

STATISTICAL AND MACHINE LEARNING METHODS FOR MULTIVARIATE
PROBLEMS 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

Sharmistha Sikdar

May 2019

© 2019 Sharmistha Sikdar

STATISTICAL AND MACHINE LEARNING METHODS FOR MULTIVARIATE
PROBLEMS IN MARKETING

Sharmistha Sikdar, Ph. D.

Cornell University 2019

DISSERTATION ABSTRACT

My dissertation investigates multivariate response problems in marketing applying both parametric and non-parametric approaches. In marketing research there are several empirical situations that examine multivariate responses. Examples include predicting customer's inter-related decision making to transact on multiple channels, multiple firms' inter-related decision to set prices or enter a market.

In my first essay, I examine a multivariate problem on customers' decision to transact with multiple channels of a firm using a parametric state-space model. I define a customer's channel engagement as a latent semi-Markovian process conditional upon which she decides her multichannel activities. I model the channel activities of website visits, and online and offline purchases using a parametric specification. My research jointly predicts the customer's online visitation behavior, and online and offline purchase propensity. Further, the framework recovers the customer's underlying engagement state with each channel and the expected duration of each state.

The second and third essays of my dissertation are motivated by the restrictions

imposed by parametric methods. In particular, when the response vector is of higher order (> 3) it becomes difficult to parametrically specify the multivariate distribution. Further, parametric methods are ineffective when the dimensionality (or number of covariates) is large and there are more complex interactions, and the response outcomes are sparse. In my second essay, I use the non-parametric multivariate random forests to develop a variable selection procedure for high dimensional problems. I develop new variable importance measures for dimensionality reduction using a recursive feature elimination strategy. In my empirical application on an ecology dataset with sparse observations I find that the proposed measures have higher prediction accuracy than the extant ones.

In my third essay I apply the proposed variable selection method for covariate extraction in a high dimensional marketing application. Here, I examine the inter-related price change decisions of multiple sellers on the Amazon marketplace. I model a series of multivariate regression models using the extracted covariates and compare their predictive performance against the embedded variable selection method of LASSO. I find that the generalized additive model trained on the extracted features outperform LASSO. Further, I provide interpretations of the underlying relationship between the predictors and the multivariate outcome.

BIOGRAPHICAL SKETCH

Sharmistha Sikdar's research interests include the development and application of statistical and machine learning methods to address important substantive questions in marketing. Her dissertation primarily focuses on modeling multivariate response problems both from the customer and firm perspectives. Her dissertation essays include usage and enhancement of methods such as hidden semi-Markov models and multivariate random forests.

She has a Bachelor's degree in Economics with Honors from University of Calcutta and an MSc degree in Quantitative Economics from Indian Statistical Institute, Calcutta. She has over 8 years of industry experience in Analytics specializing in banking and customer analytics in companies such as GE, Citi and Infosys. She is a co-inventor of a patented customer analytics solution for enterprises (patent publication number US8504408).

ACKNOWLEDGMENTS

My dissertation has been made possible by the valuable guidance, support and feedback of some exceptional individuals.

My deepest gratitude goes to two key people - my chair Prof. Vrinda Kadiyali and my mentor Prof. Giles Hooker. During the course of my dissertation, Prof. Kadiyali encouraged me to examine an interesting substantive problem on the Amazon marketplace using non-parametric methods such as random forests - which eventually formed the focus of my second and third essays.

My training under Prof. Giles began in the Fall of 2014 when I had taken his course on statistical computing. Prof. Giles guided me on my second year summer work which resulted in my first essay, and later in the methodological development of my second and third essays. A coach par excellence, Prof Giles has been my bedrock of support and the wind beneath my wings.

My sincerest gratitude goes to Prof. Young Hoon Park for settling me down in my initial years and for his selfless support with the data for my first essay. I sincerely thank Prof. Vithala Rao, Prof. Sachin Gupta and Prof. Jay Russo for their continuous feedback that improved my dissertation work immensely. I especially thank Prof. Stijn van Osselaer for the opportunity to speak in various Brown Bag seminars. I sincerely thank Shantanu Gore, MS CS '19, Cornell, for Amazon web scraping, the Cornell Lab of Ornithology for e-bird data, my PhD student colleagues and Johnson school staff Sara Ashman, Annie Johnston and Terri Whitaker for smoothing out the bumps along the way. I especially thank Prof. Jacob Bien at USC Marshall, Prof. Scott Neslin and Prof. Kusum Ailawadi at Dartmouth for their valuable feedback on my first essay.

This dissertation is dedicated to my life's first mentor- my late father and former faculty of Indian Statistical Institute, Calcutta, Prof. Kripasindhu Sikdar – a rebel of a daughter's life has come a full circle.

TABLE OF CONTENTS

Chapter1: Introduction

1.1. Background.....	12
1.2. Multivariate Response Model: Parametric Methods.....	13
1.3. Multivariate Response Model: Non-parametric Methods.....	17
1.4. Outline of Dissertation Essays.....	19
References.....	22

Chapter 2: A Multivariate Hidden semi-Markov Model of Customer-Multichannel Engagement

2.1. Background.....	24
2.2. Related Literature.....	31
2.3. Data.....	33
2.4. Conceptual Framework.....	37
2.5. Modeling Framework.....	42
2.6. Empirical Application.....	53
2.7. Managerial Implications.....	67
2.8. Contributions, Future Research Directions and Conclusions.....	70
References.....	73
Appendix to Chapter 2.....	78

Chapter 3: Variable Selection and Statistical Inference in Multivariate Random Forests

3.1. Background.....	86
----------------------	----

3.2. Multivariate Regression Trees and Multivariate Random Forests.....	92
3.3. Variable Importance Measures in Multivariate Random Forests.....	95
3.4. Variable Selection and Inference Procedures.....	102
3.5. Simulation Studies.....	105
3.6. Empirical Application.....	113
3.7. Conclusion.....	127
References.....	129

Chapter 4: Investigating Multivariate Price Dynamics: An Application to Amazon Marketplace

4.1. Background.....	132
4.2. Empirical Setting: The Amazon Marketplace.....	138
4.3. Related Literature.....	150
4.4. Modeling Framework.....	158
4.5. Empirical Application.....	164
4.6. Results and Discussion.....	171
4.7. Contributions, Future Research Directions and Conclusions.....	183
References.....	185
Appendix to Chapter 4.....	191

Chapter 5: Conclusion

5.1. Empirical Challenges and Limitations.....	199
5.2. Applications and Extensions.....	204
5.3. Final Remarks.....	206

LIST OF FIGURES

Chapter 2:

2.1. Weekly Aggregate Visit-Purchase Summary.....	36
2.2. Latent State Space Description.....	38
2.3. Hidden semi-Markov Process.....	39
2.4. Poisson versus Geometric CDF.....	46
2.5. Weekly Aggregate Online Visits of Holdout Sample.....	60
2.6. Weekly Aggregate Online Purchases of Holdout Sample.....	61
2.7. Weekly Aggregate Offline Purchases of Holdout Sample.....	62
2.8. State Duration Estimates.....	67
2.9. Online Activities by Decile Ranking.....	68
2.10. Offline Purchases by Decile Ranking.....	69

Chapter 3:

3.1. Multivariate Regression Tree.....	94
3.2. Mean Structure (F test) Importance Distribution of Top 5 Features.....	119
3.3.-3.7. Outcome Difference Importance Distribution of Top 5 Features (by species).....	120-124

Chapter 4:

4.1. Daily Price Trends on Amazon for Instant Pot and Crock-Pot.....	145
4.2. High Level Schema for the Modeling Framework.....	159

LIST OF TABLES

Chapter 2:

2.1. Customer Level Visit and Purchase Summary.....	36
2.2. State Aggregates of HMM Sub-states.....	41
2.3. State Transition Illustration.....	42
2.4. Specifications of Evaluated Models.....	56
2.5. Model Performance on Calibration Data.....	57
2.6. Predictive Ability at Individual Level on Holdout Sample.....	58
2.7. Predictive Ability at Aggregate Level on Holdout Sample.....	59
2.8. Parameter Estimates for the Emissions Model.....	63
2.9. Parameter Estimates for the Bivariate State Transition Model.....	65

Chapter 3:

3.1. Simulation Design for Explanatory and Spurious Variables.....	106
3.2. Variable Rank Ordering under Scenario 1.....	110
3.3. Variable Rank Ordering under Scenario 2.....	112
3.4. Distribution of Sightings Count by Species.....	114
3.5. Test Set Mean Squared Error for Univariate Random Forests.....	115
3.6. Test Set Mean Squared Error for Multivariate Random Forests.....	115
3.7. Top Ranked Features.....	118

Chapter 4:

4.1. Description of Data Tracked.....	141
4.2. Average Brand Characteristics.....	143
4.3. 3P Seller Features.....	144
4.4. Final Clustering Variables.....	147
4.5. Seller Distribution and Mean Cluster Profiles.....	148

LIST OF TABLES (cont'd)

Chapter 4 (cont'd):

4.6. Seller type- Brand Panel Decomposition.....	150
4.7. Variable Selection Methods and Relevant Literature.....	155
4.8. Covariates Specification.....	167
4.9. Mean Squared Error at 30 th iteration of the RFE Strategy.....	168
4.10. Mean Squared Error on Test Set for Price Change Magnitude.....	172
4.11. GAM Estimation Results for Amazon.....	174
4.12. GAM Estimation Results for Cluster 1 (“Established”).....	176
4.13. GAM Estimation Results for Cluster 2 (“Small-scale”).....	177
4.14. GAM Estimation Results for Cluster 3 (“New Entrants”).....	178
4.15. GAM Estimation Results for Cluster 4 (“Multibrand”).....	180
4.16. GAM Estimation Results for Cluster 5 (“FBA Sellers”).....	181
4.17. Significant Determinants of Price Change Magnitude.....	183

CHAPTER 1

INTRODUCTION

My doctoral dissertation focuses on extending extant and building new statistical machine learning methods for multivariate response problems in marketing. In this introductory chapter I will first briefly discuss the constituents of a multivariate response model, provide a background of various parametric approaches and the empirical applications in the marketing literature. I will then discuss some of the challenges associated with the parametric methods of estimating multivariate models and review alternative non-parametric approaches. I will conclude the chapter with a general outline of my dissertation essays that examine some of the extensions and methodological modifications to multivariate models.

1.1. Background

To motivate my research focus on multivariate problems I will begin with some application examples in marketing. First, suppose a firm wants to determine its customers' future transaction behavior. The firm wants to be able to predict the time a customer is likely to transact and the quantity or transaction size. Second, suppose a firm retails through multiple channels – online website and offline store. The firm wants to determine which of the two channels, either online or offline is a customer likely to transact next from. Third, suppose two firms retail an identical product through an ecommerce marketplace, say Amazon. Both firms want to determine how much discount to provide on the product sold such that neither loses sales to its rival. In all the three examples above, the response outcome of interest is a vector. That is, the response vector either comprises transaction time and size, or transaction on online and offline channels or discount decision making of two rival firms. Further, in all the

three scenarios above, the individual response variables, such as in the first example, transaction time and transaction amount, are likely to be correlated or explained by a common set of factors or covariates. The primary objective of a multivariate response model is two-fold – first, determine the joint occurrence or covariation of multiple response variables and second, identify factors or covariates that contribute to this covariation or co-occurrence. Most applications in marketing measure the multivariate response and the associated covariates sequentially over time in a longitudinal panel to investigate the relationship between predictors and outcome and for prediction of future outcomes. In the following sections, I will discuss some of the extant parametric specifications for multivariate response, challenges with such specifications, and discuss some of the alternative non-parametric forms.

1.2. Multivariate Response Models: Parametric Methods

A multivariate response ($K \times I$) vector comprises K response variables of interest to be modeled. These response variables could be all continuous (e.g. percentage price discount), all quantal (e.g., transact (1) or no transact (0)), count (e.g. number of visits to online website) or a combination of types (e.g. transaction time and transaction size). Further, assume we have a set of M explanatory variables or covariates that can explain or determine the covariation of the multiple response variables.

For the continuous multivariate response case, the general specification for the k^{th} outcome can be written as,

$$y_{kt} = f(\mathbf{X}_t) + \varepsilon_{kt} \quad (1)$$

where we define $\mathbf{y}_t = (y_{1t}, y_{2t}, \dots, y_{Kt})$ as the ($K \times I$) multivariate response vector and \mathbf{X}_t as the ($M \times I$) vector of explanatory variables measured at time t . The vector of

residuals is given as $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \varepsilon_{2t}, \dots, \varepsilon_{Kt})$ with appropriate distributional assumptions. The residuals are assumed to be multivariate normal, i.e., $\boldsymbol{\varepsilon}_t \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ leading to multivariate normal responses. The covariance of i^{th} and j^{th} responses are captured by the $(i,j)^{th}$ element of the matrix $\boldsymbol{\Sigma}$.

For the discrete or count response case, equation (1) can be modified as,

$$E(y_{kt} | \mathbf{X}_t) = \exp(f(\mathbf{X}_t)) \quad (2)$$

where y_{kt} is assumed to follow a Poisson process.

For quantal responses, the multivariate response model can be specified as,

$$y_{kt} = \begin{cases} 1, & \text{if } f(\mathbf{X}_t) + \varepsilon_{kt} > 0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

In the quantal case, different assumptions imposed on the residual vector $\boldsymbol{\varepsilon}_t$ can lead to alternative specifications of the response function. For instance, when $\boldsymbol{\varepsilon}_t$ is assumed to follow a multivariate normal distribution, the response function follows multivariate probit (e.g. Ashford and Sowden 1970, Lesaffre, Verbeke and Molenberghs 1994; see Edwards and Allenby 2003 and Manchanda et al. 1999 for applications in marketing). When the residual vector is assumed to follow type I extreme value or Gumbel distribution, the response function is a multivariate logit (Glonek and McCullagh 1995; see Schweidel et al. 2014 for application in marketing).

If the multivariate response vector is composed of alternative variable types, i.e., continuous, count and quantal, each response type can be specified as given in (1) – (3) and the multivariate distribution can be specified either as a product of conditional marginals or as copula (Joe 1997).

Apart from the alternative functional specifications of multivariate response models and the distributional assumptions of residuals, most applications assume a parametric relationship between the covariates and the response. That is, $f(\mathbf{X}_t) = \boldsymbol{\beta}_k^T \mathbf{X}_t$, where $\boldsymbol{\beta}_k$ is the corresponding $(M \times 1)$ vector of coefficients associated with the k^{th} response.

I make a brief digression here to differentiate between static versus dynamic models, especially since dynamic models pertain to the first essay of my dissertation (see Chapter 2). When the coefficient vector $\boldsymbol{\beta}_k$ is assumed fixed across the longitudinal panel for the k^{th} response variable the model is said to be static. If the coefficient vector is allowed to vary with time, i.e., $\boldsymbol{\beta}_{kt}$ the model is dynamic. In the marketing literature dynamic models have been examined using state-space representation such as hidden Markov models (HMMs) (e.g. Montgomery et al. 2004, Moon et al. 2007, Netzer et al. 2008). In addition to the response variables and covariates, a state-space model comprises state variables. The state variables are time-varying and depend on own values at any given time and covariates. The response outcomes in state-space models depend on the values taken by the state variables over time. Therefore, the coefficients associated with the explained part of the response model vary with the time-varying changes in the state variables. The state-space models such as HMMs consider a latent state variable construct and propose that the values of the response variables are conditional upon the value of the latent state. Multivariate extensions of such state-space models have modeled multiple response variables conditional upon a latent state variable, e.g., transaction incidence and

transaction amount conditional upon latent attrition (e.g. Schweidel and Knox 2013). However, to the best of my knowledge there has not been any study on the multivariate extension of the latent state space. In my first essay, I examine a multivariate extension of the latent state space where I jointly model two latent processes – customer’s latent engagement on online and offline channels of a multichannel firm.

A second digression is on the linearity and strictly parametric specification of covariates, i.e., $\beta_k^T X_t$. Though this specification is easier to model it can impose restrictions; especially in cases where the underlying relationship between covariates and responses is not known or potentially non-linear. This restriction along with other challenges associated with parametric specifications of multivariate response models (see section 1.2.1 below) is the motivation for my research in the second and third essays of my dissertation (Chapters 3 and 4). In section 1.3 below I discuss some of the non-parametric specifications of multivariate models, particularly, tree based ensembles such as multivariate random forests (MVRFs).

1.2.1. Some Challenges with Parametric Methods

In the marketing literature, parametric specifications of multivariate models in lower dimension (i.e., $K = 2, 3$) have been examined successfully on a variety of empirical problems (e.g. Danaher 1991, Manchanda et al. 1999, Park and Fader 2004, Danaher and Hardie 2005, Schweidel et al. 2014 etc.). However, multivariate distributions in higher dimension ($K > 3$) can be both difficult to specify and harder to estimate. The specification of multivariate distributions can be especially challenging

for higher order if the response outcomes are of different types (i.e., quantal, discrete and continuous). Some of these complex specifications for non-normal multivariate responses using copulas have been examined in the statistics literature and applied in marketing (see Joe 1997 for theoretical construction of copulas; Danaher and Smith 2011, Park and Gupta 2012 for applications of copulas in marketing). Therefore most application problems restrict examination to bivariate or trivariate response dependence.

Additionally these methods cannot identify complex non-linear patterns and relationships in the data and provide meaningful analyses if the data contains missing values among measured variables (Cutler et al. 2007). Further, for high-dimensional problems with large number and type of explanatory variables, parametric methods such as generalized linear model (GLM) may be difficult to build (De'ath and Fabricius 2000).

1.3. Multivariate Response Model: Non-Parametric Methods

In statistics and machine learning literature, non-parametric methods such as random forests (Breiman 2001) and gradient boosted trees (Friedman 2001) have been applied to high dimensional problems with higher-order and non-linear interactions. Multivariate extensions of such non-parametric specifications have been proposed in terms of multivariate random forests (or MVRF; see Segal 1992 and De'Ath 2002 for multivariate regression trees; Segal and Xiao 2011) and multivariate gradient boosted trees (or MVBT; Miller et al. 2016).

In multivariate tree-based methods, the general idea is to divide the data into sub-

groups, where each sub-group is homogeneous with respect to both the predictors and the multivariate responses (Segal and Xiao 2011). However, building a single tree to derive a set of prediction rules is subject to variance of the underlying sample. The ensemble versions of random forests and gradient boosted trees remedy this problem using bagging (Breiman 2001) and boosting (Friedman 2001) methods. In bagging (or sub-bagging; see Andonova et al. 2002, Mentch and Hooker 2016) method samples (for subbagging samples with size less than training set size) are bootstrapped from the training data. Each sample is used to build a tree and the predictions are averaged across the ensemble. In boosting, the ensemble building occurs in a stepwise iterative manner, where output from a prior tree is used to build a refined tree in sequence.

While such non-parametric methods provide more accurate predictions, unlike parametric regression based methods these are not as well suited for interpretation or explanation of relationship between covariates and outcome. The mechanism of tree build up, though motivated to obtain homogeneous subgroups, can often lead to a large number of predictors in the final prediction rule. To assess the relationship between predictors and the outcome variable and especially to determine the explanatory power of each predictor, variable importance measures (or VIMs) have been proposed and defined in a number of ways (Breiman 2001, Friedman 2001). Depending on the measure used each predictor is assigned an importance score for an ensemble and the variables are ranked in terms of the chosen importance measure. In many disciplines such as biological statistics and genomics, the VIMs are used for variable selection in high dimensional problems (e.g. Strobl et al. 2007). The variable selection methods using VIMs are typically a recursive feature elimination strategy

(Guyon et al. 2012) where the variables with lower importance scores are removed using an iterative ensemble build. The reduced variable set can then be used as inputs to parametric regression models to determine significance of coefficients and interpret the underlying relationship between covariates and outcome.

1.4. Outline of the Dissertation Essays

My dissertation comprises three essays that address and resolve alternative multivariate problems using extensions of both parametric and non-parametric approaches. A brief outline of the essays is as follows.

In my first essay, Chapter 2, I investigate a consumer side problem in multichannel marketing. Specifically, my objective is to predict the customers' online and offline channel choices conditional upon their latent engagement with each channel. I define a customer's channel engagement as a latent stochastic process that can change over a defined state space. I simultaneously model two latent engagement processes- one for each channel- to generate a bivariate state space model. Further, the outcomes that I model in this research are the customer's online website visits, and online and offline purchases. This leads to a multivariate response model. In this research, I model the multivariate response using parametric specifications and the underlying multichannel engagement as a pair of hidden semi-Markov processes. Using this framework, my research jointly predicts the customer's online visitation behavior, and online and offline purchase propensity. Further, the model recovers the customer's underlying engagement state with each channel and the expected duration of each state.

My second (Chapter 3) and third (Chapter 4) essays are motivated by the

restrictions of multivariate parametric methods. In particular, in my second essay, I examine the general case of multivariate response vector of higher order (> 3) and high dimensional data with large number of covariates. The objective of this research is to develop a variable selection procedure to remove redundant covariates and improve prediction accuracy using a non-parametric tree-based ensemble method of MVRF. A second objective is to explain the important predictors in the final prediction rule. I apply the MVRF to develop new variable importance measures using the split improvement criterion. That is, a variable's importance is measured by its ability to combine homogeneous sub-groups of the multivariate response outcome when splitting a node. In order to apply the proposed VIMs for dimensionality reduction I develop a variable selection method using a recursive feature elimination strategy. I compare the predictive performance of the proposed importance measures against some of the extant measures when used for variable selection. In my empirical examination, I apply the proposed methods on ecology (eBird) data for co-occurrence of multiple migrant bird species, some of which are rare species with sparse sightings. The multivariate response in this data is the count of sightings of multiple migrant bird species by an observer group. I find that the proposed importance measures select variables that give higher prediction accuracy than the extant measures. Further, I also propose statistical inference procedures using the proposed importance measures.

In my third essay, Chapter 4, I investigate a marketing application of a multivariate response model with high dimensionality and complex interactions. More concretely, I examine the price dynamics of sellers on an ecommerce site, such as Amazon. Here the multivariate response comprises the price change decisions of Amazon and the

third-party sellers in a chosen category. In this empirical setting, there are more than three seller groups so that the multivariate model is of dimension higher than a trivariate distribution. Additionally, the data is high-dimensional and some of the seller groups do not change prices as frequently, leading to sparsity in response outcomes. Therefore, I apply the variable selection method using the importance measures proposed and developed in my second essay. The variable selection method serves as a pre-processing step to extract the key predictor variables. In the second step, I introduce the predictor variables sequentially in a series of time-series based multivariate regression models, e.g. generalized additive model or GAM (Friedman, Hastie, and Tibshirani. 2001), VAR-X (e.g. Srinivasan et al. 2004) and compare their predictive performance against the embedded variable selection method of LASSO (e.g. Tibshirani 1996). I find that the time-series extension of the multivariate GAM with the variables selected based on the proposed importance measures outperform traditional methods such as LASSO. Further, based on the functional relationship of the covariates in the GAM, I provide interpretations of the underlying relationship between the covariates and the outcome variables.

In the concluding chapter to my dissertation, Chapter 5, I discuss the research limitations, applications and propose future research directions.

REFERENCES

- Andonova S, Elisseeff A, Evgeniou T, Pontil M (2002, July) A simple algorithm for learning stable machines. In *ECAI* (pp. 513-517).
- Ashford JR, Sowden RR (1970) Multi-variate probit analysis. *Biometrics*, 535-546.
- Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. *Ecology*, 88(11), 2783-2792.
- Danaher PJ (1991) A Canonical Expansion Model for Multivariate Media Exposure Distributions: A Generalization of the "Duplication of Viewing Law". *J. Mkt. Res.*, 361-367.
- Danaher PJ, Hardie BGS (2005) Bacon with your eggs? Applications of a new bivariate beta-binomial distribution. *American Stat.*, 59.4 : 282-286.
- Danaher PJ, Smith MS (2011) Modeling multivariate distributions using copulas: applications in marketing. *Marketing Sci.*, 30.1 (2011): 4-21.
- De'Ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81(11), 3178-3192.
- De'Ath G (2002) Multivariate regression trees: a new technique for modeling species–environment relationships. *Ecology*. 83(4), 1105-1117.
- Edwards YD, Allenby GM (2003) Multivariate analysis of multiple response data. *J. Mkt. Res.*, 40.3 (2003): 321-334.
- Friedman J, Hastie T, Tibshirani R (2001) *The elements of statistical learning*. Vol. 1. New York: Springer series in statistics.
- Glonek GF, McCullagh P (1995) Multivariate logistic models. *J R Stat. Soc.: Series B (Methodological)*, 57(3), 533-546.
- Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. *Machine learning*. 46(1-3), 389-422.
- Joe H (1997) *Multivariate models and multivariate dependence concepts*. Chapman and Hall/CRC.
- Lesaffre E, Verbeke G, Molenberghs G (1994) A sensitivity analysis of two multivariate response models. *Comp. stat. & data analysis*, 17(4), 363-391.

- Manchanda P, Ansari A, Gupta S (1999) The “shopping basket”: A model for multicategory purchase incidence decisions. *Marketing Sci.*, 18(2), 95-114.
- Mentch L, Hooker G (2016) Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.*, 17(1), 841-881.
- Montgomery AL, Li S, Srinivasan K, Liechty, JC (2004) Modeling online browsing and path analysis using clickstream data. *Marketing Sci.* 23(4): 579-595.
- Moon S, Kamakura, WA, Ledolter J (2007) Estimating promotion response when competitive promotions are unobservable. *J. Mkt. Res.*, 44(3), 503-515.
- Netzer O, Lattin JM, Srinivasan V (2008) A hidden Markov model of customer relationship dynamics. *Marketing Sci.* 27(2): 185-204.
- Park YH, Fader PS (2004) Modeling browsing behavior at multiple websites. *Marketing Sci.*, 23.3: 280-303.
- Park S, Gupta S (2012) Handling endogenous regressors by joint estimation using copulas. *Marketing Sci.*, 31(4), 567-586.
- Segal MR (1992) Tree-structured methods for longitudinal data. *J. American Stat. Assoc.* 87(418), 407-418.
- Segal M, Xiao Y (2011) Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 80-87.
- Schweidel DA, Knox G (2013) Incorporating direct marketing activity into latent attrition models. *Marketing Sci.* 32(3): 471-487.
- Schweidel DA, Park YH, Jamal Z (2014) A multiactivity latent attrition model for customer base analysis. *Marketing Sci.*, 33.2: 273-286.
- Srinivasan S, Pauwels P, Hanssens DM, Dekimpe MG (2004) Do promotions benefit manufacturers, retailers, or both? *Management Sci.*, 50.5: 617-629.
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat. Soc.: Series B (Methodological)*, 267-288.
- Zhang H (1998) Classification trees for multiple binary responses. *J. American Stat. Assoc.*, 93(441), 180-193.

CHAPTER 2

A MULTIVARIATE HIDDEN-SEMI MARKOV MODEL OF CUSTOMER- MULTICHANNEL ENGAGEMENT

2.1. Background

Customers often use multiple channels for information search on products and purchase activities. The latest business statistics on multichannel ecommerce note that around 36% of US customers use various combinations of online and in-store channels for browsing and purchase activities (Statista 2017). Marketing scientists have examined multichannel customer behavior in different empirical and substantive contexts. Studies have investigated customer channel loyalty (Danaher et al. 2003, Inman et al. 2004), effects of online search activities on aggregate online and offline sales (Biyalogorsky and Naik 2003), customer channel choices and migration patterns (e.g. Thomas and Sullivan 2005, Ansari et al. 2008), channel attribution (e.g. Li and Kannan 2014), marketing effects on online and offline sales (e.g., Zhang and Wedel 2009, Dinner et al. 2014, Lewis and Reiley 2014, Montaguti et al. 2015), multichannel customer profitability (Kushwaha and Shankar 2013), and offline showroom opening effects on online-first retailers (Bell et al. 2017).

In recent years, the business press has reported a surge in interest among marketers in improving the multichannel customer's *engagement* (e.g. Business 2017, Vocalcom 2018). The marketer's ultimate goal for improved multichannel customer engagement is tied to higher conversions, expanded relationship, loyalty, referrals, up-

sells, and renewals (e.g. Merkle 2011). Multichannel retailers thus want to understand their customers' multichannel engagement and how this translates to channel choices for future interactions with the firm. In both industry and academia, *customer engagement* has been defined under various constructs. In academia, marketing scientists have defined engagement as the psychological state of the customer in her relationship with the firm (e.g. Patterson et al. 2006, Brodie et al. 2011), the intensity of customer interaction with the firm (Vivek et al. 2012) and as an attitudinal process that manifests in the customer's interactions with the firm (Kumar 2013, Kumar and Pansari 2016).

In our research, we define the customer's "*channel engagement*" as a *latent* and *dynamic* attitude or predisposition towards the firm's channel that can vary over time with evolving experience, needs and intrinsic preferences. We propose that the customer's channel engagement is binary with "high" and "low" states, each with a different duration. The customer's channel engagement thus determines her observed channel interactions including online web visits, and online and offline purchase behavior. We explain this with an example. Suppose a customer has a positive purchase experience with the online channel of the firm, e.g., timely delivery of an online purchase. As a result on the next purchase occasion, under similar needs the customer may re-engage with the online channel. However, suppose at a different purchase occasion the customer's needs have changed, e.g., an immediate order fulfilment need for an unplanned event. Due to the change in her needs, in this purchase cycle the customer may consider engaging with the offline store. In addition, the customer's intrinsic preference for a channel may vary by products. For instance,

she may want to purchase certain products from the firm's physical store, e.g., products with "touch and feel" attributes (see Lal and Sarvary 1999). Thus, the customer's channel engagement can vary over time under alternative purchase decisions. Further, for some purchases the customer may spend more time gathering information on a channel, i.e., have a longer duration of engagement. In some other cases, there may be "spur of the moment" purchases with shorter engagement duration. Targeting such a customer for channel-specific and channel-integrated promotion decisions may depend on the following: Is the customer likely to engage with the online channel next week? If yes, is the customer expected to make online visits only or also an online purchase? If she is in a state of "high" online (or offline) engagement, how long is she expected to be in that state? Or is the customer likely to be "not" engaged with either channel? If the customer is not engaged with either channel is she likely to be at attrition risk?

Our research thus addresses two key behavioral aspects of multichannel customers – their multichannel engagement and the associated duration of engagement. Specifically we propose that a customer's channel engagement is a hidden semi-Markov process (HSM) that can vary over time between two discrete states of "high" and "low". An HSM process is a stochastic process over a discrete state space. Once the process enters a state it stays in that state for a period chosen from a pre-determined duration distribution before jumping to a new state chosen in a Markovian manner. Thus, to build a semi-Markov chain a duration distribution is assigned to each state. In the context of multichannel marketing, a customer's engagement with a channel can transition in jumps between the states of "high" and

“low”. When a customer is in a period of high engagement with a channel we are likely to observe channel specific activities. Conversely, when the customer enters a period of low engagement, she is likely to not visit or buy from that channel. Our empirical setting uses data from an Asian high-end beauty care brand that retails through its online website and physical stores. The firm tracks the customers’ visits to its websites, and purchases made both online and at the offline physical stores. Using a multivariate hidden semi-Markov (HSM) framework, we simultaneously model the customer’s latent engagement transitions on online and offline channels based on the observed activities of online visits, online and offline purchase incidences. Thus, we jointly model the customer’s state-dependent observed activities as a multivariate outcome.

An important aspect of the dynamic channel engagement process that our methodological framework addresses is the associated engagement *stickiness* or duration. Channel engagement duration can be attributed to a number of factors. These include the customer’s purchase deliberation (e.g. Putsis and Srinivasan 1994), her information needs (e.g. Lal and Sarvary 1999), her shopping strategy (e.g. Janiszewski 1998; Moe 2003; Moe and Fader 2004), her search behavior (e.g. Johnson et al. 2004), and her intrinsic preferences such as channel loyalty (Gensler et al. 2007). While inferring engagement states based on observed activities provides useful insights into a customer’s channel specific behavior, examining the duration of engagement has important consequences for marketers. First, we note that not all channel activities are observed by the firm. For instance, the firm does not record a non-purchase visit to a physical store. However, one can hypothesize that a customer is more likely to visit an

offline store when she is in a state of “high” engagement with the offline channel. Second, by our definition channel activities are not likely to be associated with the “low” engagement state. Determining the duration of “low” engagement on a channel provides the firm an understanding of the length of customer’s “downtime” period from channel related activities. This will help the firm appropriately time marketing interventions. Third, a customer with a long duration of “low” engagement on both channels is likely to be at attrition risk, i.e., may have switched over to competition.

Our research achieves the following. First, our model estimates the customer’s engagement state with online and offline channels respectively based on the multiple observation processes of online visits, and online and offline purchase incidences. Second, using the HSM framework we explicitly examine the distributional properties of engagement duration with each channel. We compare model performance based on alternative duration distributions and choose the one that best fits the observed data. Third, we provide a step-ahead forecast of the customer’s future channel activities of online website visits and online and in-store purchase incidences.

In this essay, we employ an HSMM estimation method using the notion of “sub-states”, originally developed for singular discrete time HSM processes by Langrock and Zucchini 2011 (henceforth L&Z 2011). Our application extends this method to simultaneously model two HSM processes – the customer’s engagement in the online and offline channels. A hidden semi-Markov model or HSMM (e.g. Yu and Kobayashi 2003) is an extension of the conventional hidden Markov model or HMM (e.g. Montgomery et al. 2004, Netzer et al. 2008 etc.) where the latent process is semi-

Markov. In a Markov process the state duration distribution is necessarily geometric¹ and the state transitions are assumed to occur after the modal duration of one period². HSMM relaxes this assumption and allows evaluation of alternative duration distributions. Further the duration distribution is chosen based on model fit (e.g. Russell and Cook 1987; Yu and Kobayashi 2003, Yu 2010).

We evaluate a series of models based on alternative duration distributions and nested conditions. In our model comparisons, we examine the distributional properties of state duration using two discrete distributions- Poisson and geometric. We evaluate the model fit and predictive ability of these two distributions under alternative state transition specifications. If a given distribution better describes the state duration properties, it will be able to better infer the state transitions and hence better forecast the state-dependent observation process. The proposed Poisson model consistently predicts better than the best performing geometric model both at the individual and aggregate levels. Further the week by week forecast reveals that the Poisson model is also able to capture seasonal trends better than the geometric.

Our research is useful to a multichannel firm that wants to design targeted and more effective channel interventions. By applying our method the firm can predict whether a customer is likely to visit online, her estimated number of online visits and her likelihood of making an online (or offline) purchase in the next time period. Apart from the prediction of the observed activities, the marketer will also be able to infer the customer's underlying channel engagement state to decide how to design the marketing interventions. For instance, if a customer is predicted to be high on online

¹ This follows from the memorylessness property of Markovian processes.

² Mode of the geometric distribution is 1

engagement in the next week, the firm may decide to target the customer for an online-only coupon or a multichannel promotion to improve conversion rates. Explicitly estimating the duration distribution of the customer's channel engagement will help determine the right time interval for an intervention. Knowing the offline engagement duration also helps the firm infer the likely time period of offline non-purchase visits. This can lend important insights into a customer's offline channel visit tendencies and thus target customers for in-store promotions. Finally, predicting the channel duration can also help determine the customers at attrition risk; a customer in a low engagement state for a long period is likely to be at higher attrition risk and can thus be targeted for appropriate marketing actions.

Our framework makes important methodological contributions to the marketing literature. To the best of our knowledge we are the first to propose and examine a HSMM with a multivariate state dependent emissions process conditional on two simultaneous latent processes. Based on our empirical setting, we assume that each latent channel engagement process transitions between two discrete states of "high" and "low". The assumption of two states can be flexibly relaxed and extended to more states under different empirical situations. Our work is also the first in marketing science to exploit the L&Z 2011 notion of sub-states and state-aggregates for joint modeling of state transition and state duration (see section 2.4). Additionally, our application of this estimation strategy can be easily implemented with the more familiar HMM likelihood (MacDonald and Zucchini 1997) employed by marketing scientists, while incorporating the flexibility of semi-Markovian processes. Finally, by applying the semi-Markovian property, our model can explicitly measure the mean

duration spent in each state of the latent process in addition to making inferences about the latent state and forecasting future observed channel activities.

In the remainder of the chapter, section 2.2 summarizes the related literature. In section 2.3, we discuss the data and summary statistics to motivate our research problem. In section 2.4, we provide the conceptual framework of the HSMM process and explain the state duration modeling strategy. We present the modeling framework in section 2.5 and discuss the results in section 2.6. We discuss the marketing implications in section 2.7 and conclude with future research directions in section 2.8.

2.2. Related Literature

Marketing scientists have applied hidden Markov models to study various dynamic behaviors. Some of the early works use HMMs to measure web browsing (Montgomery et al. 2004), unobserved competitor promotions (Moon et al. 2007), customer-firm relationships (e.g. Netzer et al. 2008), physicians' new drug prescriptions (Montoya et al. 2010), and customer's service portfolio choice (Schweidel et al. 2011) to name a few. In studies of dynamic latent attrition with pre-defined discrete states, HMMs have been shown to have better predictive performance than static latent class models (e.g. Schweidel and Knox 2013, Schweidel et al. 2014). Much of the marketing literature has examined HMMs with singular hidden processes defined by single Markov chains and univariate state space. Multivariate extensions of HMMs have been examined from the perspective of multivariate state dependent emissions processes. Examples of multivariate HMMs include studies on buyer-seller relationships in B2B settings (Zhang et al. 2014), direct marketing effects on multichannel customer retention (Chang and Zhang 2016), and effects of product

types on multichannel customer learning and profitability (Chang et al. 2017). Multivariate problems of studying household characteristics based on latent lifecycle changes have been examined as multinomial HMMs (e.g. Du and Kamakura 2006).

However, HMMs do not explicitly model the state duration or examine its distributional properties. By definition, a Markovian process is memoryless so that the state transition at a given time point depends only on the state at the prior time point. Thus, the state duration of a Markovian process is implicitly assumed to have a geometric decay with the modal value of duration to be one period. However, this assumption may not be consistent with the known duration distributions of the observation sequences being modeled (Johnson 2005). For instance, in the context of multichannel shopping, a customer may engage with a channel for a brief period and then disengage for a much longer duration. In case of a semi-Markovian process the state transitions can occur after some duration spent in the prior state. Thus, the assumption of memorylessness is relaxed. HSMMs can thus be employed to explicitly model state duration and study its distributional properties using any arbitrary distribution. Though duration times in marketing are of interest across multiple decision contexts (e.g. Helsen and Schmittlein 1993), there has been limited methodological research done on duration models. Some of the notable contributions are duration modeling using hazard rate models (Helsen and Schmittlein 1993), purchase deliberation duration (Putsis and Srinivasan 1994), visual attention using continuous time Markov or semi-Markov chains (e.g. Liechty et al. 2003) and website visit duration (Danaher et al. 2006).

In other disciplines, some of the noteworthy work on HSMM applications

include mobility tracking in wireless networks (Yu and Kobayashi 2003), detecting anomaly on user browsing behavior (Xie and Yu 2009), and rainfall seasonality (Sansom and Thomson 2007). However, there is a substantial computational cost associated with estimating HSMMs (e.g. Johnson 2005). Further, in the HSMM framework, covariate modeling tends to be more difficult due to separate models of state transitions and state duration (e.g. Langrock and Zucchini 2011). The expanded state HMM, or ESHMM, makes use of state aggregates, i.e., collection of sub-states that captures duration of each semi-Markovian state (e.g. Cox and Miller 1965, Russell and Cook 1987). To alleviate the computational complexities in estimating HSMMs, ESHMMs have been shown to be good alternatives to model parametric duration distributions (Johnson 2005).

We model the customer's engagement in online and offline channels as two simultaneous latent or hidden semi-Markovian processes. Since each process is associated with hidden states, the joint modeling of the two processes leads to a bivariate state space. In addition, the state dependent emissions processes of online visits, online and offline purchase incidences are modeled as a multivariate distribution. Our research thus extends the multivariate HMMs by simultaneously modeling multiple hidden processes and multivariate state-dependent emissions in a semi-Markovian framework.

2.3. Data

The data for our study come from an Asian beauty care firm that has its retail operations on both online and offline channels. We have a longitudinal panel of six

months from July 2015 until December 2015 capturing daily website browsing and online & in-store purchase activities of individual customers. For our research, the unit of time is a calendar week, which gives us a 27-week data period. We measure visits at the daily level and report the count of visits in the data period. Therefore, if a customer visits the website multiple times a day, we measure this as a single visit count. We measure purchase incidence by channel at a weekly level. Therefore, multiple purchases on a given channel in a week are aggregated as a single purchase incidence on that channel for that week. However, the number of customers making multiple purchases within a week on a given channel is very low ($< 2\%$ of customers). The data recorded by the firm were on customers who have visited the website at least once. The data provided to us has a total of 13,501 customers, who make a total of 23,833 website visits and 2,294 weekly purchases across the online and offline channels for the data period. Of the total purchases, around 70% occur via the online channel and the rest offline. We observe substantial sparsity and dispersion in the web visits and online-offline purchase data. To capture a broad spectrum of customers (from those who are online only to those who also purchase from the offline store), we sample customers who have made at least 2 website visits or 1 offline purchase in the data period. This sampling criterion yields 4,004 customers.

We provide the key summary statistics in Table 2.1. We define visit week as a week when a customer has made at least 1 website visit. The average number of visit weeks is 2.73, with a mean 1.30 visits per visit week. We note considerable sparsity at the purchase level even after imposing the sampling criterion. We observe that an average customer has made 0.28 online and 0.18 offline purchases in the entire data

period. On average, the customers make an online visit 0.78 weeks prior to an online purchase and 2.13 weeks prior to an offline purchase. Customers make on average twice the number of online visits prior to a purchase online (1.62 ~ 2 visits) than offline (1.06 visits). We find that the average inter-visit time, or the period between two concurrent website visits, is 3.42 weeks. The average inter-purchase time, or the period between two concurrent purchases on a given channel, is 5.40 weeks for online and 6.19 weeks for offline³. These initial findings are consistent with the research hypothesis that since the duration of observed channel activities can vary, the underlying channel engagement varies. There is thus a need to explicitly measure the duration of channel engagement.

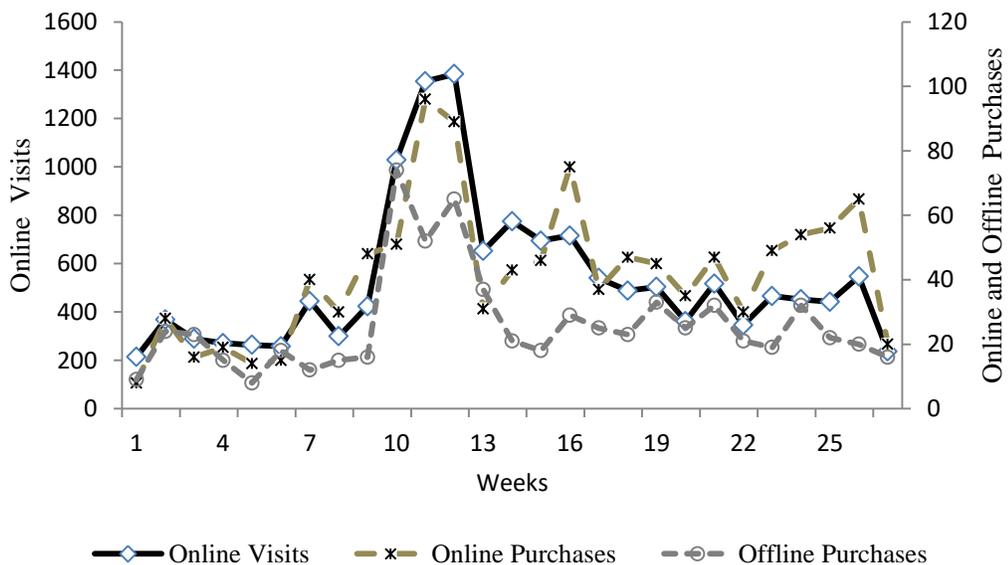
We plot the weekly aggregate online visits and online and offline purchases in Figure 2.1. We find strong evidence of a positive correlation between website visits and both online and offline purchases. Additionally, there is a seasonal spike observed in Weeks 10 through 12, which coincides with a regional festival. Based on these preliminary summary statistics, we develop the conceptual framework in section 2.4 and the model in section 2.5.

³ In appendix 1, histograms for average inter-purchase time for online and offline channels respectively provide evidence of considerable customer specific heterogeneity.

Table 2.1. Customer Level Visit and Purchase Summary

	Mean	Max.	Min.	Std. Dev.
Visit Purchase Summary				
No. of visit weeks	2.73	26.00	1.00	2.21
No. of visits per visit week	1.30	5.00	1.00	0.47
No. of online purchases	0.28	14.00	0.00	0.78
No. of offline purchases	0.18	6.00	0.00	0.42
Visit Recency (in weeks)				
Weeks since most recent website visit until online purchase	0.78	19.71	0.00	2.20
Weeks since most recent website visit until offline purchase	2.13	22.57	0.00	3.38
Visit Frequency				
Number of website visits prior to an online purchase	1.62	14.00	0.00	1.95
Number of website visits prior to an offline purchase	1.06	39.00	0.00	2.07
Inter-visit Time (in weeks)				
Inter-purchase Time by Channel (in weeks)				
Online	5.40	24.43	0.14	4.44
Offline	6.19	24.57	0.14	4.66

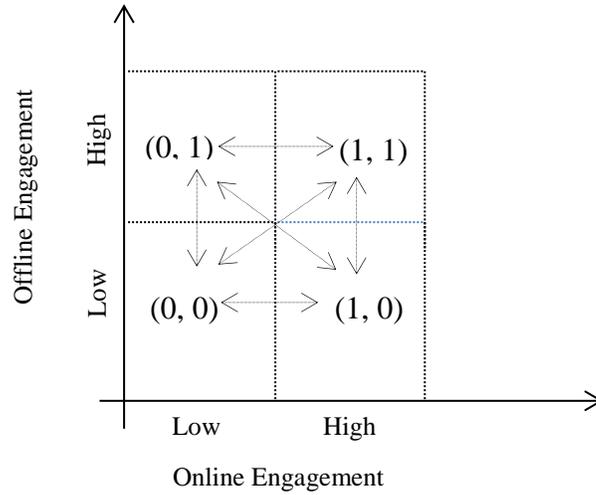
Figure 2.1. Weekly Aggregate Visit-Purchase Summary



2.4. Conceptual Framework

We assume a discrete time framework. As noted in the introduction, we propose that the customer's channel engagement follows a hidden semi-Markov (HSM) process. In an HSM process, the system stays in a state for a length of time or duration defined by a probability distribution. The process moves to a new state selected solely on its current state, i.e., in a Markovian manner. This differs from a standard HMM in that it allows an explicit specification of the duration distribution. We propose a discrete and pre-defined state space (e.g., Schweidel and Knox 2013, Schweidel et al. 2014), where a customer's latent engagement with each channel transitions between pre-defined states of "high" denoted by 1 and "low" denoted by 0 (see Figure 2.2). This assumption of two discrete states for each latent process is motivated from our empirical setting with sparse channel activities at a customer level. However, this framework can be relaxed to accommodate more states under different empirical scenarios with richer data. Mathematically, we denote s_{kt} as the engagement state at time t for channel k with $k = 1$ indicating online and $k = 2$ offline. The customer-multichannel engagement in period t is thus the state pair $s_t = (s_{1t}, s_{2t})$. The bivariate HSMM state space is given as $S = \{(1,1), (1,0), (0,1), (0,0)\}$.

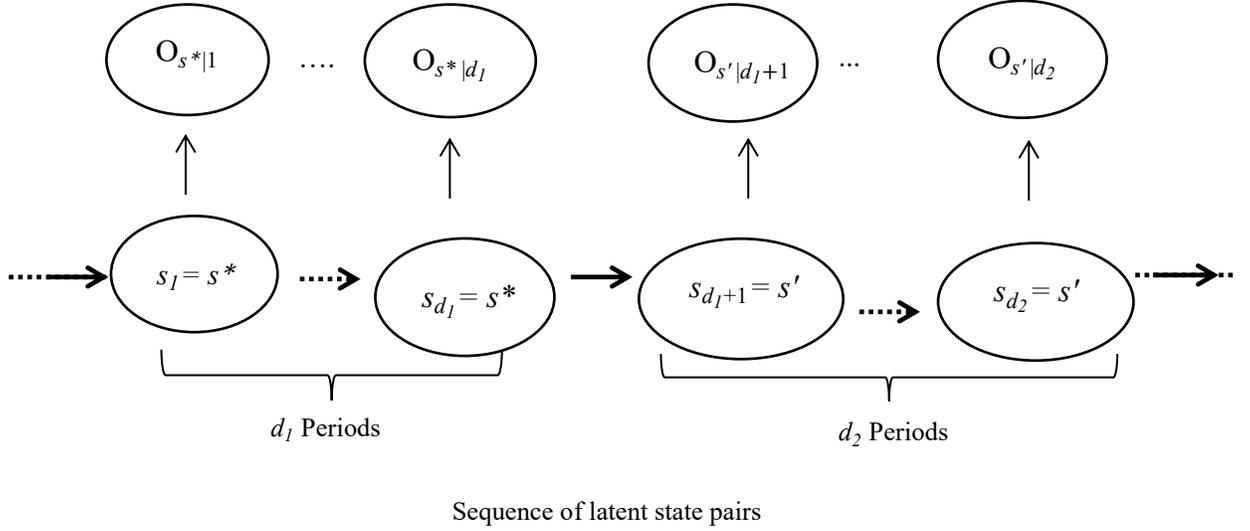
Figure 2.2. Latent State Space Description



The engagement states are indexed as High = 1, Low = 0 with online on the x-axis and offline on the y-axis. The state space is given as $\{(1, 1), (1, 0), (0, 1), (0, 0)\}$. Double arrows indicate two-way transitions between the two states

In Figure 2.3, we provide a graphical depiction of the hidden semi-Markov process where state transitions occur after some duration spent in the prior state. We observe a customer's online visits and purchases made both online and offline. We denote the observed activities conditional upon customer's multichannel engagement state pair s of duration d as $O_{s|d}$. For simplicity of notation, we ignore the time subscript here.

Figure 2.3. Hidden semi-Markov process



As shown in Figure 2.3, the customer's latent engagement state pair s^* has a duration of d_1 periods during which we observe a sequence of activities $\{O_{s^*|1}, \dots, O_{s^*|d_1}\}$. When the state pair transitions to s' , the process stays in the new state pair for a duration of d_2 periods. The sequence of observed activities in this new state pair s' is given by $\{O_{s'|d_1+1}, \dots, O_{s'|d_2}\}$. When the process enters state s^* it stays in that state for a duration described by some distribution. When the process exits state s^* , the new state s' is chosen based only on the current state. Standard Markov models imply that this duration distribution is geometric, while an HSMM allows a more flexible choice of distribution

Since each HSMM state s is a pair, for ease of notation, and as we show later in the specification of the transition probability matrix, we use the following numbering convention,

$$m(s) = \begin{cases} 1, & \text{for } s = (1,1) \\ 2, & \text{for } s = (1,0) \\ 3, & \text{for } s = (0,1) \\ 4, & \text{for } s = (0,0) \end{cases} \quad (1)$$

For the remainder of the document we will denote the HSM state as $m(s)$.

We now explain the application of sub-states and state-aggregates in the HSM framework (see L& Z 2011 for details). Conceptually, a *state-aggregate* (denoted by $I_{m(s)}$) is a set constructed by decomposing the duration spent in the corresponding HSM state $m(s)$ into a sequence of sub-states. There is a one-one mapping of an HSM state $m(s)$ and the corresponding state-aggregate ($I_{m(s)}$), numbered in the sequence given by (1). The sub-states within a state-aggregate and between state-aggregates are then numbered sequentially. Therefore, the first sub-state of the HSM state $m(s) = 2$ is sequentially numbered after the last sub-state of the HSM state $m(s) = 1$ and so on. The general specification of the state-aggregate for the HSM state $m(s)$ is thus given as,

$$I_{m(s)} = \{ n \mid \sum_{j < m(s)} l_j < n \leq \sum_{j=1}^{m(s)} l_j \} \quad (2)$$

where l_j is the maximum number of discrete HMM sub-states pre-defined for the HSM state j . From (2), combining the sequential HMM sub-states of all the state-aggregates $I_{m(s)}, m(s) \in \{1, 2, 3, 4\}$, yields the general expanded state space $\{1, 2, \dots, \sum_{m(s)=1} l_{m(s)}\}$.

The general idea of the state-aggregate construction is as follows. The HSM process sequentially transitions among HMM sub-states within a state-aggregate with a positive probability of self-transition in the final HMM sub-state of each state-

aggregate. This allows for the state pair to have a duration that exceeds the pre-defined number of HMM sub-states. The probability of self-transition in the final HMM sub-state of a state aggregate is determined by the explicit discrete distribution assumed to model the state duration. Thus, the sequential transitions among the pre-defined sub-states including the number of self-transitions into the final sub-state determines the duration spent in the state-aggregate or equivalently the HSM state. The transition to the next state-aggregate is semi-Markovian, i.e., not memoryless, since this depends on the duration spent in the current state.

In our application, we fix the maximum number of discrete sub-states of each state-aggregate a priori at five. As shown in the state-aggregate construction (2), we number the sub-states of each state-aggregate as given in Table 2.2.

Table 2.2. State Aggregates of HMM sub-states

HSM State (s)	State Pair Number	State Aggregate of HMM sub-
(1, 1)	1	$I_1 = \{1, 2, 3, 4, 5\}$
(1, 0)	2	$I_2 = \{6, 7, 8, 9, 10\}$
(0, 1)	3	$I_3 = \{11, 12, 13, 14, 15\}$
(0, 0)	4	$I_4 = \{16, 17, 18, 19, 20\}$

The state transitions within and between the state-aggregates are explained with an example as follows (see Table 2.3). Suppose at $t = 1$ the process enters the HSM state pair (1, 1), the corresponding HMM sub-state is 1. If it continues to stay in the HSM state pair (1, 1) at $t = 2$, the underlying HMM sub-state transitions to 2.

Suppose at $t = 3$ the HSM process transitions to state pair (0, 1), then the corresponding HMM sub-state transitions to sub-state 11 and so on.

Table 2.3. State Transition Illustration

HSMM State Pair (1,1)	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$
Online	1	1	1	0	0
Offline	1	1	1	1	1
Corresponding HMM Sub-state	1	2	3	11	12

The main advantage of redefining HSMM states as state-aggregates of HMM sub-states is the ability to employ standard HMM estimation methods. The transition probability matrix (TPM) can be viewed as a state-aggregate block matrix with the state duration parameters incorporated within each block (see section 2.5.4 for construction of block TPM). The likelihood for the HSMM thus reduces to the familiar HMM likelihood (MacDonald and Zucchini 1997) with the standard HMM transition matrix replaced by the state-aggregate block matrix. Post estimation one can recover both the latent state membership (see MacDonald and Zucchini 1997 on filtering approach for state membership) and estimate the state duration using this approach.

2.5. Modeling Framework

In our modeling framework, we assume T discrete time periods. As discussed in section 2.4, we propose that the customer's latent engagement with each channel is a semi-Markov process, where the time spent in each state can be explicitly modeled using any discrete distribution. The state-dependent emissions or observation

processes of the underlying semi-Markov model are the online visit counts ($V_{it|s_t}$) and purchase incidence across channels $\{Y_{kit|s_t}, k=1, 2\}$ measured at each week t for customer i . We denote online by $k=1$ and offline by $k=2$. We denote $d_{m(s)} \in \mathbb{N}$ as the duration of state $m(s) \in \{1, 2, 3, 4\}$ and p_m as its probability mass function.

2.5.1. State-dependent Emissions Model

We assume that the online visit count process ($V_{it|s_t}$) follows a Poisson distribution. In particular, when customer i is in a state of “high” online engagement, i.e., for state $m(s) \in \{1, 2\}$, the online visit count follows a non-homogenous Poisson distribution with intensity parameter $\lambda_{it|s_t}$ and when in “low” state, i.e., $m(s) \in \{3, 4\}$ zero visits are made. We thus derive partial state identification from the online visit count model. Thus, our construction of the state dependent visit count model results in an observed zero-inflated Poisson (Park et al., 2011, DeSantis and Bandyopadhyay, 2011) and is given by,

$$V_{it|s_t} = \begin{cases} \text{Poisson}(\lambda_{it|s_t}), & \text{for } s_t = m(s) \in \{1, 2\} \\ 0, & \text{for } s_t = m(s) \in \{3, 4\} \end{cases} \quad (3)$$

We parameterize the state-dependent intensity parameter $\lambda_{it|s_t}$ of the visit model as follows

$$\log(\lambda_{it|s_t}) = \alpha_{0i|s_t} + \boldsymbol{\alpha}_{1i|s_t} \mathbf{X}_{it} + \alpha_{2i|s_t} \text{Season}_t \quad (4)$$

We describe the covariates in \mathbf{X}_{it} and Season_t in Section 2.6.1.

We assume that in week t , the customer can choose to purchase online, offline or in both channels. The latent utility that the customer i derives by purchasing from

channel $k=1, 2$ in week t , conditional upon state s_t , is given by $U_{kit|s_t}$ and her corresponding purchase decision is,

$$Y_{kit|s_t} = \begin{cases} 1, & \text{if } U_{kit|s_t} \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

The latent utility is parameterized as,

$$U_{kit|s_t} = \gamma_{k0i|s_t} + \gamma_{k1|s_t} \mathbf{X}_{it} + \gamma_{k2|s_t} \text{Season}_t + \gamma_{k3|s_t} V_{it|s_t} \quad (6)$$

The state-dependent random effects intercept terms for the visit and purchase processes are respectively given by the parameters $\alpha_{0i|s_t}, \gamma_{k0i|s_t}$ ($k=1,2$). To account for the interdependence between online visits and channel purchase incidence, we introduce the state-dependent online visits made in the current week, i.e., $V_{it|s_t}$ as a covariate in the latent utility for the channel specific purchase models.

The state-dependent purchase incidence on the two channels is jointly modeled as a bivariate logit (Schweidel et al. 2014) with the parameter ϑ capturing the dependence between the processes,

$$\Pr(Y_{1it}, Y_{2it}|s_t) = \frac{\exp(Y_{1it|s_t} U_{1it|s_t} + Y_{2it|s_t} U_{2it|s_t} + Y_{1it|s_t} Y_{2it|s_t} \vartheta)}{1 + \exp(U_{1it|s_t}) + \exp(U_{2it|s_t}) + \exp(Y_{1it|s_t} U_{1it|s_t} + Y_{2it|s_t} U_{2it|s_t} + Y_{1it|s_t} Y_{2it|s_t} \vartheta)} \quad (7)$$

For $\vartheta > 0$, there is a positive correlation in purchase decisions on both channels, while for $\vartheta < 0$, the customer uses the channels as substitutes for purchase decisions.

The joint probability distribution of the state-dependent observed activities is then,

$$\Pr(V_{it} = v, \{Y_{1it}, Y_{2it}\} | s_t) = \Pr(V_{it} = v | s_t) * \Pr(\{Y_{1it}, Y_{2it}\} | s_t, v) \quad (8)$$

2.5.2. State Duration Distribution

We recall that in our framework, we measure the state duration $d_{m(s)}$ in discrete time. We evaluate the state duration model using two alternative discrete distributions – Poisson and geometric. In standard HMMs, the Markovian memoryless property implicitly assumes that the state duration follows a geometric process. The geometric model thus serves as the benchmark in our study. We define $d_{m(s)}$ as the duration spent in state $m(s)$.

We specify the geometric state duration model as follows,

$$g(d_{m(s)}) = (1 - v_{m(s)})^{d_{m(s)}} \cdot v_{m(s)}, \quad d_{m(s)} \geq 0 \quad (9)$$

where, and $v_{m(s)}$ is the geometric parameter associated with the state pair number $m(s)$.

We test the benchmark state duration geometric process against the Poisson distribution. We specify the Poisson state duration model as,

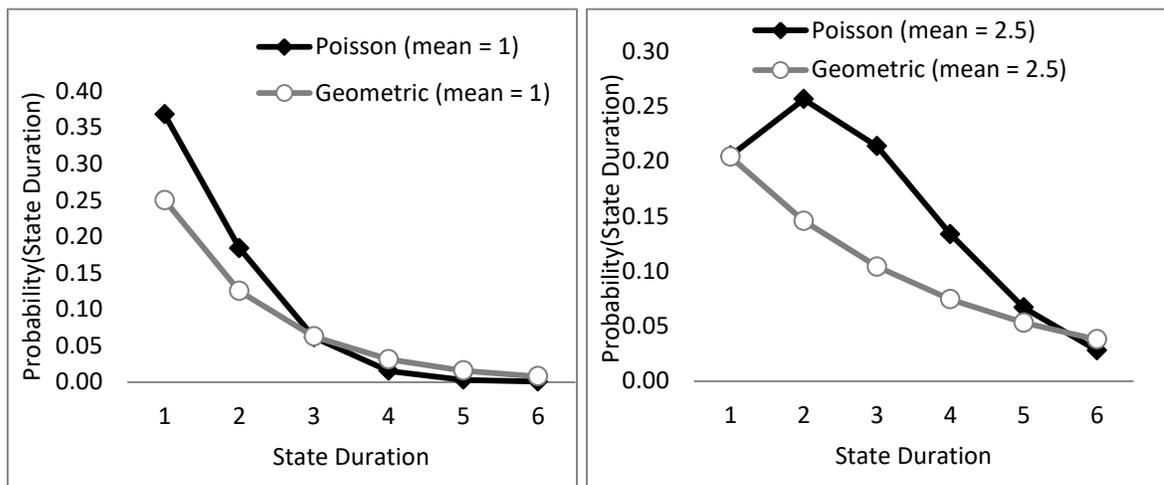
$$\rho(d_{m(s)}) = \frac{\mu_{m(s)}^{d_{m(s)}} \exp(-\mu_{m(s)})}{d_{m(s)}!}, \quad d_{m(s)} \geq 0 \quad (10)$$

We denote the Poisson parameter associated with the state pair number $m(s)$ by $\mu_{m(s)}$.

We choose the Poisson as an alternative for the following reasons: In the context of multichannel retailing, a customer's channel engagement often involves time consuming decision making of information search, deliberation and purchase. For instance, a customer's channel choice for information search may vary by product attributes (Lal and Sarvary 1999). Thus, a customer's channel engagement duration during the search and buy decision process may extend over an interval of discrete time points. For such situations, non-geometric distributions such as Poisson with

memoryless waiting times may better accommodate the channel engagement duration pattern. To illustrate the difference in distribution shape, we compare the CDFs of Poisson and geometric decay across two different scenarios in Figure 2.4. In each scenario we assume the mean duration of the process to be identical across both Poisson and geometric. As is evident from Figure 2.4, for mean duration = 1, the Poisson distribution shows higher probability associated with duration 1 and 2 than the geometric. For mean duration = 2.5, the Poisson process is associated with a modal duration of 2. Thus, the Poisson distribution has a less sharp decay than the geometric and will thus be able to account for longer duration (> 1 period) for both “high” and “low” engagement states.

Figure 2.4. Poisson versus Geometric CDF



For both the state duration distributions given in (9) and (10), we specify the failure or hazard rates $h(d_{m(s)})$ at which the $m(s)^{th}$ latent state duration ends at $d_{m(s)}+1$ as,

$$h(d_{m(s)}) = \begin{cases} \frac{p(d_{m(s)})}{1 - F(d_{m(s)}-1)}, & \text{for } F(d_{m(s)}-1) < 1 \\ 1, & \text{for } F(d_{m(s)}-1) = 1 \end{cases} \quad (11)$$

where, $p(d_{m(s)})$ and $F(d_{m(s)})$ are respectively the p.m.f and CDF of the underlying duration distribution.

We specify a no-covariates state duration model and assume homogeneous parameters of the underlying distributions. This is partly to alleviate the estimation complexity and identifiability issues of individual state duration parameters given the data sparsity noted in section 3. However this does not limit our ability to examine customer-specific behavior in determining state transitions. In section 5.3, we incorporate customer-level heterogeneity in the specification of the state transition model.

2.5.3. HSMM Transition Probability Matrix (TPM)

Customer i 's latent propensity to transition from state $s_t = (s_{1t}, s_{2t})$ in week t to $s_{t+1} = (s_{1,t+1}, s_{2,t+1})$ in week $t + 1$ can be decomposed into a bivariate channel transition with the k^{th} channel specific transition propensity denoted by the term $TR_{kit, s_{kt} \rightarrow s_{k,t+1}}$. We incorporate cross-sectional heterogeneity and non-stationarity by introducing time-varying covariates in the transition propensity model as,

$$TR_{kit, s_{kt} \rightarrow s_{k,t+1}} = \tau_{k0i, s_{kt} \rightarrow s_{k,t+1}} + \tau_{k1, s_{kt} \rightarrow s_{k,t+1}} \mathbf{Z}_{it} \quad (12)$$

Here $\tau_{k0i,s_{kt} \rightarrow s_{k,t+1}}$ denotes the customer specific random effects contributing to state transition on the k^{th} channel. The vector of coefficients associated with the time-varying covariates \mathbf{Z}_{it} for the transition model of channel k is given by $\tau_{k1,s_{kt} \rightarrow s_{k,t+1}}$. (See Section 2.6.1 for covariates description).

Based on the channel specific state transitions, we have the following four cases.

Case 1: State changes for offline but not online, i.e., $(s_{1,t+1} = s_{1t}, s_{2,t+1} \neq s_{2t})$

$$a_{it,t+1} = \frac{\exp(\text{TR}_{2it,s_{2t} \rightarrow s_{2,t+1}})}{1 + \exp(\text{TR}_{1it,s_{1t} \rightarrow s_{1,t+1}}) + \exp(\text{TR}_{2it,s_{2t} \rightarrow s_{2,t+1}})} \quad (12a)$$

Case 2: State changes for online but not offline, i.e., $(s_{1,t+1} \neq s_{1t}, s_{2,t+1} = s_{2t})$

$$a_{it,t+1} = \frac{\exp(\text{TR}_{1it,s_{1t} \rightarrow s_{1,t+1}})}{1 + \exp(\text{TR}_{1it,s_{1t} \rightarrow s_{1,t+1}}) + \exp(\text{TR}_{2it,s_{2t} \rightarrow s_{2,t+1}})} \quad (12b)$$

Case 3: State changes for both channels simultaneously, i.e., $s_{k,t+1} \neq s_{kt}, k=1,2,$

$$a_{it,t+1} = 1 - \sum_k \frac{\exp(\text{TR}_{kit,s_{kt} \rightarrow s_{k,t+1}})}{1 + \exp(\text{TR}_{1it,s_{1t} \rightarrow s_{1,t+1}}) + \exp(\text{TR}_{2it,s_{2t} \rightarrow s_{2,t+1}})} \quad (12c)$$

Case 4: Same State Transition, i.e., $(s_{1,t+1} = s_{1t}, s_{2,t+1} = s_{2t})$

Since the HSMM explicitly models duration within a state, the same state transition probability is 0 by definition. Thus, we have, $a_{it,t+1} = 0$

2.5.4. Construction of the state-aggregate block TPM

To construct the state-aggregate block TPM, we recall from the HSMM state pair numbering convention in equation (1) that the state-pair numbers are denoted by $m(s)$. Further, for every HSMM state $m(s)$, we define a state-aggregate of HMM sub-

states $I_{m(s)}$ and the expanded state space of all the HMM sub-states is given as $\{1, 2, \dots, \sum_{m(s)=1} d_{m(s)}\}$. Let $B_{it} = \{b_{it,j \rightarrow j'}\}$ be the non-stationary and individual specific TPM of the respecified state-aggregates. Here $b_{it,j \rightarrow j'} = \Pr_i(\tilde{s}_{t+1} = j' | \tilde{s}_t = j)$, is customer i 's transition probability of moving from the HMM sub-state $\tilde{s}_t = j$ in week t to sub-state $\tilde{s}_{t+1} = j'$ in week $(t+1)$, and $j, j' \in \{1, 2, \dots, \sum_{m(s)=1} d_{m(s)}\}$. By construction this matrix B_{it} can be specified as a block matrix (see L&Z 2011 for discussion) as follows,

$$B_{it} = \begin{bmatrix} B_{it,1 \rightarrow 1} & \cdots & B_{it,1 \rightarrow 4} \\ \vdots & \ddots & \vdots \\ B_{it,4 \rightarrow 1} & \cdots & B_{it,4 \rightarrow 4} \end{bmatrix} \quad (13)$$

where each diagonal block $B_{it,m(s) \rightarrow m(s)}$ is the matrix of transition probabilities within the HMM sub-states of the aggregate $I_{m(s)}$ and is of dimension $d_{m(s)} * d_{m(s)}$, $m(s) = \{1, 2, 3, 4\}$. The off-diagonal matrices $B_{it,m(s) \rightarrow m(s')}$ are the transition matrices from state-aggregate $I_{m(s)}$ to $I_{m(s')}$ where, $m(s) \neq m(s')$. These matrices each have dimension $d_{m(s)} * d_{m(s')}$. The dimension of the overall block matrix B_{it} is thus $\sum_{m(s)=1} d_{m(s)} * \sum_{m(s)=1} d_{m(s)}$, where $\sum_{m(s)=1} d_{m(s)}$ is the total sum of state durations across all 4 HSMM state-pairs. We define the block matrix of transition within the state-aggregate as,

$$B_{it,m(s) \rightarrow m(s)} = \begin{bmatrix} 0 & 1-h(1) & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1-h(d_{m(s)}-1) \\ 0 & 0 & 0 & 0 & 1-h(d_{m(s)}) \end{bmatrix} \quad (14)$$

The transition probability of being in the same state $1-h(d)$ is interpreted as the probability of survival in state $m(s)$ at the $(d+1)^{st}$ week after having survived until the d^{th} week, $d = 1, 2, \dots, d_{m(s)}$. The transition probabilities of this matrix thus reflect the transitions through the elements of the state-aggregate $I_{m(s)}$, with self-transition only in the final sub-state of the aggregate

For the off-diagonal matrices, we define

$$\mathbf{B}_{it, m(s) \rightarrow m(s')} = \begin{bmatrix} \mathbf{a}_{it, m(s) \rightarrow m(s')} h(1) & \cdots & 0 \\ \mathbf{a}_{it, m(s) \rightarrow m(s')} h(2) & 0 & 0 \\ \vdots & \ddots & \vdots \\ \mathbf{a}_{it, m(s) \rightarrow m(s')} h(d_{m(s)}) & \cdots & 0 \end{bmatrix} \quad (15)$$

where each row of the off-diagonal matrix $\mathbf{B}_{it, m(s) \rightarrow m(s')}$ is populated only for the first element and $\mathbf{a}_{it, m(s) \rightarrow m(s')} h(d)$. The off-diagonal matrix captures the between HSMM state transitions in the first column. The first element in each row thus represents the transition from state $m(s)$ to first sub-state in $m(s') \neq m(s)$ conditional upon the survival in state $m(s)$ until the d^{th} time period.

2.5.5. Initial Distribution for State-Aggregates

We define the initial distribution vector of the state-aggregates for customer i as Π_i . Following the construction of HMM sub-states from the state duration periods, we note that the dimension of this vector is $\sum_{m(s)=1} d_{m(s)}$. Following the literature on constructing the initial state distribution vector (e.g. Netzer et al. 2008), we compute the individual specific TPM at the mean of covariates $\bar{\mathbf{B}}_i$ and solve $\Pi_i = \Pi_i \bar{\mathbf{B}}_i$.

2.5.6. State Identification and Parameter Parsimony

Given the sparsity of observed activities on each channel we make a general assumption that a channel activity is observed only if the corresponding channel engagement state is high. Conversely, if the channel engagement state is low then no activity on that channel is observed. We recall from section 2.5.1 that the online visit count model assumes a Poisson process when the customer is in a state of high online engagement, i.e., states 1 and 2. This assumption provides partial identification to the online engagement state process. We make similar assumptions for the purchase models for the online and offline engagement states. More specifically, we assume that when the latent process is in states 3 and 4 (both states represent low online engagement), online purchase model parameters are set to zero. Analogously, for states 2 and 4 (states representing low offline engagement), the offline purchase model parameters are zero. As noted in the introductory section, since we do not observe offline non-purchase footfall we assume that a customer is likely to make offline visits during periods of high offline engagement. Though the identification of “high” offline engagement (states 1 and 3) is derived from the observed offline purchases, the explicit estimation of duration spent in the respective states provides an interval within which a firm is likely to observe the customer make an offline visit.

In the bivariate HSMM construction, we note that we have a large number of state dependent parameters in the emissions model to account for each state pair. Thus to make the model more parsimonious we assume homogeneous univariate state dependent parameters. That is, if the customer is in high online engagement (either state 1 or 2), the state-dependent parameters for the emissions models are assumed the

same for both states 1 and 2. This assumption is justified as long as the customer is in a state of high engagement with the online channel (either state 1 or 2) her tendencies to browse the website and purchase from the online channel would be the same under both states. In an analogous manner, for the offline purchase model, we assume homogeneity of parameters for the states 1 and 3.

2.5.7. Computational Approach

Our computational approach uses Bayesian estimation with Markov Chain Monte Carlo (MCMC) simulation methods. To ensure better convergence and mixing, we employ a Metropolis-Hastings algorithm for the proposal updates with all parameters updated in block-move steps. Denoting $\Psi_{im(s)}$, as the set of random effects and $\Theta_{m(s)}$ as the aggregate parameters associated with the state $m(s) = \{1, 2, 3, 4\}$, the customer specific state-dependent emission probability matrix in week t is written as,

$$P_{it}(V, \{Y_1, Y_2\}) = \text{diag}(\underbrace{\text{Pr}_{it}(V, \{Y_1, Y_2\}|\Psi_{i1}, \Theta_1), \dots, \text{Pr}_{it}(V, \{Y_1, Y_2\}|\Psi_{i1}, \Theta_2), \dots}_{d_1 \text{ times}}, \dots, \underbrace{\text{Pr}_{it}(V, \{Y_1, Y_1\}|\Psi_{i4}, \Theta_4), \dots, \text{Pr}_{it}(V, \{Y_1, Y_1\}|\Psi_{i4}, \Theta_4)}_{d_4 \text{ times}}) \quad (16)$$

Denoting $\Psi_i = (\Psi_{i1}, \Psi_{i2}, \Psi_{i3}, \Psi_{i4})$ and $\Theta = (\Theta_1, \Theta_2, \Theta_3, \Theta_4)$ the cross-sectional likelihood yields the familiar MacDonald and Zucchini (1997) HMM likelihood,

$$L_i(\Psi_i, \Theta) = \Pi_i P_{i1}(V, \{Y_1, Y_2\}) B_{i1} \dots \dots \dots P_{iT}(V, \{Y_1, Y_2\}) B_{iT} \mathbf{1}' \quad (17)$$

Denoting Ψ as vector of random effects across all customers, the posterior distribution for Bayesian inference for the full HSMM:

$$P(\Theta, \Psi | V, \{Y_1, Y_2\}) \propto \prod_{i=1}^n (L_i(\Psi_i, \Theta) \times \Phi(\Psi_i)) \times \varphi(\Theta) \quad (18)$$

The likelihood function across all customers is given by $\prod_{i=1}^n L_i(\Psi_i, \Theta)$ and the prior distribution for random effects and aggregate parameters respectively denoted by $\Phi(\Psi_i)$ and $\varphi(\Theta)$.

The random effects coefficients on both observed and latent models are assumed to follow a non-informative normal distribution. As noted in section 2.3, there is considerable sparsity in the online visit counts and online-offline purchase incidences. Following Gelman et al. 2008, we use a weakly informative scaled t-prior for the random effects means and fixed effects coefficients of both the state-dependent emissions and transitions processes. The logarithm of variances of the random effects are assumed to follow inverse gamma prior with suitable values for the shape and scale hyper-parameters. For the state duration Poisson and geometric distributions, we first take appropriate transformations of the parameters and assume the transformed Poisson parameters to have conjugate gamma priors and those of the geometric to follow scaled-t priors. The details of the prior specifications and the Bayesian computational steps are elaborated in the Appendix to this chapter.

2.6. Empirical Application

2.6.1. Covariates Specification

For the state dependent emissions models (equations (4) and (6)), the covariates in \mathbf{X}_{it} describe the customer's own time-varying experiences with the channels in terms of observed activities, i.e., online visits (V_{it}) and purchase

activities (Y_{kit} , $k = 1, 2$). To capture the effect of past browsing or online visit experiences on current visit and purchase behavior, we define the *Online Visit Stock* in week t as the cumulative online visit counts until the week $(t-1)$, i.e., $(\sum_{w=1}^{(t-1)} V_{iw})$. To incorporate the effect of past purchase channel experience, we define two stock variables based on the channel of purchase - *Online Purchase Stock* and *Offline Purchase Stock*. These are the cumulative purchase incidence until the last week on online and offline channels respectively, i.e., $\sum_{w=1}^{(t-1)} Y_{kiw}$; $k = 1, 2$. Based on the firm's regional festivals, the seasonality indicator $Season_t$ takes the value 1 for weeks 10, 11 and 12 and 0 otherwise.

For the covariates in the state transition model (equation (12)), we use the number of weeks since last activity for *Recency*. In the CRM literature (Fader et al. 2005, Schweidel and Knox 2013 etc.), recency reflects the inactivity period of the customer since the last observed activity. In absence of data on marketing interventions we use recency to determine the state transitions. The rationale is that a customer's state change from high to low or vice versa may depend on how far back the last activity took place. Further, recency is also likely to be correlated to a customer's attrition tendency from a channel or the firm completely and therefore may affect state transitions from high to low. In our model evaluation as discussed in the next section we test the linear and quadratic specifications of recency. We also note that for computational purposes we take logarithmic transformation of both. Therefore, the covariates in \mathbf{Z}_{it} are respectively *Ln Recency* and *Sqr. Ln Recency*.

2.6.2. Model Evaluation

Before estimating the models on the actual data, we have run a set of simulation studies to test out the modeling framework. The detailed results are provided in the Appendix to Chapter 2. In our simulation studies, we compare between two distributions for the state duration model – Poisson and geometric and impose aggregate level parameters on both. Typical model selection criteria do not distinguish the states reliably, but using the correct (Poisson) model does improve state reconstruction.

From the 4,004 customers tracked over the 27-week period July 2015-December 2015, we choose a random sample of 998 customers for model calibration and a sample of 746 customers as holdout to measure the predictive performance.

To examine the distributional properties of the state duration we consider a series of nested models – Poisson duration (Models 1, 2 and 3) and geometric duration (Models 4, 5 and 6). In Models 1 and 4, the state transition does not have any covariates except the random effects (R.E) intercept. Models 2 and 5 incorporate *Ln Recency* along with a random effects intercept. Models 3 and 6 incorporate both the linear and quadratic specifications of log transformation of recency, i.e., *Ln Recency* and *Sqr. Ln Recency*. We note that the state dependent emissions processes have identical specification for all the six models evaluated. We summarize the key difference in the modeling components of the six models in Table 2.4.

Table 2.4. Specifications of Evaluated Models

Covariates		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
State	Ln Recency		√	√		√	√
Transition	Sqr. In Recency			√			√
State	Poisson	√	√	√			
Duration	Geometric				√	√	√

Abbreviations used: Ln = Natural logarithm, sqr. = squared

For model estimation, we run MCMC chains between 100,000-200,000 iterations⁴. We used the last 20,000 iterations to calculate the model results. To reduce the autocorrelation in the MCMC chains we perform thinning at every 100th step, yielding 200 retained samples to estimate the posterior distribution.

2.6.3. Model Performance

For model performance comparison on the calibration data we compute the logarithm of marginal density (LMD), AIC, WAIC (Watanabe 2010, Gelman et al. 2014) and BIC based on the posterior distribution from the 200 retained samples. We report these in Table 2.5.

⁴ Convergence rates of the evaluated models varied. In particular, the Poisson model 3 had the slowest convergence rate.

Table 2.5. Model Performance on Calibration Data

Model	Parameters	LMD	AIC	WAIC	BIC
1: Poisson, No covariates trans.	33	-11,698	23,462	23,389	23,542
2: Poisson, Ln recency trans.	37	-11,606	23,286	23,208	23,376
3: Poisson, Ln+ sqr. ln recency trans.	41	-11,618	23,319	23,226	23,418
4: Geom., No covariates trans.	33	-11,537	23,141	23,070	23,221
5: Geom., Ln recency trans.	37	-11,577	23,229	23,144	23,319
6: Geom., Ln+ sqr. ln recency trans.	41	-11,491	23,065	23,016	23,164

Abbreviations used: trans. = transition, Geom. = geometric

To measure predictive ability of the models on out of sample data, we perform a 1-step (i.e., 1 week) ahead forecast of the online visit count, and online and offline purchase incidences on the holdout sample. We measure predictive accuracy in two ways – hit rates at individual level (Table 2.6) and mean absolute error (or MAE) at aggregate weekly level (Table 2.7). The hit rate measures the proportion of cases where the predicted matches the truth⁵. To compute the predicted values for a given model, we use the ROC curve analysis to determine the threshold probabilities (see Appendix 2.4). The threshold probability corresponds to the decile that records the maximum separation between true and false positive rates. We report the hit rates for both customer level activity and inactivity. The MAE is measured as the average error in prediction from the weekly aggregated observed behavior.

⁵ For online visits, we back out the visit activity or incidence from the estimated visit intensity parameter.

Table 2.6. Predictive Ability at Individual Level on Holdout Sample

Model	Customer Activity Hit Rates			Customer Inactivity Hit Rates		
	On Visits	On Purch.	Off Purch.	No On Visits	No On Purch.	No Off Purch.
1: Poisson, No covariates trans.	53.0%	50.5%	63.9%	78.1%	75.0%	73.5%
2: Poisson, Ln recency trans.	79.0%	57.7%	70.7%	53.2%	75.3%	75.3%
3: Poisson, Ln+ sqr. ln recency trans.	55.1%	55.2%	61.9%	78.4%	75.3%	75.2%
4: Geom., No covariates trans.	52.8%	50.5%	74.1%	78.1%	75.2%	74.6%
5: Geom., Ln recency trans.	54.3%	55.2%	78.2%	78.3%	75.3%	75.4%
6: Geom., Ln+ sqr. ln recency trans.	51.6%	53.6%	72.8%	78.0%	75.3%	75.3%

Abbreviations used: On = Online; Off = Offline; Purch. = Purchases

We observe that though the geometric duration models have better in-sample fit, the predictive ability of these models on the holdout sample vary considerably both at individual and aggregate levels. However as noted in the literature on RFM models tend to suffer from the drawback of superior in-sample fit but weaker out-of-sample prediction (see Schweidel and Knox 2013). Since the covariate specifications in all the models are based on recency and frequency measures, our results are consistent with the literature.

On the holdout prediction, the Poisson duration with linear recency transition specification (model 2) has higher hit rates than the geometric models on both online visit (79.0%) and purchase (57.7%) activities. However, at the individual level the

geometric models show superior out of sample prediction for offline purchase activity. In particular, geometric duration with linear recency transition specification (model 5) has the highest offline purchase hit rate (78.2%). Part of this could be attributed to the individual model's sensitivity to the threshold probability of visit and purchase activities (see Appendix 2.4). In terms of customer inactivity, the models perform comparably with misclassification rates in the range of 20 - 30%. However model 2 has a higher misclassification rate (46%) on online visit inactivity. This implies that model 2 tends to over-predict online visitation behavior.

Table 2.7. Predictive Ability at Aggregate Level on Holdout Sample

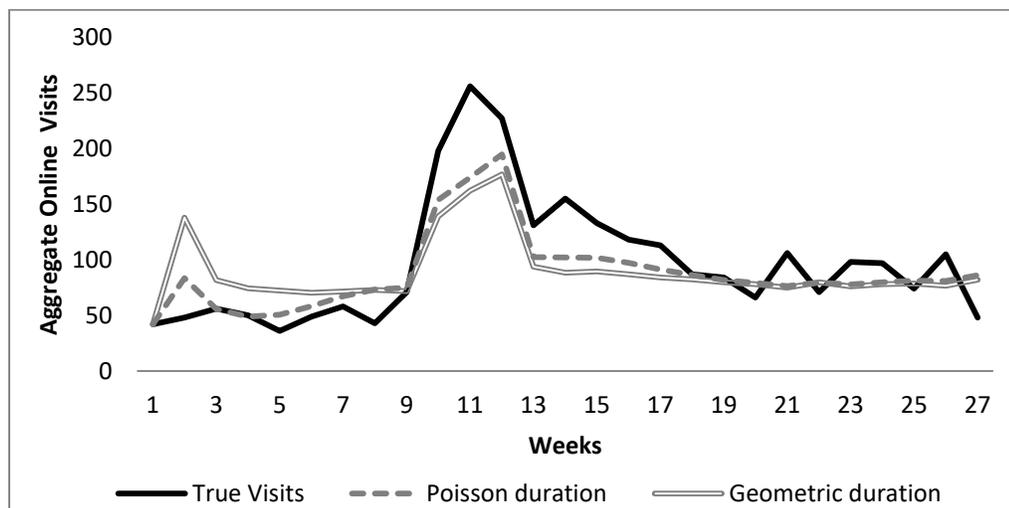
Model	Aggregate Level MAE		
	Online Visits	Online Purchase	Offline Purchase
1: Poisson, No covariates trans.	29.91	4.73	3.62
2: Poisson, Ln recency trans.	22.30	3.91	3.64
3: Poisson, Ln+ sqr. ln recency trans.	27.05	3.87	3.37
4: Geom., No covariates trans.	29.89	4.50	3.34
5: Geom., Ln recency trans.	31.56	4.55	5.00
6: Geom., Ln+ sqr. ln recency trans.	34.96	5.57	5.31

At an aggregate level, we find that the Poisson duration models show better predictive performance than geometric models. In particular, model 2 has the lowest MAE across aggregate weekly online visits (22.3) and online purchases (3.91). For aggregate weekly offline purchases, the Poisson models fare comparably to the geometric duration model 4; however significantly outperform models 5 and 6. The geometric duration models thus have higher inconsistency in predictive performance

across individual and aggregate levels. We note that model performance variation across individual and aggregate levels under sparse observation data situation has been documented in the literature (e.g., Schweidel and Knox 2013). We find that the Poisson model 2 gives the best and most consistent out-of-sample prediction at both individual and aggregate levels. In the subsequent discussions, we use Poisson model 2 as the proposed model for out-of-sample prediction and parameter estimation discussion.

To show the out-of-sample predictive ability on a week by week basis, we plot the weekly aggregate online visits and compare prediction of Poisson model 2 against the best performing geometric model 5 (lowest MAE for online visits among geometric models, Table 2.7) in Figure 2.5. We compare Poisson model 2’s performance against the best performing geometric model 4 on the weekly aggregate online and offline purchase predictions in Figures 2.6 and 2.7 respectively.

Figure 2.5. Weekly Aggregate Online Visits of Holdout Sample



We find that the Poisson model 2 captures the spikes in the seasonal trends much better than the geometric for both the online visits and online/offline purchase models. In particular, the best performing geometric duration model under-predicts changes in customer behavioral trends during seasonal weeks and over-predicts the low activity weeks. This is especially seen for online visits and online purchases in weeks 1 through 5 of low activity, where the geometric duration model significantly over-predicts both visits and purchases at an aggregate level.

Figure 2.6. Weekly Aggregate Online Purchases of Holdout Sample

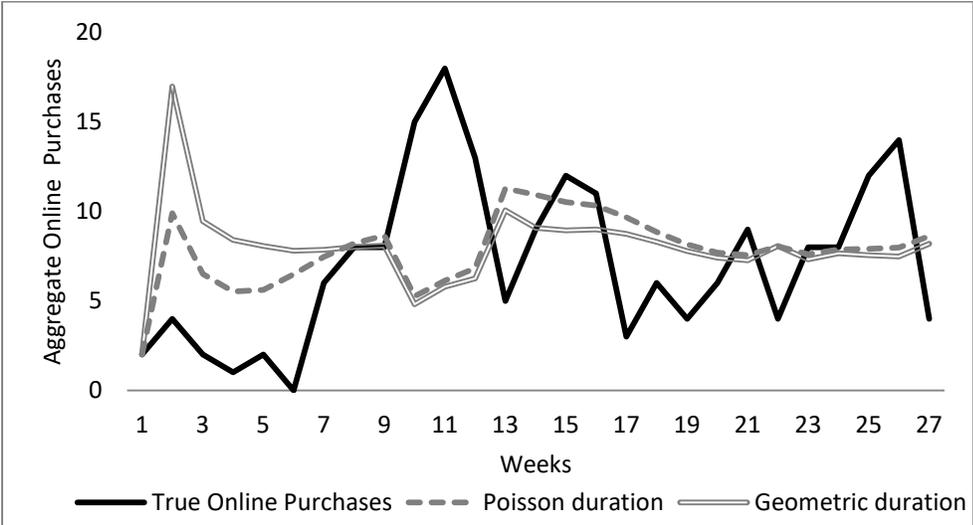
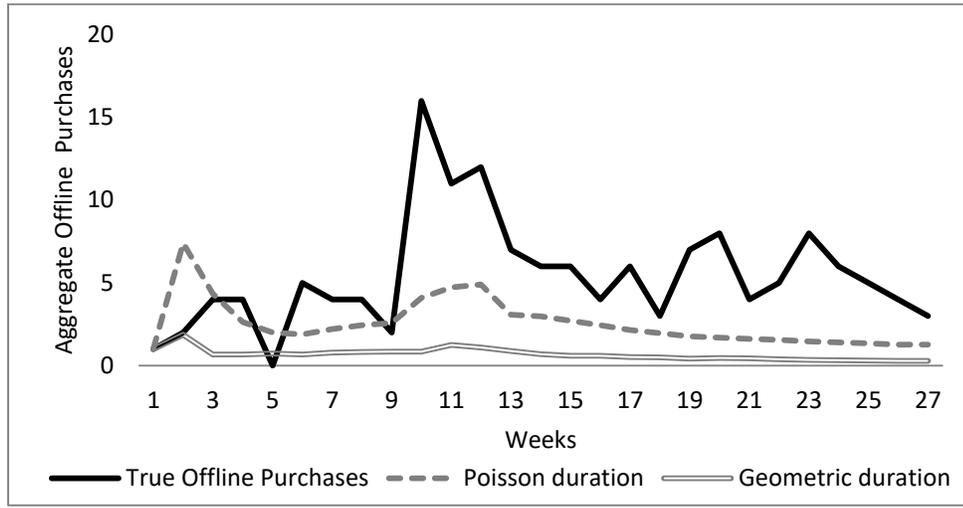


Figure 2.7. Weekly Aggregate Offline Purchases of Holdout Sample



On weekly offline purchases, we find that though the Poisson model 2 under-predicts the offline purchases it performs better than the best performing geometric model 4 (lowest MAE among geometric models, Table 2.7), consistently across all 27 weeks. These results corroborate the earlier inference that while the geometric model assumes decay in a customer’s engagement state duration beyond the observed activity, the Poisson model allows a longer duration for unobserved activities.

2.6.4. Parameter Estimates

We report the parameter estimates (posterior mean and SD) of the state-dependent emissions for the best performing Poisson model 2 in Table 2.8⁶.

⁶ The parameter estimates for the remaining models are not noted here for brevity and space constraints. These results can be provided upon request.

Table 2.8. Parameter Estimates for the Emissions Models

Emissions Process	Parameters	Channel Engagement States		
		1	1 & 2	1 & 3
Online Visit Count	RE Intercept (Mean)		-0.828 (0.002)	
	Online Visit Stock		0.026 (0.004)	
	Online Purchase Stock		0.079 (0.023)	
	Offline Purchase Stock		-0.768 (0.098)	
	Seasonality		0.668 (0.037)	
	Online Purchase Incidence	RE Intercept (Mean)		-2.998 (0.003)
Online Purchase Incidence	Online Visit Stock		-0.141 (0.019)	
	Online Purchase Stock		0.803 (0.086)	
	Offline Purchase Stock		-0.463 (0.188)	
	Online Visits in week t		0.876 (0.048)	
	Seasonality		-0.674 (0.15)	
	Offline Purchase Incidence	RE Intercept (Mean)		
Offline Purchase Incidence	Online Visit Stock			-0.673 (0.08)
	Online Purchase Stock			0.07 (0.348)
	Offline Purchase Stock			0.663 (0.296)
	Online Visits in week t			1.177 (0.088)
	Seasonality			0.302 (0.184)
	Dependence Structure of Bivariate Logit Purchase Model		-0.935 (0.333)	

Not surprisingly, we find that the posterior means of the R.E intercepts for the emissions processes are negative. This is likely with the sparse data on the observed activities. The *Online Visit Stock* has a weakly positive relationship with online visits and a negative relationship with online and offline purchase propensities. This indicates that there are many customers who browse the website but do not make purchases in the future. The *Online Purchase Stock* has a significant positive effect on online purchase propensity and weakly positive effect on online visit and offline purchase. We note that since online visits that end up in a purchase get accounted as an online purchase, the weak positive effect on online visits is attributed to the customers who visit but do not buy. Overall, this means that past online purchase experience has a significant positive effect on future online activities.

We also find that the *Offline Purchase Stock* has a significant negative effect on online visits and purchase, but a strong positive effect on offline purchase. Thus the customers who have bought from the offline store are more likely to purchase from the offline channel. However, online visit counts in the concurrent period have a strong positive effect on both online and offline purchase propensity. This implies that the customers prefer to browse online before making a purchase on either channel. The significant negative effect of the dependence term of the bivariate purchase model shows the channels are not used concurrently for purchase. Therefore, in a given purchase week the customers are likely to choose either online or offline to buy but not both. Further, seasonality plays a significant positive impact on online visits and offline purchases but a negative effect on online purchases. In the absence of marketing interventions data, we assume that this could be due to more in-store offers

during seasonal periods. That is, customers gather information on in-store offers via web search and consequently shop offline.

The parameter estimates (posterior mean and SD) of the bivariate state transition model of Poisson model 2 are provided in Table 2.9. Since Poisson model 2 outperforms Poisson model 3 with quadratic recency specification, it implies that higher orders of recency does not necessarily explain state transition better. We find that *Ln Recency* has a significant positive effect on the state transition from high to low for either channel. Further, recency also has a significantly negative effect on state transition from low to high on either channel. We note that by definition larger the value of recency further back in the past has the activity occurred. This implies that once a customer enters a period of inactivity or low engagement she is less likely to come back to either channel. From the firm’s point of view this should be seen as a concern since once a customer disengages from either channel, it is more difficult to bring her back for future activities.

Table 2.9. Parameter Estimates for the Bivariate State Transition Model

State Transitions	Parameters	Engagement State Channel	
		Online	Offline
High to Low	RE Intercept (Mean)	0.256 (0.006)	-1.428 (0.004)
	Ln Recency	1.362 (0.23)	2.962 (0.159)
Low to High	RE Intercept (Mean)	4.998 (0.002)	-0.812 (0.023)
	Ln Recency	-1.987 (0.063)	-3.33 (1.611)

2.6.5. Estimated State Duration

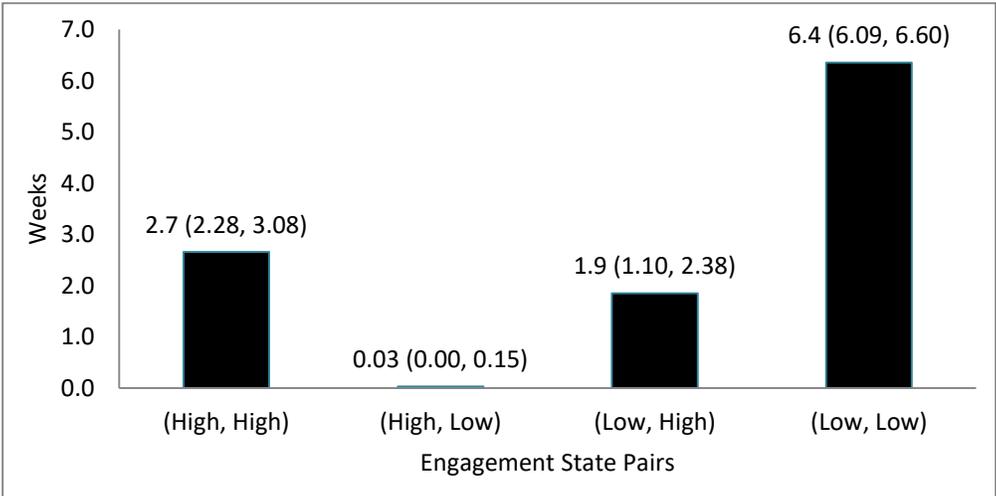
We note that the state duration parameters for the Poisson model are essentially the Poisson mean intensity or mean duration for each state. In Figure 2.8 we show a bar plot and report the posterior means of the mean duration in each state for the best performing Poisson model 2. We report the 95% credible interval in parentheses.

We note that the Poisson model estimates a nearly 3 week duration for high engagement with both channels. This is consistent with the observation that on average the most recent online website visit until an offline purchase is about 2.3 weeks (see Table 2.1). When a customer is engaged on both channels, she is likely to deliberate over a longer duration (more than a week) before she makes her purchase.

The mean engagement duration for online-only is 0.03 week ~ less than 1 day. When a customer is engaged with only the online channel she is either likely to make the purchase in the same browsing session or she visits online but does not come back again to buy. This could be specific to this high end cosmetic brand with limited number of customers returning for repeat purchase. Therefore, customers who are prompted to the online channel through firm's promotions may not be likely to return. The average duration of customers with offline-only engagement is close to 2 weeks. This indicates that on average customers tend to go back to the store after making purchases in the prior week. We note that since the offline purchase data is sparse this result tends to show the channel engagement behavior of the subset of customers who have more than 2 offline purchases in the data. We note that the average duration of low engagement on either channel is more than 6 weeks. Given that the data on

observed activities is sparse this result is again likely to be affected by the subset of customers who have made repeat purchases. Since this is a high-end cosmetic brand, this implies the firm can expect repeat customers to return after about 7 weeks of inactivity.

Figure 2.8. State Duration Estimates



The engagement state pairs are sequenced to denote (Online engagement, Offline engagement).

2.7. Managerial Implications

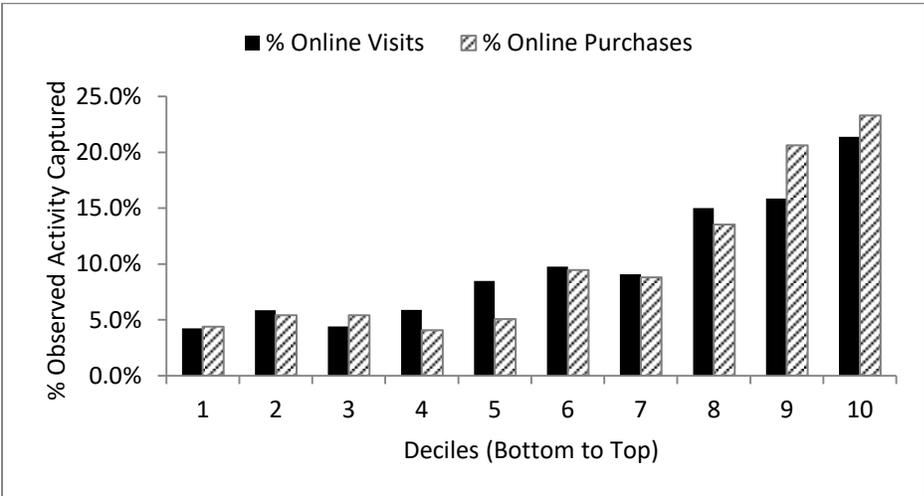
2.7.1. Forecasting Channel Specific Activities and High Engagement Customers

While the model helps in making weekly forecasts on channel specific activities of visits and purchases, the proposed framework can also be used by managers to identify the customers with high channel engagement in the forecasted period. We outline the procedure as follows. From the model we can derive 1-step or a week ahead forecasted probabilities of purchase at online and offline, and the expected online visit count at the customer level. Using the filtering approach (e.g. MacDonald

and Zucchini 1997), we can then back out the forecasted probabilities for the latent engagement states.

To determine the cut-off or threshold probability that defines a high online or offline engagement, one can use the forecasted engagement state probabilities computed on the calibration data as follows. From the calibration sample, we add the forecasted probabilities for states 1 and 2 to derive “high” online engagement probabilities. We perform a decile ranking of the derived “high” online engagement probabilities and compute the proportion of online activities captured by each decile. In Figure 2.9 we plot the proportion of true online visits and purchases captured by each decile in the calibration sample.

Figure 2.9. Online Activities by Decile Ranking

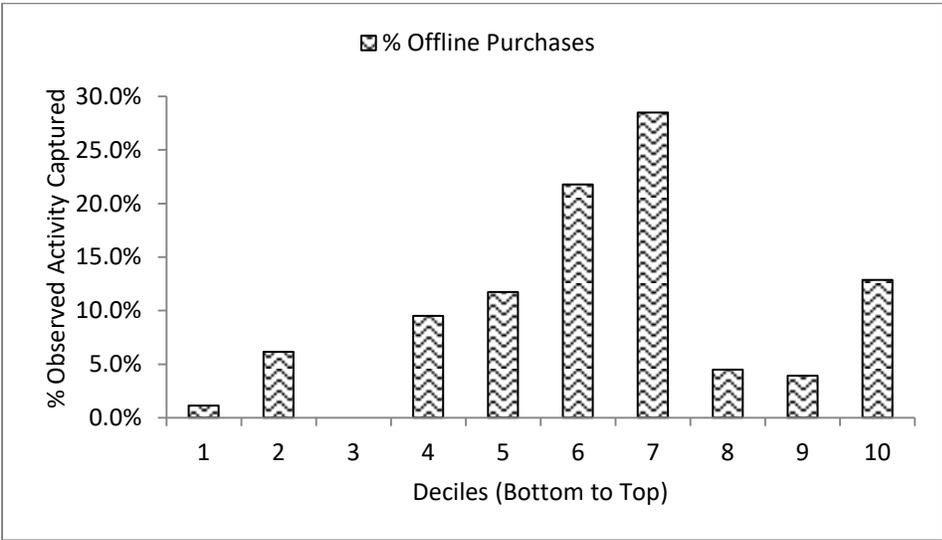


We find that the proportion of observed online activities increase as we move up the deciles. Specifically, there is a sharp improvement in the percentage of observed online activities above the 80th percentile. We can choose the probability

corresponding to the 80th percentile as the cut-off to identify high online engagement customers in the forecasted period.

A similar approach can be taken to identify “high” offline engagement customers (i.e., add probabilities for states 1 and 3 and follow similar decile ranking as above). The proportion of true offline purchases captured by the decile ranking for high offline engagement in the calibration sample is shown in Figure 2.10. Though there is a break in the ranking for offline purchase at the 80th percentile, we observe a sharp improvement in percentage observed activities between 60th and 70th percentiles. The probability cut-off at the 60th percentile can be used to identify customers in the high offline engagement in the forecasted period.

Figure 2.10. Offline Purchases by Decile Ranking



For a marketing manager this will identify customers who are predicted to be in a high engagement state on each or both channels. In particular, the prediction of

high offline engagement along with the duration estimation can enable marketers determine the time interval to expect a customer's store visit. This can help the manager determine the relevant marketing interventions for such customers including integrated channel strategies such as Buy Online, Purchase In store or BOPS (e.g. Gallino and Moreno 2014, Gallino et al. 2016), or channel specific strategies of online or in-store offers.

2.7.2. Identifying Customers at Attrition Risk

The proposed model's ability to estimate the average duration of a customer's channel engagement states can also be leveraged to gauge customer's attrition risk. The explicit estimation of state duration helps marketers to determine the mean periods of channel specific activities and inactivity for repeat customers. Therefore if a customer is forecasted to be in the low engagement state (state 4) for longer than the estimated average duration she may be deemed to be at attrition risk. From the procedure noted in the prior section 2.7.1, if the forecasted state does not fall in one of the three high engagement states, i.e., states 1, 2 and 3, the customer is likely to be in low engaged on both channels, i.e., in state 4. Depending on the profitability of customers at attrition risk marketers can take appropriate actions.

2.8. Contributions, Future Research Directions and Conclusion

This essay develops a multivariate HSMM framework that simultaneously models the customers' engagement with the online and offline channels of a firm. Our proposed modeling framework extends the marketing literature on multivariate HMMs

in two ways. First, we simultaneously model two latent processes using a bivariate state space. Second, the semi-Markov specification relaxes the distributional assumptions typically imposed on state duration in standard HMMs. We examine the distributional properties of the state duration by comparing between Poisson and geometric duration distributions under various specifications of the transition model. Our results show that the proposed Poisson model gives superior prediction over the geometric both at individual and aggregate levels. These predictions will enable marketers to target customers for appropriate channel interactions and design targeted channel promotions for higher purchase conversions. By explicitly estimating the state duration, our framework can help marketers determine the expected lengths of channel specific activities and inactivity of the customers and identify potential attrition risk customers.

Our research can be extended in a number of directions. First, we use RFM based covariates due to lack of data on marketing interventions. Incorporating marketing interventions especially in the transition models will provide more substantive insights into the marketing effects on channel engagement duration and transition. Second, we have assumed identical duration distribution on both channels. For instance, we assume Poisson to examine both online and offline engagement. However, with the Poisson model over-predicting offline purchases, there is reason to believe that the engagement duration distributions may vary by channels. Intuitively this implies a customer's engagement behavior with channels vary due to channel specific characteristics. Third, to enable state identification under sparse data conditions we have restricted the analysis to two states. Further, we have assumed a

no-covariate specification of the duration distributions with homogenous parameters. In a more data rich setting, the framework can be extended to examine > 2 states and covariates in the state duration model. Fourth, examining duration can provide important insights into channel attribution effects; a customer who is engaged longer with a certain channel is likely to gather more information and hence more likely to be influenced to buy. The literature on multichannel attribution has examined the channel touchpoints sequence to determine purchase conversion attribution (e.g. Li and Kannan 2014). Incorporating the channel touchpoint duration along with the sequence can lead to more robust measurement of channel attribution. The proposed framework can also be extended to other marketing contexts. For instance, our framework can be used to examine eye-tracking behavior and associated duration of visual attention in determining brand choices (e.g. Wedel and Pieters 2000, Wedel et al. 2008) and dynamic effects of coupon and promotion expiration dates on customer's redemption behavior (e.g. Inman and McAlister 1994, Krishna and Zhang 2000). We hope that this methodology is exploited and developed further to examine more such interesting marketing phenomena.

REFERENCES

- Ansari A, Mela CF, Neslin SA (2008) Customer channel migration. *J. Marketing Res.* 45(1): pp.60-76.
- Bell DR, Gallino S, Moreno A (2017) Offline showrooms in omnichannel retail: Demand and operational benefits. *Management Sci.* 64(4): 1629-1651.
- Brodie RJ, Hollebeek LD, Jurić B, Ilić A (2011) Customer engagement: conceptual domain, fundamental propositions, and implications for research. *J. Serv. Res.* 14(3): 252-271.
- Business (2017) 3 essential elements for winning at multi-channel customer engagement. (February 22), <https://www.business.com/articles/3-essential-elements-for-winning-at-multi-channel-customer-engagement/>
- Chang CW, Zhang JZ (2016) The effects of channel experiences and direct marketing on customer retention in multichannel settings. *J. Interact. Marketing.* 36: 77-90.
- Chang CW, Zhang JZ, Neslin SC (2017) The dynamic impact of buying 'fit products' on customer learning and profitability in multichannel settings. *Tuck School of Business Working Paper No. 2759679*
- Cox DR, Miller HD (1965) *The Theory of Stochastic Processes*. (Chapman & Hall, London).
- Danaher PJ, Wilson IW, Davis RA (2003) A comparison of online and offline consumer brand loyalty. *Marketing Sci.* 22(4): pp.461-476.
- Danaher PJ, Mullarkey GW, Essegaiier S (2006) Factors affecting web site visit duration: A cross-domain analysis. *J. Marketing Res.* 43(2): 182-194.
- DeSantis SM, Bandyopadhyay D (2011) Hidden Markov models for zero-inflated Poisson counts with an application to substance use. *Stat. Med.* 30(14): 1678-1694.
- Dinner IM, Heerde Van HJ, Neslin SA (2014) Driving online and offline sales: The cross-channel effects of traditional, online display, and paid search advertising. *J. Marketing Res.* 51(5): pp.527-545.
- Du R, Kamakura WA (2006) Household lifecycles and life styles in America. *J. Marketing Res.* 43(1): 121-132.
- Fader PS, Hardie BG, Lee KL (2005) RFM and CLV: Using iso-value curves for

- customer base analysis. *J. Marketing Res.* 42(4): 415-430.
- Gallino S, Moreno A (2014) Integration of online and offline channels in retail: The impact of sharing reliable inventory availability information. *Management Sci.* 60(6): 1434-1451.
- Gallino S, Moreno A, Stamatopoulos I (2016) Channel integration, sales dispersion, and inventory management. *Management Sci.* 63(9): 2813-2831.
- Gartner (2018) Transforming from multichannel to unified retail commerce (January 18), <https://blogs.gartner.com/robert-hetu/transforming-multichannel-unified-retail-commerce-2/>
- Gelman A, Hwang J, Vehtari A (2014) Understanding predictive information criteria for Bayesian models. *Stat. Comput.* 24: 997
- Gelman A, Jakulin A, Pittau MG, Su YS (2008) A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* 1360-1383.
- Gensler S, Dekimpe MG, Skiera B (2007) Evaluating channel performance in multi-channel environments. *J. Retailing and Consumer Svcs.* 14(1): 17-23.
- Helsen K, Schmittlein DC (1993) Analyzing duration times in marketing: Evidence for the effectiveness of hazard rate models. *Marketing Sci.* 12(4): 395-414.
- Inman JJ, McAlister L (1994) Do coupon expiration dates affect consumer behavior?. *J. Marketing Res.* 423-428.
- Inman J, Shankar V, Ferraro R (2004) The roles of channel-category associations and geodemographics in channel patronage. *J. Marketing.* 68(2): pp.51-71.
- Janiszewski C (1998) The influence of display characteristics on visual exploratory search behavior. *J. Consumer Res.* 25(3): 290-301.
- Johnson EJ, Moe WW, Fader P S, Bellman S, Lohse GL (2004) On the depth and dynamics of online search behavior. *Management Sci.* 50(3): 299-308.
- Johnson MT (2005) Capacity and complexity of HMM duration modeling techniques. *IEEE Signal Processing Letters.* 12(5): 407-410.
- Krishna A, Zhang Z J (1999) Short-or long-duration coupons: The effect of the expiration date on the profitability of coupon promotions. *Management Sci.* 45(8):1041-1056.
- Kumar V (2013) *Profitable Customer Engagement: Concept, Metrics and Strategies.* (SAGE Publications India).

- Kumar V, Pansari A (2016) Competitive advantage through engagement. *J. Marketing Res.* 53.4: 497-514.
- Kushwaha T, Shankar V (2013) Are multichannel customers really more valuable? The moderating role of product category characteristics. *J. Marketing.* 77(4): pp.67-85.
- Lal R, Sarvary M (1999) When and how is the Internet likely to decrease price competition?. *Marketing Sci.* 18(4): 485-503.
- Langrock R and Zucchini W (2011) Hidden Markov models with arbitrary state dwell-time distributions. *Comput. Stat. Data Anal.* 55(1): 715-724.
- Lewis RA, Reiley DH (2014) Online ads and offline sales: measuring the effect of retail advertising via a controlled experiment on Yahoo!. *Quant. Marketing Econom.* 12(3): pp.235-266.
- Li H, Kannan PK (2014) Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. *J. Marketing Res.* 51(1): 40-56.
- Liechty J, Pieters R, Wedel M (2003) Global and local covert visual attention: Evidence from a Bayesian hidden Markov model. *Psychometrika.* 68(4): 519-541.
- MacDonald IL, Zucchini W (1997) *Hidden Markov and Other Models for Discrete-valued Time Series* (Chapman & Hall, London).
- Merkle (2011) Multichannel customer engagement. Web report, https://www.merkleinc.com/sites/default/files/salescollateral/Multi_Channel_Engagement_0.pdf
- Moe W (2003) Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *J. Consumer Psych.* 13(1-2): 29-39.
- Moe W, Fader PS (2004) Dynamic conversion behavior at e-commerce sites. *Management Sci.* 50(3): 326-335.
- Montaguti E, Neslin SA, Valentini S (2015) Can marketing campaigns induce multichannel buying and more profitable customers? A field experiment. *Marketing Sci.* 35(2): pp.201-217.
- Montoya R, Netzer O, Jedidi K (2010) Dynamic allocation of pharmaceutical detailing

- and sampling for long-term profitability. *Marketing Sci.* 29(5): 909-924.
- Montgomery AL, Li S, Srinivasan K, Liechty, JC (2004) Modeling online browsing and path analysis using clickstream data. *Marketing Sci.* 23(4): 579-595.
- Moon S, Kamakura, WA, Ledolter J (2007) Estimating promotion response when competitive promotions are unobservable. *J. Mkt. Res.*, 44(3), 503-515.
- Netzer O, Lattin JM, Srinivasan V (2008) A hidden Markov model of customer relationship dynamics. *Marketing Sci.* 27(2): 185-204.
- Newton MA, Raftery AE (1994) Approximate Bayesian inference with the weighted likelihood bootstrap. *J R Stat. Soc. Series B.* 3-48.
- Park YH, Park CH, Ghosh P (2011) Modelling member behaviour in on-line user-generated content sites: a semiparametric Bayesian approach. *J R Stat. Soc. Series A.* 174(4): 1051-1069.
- Patterson P, Ting Y, Ruyter K (2006) Understanding Customer Engagement in Services. *Adv. Theory, Maintain. Relevance, Proceedings of ANZMAC 2006 Conference, Brisbane, 4-6 December.*
- Putsis Jr WP, Srinivasan N (1994) Buying or just browsing? The duration of purchase deliberation. *J. Marketing Res.* 393-402.
- Russell M, Cook A (1987, April) Experimental evaluation of duration modelling techniques for automatic speech recognition. *IEEE Int. Conf. on ICASSP'87*, Vol. 12: pp. 2376-2379.
- Sansom J, Thomson P (2007) On rainfall seasonality using a hidden semi-Markov model. *J. Geophys. Res. Atmos.* 112(D15).
- Schweidel DA, Bradlow ET, Fader PS (2011) Portfolio dynamics for customers of a multiservice provider. *Management Sci.* 57(3): 471-486.
- Schweidel DA, Knox G (2013) Incorporating direct marketing activity into latent attrition models. *Marketing Sci.* 32(3): 471-487.
- Schweidel DA, Park YH, Jamal Z (2014) A multiactivity latent attrition model for customer base analysis. *Marketing Sci.* 33(2): 273-286.
- Thomas JS, Sullivan UY (2005) Managing marketing communications with multichannel customers. *J. Marketing*, 69(4): 239-251.
- Statista (2017) Report, <https://www.statista.com/statistics/412099/multi-channel->

[retail-purchase-path-usa/](#)

Vivek SD, Beatty SE, Morgan RM (2012) Customer engagement: Exploring customer relationships beyond purchase. *J. Marketing Th. and Pr.* 20(2): 122-146.

Vocalcom (2018) Top 7 customer engagement trends in 2018. (January 16), <https://www.vocalcom.com/en/blog/digital-customer-engagement/top-7-customer-engagement-trends-in-2018/>

Watanabe S (2013) A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.* 14(Mar): 867-897.

Wedel M, Pieters R (2000) Eye fixations on advertisements and memory for brands: A model and findings. *Marketing Sci.* 19(4): 297-312.

Wedel M, Pieters R, Liechty J (2008) Attention switching during scene perception: how goals influence the time course of eye movements across advertisements. *J. Exp. Psychol. Appl.* 14(2): 129.

Xie Y, Yu SZ (2009) A large-scale hidden semi-Markov model for anomaly detection on user browsing behaviors. *IEEE/ACM Transactions on Networking (TON)*, 17(1): 54-65.

Yu SZ, Kobayashi H (2003) A hidden semi-Markov model with missing data and multiple observation sequences for mobility tracking. *Signal Process.* 83(2): 235-250.

Yu SZ (2010) Hidden semi-Markov models. *Artif. Intell.* 174(2):215-243.

Zhang JZ, Netzer O, Ansari A (2014) Dynamic targeted pricing in B2B relationships. *Marketing Sci.* 33(3): 317-337.

Zhang J, Wedel M (2009) The effectiveness of customized promotions in online and offline stores. *J. Marketing Res.* 46(2): pp.190-206.

Appendix 2.1. Prior Specifications

$$\begin{aligned}
 \alpha_{1||s_t \in \{1,2\}} &\stackrel{\text{iid}}{\sim} \text{scaled } t(7); \text{ scale} = 2.5 \\
 \gamma_{1k|s_t}, \gamma_{2k|s_t} &\stackrel{\text{iid}}{\sim} \text{scaled } t(7); \text{ scale} = 2.5 \\
 \tau_{1k, s_{kt} \rightarrow s_{k,t+1}} &\stackrel{\text{iid}}{\sim} \text{scaled } t(7); \text{ scale} = 2.5 \\
 \mu_{m(s)} &\sim \text{Gamma}(2;2) ; m(s) \in \{1, 2, 3, 4\} \\
 \alpha_{0i|s_t \in \{1,2\}} &\sim N(\mu_{\alpha_0}, \sigma_{\alpha_0}); \\
 \mu_{\alpha_0} &\sim \text{scaled } t(7); \sigma_{\alpha_0} \sim \text{IG}(3;1) \\
 \gamma_{0i,k|s_t} &\sim N(\mu_{\gamma_{0k|s_t}}, \sigma_{\gamma_{0k|s_t}}); \\
 \mu_{\gamma_{0k|s_t}} &\sim \text{scaled } t(7); \sigma_{\gamma_{0k|s_t}} \sim \text{IG}(3;1); k=1,2 \\
 \tau_{0i,k, s_{kt} \rightarrow s_{k,t+1}} &\sim N(\mu_{\tau_{0k, s_{kt} \rightarrow s_{k,t+1}}}, \sigma_{\tau_{0k, s_{kt} \rightarrow s_{k,t+1}}}); \\
 \mu_{\tau_{0k, s_{kt} \rightarrow s_{k,t+1}}} &\sim \text{scaled } t(7); \sigma_{\tau_{0k, s_{kt} \rightarrow s_{k,t+1}}} \sim \text{IG}(3;1); k = 1, 2
 \end{aligned} \tag{A.2.1}$$

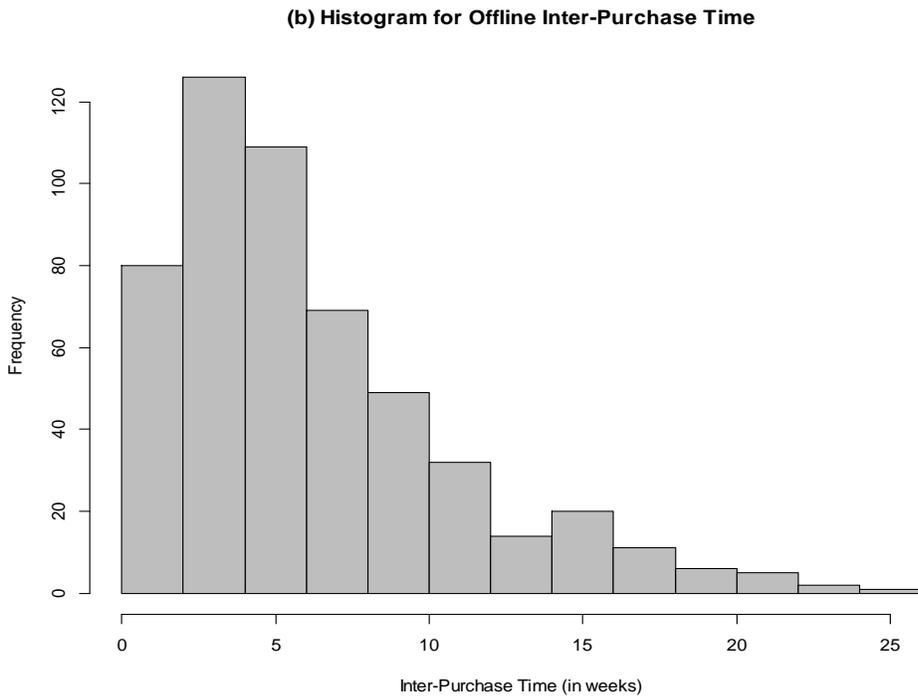
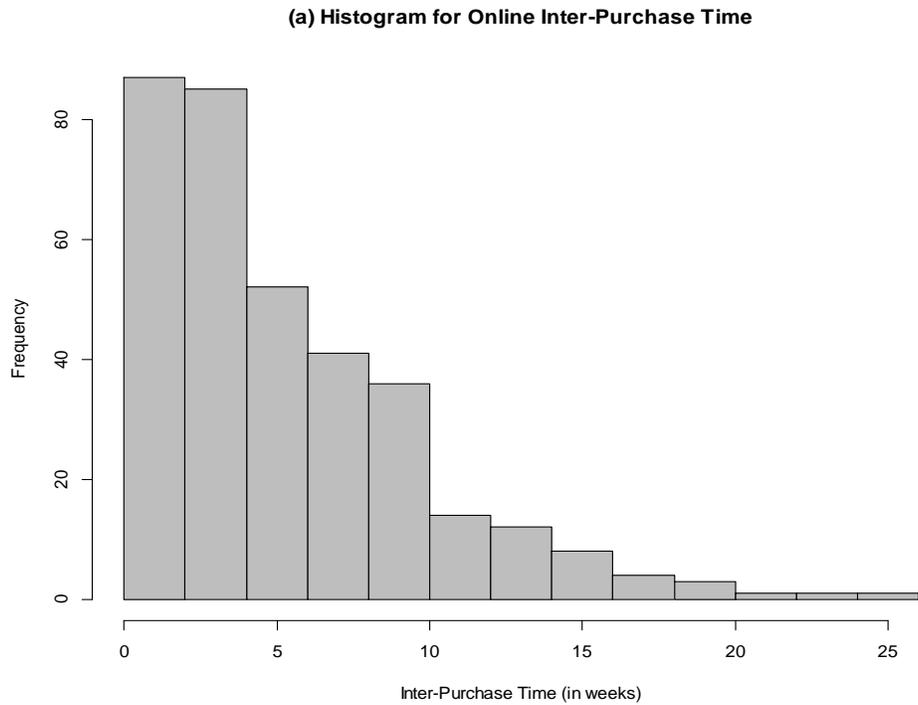
Our Bayesian computational approach follows a Metropolis-Hastings algorithm where at each update step the proposal value follows a Gaussian random walk model.

Appendix 2.2. Customer Heterogeneity in Purchase Behavior

From the data on the 4,004 customers tracked over a period of 27-weeks, we select the customers who have made at least 2 purchases on a given channel to generate histograms for the average inter-purchase time in weeks across online and offline channels respectively. It appears that most customers make two consecutive weekly online purchases while for the offline channel the modal inter-purchase time is approximately 3 weeks. The histograms provide strong evidence of customer-specific

heterogeneity in the channel usage behavior for purchases.

Figure A.2.1 Histograms of Average Inter-purchase Time in weeks



Appendix 2.3. Simulation Study

We perform a simulation study to test the hypothesis that the state duration distribution of multichannel engagement may not necessarily be geometric. We simulate a dataset of 500 customers for a panel of 27 weeks. The data generating process (DGP) assumes the specifications for the state dependent emissions processes as given in the modeling framework equations (2)-(5). We assume a no covariates random effects (R.E) intercept only model for the state transition specification and a Poisson state duration distribution. The coefficients under the DGP assume values to replicate the summary statistics of a random sample of 727 customers from the actual data (see Table A2.1).

Table A2.1 Actual versus Simulated Data

	Actual	Simulated
Sample Size	727	500
Panel Length	27 weeks	27 weeks
Total Online Visit Count	2819	1642
Total Online Purchase Count	235	229
Total Offline Purchase Count	63	48
Online Visit Rate	14.4%	12.2%
Online Purchase Rate	1.2%	1.7%
Offline Purchase Rate	0.3%	0.4%

For evaluation of state duration distribution, we compare the state predictability between two competing state duration specifications. First we estimate the model assuming the DGP specification with Poisson duration distribution. The second model evaluated has the same specifications for the state dependent and transition processes as in the DGP but with a geometric state duration distribution. For estimation, we run an MCMC chain of 20,000 iterations, where the first 10,000 used as burn-in. From the

post burn-in chain of 10,000 iterations, we thin at every 50th step to get a thinned chain of 200 iterations. We use the posterior means from the thinned chains and recover state membership using the filtering algorithm (see MacDonald and Zucchini 1997). Tables A2.2 and A2.3 summarize the confusion matrix of state prediction, where the diagonal terms describe the proportion of states that have been predicted accurately. We recall the state pair numbering convention from equation (1) in the paper. Therefore, the states 1, 2, 3 and 4 refer sequentially to the state pairs $S = \{(1, 1), (1, 0), (0, 1), (0, 0)\}$.

Table A2.2. Model under DGP Poisson duration

True State	Predicted State			
	1	2	3	4
1	20.85%	35.33%	34.18%	9.63%
2	13.64%	53.42%	24.93%	8.01%
3	13.76%	21.86%	51.18%	13.19%
4	19.88%	18.51%	53.83%	7.78%

Table A2.3. Model under Geometric duration

True State	Predicted State			
	1	2	3	4
1	20.47%	13.52%	56.89%	9.12%
2	15.91%	32.09%	44.77%	7.23%
3	1.86%	2.44%	84.52%	11.18%
4	0.73%	0.89%	94.84%	3.53%

For both models the states 1 and 4 are recovered poorly ($< 50\%$), while state 3 is recovered with $> 80\%$ accuracy by the geometric. We find that the Poisson model beats the geometric in prediction of state 2 (53.42% vs. 32.09%), and has a better prediction for state 4. In particular, as is seen from the off-diagonal terms, the

geometric model tends to consistently over-predict state 3 of Low Online, High Offline. This can be explained by the sparse offline purchase data (purchase rate 0.4%). Since there are not too many offline purchases, the geometric model with a sharper decay over-predicts the high offline engagement state.

In Table A2.4., we report the estimated mean duration in weeks of the Poisson and geometric distributions against the actual values assumed in the DGP. While we find that the pattern of over and under-prediction of state duration is consistent across states for both the Poisson and geometric models, we find that the geometric estimates are off by a larger magnitude. The geometric duration estimates leads to an over-prediction of mean duration times for the engagement states 1 (high online, high offline) and 3 (high online, low offline). Further, there is a significant under-prediction of duration for the states 2 (High Online, Low Offline) and 4 (Low Online, Low Offline). The simulation results point to the need of examining non-geometric duration distributions.

Table A2.4. Mean Duration (in weeks)

State	Actual	Poisson	Geometric
1	0.76	0.46	1.73
2	3.58	2.96	1.00
3	0.68	3.43	5.74
4	6.84	1.01	0.01

Appendix 2.4. Prediction Validity Measure: Hit Rates

In Table 2.5 of Chapter 2, we report the hit rates for customer level activities and inactivity. We briefly discuss the calculation of hit rates in this section. The hit rate is

defined as follows,

$$\text{Hit Rate} = \frac{\text{Number of Cases where observed equals predicted}}{\text{Total number of observed cases}} \quad (\text{A 2.2})$$

Therefore, the hit rate is measured for prediction of an incidence (activity or inactivity). We recall that the observed activities are the online visits, and online and offline purchase incidence at each time period. Since the online visit is modeled as a Poisson count process, we use the estimated intensity parameter to compute the visit incidence probability for the ROC of the visit model. That is, we estimate the visit incidence probability for customer i as,

$$\Pr(V_{it} > 0) = 1 - \Pr(V_{it} = 0) = 1 - e^{-\lambda_{it}} \quad (\text{A 2.3})$$

Based on the probabilities of these incidences, i.e., visit incidence, online and offline purchase incidences, we determine the probability threshold above which the model predicts the incidence to have occurred. To determine the probability threshold, we plot the ROC curves for each of these incidences. The threshold probability for each model corresponds to the maximum distance between the ROC against the null model, i.e., the maximum separation between the true and false positive rates. We plot the ROCs for online visit incidence in Figure A.2.2. We show the ROC plots for online and offline purchases in Figures A.2.3 and A.2.4 respectively.

Figure A.2.2. ROC for Online Visit Incidence

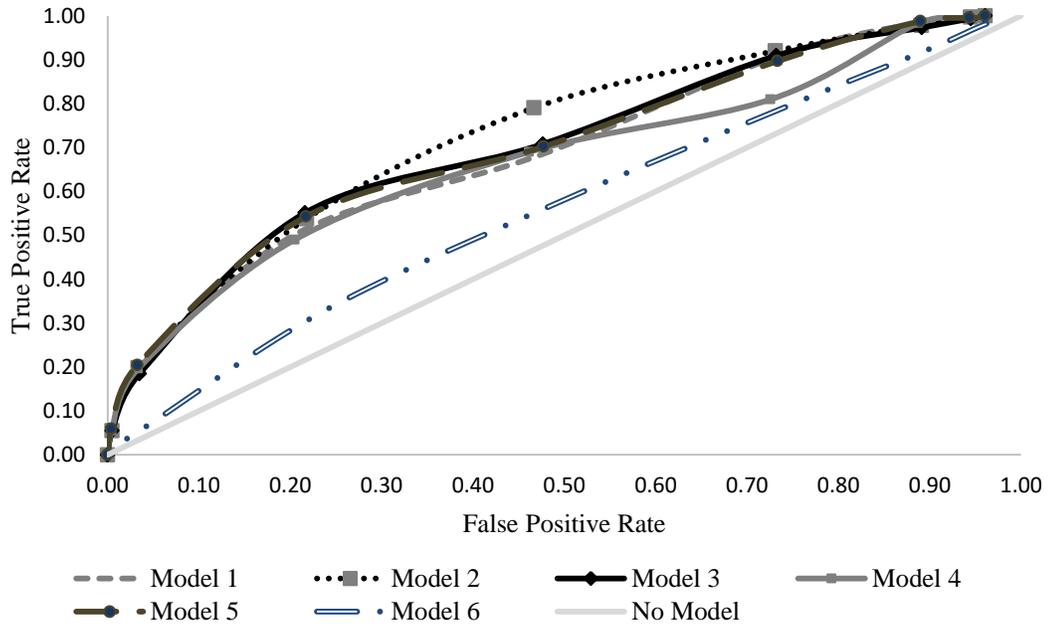


Figure A.2.3. ROC for Online Purchases

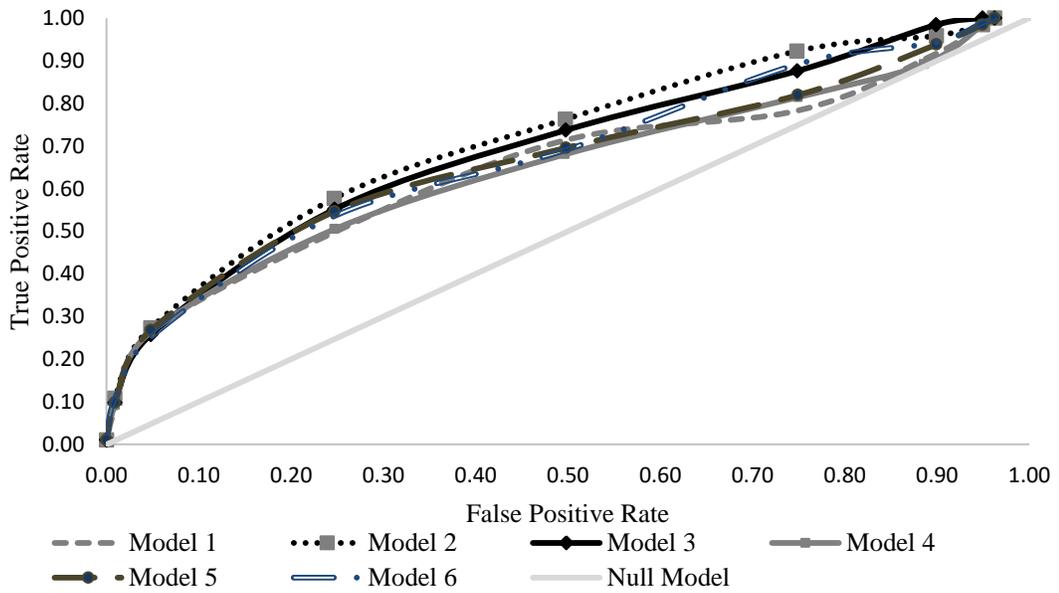
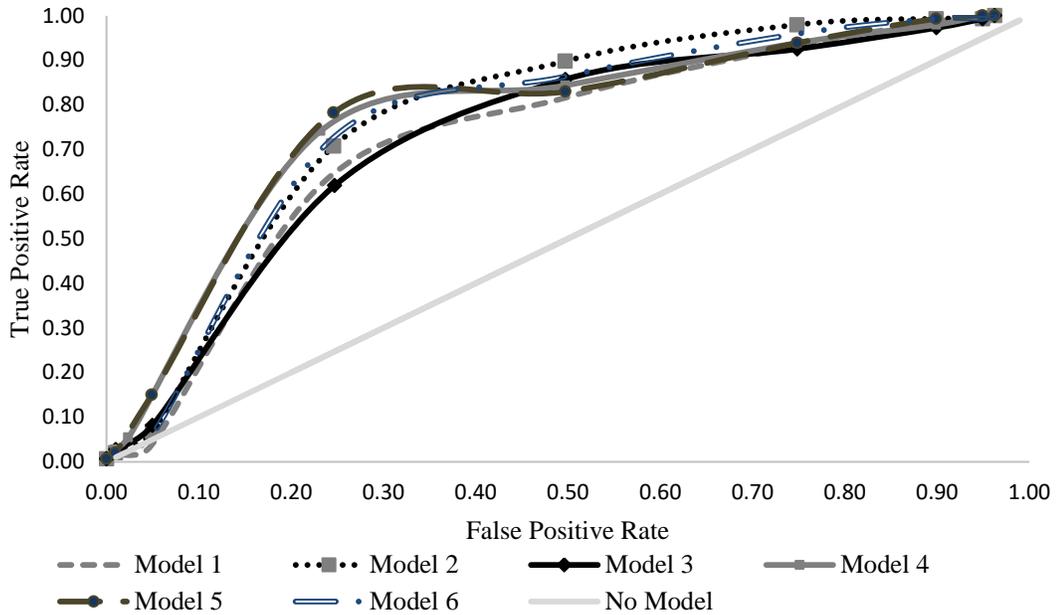


Figure A.2.4. ROC for Offline Purchases



If the predicted probability for a given observed process, i.e., visit incidence, online purchase or offline purchase, is above the threshold level, the customer is predicted to perform the activity. This gives the customer level predictions of each state-dependent process. The hit rate is then determined as the proportion of predictions that match the actual values of a given state dependent process. We note that using this threshold rule, there is a variance in the true and false positive rates across different models. An implication of this is seen in the predictive performance of the models (see section 2.6.3 of Chapter 2), the same threshold probability can produce a high hit rate on activities but low hit rate on inactivity.

CHAPTER 3

VARIABLE SELECTION AND STATISTICAL INFERENCE IN MULTIVARIATE RANDOM FORESTS

3.1. Background

Empirical studies across many disciplines are concerned with the analysis and modeling of multivariate response vectors using longitudinal data. A multivariate response is a vector of measurements taken across K different variables that are possibly jointly associated with a vector of explanatory variables or covariates (Joe 1997). The primary purpose of investigation using multivariate models is two-fold. First, the joint modeling of multiple response variables unravels the covariation or co-occurrence observed in the responses across the K different variables. Second, multivariate models determine the covariates or predictors of interest that are jointly associated with the multiple response variables. Examples of such investigations include ecological studies on coexistence or co-occurrence of multiple species in a geographic location (e.g. De'Ath 2002, Adler et al. 2018); epidemiological studies on the statistical dependency of multiple response variables such as number of emergency room visits, absenteeism and mortality count on exposure to health risk factors (e.g. Joe 1997, Lewis and Ward 2013); psychological studies on joint measurement of multiple sub-scales of psychological well-being (e.g. Miller et al. 2016); and marketing research on inter-related decision making such as consumer buying or browsing behavior across multiple categories or websites (e.g. Manchanda et al. 1999, Park and Fader 2004). In addition to this, multivariate response modeling can be

especially useful when there is high class imbalance or sparsity in some of the response outcomes. In such situations, by jointly modeling multiple responses one can borrow explanatory strength from the less sparse outcomes.

Multivariate response problems are often modeled using parametric regression structures under some stylized assumptions. Some of the more commonly used parametric models are: multivariate probit where the response function is a discretized version of a latent multivariate normal distribution (e.g. Ashford and Sowden 1970, Lesaffre, Verbeke and Molenberghs 1994), multivariate logit with the latent stochastic vector assumed to follow type 1 extreme value or Gumbel distribution (e.g. Glonek and McCullagh 1995) and copula structure which assumes uniform univariate marginals (Joe 1997). Apart from the distributional assumptions, such parametric multivariate models usually impose restrictive linear and parametric relationship with the associated covariates or predictors (see Miller et al. 2016 for discussion). Standard parametric regression methods can be hard to specify when there is large number of response outcomes, i.e., for $K > 3$. Further parametric approaches can be restrictive when the data is high-dimensional, i.e., the number of predictors is large and there are complex and non-linear interactions between predictors and response outcomes.

An alternative approach to the parametric methods is to exploit the non-parametric tree based methods (Breiman, Friedman, Olshen, and Stone 1984, Hothorn et al. 2006). The multivariate extensions of the regression tree identify strata of homogeneous outcomes across multiple response variables (e.g. Segal 1992, De'Ath 2002). In this research, I examine the following: multivariate response model for $K > 3$, sparsity in some of the response outcomes and high dimensional data with non-

linear interactions. Specifically, in this research I examine the application of a non-parametric multivariate random forest (MVRF) to model multivariate response vector of order $K > 3$. Further, I introduce variable importance measures and a variable selection procedure using MVRFs to reduce dimensionality.

In the machine learning literature tree-based ensemble methods have proven to be excellent predictors in classification and regression problems especially for higher order and non-linear interactions (Breiman 2001). Multivariate tree-based ensemble methods include MVRF (Segal and Xiao 2011) and multivariate gradient boosted trees (MVBT; Miller et al. 2016). In recent case studies of pharmaceutical drug responses, it has been demonstrated that when the outcome responses are correlated, MVRFs have higher predictive accuracy over univariate random forests (or RFs), and other ML methods such as Elastic Net and Kernelized Bayesian Multi-Task Learning (Rahman et al. 2017).

While ensemble methods have been used for their predictive accuracy, unlike parametric regression models these are “black-box” methods with limited interpretability. However, a critical factor to improving predictive accuracy is to be able to identify predictors and understand their interactions or associations with the response variable (Breiman 2001). For tree-based ensembles such as random forests a source of interpretation is provided by variable importance measures (VIMs) (Breiman 2001). A VIM is a score assigned to each variable based on its predictive ability. In many fields such as statistical genomics, VIMs have been used as a tool for variable selection and to identify predictors from a large set of candidate variables (Strobl et al. 2007, Ishwaran 2007). Some of the more commonly used methods to measure variable

importance are: prediction error by permuting a variable (Breiman 2001), node impurity in terms of mean squared error (Friedman 2001) and naïve measures such as importance based on the incidence and frequency with which a variable is used in a tree. In addition to this, variable importance has been studied using methods such as functional ANOVA decomposition to understand the interactions between subsets of variables (Hooker 2004).

In this chapter I develop new variable importance measures for variable selection using MVRFs. We propose new methods to measure variable importance based on two different split improvement (SI) criteria. The proposed VIMs score each variable by first summing the magnitude of SI across all node splits the variable is used within a tree, and then averaging across the forest ensemble. The first SI criterion measures the difference in the mean structure between parent and children nodes. This is a multivariate generalization of least squares where the magnitude of SI is the difference between the sum of squared errors at the parent splitting node and those at the children nodes. The second criterion, or the outcome difference SI, sums the magnitude of difference in outcomes of each response variable between left and right children nodes across all splitting nodes that a variable has been used in and then averages across the ensemble. Using the outcome difference SI a variable can be scored differently in its ability to split the multiple response variables. The outcome difference SI thus generates a vector of importance measures for each variable. Our implementation of MVRF uses the R package '*MultivariateRandomForest*'. The VIM currently available on the R package uses the naïve measure of count or frequency based importance. We benchmark our proposed VIMs against the naïve measures of

the average incidence and average frequency with which a variable is used across an ensemble.

To demonstrate the variable selection ability of the proposed importance measures we develop an iterative recursive feature elimination (RFE) strategy to eliminate the least important variables (e.g. Guyon et al. 2002). Our proposed RFE strategy iteratively builds MVRF using bootstrapped sub-samples (e.g. Mentch and Hooker 2016), makes predictions on a test set, computes the importance of each variable and discards the lowest-scored variables at the end of the iteration. To generate a baseline score of VIM of variable removal, we introduce a random noise or a pseudo-covariate in the training set at the start of each iteration. The VIMs are computed for all covariates including the pseudo-covariate after a forest build. All variables with a VIM lower than that of the pseudo-covariate are then discarded at the end of the iteration.

We demonstrate the validity of the proposed VIMs in recovering important covariates under four simulated data scenarios. Each of the four simulation scenarios assumes linearity of relationship between covariates and the multivariate response vector with same number of covariates but varying conditions of error correlation and data sparsity. We introduce spurious variables in the data matrix to test the ability of the proposed VIMs to recover the true covariates as most important. Under each scenario we build an MVRF on the training set and compute the proposed SI and naïve importance measures. We study the rank ordering of the true and spurious covariates using the proposed SI and naïve measures of variable importance. In all simulation scenarios the proposed methods of variable importance are able to recover the true

covariates as accurately as the naïve measures.

In our empirical application, we implement and test the predictive accuracy of the variable importance measures using the RFE strategy on an ecology (e-bird) data provided by the Cornell Lab of Ornithology on migrant bird-species. We sample a set of 5 species from the data set to model co-occurrence or joint sightings. In addition to testing the predictive accuracy of the proposed VIMs, we demonstrate statistical inference procedures using the proposed VIMs using the e-bird data. In the empirical application, we find that the proposed measures of variable importance when applied as a variable selection tool outperform the naïve measures in their prediction accuracy (in terms of mean squared errors or MSEs) and provide a more stable method of variable pruning. Further, we demonstrate inference procedures to determine the stability of the importance scores. The methods developed in this paper make important contributions to research on multivariate models and in particular to multivariate random forests.

The outline of this chapter is as follows. In section 3.2, we discuss the multivariate extension of regression trees and random forests using subbagging procedure. In section 3.3, we discuss the proposed variable importance measures using the SI criteria for the multivariate case. In section 3.4, we discuss the RFE strategy for variable selection using the proposed VIMs. We also propose the application of infinitesimal jackknife variance estimator (e.g. Wager et al. 2014) to examine the distributional properties of the proposed VIMs for retained features. We discuss the results on the simulation studies in section 3.5. In section 3.6 we discuss the robustness of the proposed VIMs using the variable selection procedure on the two

data sets. We also suggest some diagnostic studies using the proposed VIMs and the corresponding implications for association studies in multivariate random forests. We conclude with limitations and scope for future work.

3.2. Multivariate Regression Trees and MVRFs

In regression analysis, random forests can be applied to build trees where the tree predictor takes on numerical values rather than class labels (Breiman 2001). An important decision element associated with a tree-based algorithm is determining the split function. The split function at each splitting node of a multivariate regression tree exploits the between-node heterogeneity using mean and covariance for continuous outcomes (e.g. Segal 1992) and entropy for binary response (e.g. Zhang 1998).

For the multivariate case, the mean structure based split function explores the node heterogeneity by using the difference in sum of squares between the parent node and the children nodes. The covariance structure based split function replaces the sum of squares at each node with the norm of the difference between the sample covariance and hypothesized covariance matrices. Similar to the construction of forests for univariate response outcomes, in the multivariate case individual trees are grown and combined to give the multivariate forest prediction.

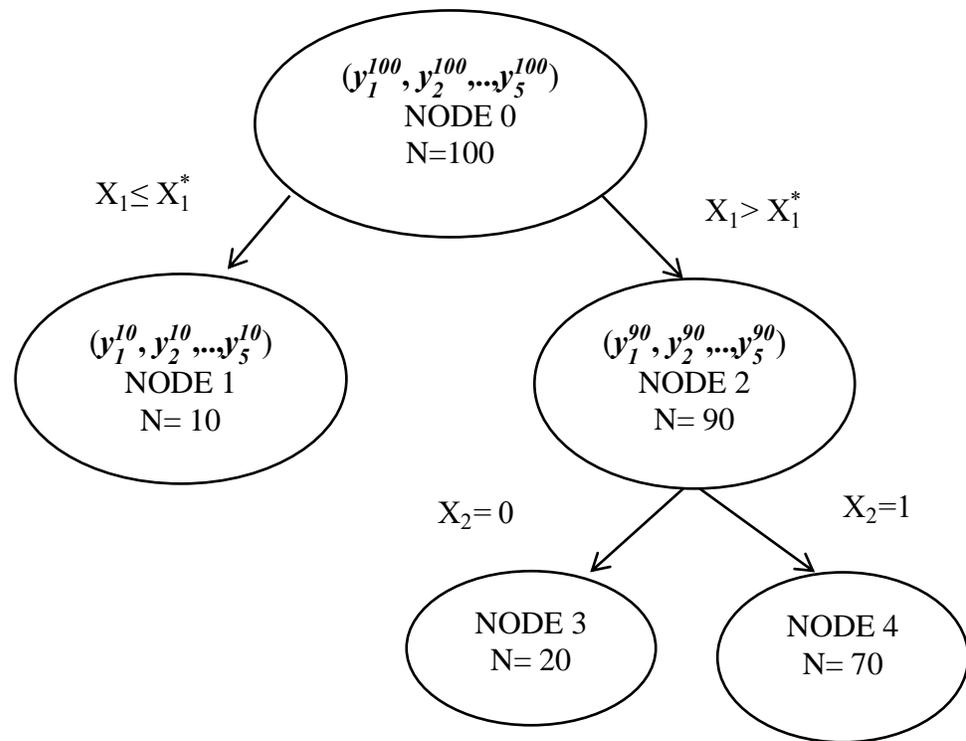
The multivariate regression tree (MVT) method for panel data is developed as follows: suppose there are K outcome variables observed over N time periods denoted by the matrix $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_K\}$, where \mathbf{y}_k is the $(N \times 1)$ vector of observations for the k^{th} outcome in the panel. Further, we assume there are P features or covariates denoted by the vector $\mathbf{X} = \{X_1, X_2, \dots, X_P\}$. A tree algorithm proceeds using a two-step approach. At each node of the tree, the algorithm first draws a random sub-sample

$L \leq P$ of covariates or predictors and examines every allowable split (s) on each predictor variable ($X_l, l = 1, 2, \dots, L$). Second, it determines the best predictor-split combination ($X_l, s(X_l)$) and splits the node into left and right children nodes according to whether $X_l < s(X_l)$. In the case of multivariate outcomes, the covariate used in each node split identifies a cluster of homogeneous multiple outcomes. This algorithm proceeds at each child node and continues until a desired tree size has been grown. The covariates can be either continuous or categorical. In case of continuous variables, the splits are the mid-points between data values. For ordered categorical variables, a split divides the categories into two groups, where the covariate values in one group are larger than those in the other. In case of unordered categorical variables, the split divides the two nodes into disjoint sets of categories.

This algorithm is illustrated in Figure 3.1. We assume the panel length is N in the training set, so that each response outcome we observe is a $(N \times 1)$ vector in the panel. For the purpose of illustration we assume $N=100$ and $K=5$. We denote $\{y_1^{100}, y_2^{100}, \dots, y_K^{100}\}$, as the (100×5) matrix of responses. We demonstrate the tree building algorithm for two node splits – the first split is made by a continuous covariate X_1 and the second by a categorical/ indicator covariate X_2 . Suppose, the first node split of the tree corresponds to the split threshold $s(X_1) = X_1^*$ for the covariate X_1 . The responses are homogeneous across all 5 variables for 10 data points in the panel for $X_1 \leq X_1^*$ and is shown as $\{y_1^{10}, y_2^{10}, \dots, y_5^{10}\}$ under Node 1. The remaining 90 sample points correspond to $X_1 > X_1^*$ is the second homogeneous sub-group, and is denoted by $\{y_1^{90}, y_2^{90}, \dots, y_5^{90}\}$ in Node 2. The next split occurs at Node 2 where the best

splitting variable X_2 is a categorical/indicator. The homogeneous responses among the 5 response variables for $X_2 = 0$ are grouped under Node 3. The tree is formed progressively with nodes getting split further into more homogeneous sub-groups, until no further splits can be found. Note that as the tree progresses, the children nodes get increasingly more homogeneous in terms of the response outcomes. For instance, Node 3 has higher homogeneity in responses across all 5 response variable than Node 2.

Figure 3.1. Multivariate Regression Tree



3.2.1. Multivariate Random Forests (MVRF) using Subbagging

In our development of the multivariate forests, we divide the data into training and testing samples. We assume a K - dimensional outcome vector denotes by

$\mathbf{Y} = \{y_1, y_2 \dots y_K\}$, and M features or predictors denoted by the vector $\mathbf{X} = \{X_1, X_2, \dots, X_M\}$. We use the subbagging algorithm (Andonova et al. 2002, Mentch and Hooker 2016) to bootstrap subsamples of the full training set. The subbagging procedure has been found in many applications to outperform traditional bagging (Zaman and Hirose 2009). To build the multivariate regression trees, we use the *build_single_tree* function on the Multivariate Random Forest R package. The tree prediction is obtained using the *single_tree_prediction* function available on the R package. The trees are built on the bootstrapped subsamples and then aggregated to get the forest prediction.

Algorithm 1: MVRF using Subbagging Procedure

Load training and testing sets, \mathbf{x} and \mathbf{x}^* respectively
 Select size of subsample l_N and number of subsamples r_N from training set of size N
for b in 1 to r_N **do**
 Select subsample of size l_N from training set \mathbf{x}
 Build tree on subsample b
 Use tree at testing set \mathbf{x}^* to get prediction vector $\hat{Y}_{N,l_N,r_N}^b = (\hat{y}_{N,l_N,r_N,1}^b, \hat{y}_{N,l_N,r_N,2}^b, \dots, \hat{y}_{N,l_N,r_N,K}^b)$
end loop for b
 Average the r_N predictions to obtain $\hat{Y}_{N,l_N,r_N} = (\hat{y}_{N,l_N,r_N,1}, \hat{y}_{N,l_N,r_N,2}, \dots, \hat{y}_{N,l_N,r_N,K})$

3.3. Variable Importance Measures in Multivariate Random Forests

3.3.1. Split Improvement Criterion

We develop variable importance measures for the multivariate case based on the split improvement (SI) criterion, i.e., the objective of maximizing either within-

node homogeneity or between-node heterogeneity at each split. This implies that a variable that achieves a higher magnitude of either within-node homogeneity or between node-heterogeneity at a split gets a higher importance. We develop variable importance measures based on SI criterion in two ways: mean structure based (Segal 1992) and the absolute difference in mean outcomes between nodes. In addition, we conduct F-test of significance of node splits and recompute the variable importance measures only for significant splits.

The general procedure to construct the variable importance measures is as follows. We build an ensemble of trees on the subsamples drawn from the training sample. We overlay the testing set on each tree and calculate the SI at each node split using the test sample. The importance assigned to a variable is equal to the magnitude of the SI obtained at a node split. If a variable is used at multiple splitting nodes in a given tree, the SI at each node is added up across all such splitting nodes to get the importance measure of the variable for that tree. The overall importance measure for the variable then simply follows the subbagging procedure for prediction, by taking average of the ensemble.

Algorithm 2: Computing SI based variable importance measures

Load training and testing sets, x and x^* respectively
Select size of subsample l_N and number of subsamples r_N from training set of size N
for b in 1 to r_N **do**
 Select subsample of size l_N from training set x
 Build tree on subsample b with number of splitting nodes Q_b
 Use tree to predict on testing set x^*
 Initialize variable importance measure vector for tree b as $VIM_0^b = 0$ vector of dimension $M \times 1$
 for j in 1 to Q_b **do**
 Calculate magnitude of SI for split j in tree b as SI_{bj}
 for m in 1 to M **do**
 if feature m is used for split j in tree b

$$VIM_{0,m}^b = VIM_{0,m}^b + SI_{bj}$$

 end loop for m
 end for loop for j
 end for loop for b
Average the r_N predictions to obtain final estimate \hat{Y}_{N,l_N,r_N}
Average the r_N calculations of variable importance vector VIM_{N,l_N,r_N}^b to get the final vector VIM_{N,l_N,r_N}^*

3.3.2. Mean Structure based SI

Segal (1992) defines the mean structure based split function $\phi_m(s, g)$ as the difference between the within parent node (g) sum of squares and the within children nodes (g_d , $d = L, R$) sum of squares. That is,

$$\phi_m(s, g) = SS(g) - SS(g_L) - SS(g_R) \quad (1)$$

in which,

$$SS(g) = \sum (\mathbf{y} - \boldsymbol{\mu}(g))^T V(\theta, g)^{-1} (\mathbf{y} - \boldsymbol{\mu}(g)) \quad (2)$$

$$SS(g_d) = \sum_d (\mathbf{y}_d - \boldsymbol{\mu}(g_d))^T V(\theta_d, g_d)^{-1} (\mathbf{y}_d - \boldsymbol{\mu}(g_d)), \quad d = L, R \quad (3)$$

Where, g is the parent node and $g_d, d = L, R$ are the children nodes. The multivariate outcome vectors are denoted by \mathbf{y} and \mathbf{y}_d for parent and children nodes respectively. We define $SS(g)$ and $SS(g_d)$ as the corresponding within node sum of squares. Further, $\boldsymbol{\mu}(g)$ and $\boldsymbol{\mu}(g_d)$ denote the vectors of mean response outcomes for the parent and children nodes respectively. The covariance matrices at the parent and children nodes are denoted by $V(\theta, g)$ and $V(\theta_d, g_d)$ respectively. The parameters are respectively denoted by θ and $\theta_d, d = L, R$. The best split is thus determined as $s^* = \text{argmax } \phi_m(s, g)$. In order to ensure that $\phi_m(s, g)$ is non-negative, the method imposes a restriction on the covariance structures, i.e., $V(\theta, g) = V(\theta_L, g_L) = V(\theta_R, g_R)$.

In the derivation of mean structure based SI importance, we use the formulation as given in (1)-(3) above to quantify the SI contributed by a variable used for a node split. At a given node split of a tree, the outcome vectors, the mean vectors and covariance matrices are replaced by the corresponding values of the testing sample outcome vectors $\mathbf{y}^*, \mathbf{y}_d^*, d = L, R$, the test set sample mean vectors $\hat{\boldsymbol{\mu}}(g), \hat{\boldsymbol{\mu}}(g_d), d = L, R$, and the covariance matrix of the overall test set residual error \hat{V}

respectively. The equivalent test set sum of squares at the parent and children nodes is given as,

$$\widehat{SS}(g) = \sum (\mathbf{y}^* - \hat{\boldsymbol{\mu}}(g))^T \widehat{\mathbf{V}}^{-1} (\mathbf{y}^* - \hat{\boldsymbol{\mu}}(g)) \quad (4)$$

$$\widehat{SS}(g_d) = \sum_d (\mathbf{y}_d^* - \hat{\boldsymbol{\mu}}(g_d))^T \widehat{\mathbf{V}}^{-1} (\mathbf{y}_d^* - \hat{\boldsymbol{\mu}}(g_d)), \quad d = L, R \quad (5)$$

Letting m denote the covariate used in the node split, its corresponding importance measure is then computed from the mean structure based SI as

$$\text{Mean VIM}_m(g) = \widehat{SS}(g) - \widehat{SS}(g_L) - \widehat{SS}(g_R) \quad (6)$$

3.3.3. Absolute Difference in Mean Outcomes based SI

In this method, the SI is defined as the absolute difference in mean outcomes between the left and right children nodes of a split. With a multivariate outcome, this measure results in a vector of absolute difference of the same dimension as the outcome vector \mathbf{y} . Similar to the mean structure based SI in the prior sub-section, we estimate the magnitude of SI on the testing sample. The importance attributed to the covariate m on splitting the k^{th} outcome at the splitting node g is computed as the absolute difference in the corresponding testing sample mean outcomes between left and right nodes,

$$\text{Outcome Difference VIM}_{m,k}(g) = |\hat{\mu}_k(g_L) - \hat{\mu}_k(g_R)| \quad (7)$$

3.3.4. SI Importance with Significance Testing of Node Splits

For both methods discussed above, we fine tune the importance measures by performing a test of significance of node separation. We use the Hotelling's T-squared

two-sample statistic for split function that optimizes node separation rather than within-node homogeneity (Segal 1992). The Hotelling's T-squared statistic is given by,

$$T^2 = \frac{n_L n_R}{(n_L + n_R)} \left(\hat{\boldsymbol{\mu}}(\mathbf{g}_L) - \hat{\boldsymbol{\mu}}(\mathbf{g}_R) \right)^T \hat{\mathbf{V}}^{-1} \left(\hat{\boldsymbol{\mu}}(\mathbf{g}_L) - \hat{\boldsymbol{\mu}}(\mathbf{g}_R) \right) \quad (8)$$

where, n_d is the number of test samples in daughter node $d = 1, 2$.

The T^2 statistic is transformed into an F statistic as follows,

$$F = \frac{n_L + n_R - K - 1}{K(n_L + n_R - 2)} T^2 \sim F_{K, n_L + n_R - K - 1} \quad (9)$$

For the null hypothesis $H_0: \boldsymbol{\mu}(\mathbf{g}_L) = \boldsymbol{\mu}(\mathbf{g}_R)$ the F statistic given in (9) follows an F distribution with K and $n_L + n_R - K - 1$ degrees of freedom. The test rejects the null at level α if the calculated F exceeds the critical value evaluated at α .

For the modifications in the SI based importance measures discussed above, we include the SIs, as given by (6) and (7) only for the splits that are significant using the two-sample F test. For the node splits, where H_0 is not rejected, the importance measure for the corresponding splitting variable takes the value 0. With this modification the general algorithm for the variable importance measure is modified to include the significance testing at each node split. The modified algorithm is provided below.

Algorithm 3: Computing SI Importance Measures for significant splits

Load training and testing sets, x and x^* respectively

Select size of bootstrap subsample l_N and number of subsamples r_N from training set of size N

for b in 1 to r_N **do**

Select subsample of size l_N from training set x

Build tree on subsample b with number of splitting nodes Q_b

Use tree to predict on testing set x^*

Initialize variable importance measure vector for tree b as $VIM_0^b = 0$ vector of dimension $M \times 1$

for j in 1 to Q_b **do**

Calculate magnitude of SI for split j in tree b as SI_{bj}

Perform F test for H_0

for m in 1 to M **do**

if feature m is used for split j in tree b

if H_0 is rejected

$$VIM_{0,m}^b = VIM_{0,m}^b + SI_{bj}$$

else $VIM_{0,m}^b = VIM_{0,m}^b$

end loop for m

end loop for j

end loop for b

Average the r_N predictions to obtain final estimate \hat{Y}_{N,l_N,r_N}

Average the r_N calculations of variable importance vector VIM_{N,l_N,r_N}^b to get the final vector VIM_{N,l_N,r_N}^*

3.4. Variable Selection and Inference Procedures

3.4.1. Variable Selection Using Recursive Feature Elimination Strategy

In this section we describe the recursive feature elimination (RFE) procedure that we have developed to use importance measures as a tool for variable selection. We compare the predictive performance (i.e., mean squared error) of each measure defined in Section 3.3 on the testing set by iteratively retraining the forests using the subbagging procedure described in Section 3.2. In this procedure, in each iteration the prediction from the resulting forest is noted and the variables with the lowest importance scores are dropped. In order to provide a benchmark for the variable importance scores we introduce a random noise term in the covariates list at each iteration of the forest and compute the importance score of the noise. For a given iteration, all covariates that yield an importance score lower than that of the random noise are dropped before the start of the next iteration. Thus, at each iteration the covariates list is pruned based on their relative importance scores and the process continues until a steady state is reached where all covariates have importance scores larger than that of the noise term.

The algorithm for the RFE strategy is explained below. We run this algorithm for each of the importance measures discussed above, i.e., mean structure based SI and absolute outcome difference based SI, computing the VIMs for both pre and post significance testing of node splits. Additionally, we run this iterative retraining procedure for the two naïve versions of importance measures discussed in the introductory paragraphs, the count of number of times and the incidence of variables used in a tree.

Algorithm 4: Iterative RFE Strategy Using Random Forest

Load training and testing sets, x and x^* respectively

Select size of bootstrap subsample l_N and number of subsamples r_N from training set of size N

Select number of iterations $niter$ for the iterative training of forests

Initialize forest iteration $iter = 1$

Generate random noise variable and append to both training and testing sets

Initialize mean variable importance measure VIM_{N,l_N,r_N}^* for first iteration to 0

for $iter$ in 1 to $niter$ **do**

for b in 1 to r_N **do**

 Select subsample of size l_N from training set x

 Build tree on subsample b

 Use tree to predict on testing set x^*

 Use tree to estimate the variable importance measure for each feature with testing set x^*

end loop for b

 Average the r_N predictions to obtain final estimate $\hat{Y}_{N,l_N,r_N,iter}$

 Compute mean squared error vector based on r_N predictions $MSE_{N,l_N,r_N,iter}$

 Average the r_N calculations of variable importance vector VIM_{N,l_N,r_N}^b to get the final vector VIM_{N,l_N,r_N}^*

for m in 1 to M **do**

 If $VIM_{N,l_N,r_N,m}^* < VIM_{N,l_N,r_N,noise}^*$ drop m from the covariates matrix and update training and testing sets

end loop for m

 Regenerate random noise variable and append to the updated training and testing sets

end loop for $iter$

3.4.2. Inference Procedure

In addition to the iterative feature elimination, we may also be interested to examine the reliability of the proposed importance measures in variable selection. This can be done by examining the distributional properties of the importance scores of the retained features. The importance measure for a feature can be viewed as a random variable that follows a distribution with mean and variance parameters. Using the subbagging procedure in section 3.3, and applying algorithms 2 and 3, we compute the sample mean importance. We note that the sample mean generated using the subbagging procedure produces is a consistent estimator of the true mean (Mentch and Hooker 2016). To estimate the variance in the tree-wise importance measures for each feature, we adopt the Infinitesimal Jackknife (IJ) estimate of variance (Efron 2013, Wager et al. 2014). As noted in the literature, the IJ estimate is a consistent estimator of the variance parameter. The IJ variance estimate of the importance measure for the m^{th} feature is,

$$\widehat{V}_m^{IJ} = \sum_{i=1}^N \text{Cov}[I_{i,l_N,r_N}^b, VIM_{N,l_N,r_N,m}^b]^2 \quad (10)$$

where, as before I_{i,l_N,r_N}^b is the number of times the i^{th} training sample is used in the b^{th} bootstrap subsample of size l_N when r_N subsamples are drawn from the training data of size N . The expression $VIM_{N,l_N,r_N,m}^b$ is the importance measure of the m^{th} feature computed from the tree generated by the corresponding bootstrap subsample. Similar to the average importance score for each feature as given by algorithms 2 and 3 in section 3.3, we compute the IJ variance in the tree-wise importance measure for each retained feature using the formula in (10).

3.5. Simulation Studies

We study the robustness of the proposed variable importance measures under four simulation scenarios. The scenarios differ in terms of assumptions made on correlation of errors in the multivariate response generation and data sparsity. In all four simulation studies we construct a $(K \times 1)$ multivariate response vector \mathbf{y} from a specified data generating model; where $K = 4$ and the data generating process has $M' = 5$ explanatory variables. We generate $M'' = 10$ spurious or non-sense covariates as additional columns in the simulated data matrix. Therefore, the first five columns of the overall data matrix \mathbf{X} contain the true explanatory variables used in generating the response vector. Further, the variables in the data matrix \mathbf{X} consist of binomial, uniform and Poisson variables. The simulation design for the full list of variables (explanatory and spurious) for the non-sparse and sparse cases is provided in Table 3.1.

Table 3.1. Simulation Design for Explanatory and Spurious Variables

Variables	Non- sparse data setting	Sparse data setting
Explanatory		
X1	Unif[0,1]	Unif[0,1]
X2	Binom(1,0.5)	Binom(1,0.5)
X3	Poisson(50)	Poisson(50)
X4	Binom(1, Unif[0,0.5])	Binom(1, 0.5)
X5	Unif[0,1]	Unif[0,1]
Spurious		
X6	Binom(1, Unif[0,0.4])	Binom(1, 0.9)
X7	Unif[0,1]	Unif[0,1]
X8	Binom(1, Unif[0,0.4])	Binom(1, 0.9)
X9	Unif[0,0.5]	Unif[0,0.5]
X10	Binom(1, Unif[0,0.3])	Binom(1, 0.9)
X11	Unif[1,1]	Unif[1,1]
X12	Binom(1, Unif[0,0.3])	Binom(1, 0.9)
X13	Unif[0,0.25]	Unif[0,0.25]
X14	Binom(1, Unif[0,0.25])	Binom(1, 1)
X15	Unif[0,0.5]	Unif[0,0.5]

We generate a dataset of size $R = 500$. The data is then split into training ($N = 300$) and testing ($N_{test} = 200$) sets. We build $numforest = 10$ multivariate random forests each with $r_N = 500$ trees. For each forest all 15 ($M' + M''$) variables are then scored based on both the proposed and naïve measures of variable importance. We note that for the proposed VIMs - mean structure based SI (with and without F test) and outcome difference SI (with and without F test), we compute the measures as given in (6) and (7). However while (6) and (7) specifically compute the SI using the test set, for the simulation studies, we compute a second set of SI measures using the actual splits made on the training trees. The naïve measures of incidence and

frequency are computed based on the individual training trees built within an ensemble. The scores of each VIM are then averaged across the $numforest = 10$ forests. The ranks of the variables are computed based on the average scores for each of the importance measures. The test of robustness of a variable importance measure is provided by the ability to recover the rank ordering of the features, i.e., true explanatory variables should get the highest importance measures.

Algorithm 5: Simulation Studies

Generate a data matrix $X = [X_E \ X_S]$, where $X_E = R \times M'$ and $X_S = R \times M''$

Generate y with specified data generating process: $y = f(X_E) + \epsilon$

Split data into training and testing sets, x and x^* respectively

Select number of random forests *numforest*

for *num* in 1 to *numforest* **do**

Select size of bootstrap subsample l_N and number of subsamples r_N from x of size N

for b in 1 to r_N **do**

Select subsample of size l_N from training set x

Build tree on subsample b

Use tree to predict on testing set x^*

for m in 1 to M **do**

 Compute VIM for mean structure based SI (with and without F test) using both b and x^*

 Compute VIM for outcome difference SI (with and without F test) using both b and x^*

 Compute VIM for incidence of m used in tree b

 Compute VIM for frequency of m used in tree b

end loop for m

end loop for b

Average the r_N calculations of the VIMs for each feature m for forest *num*

end loop for *num*

Average the *numforest* calculations of the VIMs for each feature m

Rank order features from 1 to M based on each of the variable importance measures

For brevity we provide detailed results for two of the simulation scenarios and summarize the results from the remaining two⁷.

Scenario 1: Linear Model with no sparsity and uncorrelated errors

We consider the following data generating process (DGP),

$$y_k = \sum_{m=1}^5 a_{km} X_m + \varepsilon_k,$$

where $\varepsilon_k \sim N(0, (\text{var}(\sum_{m=1}^5 a_{km} X_m))/10)$; $k = 1, 2, 3, 4$ and $m = 1, 2, \dots, 5$. The variance of the error term is chosen so that the signal to noise ratio is 10.

The coefficients of the explanatory variables are specified as:

$$A = \begin{bmatrix} 1.85 & 0.95 & -0.05 & 0.95 & -0.85 \\ 1.3 & 0.9 & 0.08 & 0.8 & -0.75 \\ 2.45 & 0.8 & 0.09 & 0.95 & -0.9 \\ 1.01 & 0.9 & -0.09 & 0.8 & 0.75 \end{bmatrix}$$

where row k represents the coefficients associated with response y_k and column m represents the contribution of X_m .

As seen by the variable rank ordering recovered by the alternative importance measures in Table 3.2, the mean structure based importance measure recovers the true explanatory variables when the SI is computed using the actual tree splits. This is true for both the cases when the including all splits (without F test) and only significant splits (with F test). However, when the test set is used to determine the SI for all splits, the mean structure fails to recover the true covariates in 3 cases. The mean structure SI when applied only for the significant splits (with F test), fails to recover 1 of the true

⁷ The results for the remaining simulations can be provided upon request.

covariates with the test set data. The outcome difference SI performs consistently across training and testing sets in recovering the true covariates under both conditions of all splits (without F test) and only significant splits (with F test) are included. An important point to note here is that while the incidence based VIM recovers the true covariates, it allocates the same rank to 4 of the 5 explanatory variables. That is, it fails to distinguish the rank ordering among the explanatory variables.

Table 3.2. Variable Rank Ordering under Scenario 1

Variables	Freq.	Incid.	Mean Structure		Mean Structure with F test		Outcome Difference		Outcome Difference with F test	
			Trn	Tst	Trn	Tst	Trn	Tst	Trn	Tst
Explanatory										
X1	2	1	1	1	1	3	1	4	1	4
X2	3	1	3	15	2	1	3	1	2	1
X3	1	1	2	14	3	15	2	3	3	2
X4	5	5	5	9	5	2	4	2	4	3
X5	4	1	4	2	4	4	5	5	5	5
Spurious										
X6	10	10	10	6	9	6	10	6	10	6
X7	8	9	8	10	8	10	8	11	8	11
X8	11	11	11	7	11	8	11	7	11	7
X9	6	6	6	12	6	12	6	12	7	12
X10	13	13	13	5	13	7	13	9	13	9
X11	15	15	15	8	15	14	15	15	15	15
X12	14	14	14	4	14	9	14	10	14	10
X13	9	8	9	11	10	11	9	14	9	13
X14	12	12	12	3	12	5	12	8	12	8
X15	7	7	7	13	7	13	7	13	6	14

Freq. = Frequency based VIM, Incid. = Incidence based VIM, Trn = SI computed on training trees, Tst = SI computed using test set on training tree splits

Scenario 2: Non-Linear Model with sparse data and uncorrelated errors

We consider a non-linear DGP to create a sparse data scenario specified as,

$$y_k = I_k * \exp(1),$$

where I_k is an indicator function generated from the binomial process

$$I_k = \text{Binom} \left(1, P \left(\text{Logistic} \left(\sum_{m=1}^5 a_{km} X_m + \varepsilon_k \right) \right) \right)$$

The response vector is non-linearly dependent on the covariates through the binomial parameter. The coefficients associated with the explanatory variables under the sparse condition are given as,

$$B = \begin{bmatrix} 4.85 & 1.5 & -0.1 & 1.45 & -0.09 \\ 5.3 & 2.01 & -0.08 & 1.02 & -0.07 \\ 4.45 & 1.24 & -0.09 & 1.02 & -0.08 \\ 3.01 & 1.05 & -0.09 & 1.02 & 0.075 \end{bmatrix}$$

where row k represents the coefficients associated with I_k and column m represents the contribution of X_m . Further, as in scenario 1, $\varepsilon_k \sim N(0, (\text{var}(\sum_{m=1}^5 a_{km} X_m))/10)$; $k = 1, 2, 3, 4$ and $m = 1, 2, \dots, 5$. All the covariates with the exception of $X_4 \sim \text{Binom}(1, 0.5)$ are generated identically as Scenario 1. The variable rank ordering results are provided in Table 3.3.

Table 3.3. Variable Rank Ordering under Scenario 2

Variables	Freq.	Incid.	Mean Structure		Mean Structure with F test		Outcome Difference		Outcome Difference with F test	
			Trn	Tst	Trn	Tst	Trn	Tst	Trn	Tst
Explanatory										
X1	1	1	1	7	2	3	1	1	1	1
X2	2	1	2	1	1	1	2	2	2	2
X3	3	3	3	9	3	4	3	3	3	3
X4	4	4	4	2	4	2	4	4	4	4
X5	6	6	5	11	6	9	7	6	7	6
Spurious										
X6	11	11	11	3	11	6	10	12	10	12
X7	5	5	6	13	5	11	5	5	5	9
X8	12	12	12	4	12	8	12	13	12	13
X9	7	7	7	12	7	10	6	7	6	7
X10	10	10	10	5	10	5	11	10	11	5
X11	14	14	14	8	14	14	14	14	14	14
X12	13	13	13	6	13	7	13	11	13	8
X13	8	8	8	14	8	12	8	9	8	11
X14	15	15	15	10	15	14	15	14	15	14
X15	9	9	9	15	9	13	9	8	9	10

Freq. = Frequency based VIM, Incid. = Incidence based VIM, Trn = SI computed on training trees, Tst = SI computed using test set on training tree splits

Under the sparse data scenario, with uncorrelated errors, all the VIMs perform similarly in terms of variable rank ordering. All six measures fail to recover X_5 as a true explanatory variable. The mean structure based SI measure when applied on the test set data performs the weakest in its ability to recover the true covariates but is the strongest on the actual training subsample trees.

For the simulation scenarios 3 and 4 (results not shown here), we replicate the DGP

of scenarios 1 and 2 respectively under correlated errors. We find that correlation of errors do not significantly alter the performance results of the VIMs from the scenarios 1 and 2. This could partly be due to the fact that the contribution of the errors and the underlying correlations assumed in these scenarios are not large relative to the explained variation.

3.6. Empirical Application

3.6.1. Application on EBird Data

Our empirical application uses an ecology data set provided by the Cornell Lab of Ornithology on observer sightings of migrant bird species. The data set contains sightings of 25 neo-tropical migrant bird species (warblers and vireos) in the North-East US for the monthly period of June 2016. This data contains 235,036 observer group row entries. Each row entry in the data set contains the count of sightings of each bird species within 0.25 km of search distance and 0.25 search hours by an observer group. Since our primary objective is to model co-occurrence of multiple species, we remove all row entries that report zero sightings across all 25 species. This reduces the data set size to 27,873 observer group entries. We sample a set of five bird species *Setaphaga Americana*, *Setaphaga Petechia*, *Vireo Gilvus*, *Vireo Olivaceus* and *Vireo Solitarius* to model the multivariate co-occurrence outcome.

We define the multivariate response as a 5×1 vector of count of sightings made by an observer group. The summary statistics for count of sightings of the selected species in the reduced data set (27,873 entries) is provided in Table 3.4. Each observer group entry records a set of observer specific features, temporal and ecological factors

associated with the sightings. These are the predictors or covariates used to model the count of sightings. We have a total of 85 predictor variables in the data set.

Table 3.4. Distribution of Sightings Count by Species

Species	No. of Sightings	Sightings as % of Observer Entries
<i>Vireo Solitarius</i>	1,788	6.4%
<i>Setophaga Americana</i>	1,813	6.5%
<i>Vireo Gilvus</i>	3,775	13.5%
<i>Setophaga Petechia</i>	11,219	40.3%
<i>Vireo Olivaceus</i>	14,579	52.3%

For model training we sample 50% observer entries (14,073) and retain the rest as holdout (13,836 entries). From the training set, we bootstrap $r_N = 500$ subsamples of size $l_N = 500$. From the holdout data, we sample 500 entries to construct the testing set. Our modeling objectives test for predictive accuracy on two counts. First, we compare the predictive accuracy of the proposed VIMs using the RFE procedure assuming independence of species sightings. That is, we first model the count of sightings of each species as a univariate outcome. The iterative RFE strategy thus builds univariate random forests (RFs) on each species. Second, we compare the predictive accuracy the proposed VIMs by modeling co-occurrence of multiple species. In the second case, the iterative RFE procedure builds multivariate trees and aggregates into an MVRF in an iteration as discussed in section 3.2. In both the univariate and multivariate cases we perform 30 iterations of recursive feature elimination for each of the VIMs. We record the test set predictions in terms of mean squared errors (MSEs) at the end of each iteration. In Tables 3.5 and 3.6 we report the

test set MSE at the end of the 30th iteration for the univariate and multivariate cases respectively.

Table 3.5. Test Set Mean Squared Error for Univariate RFs

Importance Measures	<i>Vireo Solitarius</i>	<i>Setophaga Americana</i>	<i>Vireo Gilvus</i>	<i>Setophaga Petechia</i>	<i>Vireo Olivaceus</i>
Frequency	0.0873	0.1128	0.2368	1.4267	0.6135
Incidence	0.0776	0.1053	0.2280	1.2588	0.6439
Mean Structure	0.1012	0.1172	0.2573	1.5121	0.7257
Mean Structure with F test	0.0859	0.0951	0.2150	1.3573	0.7013
Outcome Difference	0.0798	0.0836	0.2133	1.3597	0.6454
Outcome Difference with F test	0.0783	0.1060	0.1958	1.3631	0.6600

Table 3.6. Test Set Mean Squared Error for MVRFs

Importance Measures	<i>Vireo Solitarius</i>	<i>Setophaga Americana</i>	<i>Vireo Gilvus</i>	<i>Setophaga Petechia</i>	<i>Vireo Olivaceus</i>
Frequency	0.0779	0.0893	0.2134	1.3534	0.5945
Incidence	0.0817	0.0968	0.2123	1.3774	0.6398
Mean Structure	0.0860	0.0903	0.2242	1.3852	0.7374
Mean Structure with F test	0.0833	0.0883	0.2076	1.3339	0.7150
Outcome Difference	0.0822	0.0861	0.2010	1.2681	0.6577
Outcome Difference with F test	0.0818	0.0876	0.2035	1.2838	0.6725

Based on these results we make three important observations. First, the predictive accuracy of two of the rarer species *Setophaga Americana* and *Vireo Gilvus* and one of the more dominantly observed species *Setophaga Petechia* improves using a multivariate model irrespective of the importance measure selected for feature elimination. This result validates that MVRF can accurately identify features that are jointly correlated with the sightings of multiple species. Therefore in case of high class imbalance an MVRF can improve predictive ability by borrowing strength from the common factors that explain co-occurrence of both rare and dominant species. Second, for the univariate case, the RFE procedure using the proposed importance measures (with the exception of mean structure SI without F test) have higher predictive accuracy on the sightings of two of the rarer species, i.e., *Setaphaga Americana* and *Vireo Gilvus*. Finally, for the multivariate case, RFE procedure using both the mean structure based SI importance measure (with F test) and the outcome difference measures (with and without F test) perform better than the naïve measures of frequency and incidence for predictions on the three species *Setophaga Americana*, *Setophaga Petechia* and *Vireo Gilvus*.

These results provide encouraging validation to both the choice of a multivariate model for feature identification and the robustness of the proposed SI based importance measures.

3.6.2. Variable Importance Measures for Interpretation and Inference Procedures

In this section, we propose and discuss application of the variable selection method using the SI importance measures for interpretation and statistical inference.

For purposes of demonstration, we focus on the results from two of the best performing proposed importance measures— mean structure SI importance with F-test and outcome difference SI importance.

Post variable selection, we rank the retained features in order of importance towards explaining the multivariate outcome based on each measure. The ranking of the features vary based on the proposed measures used. Specifically, the mean structure based SI scores features across multivariate response outcomes. In the specific empirical context, the mean structure SI ranks features in order of importance of explaining joint occurrence of all five bird species. However, with outcome difference importance although the MVRF variable selection procedure trains on the joint sightings of species, the variables are scored separately for each species. This helps recover individual feature ranks for each outcome. In Table 3.7, we summarize the top 5 features across all bird species by the mean structure with F test and the species specific top 5 features using the outcome difference measures respectively. We apply the IJ estimator for variance as given in Section 3.4.2 to estimate the variance of the variable importance scores, and construct box-plots and confidence intervals (CIs). The importance score distribution of the top 5 features selected by the mean structure based SI (with F-test) across all species is shown in Figure 3.2 below. The species specific importance score distribution of the top 5 features selected by the outcome difference SI are given in Figures 3.3 – 3.7.

Table 3.7. Top 5 Ranked Features

Rank	Mean Structure Rank across Species	Outcome Difference SI Rank by Species				
		<i>Vireo Solitarius</i>	<i>Setophaga Americana</i>	<i>Vireo Gilvus</i>	<i>Setophaga Petechia</i>	<i>Vireo Olivaceus</i>
1	Effort Distance	I.Stationary	Effort Hours	Day	Day	Day
2	Shallow Ocean ED	Day	Day	Time	Effort Hours	Effort Hours
3	No. Observers	Time	Elevation	Effort Distance	Effort Distance	Time
4	Shallow Ocean LPI	Effort Hours	Time	Effort Hours	Shallow Ocean PD	I.Stationary
5	Shallow Ocean PD	Effort Distance	No. Observers	I.Stationary	Time	Effort Distance

The table reports the top 5 features across all species using mean structure SI and species specific top 5 features using outcome difference SI

Figure 3.2. Mean Structure (F test) Importance Distribution of Top 5 Features

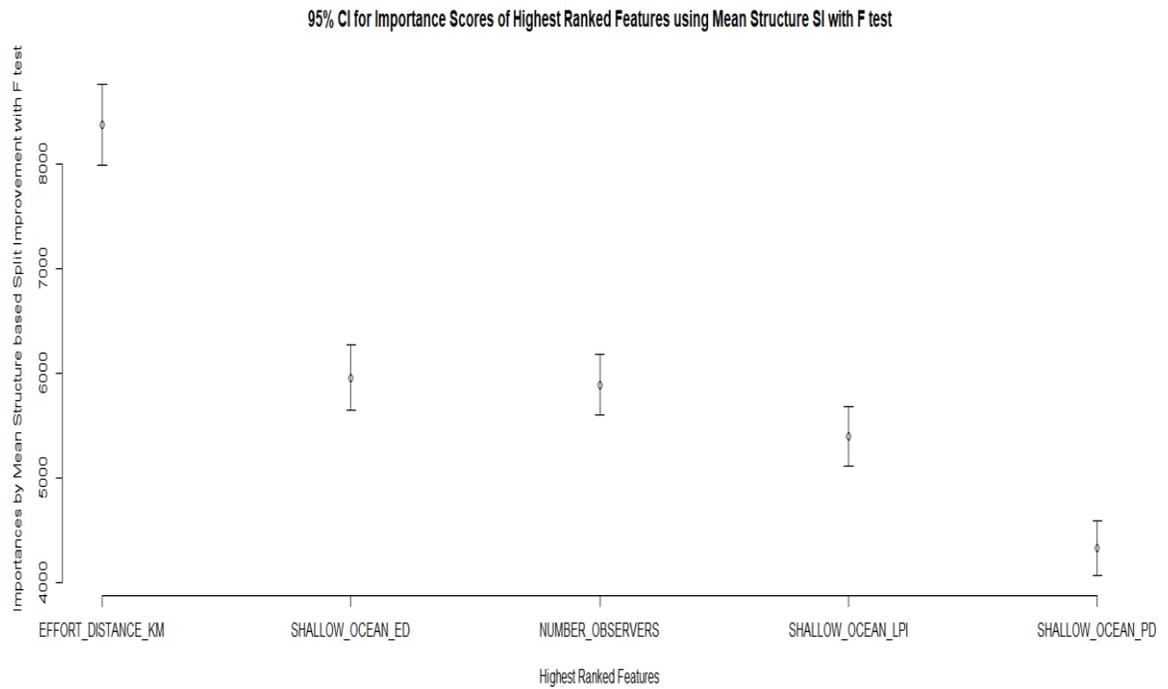
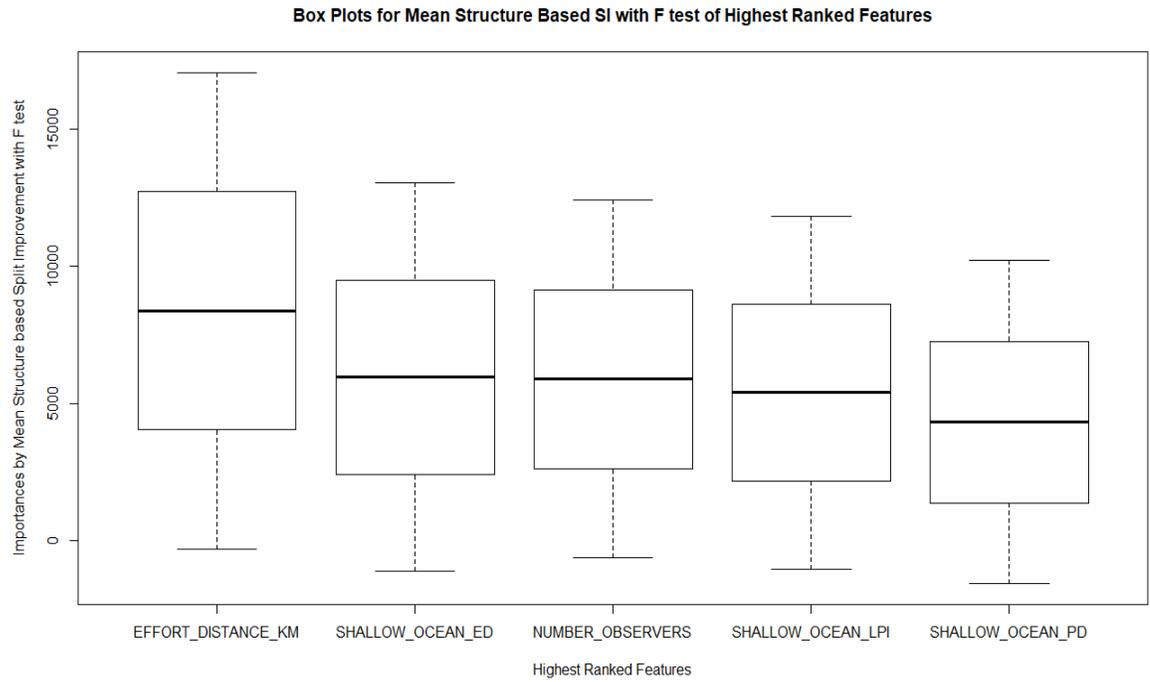


Figure 3.3. Outcome Difference Importance Distribution of Top 5 Features (V. Solitarius)

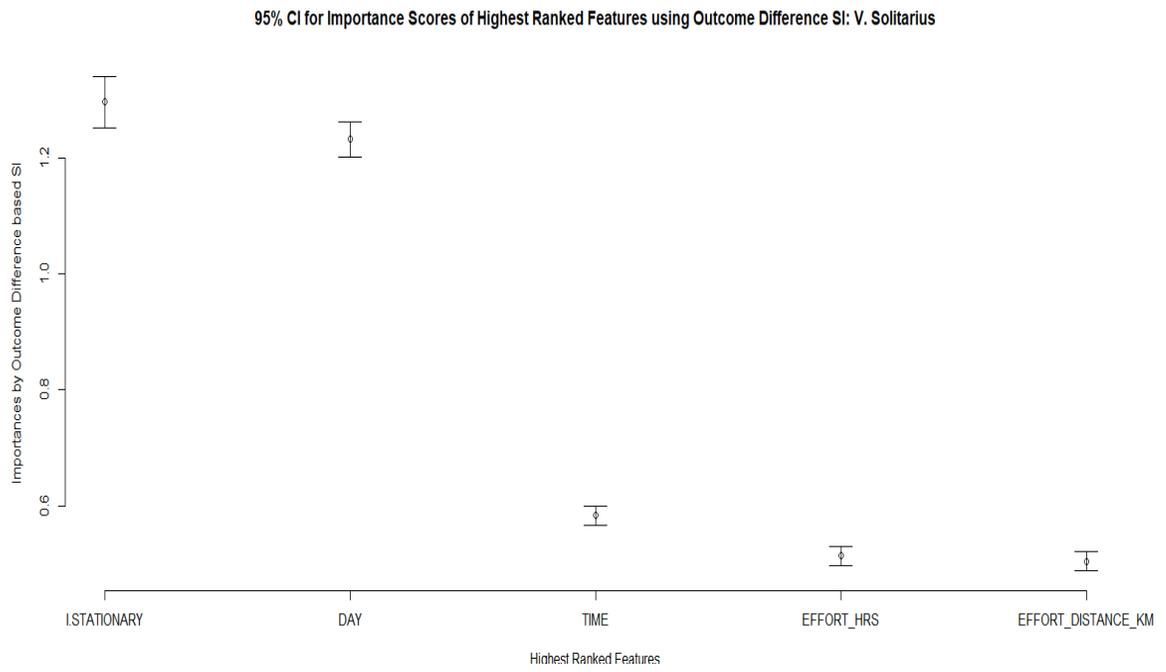
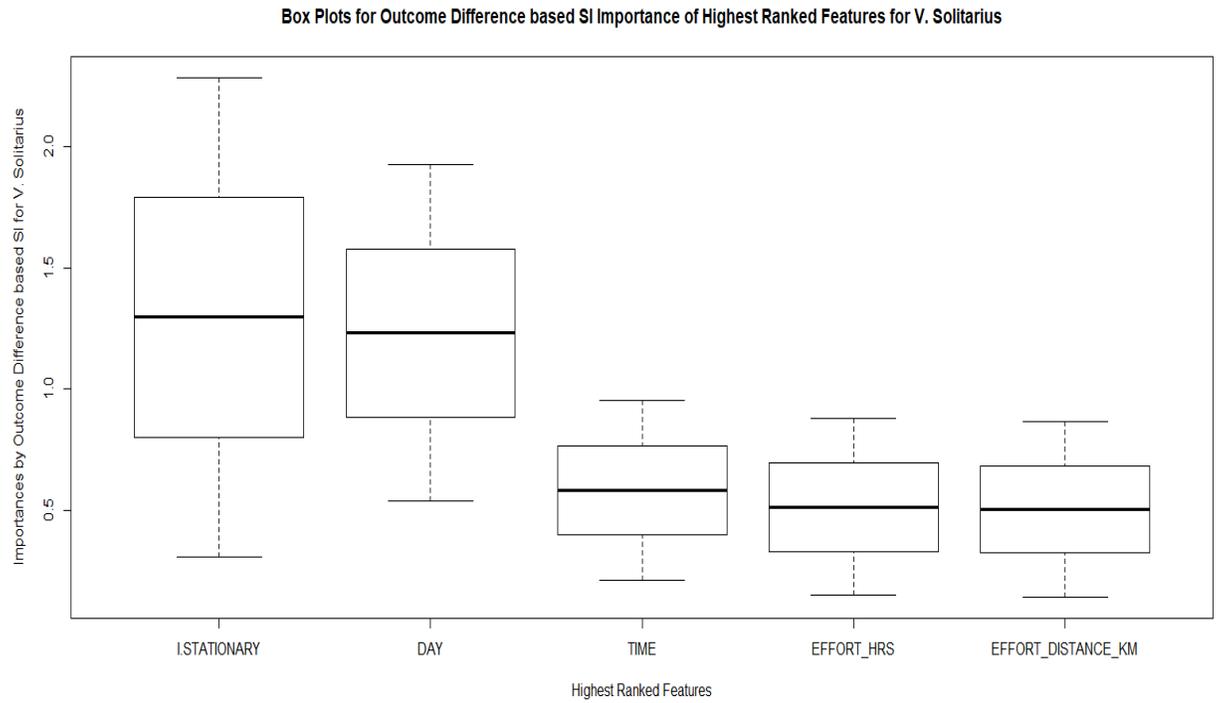


Figure 3.4. Outcome Difference Importance Distribution of Top 5 Features (*S. Americana*)

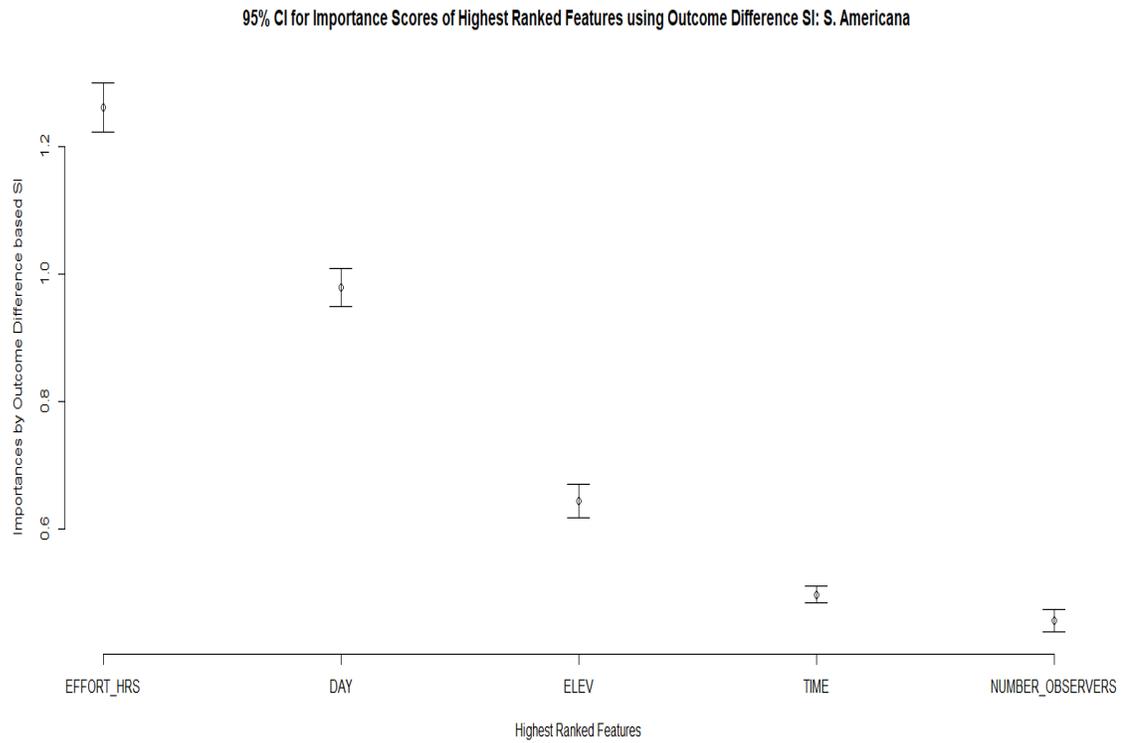
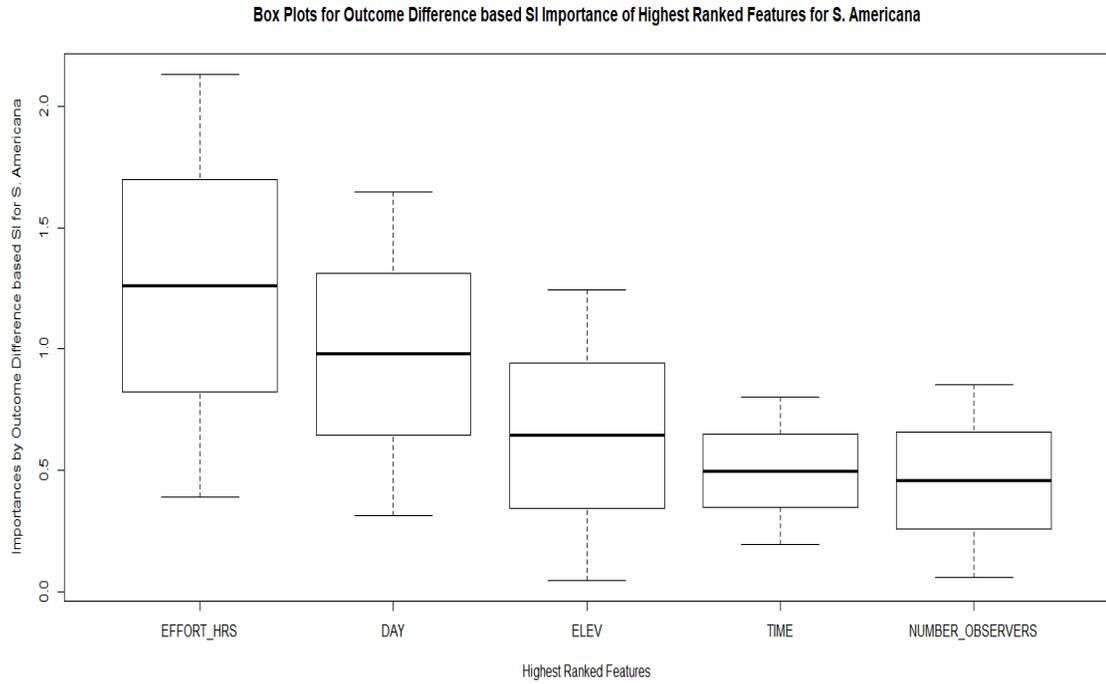


Figure 3.5. Outcome Difference Importance Distribution of Top 5 Features (V. Gilvus)

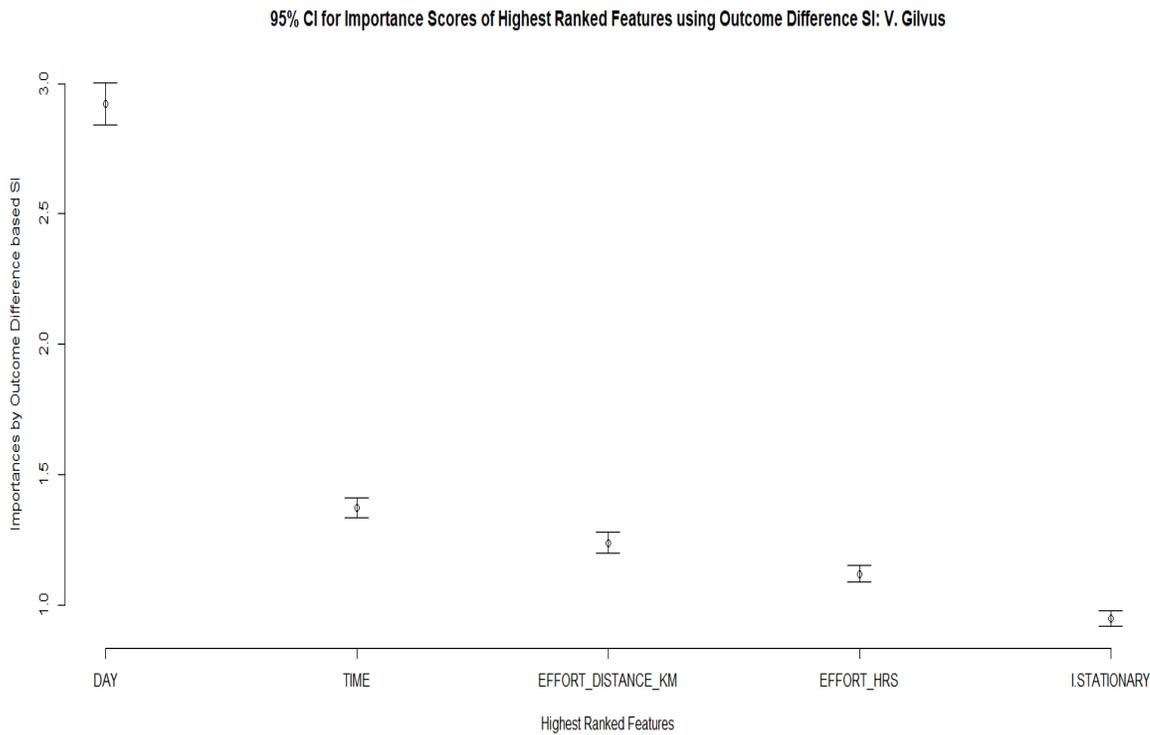
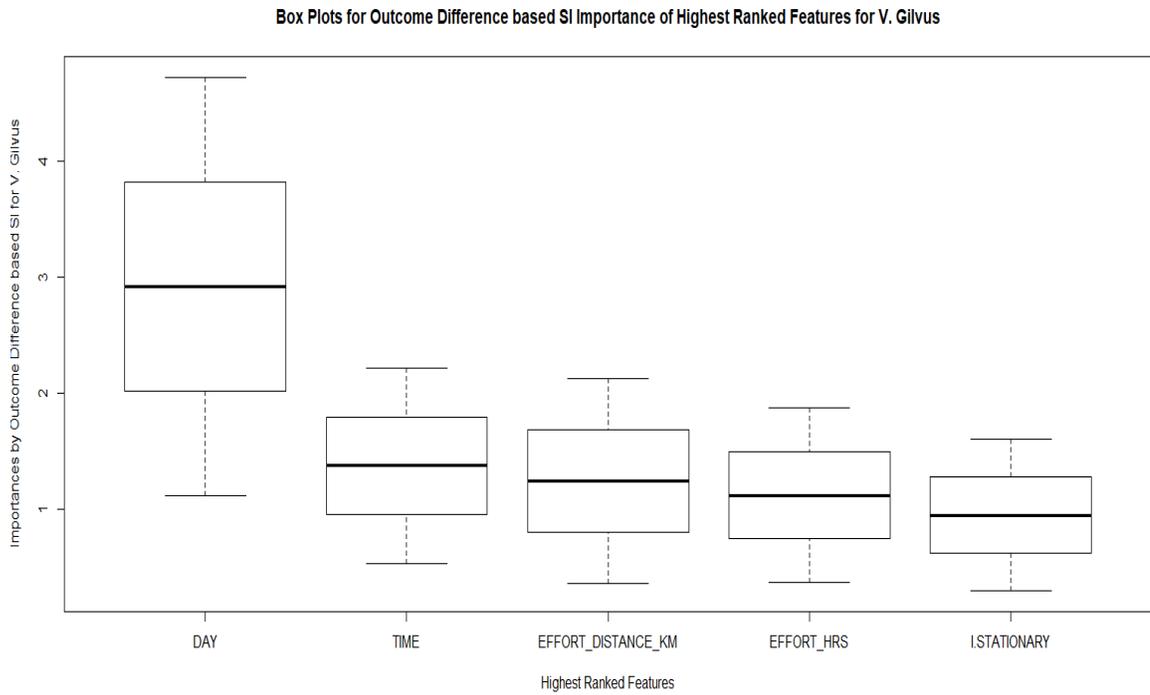


Figure 3.6. Outcome Difference Importance Distribution of Top 5 Features (*S. Petechia*)

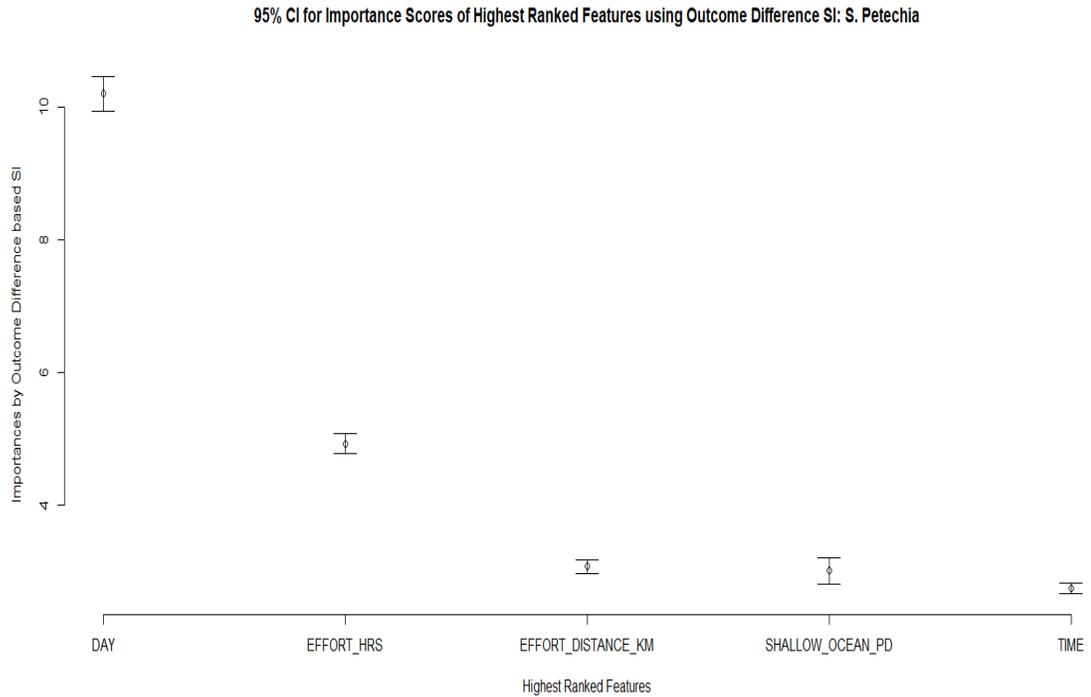
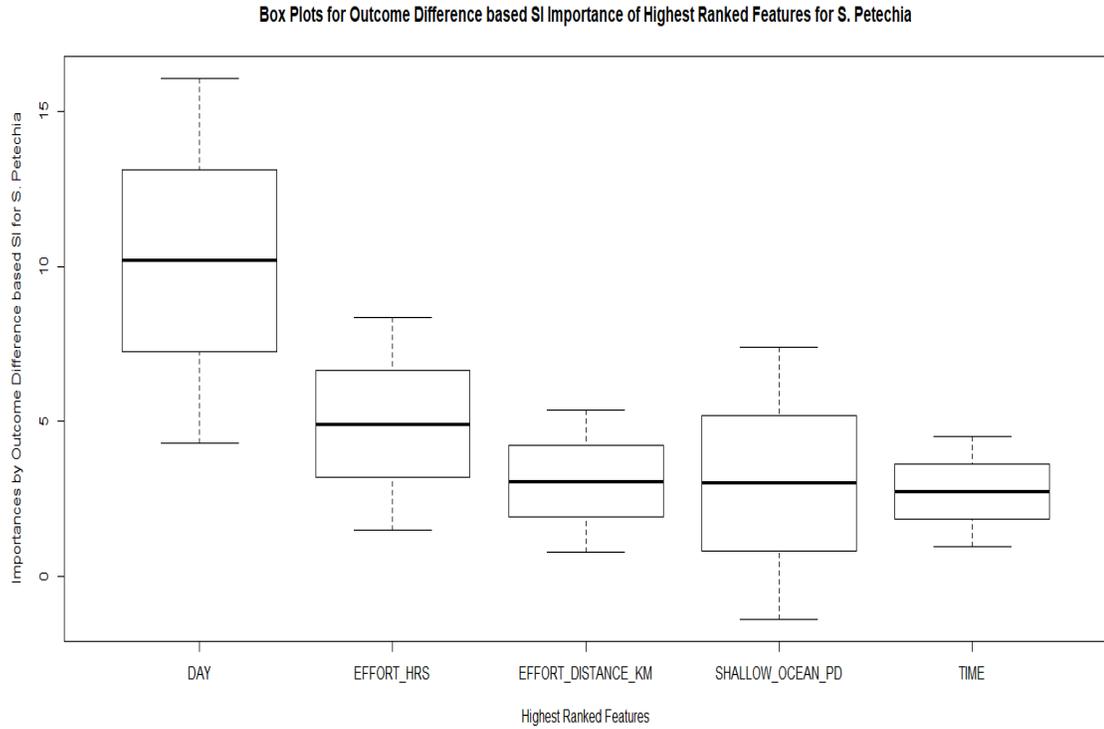
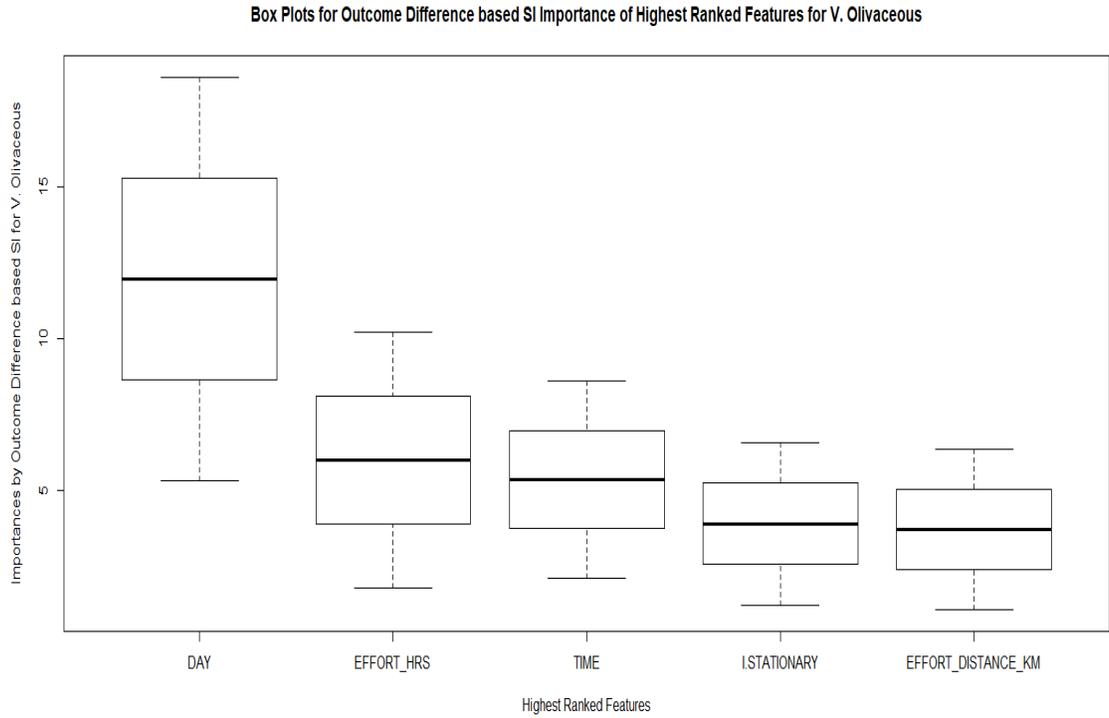
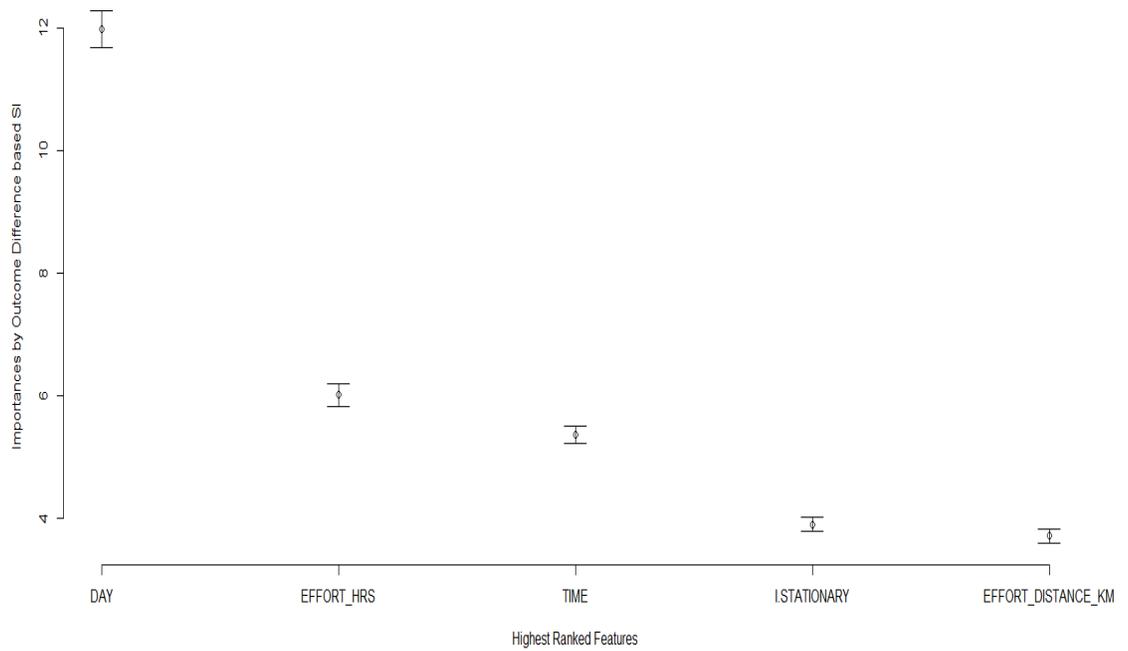


Figure 3.7. Outcome Difference Importance Distribution of Top 5 Features (V. Olivaceous)



95% CI for Importance Scores of Highest Ranked Features using Outcome Difference SI: V. Olivaceous



From Table 3.7 and the figures 3.2-3.7 we make two observations on the application of the proposed importance measures for variable selection. First, out of the top five high ranked features selected by the mean structure based SI importance measure three of these are ecological predictors. Recall that the mean structure SI calculates the difference in the generalized sum of squares among the nodes (parent and children nodes). Thus, a variable assigned a higher score using this measure has a higher ability to jointly split among multiple response outcomes. Since the co-occurrence of multiple species often depends on ecological factors, the mean structure SI measure gives highest ranks to some of these features. The outcome difference SI measure calculates the outcome specific absolute difference between the children nodes. Thus, a variable assigned a higher score for a specific outcome (in this case species) is better able to separate responses associated with that outcome variable. In this application, except for the sightings of the species *Setaphaga Petechia*, the outcome difference SI measure gives the highest ranks to observer group specific predictors and temporal predictors such as day and time of sightings. This implies that for individual species level sightings, more focused effort by observer groups and specific times are more relevant than ecological factors.

Second, the spread or inter-quartile (IQ) range of the box plots and the width of the CIs are determinants of the reliability of the importance measure and rank ordering produced by it. As an overall goal, the variable selection procedure using the proposed measures will be reliable if we can recover the same set of high ranked features and preferably in the same rank order using different samples from the population. A

lower inter-quartile range will indicate lower variability and higher stability in the importance score of a feature across multiple samples. Shorter CIs will indicate higher precision of the importance scores assigned to a feature. Further, for two closely ranked features we would want the CIs to be non-overlapping to ensure the rank ordering is preserved under multiple sampling scenarios. From inspection of the Figures 3.2-3.7, we find that the mean structure based SI importance has wider IQ range for all the top ranked features relative to the outcome difference SI measure. Further evidence of this is found in the CIs computed using both mean structure and outcome difference measures. We find that the CIs produced by the mean structure SI are wider and three of the adjacent ranked features (*Shallow Ocean ED*, *Number of observers* and *Shallow Ocean LPI*) have overlapping intervals. The species specific CIs produced by the outcome difference SI measure are much tighter indicating higher precision. Further, with the exception of *Vireo Solitarius* (for features *Stationary* and *Day*) and *Setaphaga Petechia* (for features *Effort distance* and *Shallow Ocean PD*) the species specific CIs across adjacent ranked features are non-overlapping. This indicates that the variable rank ordering is more stable when using the outcome difference SI measure.

Therefore, while we have shown in section 3.6.1 that both the proposed measures discussed perform better than naïve measures in predicting sightings of three among the five species examined, the variables isolated by these measures are different. Based on the overall research goal, that is, either to identify features that jointly explain the multivariate response outcome or to identify features specific to a response outcome one can employ one of the two measures. Further, the reliability of the

variable ranking may differ based on the measure used. We find that the outcome difference measure gives a more stable rank ordering of the features.

For interpretation of the underlying relationship between features and outcome, the variable selection procedure using either of the proposed importance measures can be used as a pre-processing step in high-dimensional multivariate problems to extract high-ranked features. The extracted features can then be used in standard parametric or non-parametric multivariate regression analysis to investigate the nature of interaction, linearity of relationship, and significance of coefficients in parametric specifications. We demonstrate this in the empirical application using Amazon marketplace data in Chapter 4.

3.7. Conclusion

In this chapter, we examine and propose novel methods of measuring variable importance for variable selection and inference in multivariate random forests. Our proposed methods exploit the split improvement criterion and node heterogeneity in determining the importance scores. We proposed two methods based on split improvement – mean structure based SI and outcome difference based SI measures. These measures when used as tools for variable selection give higher predictive accuracy than the naïve measures in multivariate outcome problems. Further, we examine the distributional properties of the importance measures developed and discuss the reliability of variable ranking produced by the proposed measures. We propose that the choice of the importance measure will depend on the research goal. Though more reliable in feature ranking, the outcome difference SI measure isolates

outcome specific predictors from a multivariate response model. The mean structure SI isolates predictors that jointly determine the multivariate response. We further propose that the proposed measures and the variable selection procedure (RFE strategy) can be applied to reduce features in high dimensional multivariate response problems. The high ranked features can then be examined using standard parametric or non-parametric multivariate regression setting to examine the underlying nature of relationship between outcomes and predictors. The proposed method for feature extraction in multivariate models will be useful to researchers in fields of ecology, marketing, economics, computational biology, genomics and biological statistics.

REFERENCES

- Andonova S, Elisseeff A, Evgeniou T, Pontil M (2002, July) A simple algorithm for learning stable machines. In *ECAI* (pp. 513-517).
- Adler PB, Kleinhesselink A, Hooker G, Taylor JB, Teller B, Ellner SP (2018) Weak interspecific interactions in a sagebrush steppe? Conflicting evidence from observations and experiments. *Ecology*. 99(7), 1621-1632.
- Ashford JR, Sowden RR (1970) Multi-variate probit analysis. *Biometrics*, 535-546.
- Athey S, Tibshirani J, Wager S (2016) Generalized Random Forests. *arXiv preprint arXiv:1610.01271*.
- Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. *Wadsworth Int. Group*. 37(15), 237-251.
- Breiman L (1996) Bagging predictors. *Machine learning*. 24(2), 123-140.
- Breiman L (2001) Random forests. *Machine learning*. 45(1), 5-32.
- De'Ath G (2002) Multivariate regression trees: a new technique for modeling species–environment relationships. *Ecology*. 83(4), 1105-1117.
- Efron B (2014) Estimation and accuracy after model selection. *J. American Stat Assoc*. 109(507), 991-1007.
- Ghosal I, Hooker G (2018) Boosting Random Forests to Reduce Bias; One-Step Boosted Forest and its Variance Estimate. *arXiv preprint <https://arxiv.org/abs/1803.08000>*
- Glonek GF, McCullagh P (1995) Multivariate logistic models. *J R Stat. Soc.: Series B (Methodological)*, 57(3), 533-546.
- Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. *Machine learning*. 46(1-3), 389-422.
- Hothorn T, Hornik K, Zeileis A (2006) Unbiased recursive partitioning: A conditional inference framework. *J Comp. and Graph. stat*. 15(3), 651-674.
- Hooker G (2004, August) Discovering additive structure in black box functions In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 575-580). ACM.

- Ishwaran H (2007) Variable importance in binary regression trees and forests. *Electronic J. Stat.* 1: 519-537.
- Joe H (1997) *Multivariate models and multivariate dependence concepts*. Chapman and Hall/CRC.
- Lesaffre E, Verbeke G, Molenberghs G (1994) A sensitivity analysis of two multivariate response models. *Comp. stat. & data analysis*, 17(4), 363-391.
- Lewis FI, Ward MP (2013) Improving epidemiologic data analyses through multivariate regression modelling. *Emerging themes in epidemiology*, 10(1), 4.
- Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P (2004) Screening large-scale association study data: exploiting interactions using random forests. *BMC genetics*, 5(1), 32.
- Manchanda P, Ansari A, Gupta S (1999) The “shopping basket”: A model for multicategory purchase incidence decisions. *Marketing Sci.*, 18(2), 95-114.
- Mentch L, Hooker G (2016) Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learning Res.*, 17(1), 841-881.
- Miller PJ, Lubke GH, McArtor DB, Bergeman CS (2016) Finding structure in data using multivariate tree boosting. *Psych. methods*. 21(4), 583.
- Park YH, Fader PS (2004) Modeling browsing behavior at multiple websites. *Marketing Sci.* 23(3), 280-303.
- Rahman R, Otridge J, Pal R (2017) IntegratedMRF: random forest-based framework for integrating prediction from different data types. *Bioinformatics* 33.9: 1407-1410.
- Segal MR (1992) Tree-structured methods for longitudinal data. *J. American Stat. Assoc.* 87(418), 407-418.
- Segal M, Xiao Y (2011) Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 80-87.
- Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*. 8(1), 25.
- Wager S, Hastie T, Efron B (2014) Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *J. Mach. Learning Res.* 15(1), 1625-1651.

Zaman F, Hirose H (2009, December) Effect of subsampling rate on subbagging and related ensembles of stable classifiers. In *International Conf. on Pattern Recognition and Machine Intelligence* (pp. 44-49). Springer, Berlin, Heidelberg.

Zhang H (1998) Classification trees for multiple binary responses. *J. American Stat. Assoc.*, 93(441), 180-193.

CHAPTER 4

INVESTIGATING MULTIVARIATE PRICE DYNAMICS: AN APPLICATION TO AMAZON MARKETPLACE

4.1. Background

“There are two kinds of companies, those that work to try to charge more and those that work to charge less. We will be the second” – Jeff Bezos, Amazon CEO

The philosophy of Amazon CEO Jeff Bezos has been to always provide Amazon customers with lowest prices. However, the business press reports that for any item or category, Amazon does not necessarily provide the lowest price relative to rivals and nor is the price steady (ProPublica 2016, Splinter News 2016). Amazon is reported to use dynamic or algorithmic pricing- prices that are set based on computer algorithms (Chen et al. 2016) and that vary over time (Splinter News 2016). An important feature in Amazon is the presence of independent sellers. The Amazon Marketplace was launched in 2000 to enable independent or third-party (3P hereon) retailers to sell alongside Amazon. Currently 3P sellers account for 52% of all paid units sold on Amazon (Statista 2018). Amazon’s algorithmic pricing is likely to affect 3P sellers’ pricing since these sellers often sell same or similar items as Amazon. In an attempt to offer competitive prices, many 3P sellers on the marketplace are also reported to follow pricing strategies similar to Amazon (Chen et al. 2016). Therefore, 3P seller pricing strategies are an integral part of examining pricing on Amazon marketplace.

Despite the extensive business press coverage of the Amazon price dynamics, there do not appear to be significant academic studies of this phenomenon or other aspects of the Amazon marketplace (Chen et. al. 2016 and Bajari et al. 2018 are notable exceptions). 3P sellers on Amazon can observe prices and features of all sellers and only their own sales. Our primary research objective is to identify the key drivers of price changes for Amazon and the 3P sellers from this observed set of features. Since Amazon and 3P sellers sell similar products, we expect dependencies in their pricing. We jointly model the price changes of Amazon and 3P sellers as a multivariate price change response model. To account for the complexities of the marketplace, our model allows for non-linear and non-parametric relationships between price changes and such covariates.

There are several empirical challenges in the joint modeling of price change decisions of Amazon and 3P sellers. First, the marketplace allows 3P sellers to enter and exit a category or an item with minimal cost and effort. Therefore, for any category, we observe a large number of 3P sellers with discontinuous and sparse data at an individual seller level. Other changing marketplace conditions (e.g. new category and brand introductions, changing stock market pressures on competitors) also are likely to result in price change decisions of the sellers. This is an additional reason why long panel length might be unsuitable. Second, price changes can be triggered by several observed factors. For each brand, we observe a list of attributes such as star rating, sales rank, and customer reviews. Similarly, for each 3P seller we observe several descriptors such as seller star rating, percentage positive rating, free shipping conditions and prices offered. These two empirical challenges lead to the classical

“small n (number of observations) large p (covariates)” problem- too few observations relative to the covariate space. Potential non-linear relationships between these covariates and outcome variables create further complexities in modeling. We address the problem of “small n large p ” and complex and non-linear interactions of covariates with the multivariate response model using a novel variable selection algorithm.

In the statistics and machine learning (ML) literatures, several methods of variable selection have been proposed. These methods have been classified as filter, wrapper, embedded (e.g. LASSO, Tibshirani 1996) and ensemble techniques⁸ (Abeel et al. 2009, Meinshausen and Bühlmann 2010). Several scientific disciplines such as computational biology, genomics and biological statistics employ ML ensemble variable selection methods to deal with “small n , large p ” problems. Methods include tree-based ensemble methods such as random forests (Breiman 2001) and gradient boosting (Friedman 2001). In marketing, there has been limited research on “small n , large p ” problems (see Trusov, Bodapati and Bucklin 2010 for an early mention). Furthermore, marketing scientists have used tree based algorithms only in the context of prediction problems (Lemmens and Croux 2006, Neslin et al. 2006, Rafieian and Yoganarasimhan 2017, Yoganarasimhan 2016).

In this chapter we propose the random forest based recursive feature elimination (RFE) algorithm developed in Chapter 3 to solve this “small n , large p ” problem. We exploit the properties of node splits in multivariate regression trees (or MVTs; Segal 1992) and resulting ensemble of multivariate random forests (or MVRFs hereon). We apply two of the proposed split improvement based VIMs introduced in

Chapter 3 for variable selection.

Post the variable selection based on the iterative RFE strategy, we train a series of multivariate regression models with varying degrees of flexibility in parametric and non-parametric covariates specification. First is the most flexible, the non-linear and non-parametric MVRF itself. The second model is the linear parametric vector autoregressive model with exogenous covariates (VAR-X, e.g. Srinivasan et al. 2004, Steenkamp et al. 2005). The third is the generalized additive model (GAM, e.g. Friedman, Hastie, and Tibshirani. 2001, Wood 2006) which allows for both linear parametric and non-parametric specifications. We benchmark the predictive performance of these multivariate models against the null vector autoregressive (VAR) model with no exogenous covariates (e.g. Pesaran and Smith 1998) and the embedded variable selection method of LASSO (e.g. Tibshirani 1996).

The data for this research is scraped from Amazon's website for the electric cooker category with a sample of fourteen brands from September 2017 through April 2018. We choose Amazon's best-selling electric cooker brand Instant Pot as the focal brand for our analysis. We confine our examination to price change of new items. Since there are a large number of heterogeneous 3P sellers, we form five clusters of relatively homogeneous sellers and develop our price change outcome measures. The resulting five clusters are based on mean seller characteristics such as seller reputation as reported on Amazon reviews, tenure, number of unique brands sold, and terms of shipping and order fulfilment. Our covariates include cluster-level seller characteristics, brand characteristics and past price changes made by Amazon and the

⁸ Haury et al. 2011 and Lazar et al. 2012 provide a survey of the methods; see section 3 for detailed

3P seller clusters. Additionally, we include the price changes observed on used items on both focal and non-focal brands in the covariates space to account for effects of used items on price change decisions of new items. We also include characteristics of the 3P sellers who sell used Instant Pot and other brands.

We estimate the multivariate outcome for magnitude of price change⁹. We find that by using variables selected by the proposed variable selection algorithm, the GAM model yields the best predictive performance. The significant coefficients from the GAM estimation provide insights into the factors that trigger price changes among sellers on the marketplace. The superiority of GAM in predictive performance highlights the non-linearity of association between the covariates and the price change outcome. Our variable selection methodology works as a pre-processing step to filter important predictors into the multivariate regression models. Furthermore, the model selection is based on predictive performance. Thus this methodology helps in attaining the dual objective of predictive accuracy and interpretability. In the statistical machine learning literature, model interpretability is often achieved at the expense of predictive ability. Our work is one of the few studies that suggest otherwise (see Tan et al. 2018 for another notable exception).

A brief preview of the results is as follows. We find some non-linear effects of covariates on price changes. This provides support for using a flexible functional form. For example, we find that as small-scale seller ratings improve up to a threshold, Amazon's price changes are more likely. Past this threshold, Amazon's

discussion and literature review.

⁹ We also estimate the multivariate outcome of price change with direction included. These results and discussion are in Appendix A4.

price changes are less driven by changes in small sellers' reputation. We also find differential impact of category variables on price changes of various sellers. For example, Amazon and established 3P sellers' price changes are less driven by category variables; price changes for smaller and newer 3P sellers are more dependent on category variables. This suggests the possibility that Amazon and more established 3P sellers rely on other sources of information, e.g. cross-category relationships in their price change decisions. The less established sellers' pricing is more dependent on within-category pricing dynamics.

This research is of interest to marketing scholars for the following reasons. Our most important methodological contribution is the development of a new ensemble variable selection technique in multivariate modeling with "small n large p". We have demonstrated variable selection as inputs to a GAM model, and found good out-of-sample robustness including relative to the more frequently used (embedded method) LASSO. As discussed in Chapter 3, we have found the proposed split improvement based VIMs to give better out-of-sample predictions than the naïve measures found in the multivariate random forest R package (*MultivariateRandomForest.R*). The RFE strategy using the proposed VIMs is an extension of the ensemble variable selection technique. This can be used as a pre-processing step for any high-dimensional non-parametric modeling, e.g. for design of field and laboratory experiments, and other statistical and econometric applications.

Substantively, the findings of price dynamics on Amazon in this chapter are one of the first in marketing, economics, and information management. The model and research findings have important implications for the business community especially

the 3P retailers, manufacturers and regulatory authorities. Existing and new 3P sellers and Amazon can use our methodology to identify key variables that affect pricing of other firms in the marketplace, and appropriately alter their own strategies for pricing, entry and exit. Manufacturers can use this information to understand what triggers price changes among 3P sellers and Amazon. Regulators might also find our methodology and results of interest. There has been increasing business press coverage (Wall Street Journal 2017, Investor's Business Daily 2017, Axios 2018) on concerns of growing dominance of a handful of technology companies including Amazon (FAANG- Facebook, Apple, Amazon, Netflix and Google). Our methodology can be used by regulators to monitor pricing competition on Amazon marketplace with data that can be scraped publicly from its website.

The rest of the chapter is organized as follows. Section 4.2 provides an overview of the Amazon marketplace landscape, data description and highlights the data challenges. In section 4.3, I review the relevant literature. I provide the methodological approach using MVRF as variable selection tool in section 4.4 and lay out the modeling framework in section 4.5. I present the empirical application of our modeling framework in section 4.6 and discuss the key results and findings in section 4.7. I conclude with a discussion of limitations and scope for future research work.

4.2. Empirical Setting: The Amazon Marketplace

4.2.1. Amazon and Third Party Sellers

The Amazon marketplace sells products in multiple categories (or departments) such as books, electronics, kitchen and dining, and furniture categories.

On the product page of an item, Amazon displays information on its features, its sales rank within its category, default or buy box seller for the item, shipping options under the default seller, other sellers who offer the item, sponsored listings and other similar items.

For its 3P sellers Amazon offers professional and individual selling plans. Under the professional plan, a seller pays \$39.99 monthly subscription fee in addition to per item sales fee and can sell more than 40 items per month. Sellers opting to sell fewer than 40 items per month can choose the individual plan and pay a fee of \$0.99 per item sold. Additionally, depending on the nature of the category, the sellers pay variable referral and closing fees.

The 3P sellers have an opportunity to sell and fulfil shipment on their own, i.e., fulfilment by merchant (or FBM), or utilize Amazon's services to warehouse and fulfil shipments, i.e., fulfilment by Amazon (FBA) for a fee. Amazon maintains and displays records of seller ratings and customer reviews for the sellers of each item sold on its marketplace. Both Amazon and 3P sellers have the option to delist their items temporarily from Amazon. The temporary delisting of items can occur due to a variety of factors such as inventory running out, seller away on holiday, etc. De novo entry and exit also affect item availability. Though Amazon allows its customers to comparison-shop by listing all sellers of the product, around 82% of sales happen through the coveted "buy box" or the "Add to Cart" box that quotes price of the buy box winning seller (Chen et al. 2016).

4.2.2. Data and Descriptive Statistics

We analyze price variations of Amazon and 3P sellers¹⁰ for electric cookers. We examined price fluctuations across several items from multiple categories on commercial websites such as keepa.com and camel.com. Based on this, we found that the electric cookers category had tractability both in terms of number of brands and 3P sellers. That is, this category sells a range of reputed brands. Additionally, since the marketplace has low barriers to entry it can potentially allow any establishment or individual to enlist as a seller (see business press reports on Amazon fake sellers, e.g., Forbes 2017, The Penny Hoarder 2017). We find that in the electric cooker category the number and variance in 3P seller types is lower (e.g. average star rating is equal across clusters) than in other categories such as apparel, electronics etc. Thus, the 3P sellers in this category are likely to be legitimate establishments.

We scraped the data from Amazon's website daily from September 13th 2017 through April 30th 2018. Due to server failures on some days during data scraping, the final tracking sample is 222 days. For our analysis, we selected a sample of 14 brands of electric cookers based on Amazon's top suggestions through keyword searches. Of the 14 brands, we denote Amazon's brand of choice or best-selling brand Instant Pot as the focal brand and the rest as non-focal brands. For the focal brand Instant Pot,

¹⁰ 3P sellers can sell across multiple categories within the Amazon marketplace. One approach is to select products and then track the 3P sellers. The alternative is to first select 3P sellers and then track products they sell. However, the products sold by a 3P seller are across multiple pages and not all products are offered by the seller at all times. This makes it harder to collect information with this approach. We take the former approach of taking a sample of products from a category and track

there is a range of similarly priced and functionally similar SKUs that differ only slightly in terms of physical dimensions, color scheme etc. We choose 7 such SKUs for the focal brand Instant Pot. For each non-focal brand, we select 1 functionally similar SKU to Instant Pot SKUs. Our data scraping involves scraping of the product characteristics from the product page of each SKU tracked, followed by scraping the first page view of the 3P seller features listed on “Other Sellers on Amazon”. Table 4.1 reports the brands and the number of 3P sellers observed in the marketplace during the tracking period for each brand. We indicate by (\checkmark/\times) whether or not Amazon sells the reported brands. We also note that a seller can sell multiple brands.

Table 4.1. Description of Data Tracked

Brands	No. of 3P Sellers	Amazon (\checkmark/\times)
Instant Pot	35	\checkmark
Crock-Pot	52	\checkmark
Hamilton Beach	33	\checkmark
Oyama	1	\times
Avalon Bay	5	\checkmark
Cusimax	2	\checkmark
Cuisinart	50	\checkmark
Elechomes	2	\checkmark
Pressure Pro	14	\times
Magic Mill	8	\checkmark
Geek Chef	13	\checkmark
Cosori	7	\checkmark
Elite Platinum	41	\checkmark
Breville	46	\checkmark

Each row entry reports the number of sellers offering the given brand in the tracking period and a seller may be accounted in multiple entries.

information on the sellers and prices. We discuss how results for 3P sellers (and Amazon, which also sells across multiple categories) might be affected by this focus on single-category data.

In our data, we find 236 unique sellers observed in the marketplace during the tracking period. Amazon marketplace allows sales of both new and used items of a given SKU. There are several instances through the tracking period where both Amazon and 3P sellers offer items in either condition. In our data, 119 third-party sellers sell used items of Instant Pot and 25 sellers sell used items of non-focal brands at different points in the tracking period. We note that the two sets of used item 3P sellers are not mutually exclusive, i.e., some of these 3P sellers sell used items of both Instant Pot and non-focal brands¹¹.

Table 4.2 summarizes the brand characteristics. The average buy box price shows that Crock-Pot and Hamilton Beach are at the lower price range of \$25-\$35 and Breville at the highest price range of above \$200. In terms of brand reputation as denoted by product star rating (≥ 4.4), the top brands are Instant Pot, Magic Mill, Hamilton Beach and Crock-Pot. The sales rank reported on Amazon's marketplace is directly related to the sales volume, and given Crock-Pot has the highest within category sales rank implies that it has the highest sales among the brands in this category. However, since Amazon's choice (see Business Insider 2018 for overview of Amazon's choice) in the electric cookers category is Instant Pot, we use this brand as the focal brand.

¹¹ In particular, three 3P sellers offer both focal and non-focal used items in the tracking period. There

Table 4.2. Average Brand Characteristics

Brands	Buy Box Price	No. of Answered Questions	Star Rating	No. of Reviews	Sales Rank
Instant Pot	\$116.21	490	4.5	6,627	6,003
Crock-Pot	\$26.55	220	4.4	2,522	318
Hamilton Beach	\$30.39	114	4.4	974	5,380
Oyama	\$49.99	3	3.7	3	37,927
Avalon Bay	\$71.19	10	3.8	26	15,578
Cusimax	\$73.73	9	3.7	15	32,988
Cuisinart	\$79.49	244	4.3	1,453	6,069
Elechomes	\$79.67	76	4.2	143	23,288
Pressure Pro	\$82.01	50	4	79	30,886
Magic Mill	\$83.43	53	4.5	73	37,680
Geek Chef	\$83.46	137	3.8	137	34,788
Cosori	\$87.09	137	4.3	394	14,113
Elite Platinum	\$92.90	271	4.2	764	22,795
Breville	\$236.75	135	4.2	190	8,733

Table 4.3 reports the characteristics of 3P sellers observed in the tracking period. On average, each 3P seller sells 1 unique SKU with a standard deviation of 1, and the maximum number of SKUs sold per seller is 8. Only 17% of the sellers have an FBA arrangement with Amazon. Nearly 94% of the 3P sellers provide free shipping which includes the sellers opting for FBA free shipping. The mean number of customer ratings is 7,432 with a high standard deviation of 12,919. Since the number of customer ratings is cumulative over the tenure of the seller on Amazon marketplace, the standard deviation indicates that many of the sellers are new entrants. The number of days a seller is seen on the marketplace is on average 57 in the 222 days of tracking. Additionally, we note that an average seller changes prices on only 7

are exactly 141 distinct 3P sellers that sell used items in either focal or non-focal brands.

days. We also find that 39% ~ 93 third-party sellers do not make any price changes at all in the tracking period. The summary statistics indicate that there is a large number of 3P sellers with sufficient heterogeneity in terms of number of unique SKUs sold, tenure on Amazon marketplace as given by the total number of ratings, number of days present on the marketplace, and price change frequency.

Table 4.3. 3P Seller Features

Features	Mean	Max.	Min.	Std. dev.
Fulfilment by Amazon (FBA)	17%	100%	0%	37%
Free ship Offers	94%	100%	0%	24%
% Positive Rating	92%	100%	0%	8%
Star Rating	4.68	5.00	1.00	0.36
Total No. of Ratings	7,432	76,495	1	12,919
No. of unique items sold	1	8	1	1
No. of days seen on marketplace	57	222	1	46
No. of daily price changes made	7	120	0	16

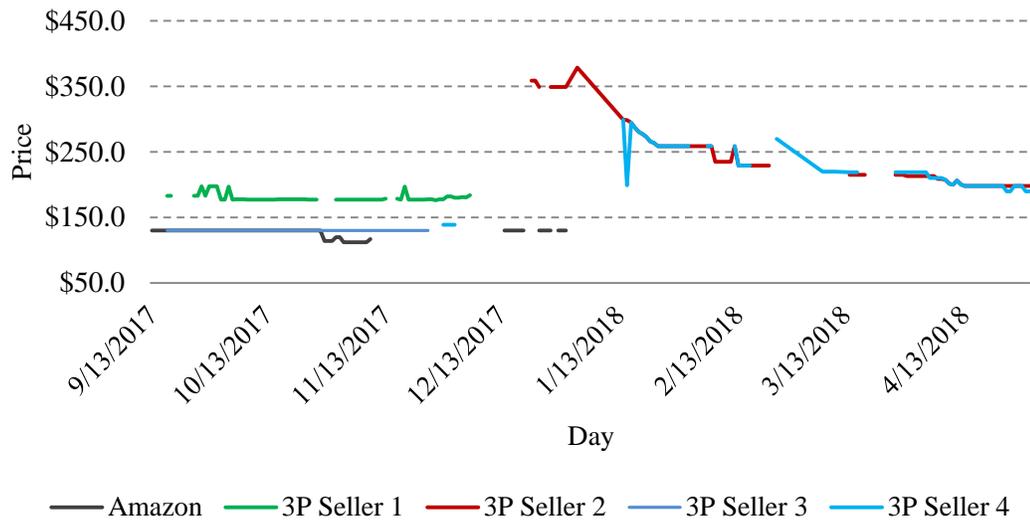
The features FBA and Free ship offers are 1/0 indicators. The mean represents the percentage sellers for these two features.

4.2.3. Data Challenges and Implications for Modeling

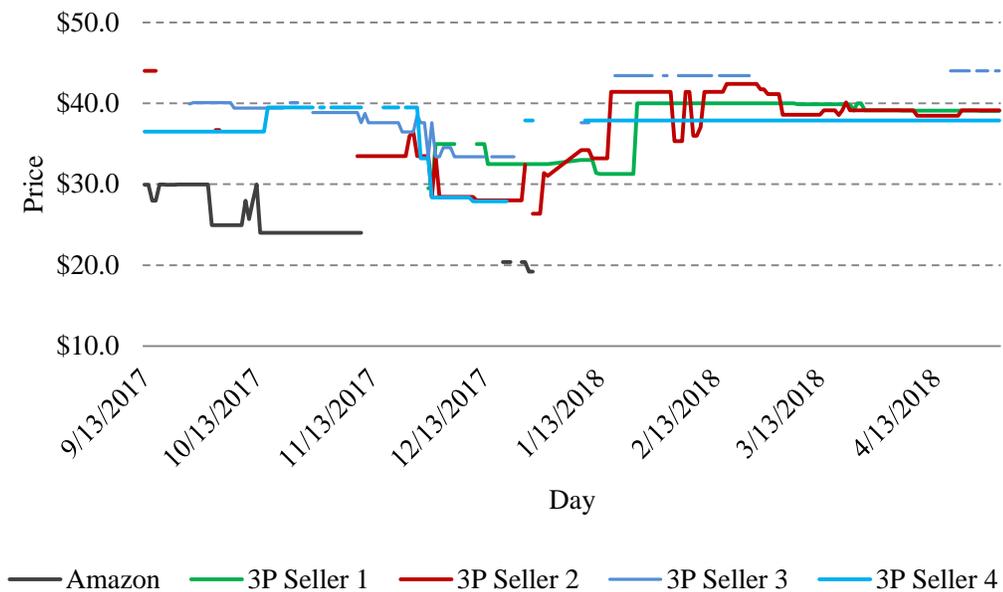
We discuss features of the marketplace that have important modeling implications. First, due to entry, exit, and temporary delisting we observe a large number of 3P sellers with very sparse data at the individual seller level. In table 4.3, we note that the standard deviation of the number of days a 3P seller is seen on the marketplace is 46, with minimum being 1 and maximum 222. We illustrate this in the

daily price change trends for Instant Pot and Crock-Pot in Figures 4.1 (a) and 4.1 (b) respectively.

Figure 4.1. Daily Price Trends on Amazon for Instant Pot and Crock Pot



a) Instant Pot



b) Crock-Pot

The graphs report the price trends of new items of the respective brands sold by Amazon and top 4 third-party sellers (based on number of days seen on the marketplace) in either brand. The price change trend lines show discontinuities at an individual seller level where sellers enter or exit the marketplace. As a result, the marketplace for a given category has a changing roster of 3P sellers, with sparse data per seller.

Second, the marketplace provides a large amount of information on the 3P sellers and products sold. We observe the consumer reviews and ratings and sales rank at individual product level and sales reputation of individual sellers. The pricing behavior of sellers can be triggered by a large number of factors in the marketplace. For instance, a low consumer rating of a brand may trigger sellers to reduce prices for it. Alternatively, a price change triggered by a seller with high reputation may trigger a similar response from others. The many observed features of the marketplace lead to a large number of covariates.

Third, the marketplace landscape may change with new brand introductions, changes in brand features, and the introduction of substitute categories. As a result, price changes may be triggered by changing sets of factors over time. Therefore, unlike empirical studies of steady competitive environments the length of the panel may need to be restricted to effectively identify the predictors of price dynamics.

These data challenges lead to the following modeling considerations. Sparse data on 3P sellers makes empirically examining individual seller behavior difficult. We address this issue by seller aggregation or clustering in section 2.4. Second, we have chosen a data collection period where there appear to be no large changes in the

competitive landscape. In this time period, the dual challenge of restricted panel length and a high dimensional features space remains. In the literature review, we discuss how these data challenges make it difficult to model price changes using traditional economic theory and then provide our methodological resolution in sections 4.4 and 4.5.

4.2.4. 3P Seller Aggregation and Profiling

We use the K-means clustering algorithm (MacQueen 1967, Pollard 1981) to form homogeneous seller clusters. We use observed individual seller level characteristics and compute summary statistics per seller across the panel. We provide a description of the seller characteristics used in the clustering algorithm in Table 4.4 (see Appendix 4.2 for details).

Table 4.4. Final Clustering Variables

Seller Characteristics		Description
Total No. of Ratings	% Positive	Mean value for the seller across the panel
Focal-Brand Indicator		Mean value for the seller across the panel
FBA		1 if seller offered Instant Pot once in the panel, 0 otherwise
Free ship Offers		Proportion of FBA offers among the selected SKUs made by the seller across the panel
Sales Presence		Proportion of free ship offers among the selected SKUs made by the seller across the panel
		Number of days seller observed on the marketplace

We report the seller distribution by clusters and the mean cluster profiles in Table 4.5 and brand decomposition of seller clusters and Amazon across the tracking panel in Table 4.6.

Table 4.5. Seller Distribution and Mean Cluster Profiles

Clusters	No. of Sellers	Star Rating	No. of Ratings	% +ve Ratings	% FBA	% FS	No. Unique ASINs	Days on market*	No. of Price Δ s
1	17	4.66	44,197	92.6	2.5	94.1	1.35	52	2
2	30	4.26	2,669	82.1	0.0	1.8	1.13	30	1
3	91	4.56	2,498	89.6	0.0	99.8	1.16	25	3
4	57	4.73	4,470	93.7	4.9	96.5	2.21	65	21
5	41	4.87	3,678	95.9	98.3	100.0	1.07	70	4

* Total number of days in tracking period = 222

No. of Ratings = Cumulative Ratings for the seller ever since it joined the marketplace, +ve

Ratings = Positive Customer Ratings, FS= Free Shipping conditions, Price Δ s = Price

changes

Cluster 3 is the largest cluster with 91 sellers with an average of only 2,498 ratings and an average presence on the marketplace for 25 days in the 222 day tracking period. This indicates that this cluster is composed of relatively new entrants to the marketplace. Cluster 1, on the other hand, is composed of only 17 sellers but an average seller of this cluster has 44,197 cumulative ratings. While they appear to be present for only part of the data collection time period, their high cumulative ratings indicate that they likely temporarily exited the category and re-entered during our observation period. (Ratings gather cumulatively over the entire period on Amazon, i.e. do not start fresh with each delist/ relist occasion). This indicates that cluster 1 has a more “established” group of sellers. Cluster 2 comprises 30 sellers and though

similar to cluster 3 has a low average of 2,669 ratings and present on average for 30 days, this set of sellers does not free ship (~ 1.8%). This suggests that cluster 2 sellers are probably “small-scale” retailers. Cluster 4 has 57 sellers who on average sell more than 1 unique item in the category. These sellers are thus likely to be “multi-brand” players. We find further evidence of this in Table 4.6 below, where 40.1% of offers from this seller cluster are made on Instant Pot. Since we track 7 SKUs for the focal brand Instant Pot and 1 SKU per non-focal brand in our data, it is likely that the sellers in cluster 4 offer one item of Instant Pot and the other of a non-focal brand. In Cluster 5 around 98.3% of sellers use Amazon’s order fulfilment services or FBA. We thus call this group as the “FBA sellers”.

Table 4.6 highlights two key points. First, not all brands are sold by all seller types. Second, of the brands offered a seller type may not sell these throughout the tracking period. For instance, slightly more than half of Amazon offers are on Instant Pot. This implies, Amazon does not always offer Instant Pot even when it is present in the electric cooker marketplace. An important implication is that this leads to considerable discontinuities in brand-seller type longitudinal panel data. This prevents the ability to track price changes at a seller type-brand level. We discuss the modeling implications and consequences of this limitation in section 4.5 and Appendix 4.4 respectively.

Table 4.6. Seller Type - Brand Panel Decomposition

Brands	Amazon	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
		Established Sellers	Small Sellers	New Entrants	Multi-Brand Sellers	FBA Sellers
Instant Pot	55.5%	1.4%	3.3%	0.3%	40.1%	1.3%
Geek Chef	0.0%	0.0%	3.3%	5.5%	1.0%	7.3%
Cosori	5.1%	0.0%	3.3%	2.3%	0.1%	4.9%
Oyama	0.0%	0.0%	0.0%	0.0%	1.8%	0.0%
Elechomes	0.0%	0.0%	0.0%	1.1%	0.0%	2.4%
Avalon Bay	0.0%	0.0%	0.0%	3.3%	0.0%	4.9%
Hamilton Beach	7.9%	14.3%	5.5%	12.6%	14.9%	0.0%
Cusimax	0.0%	0.0%	0.0%	0.0%	0.0%	2.4%
Pressure Pro	0.0%	0.0%	7.3%	8.4%	0.7%	2.4%
Crock-Pot	6.5%	14.5%	14.1%	27.6%	11.8%	4.9%
Magic Mill	0.0%	0.0%	16.7%	0.0%	0.9%	2.4%
Cuisinart	8.3%	37.3%	15.1%	16.0%	5.7%	8.2%
Breville	8.7%	23.5%	13.3%	7.7%	11.3%	54.0%
Elite Platinum	8.0%	9.0%	18.1%	15.2%	11.7%	4.9%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Each column entry denotes the proportion of times the seller offered a given brand in the tracking period. Each column adds up to 100%

4.3. Related Literature

Methodologically this research borrows from and builds on the random forest based methods for variable selection that I presented in Chapter 3. Substantively, this work is related to the literature on retail pricing for both online and offline, and strategic competition among firms. I discuss some of the streams of literature

pertaining to the empirical setting of the current study below.

The closest substantive area is the retail pricing literature studied in both online and offline contexts. In the offline context, studies have examined factors that affect retailer pricing strategies (e.g., Shankar and Bolton 2004, Nijs et al. 2007), effect of pricing formats such as EDLP or Hi-Lo on sales (e.g. Hoch et al. 1994), consumer and competitor response to changes in marketing-mix strategy (e.g. Ailawadi, Lehmann and Neslin 2001), effect of price discounting on sales (Kopalle et al. 1999). These studies examine retail pricing as a function of sales data, store type, market type, category and competition variables. These studies are directly useful for our research in providing a checklist of the covariate types that might influence price dynamics. An important distinction between this work and ours is that we do not observe sales data. The ecommerce marketplace also differs from offline retail with variables like seller and brand ratings, free shipping offers. Some variables like brand and seller characteristics are measured differently in the Marketplace than in offline retail stores.

This essay is also broadly related to work in online pricing. First, from the early days of online retailing, a question of interest has been price dispersion in online markets. The expectation was that as search costs for consumers and other frictions were lower in online markets, there would be lower price dispersion for similar items sold by different retailers. Even early studies (Bailey 1998, Brynjolfson and Smith 2000) showed that this “law of one price” (Varian 1980) does not hold. This can be because identical item are sold differently (e.g. with free shipping versus not). Price can also vary because there are market frictions like seller listings making

comparability of items harder (Bodoh-Creed et al. 2018), or strategic behavior by sellers causing varied pricing even for homogenous products. Bodoh-Creed et al. (2018) build a theory model of competition where there is very low price dispersion as long as sellers are rational profit maximizers, patient and have low storage costs. However, as they show several of these assumptions are likely not valid. Their theory model has important implications for the interpretability of our work. We find differing price levels and price changes for similar items. This could be because both Amazon and 3P sellers deviate from assumptions of profit maximization, high patience and low storage costs. The business press has discussed Amazon's interest in market share rather than profits. Some 3P sellers are also alleged to have reserve prices, or lowest prices at which they will sell, based on their own purchase price of products. Some smaller 3P sellers who do not use FBA service might have higher storage costs than Amazon with its large warehouses. Therefore, even within the Amazon Marketplace where price comparisons might be easier for consumers, we can expect deviations from the "law of one price".

Another stream of literature relevant to this current study explores the reasons for price stickiness- or differences in prevalence of price changes by retailers on Amazon. While the price stickiness literature is a large and old stream of literature (see Mankiw 1985), we discuss here the most directly relevant papers to the Amazon marketplace setting. Menu costs or the costs of changing prices (e.g. Dutta et. al. 1999) as a source of price stickiness are not relevant to our context. The more plausible reasons for price stickiness in our context are managerial inertia or managerial inattention (Goldfarb and Xiao 2011, Ellison, Snyder and Zhang 2016).

Based on business press articles, it appears more plausible that smaller 3P sellers on Amazon might have managerial inertia or managerial inattention.

Another important and related substantive area is the strategic competition among firms examined in the empirical industrial organization literature. This includes both static and dynamic structural models of competition. Since this literature is vast, I highlight a small number of directly relevant studies. These include strategic decisions of entry and exit (e.g. Aguirregabiria and Mira 2007), pricing (Slade 1992, Ellickson and Misra 2008, 2012), and product assortment strategies (Draganska et al. 2009). These structural models elegantly capture firm behavior based on the assumption of profit maximization by firms. The covariates such as local market and firm characteristics in these modeling specifications are determined on the basis of economic theory.

The structural models of firm competition are not suitable for our research context for the following reasons. First, Amazon's objective function in any one category is hard to characterize. Business press articles suggest that in all likelihood the company aims to win long-term market share over profits. Second, business press reports that Amazon uses algorithmic pricing across multiple categories, and often runs field experiments (A/B testing). It is unclear what structural model of pricing can capture such behavior. This problem is compounded by the large number of potential covariates available to us as researchers, and no research to guide us about which of these covariates might be important to capture pricing dynamics. Third, Amazon has far more information than 3P sellers on demand for its products and those of 3P sellers. To the best of my knowledge, there are no structural models of such

asymmetric competitive information. Note we also do not have any data on sales. This makes it impossible to estimate any model of competitive price dynamics. For these reasons, we do not model our price competition using structural industrial organization methods.

The methodological focus of this work builds on the variable selection method examined in Chapter 3. In this chapter, I briefly discuss the literature related to extant variable selection methods. Table 4.7 summarizes the extant approaches for variable selection. In the filter techniques, the variables are ranked based on their discriminative power. The variable selection happens as a pre-processing step independent of the prediction problem at hand (Lazar et al. 2012, Gregorutti et al. 2016). In the wrapper methods, the variables are selected during the model training and therefore are chosen based on their predictive power. The wrapper methods heuristically introduce and eliminate variables through forward or backward algorithms (e.g. Guyon et al. 2002, Svetnik et al. 2004, Díaz Uriarte and Alvarez de Andrés 2006). A drawback of the wrapper methods is that these are often computationally very demanding (Lazar et al. 2012). Similar to the wrapper methods, the embedded methods such as LASSO (Tibshirani 1996) for regression problems and decision trees (Breiman et al. 1984) select variables during the learning process. However, unlike the wrapper methods, the embedded methods have much smaller computation time. A common problem in these three variable selection methods is that of instability. Any change in the training sample can change the set of selected variables. The ensemble methods are better equipped to deal with the instability issues of the classical variable selection methods (Abeel et al. 2009, Meinshausen and

Bühlmann 2010, Haury et al. 2011). Instead of using one training sample the ensemble technique trains on several bootstrap samples or subsamples of the training data. The variables are then selected based on the average ranking across the ensemble.

Table 4.7. Variable Selection Methods and Relevant Literature

Methods	Related Papers
Filter	Weston et al. 2003, Torkolla 2003
Wrapper	Kohvani & John 1997, Guyon et al. 2002, Svetnik et al. 2004
Embedded	Tibshirani 1996, Breiman et al. 1984
Ensemble	Abeel et al. 2009, Meinshausen and Bühlmann 2010

Tree-based ensemble methods can be broadly classified into “bagging” and “boosting” types. The “bagging” type methods such as random forests combine trees that are obtained from identical randomized processes while the “boosting” type methods grow the trees sequentially, with one tree depending on the output of the prior tree (Ghosal and Hooker 2018). For the multivariate extensions of tree-based ensemble methods gradient boosted regression trees (Miller et al. 2016) and one-step boosted forests (Ghosal and Hooker 2018) have been examined in the literature. The relative advantage of “bagging” type methods such as random forests is that these do not require additional tuning and shrinkage parameters that are needed for boosted regression tree techniques.

In statistics and machine learning literature, random forests are a popular ensemble technique using tree-based algorithms. Random forests have been widely used as a tool for prediction problems involving higher order interactions and non-linear effects between covariates and the outcome variable. However, there has also

been an equal focus on using random forests for variable selection and dimensionality reduction (Ishwaran 2007, Strobl et al. 2007). The variable selection algorithms in these methods employ variable importance measure (VIM) for feature selection with either the non-recursive feature elimination (NRFE) or recursive feature elimination (RFE) strategies. The NRFE strategy eliminates variables based on a static VIM ranking computed at the start of the algorithm. This strategy uses an iterative algorithm and eliminates the lowest ranked variables after every iteration in the ensemble training phase (e.g. Svetnik et al. 2004). The RFE strategy uses a similar algorithm of iterative elimination of lowest ranked variables (e.g. Guyon et al. 2002). However, in the RFE strategy the VIM scores or ranks of the variables are recomputed at every iterative training of the ensemble. The applications have been to sift covariates when the true nature of association between covariates and outcomes are not known (Strobl et al. 2007, Ishwaran 2007). In clinical case control studies it has been shown that random forests are much better suited than traditional statistical methods such as logistic regression for prediction and variable selection in “small n large p ” problems (Strobl et al. 2007).

Several VIMs have been proposed in the literature, and some of which are available in canned statistical packages. A naïve VIM is to count the average number of times or the frequency a covariate is used in building a tree in an ensemble. Another naïve measure is the average incidence of a covariate’s use in an ensemble of trees. Breiman (2001) proposed the “permutation accuracy” method of variable importance. In this method, the forests are first constructed based on true values of the variables. To test the predictive ability of a variable and score its importance, it is first permuted

randomly and these permuted values are used to make predictions for the observed outcomes. The difference in prediction accuracy between the forest constructed with the true values and that with the permuted values of the variable is used to measure its importance. Another VIM in univariate forests is to measure the split improvement (SI) made by the variable at the splitting node. In regression trees, the split improvement is measured in terms of mean squared error and in classification problems this is measured in terms of Gini index (Ishwaran 2007).

I apply two of the proposed SI based VIM (mean structure based SI with F test and outcome difference SI) as introduced in Chapter 3 using an MVRF (Xiao and Segal 2009, Segal and Xiao 2011) based recursive feature elimination strategy to identify predictors for a multivariate price change model. I jointly model price change response of Amazon and 3P sellers as a non-linear function of predictors from a large feature space.

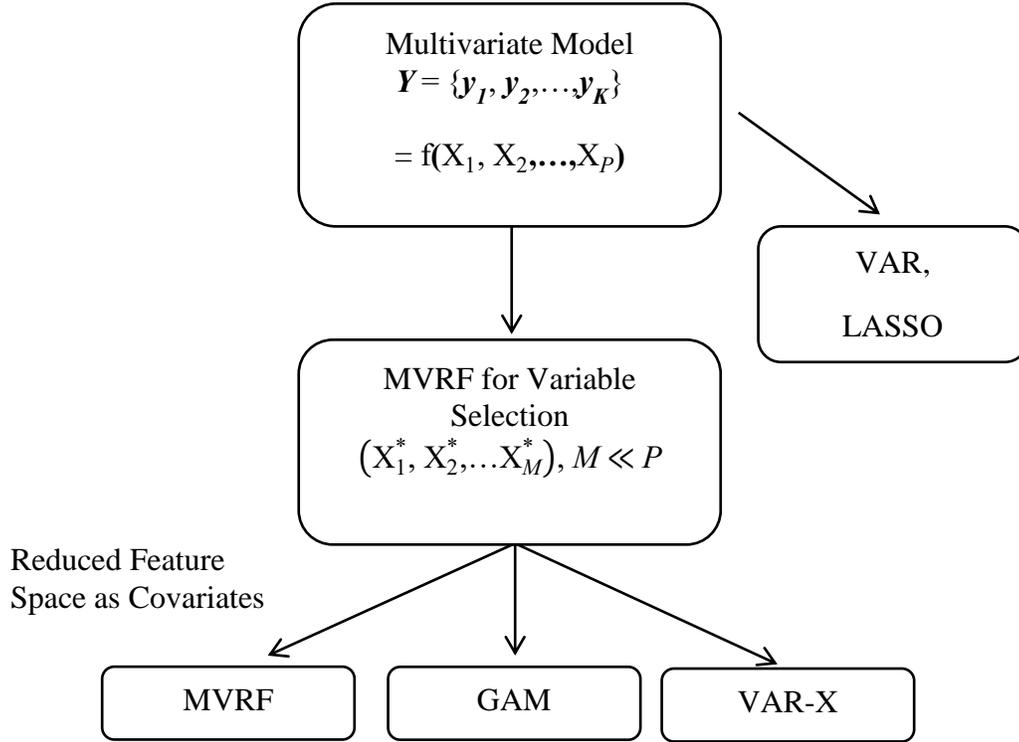
Methodologically, this research is also related to time series techniques explored in marketing. In marketing, time series models have been used to examine sales-advertising dynamics, price change and price setting (e.g. DeSarbo et al. 1987, Dekimpe and Hanssens 2000). Previous marketing research with time series techniques have used methods such as Granger causality tests to determine leader-follower behavior in price settings and identify competitive reaction patterns (Roy et al. 1994, Hanssens 1980), multivariate models to uncover drivers of retailer pricing tactics (e.g. Nijs et al. 2007). Marketing scholars have studied VAR and VAR-X models for studying price promotions and dynamic interactions among prices across brands (e.g., Pesaran and Smith 1998, Srinivasan et al. 2004). In our paper, we apply

MVRF for variable selection to reduce the features space. We run a series of multivariate time-series parametric and non-parametric models such as VAR-X and GAM to compare predictive performance. We use VAR as a null model to provide a benchmark comparison for the variable selection method. The results from the best performing time-series model are used to provide an interpretation of the nature of association between the covariates and the multivariate outcome. This investigation enables us to identify the drivers of strategic interactions among Amazon and other third party sellers.

4.4. Modeling Framework

In this section, I discuss the construction of the multivariate price change outcome, the variable importance measures used for the variable selection and propose alternative multivariate regression models for modeling the outcome. The high-level schema for the theoretical framework is presented in Figure 4.2. The discussion in Chapter 3 on the recursive feature elimination strategy using the proposed SI based VIMs on MVRF forms the middle layer in the schema.

Figure 4.2. High Level Schema for the Modeling Framework



4.4.1. Multivariate Price Response Model

In our theoretical framework, we assume there are K seller types in a marketplace selling at least one of B brands each. We assume a discrete time framework with T time periods. We assume p_{bkt} to denote the price of brand b charged by seller type k at time t . The magnitude of price change response of the k^{th} seller type at time t for brand $b=1, 2, \dots, B$ is given by,

$$|y_{bkt}| = \left| \frac{p_{bkt} - p_{bk,t-1}}{p_{bk,t-1}} \right| \quad (1)$$

At any time period t , the k^{th} seller type can choose to either change its price on brand b or keep it unchanged. We define the price change response for the k^{th} seller type at time t as,

$$y_{kt} = y_{bkt}, \text{ where } |y_{bkt}| = \text{Max}(|y_{1kt}|, |y_{2kt}|, \dots, |y_{Bkt}|) \text{ for } b = 1, 2, \dots, B \quad (2)$$

We denote the vector $\mathbf{Y}_t = \{y_{1t}, y_{2t}, \dots, y_{Kt}\}$ as the vector of price responses of all K seller types at time t . The data across all T time periods gives the multivariate price response matrix,

$\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_K\}$, where, \mathbf{y}_k is the $(T \times 1)$ vector of price change responses across the data period for the k^{th} seller type.

From the discussion in section 4.2.4 (Table 4.6), we note that the seller types (whether Amazon or 3P seller clusters) do not offer the same brand throughout the tracking period. This justifies the brand level aggregation in equation (6). However, this also implies that the price change response of a seller type correspond to different brands over time. From equation (6), we note that the price change outcome can be modeled in two ways. We could model simply the magnitude of price change as the response outcome, i.e., $y_{kt} = |y_{bkt}|$ or model it as price change with sign, i.e., $y_{kt} = y_{bkt}$. Since the brand may vary over time, the model for price change with sign is not meaningful. However, what matters is the price change reaction (i.e., propensity of price change and magnitude) of a seller type. In our empirical application we model the magnitude of price change¹².

¹² I also model for the price change with sign included for robustness test of the method. Refer to Appendix 4.4 for estimation results and discussion.

4.4.2. Variable Selection Using MVRF

In Chapter 3, I have introduced measures of variable importance for MVRFs based on the split improvement criterion. The first is a mean structure based SI importance derived from the mean based split function (Segal 1992) for multivariate regression trees (equation (6) in Chapter 3). In this measure, the split improvement is defined as the difference between the sum of squares at the parent node and the sum of squares at the left and right children nodes. The sums of squares calculation at both the parent and children nodes can incorporate the covariance structure between the multivariate response outcomes. We compute the mean structure SI importance in two ways. First, we account for all splits made by a variable, cumulated the SIs in a tree and took the mean across the ensemble. Second, we account for the SI of only the significant splits (determined using an F-test).

The second SI based measure introduced in Chapter 3 is the outcome difference based split improvement (equation (7) in Chapter 3). Here we first compute the sample mean of each outcome variable for both left and right children nodes and take the absolute difference as a measure of node heterogeneity or split improvement. The outcome difference SI VIM exploits the property of “between node” separation so that higher magnitudes of separation or heterogeneity between left and right nodes result in larger values of the computed variable importance. Similar to the mean structure SI, we account for the outcome difference SI for all splits and only for the significant splits (with F test).

In the empirical application using e-bird data, we found that the proposed VIMs especially the mean structure SI with F test and the outcome difference across

all splits perform the best against the naïve frequency and incidence based importance measures. As an additional validation, we repeat the exercise with Amazon data to test the predictive performance using the RFE strategy discussed in Chapter 3 with the proposed and the naïve importance measures. Specifically we apply the RFE strategy using the best performing VIMs to model the multivariate price response outcome as given in equation (2) to select P variables from the overall set of M variables.

4.4.3. Multivariate Regression Models

Post variable selection, I examine the relationship between the selected covariates and the multivariate price response outcome with a series of alternative regression specifications. The specifications differ in the covariates used as inputs to the regression models.

The first specification is the time-series extension of the generalized additive model or GAM (Friedman, Hastie and Tibshirani 2001). The seller type k 's price response at time t using a semi-parametric specification is given by

$$y_{kt} = \alpha_{k,gam} + \boldsymbol{\beta}_{\mathbf{k},gam} \mathbf{I}_t + \gamma_{k,gam} Season_t + \sum_m f(X_m) \quad (3)$$

where, $\alpha_{k,gam}$ is the seller type k specific intercept for the GAM specification. The reduced features space from the MVRF variable selection method is split into two groups of covariates – indicators and continuous covariates. The indicator variables denoting the market dynamics or actions of all seller types in the current and lagged periods is included in the matrix \mathbf{I} and is introduced as a linear additive specification. The continuous covariates from the reduced features set are specified through a non-parametric smooth function denoted by $f(X_m)$. The seasonality in price response is

accounted for by the indicator $Season_t$. The covariates are explained in section 4.5.2.

The other benchmark specifications are the multivariate time-series regression models, VAR and VAR-X (Pesaran and Smith 1998, Srinivasan et al. 2004). The VAR model is specified as follows,

$$Y_t = \alpha_{\text{var}} + \delta_{1,\text{var}} Y_{t-1} + \delta_{2,\text{var}} Y_{t-2} + \dots + \delta_{r,\text{var}} Y_{t-r} \quad (4)$$

where, α_{var} is the $(K \times 1)$ vector of seller type specific intercepts for the VAR model. $\{Y_{t-1}, Y_{t-2}, \dots, Y_{t-r}\}$ represent the r -order autoregressive vectors with the corresponding coefficient vectors denoted by $\{\delta_{1,\text{var}}, \delta_{2,\text{var}}, \dots, \delta_{r,\text{var}}\}$.

The VAR-X specification is given as follows,

$$Y_t = \alpha_{\text{varx}} + \delta_{1,\text{varx}} Y_{t-1} + \delta_{2,\text{varx}} Y_{t-2} + \dots + \delta_{r,\text{varx}} Y_{t-r} + \beta_{k,\text{varx}} \mathbf{I}_t + \gamma_{k,\text{varx}} Season_t + \sum_m \vartheta_m X_m \quad (5)$$

The specification here incorporates the intercept and the r -order autoregressive terms similar to the VAR model. All indicators enter in a linear additive specification similar to GAM. However, unlike in the GAM specification, the continuous exogenous covariates X_m from the reduced features space enter as a linear additive specification.

Similar to the linear VAR-X specification, I include a variable selection benchmark model of LASSO (Tibshirani 1996). As discussed in section 4.3, the LASSO is an embedded variable selection technique where the variables are selected during model training. Similar to the OLS estimation, the LASSO objective is to minimize the residual sum of squares. However, LASSO is a constrained minimization

problem where the sum of the absolute value of the coefficients is set less than a constant. This constraint results in some coefficients being shrunk exactly to 0 and hence gives a reduced set of covariates as a model outcome.

In addition to the multivariate regression models, I run an MVRF on the reduced features space post variable selection as an additional benchmark to compare against the regression models.

4.5. Empirical Application

4.5.1. Price Change Outcome for 3P Seller Clusters

From the theoretical framework in section 4.4, we assume that there are $(K-1)$ 3P seller types (K^{th} being Amazon as a standalone seller type). From the K-means clustering algorithm discussed in section 4.2.4, each seller cluster represents a type of 3P seller. The unit of time is day and the data is tracked for T days. The k^{th} third-party seller type or equivalently cluster's representative price for brand b on day t is assumed to be the minimum price offered in the cluster.

If there are n_k sellers in cluster k , this implies,

$$p_{bkt} = \text{Min} \left(p_{bkt}^1, p_{bkt}^2, \dots, p_{bkt}^{n_k} \right) \quad (6)$$

The minimum price seller for a brand forms a good cluster representative in this specific marketplace context. This is due to the fact that it is easy for customers to search for low price and to make comparisons of terms in an online environment. Therefore, the minimum price is likely to influence demand. The minimum price offered by a homogeneous group of sellers is likely to be closely monitored by rival sellers, and is thus a reasonable statistic to also capture competitive factors in pricing.

Further, we note that equation (2) leads to a brand-level aggregation. Therefore, the price change outcome that is modeled for a cluster corresponds to the brand with the maximum absolute price change.

4.5.2. Defining the Covariates

The 3P seller clustering leads to two sources of variation in the seller characteristics – within cluster variation and across time variation. To account for both sources of variations, we consider the minimum, maximum and mean of each seller characteristic within a cluster and over time. The 3P sellers do not have information on overall sales of a product at the Amazon marketplace. Further, both Amazon and 3P sellers do not observe the cost or demand information of their competitor 3P sellers. The observable seller characteristics such as star ratings, cumulative number of ratings, shipping and delivery options etc. are the sources of information to monitor how their competitors are performing in the marketplace. Any changes in these performance metrics or shipping/delivery offers may trigger changes in consumer buy behavior and are thus likely to affect price changes among these sellers. We thus consider 1-period lagged values for the summary measures of these seller characteristics for each cluster. We also note that since some of the 3P sellers offer used items in the tracking period, we capture the summary measures of the characteristics of these sellers as an additional set of covariates.

To account for time variation of brand characteristics such as star rating, number of positive ratings etc., we compute the minimum, maximum and mean values to date for characteristics of each brand. Similar to seller characteristics, we consider 1-period lagged values of the summary measures of each brand characteristic.

We include lagged values of the multivariate price response outcomes, i.e., the price change responses made by the 3P seller clusters or Amazon in the past periods. We also include brand dummy indicators to capture the brands with the said price changes. Additionally, to account for the effects of price changes made on used items on the prices of new items, we also include lagged values of price changes made on used items. We group these covariates as lagged price changes in focal-brand used items and those in non-focal brand used items. The vector of brand indicators of past price changes by seller clusters and Amazon make up \mathbf{I}_t in equations (3) and (5). For the variable selection method, in our initial set of covariates we consider up to 3 period lagged values of past actions of players.

Finally, we include controls for seasonality. In Table 4.8, we provide details of the covariates used as initial inputs in the variable selection method. After incorporating all the time varying distribution by clusters and brands, and past actions of sellers, we have a total of 892 covariates excluding seasonality with a panel of 222 data points. This highlights the high dimensionality problem of “small n , large p ” and additionally the non-linear nature of the relationship among predictors and the multivariate outcome model.

While our list of covariates is extensive, we do not have data on several variables that likely influence pricing. These include other online and offline rivals, prices of accessories, prices of other categories where these sellers sell (e.g. if sellers profit maximizing across multiple categories), etc. Due to the complexity in data scraping across multiple pages on Amazon, we limit our analysis to the data collected for the brands within the chosen category and for the main product (electric cooker).

Table 4.8. Covariates Specification

Variables	Description
<i>Seller Cluster Characteristics</i>	
Star Rating	Maximum, Minimum and Mean till date per cluster at 1 period lagged values
% Positive Rating	
Number of Ratings	
% FBA Offers	
% Free ship Offers	
Number of unique items sold	
Number of days present**	
<i>Brand Characteristics</i>	
Star Rating	Maximum, Minimum and Mean till date per brand at 1 period lagged values
Number of Answered Questions	
Number of Customer Reviews	
Buy Box Price	
Sales Rank in own category	
<i>Market Dynamics</i>	
Price response by Amazon	
Price response by 3P seller clusters	Lagged values up to 3 periods
Brand Indicators of Price response by Amazon	Lagged values up to 3 periods
Brand Indicators of Price response by 3P seller clusters	
<i>Seasonality</i>	Indicator (1) for each day of the seasonal period November 2017- mid-January 2018

**Number of days seller is cumulatively present on Amazon marketplace

4.5.3. Variable Selection and Estimation Strategy

As shown in Figure 4.2, the estimation proceeds in two steps. The first step runs the RFE strategy on the MVRFs introduced in Chapter 3. The validation check to determine the best performing importance measures in the empirical application for Amazon is presented in Table 4.9. To select the best performing importance measures, we run the RFE strategy for 30 iterations with an ensemble of 500 trees built in each iteration. We compute the mean squared error or MSE as the average of the squared

difference in the true price change and predicted price change on the test set at the end of each iteration. We report the MSE computed at the end of the 30th iteration of the RFE strategy in Table 4.9. Since the price change is calculated as a percentage, the MSE is represented as such.

Table 4.9. MSE at 30th Iteration of the RFE strategy on Amazon data

Importance Measures	<i>Amazon</i>	<i>Seller Cluster 1</i>	<i>Seller Cluster 2</i>	<i>Seller Cluster 3</i>	<i>Seller Cluster 4</i>	<i>Seller Cluster 5</i>
Frequency	0.14%	0.27%	7.15%	46.35%	7.53%	0.24%
Incidence	0.14%	0.28%	9.15%	39.02%	4.70%	0.23%
Mean Structure	0.39%	0.28%	7.90%	4.97%	8.34%	0.92%
Mean Structure with F test	0.42%	0.29%	8.81%	4.96%	7.08%	0.89%
Outcome Difference	0.18%	0.27%	7.22%	9.89%	5.39%	0.36%
Outcome Difference with F test	0.25%	0.30%	7.04%	11.10%	5.94%	0.47%

While the frequency and incidence measures seem to be performing the best in predicting Amazon’s price change decisions, the MSEs for these measures are 10 times higher than those of the proposed measures for seller cluster 3 or “new entrants”. Recall from Table 4.5 that 3P seller cluster 3 is the largest cluster with 91 sellers (38% of 3P sellers in our data). Further this cluster on average makes 3 price changes in the data period. Therefore cluster 3 is an important representation of the 3P seller clusters and predicting its price changes accurately is all the more relevant. The mean structure SI when computed on the significant splits (with F test) and the outcome difference SI when computed across all splits perform better than the naïve measures for cluster 3, and show similar performance (< 1% difference in MSE in most cases) as the naïve measures for Amazon and clusters 1, 2 and 5. For cluster 4 or “multi-

brand sellers”, we find that the mean structure SI with F test and the outcome difference SI outperform the frequency importance, but slightly worse than the incidence importance. At an overall level, the mean structure SI with F test and the outcome difference SI measures seem to be most consistent in the predictive performance across outcomes. We thus use these two measures for variable selection for the Amazon data.

For brevity, the results from the variable selection algorithm using the mean structure SI (with F test) and the outcome difference SI are provided in Appendix 4.3. We provide graphs for the iterative performance of MVRF for the variable selection algorithms run using the two proposed VIMs. We have an initial set of 892 covariates. After 40 iterations of the RFE strategy, the mean structure based SI VIM (with F test) used for variable selection reduces the covariates set to 246, while the outcome difference SI VIM reduces it to 58. Since the construction of the importance measures vary by definition, the variables selected by the two VIMs will differ. Furthermore, the scores for the variables are outcome specific for the outcome difference SI VIM, while the scores are homogeneous across outcomes for the mean based SI VIM. That is to say, the variables selected as important by the outcome difference SI VIM will have a different score for Amazon and each of the 3P seller clusters.

There are four key points to note in the estimation strategy following the variable selection algorithm. First, the MVRF serves as a benchmark predictive model itself. I run the MVRF using all the covariates from the reduced set and record the out of sample performance in terms of mean squared error (MSE).

Second, the variable selection algorithm may select collinear predictors in the

reduced covariates set. However, as the algorithm works as a pre-processing step (rather than an embedded variable selection technique such as LASSO), we can perform multicollinearity checks post variable reduction. In our selection among collinear predictors for the regression models (GAM and VAR-X), we select the predictor with highest score as given by the respective VIM.

Third, for the VAR-X and GAM, I sequentially introduce the covariates from the reduced covariates list according to the variable importance scores. I record the out-of-sample predictive performance of the regression models for each new variable introduction and determine the cut-off point for variable inclusion when the performance begins to deteriorate.

Finally, for the VAR-X model, I include only the covariates related to brand and seller characteristics and indicators of past seller actions on brands and do not include any autoregressive variables, i.e., the lagged price response outcomes of sellers and Amazon that are identified as important by the variable selection scheme. Rather, I allow the VAR-X function to decide the order of lags for the endogenous variables. For the GAM, I include lagged endogenous variables, the indicators and the purely exogenous brand and seller characteristics that are ranked as important by our variable selection scheme. Since the GAM specification has the flexibility of introducing covariates both as linear parametric and non-linear semi-parametric functions, the covariates specifications can differ by outcome, i.e., whether Amazon or 3P seller cluster. The choice between parametric or semi-parametric specification is based on the nature of the covariate. For instance, all indicator covariates are introduced as linear terms and all continuous covariates with sufficient variations are

included as smooth functions. For continuous covariates that do not have enough variations over time, we model these using a linear specification.

The benchmark models to the proposed variable selection method are the VAR model with no exogenous covariates and the LASSO. For VAR we again allow the function to decide the order of autoregressive lags. The LASSO specification includes the autoregressive terms upto three-period lags and the full covariates space.

4.6. Results and Discussion

4.6.1. Model Performance

The model performance results on the test set are presented in Table 4.10. I report the mean squared error rates (or MSE) and standard deviation of squared errors (in parentheses).

The VAR and LASSO give comparable predictive performance with LASSO doing worse for Amazon and cluster 4 (“multi-brand” 3P sellers). The MVRF and VAR-X show varying degrees of performance relative to both these benchmark models. For instance, MVRF and VAR-X predict Amazon’s response better when using covariates from either variable selection method. However, for 3P seller clusters specifically clusters 4 and 5 (“FBA sellers”), MVRF’s predictive performance is weaker than that of VAR and LASSO. On the other hand, VAR-X does better than LASSO and VAR across all seller clusters and Amazon using the mean based split improvement variable selection method. However, the best predictive performance is seen in the GAM when run using covariates from either selection method. Particularly, GAM’s predictive performance using variables selected by the outcome difference split improvement method shows a significant improvement over the VAR and

LASSO across all seller types. For instance, GAM gives an MSE of 0.18% against 0.39% for VAR and 0.44% for LASSO in Amazon’s prediction. Similarly, for Cluster 3 (“new entrants”), the MSEs are 4.29% for VAR, 4.13% for LASSO and 1.77% for GAM with variables selected using the outcome difference method.

Table 4.10. Mean Squared Error Rate on Test Set for Price Change Magnitude

	Amazon	3P Seller Clusters				
		1 Estd. sellers	2 Small sellers	3 New Entrants	4 Multi- Brand	5 FBA Sellers
VAR	0.39% (0.3%)	0.29% (1.1%)	7.81% (49.0%)	4.29% (16.0%)	3.74% (10.0%)	0.40% (1.0%)
LASSO	0.44% (0.2%)	0.27% (1.0%)	7.49% (50.0%)	4.13% (15.0%)	4.91% (7.0%)	0.34% (0.7%)
<i>Multivariate Models with covariates selected from proposed variable selection algorithm</i>						
MVRF						
Mean Structure SI	0.37% (0.2%)	0.29% (1.0%)	8.48% (46.0%)	4.99% (13.0%)	7.64% (5.0%)	0.88% (0.5%)
Outcome Difference SI	0.25% (0.3%)	0.27% (1.0%)	7.81% (51.0%)	3.94% (11.0%)	6.15% (6.0%)	1.30% (0.8%)
VAR-X						
Mean Structure SI	0.35% (0.3%)	0.26% (1.0%)	7.22% (50.0%)	4.19% (15.0%)	3.33% (9.8%)	0.29% (1.0%)
Outcome Difference SI	0.28% (0.3%)	0.29% (1.0%)	7.69% (50.0%)	3.96% (10.4%)	5.18% (14.0%)	1.14% (4.0%)
GAM						
Mean Structure SI	0.37% (0.3%)	0.27% (1.0%)	7.22% (51.0%)	3.65% (15.0%)	3.41% (10.0%)	0.23% (1.0%)
Outcome Difference SI	0.18% (0.4%)	0.22% (1.0%)	7.22% (51.0%)	1.77% (5.0%)	3.63% (11.0%)	0.23% (1.0%)

Standard deviation of squared error on test set reported in parentheses

The standard deviation measures of the squared errors are comparable across all models and are fairly tight for Amazon, cluster 1 (“established” 3P sellers) and cluster 5. The largest deviations are seen in the prediction errors of cluster 2 (“small-scale” 3P sellers). This is intuitive since as seen in Table 4.5 the average number of price change in cluster 2 is only one. This implies the price change data for cluster 2 does not have sufficient variability for modeling and robust predictions.

4.6.2. Determinants of Price Change

I discuss the estimation results from the best predictive model, i.e., the GAM when the outcome difference SI VIM is used for variable selection. The estimation results are presented in Tables 4.11 through 4.16. For brevity I will focus the discussion only on the significant covariates for each outcome of price change.

For Amazon’s price response (Table 4.11), among the covariates with linear parametric specification, the past period price change in used items of Instant Pot made by 3P sellers is significant and negatively related to Amazon’s current period price change. This implies the larger the magnitude of price change in used items of Instant Pot is the smaller is the magnitude of Amazon’s price change.

Table 4.11. GAM Estimation Results for Amazon

Covariates with Parametric Specification	Mean	SD	Signif.

<i>Intercept</i>	-2.62	0.18	
<i>Price Dynamics</i>			
3 pd. lagged Cluster 4 Price change	0.02	0.43	
1 pd. lagged Cluster 5 price change indicator for Magic Mill	-0.42	1.39	
1 pd. lagged Cluster 5 Price change	0.39	0.79	*
1 pd. lagged 3P used Instant Pot Sellers' price change	-0.76	0.38	
<i>Brand Characteristics</i>			
1 pd. lagged maximum till date star rating of Pressure Pro	-0.08	0.06	
<i>Seasonal Factors</i>			
Seasonality	0.15	0.23	
	p-value	Significance	
Approximate significance of smooth terms			
<i>Seller Characteristics</i>			
1 pd. lagged maximum no. of seller ratings of used non-focal brand 3P sellers	0.18		
1 pd. lagged mean % positive seller ratings of Cluster 2	0.00	***	
<i>Brand Characteristics</i>			
1 pd. lagged maximum Buy box price of Elite Platinum	0.31		

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Some seller and brand characteristics enter the Amazon model as non-linear smooth specifications, of which the mean percentage positive seller rating of cluster 2 or “small-scale” 3P retailers is highly significant (see Appendix Figure A.4.5 for functional form). The functional form suggests an inverted U-shaped relationship between mean percent seller rating (in the range of 40% - 100%) of small-scale sellers and Amazon’s price change magnitude. Between 40-80%, any increase in mean percent rating of small-scale sellers leads to increase in Amazon’s price change

magnitude. Beyond the threshold, the mean percent seller-rating is negatively related to Amazon's price change magnitude. In other words, upto a certain threshold as reputation of the small-scale sellers improves Amazon is at a risk of losing its customers and thus responds by increasing magnitude of price change. The intercept term is highly significant and negatively related to Amazon's price change magnitude. This indicates that Amazon's price changes are negatively affected by factors outside the focal category. This is unsurprising since Amazon sells in multiple categories, and is said to employ sophisticated maximands across categories.

For cluster 1 or "established" group of 3P sellers, see Table 4.12. The cluster 1's own past period price change indicator on Hamilton Beach is significant and positively related to its current period price change magnitude. Similar to Amazon, the intercept for the cluster 1 model is highly significant. The GAM results provide empirical evidence that the established sellers on Amazon marketplace do not change prices based on other sellers' price change actions. Rather these sellers depend on own past price changes. Furthermore similar to Amazon, the intercept term is significant and negatively related to the price change magnitude of these sellers. This implies changes in factors outside the category may reduce their propensity to change prices in the focal category.

Table 4.12. GAM Estimation Results for Cluster 1 (“Established”)

Covariates with Parametric Specification	Mean	SD	Signif.
<i>Intercept</i>	-4.58	0	***
<i>Price Dynamics</i>			
1 pd. lagged Cluster 1 price change indicator for Hamilton Beach	1.78	7	***
<i>Seasonal Factors</i>			
Seasonality	-0.29	2	

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

The GAM results for cluster 2 of “small-scale” 3P sellers (see Table 4.13) show no significant covariate other than the intercept. Though this finding may initially cast doubt on the proposed variable selection method, however upon closer inspection this result is actually fairly intuitive. Based on the profile characteristics, this cluster comprises sellers who have been present on average of only 30 days in the marketplace and make only 1 price change on average in the 222 day tracking period (see Table 4.5). Therefore it is not surprising that the model coefficients are all insignificant. Furthermore, the intercept is significant and negatively related to price change magnitude of this seller group. We recall that this seller group has limited presence in the Amazon marketplace (as shown by the cumulative seller ratings in Table 4.5). There may be factors outside of the Amazon marketplace such as own cost structure or reserve prices that affect their price change decisions.

Table 4.13. GAM Estimation Results for Cluster 2 (“Small-scale”)

Covariates with Parametric Specification	Mean	SD	Signif.
<i>Intercept</i>	-3.74	2	**
<i>Price Dynamics</i>			
1 pd. lagged Cluster 5 price change indicator for Magic Mill	-0.06	8	
1 pd. lagged Cluster 3 price change indicator for Elite Platinum	-1.69	4	
2 pd. lagged Amazon price change	-4.18	2	
<i>Seasonal Factors</i>			
Seasonality	1.94	4	

Significance codes: 0 ‘****’ 0.001 ‘***’ 0.01 ‘**’ 0.05 ‘.’ 0.1

For cluster 3 or “new entrant” 3P sellers, see Table 4.14. The variables on price change enter the model as linear parametric specification and we find that the cluster’s own past period price change indicator on brand Elite Platinum is highly significant and positively related (similar to cluster 1 or established sellers’ past price change on Hamilton Beach). Among other players’ past period actions, cluster 4 or “multi-brand” 3P sellers’ 2-period lagged price change is positively related while Amazon’s 2-period lagged price change is negatively related to new entrants’ price change magnitude. This implies that though the new entrants change prices by taking cues of price changes made by Amazon and multi-brand 3P sellers, they are more likely to positively correlate the magnitude of change with those of the multi-brand sellers.

Table 4.14. GAM Estimation Results for Cluster 3 (“New Entrants”)

Covariates with Parametric Specification	Mean	SD	Signif
<i>Intercept</i>	-3.80	1.09	***
<i>Price Dynamics</i>			
1 pd. lagged Cluster 3 price change indicator for Elite Platinum	2.15	0.32	***
2 pd. lagged Cluster 4 price change	2.37	0.49	***
1 pd. lagged Cluster 5 price change	-2.31	2.97	
2 pd. lagged Amazon price change	-5.08	2.54	*
1 pd. lagged 3P used Instant Pot sellers’ price change	-0.67	0.39	.
<i>Seller Characteristics</i>			
1 pd. lagged max. no. of unique products sold till date by Cluster 4	0.14	0.13	
<i>Brand Characteristics</i>			
1 pd. lagged minimum till date sales rank of Instant Pot	0.09	0.12	
<i>Seasonal Factors</i>			
Seasonality	-0.73	0.66	
	p-value	Signif	.
<i>Approximate significance of smooth terms</i>			
<i>Brand Characteristics</i>			
1 pd. lagged mean Buy box price till date of Instant Pot	0.00	***	
1 pd. lagged maximum price till date of Cuisinart	0.00	***	

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Further, among the brand characteristics that enter the model as smooth functions (see Appendix Figure A.4.6 for plots), the past mean buy box price of Instant Pot and 1-period lagged maximum price of Cuisinart (one of the top non-focal brands) are significant. From the smooth plots we find that the past mean buy box price of Instant Pot has a U-shaped relationship with the price change magnitude of new entrants with a fairly flat base. In particular, for the Instant Pot mean buy box price ranges between \$110 and \$130, the new entrants appear to be making not much

change in its price. However, for prices below \$110 and above \$130, they respond by increasing their price change magnitude. Thus new entrant sellers, irrespective of the brand sold by them, monitor buy box price changes of the most popular brand in the category, Instant Pot. The maximum price till date of Cuisinart (in the range of \$80 - \$140) has an inverted U-shaped relationship with new entrants' price change magnitude. That is, upto a threshold with increase in price of Cuisinart the new entrants respond with increase in magnitude of price change. From Table 4.6 recall that the most frequently sold brands of the new entrants are Crock-Pot (27.6%), Cuisinart (16%), Elite Platinum (15.2%) and Hamilton Beach (12.6%). The relationship between the brand Cuisinart's maximum price till date and new entrants' price change behavior indicate possible brand competition between Cuisinart and brands such as Crock-Pot, Elite Platinum and Hamilton Beach. Similar to the small-scale sellers (see Table 4.5), this group is relatively less established in the Amazon marketplace. The negative and highly significant intercept term indicates that there are factors outside the marketplace such as cost structure that affects their price change.

For cluster 4 ("multibrand" seller cluster), the estimation results (Table 4.15) show that all of the covariates for past price changes are highly significant and positively correlated to its price change magnitude. The past period price change indicator of cluster 5 or "FBA 3P sellers" on brand Magic Mill and Amazon's 2 period lagged price change are significant predictors of price change magnitude of multibrand sellers. Most of these multi-brand sellers sell the focal brand Instant Pot and (at least) and one non-focal item. Given that Amazon sells mostly Instant Pot, it is thus not surprising Amazon's price changes can trigger similar response from this seller

cluster. Another interesting result is that the seasonality indicator is significant only for this cluster. This indicates that this cluster has a stronger intra-category presence and offers more promotions during seasonal periods than Amazon and established sellers with multicategory presence. The multibrand sellers also show a negative and highly significant intercept term. Since these sellers offer multiple brands, they are likely to get affected by the cross-category products of the brands they sell.

Table 4.15. GAM Estimation Results for Cluster 4 (“Multibrand”)

Covariates with Parametric Specification	Mean	SD	Signif.
<i>Intercept</i>	-2.76	0.54	***
<i>Price Dynamics</i>			
1 pd. lagged Cluster 5 price change indicator for Magic Mill	1.08	0.22	***
2 pd. lagged Amazon price change	3.17	0.84	***
<i>Seller Characteristics</i>			
1 pd. lagged max. no. of unique products sold till date by Cluster 4	0.08	0.11	
<i>Seasonal Factors</i>			
Seasonality	1.09	0.27	***

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

The GAM results for cluster 5 or “FBA sellers” (Table 4.15) show that among price change covariates the past period price change indicator of cluster 1 or established sellers on brand Hamilton Beach is significant and positively correlated. Among the covariates selected from brand characteristics, the past period mean buy box price of Instant Pot is highly significant and negatively related to price change magnitude of this cluster. That is, as the average buy box price of Instant Pot products increase, the FBA sellers are likely to decrease the price change magnitude on own products. Since Amazon stands to gain from sales through FBA sellers, these sellers

are more likely to be selected by Amazon for the buy-box. It is thus likely that when the buy box price of a major brand like Instant Pot increases, the FBA sellers (and potentially winners of buy box on the respective brands they sell) are likely to maintain prices in the brands. This explains why changes in the mean buy box price of Instant Pot can predict price changes in this cluster. The intercept for the FBA sellers is significant and positively related to their price change magnitude. Since these sellers are likely buy box winners they are likely to change prices based on cross-category factors such as buy box prices of brands in other related categories.

Table 4.16. GAM Estimation Results for Cluster 5 (“FBA Sellers”)

Covariates with Parametric Specification	Mean	SD	Signif.
<i>Intercept</i>	7.68	3.38	*
<i>Price Dynamics</i>			
1 pd. lagged Cluster 1 price change indicator for Hamilton Beach	0.88	0.36	*
2 pd. lagged Cluster 2 price change	-0.09	0.61	
<i>Brand Characteristics</i>			
1 pd. lagged mean Buy Box price of Instant Pot products	-0.10	0.03	***
<i>Seasonal Factors</i>			
Seasonality	1.26	1.27	

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

I summarize the key determinants (significant covariates) of price change in the electric cooker category by seller groups in Table 4.17. Here are the key points to note. First, both linear and non-linear significant effects are found in the best-fitting GAM model. This supports the use of a flexible non-parametric functional form in variable selection (MVRF). Second, I find that of the several variables chosen as being important, very few are statistically significant in the GAM specification.

However, the GAM specification gives the best overall predictive performance, beating even the MVRF. The coefficients of the significant parametric effects and the functional form of the significant non-linear effects enable us to interpret the type of association and functional relationship between outcome and covariates. Furthermore, the proposed RFE strategy selects the most important predictors through the learning algorithm of the MVRF. In the downstream application on the regression models, the significant covariates provide interpretability. Thus, this approach provides interpretability without sacrificing predictive ability.

Third, entrants and multibrand 3P sellers' price changes are triggered by other sellers' price change actions. In contrast, Amazon and established 3P seller pricing appears to be at the most very narrowly a function of other players' actions. This might be a result of Amazon and 3P sellers focusing more on their cross-category maximands. It might also be a result of them having their price strategies being driven by their costs structures and demand factors rather than by competitor actions. It is less likely that either Amazon or larger 3P sellers suffer from managerial inattention or managerial inertia in responding to competition in the category. An important piece of model-free evidence to note is that small-scale sellers are stickier in their prices. Additionally, these sellers have limited presence in the Amazon marketplace. However, the significant and negative intercept indicates that there are factors outside of the Amazon marketplace that may reduce their ability to change prices. This is consistent with business press reports of them having reserve prices.

Table 4.17. Significant Determinants of Price Change Magnitude

Sellers	Labels	Past Actions	Brand Variables	Seller Variables
Amazon		Used Instant Pot 3P sellers		Mean % positive ratings of sellers in Cluster 2
Cluster 1	Established Sellers	Own	Price change on Hamilton Beach	
Cluster 2	Small Scale Players	Amazon, Clusters 3, 5		
Cluster 3	New Entrants	Amazon, Own, Cluster 4	Buy box price of Instant Pot, Maximum price till date of Cuisinart	
Cluster 4	Multi-Brand Sellers	Amazon and Cluster 5	Price change on Magic Mill, Hamilton Beach	
Cluster 5	FBA Sellers	Cluster 1	Buy box price of Instant Pot, Price change on Hamilton Beach	

4.7. Contributions, Future Research Directions and Conclusions

In this chapter, I examine price dynamics on Amazon marketplace using a machine learning approach. The multivariate objective function is to jointly model price changes by Amazon and 3P sellers. The substantive findings from this research provide 3P sellers and regulators key insights into the factors that drive price changes in the selected category. Especially, the data as scraped from the Amazon website are all information that any seller on the marketplace has about other sellers. Recall each seller has additional information on its own sales only, and Amazon has information on each seller's sales since it hosts the Marketplace and doesn't share sales data with sellers. Therefore, any 3P seller and manufacturer can use this methodology and this

data (of shared information across all sellers) to understand key price change drivers.

The findings reveal some seemingly unexpected results for predictors of price change decisions by Amazon and 3P sellers. For instance, Amazon's price change decisions are affected by seller reputation of small-scale sellers who do not offer free shipping and might consequently have lower posted price. These unexpected results showcase the importance of variable selection; with intuition or economic principles alone, such drivers of price dynamics would likely not been included in the model.

Due to the data limitations and complexity of the marketplace we make some aggregation assumptions. The seller clustering leads to a loss of individual seller level information on price changes. Further, this aggregation also reduces our ability to predict the direction of price change movement as effectively as observed in the model results. However, these data-driven limitations can be easily relaxed in a data context with different structure e.g. longer spells of seller presence and fewer sellers. Another limitation of this study is that we examine pricing on Amazon's marketplace, with no analysis of rival websites and offline retailers. This is a promising area for future research, especially as Walmart has been trying hard to improve its ecommerce offerings. In conclusion, in this study I examine a multivariate outcome model with "small n and large p ", and apply a RFE strategy using MVRF. The proposed variable importance measures as introduced in Chapter 3 and the variable selection method will be a useful addition to the random forest based machine learning tool kit of researchers across multiple fields.

REFERENCES

- Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y (2009) Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3), 392-398.
- Aguirregabiria V, Mira P (2007) Sequential estimation of dynamic discrete games. *Econometrica*, 75.1 : 1-53.
- Ailawadi KL, Lehmann DR, Neslin SA (2001) Market response to a major policy change in the marketing mix: Learning from Procter & Gamble's value pricing strategy. *J. Marketing*, 65(1), 44-61.
- Andonova S, Elisseeff A, Evgeniou T, Pontil M (July 2002) A simple algorithm for learning stable machines. In *ECAI*, (pp. 513-517).
- Axios (2018) <https://www.axios.com/policing-the-power-of-tech-giants-1513302767-acd31a83-d517-463f-8a97-a3ed645caa36.html>
- Bailey JP (1998) Electronic commerce: prices and consumer issues for three products: books, compact discs, and software. *Organisation for Economic Co-Operation and Development*, OCDE/GD (98), 4.
- Bajari P, Chernozhukov V, Hortaçsu A, Suzuki J (2018) The impact of big data on firm performance: An empirical investigation (No. w24334). *NBER*.
- Ball L, Mankiw NG (1992) Asymmetric price adjustment and economic fluctuations (No. w4089). *NBER*.
- Bodoh-Creed AA, Boehnke J, Hickman B (2018). Using Machine Learning to predict price dispersion. *Working Paper*.
- Breiman L (2001) Random forests. *Machine learning*, 45(1), 5-32.
- Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and Regression Trees. *Wadsworth Advanced Books and Software*
- Brynjolfsson E, Smith MD (2000) Frictionless commerce? A comparison of Internet and conventional retailers. *Management Sci.*, 46(4), 563-585.
- BusinessInsider (2018) <https://www.businessinsider.com/what-is-amazons-choice-2018-5>
- Chen L, Mislove A, Wilson C (2016) An empirical analysis of algorithmic pricing on

- Amazon marketplace. *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee.
- Dekimpe MG, Hanssens DM (2000) Time-series models in marketing: Past, present and future. *International J. Res. Marketing*, 17(2-3), 183-193.
- DeSarbo WS, Rao VR, Steckel, JH, Wind J, Colombo R (1987). A friction model for describing and forecasting price changes. *Marketing Sci.*, 6(4), 299-319
- Díaz-Uriarte R, De Andres SA (2006) Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1), 3.
- Draganska M, Mazzeo M, Seim K (2009) Beyond plain vanilla: Modeling joint product assortment and pricing decisions. *QME*, 7.2: 105-146.
- Dutta S, Bergen M, Levy D, Venable R (1999) Menu costs, posted prices, and multiproduct retailers. *J. Money, Credit, Banking*, 683-703.
- Ellickson PB, Misra, S (2012) Enriching interactions: Incorporating outcome data into static discrete games. *QME*, 10.1: 1-26.
- Ellickson PB, Misra, S (2008) Supermarket pricing strategies. *Marketing Sci.*, 27.5: 811-828.
- Ellison SF, Snyder CM, Zhang H (December 2016) Costs of Managerial Attention and Activity as a Source of Sticky Prices: Structural Estimates from an Online Market. *CESifo Working Paper Series No. 6285*. Available at SSRN: <https://ssrn.com/abstract=2914249>
- Forbes (2017) <https://www.forbes.com/sites/wadeshepard/2017/01/02/amazon-scams-on-the-rise-in-2017-as-fraudulent-sellers-run-amok-and-profit-big/#5487b8e93ea6>
- Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, 1189-1232
- Friedman JH, Hastie T, Tibshirani R (2001) *The elements of statistical learning*. Vol. 1. New York: Springer series in statistics.
- Ghosal I., Hooker G (2018) Boosting Random Forests to Reduce Bias; One-Step Boosted Forest and its Variance Estimate. *arXiv preprint arXiv:1803.08000*.
- Goldfarb A, Xiao M (2011) Who thinks about the competition? Managerial ability and strategic entry in US local telephone markets. *AER*, 101(7), 3130-61.

- Gregorutti B, Michel B, Saint-Pierre P (2017) Correlation and variable importance in random forests. *Stat. Comput.*, 27(3), 659-678.
- Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3(Mar), 1157-1182.
- Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3), 389-422.
- Hanssens DM (1980) Bivariate time-series analysis of the relationship between advertising and sales. *Applied Economics*, 12(3), 329-339.
- Hanssens DM (1980) Market response, competitive behavior, and time series analysis. *J. Marketing Res.*, 470-485.
- Haury AC, Gestraud P, Vert JP (2011) The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS one*, 6(12), e28210.
- Hoch SJ, Dreze X, Purk ME (1994) EDLP, Hi-Lo, and margin arithmetic. *J. Marketing*, 16-27.
- Hooker G (August 2004) Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 575-580). ACM.
- Investor's Business Daily (2017)
<https://www.investors.com/news/technology/amazon-monopoly-problem-antitrust-action-vs-amazon-facebook-google/>
- Ishwaran H (2007) Variable importance in binary regression trees and forests. *Electron. J. Stat.*, 1: 519-537.
- Kopalle PK, Mela CF, Marsh L (1999) The dynamic effect of discounting on sales: Empirical analysis and normative pricing implications. *Marketing Sci.*, 18(3), 317-332.
- Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, de Schaetzen V, Duque R, Nersini H, Nowe A (2012) A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(4), 1106-1119.
- Lemmens A, Croux. C (2006) Bagging and boosting classification trees to predict

- churn. *J. Marketing Res.*, 43.2 (2006): 276-286.
- MacQueen J (June 1967) Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).
- Mankiw NG (1985) Small menu costs and large business cycles: A macroeconomic model of monopoly. *QJE*, 100(2), 529-537.
- Meinshausen N, Bühlmann P (2010) Stability selection. *J. R. Stat. Soc.: Series B (Stat. Methodol.)*, 72(4), 417-473.
- Mentch L, Hooker G (2016) Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.*, 17(1), 841-881.
- Miller PJ, Lubke GH, McArtor DB, Bergeman CS (2016) Finding structure in data using multivariate tree boosting. *Psychol. methods*, 21(4), 583.
- Muth JF (1961) Rational expectations and the theory of price movements. *Econometrica*, 315-335.
- Neslin SA, Gupta S, Kamakura W, Lu J, Mason CH (2006) Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *J. Marketing Res.*, 43, no. 2: 204-211.
- Nijs VR, Srinivasan S, Pauwels K (2007) Retail-price drivers and retailer profits. *Marketing Sci.*, 26(4), 473-487.
- Pesaran MH, Smith RP (1998) Structural analysis of cointegrating VARs. *J. Econ. Surveys*, 12(5), 471-505.
- Pollard D (1981) Strong consistency of k-means clustering. *Ann. Stat.*, 9(1), 135-140.
- ProPublica (2016) <https://www.propublica.org/article/amazon-says-it-puts-customers-first-but-its-pricing-algorithm-doesnt>
- Rafieian O, Yoganarasimhan H (2017) The Value of Information in Mobile Ad Targeting. *Working Paper*.
- Roy A, Hanssens DM, Raju JS (1994) Competitive pricing by a price leader. *Management Sci.*, 40(7), 809-823
- Segal MR (1992) Tree-structured methods for longitudinal data. *J. American Stat. Assoc.* 87.418: 407-418.

- Segal MR, Xiao Y (2011) Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.1: 80-87.
- Shankar V, Bolton RN (2004) An empirical analysis of determinants of retailer pricing strategy. *Marketing Sci.*, 23(1), 28-49.
- Slade ME (1992) Vancouver's gasoline-price wars: An empirical exercise in uncovering supgame strategies. *Rev. Econ. Stud.*, 59(2), pp.257-276.
- Splinter News (2016) <https://splinternews.com/how-to-not-get-screwed-on-amazon-1793862519>
- Srinivasan S, Pauwels K, Hanssens DM, Dekimpe MG (2004) Do promotions benefit manufacturers, retailers, or both? *Management Sci.*, 50.5: 617-629.
- Statista (2018) <https://www.statista.com/statistics/259782/third-party-seller-share-of-amazon-platform/>
- Steenkamp JBE, Nijs VR, Hanssens DM, Dekimpe MG (2005) Competitive reactions to advertising and promotion attacks. *Marketing Sci.*, 24(1), 35-54.
- Strobl C, Boulesteix A, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics* 8, no. 1: 25.
- Svetnik V, Liaw A, Tong C, Wang T (June 2004) Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. *In International Workshop on Multiple Classifier Systems* (pp. 334-343). Springer, Berlin, Heidelberg.
- Tan S, Caruana R, Hooker G, Lou Y (2018) Transparent Model Distillation. *arXiv preprint arXiv:1801.08640*.
- The Penny Hoarder (2017) <https://www.thepennyhoarder.com/smart-money/amazon-fake-sellers/>.
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B (Method.)*, 267-288.
- Trusov M, Bodapati AV, Bucklin RE (2010) Determining influential users in internet social networks. *J. Marketing Res.*, 47(4), 643-658.
- Varian HR (1980) A model of sales. *AER*, 70(4), 651-659.
- Wager S, Hastie T, Efron B (2014) Confidence intervals for random forests: The

- jackknife and the infinitesimal jackknife. *J. Mach. Learn. Res.*15(1), 1625-1651.
- Wall Street Journal (2017) <https://www.wsj.com/articles/watch-out-for-fang-inc-1497000604>
- Wood SN (2006) Generalized additive models: an introduction with R. Chapman and Hall/CRC.
- Xiao Y, Segal MR (2009) Identification of yeast transcriptional regulation networks using multivariate random forests. *PLoS computational biology* 5.6: e1000414.
- Yoganarasimhan H (2016) Search personalization using machine learning. *Working Paper*.
- Zaman F, Hirose H (December 2009) Effect of subsampling rate on subbagging and related ensembles of stable classifiers. In *International Conference on Pattern Recognition and Machine Intelligence* (pp. 44-49). Springer, Berlin, Heidelberg.
- Zhang H (1998) Classification trees for multiple binary responses. *J. American Stat. Assoc.*, 93.441: 180-193.

CHAPTER 4: APPENDIX

Appendix 4.1. K-means Clustering Results for 3P Sellers

I use the SAS software for the clustering algorithm. The initial list of variables input in the SAS variable clustering procedure to identify correlated variables is in Table A.4.1. The final clustering variables as described in Table 4.5 are used in the SAS FASTCLUS procedure with K-means algorithm to get 5 clusters.

Table A.4.1. Variables considered for K-means Clustering

Seller Characteristics	Description
Total No. of Ratings	Min, Max, Mean for the seller across the panel
% Positive Ratings	Min, Max, Mean for the seller across the panel
Seller Star Rating	Min, Max, Mean for the seller across the panel
Focal Brand Indicator	Value 1 if seller offered Instant Pot any time in the panel, 0 otherwise
Non-focal brand Indicators	Value 1 if seller offered respective Non-focal brand any time in the panel, 0 otherwise
FBA	Proportion of FBA offers made by the seller across the panel
Free ship Offers	Proportion of free ship offers made by the seller across the panel
No. of unique items sold	Count of unique items sold by a seller across the panel
Sales Presence	Number of days seller observed on the marketplace

Appendix 4.2. Variable Selection Algorithm and Results

I briefly describe the variable selection algorithm in our empirical application. I run two separate variable selection algorithms, one each for the mean structure based SI VIM (with F test) and the outcome difference based SI VIM. The variable selection algorithm is as follows.

1. Split data into training (n_{train}) and testing (n_{test}) sets.
2. On both the training and testing sets, introduce a uniform random noise pseudo-variable.
3. Bootstrap B subsamples from the training set.
4. At each iteration, run a MVT on each bootstrapped subsample $b = 1, 2, \dots, B$.
5. Compute the importance score of each covariate, including the random noise variable based on each MVT output on the test set.
6. Record the prediction of each MVT on the test set.
7. Average across the B MVTs to compute the MVRF prediction error (MSE) and the VIM score for each variable.
8. Remove covariates with VIM lower than that of the pseudo random noise variable.
9. Repeat steps 2-6 $niter$ times or until steady state of covariates is observed.

In this application, $n_{train}:n_{test} = 70:30$, $B = 5000$, $niter = 40$. The size of each bootstrapped subsample is approximately two-thirds that of the training set. At the end of 40 iterations, the mean SI VIM retains 246 covariates, while the outcome difference SI VIM retains 58 out of the original 892 covariates.

Figure A.4.1. MSE trend using mean SI VIM for price change magnitude

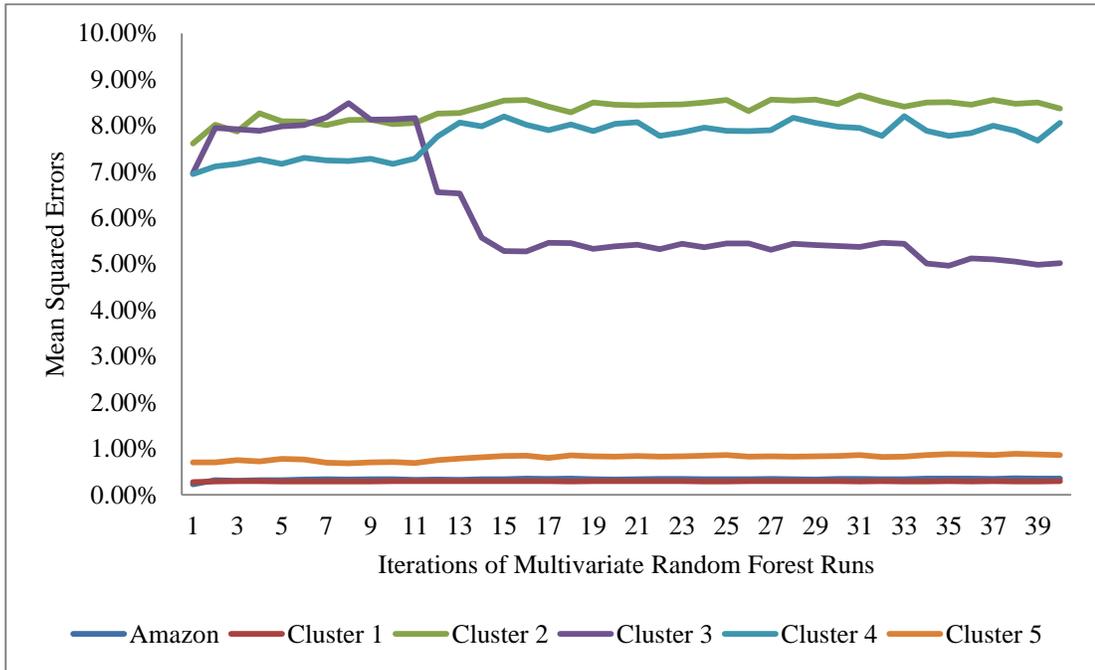


Figure A.4.2. MSE trend using outcome difference SI VIM for price change magnitude

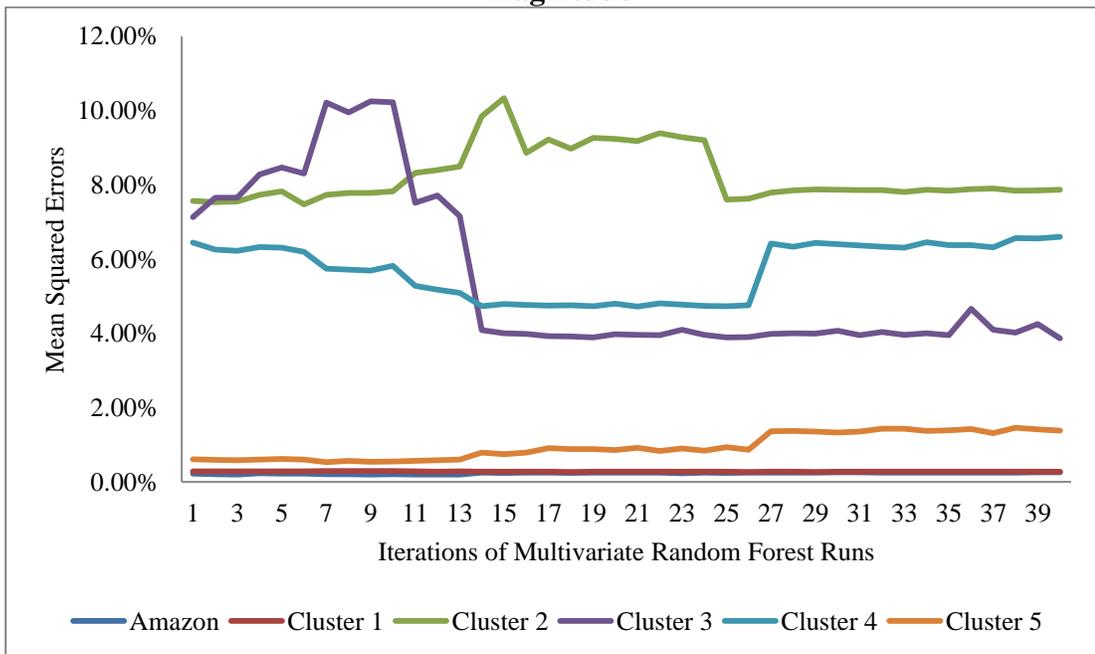


Figure A.4.3. MSE trend using mean SI VIM for price change with sign included



Figure A.4.4. MSE trend using outcome difference SI VIM for price change with sign included



Appendix 4.3. Estimation Results of Price change with sign included

Table A.4.2. Mean Squared Error Rate on Test Set for Price Change with sign included

	Amazon	3P Seller Clusters				
		1 Estd.	2 Small-scale	3 New Entrants	4 Multi-Brand	5 FBA Sellers
VAR	0.15%	0.29%	7.44%	4.29%	4.80%	0.28%
	(0.7%)	(1.0%)	(52.0%)	(19.0%)	(11.0%)	(1.0%)
LASSO	0.11%	0.29%	7.38%	4.24%	5.05%	0.26%
	(0.6%)	(1.0%)	(52.0%)	(19.0%)	(12.0%)	(1.0%)
<i>Multivariate Models with covariates selected post variable reduction</i>						
MVRF						
Mean Based SI	0.13%	0.30%	7.85%	4.61%	5.37%	0.31%
	(0.7%)	(1.0%)	(51.0%)	(18.0%)	(10.0%)	(1.0%)
Outcome Difference SI	0.12%	0.29%	7.53%	8.54%	5.52%	0.54%
	(0.7%)	(1.0%)	(53.0%)	(18.0%)	(12.0%)	(1.0%)
VAR-X						
Mean Based SI	0.23%	0.29%	7.52%	4.30%	4.15%	0.28%
	(0.7%)	(1.0%)	(53.0%)	(19.0%)	(11.0%)	(1.0%)
Outcome Difference SI	0.27%	0.31%	7.92%	4.38%	7.37%	0.46%
	(0.8%)	(1.0%)	(54.0%)	(19.0%)	(17.0%)	(1.0%)
GAM						
Mean Based SI	0.17%	0.29%	7.43%	4.24%	4.66%	0.27%
	(0.7%)	(1.0%)	(53.0%)	(19.0%)	(13.0%)	(1.0%)
Outcome Difference SI	0.16%	0.28%	7.45%	4.23%	4.59%	1.38%
	(0.8%)	(1.1%)	(53.0%)	(19.0%)	(14.0%)	(4.0%)

Standard deviation of squared error on test set reported in parentheses

As can be seen in Table A.4.2, for the model on price change with sign included the relative predictive power of the models is less certain. The GAM with variables selected by the mean based split improvement method is at par or beats in some cases both VAR and LASSO for the 3P seller clusters but is slightly worse off for Amazon. Based on the results, it appears that for the price change with sign

included there is no clear winner model.

The brands offered by each seller type vary in the tracking period (see Table 4.6). Therefore, the longitudinal data on price changes of a seller type at an individual brand is too sparse to be modeled. To work around this data limitation, I employed the aggregation strategy to derive the final outcome for a seller type. First, I select the minimum price seller of a brand for each 3P seller type or cluster (see equation (7)) as the representative brand price; note that for Amazon, this step is irrelevant. The identity of the minimum price 3P seller can change over time. The representative brand price for a cluster will thus reflect prices offered by different sellers over time. Second, from equation (2), I choose the maximum price change across all brands as the price change response of the seller type. This second step holds even for Amazon. It is likely that the maximum price change will occur on different brands across time. Given the two levels of aggregation considered, it is highly likely that the outcome variable of a given seller type will track price changes of varying brand-seller combinations over time. Therefore, as the brand-seller combination changes across the longitudinal panel, the underlying predictors associated with the changes will vary. This makes it harder to determine significant predictors for price change with sign included. This is seen across all the models in terms of comparable predictive performance for the outcome with sign of price change included. We note that despite the aggregations, we are more likely to identify the significant patterns in the data for the price change magnitude model. This can be explained by the fact that for a seller type, there is a degree of homogeneity in reactions and responses. Therefore, for a given seller type the magnitude of response and the factors driving the response may

be similar.

Appendix 4.4. Estimation Plots of the Significant Non-Linear (Smooth) Covariates

Figure A.4.5. Amazon

Past Period Mean Rating of Small-Scale sellers

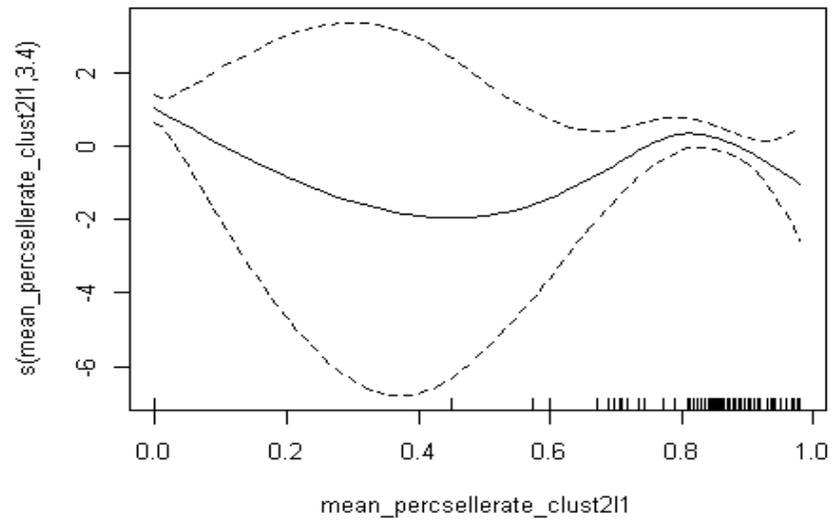
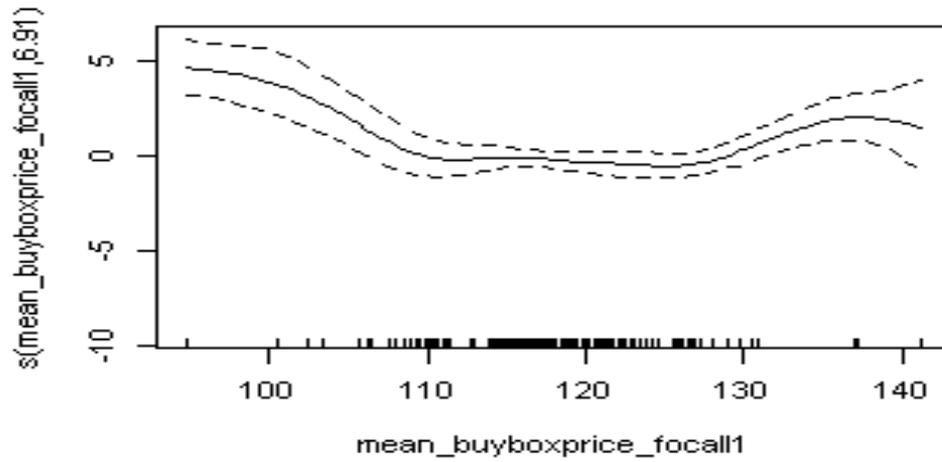


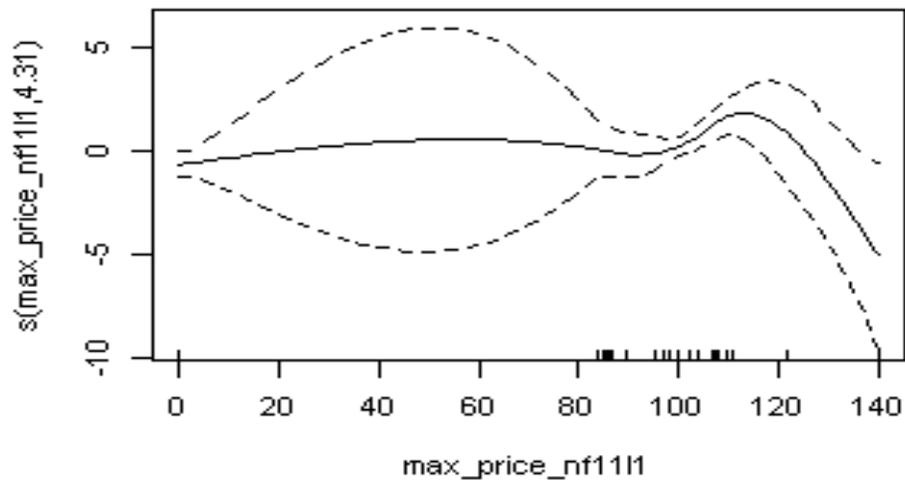
Figure A.4.6. Cluster 3 (“New Entrants”)

Past Period Mean Buy box Price of Instant Pot



(a)

Past Period Maximum Price of Cuisinart



(b)

CHAPTER 5

CONCLUSION

In this concluding chapter I will briefly discuss the unique challenges with the respective empirical settings that motivated me towards the methodological framework examined in the three essays. I will discuss some of the limitations of the methodologies used across all three essays and then specifically talk about the substantive shortcomings of the marketing applications (Chapters 2 and 4). Finally, I will discuss some of the application areas and extensions of my dissertation research.

5.1. Empirical Challenges, Methodological Motivation and Limitations

In the empirical settings for all three essays a critical challenge was data sparsity. The motivation to examine these as multivariate response models was partly driven by the sparse response outcomes. The idea was to borrow explanation strength from response variables that are less sparse but likely to be correlated with the sparser outcomes.

In essay 1 (Chapter 2), there were considerable data sparsity in the response outcomes of website visits and online and in-store purchase incidences. The average number of weeks that a website visit was made in the 27 week data period was only 2.73, while the average number of online and offline purchases were less than 1 (0.28 for online and 0.18 for offline). The sparsity of visits and purchase behavior is a typical phenomenon for most retailers specializing in high-end beauty care and apparel products. Given that the objective was to find the underlying latent engagement states of the customers with online and offline channels, the data sparsity led to issues with

model identification. In order to work around this, I treated the state-dependent website visit count process as similar to a zero-inflated Poisson, where zeroes are observed for low online engagement states. The state-dependent online and offline purchases were modeled as a bivariate logit process with homogeneous parameters for high engagement states on a channel. This implies that when a customer is in a high engagement state with say online channel, her actual observed online behavior does not depend on her offline channel engagement state. While these assumptions helped reduce the dimensionality of the overall model, these are unable to differentiate between customer's channel specific behaviors when she is engaged with only the focal channel versus when she is engaged with both channels. Therefore, this limits the ability to examine cross-channel latent engagement effects on focal channel behavior.

In the second and third essays (Chapters 3 and 4), the problem of data sparsity in response outcomes was compounded by the fact that the response vector was of higher order (>3). Further, both the empirical settings of e-bird (Chapter 3) and Amazon marketplace (Chapter 4) were high dimensional with possible non-linear and complex interactions among predictors and outcome variables. This motivated me to apply non-parametric methods, e.g., random forests, to model the multivariate outcome. Additionally, since multivariate random forests can model higher order response vectors more flexibly than parametric methods, this was computationally a more feasible alternative. To tackle the issue of high dimensionality, the MVRFs were applied as a variable selection tool to recursively remove features that were deemed unimportant using a predefined importance measure. In order to define a variable's importance in a multivariate response setting, I was motivated to look beyond what was already available in standard statistical packages (e.g. R package on *MultivariateRandomForest* uses frequency based importance). Some of these new

variable importance measures are proposed and examined in Chapter 3. The proposed measures build on a variable's ability to separate between children nodes when used as a node-splitting variable. Therefore, higher the magnitude of separation or split improvement at a node split on average, higher is the variable's importance in predicting the outcome. Further, I compared the predictive performance of the selected variables based on proposed and extant importance measures. In my application on e-bird data in Chapter 3, the proposed measure of mean structure based SI on significant splits (using F tests) and the outcome difference SI performed better than the extant ones for 3 out of the 5 bird species.

A limitation of the variable selection procedure introduced in Chapter 3 is in the stability of variable ranking. While the introduction of the random noise pseudo-variable provides a benchmark for feature elimination, the rank ordering of the retained features may not be stable over different training samples. In the simulation exercise, part of this is controlled for by building multiple random forests and taking the average measure across the forests. However we do not account for this in the actual empirical application of the variable selection procedure for either the e-bird data (Chapter 3) or the Amazon case study (Chapter 4). This limitation can be an issue especially when using the proposed variable selection procedure as a pre-processing step for multivariate regression analysis. In the Amazon application, the retained features were sequentially introduced into the GAM and VAR-X models based on the rank ordering produced by the proposed VIMs. If the rank ordering is unstable, the features introduced in the actual regression analysis may change. This will affect both the interpretation of the underlying relationship between predictors and outcome and the model's predictive performance.

Another methodological limitation pertaining to my essays is in the computation speed of the statistical software R. This is especially true for essay 1

where the computational time required in estimating the HSMM using hierarchical Bayesian methods was painfully long. As discussed in Chapter 2, the modeling framework is inherently complex due to the multivariate components in both state transition and state emissions models. The state transition block matrix is built on a bivariate state space and the state-dependent emissions model is a multivariate distribution of website visits and online and offline purchases. The estimation of this model using maximum likelihood estimation assuming homogeneous parameters was in itself fairly time consuming (> 2 days for the fully specified models when run on 10% of the training sample size 1000). The Bayesian estimation using R's *mcmc* package proved to be a very slow and inefficient method for such a complex model (100K iterations on 1000 sample size took nearly 4 months!! my dissertation essay is yet another documentation of R's slowness when dealing with *for loops*). The MVRF and other multivariate regression models examined in essay 3 (Chapter 4) were manageable with smaller data. However, the computational time is a limitation even for the MVRF models run on larger data set. For instance, the e-bird training data size is 14,073. However, to efficiently run the MVRF I had to bootstrap sub-samples of size 500 (~3.5% of training set size). The RFE strategy when implemented on bootstrap subsamples of size 500 to create an MVRF ensemble of 500 trees took more than 2 hours per iteration (I ran a total of 30 iterations). Based on this limitation, a possible alternative is to modify the code for other languages such as Python. I am yet to exploit these other programming languages and do a comparison of estimation time against R.

From a purely substantive standpoint I will now summarize some of the limitations specific to the marketing applications examined in essays 1 and 3. In the customer-multichannel behavior application in essay 1, an important factor that can influence the customer's latent state is external actions made by the firm. For instance,

a marketing intervention of email ad or in-mail discount coupon can influence a customer's channel engagement state change from "low" to "high". However in the data provided by the multichannel retailer, I do not observe any promotional activities made by the firm. This is an important drawback which may adversely affect both the predictive accuracy of state transition and the interpretation of the state transition parameters.

In the Amazon marketplace application examined in essay 3, the most severe limitation is the lack of generalizability of the substantive findings. This is partly rendered due to the uniqueness of the Amazon marketplace. This non-traditional marketplace is more volatile and susceptible to environmental changes such as introduction of new brands, entry and exit of 3P sellers and other policy changes made by Amazon itself. Therefore, while the results of the multivariate GAM provide interesting insights into the current marketplace competition, these may change with changes in the marketplace environment (which can be as frequent as 1-2 months). This implies that in order to effectively study the marketplace competition and the factors that drive price changes, one may need to retrain the model as frequently as every 2 months using a rolling window method.

Further, the results from one category may not hold for another. In our application, we have examined the category of programmable pressure cookers. The category has its own unique set of products/brands and 3P seller composition. This composition may widely vary across categories, especially in some categories the 3P sellers may be less reputable (or even fake) vendors operating on the marketplace. This is also true of the products and brands sold. We have selected a sample of national brands from the electric cooker category. In some of the categories such as electronics the Amazon marketplace is flooded with many lesser known brands and manufacturers. This poses a real challenge in generalizing the substantive findings

across categories.

5.2. Applications and Extensions

I will briefly discuss some of the empirical applications and the possible extensions of the methods developed in my dissertation. I will first focus on the applications for the HSMM framework examined in essay 1 and then talk about the applications of the MVRF variable selection method examined in essays 2 and 3.

In the HSMM framework developed in essay 1, I have examined latency in customer's multichannel engagement. This framework helps in inferring both the underlying latent state as well as the duration spent in a state. A possible marketing application of this research is in eye tracking behavior. In eye-tracking research, an important behavior to predict is the duration of a customer's visual attention on a brand or product. The duration of visual attention is directly correlated to the customer's consideration to purchase the product. A customer's visual attention is a latent response of the customer's sensitivity to the product and its memory recall. Therefore, one can potentially model customer's visual attention as a latent semi-Markovian process. Since eye-movements and visual attention can change in instantaneous time, the proposed HSMM framework can be extended under a continuous time setting.

Further, the HSMM framework can be extended and applied in research on customer's web browsing or online media consumption behavior. The motivation for customer's browsing behavior, e.g., browsing a website for education or knowledge gaining, or online media consumption, e.g. watching movies online on Netflix, is not directly known to the researcher. Further, a customer may browse through multiple pages of the website and the duration spent on each website may be directly related to the underlying motivation. For instance, a customer may search multiple webpages within Netflix, read movie reviews across multiple genres before deciding on a specific movie. The latent motivation for such online media consumption and web-browsing can be modeled using a HSMM. As noted in the earlier section, the data setting of essay 1 was very sparse, which motivated me to examine a variant of the zero-inflated Poisson to model the state-dependent visitation process. Under a richer

data setting, one can apply alternative processes, e.g. Hawkes process, to examine the customer browsing and consumption behavior.

Moving over to the non-parametric MVRF methods examined in essays 2 and 3, I will first discuss a marketing application and then suggest extensions of the proposed variable selection method. An overall comment for the proposed variable selection procedure is that this can be applied across multiple disciplines. The general paradigm for the application is multivariate response with sparsity and high dimensional data. More specifically in marketing, an application of this research is in predicting customer's online multicategory purchase decisions. For instance, while shopping online, customers are often exposed to multiple external stimuli, such as product recommendations, related items, website features, and focal search product features. Further when customers reach a purchase decision, they may buy from multiple categories. For marketplaces such as Amazon or eBay, where the product assortments cater to large number of categories (> 3), modeling a customer's multicategory purchase decision may be daunting under traditional parametric methods. In such cases, MVRFs can be applied to model customer's multicategory purchase decisions. Further the proposed variable selection procedure can be applied to extract the relevant external stimuli or features that predict the customer's multicategory purchase decision.

In the earlier section on limitations, I noted that the stability of variable rank ordering may be an issue with the proposed variable selection procedure. A possible improvement and extension is to quantify the uncertainty in variable ranking of the proposed importance measures. One approach is to do a post-variable selection check by examining the retained features pairwise. Here one can determine the difference in importance measures for a pair of features within a tree and then compute the variance of difference across the MVRF ensemble. A stable rank ordering would indicate a smaller variance in the importance measure difference for a pair of retained features. In addition to the prediction check at the end of the RFE strategy, this variance of difference in variable ranking can determine robustness of an importance measure in a given empirical context. A second extension is to employ the mean and variance of the

prediction at each iteration as part of the decision rule of feature elimination. In the proposed RFE algorithm, the feature elimination is conducted on the basis of comparison against a benchmark pseudo-variable. An enhancement to this strategy would be to in addition benchmark against the mean and variance in prediction of the prior iteration. If the mean prediction error improves, e.g., MSE reduces, and variance in prediction reduces, the algorithm can proceed to remove the features suggested by the pseudo-variable benchmark.

5.3. Final Remarks

I am a data scientist at heart and firmly believe that the merit of a method is in its applicability to solve real problems in the industry irrespective of the vertical, i.e., marketing, financial services, healthcare etc. While there are computational difficulties in applying some of the theoretical methods of my research in exactness, simpler variants can be readily used and applied. For instance, the HSMM framework using hierarchical Bayesian estimation may not be feasible for industry application due to computational cost. However assuming parameter homogeneity, one can apply maximum likelihood estimation and still infer latent states of customer-multichannel behavior with fairly high precision and modeling sophistication. Similarly, the variable selection procedure developed using MVRFs is amenable to computational simplifications e.g., lesser number of trees or smaller sub-sample size, and can be used for many industry problems with high dimensional and multivariate response data. I sincerely hope that with my dissertation research I have been able to make some impactful contribution in solving such real world phenomena.