lead conversion, what is the ROI on various online and offline marketing activities pursued by the firm? ; and
4. Given multiple leads belonging to the same company, how does the firm prioritize which leads to pursue?

4.3 Data

We use data from a collaborating firm which is a mid-sized B2B firm, that offers a unique software to implement the PAXOS algorithm [65, 66]. A practical implementation of this algorithm is considered impossible by computer science academics even today. Their software is used mainly for Application Lifecycle Management (ALM), though in the recent past, the firm has also offered reliable solutions for replication and backup of big data. The ALM software industry currently represents a $ 5.45 billion dollar market [3], with a Compound Annual Growth Rate (CAGR) of 4.51% [4]. Big Data, on the other hand, is a $ 125 billion dollar market, steadily growing in size each year (Press 2014). The firm collects

[3] investors.rallydev.com/download/Rally+Q1+Investor+Presentation-3.pdf

[4] http://www.sandlerresearch.org/global-application-lifecycle-management-market-2014-2018.html

data on a set of prospective customers called "leads" whose basic information is known (demographic and professional). They have weblogs of lead visits, and also information on their offline interactions with the firms sales and marketing activities For example, they collect data on whether these leads attended webinars conducted by the company, visited the firm information booth at tech conferences, filled information request forms and called for product information, etc. Hence, in addition to online information, the firm also has data on the offline "path" of their leads. Figure 4.1 gives a breakdown of lead activities data to which the firm has access. Data were obtained for leads for the period

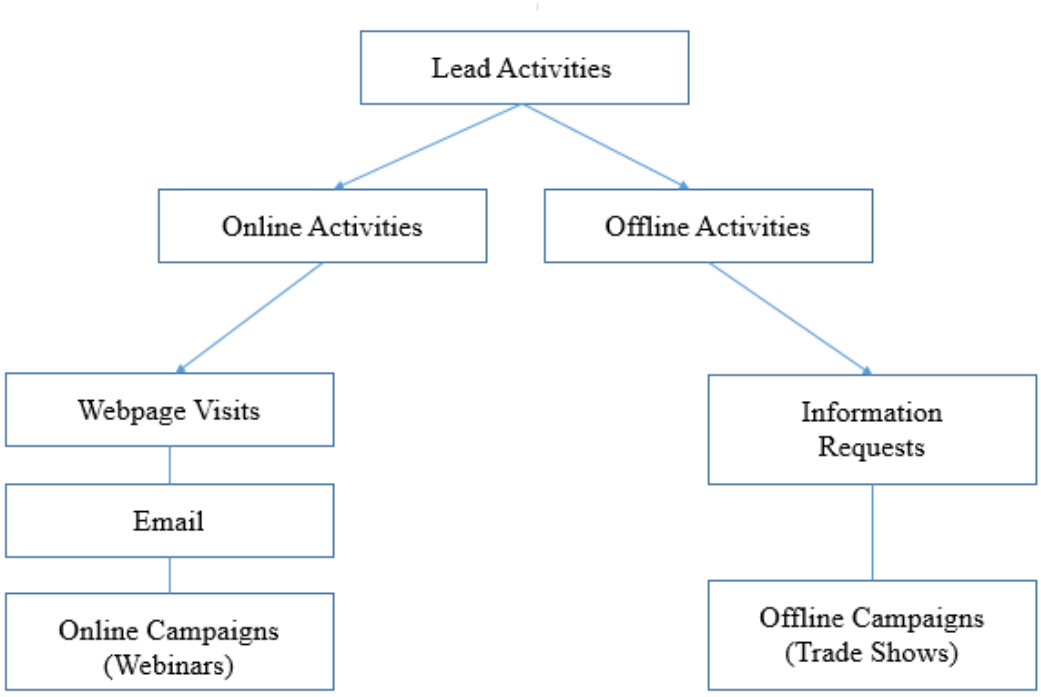

Figure 4.1: A Breakdown of Lead Activity Data

28^{th} June 2012 to 19^{th} September 2015. The database consisted of 0.54 million leads, belonging to $10,000$ firms around the world. After cleaning the data to remove inactive leads (leads who havent shown any interest in the marketing

communications of the firm in the period of observation) and duplicated entries (the lead database possessed by the firm was put together from different lead sources) the number of leads boils down to 135,245 [5] . Table 4.2 shows a sample of the online activity information that the company has on a particular lead during a week in the month of July 2015. In this particular data sample, the firm keeps track of the fact that the lead opened an email (sent by the firm), and registered for a webinar which was scheduled to be held a week later. The firm also keeps a record of the leads attendance of the webinar. Similarly, the collaborating firm also has information on lead attendance at trade shows and offline information requests, to name a few. In order to build a comprehensive model, we study both the online and offline data for model free evidence. The final model also needs to take into the account the selling process employed by the firm being studied. We begin by analyzing the clickstream data of the leads in the data set, a sample of which was shown in Table 4.2. Following the convention in Montgomery et al. (2004), we classify the various webpages present into 4 broad categories: Home, Account, Downloads and Information. The category Home refers to the home page of the company. Account refers to all the personalized webpages that appear once a lead logs in using his/her account information on the company webpage. Download refers to all the webpages where a free download of a trial version of the software product (binary henceforth) is available (in the data, a visit to a download page resulted in a binary download 94% of the time). Information refers to all the webpages containing software descriptions, news, blogs, etc. The categorization of webpages differs a little from that of Montgomery et al. (2004), since the firm we are studying does not sell its product directly on its webpage, nor does it sell different cate-

[5]The loss of a large number of leads can be explained due to the fact that most of the data in the company database were outdated since they were purchased from another lead database firm in 2010.

Table 4.2: A Sample of Activities for a Given Lead during the month of July 2015 (Campaign Name was anonymized by company request)

Date-Time	Campaign Name	Activity Type
2015-07-09 15:09:51	Email Batch Program-2076-send-email-campaign	Visit Web Page
2015-07-09 15:26:12	Interesting Moments.Opens Email	Interesting Moment
2015-07-09 15:26:16	Interesting Moments.Clicks Link in Email	Interesting Moment
2015-07-09 23:07:17		Add to YYYY Campaign
2015-07-09 23:07:33	07162015 - Git - Webinar - Modernizing Development Workflow with Git.01 - Registers	Interesting Moment
2015-07-16 23:05:53	07162015 - Git - Webinar - Modernizing Development Workflow with Git.03 - Attends On Demand	Interesting Moment

gories of products (unlike in their paper), hence the lack of categories such as shopping cart, order or category. To understand what pages on the company website drive traffic, the monthly average number of page hits was computed, and a plot of the same is shown in Figure 4.2 (with the links anonymized). The category-wise information for (monthly) average webpage traffic is also given

in Table 4.3. Table 4.3 also contains category-wise information for average webpage traffic for the leads who converted, and for those who didn't.

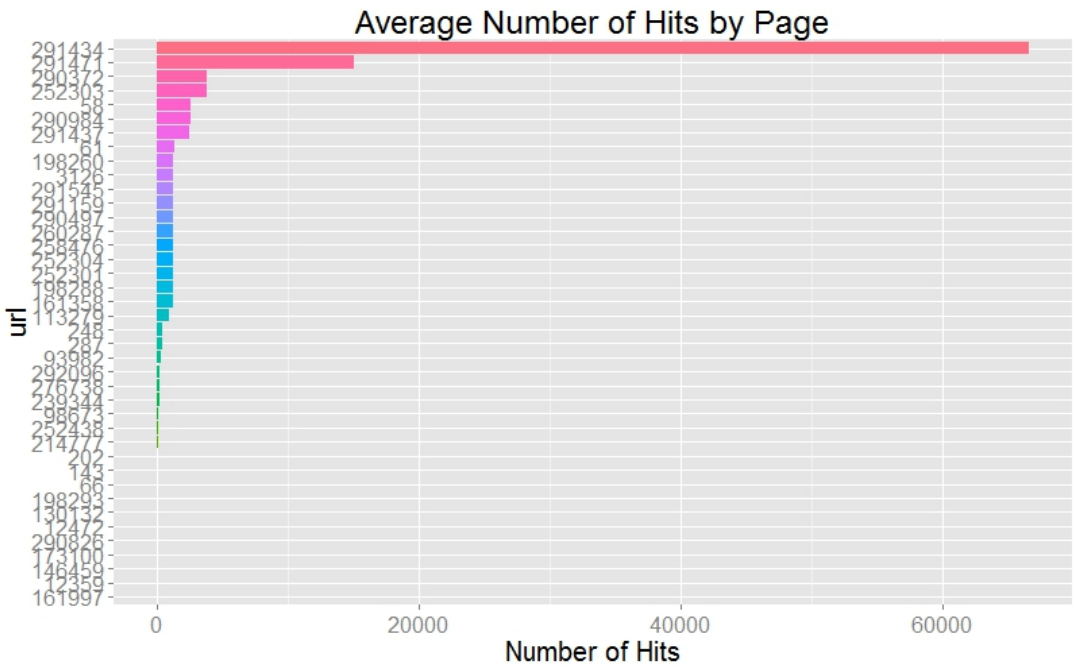

Figure 4.2: Average Number of Webpage Hits

Table 4.3: Average Monthly Webpage Traffic in the Data

Category	Abbreviation	Page Requests (%)	Page Requests (Converted) (%)	Page Requests (Not Converted) (%)
Home	H	7.79	1.9	8.3
Account	A	9.89	7.13	10.13
Downloads	D	11.32	16.71	10.85
Information	I	71	74.26	70.72

From Table 4.3, it is evident that pages in the information category tend

to dominate the other categories in terms of the average number of visits per month. The leads in the database appear to be more interested in reading about the product first, followed by downloading a trial version of the software offered by the company.

In addition, we also check if the page visit behavior of leads who became clients of the firm differed from those who didnt. The third column of Table 4.3 shows the average monthly webpage visits of leads who eventually became clients of the firm in the sample period contained in the data, while the fourth column shows the same for those who didn't. Comparing the different columns, we notice that for the most part, the percentage of page requests to pages in the Information and Account categories doesnt change by much. However, for the leads who converted, the percentage of downloads is significantly higher. At the same time, the percentage of page requests in the home category is significantly lower for the same leads. This comparison appears to indicate that the leads who converted spend more time on the downloads page than on the home page, and that the number of downloads could be an antecedent of lead conversion.

In addition to the average page hits, we also explored the individual page visits of each of the leads, to see if there was any web browsing path information, which could be exploited for more insights [55, 78]. However, the average number of distinct page categories visited by leads per session was 2, with a standard deviation of 0.2497, indicating that there is not much path information to be exploited in this case. A possible explanation for the low number of distinct category page views could be due to the design of the website [78], which may not be conducive to leads exploring other parts of the website. Another explanation could be that the leads in our data knew exactly what they were

looking for (given that the firm sells a product that needs a certain level of skill with programming software), and hence there is not much exploration on the company website.

Another important difference of this setting from that of Montgomery et al. [78], has to do with the website "form-filling" activity that all of the leads have to do, to download a binary. The leads interested in downloading the product need to fill out an information form online (i.e. professional information such as employment, contact details, etc.). This helps the company store new lead information and track future visits by this lead to the website. This is also useful in collecting information on those leads who havent created an account with the company website.

As mentioned earlier, the firm also has offline data of lead activities. These activities include trade show attendance, conferences and offline information requests. To understand which of these activities are important in terms of driving product sales, we again compare the offline activities of the leads who converted in our data set to those leads who didnt. Table 4.4 shows the mean (yearly) participation of these leads in the various offline activities offered by the company. When mean attendance is compared between the two types of leads, it is evident that the leads more likely to convert put in the effort needed to gather more information (i.e. attending the trade show at a different city and seek out the company's information desk). While the mean attendance of the leads who became clients is understandably higher, it is also instructive to take a closer look at the mean information requests. The leads who eventually became clients tended to request information twice as much as those who didnt. But we also note that when moving from the leads who converted to those who

Table 4.4: Average Yearly Offline Activity Participation

Offline Activity	Mean Attendance (Converted)	Standard Deviation (Converted)	Mean Attendance (Not Converted)	Standard Deviation (Not Converted)
Trade Shows	7.21	1.14	1.36	0.33
Conferences	4.21	2.23	2.3	0.99
Information Requests	23.85	12.62	11.08	4.1

do not, the relative reduction in the mean number of information requests is less when compared to the mean attendance numbers. This can be attributed to the fact that information requests are easier to carry out (either over the phone or by setting up a time to talk via email), when compared to attending trade shows and conferences. These metrics show that trade show attendance could be a good indicator of a lead becoming a prospective client. We find confirmation of this finding in previous literature on the value of trade shows. Gopalakrishna, Lilien, Williams and Sequeira [49] find that trade shows help in generating product awareness and interest, while providing positive economic returns to the firm. Gopalakrishna and Lilien [48] also note that trade shows are more effective than advertising or personal selling in generating product awareness. These findings indicate the importance (to the firm) of trade show attendance by leads, so as to spread awareness of the product in the early stages.

The model free evidence guides our decision to include certain variables in our model (the number of binary downloads, trade show and conference attendance and information requests). In addition, we also need to take into consideration the selling process followed by the firm being studied, since that also affects the choice of the model to be used for this application. The selling pro-

cess adopted by the firm being studied helps us understand how the leads were targeted, which in turn informs us of the modeling approach to be used. The next section describes the selling process.

4.4 B2B selling process

The collaborating firm stores all the information they have on their leads in a database managed by a well-known marketing automation software, which assigns "lead scores" to all the leads present in their database. Based on a certain threshold value, any leads with a score above this threshold are targeted first by the sales force. These leads are called "Marketing Qualified Leads" or MQL. Once these leads are exhausted, the sales force shifts attention to the leads with lower score values. Of these leads, the leads that make it to the next Stage are called the "Sales Qualified Leads" or SQL. These leads usually request a demonstration of the product, after which they take a call on whether to buy the product or not. If they (i.e. the firm they work for) choose to buy the product, they request a few changes and then negotiate a buying price (unlike in the B2C context, prices in a B2B context vary by the customer firm [119]. Figure 4.3 shows the multiple Stages involved in the conversion of a lead into a customer of the firm. The firm uses a marketing automation software to develop lead attractiveness scores to enable a customer's "conversion path". However, in the past, the firm has noticed that the lead scores were not accurate, with sales force calling on leads with higher lead scores only to discover that they were not really interested, and leads with lower lead scores actually converting to customers (this is also the reason the sales team calls leads with a lower lead score once theyve exhausted all the leads with a higher lead score). In between these Stages, all

of these leads may indulge in any of the online or offline activities described before.

Given the selling process, it is clear to see that for a given company (employing the lead in question), there are many people involved in the decision making process leading to a purchase (Srinivasan 2012). The data also include a ranking for the lead in the firm on a scale of 1-5 (1: position of higher decision making power, and 5: position of least power in the company), which we include in the covariates to be used in the model. Finally, in the absence of price information, we use firm size information as a proxy (our reasoning being that a larger client firm might be willing to pay a higher price, given the uniqueness of the product, than a smaller client firm, which is also consistent with the information we received from the data provider).

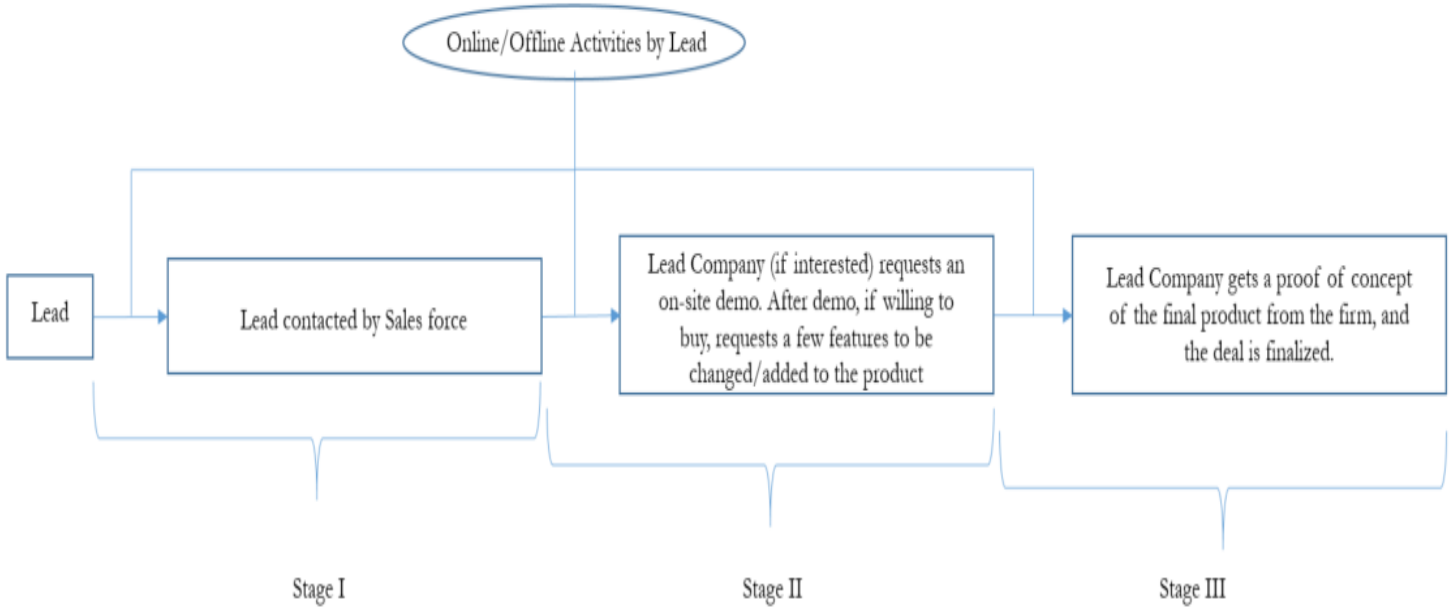

Figure 4.3: Different Stages in the Conversion of a Lead Company

4.5 Model Description

The challenge in modeling this lead conversion behavior is accounting for the different Stages in the conversion path, while incorporating the insights on the antecedents of conversion behavior from online and offline lead activity data. Given the detailed information we have on lead activities, we break the conversion path into three Stages (as shown in Figure 4.3). In each Stage, we have a record of the online and offline activities of the lead, and the time spent by the lead in between Stages (in days). We model the outcome of moving from one Stage to the next as a binary Probit, with an outcome of 1 indicating that a lead has successfully moved on to the next Stage of the conversion path. In addition, we also model the time spent by the lead in each Stage. The time spent in each Stage is non-zero if the lead makes it to that Stage (or is zero otherwise), and is a continuous variable. As an illustrative example, consider a particular Stage in the selling process, as shown in Figure 4.4. Initially, a binary Probit models the chances of a lead getting past Stage 1. Once the lead gets past Stage 1 (i.e. his/her company shows initial interest in buying the product), we model the time spent between Stages 1 and 2 as a Tobit regression (since the observed variable for the leads who do not make it to Stage 2 is the maximum time in days as observed in the data these leads might convert sometime in the future, given the broad applicability of the product. These leads also participate in various online and offline activities, but in our sample, do not transition between Stages).

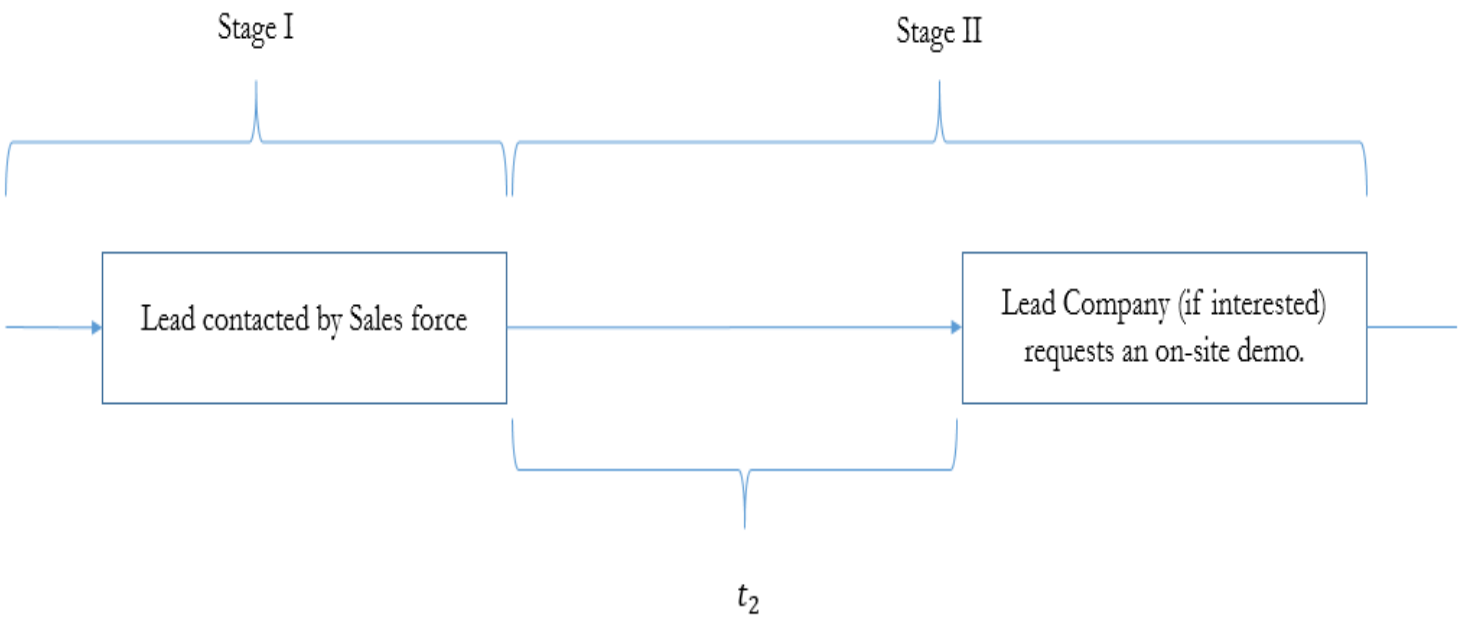

Figure 4.4: Stage to Stage Transition Structure

In model form, for a lead i, this would be:

$$y_{i1} = 1(X_{i1}\beta_i + \epsilon_{i1} > 0)$$

$$t_{i2} = X_{i2}\beta_i + \epsilon_{i2} \; if \; y_{i1} = 1, \; 0 \; otherwise$$

Where X_{i1} is the covariate matrix that includes information on lead activity for lead i (number of binary downloads, trade show and conference attendance, information requests, job position of lead in company, price charged) before Stage 1, and X_{i2} contains the same information, but for the time period between Stage 1 and Stage 2. $\epsilon_{i1}, \epsilon_{i2}$ are the error terms, where $\epsilon_{i1} \sim N(0, 1)$ (for identification purposes) and $\epsilon_{i2} \sim N(0, \sigma^2)$ (where σ is to be estimated). We build similar models for each Stage of the conversion path due to availability of lead activity data at each Stage (given a lead made it to that particular Stage). These models belong to a class of discrete-continuous choice models, first discussed by Hanemann [52]. Though in his work, the continuous variable was used to model the quantity of a product purchased, given it was chosen in the first place. Models of discrete-continuous choice have also been applied in marketing [96] and transportation research [72, 38]. We chose this modeling approach over other benchmark time series models used in this literature for three reasons:

1. A lack of time series persistence in the lead behavior (i.e. no clear indication of a time series dependence in lead activities as discussed in the data section);
2. Sparsity of activities (covariates), which renders a linear model (as is the norm in B2C literature to model conversion behavior using clickstream data) to be inappropriate for our application unless copulas are used to capture contemporaneous correlations [97]. Given the sparsity of time series data in the model, we compute aggregate values for the variables to

be used in the model for each of the leads, in each Stage. These in turn are used as the covariates in the discrete continuous choice model implemented; and

3. The nuances of the specific selling process employed by the firm being studied cannot be captured by the dynamic time series models (VAR models) used in this stream of literature.

Certain features of the B2B setting and the nature of the product being sold by the collaborating firm guide our choice of estimation of this model. Firstly, in a B2B setting, buying situations vary considerably [98]. In addition, there is heterogeneity in the customers documented usage of ALM software[6]. Since the collaborating firm provides software to a global audience, it caters to a lot of geographical markets. As an example, Figure 4.5 shows a plot of all the leads in the USA, and their interest in the intended usage of the firms product. Given (1) the varied usage of the firm's product (2) the many markets all over the world the firm caters to and (3) the varied buying situations, the coefficients β in the equations above cannot be assumed to be the same across leads. One approach to account for this heterogeneity is to use a finite mixture model approach [4]. However, it is not clear how to choose the number of mixture components before estimation. Early work in the field of Bayesian Nonparametric models has tackled this issue by placing a Dirichlet Process Prior on the heterogeneous quantity of interest [35, 108, 63, 23, 69]. The idea is to model the coefficients as a draw form an infinite mixture model [57, 35], without having to specify the number of underlying clusters beforehand. Irrespective of the convenience of placing a Dirichlet Process Prior on coefficients in our model, researchers in

[6]Singh, M (2014), "The Other (and very real) Benefits of Application Lifecycle Management!" http://blog.digite.com/the-other-and-very-real-benefits-of-application-lifecycle-management/

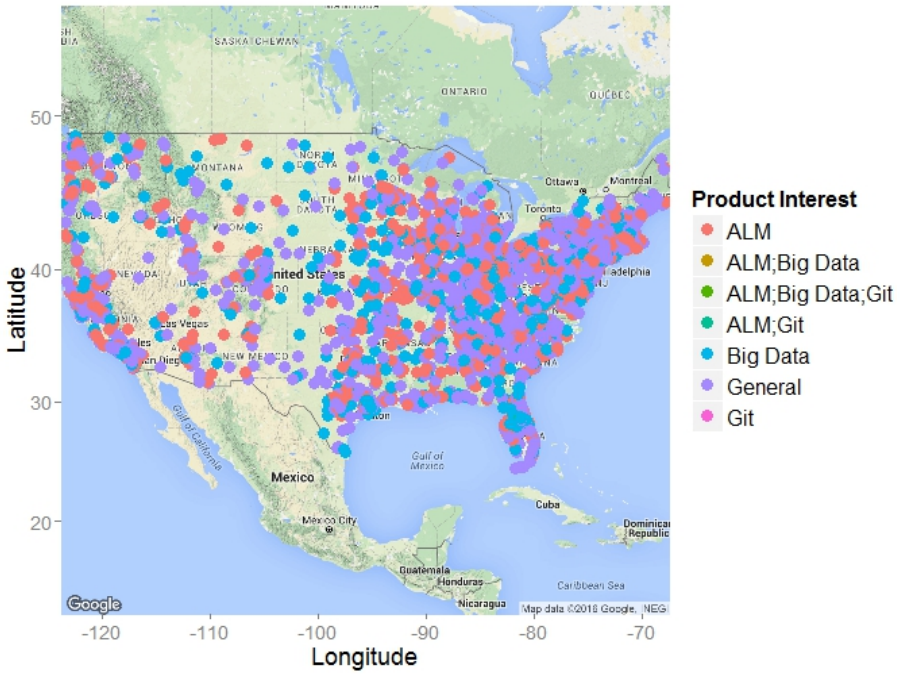

Figure 4.5: A Map Showing Lead Locations in the US, and their Product Interests (ALM; Big Data; Git)

the past have stated that modeling mixture data using Bayesian Nonparametric methods is inherently misspecified [84], since the model might "decide" to create extra latent clusters which might not exist in the given data [63]. In addition, previous applications using the Dirichlet Process Prior to model Heterogeneity have only focused on settings in a single market [23, 69]. The challenge we face here is to model the coefficients so as to account for Heterogeneity across and within markets. So as to tackle this issue, we model the coefficients as draws from a Hierarchical Dirichlet Process Mixture [105], (HDP) and estimate the discrete-continuous choice model of lead conversion behavior using a Bayesian framework. The advantages of using a HDP Mixture to model heterogeneity are three fold:

1. A HDP model imposes a structure on the coefficients in the model, such that they reflect the different preferences in the data and not make the model too unwieldy to estimate at the same time [105].
2. A HDP corrects for the case where we might not have enough information on lead activities in a particular Stage of the conversion path. It automatically "checks" if the lead in question exhibits similar choice patterns to other existing leads (on whom we have enough information) and uses that information to fit the model. In other words, it pools information appropriately where necessary, so as to "share statistical strength" [105]
3. As more data become available, the HDP Mixture framework keeps adjusting the ideal number of component mixtures based on model fit (i.e. it "learns" with newer data).

Fang [38] estimates a Bayesian discrete-continuous choice model of household vehicle usage, but does not account for household level heterogeneity. We build on the framework she uses and account for the heterogeneity of lead behavior. The modeling novelty in this work is to estimate this discrete-continuous choice model, while accounting for lead heterogeneity across markets using a HDP mixture; these features, to the best of our knowledge, have not been attempted earlier in B2B lead conversion models.

A DAG which represents the Probit model implemented in the paper is shown in Figure 4.6 (the graph is similar for the continuous part of the model). H is the base distribution (assumed to be a diffuse Normal Inverse Wishart distribution) and α and γ are the concentration parameters (both are set to 1 in this case, following past literature [69, 105]. A Bayesian procedure is used to estimate this model for two reasons:

1. When modeling discrete choice models with individual level heterogeneity, a Bayesian procedure tends to "pick-out" individual level parameters during the estimation process, since that makes the estimation process easier [110].
2. The computational cost of evaluating multiple integrals in the Probit model can be avoided by introducing latent variables that need to be estimated [38, 3, 24].

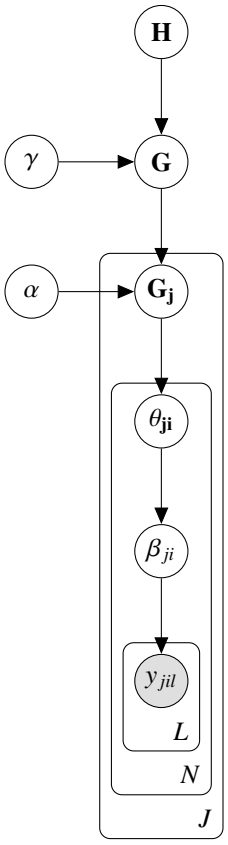

Figure 4.6: A Probit with a Hierarchical Dirichlet Process Mixture Prior on the Coefficients

The DAG in Figure 4.6 succinctly represents the dependency structure in the model. The HDP models heterogeneity across markets by specifying different "market types" (j), based on the variation present in the data being mod-

eled (Teh et al. [105], analogously use the HDP to model different "document types"). This serves as an alternative to either (i) ignoring heterogeneity across multiple markets (which might give incorrect results), or (ii) modeling the heterogeneity in each market separately (which could be cumbersome). As can be noted from the DAG in Figure 4.6, in a given Stage, we have 1 observation for each individual (lead) belonging to a market type j (since we aggregate data in a Stage due to sparsity). Since there are multiple leads belonging to a firm, we have L such observations. Econometrically, the model would be unidentified if we were to place an individual level coefficient on each of the leads (due to only 1 observation per lead, per Stage). Instead, we treat all the leads belonging to the same company i to "share" the same distribution of the coefficient (β_{ji} in the graph). From a theoretical standpoint in B2B literature, this approach (of modeling heterogeneity at the firm level) has been used previously to model tradeshow attendance [49, 48] and pricing decisions in a B2B context [119].

Given the firm level coefficient β_{ji}, the likelihood is computed for all leads L belonging to the company. For a given market type j , the likelihood contribution of a lead l belonging to firm i is computed using the approach in Greenberg [50]. To introduce conjugacy and avoid the integrations involved in computing the probabilities in the Probit model, Greenberg [50] introduces latent variables, which can be treated as parameters to be estimated (and later discarded). An outline of the Gibbs sampling scheme is provided below (the market type subscript j is suppressed for clarity). Let y_{il} be the observed data for a lead l in company i , in a given Stage. Introducing latent data y^*_{il} and following the procedure in Greenberg [50], the likelihood contribution of this lead in the given

Stage is:

$$p(y_{il}|y^*_{il}) = 1(y_{il} = 0)1(y^*_{il} \leq 0) + 1(y_{il} = 1)1(y^*_{il} \geq 0) \tag{4.1}$$

$$y^*_{il} = X_{il}\beta_{il} + \epsilon_{il} \tag{4.2}$$

Assuming a normal prior on β_{il}, we get the following expression for the posterior:

$$\pi(\beta_{il}, y^*_{il}|y_{il}) = \prod_l \{1(y_{il} = 0)1(y^*_{il} \leq 0) + 1(y_{il} = 1)1(y^*_{il} \geq 0)\} f(y^*_{il}|\beta_{il})\pi(\beta_{il}) \tag{4.3}$$

Where $y^*_{il}|_{il} \sim N(X_{il}\beta_{il}, 1)$ (which follows from equation 4.2). The indicator functions in equation 3 serve the purpose of truncating the normal appropriately, based on the observed data y_{il} . To model heterogeneity using the HDP, we place a HDP mixture prior on the parameters of the normal prior for β_{il}. If $\beta_{il} \sim N(\mu, \Sigma)$, then $(\mu, \Sigma) \sim HDP(\alpha, \gamma, H)$. Hence, in the Gibbs algorithm, once we obtain draws for the parameters for the distribution of β_{il} , these become the "data" to be modeled using a HDP mixture model. The algorithm for fitting these data using the HDP mixture model is described in Teh et al. [105].

Based on the above discussion, the Gibbs sampling steps are (starting with an initial draw β^0_{ji}):

1. Draw $y^{*k}_{il} \sim 1(y_{il} = 0)TN_{(-\infty,0]}(X_{il}\beta^{k-1}_{il}, 1) + 1(y_{il} = 1)TN_{[0,\infty)}(X_{il}\beta^{k-1}_{il}, 1)$
2. Draw $\beta^k_{il} \sim N(\mu_n, \Sigma_n)$; where $\Sigma_n = (X_i'X_i + \Sigma^{-1})^{-1}$ and $\mu_n = \Sigma_n(X_i'y^{*k} + \Sigma^{-1}\mu)$
3. Model $(\mu_n, \Sigma_n) \sim HDP(\alpha, \gamma, H)$

Where k is the number of iterations in the Gibbs sampling algorithm. (The Gibbs algorithm for the Tobit regression was also implemented similarly using the technique described in Greenberg [50]).To check if using a HDP Mixture to

model heterogeneity (referred to as model III) actually fits the data better, we estimate a different version of the model, where step (iii) in the Gibbs sampling algorithm now is a Dirichlet Process Mixture (i.e. $(\mu_n, \Sigma_n) \sim DP(\alpha, H)$), referred to as model II. Another model was estimated without accounting for heterogeneity (i.e. ignoring step (iii)), which is referred to as model I. All three models were "trained" using the lead activity data for the first two and a half years in the data (28^{th} June 2012 to 31^{st} Dec 2014), and the predictive validity of the model was checked on the remaining data. A zero inflated probit model was also estimated, but the results didn't change by much. The results aren't included here, and are available upon request. The models were estimated for each Stage, with $40,000$ iterations of the Gibbs sampler each (the Raftery and Lewis [90] diagnostic suggested convergence was reached at this number of iterations).

4.6 Results and Conclusion

4.6.1 Model Fit

Table 4.5 shows a comparison of the fit of different models estimated on the data for each Stage. For all three Stages, the discrete continuous Probit model with a HDP mixture prior on the coefficients (model III) is a better fit (according to the BIC criterion).

Table 4.5: Log-Likelihood (LL) and BIC comparison for Models in all Stages. Model I: No Heterogeneity, Model II: DP prior for Heterogeneity, Model III: HDP prior for Heterogeneity

Criterion	Model I	Model II	Model III
Stage I			
LL	-5,637,773	-5,092,328	-4,881,675
BIC	11,275,568	10,184,687	9,763,405
Stage II			
LL	-2,774,099	-2,724,232	-2,618,174
BIC	5,548,256	5,448,484	5,236,413
Stage III			
LL	-231,945.4	-213,744	-206,564.5
BIC	463,907.7	427,577	413,226.6

4.6.2 Model Estimates

Tables 4.6, 4.7 and 4.8 show the results of the discrete continuous choice model estimation for each Stage, by segment (the coefficients are provided with their 95% credible intervals in brackets). Model III recovers $j = 4$ "market types" and each of these market types has three segments. Remarkably, the nature of the coefficients is the same for all segments, except for in segment III, where the ceofficient of conference attendance in not significant in Stage I, but is significant in Stage II, unlike in the other segments. We now interpret the coefficients in each segment, across all Stages and then study the market types recovered by model III.

Segments I and II

Segments I and II are examined here together, since similar coefficients are significant in each Stage for both these segments. In Stage I, across both segments, the number of downloads appears to increase the chances of a lead to transi-

tion towards the next Stage and also shortens the time spent in the Stage, but in the later Stages, it doesnt seem to affect the chances of a lead moving further along. A possible explanation for this is due to the selling process of the firm being studied: in later Stages, the product is demonstrated to the leads' firm and feedback is obtained regarding changes to be made to the software, and this event takes precedence over a lead downloading a binary. Tradeshows and conferences appear to matter in Stage I, which confirms the results in Gopalakrishna et al. and Gopalakrishna and Lilien [49, 48], in that they are important for spreading awareness about the firms product. The coefficient on information gathering is positive and significant only in Stage II, indicating that the process of information gathering might not be that critical in the early and late Stages of the selling process, but is very important in the middle Stage. However, they dont seem to matter in later Stages. This can be attributed to the fact that in Stage II, the process of information gathering is more prevalent, and once enough information is obtained on the product, attendance of tradeshows or conferences is unnecessary. The decision making power (as measured by the rank in the firm) of a lead does not appear to matter for Stage I and II, but does make a difference in Stage III (the higher the rank (lower in magnitude), the higher the chances of a lead converting). A possible explanation for this could be that in the later Stages, a higher ranking executive in the lead firm can take a call on buying the product, rather than wait for his/her superior to make a decision on the purchase (in the case of a lower ranking executive in the lead firm). Finally, segment II is more price sensitive than segment I. While the coefficient of price negatively affects the outcome of a deal being closed in Stage III in both segments, it does not affect the transition of a lead in the earlier Stages. Since the discussion of price for the product is set in Stage III of the selling process, it

appears to be irrelevant in the earlier Stages. Another key fact to note here is the positive coefficient of the price term in the continuous model of Stage III. This indicates that the higher the negotiated price, the longer it takes for a lead firm to convert into a client of the company.

Segment III

Segment III is the most price sensitive segment, since the magnitude of its price coefficient is the highest in Stage III, when compared to the price coefficients in segments I and II. Another key difference is that in Stage I, the coefficient of conference attendance is not significant, but it is significant in Stage II. This segment could contain cash-strapped client firms (due to their high price sensitivity), who might not be able to attend both conferences and trade shows in the first Stage (maybe due to high registration fees). However, once there is an interest in the product, only then do they decide to attend conferences to learn more about the product, and this happens in Stage II. All the other coefficients in all Stages are similar in nature to those of segments I and II, so the same inferences from before apply here.

4.6.3 Market Types

Table 4.9 shows the different market types recovered from the data by model III. We see a marked difference of segment probabilities across market types. There are more markets of type I (market type probability 0.514), in which segments II and III share the same probability. This tells us that there are more market types with price sensitive segments. Indeed, all other market types that place a

higher probability on segment I are smaller in number. As mentioned in earlier sections, the HDP mixture prior on the coefficients of model III allows us to compare segments across market types, something that would be hard to do when modeling each market separately.

Table 4.6: Segment I Coefficients from Model III for all Stages (with 95% Credible Intervals)

	Stage I		Stage II		Stage III	
Variable	Discrete	Continuous	Discrete	Continuous	Discrete	Continuous
Binary Downloads	**2.26**	**-7.23**	1.04	-1.65	0.89	-1.65
	[1.23, 3.29]	[-9.21, -5.25]	[-0.34, 2.42]	[-3.75, 0.45]	[-1.53, 3.31]	[-5.21, 1.99]
Tradeshow Attendance	**4.12**	**-13.27**	1.26	-2.47	0.71	-1.12
	[2.93, 5.31]	[-16.29, -10.25]	[-0.52, 3.04]	[-6.17, 1.23]	[-0.88, 2.3]	[-3.67, 1.43]
Conference Attendance	**3.13**	**-15.42**	2.85	-1.87	0.94	-1.38
	[1.69, 4.57]	[-18.37, -12.47]	[-1.15, 6.85]	[-5.01, 1.27]	[-1.62, 3.5]	[-5.91, 3.15]
Information Request	1.15	-2.67	**3.78**	**-17.27**	0.64	-1.59
	[-0.91, 3.27]	[-6.37, 1.03]	[2.21, 5.35]	[-19.56, -14.98]	[-0.47, 1.75]	[-4.46, 1.28]
Job Position	0.62	-0.33	0.41	-0.33	**-2.21**	**1.65**
	[-2.33, 3.57]	[-1.24, 0.58]	[-2.92, 3.74]	[-1.63, 0.97]	[-3.56, -0.86]	[0.67, 2.63]
Price	-0.45	0.82	-0.63	0.45	**-5.61**	**13.25**
	[-1.65, 0.79]	[-2.41, 4.05]	[-1.82, 0.56]	[-2.19, 3.09]	[-7.52, -3.6]	[11.41, 15.09]

Table 4.7: Segment II Coefficients from Model III for all Stages (with 95% Credible Intervals)

	Stage I		Stage II		Stage III	
Variable	Discrete	Continuous	Discrete	Continuous	Discrete	Continuous
Binary Downloads	**0.69**	**-2.64**	0.39	-0.39	0.37	-0.54
	[0.38, 1.02]	[-3.37, -1.91]	[-0.13, 0.90]	[-0.90, 0.11]	[-0.64, 1.38]	[-1.69, 0.65]
Tradeshow Attendance	**1.83**	**-4.26**	0.33	-0.54	0.29	-0.56
	[1.3, 2.36]	[-5.23, -3.29]	[-0.14, 0.79]	[-1.35, 0.27]	[-0.35, 0.93]	[-1.83, 0.71]
Conference Attendance	**1.18**	**-4.64**	1.13	-0.62	0.44	-0.59
	[0.64, 1.73]	[-5.53, -3.75]	[-0.45, 2.71]	[-1.67, 0.42]	[-0.76, 1.65]	[-2.57, 1.37]
Information Request	0.57	-0.89	**1.74**	**-8.51**	0.30	-0.37
	[-0.45, 1.63]	[-2.14, 0.35]	[1.02, 2.46]	[-9.64, -7.38]	[-0.22, 0.82]	[-1.03, 0.30]
Job Position	0.16	-0.12	0.17	-0.15	**-0.59**	**0.74**
	[-0.62, 0.95]	[-0.45, 0.21]	[-1.22, 1.56]	[-0.76, 0.45]	[-0.95, -0.23]	[0.30, 1.17]
Price	-0.13	0.35	-0.19	0.11	**-8.23**	**15.29**
	[-0.49, 0.23]	[-1.03, 1.74]	[-0.56, 0.17]	[-0.55, 0.77]	[-9.41, -7.05]	[13.37, 17.21]

Table 4.8: Segment III Coefficients from Model III for all Stages (with 95% Credible Intervals)

	Stage I		Stage II		Stage III	
Variable	Discrete	Continuous	Discrete	Continuous	Discrete	Continuous
Binary Downloads	**3.35**	**-11.38**	1.65	-2.26	1.48	-2.49
	[1.82, 4.87]	[-14.50, -8.26]	[-0.54, 3.85]	[-5.13, 0.62]	[-2.54, 5.50]	[-7.86, 3.00]
Tradeshow Attendance	**7.04**	**-19.93**	1.76	-3.29	1.16	-2.02
	[5.00, 9.07]	[-24.47, -15.40]	[-0.73, 4.26]	[-8.21, 1.64]	[-1.44, 3.77]	[-6.60, 2.57]
Conference Attendance	1.05	-2.64	**4.63**	**-2.85**	1.65	-2.33
	[-3.74, 5.84]	[-6.94, 1.66]	[2.23, 7.03]	[-3.27, -2.43]	[-2.84, 6.13]	[-9.99, 5.33]
Information Request	2.07	-4.07	**6.56**	**-30.88**	1.12	-2.15
	[-1.64,5.89]	[-9.72, 1.57]	[3.83, 9.28]	[-34.98, -26.79]	[-0.82, 3.06]	[-6.03, 1.73]
Job Position	0.87	-0.52	0.68	-0.58	**-3.12**	**2.82**
	[-3.28, 5.03]	[-1.95, 0.91]	[-4.85, 6.22]	[-2.84, 1.69]	[-5.02, -1.21]	[1.15, 4.50]
Price	-0.66	1.38	-0.93	0.62	**-7.60**	**20.44**
	[-2.40, 1.15]	[-4.05, 6.81]	[-2.69, 0.83]	[-3.03, 4.28]	[-10.19, -4.88]	[17.61, 23.28]

Table 4.9: Segment Probabilities for Different Market Types

Market Type (Type Size)	Segment I	Segment II	Segment III
Market Type I (0.514)	0.26	0.37	0.37
Market Type II (0.122)	0.55	0.29	0.16
Market Type III (0.214)	0.41	0.24	0.35
Market Type IV (0.15)	0.63	0.27	0.10

4.6.4 Validation

Finally, in a check of the three different models on the validation data sample, the proposed model (model III), does better in terms of predicting lead conversion and the time spent between Stages, as shown in Figures 4.7,4.8 and 4.9. A comparison of the Root Mean Squared Error (RMSE) for these predictions is shown in Table 4.10. The Figures and the Table show that the predictions from model III are closer to the actual values, when compared to predictions from models I and II.

The results of this analysis can be used to decide which activities deserve a higher share of the marketing budget in a company. If the goal is to attract new leads, the model recommends investing in tradeshows and conferences to

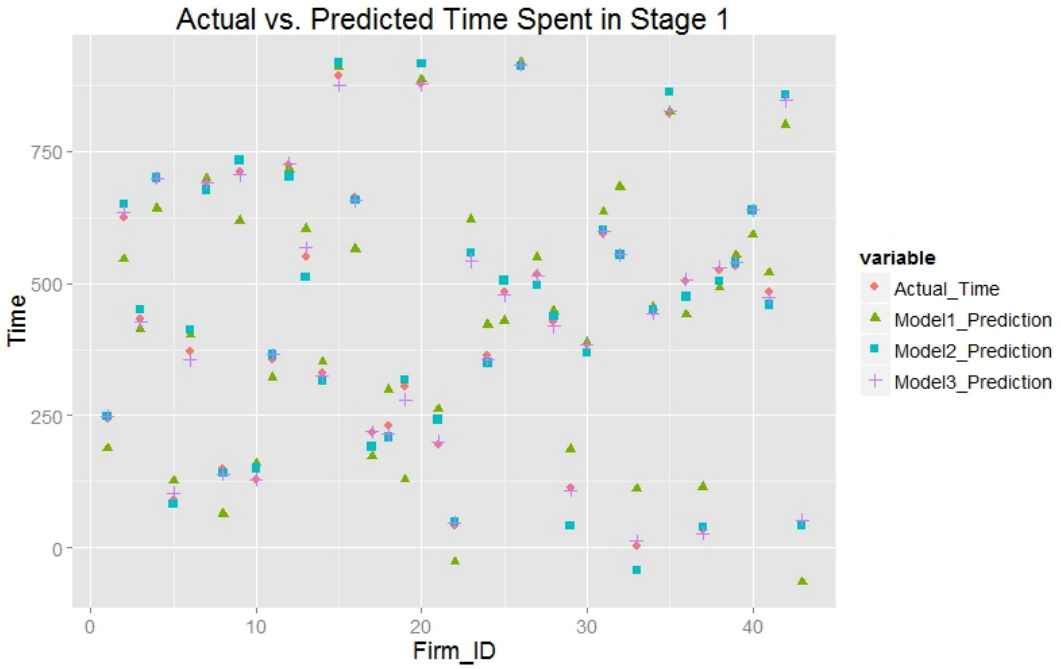

Figure 4.7: Actual and Predicted Times spent in Stage I

Table 4.10: RMSE of Predicted Conversion Times

Model	Stage I	Stage II	Stage III
Model I	62.99	61.31	47.11
Model II	23.85	30.26	29.15
Model III	8.81	12.33	4.42

spread awareness. The model also helps the manager focus on the leads that warrant his/her attention. For example, in Stage I, the manager would want to focus on leads who have downloaded a binary. In Stage II, he/she would be well served to study leads who are actively gathering information through offline channels. In the last Stage, the manager would want to give preference to the lead who is the higher ranking official in a potential client firm. It should

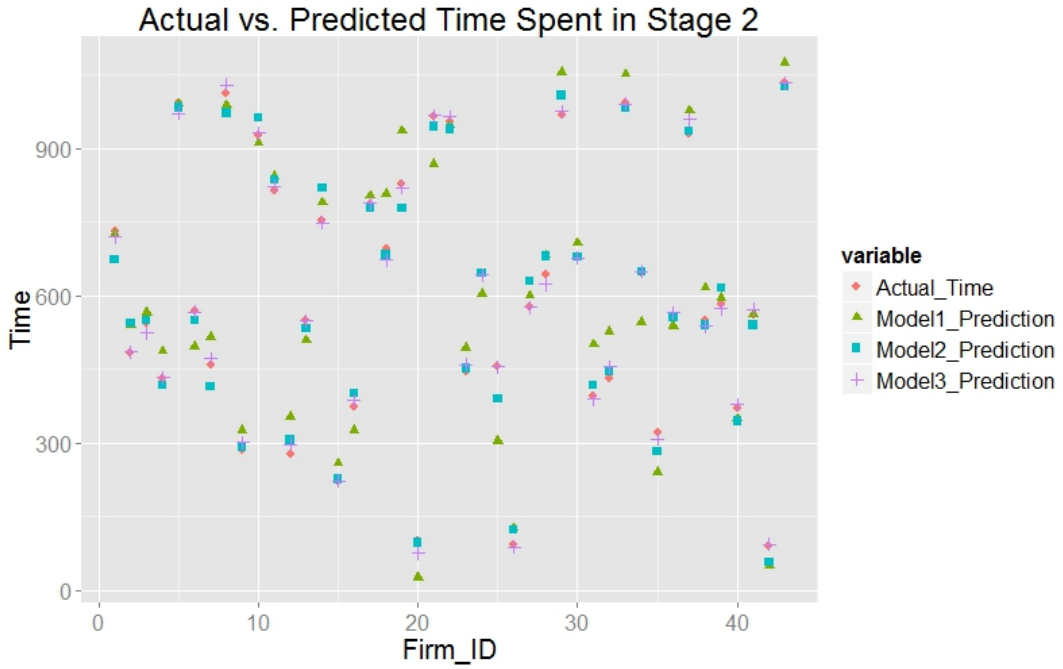

Figure 4.8: Actual and Predicted Times spent in Stage II

be noted that the results obtained in this paper indicate the importance of incorporating the actual selling process employed in a B2B context. These insights couldn't have been gained by disregarding the selling process altogether.

4.7 Future Research

We contribute methodologically to the customer acquisition analysis literature, by introducing a data driven way to model lead heterogeneity and modeling the time spent between Stages in a conversion path. This research project uses novel data on leads collected by our collaborating firm, to optimize marketing spending and help the sales force target the leads more likely to convert, in a B2B context. This work will be useful for B2B firms similar to our collaborating firm

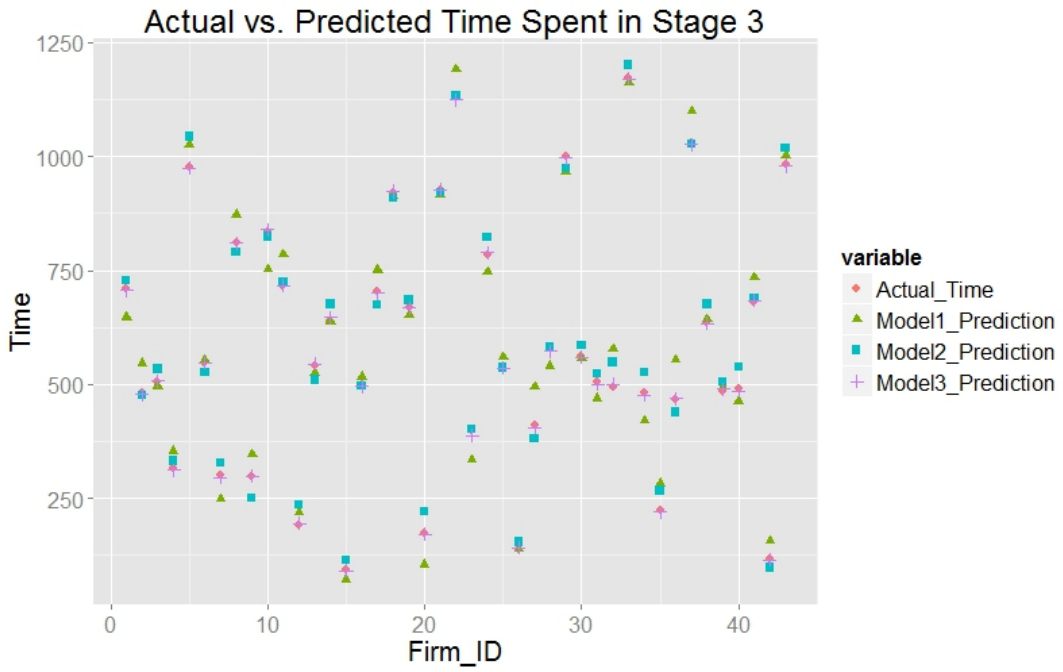

Figure 4.9: Actual and Predicted Times spent in Stage III

as it will help them reduce wasteful marketing spending and efficiently target leads more likely to become customers. However, some choices in modeling were driven by the sparsity in the data, and this leaves a lot of room for model improvement once richer data are available.

We use a continuous representation of heterogeneity in this model, which is a better fit to the data when compared to discrete representations [7]. The model used in this chapter can easily be modified and applied to other B2B and B2C contexts, with the added advantage of modeling for heterogeneity across multiple markets. Using the HDP to model heterogeneity in consumer search and purchase for B2C contexts across multiple markets could also offer additional insight. Another extension of the model used in this chapter is to incorporate time varying preferences of leads in a company. As of now, these preferences

(captured by the coefficients in the model) are all fixed over time, and this might be restrictive in certain applications.

APPENDIX A

SELECT PROPERTIES OF THE DIRICHLET DISTRIBUTION

The Dirichlet Process was formally introduced by Ferguson [40], where he showed that this process becomes a finite dimensional distribution over a particular partition of the probability space over which it (the process) is defined. Since a majority of the properties of the Dirichlet Process are derived from the properties of the Dirichlet distribution,this chapter will derive the relevant properties of this distribution.

A.1 The Dirichlet Distribution

The intent of this section is to walk the reader through some of the basic properties of the Dirichlet Distribution (its infinite dimensional counterpart, is the dirichlet process), accompanied by their derivations. These properties will be used in the later sections of this chapter. This is in no way a comprehensive study of this distribution, and the interested readers are encouraged to refer to Ghosh and Ramamoorthi [45] for more.

A.2 Derivation from the Gamma distribution

A dirichlet distributed random variable can be derived from Gamma distributed random variables. For example, in the two dimensional case, if Z_1 & Z_2 are IID random variables such that $Z_1 \sim Gamma(\alpha_1, 1)$ & $Z_2 \sim Gamma(\alpha_2, 1)$ (Z_1 and Z_2 are > 0) and we define $Y = \frac{Z_1}{Z_1+Z_2}$ (by definition, $0 \leq Y \leq 1$), then we

have $Y \sim Beta(\alpha_1, \alpha_2)$ (by definition, the Dirichlet is the multivariate generalization of the Beta distribution).

A.2.1 Proof

The Two Dimensional Case

The strategy here is to work backwards from the distribution function of Y. The probability that the random variable $Y \leq y$ is given by:

$$
\begin{aligned}
& P(Y \leq y) = P\left(\frac{Z_1}{Z_1 + Z_2} \leq y\right) = P\left(Z_1 \leq \frac{yZ_2}{1-y}\right) \\
\implies & P(Y \leq y) = \int_0^{\infty} \int_0^{\frac{yZ_2}{1-y}} \frac{1}{\Gamma(\alpha_2)} z_2^{\alpha_2 - 1} e^{-z_2} \frac{1}{\Gamma(\alpha_1)} z_1^{\alpha_1 - 1} e^{-z_1} dz_1 dz_2 \\
\implies & P(Y \leq y) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^{\infty} z_2^{\alpha_2 - 1} e^{-z_2} \left[\int_0^{\frac{yZ_2}{1-y}} z_1^{\alpha_1 - 1} e^{-z_1} dz_1 \right] dz_2
\end{aligned}
$$

The probability density function of Y is obtained by taking the derivative of $P(Y \leq y)$ with respect to y (all regularity conditions needed for this operation to be valid are satisfied here). The Leibniz rule is used to obtain the derivative under an integral. Formally, if $f(x, t)$ is a function such that the partial derivative of f with respect to t exists, and is continuous, then:

$$
\frac{d}{dt}\left(\int_{a(t)}^{b(t)} f(x,t)dx\right) = \int_{a(t)}^{b(t)} \frac{\partial f}{\partial t} dx + f(b(t), t) \cdot b'(t) - f(a(t), t) \cdot a'(t)
$$

Taking the derivative of $P(Y \leq y)$ with respect to y and using Leibniz's rule, we

have:

$$\frac{dP(Y \leq y)}{dy} = f_Y(y) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^\infty z_2^{\alpha_2 - 1} e^{-z_2} \frac{\partial}{\partial y} \left[\int_0^{\frac{yZ_2}{1-y}} z_1^{\alpha_1 - 1} e^{-z_1} dz_1 \right] dz_2$$

$$\implies f_Y(y) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^\infty z_2^{\alpha_2 - 1} e^{-z_2} \left[\left(\frac{yz_2}{1-y} \right)^{\alpha_1 - 1} e^{-\frac{yZ_2}{1-y}} \frac{1}{(1-y)^2} z_2 \right] dz_2$$

$$\implies f_Y(y) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^\infty z_2^{\alpha_1 + \alpha_2 - 1} \left(\frac{y}{1-y} \right)^{\alpha_1 - 1} e^{-\frac{z_2}{1-y}} \frac{1}{(1-y)^2} dz_2$$

Setting $x = \frac{z_2}{1-y}$, ($dx = \frac{dz_2}{1-y}$) and changing the limits of integration appropriately, we get

$$f_Y(y) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^\infty (1-y)^{\alpha_1 + \alpha_2 - 1} x^{\alpha_1 + \alpha_2 - 1} \left(\frac{y}{1-y} \right)^{\alpha_1 - 1} e^{-x} \frac{1}{(1-y)^2} (1-y) dx$$

$$\implies f_Y(y) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^\infty (1-y)^{\alpha_2 - 1} x^{\alpha_1 + \alpha_2 - 1} y^{\alpha_1 - 1} e^{-x} dx$$

$$\implies f_Y(y) = \frac{y^{\alpha_1 - 1}(1-y)^{\alpha_2 - 1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^\infty x^{\alpha_1 + \alpha_2 - 1} e^{-x} dx \tag{A.1}$$

Where the definite integral in A.1 is a standard Euler integral of the second kind [29], and equals $\Gamma(\alpha_1 + \alpha_2)$. Putting this back in the previous expression gives

$$f_Y(y) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} y^{\alpha_1 - 1}(1-y)^{\alpha_2 - 1} \tag{A.2}$$

Which is the $Beta(\alpha_1, \alpha_2)$ density.

A Change of Variables Approach

We can prove the above result using a change of variables approach. We define random variables U and V such that

$$U = \frac{Z_1}{Z_1 + Z_2}$$

$$V = Z_1 + Z_2$$

Where by definition, $0 \leq U \leq 1$ and $0 \leq V \leq \infty$. Solving for Z_1 and Z_2 in terms of U and V gives

$$Z_1 = UV \tag{A.3}$$

$$Z_2 = V(1 - U) \tag{A.4}$$

The change of variables approach gives us the density of the joint distribution of U and V from the following formula

$$f_{U,V}(u, v) = f_{Z_1,Z_2}(z_1, z_2)|J_{(z_1,z_2)\to(u,v)}| \tag{A.5}$$

Where $J_{(z_1,z_2)\to(u,v)}$ is the Jacobian matrix, given by

$$J_{(z_1,z_2)\to(u,v)} = \begin{bmatrix} \frac{\partial z_1}{\partial u} & \frac{\partial z_1}{\partial v} \\ \frac{\partial z_2}{\partial u} & \frac{\partial z_2}{\partial v} \end{bmatrix}$$

Given A.3 and A.4, the Jacobian matrix in this case becomes

$$J_{(z_1,z_2)\to(u,v)} = \begin{bmatrix} v & u \\ -v & 1 - u \end{bmatrix}$$

$$\implies |J_{(z_1,z_2)\to(u,v)}| = v(1 - u) + vu = v \tag{A.6}$$

Substituting for $|J_{(z_1,z_2)\to(u,v)}|$ in A.5 with the result in A.6 gives

$$f_{U,V}(u, v) = f_{Z_1,Z_2}(z_1, z_2) \cdot v \tag{A.7}$$

Since Z_1 and Z_2 are IID $Gamma(\alpha_1, 1)$ and $Gamma(\alpha_2, 1)$ respectively, we have

$$f_{Z_1,Z_2}(z_1, z_2) = f_{Z_1}(z_1) \cdot f_{Z_2}(z_2) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} z_1^{\alpha_1 - 1} z_2^{\alpha_2 - 1} e^{-(z_1 + z_2)}$$

Substituting this in A.7, and using relations A.3 and A.4, we get

$$f_{U,V}(u, v) = f_{Z_1,Z_2}(z_1, z_2) \cdot v = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} z_1^{\alpha_1 - 1} z_2^{\alpha_2 - 1} e^{-(z_1 + z_2)} \cdot v$$

$$\implies f_{U,V}(u, v) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} (uv)^{\alpha_1 - 1} (v(1 - u))^{\alpha_2 - 1} e^{-v} \cdot v$$

$$\implies f_{U,V}(u, v) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-v} (uv)^{\alpha_1 - 1} v^{\alpha_2 - 1} (1 - u)^{\alpha_2 - 1} \cdot v$$

$$\implies f_{U,V}(u, v) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-v} v^{\alpha_1 + \alpha_2 - 1} (u)^{\alpha_1 - 1} (1 - u)^{\alpha_2 - 1} \tag{A.8}$$

Note that the above joint distribution A.8 can be factorized into a product of functions which depending only on u and v

$$f_{U,V}(u,v) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-v} v^{\alpha_1+\alpha_2-1} (u)^{\alpha_1-1} (1-u)^{\alpha_2-1}$$

$$\implies f_{U,V}(u,v) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} (u)^{\alpha_1-1} (1-u)^{\alpha_2-1} \cdot \frac{1}{\Gamma(\alpha_1+\alpha_2)} e^{-v} v^{\alpha_1+\alpha_2-1} \tag{A.9}$$

$$\implies f_{U,V}(u,v) = f_U(u) \cdot f_V(v) \tag{A.10}$$

Equation A.9 gives us the result we need and more:

- $f_U(u) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} (u)^{\alpha_1-1} (1-u)^{\alpha_2-1}$ is the $Beta(\alpha_1, \alpha_2)$ density, which gives us the result we wanted to prove
- U and V are independent random variables
- $f_V(v) = \frac{1}{\Gamma(\alpha_1+\alpha_2)} e^{-v} v^{\alpha_1+\alpha_2-1}$, hence $V \sim Gamma(\alpha_1+\alpha_2, 1)$

As the next few sections will show, these results are critical to a lot of proofs that follow, and will be used in the empirical work to be discussed in this thesis.

The N-Dimensional Case

The same results will be dervied for the N-Dimensional Dirichlet Distribution. Essentially, we want to prove the following:

Theorem 1 *Given $Z_i \sim Gamma(\alpha_i, 1)$, are IID, ($\alpha_i > 0$ and $1 \leq i \leq n$) and we define the variables $U_i = \frac{Z_i}{\sum_{i=1}^{N} Z_i}$ and $V = \sum_{i=1}^{N} Z_i$. Then:*

- $(U_1, U_2, \ldots, U_N) \sim Dir(\alpha_1, \alpha_2, \ldots, \alpha_N) = \frac{\Gamma(\sum \alpha_i)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_N)} u_1^{\alpha_1 - 1} u_2^{\alpha_1 - 1} \ldots u_N^{\alpha_N - 1} = \frac{\Gamma(\sum \alpha_i)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_N)} u_1^{\alpha_1 - 1} u_2^{\alpha_1 - 1} \ldots (1 - \sum_{i=1}^{N-1} u_i)^{\alpha_N - 1}$
- $V \sim Gamma(\sum_{i=1}^{N} \alpha_i, 1)$
- *$U = (U_1, U_2, \ldots, U_N)$ and V are independent*

An important point to note here is that by definition, $\sum_{i=1}^{N} U_i = 1$. The N^{th} term is redundant (in the sense that it can be obtained from the remaining $N - 1$ terms). Extending the change of variables approach to the N-dimensional case, we now set out to find the joint distribution of $(U_1, U_2, \ldots, U_{N-1}, V)$. We represent the Z_i's as functions of U_i and V to get:

$$Z_i = U_i V \qquad 0 \leq i \leq N - 1 \tag{A.11}$$

$$Z_N = (1 - \sum_{i=1}^{N-1} U_i) V \tag{A.12}$$

By definition, the joint density is given by

$$f_{U_1, U_2, \ldots, U_{N-1}, V}(u_1, u_2, \ldots, u_{N-1}, v) = f_{Z_1, \ldots, Z_N}(z_1, \ldots, z_N) |J_{(z_1, z_2, \ldots, z_{N-1}, z_N) \to (u_1, u_2, \ldots, u_{N-1}, v)}| \tag{A.13}$$

The Jacobian matrix is given by

$$J_{(z_1, z_2, \ldots, z_{N-1}, z_N) \to (u_1, u_2, \ldots, u_{N-1}, v)} = \begin{bmatrix} \frac{\partial z_1}{\partial u_1} & \frac{\partial z_1}{\partial u_2} & \cdots & \frac{\partial z_1}{\partial u_{N-1}} & \frac{\partial z_1}{\partial v} \\ \frac{\partial z_2}{\partial u_1} & \frac{\partial z_2}{\partial u_2} & \cdots & \frac{\partial z_2}{\partial u_{N-1}} & \frac{\partial z_2}{\partial v} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\partial z_{N-1}}{\partial u_1} & \frac{\partial z_{N-1}}{\partial u_2} & \cdots & \frac{\partial z_{N-1}}{\partial u_{N-1}} & \frac{\partial z_{N-1}}{\partial v} \\ \frac{\partial z_N}{\partial u_1} & \frac{\partial z_N}{\partial u_2} & \cdots & \frac{\partial z_N}{\partial u_N} & \frac{\partial z_N}{\partial v} \end{bmatrix} \tag{A.14}$$

Given A.11 & A.12, the partial derivatives in the Jacobian matrix A.14 are given by

$$\frac{\partial z_i}{\partial u_i} = v \qquad 0 \leq i \leq N-1 \tag{A.15}$$

$$\frac{\partial z_i}{\partial u_j} = 0 \qquad j \neq i, 0 \leq j \leq N-1 \tag{A.16}$$

$$\frac{\partial z_i}{\partial v} = u_i \qquad 0 \leq i \leq N-1 \tag{A.17}$$

$$\frac{\partial z_N}{\partial u_i} = -v \qquad 0 \leq i \leq N-1 \tag{A.18}$$

$$\frac{\partial z_N}{\partial v} = 1 - \sum_{i=1}^{N-1} u_i \tag{A.19}$$

Substituting A.15, A.16, A.17, A.18 & A.19 in A.14, we get

$$J_{(z_1,z_2,\ldots,z_{N-1},z_N)\to(u_1,u_2,\ldots,u_{N-1},v)} = \begin{bmatrix} v & 0 & \ldots & 0 & u_1 \\ 0 & v & \ldots & 0 & u_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & v & u_{N-1} \\ -v & -v & \ldots & -v & 1 - \sum_{i=1}^{N-1} u_i \end{bmatrix} \tag{A.20}$$

We need the determinant of the Jacobian matrix, which can be computed easily after making a few elementary row transformation opertaions

$$|J_{(z_1,z_2,\ldots,z_{N-1},z_N)\to(u_1,u_2,\ldots,u_{N-1},v)}| = \begin{vmatrix} v & 0 & \ldots & 0 & u_1 \\ 0 & v & \ldots & 0 & u_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & v & u_{N-1} \\ -v & -v & \ldots & -v & 1 - \sum_{i=1}^{N-1} u_i \end{vmatrix} \tag{A.21}$$

Apply the transformation $R_N \rightarrow R_N + \sum_{i=1}^{N-1} R_i$ (where R_i is the i^{th} row of the determinant) to get

$$|J_{(z_1,z_2,\ldots,z_{N-1},z_N)\rightarrow(u_1,u_2,\ldots,u_{N-1},v)}| = \begin{vmatrix} v & 0 & \ldots & 0 & u_1 \\ 0 & v & \ldots & 0 & u_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & v & u_{N-1} \\ 0 & 0 & \ldots & 0 & 1 \end{vmatrix} \tag{A.22}$$

Using the result in A.22 in A.13, we get

$$f_{U_1,U_2,\ldots,U_{N-1},V}(u_1,u_2,\ldots,u_{N-1},v) = f_{Z_1,Z_2,\ldots,Z_{N-1},Z_N}(z_1,z_2,\ldots,z_{N-1},z_N)v^{N-1} \tag{A.23}$$

Since Z_i are IID $Gamma(\alpha_i, 1)$ distributed

$$f_{Z_1,\ldots,Z_N}(z_1,\ldots,z_N) = \prod_{i=1}^{N} f_{Z_i}(z_i) = \prod_{i=1}^{N} \frac{1}{\Gamma(\alpha_i)} z_i^{\alpha_i-1} e^{-z_i} = \left[\prod_{i=1}^{N} \frac{1}{\Gamma(\alpha_i)} z_i^{\alpha_i-1}\right] e^{-\sum_{i=1}^{N} z_i} \tag{A.24}$$

Substituting A.24 into A.23, and using the relations in A.11 & A.12

$$\begin{aligned} & f_{U_1,\ldots,U_{N-1},V}(u_1,\ldots,u_{N-1},v) = \left[\prod_{i=1}^{N} \frac{1}{\Gamma(\alpha_i)} z_i^{\alpha_i-1}\right] e^{-\sum_{i=1}^{N} z_i} \cdot v^{N-1} \\ \implies & f_{U_1,\ldots,U_{N-1},V}(u_1,\ldots,u_{N-1},v) = \left[\prod_{i=1}^{N-1} \frac{1}{\Gamma(\alpha_i)} z_i^{\alpha_i-1}\right] \frac{1}{\Gamma(\alpha_N)} z_N^{\alpha_N-1} e^{-v} v^{N-1} \\ \implies & f_{U_1,\ldots,U_{N-1},V}(u_1,\ldots,u_{N-1},v) = \left[\prod_{i=1}^{N-1} \frac{1}{\Gamma(\alpha_i)} (u_i v)^{\alpha_i-1}\right] \frac{1}{\Gamma(\alpha_N)} (v(1-\sum_{i=1}^{N-1} u_i))^{\alpha_N-1} e^{-v} v^{N-1} \end{aligned} \tag{A.25}$$

Since $U_N = 1 - \sum_{i=1}^{N-1} U_i$, we put this in A.25 to get

$$f_{U_1,\ldots,U_{N-1},V}(u_1,\ldots,u_{N-1},v) = \left[\prod_{i=1}^{N-1}\frac{1}{\Gamma(\alpha_i)}(u_i v)^{\alpha_i-1}\right]\frac{1}{\Gamma(\alpha_N)}(u_N v)^{\alpha_N-1}e^{-v}v^{N-1}$$

$$\implies f_{U_1,\ldots,U_{N-1},V}(u_1,\ldots,u_{N-1},v) = \left[\prod_{i=1}^{N}\frac{1}{\Gamma(\alpha_i)}(u_i v)^{\alpha_i-1}\right]e^{-v}v^{N-1}$$

$$\implies f_{U_1,\ldots,U_{N-1},V}(u_1,\ldots,u_{N-1},v) = \left[\prod_{i=1}^{N}\frac{1}{\Gamma(\alpha_i)}(u_i)^{\alpha_i-1}\right]v^{\left(\sum_{i=1}^{N}\alpha_i-N\right)}e^{-v}v^{N-1}$$

$$\implies f_{U_1,\ldots,U_{N-1},V}(u_1,\ldots,u_{N-1},v) = \left[\prod_{i=1}^{N}\frac{1}{\Gamma(\alpha_i)}(u_i)^{\alpha_i-1}\right]v^{\left(\sum_{i=1}^{N}\alpha_i-1\right)}e^{-v}$$

$$\implies f_{U_1,\ldots,U_{N-1},V}(u_1,\ldots,u_{N-1},v) = \Gamma\left(\sum_{i=1}^{N}\alpha_i\right)\left[\prod_{i=1}^{N}\frac{1}{\Gamma(\alpha_i)}(u_i)^{\alpha_i-1}\right]\frac{1}{\Gamma\left(\sum_{i=1}^{N}\alpha_i\right)}v^{\left(\sum_{i=1}^{N}\alpha_i-1\right)}e^{-v}$$

$$\implies f_{U_1,\ldots,U_{N-1},V}(u_1,\ldots,u_{N-1},v) = \frac{\Gamma\left(\sum_{i=1}^{N}\alpha_i\right)}{\prod_{i=1}^{N}\frac{1}{\Gamma(\alpha_i)}}\left[\prod_{i=1}^{N}(u_i)^{\alpha_i-1}\right]\frac{1}{\Gamma\left(\sum_{i=1}^{N}\alpha_i\right)}v^{\left(\sum_{i=1}^{N}\alpha_i-1\right)}e^{-v}$$

$$\implies f_{U_1,\ldots,U_{N-1},V}(u_1,\ldots,u_{N-1},v) = Dir(\alpha_1,\alpha_2,\ldots,\alpha_N)\cdot Gamma(\sum_{i=1}^{N}\alpha_i,1) \tag{A.26}$$

Which establishes (1) $U = (U_1, U_2, \ldots, U_{N-1}, U_N)$ & V are independent, (2) $(U_1, U_2, \ldots, U_{N-1}, U_N) \sim Dir(\alpha_1, \alpha_2, \ldots, \alpha_N)$ & $V \sim Gamma(\sum_{i=1}^{N}\alpha_i, 1)$. This completes the proof of theorem 1.

We also prove the following useful property of the Dirichlet Distribution:

Theorem 2 *If* $U = (U_1, \ldots, U_i, U_{i+1}, \ldots, U_N) \sim Dir(\alpha_1, \ldots, \alpha_i, \alpha_{i+1}, \ldots, \alpha_N)$ *then*

$U' = (U_1, U_2, \ldots, U_i + U_{i+1}, \ldots, U_N) \sim Dir(\alpha_1, \alpha_2, \ldots, \alpha_i + \alpha_{i+1}, \ldots, \alpha_N)$

In general, $U'' = (\sum_{i=1}^{k_1} U_i, \sum_{i=k_1+1}^{k_2} U_i, \ldots, \sum_{i=k_j+1}^{N} U_i) \sim Dir(\sum_{i=1}^{k_1}\alpha_i, \sum_{i=k_1+1}^{k_2}\alpha_i, \ldots, \sum_{i=k_j+1}^{N}\alpha_i)$

We prove the first part of theorem 2 here. The rest follows by simply extending the proof. The strategy here is to construct IID Gamma distributed variables

with appropriate parameters, and then derive the distribution of U' using the results in theorem 1. Assume Z_i's are IID $\sim Gamma(\alpha_i, 1)$. From theorem 1, I have $Z_{i,i+1} = Z_i + Z_{i+1} \sim Gamma(\alpha_i + \alpha_{i+1}, 1)$. By definition, $Z_{i,i+1}$ is independent of Z_j's for $1 \leq j \leq N, j \neq i, i+1$. We now define $U' = (U_1, U_2, \ldots, U_i + U_{i+1}, \ldots, U_N) = (U_1, U_2, \ldots, U_{i,i+1}, \ldots, U_N)$, where $U_j = \frac{Z_j}{\sum_{j \neq i,i+1}^{N} Z_j + Z_{i,i+1}}$. From theorem 1, it follows that $U' \sim Dir(\alpha_1, \alpha_2, \ldots, \alpha_i + \alpha_{i+1}, \ldots, \alpha_N)$. The second part of theorem 2 follows from similar reasoning.

Theorem 2 can be used to get the marginal distributions of the U_i's, where $U = (U_1, \ldots, U_i, \ldots, U_N)$. Define $U' = (U_i, \sum_{j \neq i}^{N} U_j)$. From theorem 2, this is distributed $Dir(\alpha_i, \sum_{j \neq i}^{N} \alpha_j) = Beta(\alpha_i, \sum_{j \neq i}^{N} \alpha_j)$. Since $\sum_{j \neq i}^{N} U_j = 1 - U_i$, the density function can be expressed purely as a function of U_i, which is the marginal distribution of U_i. Formally:

Theorem 3 *If* $U = (U_1, \ldots, U_N) \sim Dir(\alpha_1, \ldots, \alpha_N)$ *then* $U_i \sim Beta(\alpha_i, \sum_{j \neq i}^{N} \alpha_j)$

Note that theorem 2 is a more general version of theorem 3. Another key property of the Dirichlet Distribution is discussed below:

Theorem 4 *If* $U = (U_1, \ldots, U_N) \sim Dir(\alpha_1, \ldots, \alpha_N)$ *and when* $k < N$, *then*

$\frac{1}{(1 - \sum_{j=k+1}^{N} U_j)}(U_1, \ldots, U_k) \sim Dir(\alpha_1, \ldots, \alpha_k)$

Here, we note the term that divides each component of the vector $(U_1, \ldots, U_k)$ is $(1 - \sum_{j=k+1}^{N} U_j)$, which is nothing but $\sum_{j=1}^{k} U_j$, since $\sum_{j=1}^{N} U_j = 1$. Assume $\{Z_i\}_{i=1}^{N}$, are N IID Gamma distributed random variables, with $Z_i \sim Gamma(\alpha_i, 1) \; \forall \; i = 1, \ldots, N$. In theorem 1, it was shown that when $U_i = Z_i / (\sum_{j=1}^{N} Z_j) \; \forall \; i = 1, \ldots, N$, then $U =$

$(U_1, \ldots, U_N) \sim Dir(\alpha_1, \ldots, \alpha_N)$. With this construction, we now have:

$$X_i = \frac{U_i}{1 - \sum_{j=k+1}^{N} U_j} = \frac{U_i}{\sum_{j=1}^{k} U_j} \; \forall \, i = 1, \ldots, k \tag{A.27}$$

Since $U_i = Z_i / (\sum_{j=1}^{N} Z_j) \; \forall \, i = 1, \ldots, N$, we put this in A.27 to get:

$$X_i = \frac{U_i}{\sum_{j=1}^{k} U_j} = \frac{Z_i}{\sum_{j=1}^{k} Z_j} \; \forall \, i = 1, \ldots, k \tag{A.28}$$

From theorem 1, we know that $X = (X_1, \ldots, X_k) \sim Dir(\alpha_1, \ldots, \alpha_k)$, which completes the proof. This property of the Dirichlet Distribution is also called the *complete neutral property* [2].

With theorem 3, we can easily compute the expected values and the variances of each of the U_i's. Setting $\sum_{j \neq i}^{N} \alpha_j = \alpha_{-i}$, U_i is distributed $Beta(\alpha_i, \alpha_{-i})$. We compute the expected value as follows:

$$\begin{aligned}
& E(U_i) = \int_0^1 \frac{\Gamma(\alpha_i + \alpha_{-i})}{\Gamma(\alpha_i)\Gamma(\alpha_{-i})} u_i^{\alpha_i - 1} (1 - u_i)^{\alpha_{-i} - 1} u_i \, du_i \\
\implies & E(U_i) = \int_0^1 \frac{\Gamma(\alpha_i + \alpha_{-i})}{\Gamma(\alpha_i)\Gamma(\alpha_{-i})} u_i^{\alpha_i} (1 - u_i)^{\alpha_{-i} - 1} \, du_i \\
\implies & E(U_i) = \frac{\Gamma(\alpha_i + \alpha_{-i})}{\Gamma(\alpha_i)\Gamma(\alpha_{-i})} \int_0^1 u_i^{\alpha_i} (1 - u_i)^{\alpha_{-i} - 1} \, du_i
\end{aligned} \tag{A.29}$$

The integral in A.29 takes the form of another integral studied by Euler [29], and is given by $\frac{\Gamma(\alpha_i + 1)\Gamma(\alpha_{-i})}{\Gamma(\alpha_i + \alpha_{-i} + 1)}$. Also using the fact that when $n \in \mathbb{R}^+, \Gamma(n+1) = n\Gamma(n)$ in A.29, we get:

$$\begin{aligned}
& E(U_i) = \frac{\Gamma(\alpha_i + \alpha_{-i})}{\Gamma(\alpha_i)\Gamma(\alpha_{-i})} \cdot \frac{\Gamma(\alpha_i + 1)\Gamma(\alpha_{-i})}{\Gamma(\alpha_i + \alpha_{-i} + 1)} \\
\implies & E(U_i) = \frac{\Gamma(\alpha_i + \alpha_{-i})}{\Gamma(\alpha_i)\Gamma(\alpha_{-i})} \cdot \frac{\alpha_i \Gamma(\alpha_i)\Gamma(\alpha_{-i})}{(\alpha_i + \alpha_{-i})\Gamma(\alpha_i + \alpha_{-i})} \\
\implies & E(U_i) = \frac{\Gamma(\alpha_i + \alpha_{-i})}{\Gamma(\alpha_i)\Gamma(\alpha_{-i})} \cdot \frac{\alpha_i \Gamma(\alpha_i)\Gamma(\alpha_{-i})}{(\alpha_i + \alpha_{-i})\Gamma(\alpha_i + \alpha_{-i})} \\
\implies & E(U_i) = \frac{\alpha_i}{\alpha_i + \alpha_{-i}}
\end{aligned} \tag{A.30}$$

The variance of U_i is computed similarly. Since $Var(U_i) = E(U_i^2) - E(U_i)^2$, we compute $E(U_i^2)$ first, and then compute the variance, as shown below:

$$
\begin{aligned}
E(U_i^2) &= \int_0^1 \frac{\Gamma(\alpha_i + \alpha_{-i})}{\Gamma(\alpha_i)\Gamma(\alpha_{-i})} u_i^{\alpha_i - 1}(1-u_i)^{\alpha_{-i}-1} u_i^2 du_i \\
\implies E(U_i^2) &= \int_0^1 \frac{\Gamma(\alpha_i + \alpha_{-i})}{\Gamma(\alpha_i)\Gamma(\alpha_{-i})} u_i^{\alpha_i + 1}(1-u_i)^{\alpha_{-i}-1} du_i \\
\implies E(U_i^2) &= \frac{\Gamma(\alpha_i + \alpha_{-i})}{\Gamma(\alpha_i)\Gamma(\alpha_{-i})} \int_0^1 u_i^{\alpha_i + 1}(1-u_i)^{\alpha_{-i}-1} du_i \\
\implies E(U_i^2) &= \frac{\Gamma(\alpha_i + \alpha_{-i})}{\Gamma(\alpha_i)\Gamma(\alpha_{-i})} \int_0^1 u_i^{\alpha_i + 1}(1-u_i)^{\alpha_{-i}-1} du_i \\
\implies E(U_i^2) &= \frac{\Gamma(\alpha_i + \alpha_{-i})}{\Gamma(\alpha_i)\Gamma(\alpha_{-i})} \cdot \frac{\Gamma(\alpha_i + 2)\Gamma(\alpha_{-i})}{\Gamma(\alpha_i + \alpha_{-i} + 2)} \\
\implies E(U_i^2) &= \frac{\Gamma(\alpha_i + \alpha_{-i})}{\Gamma(\alpha_i)\Gamma(\alpha_{-i})} \cdot \frac{\alpha_i(\alpha_i + 1)\Gamma(\alpha_i)\Gamma(\alpha_{-i})}{(\alpha_i + \alpha_{-i})(\alpha_i + \alpha_{-i} + 1)\Gamma(\alpha_i + \alpha_{-i})} \\
\implies E(U_i^2) &= \frac{\alpha_i(\alpha_i + 1)}{(\alpha_i + \alpha_{-i})(\alpha_i + \alpha_{-i} + 1)} \\
Var(U_i) &= E(U_i^2) - E(U_i)^2 = \frac{\alpha_i(\alpha_i + 1)}{(\alpha_i + \alpha_{-i})(\alpha_i + \alpha_{-i} + 1)} - \left(\frac{\alpha_i}{\alpha_i + \alpha_{-i}}\right)^2 \\
\implies Var(U_i) &= \frac{\alpha_i(\alpha_i + 1)(\alpha_i + \alpha_{-i}) - \alpha_i^2(\alpha_i + \alpha_{-i} + 1)}{(\alpha_i + \alpha_{-i})^2(\alpha_i + \alpha_{-i} + 1)} \\
\implies Var(U_i) &= \frac{\alpha_i \alpha_{-i}}{(\alpha_i + \alpha_{-i})^2(\alpha_i + \alpha_{-i} + 1)}
\end{aligned} \tag{A.31}
$$

The derivation of the covariance between two components of an N-dimensional Dirichlet random vector $U = (U_1, \ldots, U_N)$ is a little more involved. From theorem 2, we know that $(U_i, U_j, \sum_{k \neq i,j} U_k) \sim Dir(\alpha_1, \alpha_2, \sum_{k \neq i,j} \alpha_k)$. Let:

$$
\begin{aligned}
V &= \sum_{k \neq i,j} U_k \\
\beta &= \sum_{k \neq i,j} \alpha_k
\end{aligned}
$$

We then have (from theorem 2):

$$
(U_i, U_j, V) \sim Dir(\alpha_i, \alpha_j, \beta) \tag{A.32}
$$

We know $Cov(U_i, U_j) = E(U_i U_j) - E(U_i)E(U_j)$. We know $E(U_i)$ from A.30 and

$E(U_iU_j)$ by definition is given by:

$$E(U_iU_j) = \int_0^1 \int_0^{1-u_j} \frac{\Gamma(\alpha_i+\alpha_j+\beta)}{\Gamma(\alpha_i)\Gamma(\alpha_j)\Gamma(\beta)} u_i^{\alpha_i-1} u_j^{\alpha_j-1} (1-u_i-u_j)^{\beta-1} u_i u_j du_i du_j$$

$$\implies E(U_iU_j) = \int_0^1 \frac{\Gamma(\alpha_i+\alpha_j+\beta)}{\Gamma(\alpha_i)\Gamma(\alpha_j)\Gamma(\beta)} u_j^{\alpha_j} \left[\int_0^{1-u_j} u_i^{\alpha_i} (1-u_i-u_j)^{\beta-1} du_i \right] du_j \tag{A.33}$$

Where we used the fact that $V = 1 - U_i - U_j$. Denoting the inner integral in A.33 by I and letting $a = 1 - u_j$, we have:

$$I = \int_0^a u_i^{\alpha_i} (a-u_i)^{\beta-1} du_i = a^{\alpha_i+\beta-1} \int_0^a \left(\frac{u_i}{a}\right)^{\alpha_i} \left(1-\frac{u_i}{a}\right)^{\beta-1} du_i \tag{A.34}$$

Set $x_i = u_i/a$, so that $du_i = a dx_i$ and the limits of the integral I change from $0 \to a$ to $0 \to 1$. Equation A.34 now becomes:

$$I = a^{\alpha_i+\beta-1} \int_0^1 x_i^{\alpha_i} (1-x_i)^{\beta-1} a dx_i = a^{\alpha_i+\beta} \int_0^1 x_i^{\alpha_i} (1-x_i)^{\beta-1} dx_i \tag{A.35}$$

As noted before, the integral in A.35 takes the form of another integral studied by Euler [29], and is given by $\frac{\Gamma(\alpha_i+1)\Gamma(\beta)}{\Gamma(\alpha_i+\beta+1)}$. Putting this back in A.35, we get:

$$I = a^{\alpha_i+\beta} \int_0^1 x_i^{\alpha_i} (1-x_i)^{\beta-1} dx_i = a^{\alpha_i+\beta} \frac{\Gamma(\alpha_i+1)\Gamma(\beta)}{\Gamma(\alpha_i+\beta+1)} = (1-u_j)^{\alpha_i+\beta} \frac{\Gamma(\alpha_i+1)\Gamma(\beta)}{\Gamma(\alpha_i+\beta+1)} \tag{A.36}$$

Since $a = (1-u_j)$ by definition. Putting A.36 back in A.33 we get:

$$E(U_iU_j) = \int_0^1 \frac{\Gamma(\alpha_i+\alpha_j+\beta)}{\Gamma(\alpha_i)\Gamma(\alpha_j)\Gamma(\beta)} u_j^{\alpha_j} \left[(1-u_j)^{\alpha_i+\beta} \frac{\Gamma(\alpha_i+1)\Gamma(\beta)}{\Gamma(\alpha_i+\beta+1)} \right] du_j$$

$$\implies E(U_iU_j) = \frac{\Gamma(\alpha_i+\alpha_j+\beta)}{\Gamma(\alpha_i)\Gamma(\alpha_j)\Gamma(\beta)} \frac{\Gamma(\alpha_i+1)\Gamma(\beta)}{\Gamma(\alpha_i+\beta+1)} \int_0^1 u_j^{\alpha_j} (1-u_j)^{\alpha_i+\beta} du_j$$

$$\implies E(U_iU_j) = \frac{\Gamma(\alpha_i+\alpha_j+\beta)}{\Gamma(\alpha_j)} \frac{\alpha_i}{\Gamma(\alpha_i+\beta+1)} \int_0^1 u_j^{\alpha_j} (1-u_j)^{\alpha_i+\beta} du_j \tag{A.37}$$

Since $\Gamma(\alpha_i+1) = \alpha_i\Gamma(\alpha_i)$. We find an integral in A.37 which we saw before, and by the same standard result [29], this integral equals $\frac{\Gamma(\alpha_j+1)\Gamma(\alpha_i+\beta+1)}{\Gamma(\alpha_i+\alpha_j+\beta+2)}$. Putting this

back in A.37, we get:

$$
\begin{aligned}
E(U_iU_j) &= \frac{\Gamma(\alpha_i+\alpha_j+\beta)}{\Gamma(\alpha_j)}\frac{\alpha_i}{\Gamma(\alpha_i+\beta+1)}\frac{\Gamma(\alpha_j+1)\Gamma(\alpha_i+\beta+1)}{\Gamma(\alpha_i+\alpha_j+\beta+2)} \\
\implies E(U_iU_j) &= \frac{\alpha_i\Gamma(\alpha_j+1)\Gamma(\alpha_i+\alpha_j+\beta)}{\Gamma(\alpha_j)\Gamma(\alpha_i+\alpha_j+\beta+2)} \\
\implies E(U_iU_j) &= \frac{\alpha_i\alpha_j}{(\alpha_i+\alpha_j+\beta)(\alpha_i+\alpha_j+\beta+1)} = \frac{\alpha_i\alpha_j}{\left(\sum_{k=1}^{N}\alpha_k\right)\left(\sum_{k=1}^{N}\alpha_k+1\right)}
\end{aligned} \tag{A.38}
$$

Since $\beta = \sum_{k\neq i,j}\alpha_k$ by definition, we have $\alpha_i+\alpha_j+\beta = \sum_{k=1}^{N}\alpha_k$. Let $\alpha = \sum_{k=1}^{N}\alpha_k$. The covariance is now given by:

$$
Cov(U_i,U_j) = E(U_iU_j) - E(U_i)(U_J) = \frac{\alpha_i\alpha_j}{(\alpha)(\alpha+1)} - \frac{\alpha_i}{\alpha}\frac{\alpha_j}{\alpha} = -\frac{\alpha_i\alpha_j}{(\alpha^2)(\alpha+1)} \tag{A.39}
$$

The three results (expected value, variance and covariance) are formally stated below:

Theorem 5 *If $U_i = (U_1,\ldots,U_N) \sim Dir(\alpha_1,\ldots,\alpha_N)$, then*

$$E(U_i) = \frac{\alpha_i}{\alpha} \tag{A.40}$$

$$Var(U_i) = \frac{\alpha_i\alpha_{-i}}{(\alpha_i+\alpha_{-i})^2(\alpha_i+\alpha_{-i}+1)} \tag{A.41}$$

$$Cov(U_i,U_j) = -\frac{\alpha_i\alpha_j}{(\alpha^2)(\alpha+1)} \tag{A.42}$$

Where $\alpha = \sum_{k=1}^{N}\alpha_k$ and $\alpha_{-i} = \alpha - \alpha_i$ for $i = 1,2,\ldots,N$

BIBLIOGRAPHY

[1] Victor Aguirregabiria and Pedro Mira. Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156(1):38–67, 2010.

[2] Isabelle Albert and Jean-Baptiste Denis. Dirichlet and multinomial distributions: properties and uses in JAGS. *Analysis*, 31:1141–1155, 2011.

[3] James H Albert and Siddhartha Chib. Bayesian Analysis of Binary and Polychotomous Response Data. 88(88:422):669–679, 1993.

[4] Greg M Allenby, Neeraj Arora, and James L Ginter. On The Heterogeneity of Demand. *Journal of Marketing Research*, 35(3):384–389, 1998.

[5] Greg M. Allenby and Peter E. Rossi. Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89(1-2):57–78, 1998.

[6] James C Anderson and James A Narus. *Business Market Management: Understanding, Creating, and Delivering Value*, volume 3rd. Upper Saddle River, NJ: Pearson Prentice Hall, Pearson Education International, 2004.

[7] Rick L Andrews, Andrew Ainslie, and Imran S Currim. An Empirical Comparison of Logit Choice Models with Discrete Versus Continuous Representations of Heterogeneity. *Journal of Marketing Research*, 39(4):479–487, 2002.

[8] Charles E Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[9] Patrick Bajari, Han Hong, John Krainer, and Denis Nekipelov. Estimating Static Models of Strategic Interactions. *Journal of Business & Economic Statistics*, 28(January 2015):469–482, 2010.

[10] AG Barto. *Reinforcement learning: An Introduction*. 1998.

[11] Richard Ernest Bellman. *Dynamic programming*. Rand research study. Princeton Univ. Press, Princeton, NJ, 1957.

[12] Kristin P Bennett. The Interplay of Optimization and Machine Learning Research. *Journal of Machine Learning Research*, 7:1265–1281, 2006.

[13] Arier Beresteanu, Ilya Molchanov, and Francesca Molinari. Sharp Identification Regions in Models With Convex Moment Predictions. *Econometrica*, 79(6):1785–1821, 2011.

[14] Dimitri P Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific Belmont, MA, 1995.

[15] Nikhil Bhat, Vivek Farias, and Ciamac Moallemi. Non-parametric Approximate Dynamic Programming via the Kernel Method. 2012.

[16] Christopher M Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[17] David Blackwell. Discreteness of Ferguson Selections. *The Annals of Statistics*, 1(2):356–358, 1973.

[18] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[19] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2009.

[20] Sébastien Bubeck. Theory of Convex Optimization for Machine Learning. *arXiv*, 2014.

[21] Randolph E Bucklin and Sunil Gupta. Brand Choice, Purchase Incidence, and Segmentation: An Integrated Modeling Approach. *Journal of Marketing Research*, 29(2):201–215, 1992.

[22] Randolph E. Bucklin and Catarina Sismeiro. Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing. *Journal of Interactive Marketing*, 23(1):35–48, 2009.

[23] Martin Burda, Matthew Harding, and Jerry Hausman. A Bayesian mixed logit-probit model for multinomial choice. *Journal of Econometrics*, 147(2):232–246, 2008.

[24] Siddhartha Chib and Edward Greenberg. Analysis of multivariate probit models. *Biometrika*, 85(2):347–361, 1998.

[25] Pradeep K. Chintagunta. Heterogeneous Logit Model Implications for Brand Positioning. *Journal of Marketing Research*, 31(2):304–311, 1994.

[26] Pradeep K Chintagunta, Dipak C Jain, and Naufel J Vilcassim. Investigating Heterogeneity in Brand Preferences in Logit Models for Panel Data. *Journal of Marketing Research*, 28(4):417–428, 1991.

[27] Melanie Coggan. Exploration and exploitation in reinforcement learning. Technical report, 2004.

[28] Dapeng Cui and David Curry. Prediction in Marketing Using the Support Vector Machine. *Marketing Science*, 24(January 2015):595–615, 2005.

[29] Philip J Davis. Leonhard Euler's Integral: A Historical Profile of the Gamma Function: In Memoriam: Milton Abramowitz. *American Mathematical Monthly*, pages 849–869, 1959.

[30] Mj Dickstein. Efficient provision of experience goods: Evidence from antidepressant choice. 2014.

[31] D. Dzyabura and J. R. Hauser. Active Machine Learning for Consideration Heuristics. *Marketing Science*, 30:801–819, 2011.

[32] P. B. Ellickson and S. Misra. Structural Workshop Paper–Estimating Discrete Games. *Marketing Science*, 30(January 2015):997–1010, 2011.

[33] Paul B Ellickson, Sanjog Misra, and Harikesh S Nair. Repositioning Dynamics and Pricing Strategy. *Journal of Marketing Research*, 2437:1–23, 2012.

[34] Richard Ericson and A Pakes. Markov-perfect industry dynamics: A framework for empirical work. *The Review of Economic Studies*, 62(1):53–82, 1995.

[35] Michael D. Escobar and Mike West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

[36] Michael D Escobar and Mike West. *Computing Nonparametric Hierarchical Models*. Springer, 1998.

[37] Theodoros Evgeniou, Massimiliano Pontil, and Olivier Toubia. A Convex

Optimization Approach to Modeling Consumer Heterogeneity in Conjoint Estimation. *Marketing Science*, 26(6):805–818, 2007.

[38] Hao Audrey Fang. A discretecontinuous model of households' vehicle choice and usage, with an application to the effects of residential density. *Transportation Research Part B: Methodological*, 42(9):736–758, 2008.

[39] Vivek Farias, Denis Saure, and Gabriel Y Weintraub. An approximate dynamic programming approach to solving dynamic oligopoly models. 43(2):253–282, 2012.

[40] Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[41] George M. Furnival and Robert W Jr. Wilson. Regressions by Leaps. *Technometrics*, 16(4):499–511, 1974.

[42] A Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2014.

[43] Samuel J. Gershman and David M. Blei. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, feb 2012.

[44] John Geweke and Michael P Keane. Bayesian inference for dynamic discrete choice models without the need for dynamic programming. *Simulation-Based Inference in Econometrics: Methods and Applications, eds. RS Mariano, T. Schuermann and M. Weeks, Cambridge: Cambridge University Press*, pages 100–131, 2000.

[45] Jayanta K Ghosh and R V Ramamoorthi. *Bayesian nonparametrics*. Springer Science & Business Media, 2003.

[46] JC Gittins. Bandit processes and dynamic allocation indices. In *Statistics*, volume 41, pages 148–177, 1979.

[47] Jc Gittins and Dm Jones. A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, 66(3):561–565, 1979.

[48] S. Gopalakrishna and G. L. Lilien. A Three-Stage Model of Industrial Trade Show Performance. *Marketing Science*, 14(1):22–42, 1995.

[49] Srinath Gopalakrishna, Gary L. Lilien, Jerome D. Williams, and Ian K. Sequeira. Do trade shows pay off? *Journal of Marketing*, 59(3):75, 1995.

[50] Edward Greenberg. *Introduction to Bayesian econometrics*. Cambridge University Press, 2012.

[51] Sachin Gupta and Pradeep K Chintagunta. On Using Demographic Variables to Determine Segment Membership in Logit Mixture Models. *Journal of Marketing*, 31(1):128–136, 1994.

[52] W. Michael Hanemann. Discrete-Continuous Models of Consumer Demand. *Econometrica*, 52(3):541–561, 1984.

[53] H. Hindi. A tutorial on convex optimization. *Proceedings of the 2004 American Control Conference*, 4:3252–3265, 2004.

[54] H. Hindi. A tutorial on convex optimization II: duality and interior point methods. *2006 American Control Conference*, (1), 2006.

[55] Sam K Hui, Peter S Fader, and Eric T Bradlow. Path data in marketing: An integrative framework and prospectus for model building. *Marketing Science*, 28(2):320–335, 2009.

[56] Susumu Imai, Neelam Jain, and Andrew Ching. Bayesian Estimation of Dynamic Discrete Choice Models. *Econometrica*, 77(6):1865–1899, 2009.

[57] Hemant Ishwaran and Lancelot F James. Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[58] Hemant Ishwaran and Mahmoud Zarepour. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390, 2000.

[59] W. A. Kamakura, B.-D. Kim, and J. Lee. Modeling Preference and Structural Heterogeneity in Consumer Choice. *Marketing Science*, 15(2):152–172, 1996.

[60] Wagner A. Kamakura and Gary J. Russell. A Probabilistic Choice Model for Market Segmentation and Elasticity Structure. *Journal of Marketing Research*, 26(4):379–390, 1989.

[61] Wagner A. Kamakura and Gary J. Russell. Measuring brand value with scanner data, 1993.

[62] Wagner A. Kamakura, Michel Wedel, and Jagadish Agrawal. Concomitant variable latent class models for conjoint analysis. *International Journal of Research in Marketing*, (11):451–464, 1994.

[63] Jin Gyo Kim, Ulrich Menzefricke, and Fred M. Feinberg. Assessing Heterogeneity in Discrete Choice Models Using a Dirichlet Process Prior. *Review of Marketing Science*, 2, 2004.

[64] Roger Koenker and Ivan Mizera. Convex optimization in R. *J. Stat. Softw.*, VV(5), 2013.

[65] Leslie Lamport. The part-time parliament. *ACM Transactions on Computer Systems (TOCS)*, 16(2):133–169, 1998.

[66] Leslie Lamport. Paxos made simple. *ACM Sigact News*, 32(4):18–25, 2001.

[67] Peter S H Leeflang, Jaap E Wieringa, Tammo Hendrik Anthonie Bijmolt, and Koen H Pauwels. *Modeling Markets: Analyzing Marketing Phenomena and Improving Marketing Decision Making*. Springer, 2014.

[68] Qi Li and Jeffrey Scott Racine. *Nonparametric econometrics: theory and practice*. Princeton University Press, 2007.

[69] Yang Li and Asim Ansari. A Bayesian Semiparametric Approach for Endogeneity and Heterogeneity in Choice Models. *Management Science*, 60(5):1161–1179, 2014.

[70] Gary L Lilien and Rajdeep Grewal. *Handbook of Business to Business Marketing*. Edward Elgar Publishing, 2012.

[71] Song Lin, Juanjuan Zhang, and John R. Hauser. Learning From Experience, Simply. *Minnesota medicine*, (September):1–19, 2014.

[72] Yangwen Liu and Cinzia Cirillo. An integrated model for discrete and continuous decisions with application to vehicle ownership , type and usage choices. *Transportation Research Board 93rd Annual Meeting. January 12-16, Washington, D.C.*, 12(January 2014):12–16, 2014.

[73] Sanjog Misra and Harikesh S. Nair. A structural model of sales-force compensation dynamics: Estimation and field implementation. *Quantitative Marketing and Economics*, 9:211–257, 2011.

[74] Wendy W Moe. An Empirical Two-Stage Choice Model with Varying Decision Rules Applied to Internet Clickstream Data. *Journal of Marketing Research*, 43(4):680–692, 2006.

[75] Wendy W. Moe and Peter S. Fader. Capturing evolving visit behavior in clickstream data. *Journal of Interactive Marketing*, 18:5–19, 2004.

[76] Wendy W. Moe and Peter S. Fader. Dynamic Conversion Behavior at E-Commerce Sites. *Management Science*, 50(January 2015):326–335, 2004.

[77] Francesca Molinari and Haim Bar. Computational Methods for Partially Identified Models via Data Augmentation and Support Vector Machines. 2015.

[78] Alan L Montgomery, Shibo Li, Kannan Srinivasan, and John C Liechty. Modeling Online Browsing and Path Analysis using Clickstream Data. *Marketing Science*, 23(4):579–595, 2004.

[79] Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[80] Harikesh Nair. Intertemporal price discrimination with forward-looking consumers: Application to the US market for console video-games. *Quantitative Marketing and Economics*, 5:239–292, 2007.

[81] Radford M Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[82] Andriy Norets. Inference in Dynamic Discrete Choice Models With Serially Correlated Unobserved State Variables. *Econometrica*, 77(5):1665–1682, 2009.

[83] Andriy Norets. Estimation of Dynamic Discrete Choice Models Using Artificial Neural Network Approximations. *Econometric Reviews*, 31(January 2015):84–106, 2012.

[84] Peter Orbanz. Lecture Notes on Bayesian Nonparametrics. Technical report, Columbia University, 2014.

[85] Ariel Pakes. Patents as Options: Some Estimates of the Value of Holding European Patent Stocks. *Econometrica*, 54:755–784, 1986.

[86] Ariel Pakes and Paul McGuire. Computing Markov perfect Nash equilibria: Numerical implications of a dynamic differentiated product model. *The RAND Journal of Economics*, 25(4):555–589, 1994.

[87] Young-Hoon Park and Peter S. Fader. Modeling Browsing Behavior at Multiple Websites. *Marketing Science*, 23(January 2015):280–303, 2004.

[88] Warren B Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*, volume 703. John Wiley & Sons, 2007.

[89] David V. Pritchett and Robert E Lucas Jr. Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, volume 1, pages 19–46. Elsevier, 1976.

[90] Adrian E Raftery and Steven Lewis. How many iterations in the Gibbs sampler? *Bayesian Statistics*, pages 763—-773, 1992.

[91] Peter E Rossi, Greg M Allenby, and Rob McCulloch. *Bayesian statistics and marketing*. John Wiley & Sons, 2012.

[92] John Rust. Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica: Journal of the Econometric Society*, 55(5):999–1033, 1987.

[93] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[94] Catarina Sismeiro and Randolph E Bucklin. Modeling Purchase Behavior at an E-Commerce Web Site: A Task-Completion Approach. *Journal of Marketing Research*, 41(3):306–323, 2004.

[95] Inseong Song and Pradeep K. Chintagunta. A Micromodel of New Product Adoption with Heterogeneous and Forward-Looking Consumers: Application to the Digital Camera Category. *Quantitative Marketing and Economics*, 1(1969):371–407, 2003.

[96] Inseong Song and Pradeep K. Chintagunta. A DiscreteContinuous Model for Multicategory Purchase Behavior of Households. *Journal of Marketing Research*, 44(November):595–612, 2007.

[97] Yicheng Song, Nachiketa Sahoo, Shuba Srinivasan, and Chrysanthos Dellarocas. Uncovering Characteristic Paths to Purchase of Consumers. 2016.

[98] Raji Srinivasan. Marketing metrics for B2B firms. In Gary L Lilien and Rajdeep Grewal, editors, *Handbook of Business to Business Marketing*, chapter 38, pages 715–730. 2012.

[99] S. Sriram, P. K. Chintagunta, and M. K. Agarwal. Investigating Consumer Purchase Behavior in Related Technology Product Categories. *Marketing Science*, 29(January 2015):291–314, 2010.

[100] Jan-Benedict E.M Steenkamp and Frenkel Ter Hofstede. International market segmentation: issues and perspectives. *International Journal of Research in Marketing*, 19(3):185–213, 2002.

[101] Erik B Sudderth. *Graphical Models for Visual Object Recognition and Tracking*. PhD thesis, MIT, 2006.

[102] Pl Sundsøy, Johannes Bjelland, Asif M. Iqbal, Alex Pentland, and Yves Alexandre De Montjoye. Big data-driven marketing: How machine learning outperforms marketers' gut-feeling. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8393 LNCS:367–374, 2014.

[103] David M Szymanski, Sundar G Bharadwaj, and P Rajan Varadarajan. Standardization versus Adaptation of International Marketing Strategy: An Empirical Investigation. *Journal of Marketing*, 57(4):1–17, 1993.

[104] Yee Whye Teh. Dirichlet Process. *Encyclopedia of Machine Learning*, pages 280–287, 2010.

[105] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[106] Seshadri Tirunillai and Gerard Tellis. Extracting Dimensions of Consumer Satisfaction with Quality From Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. 2437:1–61, 2014.

[107] George Tomlinson and Michael Escobar. *Analysis of densities*. Citeseer, 1999.

[108] George Andrew Tomlinson. *Analysis of Densities*. PhD thesis, University of Toronto, 1998.

[109] Olivier Toubia, Theodoros Evgeniou, and John Hauser. Optimization-based and machine-learning methods for conjoint analysis: Estimation and question design, 2007.

[110] Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

[111] John N. Tsitsiklis and Benjamin Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 22:59–94, 1996.

[112] Dirk Van Den Poel and Wouter Buckinx. Predicting online-purchasing behaviour. *European Journal of Operational Research*, 166:557–575, 2005.

[113] Hal R. Varian. Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, pages 1–36, 2014.

[114] Michel Wedel and Wagner A. Kamakura. *Market Segmentation: Conceptual and Methodological Foundations*, volume 37. Springer Science & Business Media, 2000.

[115] Thorsten Wiesel, Koen Pauwels, and Joep Arts. Practice Prize Paper-Marketing's Profit Impact: Quantifying Online and Off-line Funnel Progression. *Marketing Science*, 30(4):604–611, 2011.

[116] Yoram Wind. Issues and Advances in Segmentation Research. *Journal of Marketing Research*, 15(3):317–337, 1978.

[117] Kenneth I. Wolpin. An Estimable Dynamic Stochastic Model of Fertility and Child Mortality. *Journal of Political Economy*, 92(5):852, 1984.

[118] Mingan Yang, David B. Dunson, and Donna Baird. Semiparametric Bayes hierarchical models with mean and variance constraints. *Computational Statistics and Data Analysis*, 54(9):2172–2186, 2010.

[119] Jonathan Z Zhang, Oded Netzer, and Asim Ansari. Dynamic Targeted

Pricing in B2B Relationships. *Marketing Science*, 33(October):317–337, 2014.